2-11-2021 2:00 PM

Visual Analytics for Performing Complex Tasks with Electronic Health Records

Neda Rostamzadeh, *The University of Western Ontario*

Supervisor: Sedig, Kamran, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Computer Science

Abstract

Electronic health record systems (EHRs) facilitate the storage, retrieval, and sharing of patient health data; however, the availability of data does not directly translate to support for tasks that healthcare providers encounter every day. In recent years, healthcare providers employ a large volume of clinical data stored in EHRs to perform various complex data-intensive tasks. The overwhelming volume of clinical data stored in EHRs and a lack of support for the execution of EHR-driven tasks are, but a few problems healthcare providers face while working with EHR-based systems. Thus, there is a demand for computational systems that can facilitate the performance of complex tasks that involve the use and working with the vast amount of data stored in EHRs. Visual analytics (VA) offers great promise in handling such information overload challenges by integrating advanced analytics techniques with interactive visualizations. The user-controlled environment that VA systems provide allows healthcare providers to guide the analytics techniques on analyzing and managing EHR data through interactive visualizations.

The goal of this research is to demonstrate how VA systems can be designed systematically to support the performance of complex EHR-driven tasks. In light of this, we present an activity and task analysis framework to analyze EHR-driven tasks in the context of interactive visualization systems. We also conduct a systematic literature review of EHR-based VA systems and identify the primary dimensions of the VA design space to evaluate these systems and identify the gaps. Two novel EHR-based VA systems (SUNRISE and VERONICA) are then designed to bridge the gaps. SUNRISE incorporates frequent itemset mining, extreme gradient boosting, and interactive visualizations to allow users to interactively explore the relationships between laboratory test results and a disease outcome. The other proposed system, VERONICA, uses a representative set of supervised machine learning techniques to find the group of features with the strongest predictive power and make the analytic results accessible through an interactive visual interface. We demonstrate the usefulness of these systems through a usage scenario with acute kidney injury using large provincial healthcare databases from Ontario, Canada, stored at ICES.

Keywords

Visual Analytics; Analytics; Visualization; Interaction; Human-Data Interaction; Human-Computer Interaction; Electronic Health Records; Machine Learning; Frequent Itemset Mining; Extreme Gradient Boosting; Health-Related Tasks; Prediction Models; Data-Driven Healthcare; Precision Medicine

Summary for Lay Audience

Many medical organizations adopt electronic health record systems (EHRs) to replace traditional paper-based patient records as they modernize their operations. EHR data includes patients' medical history, medications, diagnoses, treatment plans, and laboratory test results. Healthcare professionals use EHR-based systems to perform various tasks that involve the use and working with a vast amount of data stored in EHRs. Such tasks include identifying patients at high risk of developing diseases, monitoring a patient's condition, and studying the effect of treatments, among others. Despite the benefits of EHR systems, they fail to meet the healthcare professional's computational needs. Therefore, it seems like there is a need for computational tools that can support the execution of various tasks on large bodies of data in EHRs. This research aims to prove the usefulness of computational tools, known as visual analytics, in performing different tasks on EHRs. VA combines the strength of data analytics techniques with interactive visualizations to allow healthcare professionals to explore and analyze the clinical data interactively. We first identify the gaps in support of tasks performed by EHR-based systems using a proposed framework. We then provide a comprehensive overview of EHR-based VA systems through a systematic literature review. We evaluate these systems based on the tasks, analytics, visualizations, and interactions they support and identify the areas with little prior work. We develop two novel VA systems (SUNRISE and VERONICA) to show how the VA approach can be used to address the challenges of EHRs. SUNRISE is designed to help healthcare professionals to identify relationships between laboratory test results and a disease. VERONICA uses several analytics techniques to find the best representative group of features in identifying high-risk patients. We show how these VA systems can be used to solve real-world problems using the healthcare datasets from Ontario, Canada, stored at ICES.

Co-Authorship Statement

Chapter 1 is my original work in introducing the dissertation and explaining the connections between chapters. **Chapters 2 and 3**, which focus on systematic reviews, proposed framework, and design space, were a collaborative effort with my supervisor, Kamran Sedig, and a graduate student in our research group, Sheikh Abdullah.

Chapters 4, 5 were a collaborative effort with my supervisor, Kamran Sedig, and three colleagues, Sheikh Abdullah, Amit Garg, and Eric McArthur. It should be noted that all the systems presented in this dissertation use healthcare databases stored at ICES. It is a mandatory requirement that all publications that incorporate ICES data must include responsible ICES scientists as co-authors. Amit is the program lead of the Kidney, Dialysis & Transplantation Program at ICES. Eric is a Local lead analyst at ICES Western. Amit and Eric were responsible for providing us access to the data and making sure the published results comply with the privacy guidelines at ICES. **Chapter 6** is my original work, to summarize the chapters and outline limitations and future research areas.

In developing the framework and the design space introduced in Chapters 2 and 3, I was responsible for conducting the literature reviews, summarizing the systems, and writing the original draft. My supervisor and Sheikh helped me in integrating the concepts into a unified conceptual foundation by creating common terminology, upon which the framework and the design space were built. I was also primarily responsible for the design, implementation, and writing of the original draft for visual analytics systems presented in Chapters 4 and 5. My supervisor and Sheikh helped me with the conceptualization and revision of the papers.

Acknowledgments

First and foremost, I would like to thank my supervisor, Kamran Sedig. I am eternally grateful for your dedication, patience, and mentorship. No one can hope for a better supervisor.

I would like to thank Dr. Amit Garg for all his guidance and support. I would also like to thank the members of Insight lab, especially Sheikh Abdullah. Thank you for helping me throughout my Ph.D. journey. My appreciation goes out to Janice Wiersma and Sean Leonard for answering my questions and being a constant support. I would like to express my gratitude to Eric McArthur, Rey Acedillo and Flory Tsobo Muanda, whose suggestions, reviews, and comments have been invaluable throughout this project.

A special note of thanks goes to my husband, Ramtin. Thank you for your love and patience. You are my best friend and my greatest supporter.

Finally, I offer my sincerest thanks to my parents and my sisters, without whom I would not be where I am today. Thank you for your constant support over many years of my education. Thanks for loving and always believing in me.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Introduction

1.1 Motivation

The increasing adoption of electronic health records (EHRs) by medical organizations has created an untapped resource with the power to improve and change healthcare (Caban & Gotz, 2015; Harerimana et al., 2019; Yazhini & Loganathan, 2019). The EHR data includes patients' medical history and diagnoses, laboratory test results, medications, procedures, immunization history, allergies, treatment plans, and symptoms, among others (Lee et al., 2017). EHR systems have made patient information easily accessible to healthcare facilities for various basic healthcare operations. While initially developed for archiving patient records and supporting administrative tasks, researchers have recognized the inevitable application of EHRs to clinical research (Shickel et al., 2018). EHR-based systems offer an opportunity to utilize information derived from real-world patient data to better guide clinical decisions regarding patients (Murphy et al., 1999; Reisman, 2017). Healthcare providers utilize these systems to perform various EHR-driven tasks such as monitoring patient progression (Doupi, 2012), studying the effectiveness of treatments and medications (Cowie et al., 2017; Feng et al., 2019), detecting adverse clinical events (Medicine & America, 2000), and ultimately improving quality of care (Ali et al., 2007; Christensen & Grimsmo, 2008; Tang & McDonald, 2006). Despite the many hopes that access to more information through EHR-based systems would lead to better decisions, access to huge volumes of clinical data has made some analytical processes more challenging (Amarasingham et al., 2014). This problem is referred to as information overload and is very common in the healthcare domain. Information overload arises when trying to analyze a large number of variables that exceed human cognition's limits (Halford et al., 2005). It results in individuals misinterpreting, ignoring, or overlooking important information. In healthcare settings, information overload can lead to erroneous diagnoses, incorrect treatment decisions, and wrong interpretation of the clinical data (Caban & Gotz, 2015). Thus, there is a need for data analytics techniques to keep pace with the large volumes of complex EHR data.

Data analytics techniques incorporate methods and algorithms from various fields, such as statistics, machine learning, and data mining, to facilitate clinical decision making and to examine condition-specific clinical process outcomes (Han et al., 2011). In the past years, several data analytics techniques have been developed to support complex analytical tasks such as identifying patient cohorts, risk prediction, and biomarker discovery, studying the effect of treatments, and detecting adverse drug events (W. Sun et al., 2017; Yadav et al., 2018). However, while data analytics techniques are capable of processing a huge amount of data, they are not equipped to handle noisy and heterogeneous EHR data efficiently. They are also not capable of managing ill-defined clinical tasks that require human judgment (D. A. Keim et al., 2010). Since these techniques mostly hide the intermediary steps in the analysis process, healthcare providers can only be minimally involved in the process. Complementing data analytics, interactive visualizations display the clinical data in a visual form, enable healthcare providers to control the flow of the data, and let them customize representations to fulfill their cognitive needs. Several interactive visualization tools have been developed to explore and query EHRs (Rind et al., 2013). While beneficial, these tools fell short when confronted with problems requiring computational analysis, such as identifying patients at risk of certain diseases or detecting nephrotoxic medications (D. A. Keim et al., 2010). Thus, there is an increasing demand for an approach that integrates interactive visualizations with data analytics techniques to address the cognitive and computational needs of healthcare providers.

Visual analytics systems (VA) have the potential to transform raw EHR data into actionable insights by combining the strengths of data analytics and interactive visualizations (Caban & Gotz, 2015; D. Keim, Andrienko, et al., 2008; Ola & Sedig, 2014). VA systems support the execution and performance of a wide variety of complex EHR-driven tasks. For instance, VA could help researchers perform population-based analysis and gain insights from the huge amount of clinical data. Patients could use VA to understand personalized wellness plans and compare their health status and measurements against similar patients. Healthcare administrators could be supported in understanding the productivity of an organization, outcomes measurements, gaps in care, and patient satisfaction. Physicians can use VA to explore patients' trajectories and determine how a group of patients with chronic diseases can develop other comorbidities over time (Goldsmith et al., 2010; Perer & Sun, 2012; Rajwan et al., 2013). When using VA systems, tasks can be distributed between the healthcare provider and the system. In other words, the healthcare provider and the system work together to accomplish the task (Didandeh & Sedig, 2016; Sedig et al., 2012). For instance, a physician responsible for identifying the effect of a treatment on a patient might choose to delegate the computational sub-task of finding similar patients to the VA system. From observing the analytics results in the visualization, the physician can examine how similar patients responded to the treatment. In this scenario, both the VA system and the physician collaborate to determine the effect of a treatment on a patient. As EHR-driven tasks are usually ill-defined and domain-knowledge intensive, this user-guided discourse is very beneficial (Kamal, 2014; D. Keim et al., 2010). Although VA systems have shown great promise in supporting various EHR-driven tasks, to date, health lags behind other fields in the design and development of VA systems. The design of VA systems for EHRs is a non-trivial endeavor that requires a deeper understanding of individual components of VA systems and how to best develop and integrate them. There are several decisions that need to be made by the designer. For instance, the designer needs to determine how to organize and encode data items in the visualization, support and facilitate healthcare providers' tasks, and decide which data analytics technique to choose. There is currently a lack of direction on the effective and systematic design of VA systems for EHRs (Amarasingham et al., 2014; Shah, 2014; Silow-Carroll et al., 2012).

This research aims to show how VA systems can be designed and developed systematically for EHRs. We first conduct a systematic literature review of all the EHR-based interactive visualization tools that support clinical decision-making. We then present a framework to analyze EHR-driven tasks and activities in the context of interactive visualization tools. The framework helps us to evaluate the existing tools and identify the gaps in support of activities performed by these tools. We also conduct a comprehensive review of existing EHR-based visual analytics systems. These systems are evaluated based on four key dimensions: visual analytics tasks, analytics, visualizations, and interactions. We then identify which challenges remain insufficiently addressed. In light of this, we design and develop two novel EHR-based VA systems-namely, SUNRISE and VERONICA. These VA are designed for epidemiologists and clinical researchers at ICES-KDT. ICES is a not-for-profit, independent, world-leading research corporation that uses population-based health data and clinical and administrative databases to produce reliable and generalizable knowledge on a wide range of healthcare problems. KDT refers to the provincial Kidney Dialysis and Transplantation research program located in London, Ontario, Canada. We demonstrate the utility of these systems through usage scenarios with acute kidney injury (AKI) by investigating the process of exploring large provincial healthcare databases from Ontario, Canada, stored at ICES. SUNRISE and VERONICA allow healthcare providers at ICES to gain novel and actionable insights into the data and accomplish various EHR-driven tasks. These tasks include investigating the risk prediction result, tracking the decision path leading to the prediction, identifying the best representative features in predicting patients at high risk, examining the relationship between laboratory test results and a disease outcome, and conducting what-if analysis by testing hypothetical scenarios on patients.

One of the primary contributions of this research is developing an activity and task analysis framework that can help researchers and designers conceptualize functionalities of EHR-based interactive visualization tools in an organized manner. This research can help with the evaluation and systematic design of EHR-based interactive visualization tools using this framework. Furthermore, this research offers a comprehensive review and characterization of the state-of-the-art in VA for EHRs. It also identifies the key dimensions of EHR-based VA design space that unify prior work through the collection and analysis of the literature on EHR-based VA systems. Moreover, this research identifies the gaps and challenges of VA’s use in EHRs that remain insufficiently addressed. When developing EHR-based VA systems, there are several challenges that a designer might face. These challenges include providing busy clinicians with timely information in the proper format, identifying and visualizing cause-and-effect relationships, scaling VA systems to billions of patient records, and validating and refining the VA systems, among others. This research addresses these challenges by combining machine learning techniques, data mining algorithms, visualization, and human-data interaction. This research shows how VA systems can be designed systematically. It provides a comprehensive description of different components of VA systems in an organized manner and explains how these components work together to support the execution of complex data-driven tasks. This research then discusses the design decisions that need to be considered while developing an optimized and efficient EHR-based VA system. Finally, through the development of two unique VA systems, this research provides the healthcare domain with evidence of VA's efficiency for exploring EHRs.

1.2 Structure of Dissertation

The rest of this dissertation is broken into five chapters, as follows:

Chapter 2 provides a framework for identifying and analyzing EHR-driven tasks and activities in the context of interactive visualization tools—that is, all the activities, sub-activities, tasks, and sub-tasks that are and can be supported by EHR-based tools. To do so, we conduct a systematic literature review to collect all the research papers that describe the design, implementation, and evaluation of interactive visualization tools that support exploring and querying of EHR data. We provide an overview of each tool's overall purpose, describe its visualization, and analyze how different sub-activities, tasks, and sub-tasks combine to achieve the tool’s main higher-level activities of predicting, monitoring, and interpreting. The frameworks can be used to evaluate the existing EHR-based tools and design new tools systematically. It identifies the gaps in support of some higher-level activities that are supported by these tools.

In Chapter 3, we conduct a systematic literature review to gather articles describing the design and implementation of EHR-based VA systems and provide a comprehensive overview of these systems. This review also proposes a design space, including four primary dimensions used to characterize and evaluate the state-of-the-art EHR-based VA systems. These key dimensions include VA tasks, analytics, visualizations, and interactions. This review illustrates the major application of analytics, visualizations, and interactions in supporting the EHR-driven VA tasks. We also connect and unify the existing work using the dimensions identified in the design space with this review. Finally, we discuss the remaining challenges, areas of little prior work, and identify promising future research directions.

Chapter 4 presents a novel proof of concept VA system called SUNRISE that utilizes laboratory test result data to develop disease prediction models. SUNRISE allows healthcare providers to interactively explore the associations between laboratory test results and a disease outcome. It integrates frequent itemset mining with extreme gradient boosting (XGBoost) to create specialized prediction models. SUNRISE also includes interactive visualizations to enable the user to interact with the prediction model, track the decision process, and conduct what-if analysis by generating hypothetical input and observing how the model responds. It improves the user's confidence in the generated predictions by illustrating models' underlying working mechanisms through visualization representations. We demonstrate SUNRISE's utility through a usage scenario of exploring the relationships between laboratory test results and acute kidney injury using large provincial healthcare databases from Ontario, Canada stored at ICES.

Chapter 5 presents another novel VA system, called VERONICA, that takes advantage of the natural group structure of features in EHRs to find the group of features with the strongest predictive power. VERONICA incorporates several machine learning techniques —namely, classification and random forest, regression and classification tree, C5.0, support vector machines, and naive Bayes to allow the analysis of EHRs from different perspectives. It then enables the user to compare the risk prediction models in a systematic way through an interactive visual interface by integrating different sampling strategies, analytics algorithms, visualization techniques, and human-data interaction. To demonstrate the usefulness and utility of this VA system, we use the clinical dataset stored at ICES to identify the best representative feature groups in detecting patients who are at high risk of developing acute kidney injury.

In **Chapter 6**, we summarize the conclusions drawn from the research reported in the preceding chapters, explain the contributions of this research to the wider scientific community, and discuss some future research areas.

It should be noted that the chapters of this dissertation are self-sufficient and can be read individually or sequentially. Chapter 2 has been published; Chapters 3,4, and 5 have been submitted. This dissertation is written in an integrated article format, so Chapters 2 through 5 are self-contained.

Chapter 2

2 Data-Driven Activities Involving Electronic Health Records: An Activity and Task Analysis Framework for Interactive Visualization Tools

This chapter has been published as N. Rostamzadeh, S.S. Abdullah, and K. Sedig, "Data Driven Activities Involving Electronic Health Records: An Activity and Task Analysis Framework for Interactive Visualization Tools" in the Multimodal Technologies Interact. Journal, 4(1), 7; February 2020.

Please note that the format has been changed to match the format of the dissertation. Figure, Section, and Table numbers mentioned herein are relative to the chapter number. For instance, "Figure 1" corresponds to Figure 2-1. Additionally, when the term "paper", "work", or "research" is used, it refers to this particular chapter.

2.1 Introduction

An electronic health record (EHR) contains patient data, such as demographics, prescriptions, medical history, diagnosis, surgical notes, and discharge summaries. Healthcare providers use EHRs to make critical decisions, study the effects of treatments, determine the effectiveness of treatments, and monitor patient improvement after a particular treatment. In addition to these benefits, EHRs can potentially aid clinical researchers in detecting hidden trends and missing events, revealing unexpected sequences, reducing the incidence of medical errors, and establishing quality control (Christensen & Grimsmo, 2008; Tang & McDonald, 2006). Recently, several healthcare organizations have used systems that incorporate EHR data to improve the quality of care; these systems are intended to replace traditional paper-based medical records (Boonstra et al., 2014). However, a few studies (Himmelstein et al., 2010; Stead & Lin, 2009) reveal that these EHR-based systems hardly improve the quality of care. One of the reasons for this is that they do not allow for human–data interaction in a manner that fits and supports the needs of healthcare providers (Himmelstein et al., 2010; Rind et al., 2013). A set of technologies and techniques that can improve the efficacy and utility of these EHR-based systems can be found in information visualization (Rind et al., 2013), or broadly speaking interactive visualization tools (IVTs).

IVTs can be defined as computational technologies that use visual representations (i.e., visualizations) to amplify human cognition when working with data (Sears & Jacko, 2007; Sedig & Parsons, 2016). IVTs can help people who use them gain better insight by providing the means to explore the data at various levels of granularity and abstraction. An important feature of IVTs that makes them suitable for the exploration of EHRs is the ability to show relevant data quickly by mapping it to visualizations (Rind et al., 2013). Another feature is interaction. Making the visualization interactive allows healthcare providers to perform various data-driven tasks and activities. Interaction helps users accomplish their overall goals by dynamically changing the mapping, view, and scope of EHR data. In recent years, a number of EHR-based IVTs have been developed and deployed to support healthcare providers in performing data-driven activities.

To provide a clear and systematic approach in examining EHR-based IVTs for clinical decision support, this paper provides a framework for analyzing tasks and activities supported by these tools. To do so, we will first provide a brief survey of some of the existing IVTs that support the exploration and querying of EHR data and examine overall patterns in these tools. This survey does not include EHR-based IVTs that are designed for clinical documentation, administration, and billing processes.

There are a few studies that review EHR-based IVTs and their applications. Rind et al. (Rind et al., 2013) reviewed and compared state-of-the-art information visualization tools that involve EHR data using four criteria: (1) data types that they cover, (2) support for multiple variables, (3) support for one versus multiple patient records, and (4) support for user intents. Lesselroth and Pieczkiewicz (Lesselroth & Pieczkiewicz, 2011) surveyed different visualization techniques for EHRs. They cover a large number of visualization tools (e.g., Lifelines, MIVA, WBIVS, and VISITORS). Their survey is organized into five sections: (1) multimedia, (2) smart dashboards to improve situational awareness, (3) longitudinal and problem-oriented views to tell clinical narratives, (4) iconography and context links to support just-in-time information, and (5) probability analysis and decision heuristics to support decision analysis and bias identification. Combi et al. (Combi et al., 2010) reviewed a few visualization tools (e.g., IPBC, KHOSPAD, KNAVE II, Paint Strips, and VISITORS) and described them based on the following features: subject cardinality (single/multiple patients), concept cardinality (single/multiple variables), abstraction level (raw data, abstract concepts, knowledge), and temporal granularity (single, single but variable, multiple). Finally, in a book chapter, Aigner et al. (Aigner et al., 2008) described strategies to visualize (1) clinical guidelines seen as plans (e.g., GEM Cutter, DELT/A), (2) patients' data seen as multidimensional information space (e.g., Midgaard, VIE-VISU, Gravi++), and (3) patients' data related to clinical guidelines (e.g., Tallis Tester, CareVis).

A careful examination of the above surveys shows that a systematic analysis of IVTs with a focus on how they support EHR-data-driven tasks and activities is lacking. The purpose of the current paper is to fill this gap. Here, we present a framework for analyzing how IVTs can support different EHR-based tasks and activities. The framework can help designers and researchers to conceptualize the functionalities of EHR-based IVTs in an organized manner. In addition, this paper is suggestive of how this framework can be used to evaluate existing EHR-based IVTs and design new ones systematically. This paper also leads to the development of best practices for designing similar frameworks in similar areas.

The rest of this paper is organized as follows. Section 2 discusses how the proposed framework is formed and examines the relationships among the three concepts of activities, tasks, and low-level interactions in the context of the framework. Section 3 presents our strategy for searching relevant literature and explains our selection criteria. Section 4 provides a brief survey of a set of IVTs and outlines their main goal(s). In this section, using the proposed analytical framework, we identify the tasks and activities that IVTs support. Finally, Section 5 discusses how the framework can be used to evaluate the surveyed EHR-based IVTs.

2.2 A Proposed Activity and Task Analysis Framework

In the context of IVTs, user-tool interaction can be conceptualized as actions that are performed by users and consequent reactions that occur via the tool’s interface. This bi-directional relationship between the user and the tool supports the flow of information between the two. Interaction allows for human–information discourse (Ola & Sedig, 2018). Furthermore, it allows users to adjust different features of the IVT to suit their analytical needs. Interaction can be characterized at different levels of granularity (Sedig & Parsons, 2013, 2016). As displayed in Figure 1, an activity can be conceptualized at the highest level, where it is composed of multiple lower-level tasks (e.g., ranking, categorizing, and identifying) that work together to accomplish the activity's overall goal. An activity and a task can consist of multiple sub-activities and sub-tasks, respectively. At the lower level, tasks can be considered to have visual and interactive aspects; tasks that are supported by visual processing are called visual tasks. For instance, consider a scenario in which a user is working with a stacked bar chart that aggregates laboratory test results. The user needs to understand the distribution of a specific test of a collection of patients after surgery over time. Some of the visual tasks that the user may need to perform can include detecting the time when the test is at its peak and observing the average test result at different times. Interactive tasks require users to act upon visualizations. For instance, in the example above, the user may want to cluster the test results based on different time granularities (e.g., over an hour, over a day, or over a month). Each interactive task is made up of a number of lower-level actions (i.e., interactions) that are carried out to complete the task.

In most complex situations, activities, sub-activities, tasks, and sub-tasks are combined to support users in accomplishing their overall goal. It is important to note two perspectives from which we can view human–data discourse. From a top-down perspective, users’ goals flow from higher-level activities that need to be accomplished. From here, we go down to a number of tasks and sub-tasks (visual and interactive), and then to a set of low-level interactions. From a bottom-up perspective, the performance of a series of low-level interactions that users perform with visual representations gives emergence to tasks.

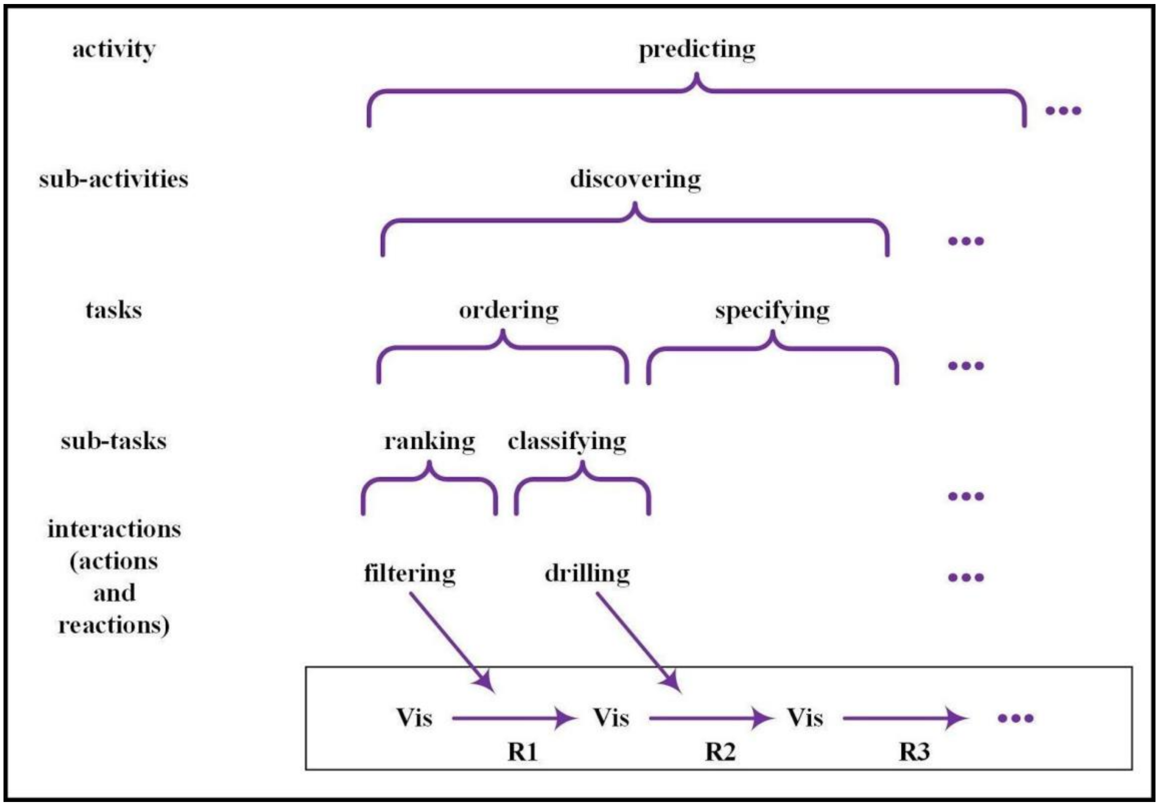


Figure 2-1: Relationships among activities, tasks, and interactions. Top-down view: activity is made up of sub-activities, tasks, sub-tasks, and interactions. Bottom-up view: activity emerges over time, through performance of tasks and interactions. Visualizations are depicted as Vis and reactions as R_x. Source: adapted from (Sedig & Parsons, 2016).

Similarly, the performance of a sequence of tasks gives emergence to activities all the way up until an overall goal is accomplished.

In this paper, we present an activity and task analysis framework for examining EHR-based IVTs (i.e., ones that involve EHRs as their main source of data with which users perform data-driven tasks and activities). To identify what activities, sub-activities, tasks, and sub-tasks are supported in EHR-based ITVs, we have examined a number of such tools that have been developed by different researchers and have been reported in the literature (see Wang et al. (Taowei David Wang et al., 2008); Wongsuphasawat et al. (Krist Wongsuphasawat et al., 2011); Wongsuphasawat and Gotz (K. Wongsuphasawat & Gotz, 2012); Malik et al. (Malik et al., 2014); Fails (Fails et al., 2006); Klimov et al.

(Klimov et al., 2010a); Wongsuphasawat (Krist Wongsuphasawat, 2009); Monroe et al. (Monroe et al., 2013); Brodbeck et al. (Brodbeck et al., 2005); Chittaro et al. (Chittaro et al., 2003); Rind et al. (Rind, Aigner, Miksch, Wiltner, Pohl, Drexler, et al., 2011); Plaisant et al. (Plaisant et al., 1998); Faiola and Newlon (Faiola & Newlon, 2011); Pieczkiewicz et al. (Pieczkiewicz et al., 2007); Bade et al. (Bade et al., 2004); Hinum et al. (Hinum et al., 2005); Rind et al. (Rind, Aigner, Miksch, Wiltner, Pohl, Turic, et al., 2011); and Ordonez et al. (P. Ordonez et al., 2012); Gresh et al. (Gresh et al., 2002); Horn et al. (Horn et al., 2001)). To conceptualize and develop the elements of the framework, our focus is the identification of activities and tasks that are independent of any specific technology or platform. To be consistent, we re-interpret how activities and tasks are named by the authors of the afore-listed sources in light of the unified language of our proposed framework. The activity and task terms we use might differ from the language of the existing literature since the authors have described their tools using their own vocabulary. Unfortunately, the language that different authors use is not consistent. Such inconsistency makes it difficult to analyze how well and comprehensively such tools support EHR-based tasks and how they can be improved. In the next section, we define and categorize the higher-level activities that result from interaction and combination of different sub-activities, tasks, and sub-tasks.

2.2.1 Higher-Level Activities: Interpreting, Predicting, and Monitoring

After reviewing numerous papers, we have concluded that, broadly speaking, all EHR-data-driven healthcare activities can be organized under three main categories: interpreting (Auffray et al., 2016; Groves et al., 2003; Komaroff, 1979; M. Kumar et al., 2007; Låg et al., 2014), predicting (Amarasingham et al., 2014; Cohen et al., 2014; Kankanhalli et al., 2016; Raghupathi & Raghupathi, 2014; Allan F. Simpao et al., 2014; Y. Wang et al., 2018), and monitoring (Anderson et al., 2015; Hauskrecht et al., 2013; Kho et al., 2007; Li & Wang, 2016; Saeed et al., 2002; Tia Gao et al., 2005). Interpreting refers to the activity of detecting patterns from patients' medical records and making sense of the relationships among different features. Predicting refers to the activity of anticipating patient outcomes and creating new hypotheses by analyzing patient history and status (Siegel, 2013). Lastly, monitoring refers to the activity of repetitive testing with the aim of adjusting and guiding the management of recurrent or chronic diseases (Glasziou et al., 2005).

2.2.2 Hierarchical Structure of Activities, Sub-Activities, Tasks, and Sub-Tasks

In this section, we identify sub-activities, tasks, and sub-tasks that blend and combine together to give rise to the three activities of ***interpreting***, ***predicting***, and ***monitoring***. ***Interpreting***, as a higher-level activity, can be comprised of four sub-activities: (i) *understanding* (e.g., gaining insight into patient medical records), (ii) *discovering* (e.g., finding patients with interesting medical event patterns), (iii) *exploring* (e.g., observing patient data in different temporal granularities), and (iv) *overviewing* (e.g., providing compact visual summaries of all event sequences found in the data). Likewise, ***predicting*** can be comprised of two sub-activities: (i) *learning* (e.g., generating new hypotheses from the data), and (ii) *discovering* (e.g., recognizing the deterioration of the disease). Finally, ***monitoring*** is composed of (i) *investigating* (e.g., examining the development of a patient after treatment), (ii) *analyzing* (e.g., studying the aggregated event sequences for quality assurance), and (iii) *evaluating* (e.g., assessing the quality of care based on clinical parameters). At the next level of the hierarchy, as shown in Figure 2, each sub-activity can be composed of a number of visual (e.g., *specifying*, *recognizing*, and *detecting*) as well as interactive tasks (e.g., *locating*, *ordering*, *querying*, and *clustering*). Moreover, as shown in Table 1, each task consists of different sub-tasks; for instance, *ordering* can be carried out by a combination of sub-tasks such as *ranking*, *aggregating*, *identifying*, and *classifying*.

2.3 Methods

2.3.1 Search Strategy

We conducted an electronic literature search in order to collect the research papers that describe the design, implementation, or evaluation of EHR-based IVTs. In order to assure a comprehensive document search, we included all the keywords that are relevant to the goal of the research and also covered all the synonyms and related terms, both for EHRs

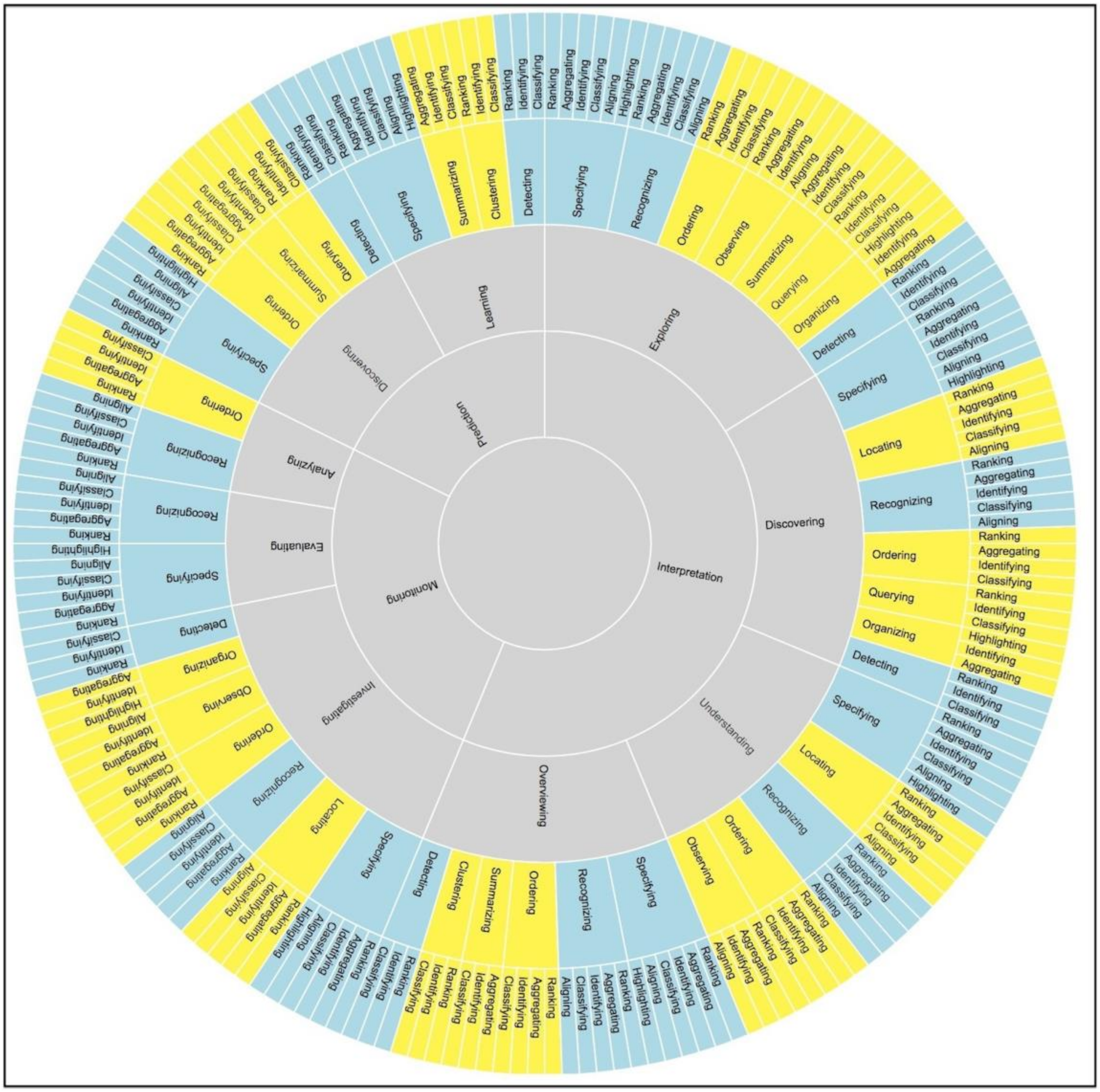


Figure 2-2: Overview of the proposed activity and task analysis framework. The visual tasks are represented as blue and interactive tasks are represented as yellow.

and visualization tools. We further broadened our search by adding an * to the end of a term to make sure the search engines picked out different variations of the term. We also added quotation marks around phrases to ensure that the exact sequence of words is found. To ensure that relevant papers were not missed in our search, we used a relatively large set of keywords. We used two categories of keywords. The first category concerned visualization tools and included the following terms: “visualization*”, “visualization tool*”, “information visualization*”, “interactive visualization*”, “interactive

Table 2-1: Shows the breakdown of the interactive and visual tasks.

	Task	Sub-tasks
Interactive	Ordering	Aggregating, Classifying, Identifying, Ranking
	Locating	Aggregating, Aligning, Classifying, Identifying, Ranking
	Querying	Classifying, Identifying, Ranking,
	Organizing	Aggregating, Classifying, Identifying, Highlighting
	Summarizing	Aggregating, Classifying, Identifying
	Clustering	Classifying, Identifying, Ranking
	Observing	Aggregating, Aligning, Identifying, Ranking
Visual	Recognizing	Aggregating, Aligning, Classifying, Identifying, Ranking
	Specifying	Aggregating, Aligning, Classifying, Identifying, Highlighting, Ranking
	Detecting	Classifying, Identifying, Ranking

visualization tool*”, “visualization system*”, and “information visualization system*”. For the second category, EHR, we used the following terms: “Health Record*”, “Electronic Health Record*”, “EHR*”, “Electronic Patient Record*”, “Electronic Medical Record*”, “Patients Record*”, and “Patient Record*”. As we were looking for papers about EHR-based visualization tools, we used the keywords shown in Table 2. We used the following search engines based on their relevance to the field: PubMed, the ACM Digital Library, the IEEE Library, and Google Scholar. We also looked for relevant papers in two medical informatics journals (International Journal of Medical Informatics and Journal of the American Medical Informatics Association). Furthermore, additional papers were collected in conference proceedings (e.g., IEEE Conference on Visual Analytics Science and Technology (VAST), HCIL Workshop 2015, and IEEE VisWeek Workshop on Visual Analytics in Health Care) that were published in 2007 and later. We then manually reviewed the reference lists of the papers that met the selection criteria to find other relevant studies that had not been identified in the database search. All the studies included in this survey were published from 1998 until 2015. We reviewed all of the abstracts, removed the duplicates, and shortlisted abstracts for a more detailed assessment.

Table 2-2: Overview of the search terms used.

Terms Used
“Visualization*” +“Health Record*”
“Visualization*” + “Electronic Health Record*”
“Visualization*” + “EHR*”
“Visualization*” + “Electronic Patient Record*”
“Visualization*” + “Electronic Medical Record*”
“Visualization*” + “Patients Record*”
“Visualization*” + “Patient Record*”
“Visualization tool*” +“Health Record*”
“Visualization tool*” + “Electronic Health Record*”
“Visualization tool*” + “EHR*”
“Visualization tool*” + “Electronic Patient Record*”
“Visualization tool*” + “Electronic Medical Record*”
“Visualization tool*” + “Patients Record*”
“Visualization tool*” + “Patient Record*”
“Information visualization*” +“Health Record*”
“Information visualization*” + “Electronic Health Record*”
“Information visualization*” + “EHR*”
“Information visualization*” + “Electronic Patient Record*”
“Information visualization*” + “Electronic Medical Record*”
“Information visualization*” + “Patients Record*”
“Information visualization*” + “Patient Record*”
“Interactive visualization*” +“Health Record*”
“Interactive visualization*” + “Electronic Health Record*”
“Interactive visualization*” + “EHR*”
“Interactive visualization*” + “Electronic Patient Record*”
“Interactive visualization*” + “Electronic Medical Record*”
“Interactive visualization*” + “Patients Record*”

“Interactive visualization*” + “Patient Record*”
“Interactive visualization tool*” +“Health Record*”
“Interactive visualization tool*” + “Electronic Health Record*”
“Interactive visualization tool*” + “EHR*”
“Interactive visualization tool*” + “Electronic Patient Record*”
“Interactive visualization tool*” + “Electronic Medical Record*”
“Interactive visualization tool*” + “Patients Record*”
“Interactive visualization tool*” + “Patient Record*”
“Visualization system*” + “Health Record*”
“Visualization system*” + “Electronic Health Record*”
“Visualization system*” + “EHR*”
“Visualization system*” + “Electronic Patient Record*”
“Visualization system*” + “Electronic Medical Record*”
“Visualization system*” + “Patients Record*”
“Visualization system*” + “Patient Record*”
“Information visualization system*” + “Health Record*”
“Information visualization system*” + “Electronic Health Record*”
“Information visualization system*” + “EHR*”
“Information visualization system*” + “Electronic Patient Record*”
“Information visualization system*” + “Electronic Medical Record*”
“Information visualization system*” + “Patients Record*”
“Information visualization system*” + “Patient Record*”

2.3.2 Selection Criteria

Out of all the studies that survived the initial filtering, we only included those that described an interactive visualization tool and provided a detailed description of the tool’s visualization and its interaction design in order to analyze how the tool can support different EHR-data-driven tasks and activities. All the papers related to the visualization of any administrative tasks with patient data, medical guidelines, genetics data, and syndromic surveillance were excluded from our survey as we only focused on clinical

EHR data. We also excluded the studies that were solely focused on the visualization of free text (e.g., the patient's progress notes) and medical images (e.g., magnetic resonance imaging, and X-ray images).

2.3.3 Results

A total of 912 articles were identified from our initial search of electronic databases. A search of the gray literature and manually searching references from articles resulted in an additional 34 papers. We removed a total number of 205 duplicates that were included in the 946 articles, both within and between search engines. We then reviewed all the abstracts and excluded 685 further articles. Next, we read the full text of 56 remaining articles and excluded the ones that did not meet the selection criteria. Finally, 24 studies remained for the analysis. The results of the selection procedure are displayed in the flow diagram in Figure 3.

2.4 Survey of the Interactive Visualization Tools

In this section, we provide a survey of 19 IVTs that are described in the chosen articles and use our proposed activity and task framework to analyze them. The survey includes an overview of the goal of the IVT, a brief description of its visualization, and an analysis of how sub-activities, tasks, and sub-tasks blend and combine to accomplish the tool's main higher-level activities of ***interpreting***, ***predicting*** and, ***monitoring***. A very important criterion to differentiate IVTs is whether they support activities that involve multiple patient records or exploration of an individual patient. We divide our survey into two different types of IVTs based on this criterion: population-based tools and single-patient tools. Initially, studies were focused on single-patient tools, but since 2010, most of the IVTs are developed to support large numbers of patient records. Our survey includes more population-based tools, as it seems that these are more prevalent than single-patient tools. For the first type, we survey 14 tools, and, for the second type, we survey five tools.

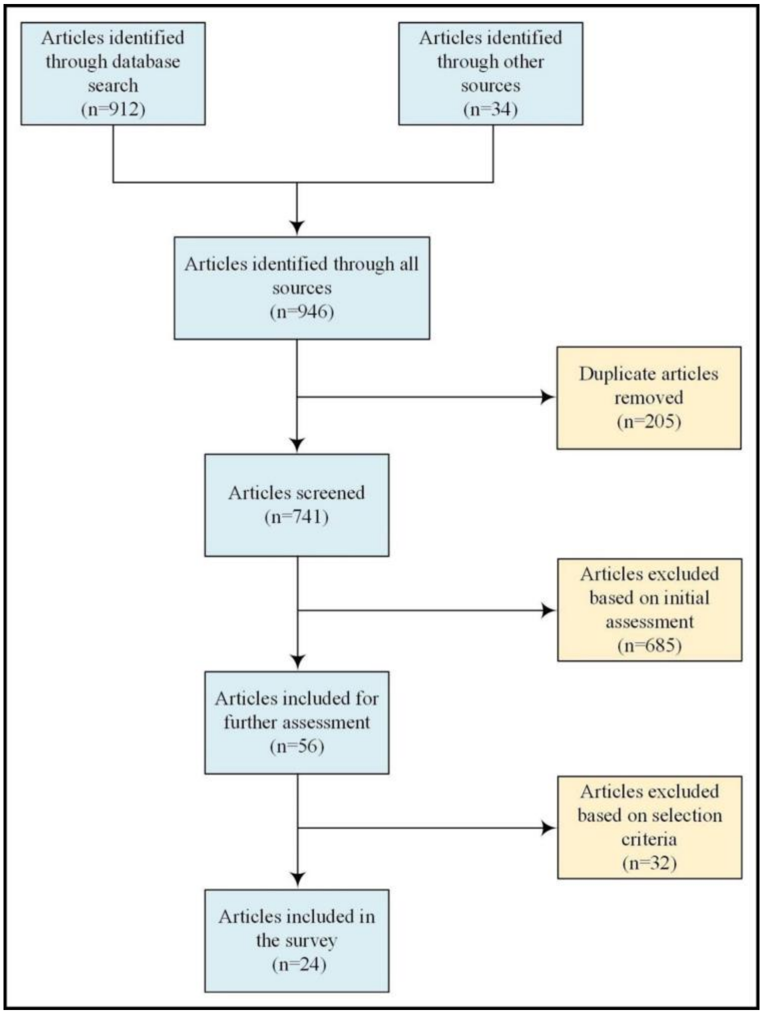

Articles identified through database search (n=912)
Articles identified through other sources (n=34)
Articles identified through all sources (n=946)
Duplicate articles removed (n=205)
Articles screened (n=741)
Articles excluded based on initial assessment (n=685)
Articles included for further assessment (n=56)
Articles excluded based on selection criteria (n=32)
Articles included in the survey (n=24)

Figure 2-3: Search results and how we selected the 24 articles that described 19 IVTs.

2.4.1 Population-Based Tools

Population-based IVTs support data-driven activities that involve multiplicity of patient records in an aggregate form. Although these types of tools display fewer details about a particular patient, they provide users with the ability to recognize patterns, detect anomalies, find desired records, and cluster and aggregate records into different groups. In this section, we survey fourteen population-based IVTs.

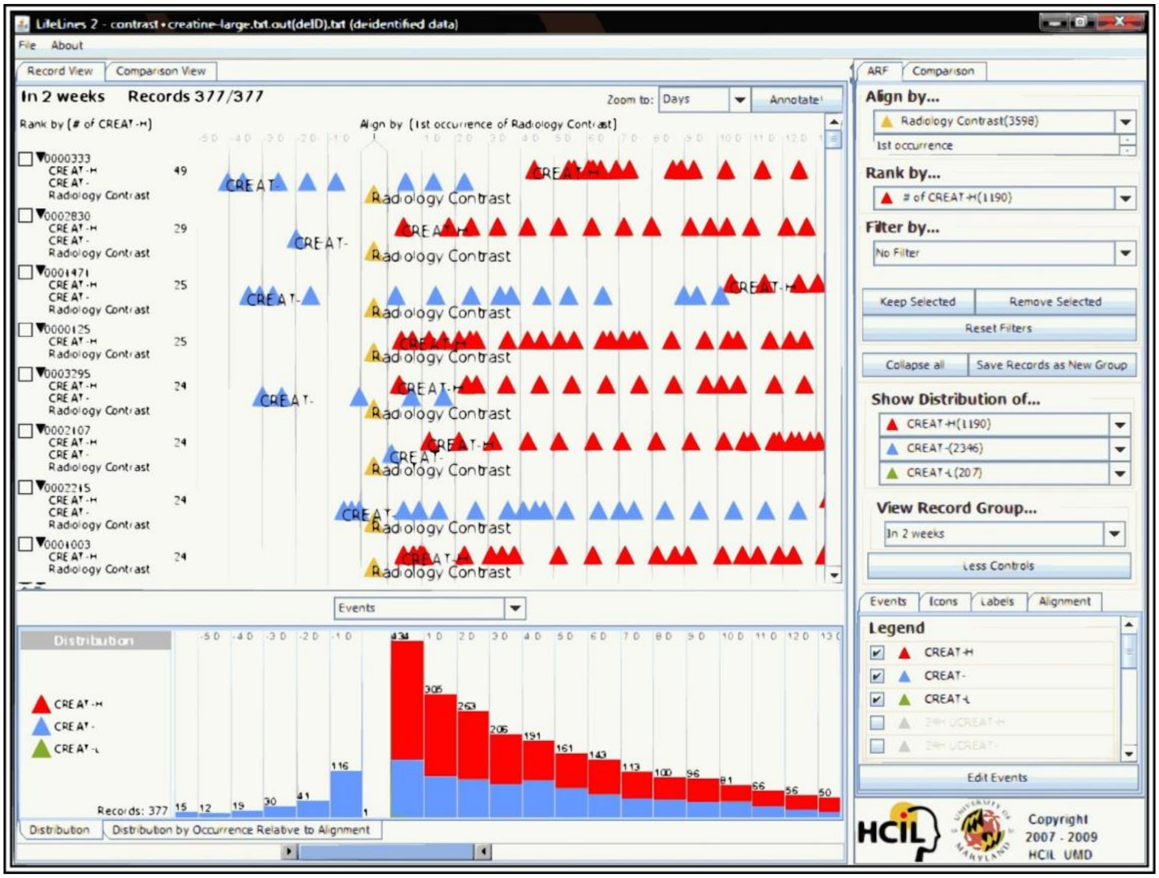
2.4.1.1 Lifelines2

Lifelines2 (T. D. Wang et al., 2009; Taowei David Wang et al., 2008) enables users to explore and analyze a set of temporal categorical patient records interactively. As shown in Figure 4, each record is represented by a horizontal strip containing patient ID and multiple events in patient history that occur at various times. Each event shows up as a color-coded triangle icon on a horizontal timeline. Lifelines2 allows the detection of temporal patterns and trends across EHRs to facilitate hypothesis generation and identify cause-and-effect relationships between patient records.

This tool supports the activity of ***interpreting*** by allowing users to get a better *understanding* of clinical problems and *discovering* patients with interesting medical event patterns. It also supports ***monitoring*** by *investigating* the impact of hospital protocol changes in patient care. It allows for temporal *ordering* of event sequences, *observing* the distribution of temporal events, and *locating* records with particular event sequences. These tasks (*ordering, observing, locating)* are supported by sub-tasks such as *ranking, aggregating,* and *identifying.*

2.4.1.2 Lifeflow

Lifeflow (Guerra Gómez et al., 2011; Krist Wongsuphasawat et al., 2011) provides a visual summary of the exploration and analysis of event sequences in EHR data. While in Lifelines2, due to limited screen space, it is not possible to see all records simultaneously; Lifeflow gives users the ability to answer questions that require an overview of all the records. To convert from Lifelines2 view to Lifeflow, a data structure called “tree of sequences” is created by aggregating all the records. This structure is then converted into a Lifeflow view with each node representing an event bar. Figure 5 shows Lifeflow visualization where all the records are vertically stacked on the horizontal timeline and all the events are represented using color-coded triangles.



LifeLines 2 - contrast+creatine-large.txt.out(deID).txt (deidentified data)

In 2 weeks Records 377/377

Zoom to: Days

Rank by (# of CREAT-H)

Align by (1st occurrence of Radiology Contrast)

Align by... Radiology Contrast(3598) 1st occurrence

Rank by... # of CREAT-H(1190)

Filter by... No Filter

Show Distribution of... CREAT-H(1190) CREAT-(2346) CREAT-L(207)

View Record Group... In 2 weeks

Records: 377

Figure 2-4: Lifelines2: Interactive visualization tool for temporal categorical data. Source: Image courtesy of the University of Maryland Human–Computer Interaction Lab, http://hcil.umd.edu.

In this IVT, the sub-activities of *exploring* and *overviewing* medical events support the activity of ***interpreting***, while *analyzing* aggregated event sequences for quality assurance supports the activity of ***monitoring***. *Recognizing* patterns and temporal *ordering* of aggregated event sequences are two tasks that enable Lifeflow to support *exploring*, *overviewing*, and *analyzing* sub-activities. Finally, sub-tasks such as *aggregating*, *identifying*, and *classifying* work together to accomplish higher-level tasks.

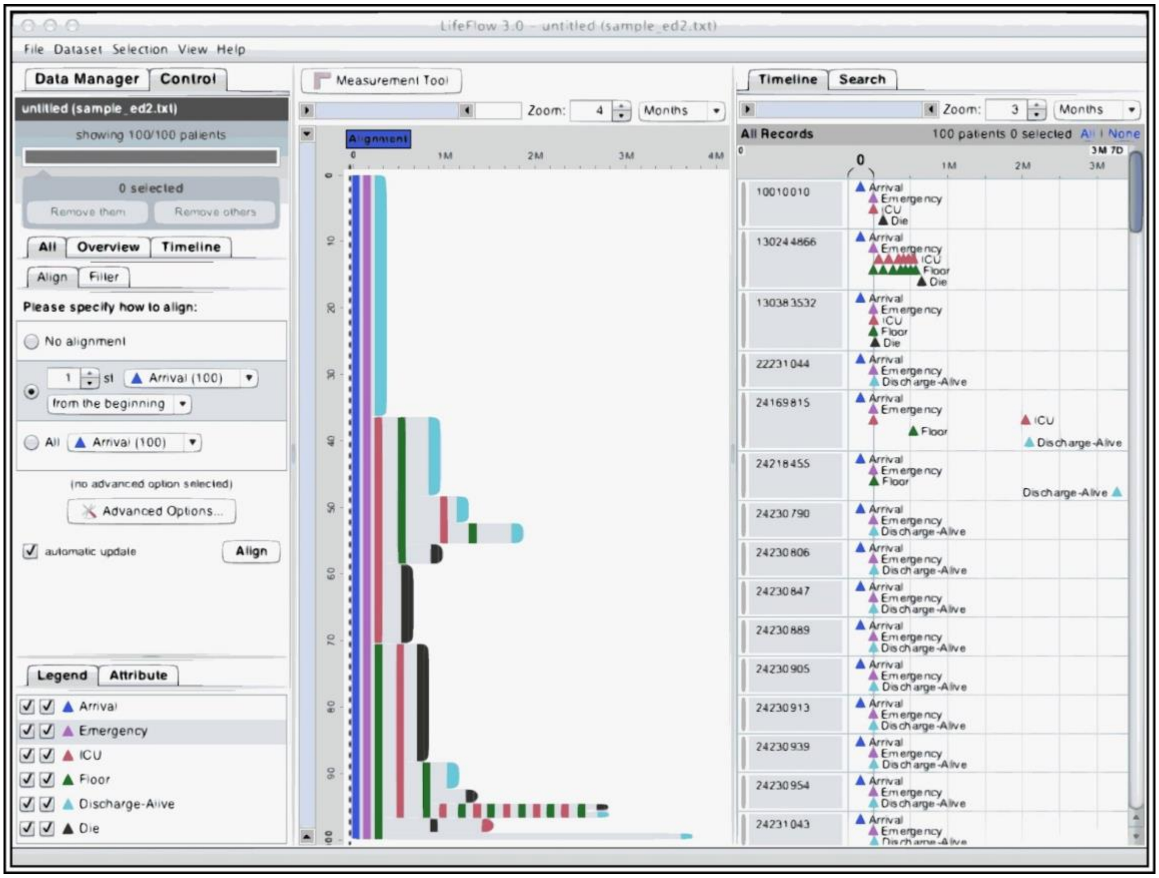
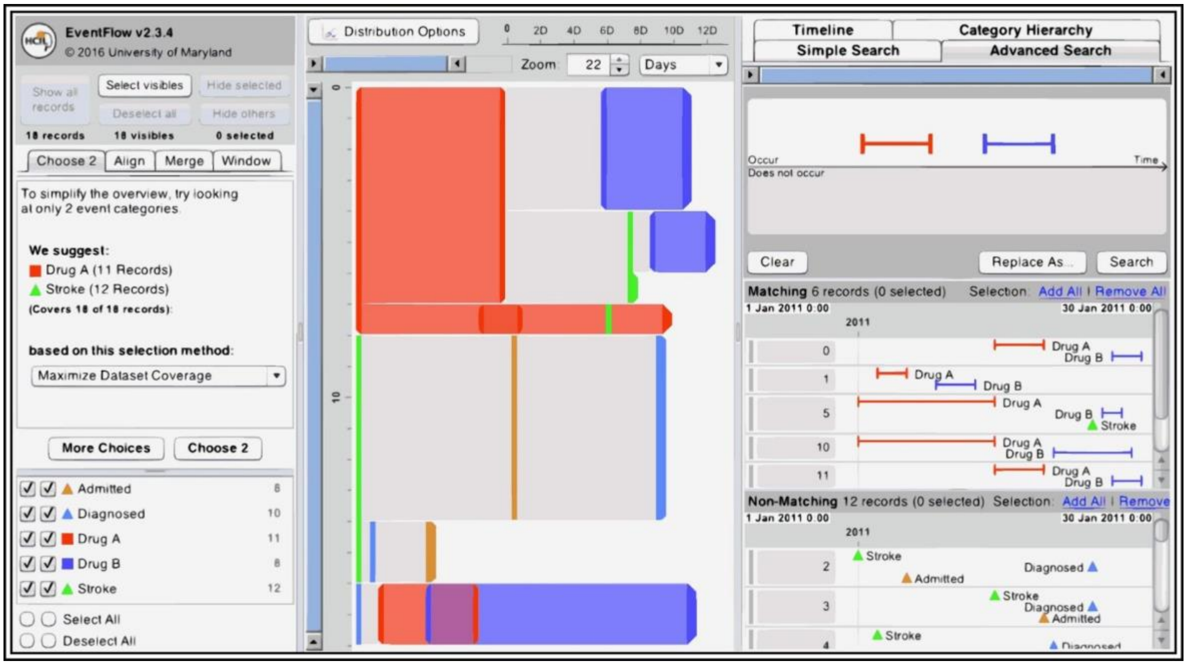


Figure 2-5: Lifeflow: Interactive visualization tool that provides an overview of event sequences. Source: Image courtesy of the University of Maryland Human–Computer Interaction Lab, http://hcil.umd.edu.

2.4.1.3 Eventflow

Eventflow (Monroe et al., 2013) provides users with the ability to query, explore, and visualize interval data interactively. It allows pattern recognition by visualizing events in both a timeline that displays all individual records and an aggregated overview that shows common and rare patterns. As displayed in Figure 6, all the records are shown on a scrollable timeline browser. On the horizontal timeline, point-based events are displayed as triangles, while interval events are represented by the connected rectangles. In the center, an aggregated display gives users an overview of all event sequences in EHR data. The aggregation method works exactly like the one in Lifeflow, but it has been extended to work for interval events in the Eventflow. All the records with the same event sequence are aggregated into a single bar and the average time between two events among the records in the group is represented by the horizontal gap between two bars.

This tool supports ***interpreting*** by providing an *overview* of all event sequences found in the data and *exploring* medical events (point-based events as well as interval events). The *overview*ing and *exploring* sub-activities can be accomplished by *recognizing* temporal patterns and *simplifying* temporal event sequences. ***Monitoring*** can be accomplished by *investigating* aggregated event sequences. The *investigating* sub-activity is supported by *detecting* anomalies in the data. Eventflow supports ***predicting*** by *learning* new hypotheses where this sub-activity can be carried out by tasks such as *specifying* temporal patterns and *simplifying* temporal event sequences. *Aggregating*, *identifying*, *classifying* are the lowest-level sub-tasks for Eventflow.



To simplify the overview, try looking at only 2 event categories.

We suggest:
Drug A (11 Records)
Stroke (12 Records)
(Covers 18 of 18 records)

based on this selection method: Maximize Dataset Coverage

Matching 6 records (0 selected)

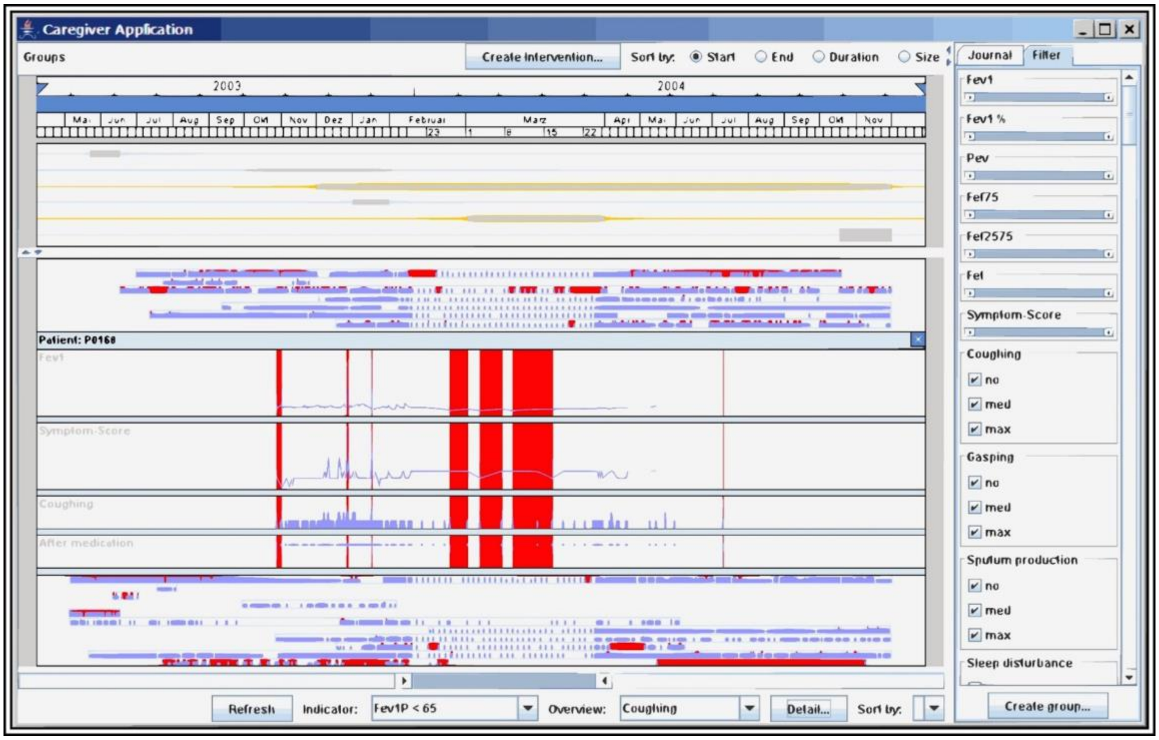
Non-Matching 12 records (0 selected)

Figure 2-6: Eventflow: Interactive visualization tool for analysis of event sequences for both point-based and interval events. Source: image courtesy of the University of Maryland Human–Computer Interaction Lab, http://hcil.umd.edu.

2.4.1.4 Caregiver

Caregiver (Brodbeck et al., 2005) is an IVT that supports therapeutic decision making, intervention, and monitoring. As displayed in Figure 7, the tool has three different views where the upper view displays the duration and size of the patient groups that are chosen by physicians to receive interventions. A common timeline for each patient is shown in the lower view of the chosen attributes. Caregiver allows users to create new cohorts from the search results based on a combination of values of any number of variables.

In this tool, the activity of ***interpreting*** can be accomplished by *discovering* trends, critical incidents, and cause–effect relationships. Caregiver also supports ***predicting*** by allowing users to *learn* about the deterioration in the status of a disease. It supports these sub-activities (*discovering* and *learning*) by *specifying* temporal relationships and *clustering*. *Specifying* and *clustering* can be carried out by sub-tasks such as *identifying*, *classifying*, and *ranking*.



Sort by: ● Start ○ End ○ Duration ○ Size

Coughing ☑ no ☑ med ☑ max

Gasping ☑ no ☑ med ☑ max

Sputum production ☑ no ☑ med ☑ max

Indicator: Fev1P < 65

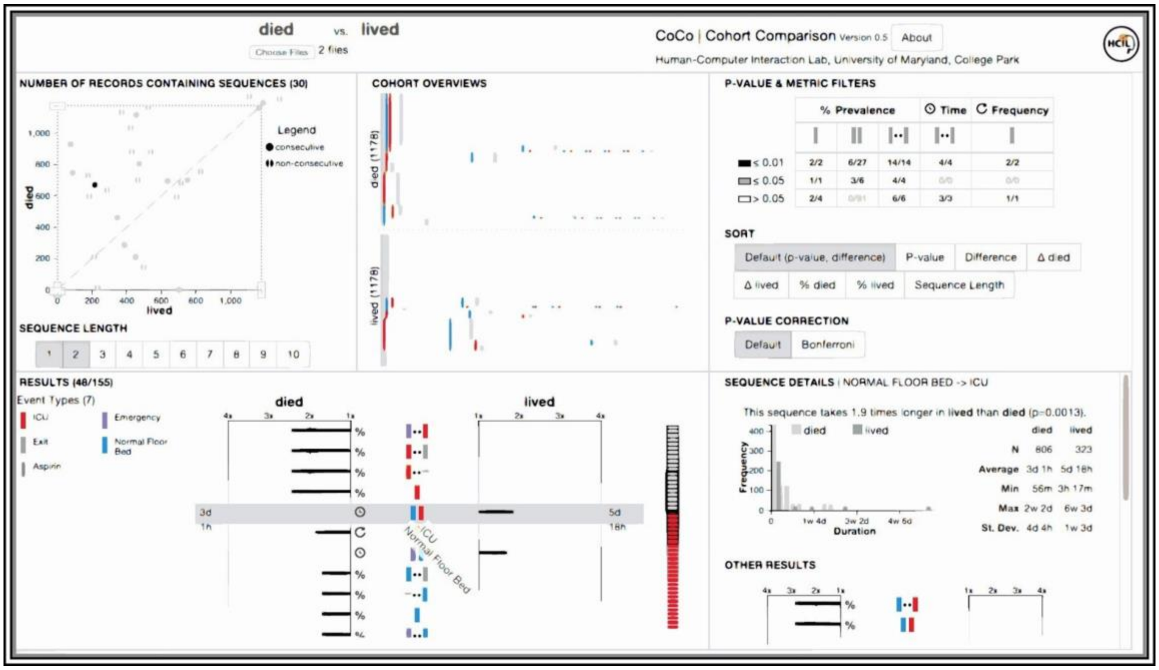
Overview: Coughing

Figure 2-7: Caregiver: Interactive visualization tool for visualization of categorical and numerical data. Source: Image courtesy of Dominique Brodbeck.

2.4.1.5 CoCo

CoCo (Malik et al., 2014, 2015) is an IVT for comparing cohorts of sequences of events recorded in EHRs. It provides users with overview and event-level statistics of the chosen dataset along with a list of available metrics to generate new hypotheses. It consists of a file manager pane, a dataset statistics pane, an event legend, a list of available metrics, the main window, and options for filtering and sorting the results (as shown in Figure 8). The summary panel includes high-level statistics containing the total number of records and events in each record.

CoCo supports the activity of ***interpreting*** by allowing users to *explore* and *investigate* two groups of temporal event sequences simultaneously. The activity of ***predicting*** can be accomplished by *learning* new hypotheses from the statistical analysis while comparing the event sequences (i.e., *detecting* differences among groups of patients). *Ranking*, *classifying*, and *identifying* are the lowest-level sub-tasks in CoCo.



	% Prevalence			Time	Frequency
■ ≤ 0.01	2/2	6/27	14/14	4/4	2/2
▨ ≤ 0.05	1/1	3/6	4/4	0/0	0/0
□ > 0.05	2/4	0/91	6/6	3/3	1/1

This sequence takes 1.9 times longer in **lived** than **died** (p=0.0013).

	died	lived
N	806	323
Average	3d 1h	5d 18h
Min	56m	3h 17m
Max	2w 2d	6w 3d
St. Dev.	4d 4h	1w 3d

Figure 2-8: CoCo: Interactive visualization tool for comparing cohorts of event sequences. Source: image courtesy of the University of Maryland Human–Computer Interaction Lab, http://hcil.umd.edu.

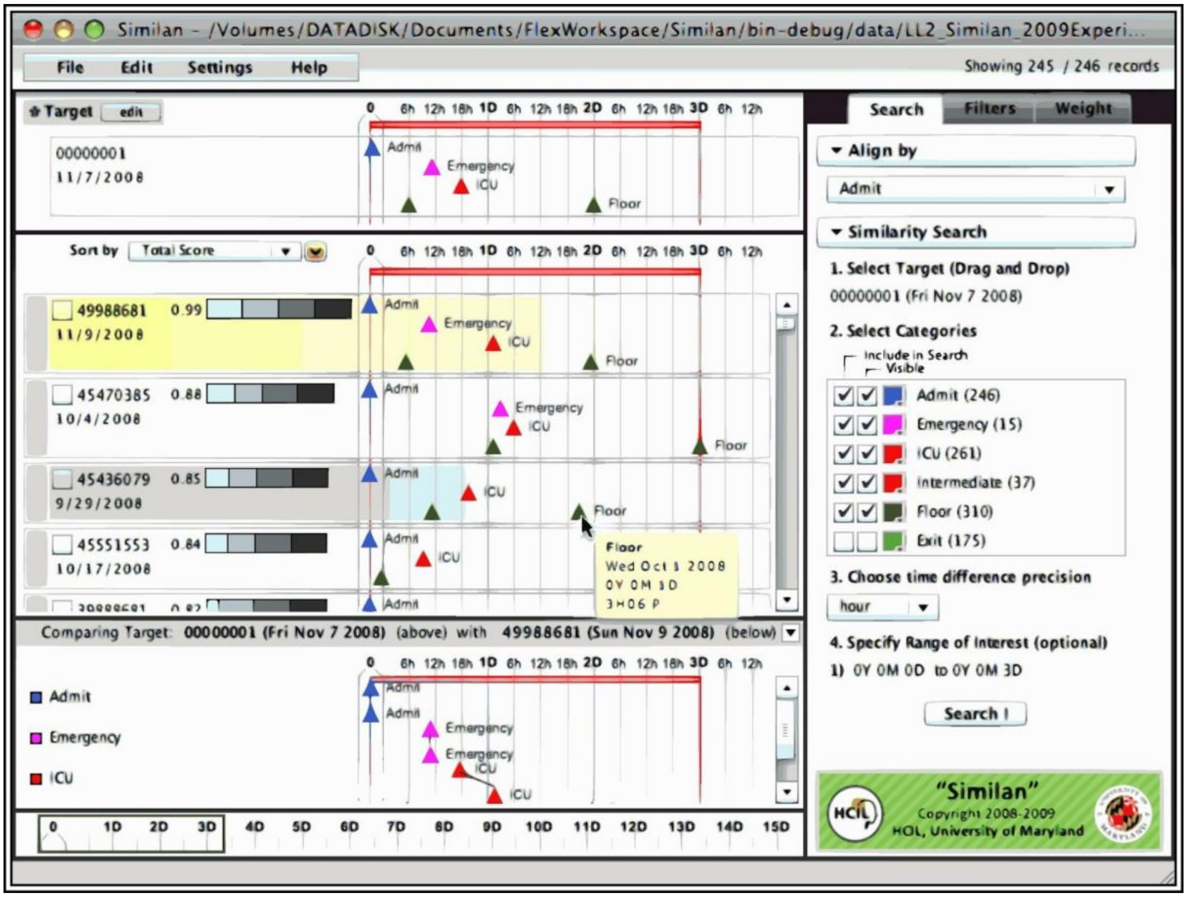
2.4.1.6 Similan

Similan (Krist Wongsuphasawat, 2009) is a tool that provides users with the ability to discover and explore similar records in the temporal categorical dataset. Records are ranked by their similarity to a target record that can be either a reference record or a user's specified sequence of events. The similarity measure considers the transposition of events, addition, removal, and temporal differences of matching to estimate the similarity of temporal sequences. Simian lets users to visually compare the selected target with a set of records and rank those records based on the matching score, as shown in the left side middle panel in Figure 9.

In this IVT, ***interpreting*** can be carried out by *exploring* and *discovering* similar records in temporal categorical data where these sub-activities themselves are supported by *detecting* (calculating similarity measure among records) and *recognizing* similarity among records. ***Predicting*** is accomplished by *discovering* patients with similar symptoms to a certain target patient. The sub-activity *discovering* can be carried out by tasks such as temporal *ordering* and dynamic *query*. Finally, sub-tasks such as *ranking*, *identifying*, and *classifying* work together to accomplish higher-level tasks.

2.4.1.7 Outflow

Outflow (K. Wongsuphasawat & Gotz, 2012; Krist Wongsuphasawat & Gotz, 2011) is a graph-based visualization that shows the eventual outcome across the event sequences in patient records. It aggregates and displays event progression pathways and their corresponding properties, such as cardinality, outcomes, and timing. The tool allows users to interactively analyze the event sequences and detect their correlation with external factors (e.g., beyond the collection of event types that specify an event sequence). The tool is a state transition diagram, which is represented by a directed acyclic graph. The states (nodes) are unique combinations of patient symptoms that are mapped to rectangles, where the height of each rectangle is proportional to the number of patients. The graph is divided into different layers vertically, where layer *i* consists of all states in the graph with *i* symptoms. These layers are arranged from left to right, displaying patient history from past to future. Edges display transitions among symptoms



Similan - /Volumes/DATADISK/Documents/FlexWorkspace/Similan/bin-debug/data/LL2_Similan_2009Experi...

Showing 245 / 246 records

Align by: Admit

1. Select Target (Drag and Drop): 00000001 (Fri Nov 7 2008)

2. Select Categories

Include in Search	Visible	Category
☑	☑	Admit (246)
☑	☑	Emergency (15)
☑	☑	ICU (261)
☑	☑	Intermediate (37)
☑	☑	Floor (310)
□	□	Exit (175)

3. Choose time difference precision: hour

4. Specify Range of Interest (optional): 1) 0Y 0M 0D to 0Y 0M 3D

Sort by: Total Score

Record	Date	Score
49988681	11/9/2008	0.99
45470385	10/4/2008	0.88
45436079	9/29/2008	0.85
45551553	10/17/2008	0.84
30000581		0.82

Floor
Wed Oct 1 2008
0Y 0M 1D
3H06 P

Comparing Target: 00000001 (Fri Nov 7 2008) (above) with 49988681 (Sun Nov 9 2008) (below)

Figure 2-9: Similan: interactive visualization tool for the exploration of similar records in the temporal categorical data. Source: image courtesy of the University of Maryland Human–Computer Interaction Lab, http://hcil.umd.edu.

where each edge encodes the number of patents that are involved in the transition and the average time interval between different states. The end state that is represented by a trapezoid followed by a circle is used to mark points where the patient paths have ended. Finally, the color of the edges and end states represents the average outcome for the corresponding group of patients.

In this tool, sub-activities of *exploring* and *overviewing* event sequences work together to accomplish the activity of ***interpreting***. Outflow also supports ***predicting*** by allowing users to *discover* the progression of temporal event sequences. The sub-activities of *exploring*, *overviewing*, and *discovering* can be accomplished by *summarizing* temporal event sequences, *specifying* temporal relationships, and *detecting* patterns from statistical summaries. Finally, *aggregating*, *identifying*, and *classifying* are the lowest-level sub-tasks.

2.4.1.8 IPBC

IPBC (Chittaro et al., 2003) (interactive parallel bar charts) is an interactive 3D visualization of temporal data. IPBC applies visual data mining to a real medical problem such as the management of multiple hemodialysis sessions. It provides users with the ability to make various decisions regarding such things as therapy, management, and medical research. Each time series is displayed as a 3D bar chart where one of the horizontal axes shows time and the vertical axis represents the value, as displayed in Figure 10. Lined up bar charts on the second horizontal axis enable users to view all the series simultaneously.

IPBC supports ***interpreting*** by allowing users to *explore* patient data interactively. ***Monitoring*** can be carried out by *evaluating* the quality of care based on certain clinical parameters. The sub-activities of *exploring* and *evaluating* are supported by *specifying* temporal relationships and *recognizing* similar patterns where these tasks themselves can be accomplished by sub-tasks such as *identifying*, *classifying*, and *ranking*.

2.4.1.9 Gravi++.

Gravi++ (Hinum et al., 2005) allows users to explore and analyze multiple categorical variables using interactive visual clustering. This tool uses a spring-based layout to place both patient and variable icons across the visualization, where the value of a variable for a patient identifies the distance between that patient's icon and the variable’s icon. Gravi++ provides users with the ability to detect clusters since patients with similar values are placed together on screen. In order to visualize the exact values of each variable for each patient, the tool shows each patient’s value as a circle around variables. The patient icons are represented by spheres while the variable icons are encoded by squares. Moreover, the tool can encode different patient attributes using patient icons; for instance, the size of the sphere can be mapped to the body mass index of the patient and its color can encode the patient’s gender or therapeutic outcome.

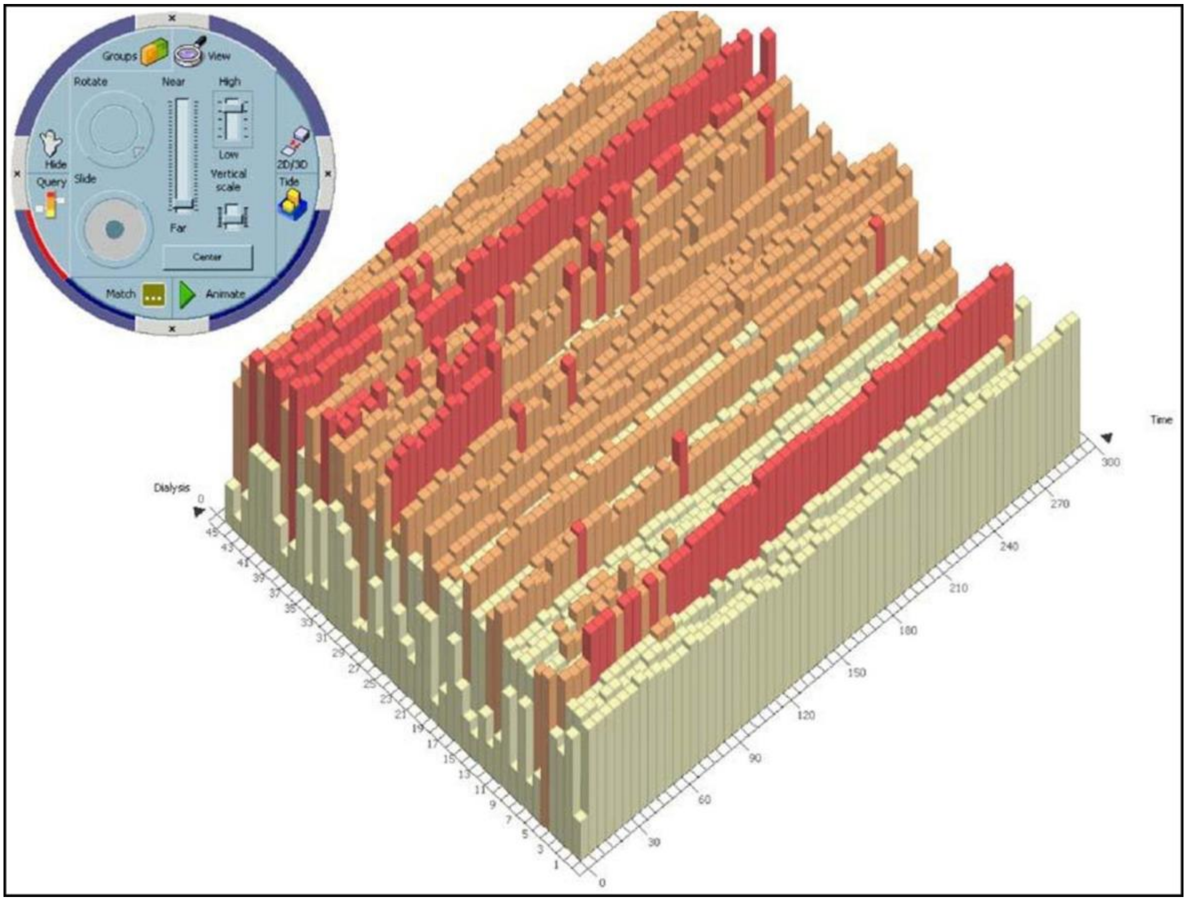


Figure 2-10: IPBC: 3D visualization tool for analysis of numerical data from multiple hemodialysis sessions. Source: reprinted from Journal of Visual Languages & Computing, 14, Chittaro L, Combi C, Trapasso G, Data mining on temporal data: a visual approach and its

This tool supports the activity of ***interpreting*** by allowing users to *explore* patient data and *discover* clusters of similar patients. ***Monitoring*** can be accomplished by *investigating* the development of a patient after a certain treatment. The sub-activities of *exploring*, *discovering*, and *investigating* are supported by tasks such as *recognizing* patterns and *specifying* temporal relationships. Finally, *identifying* and *classifying* are the lowest-level sub-tasks that are supported by the tool.

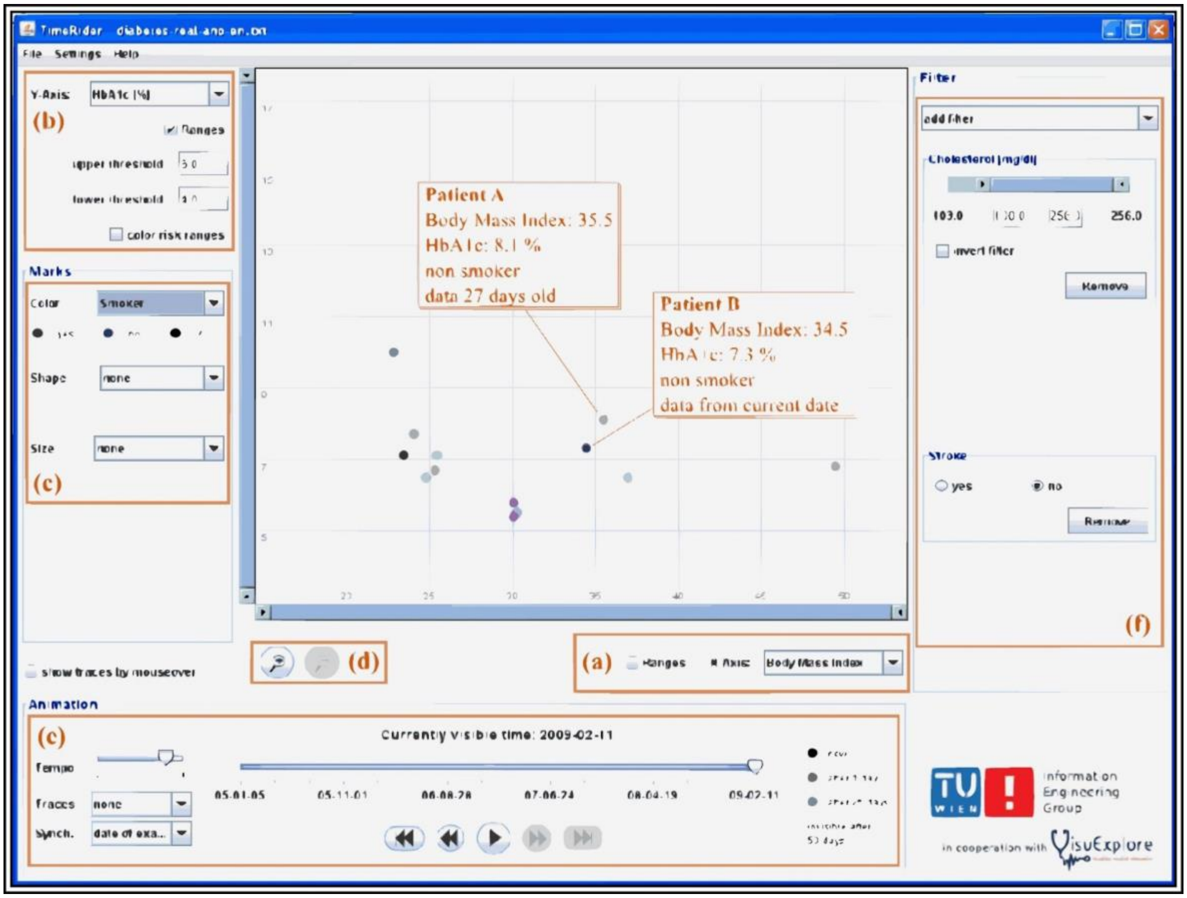
2.4.1.10 PatternFinder

PatternFinder (Fails et al., 2006) is a query-based tool for data visualization and visual query that can help users search and discover temporal patterns within multivariate categorical data. PatternFinder allows users to specify queries for temporal events with time span and value constraints and enables them to look for temporally ordered events/values/trends as well as the existence of events. Also, users can set a range of possible time spans among the events to specify how far apart the events are from each other. The tool has two main panels: the pattern design and query specification panel and the result visualization panel. The leftmost part of the pattern design panel is the Person/People panel that enables users to limit the types of patients by name, by choosing from a list of patients, or by typing a text string. Any modifications that are done in this panel are dynamic queries that lead to an immediate update of the results in the result visualization panel. The temporal panel that is placed to the right of the Person/People panel enables users to form temporal pattern queries by chaining the events together. Users are able to search for the presence of events, the temporal sequence of events (e.g., an emergency doctor's visit followed by a hospitalization), the temporal sequence of values (e.g., 200 or below cholesterol followed by 240 or higher), and the temporal value patterns (e.g., monotonically decreasing). The result visualization panel displays a graphical table of all the matches where each row shows a single pattern match for one patient. Pattern matches are represented as a timeline in a "ball-and-chain" visualization fashion where the event points are shown as circles and time spans are displayed by blue bars between the events. The color of the event point in the result visualization panel matches the color of the associated event in the query specification panel. All the events that match the query pattern specified by users are linked together by horizontal lines.

In this tool, the activity of ***interpreting*** is supported by *discovering* patterns and *exploring* patient data dynamically, where these sub-activities themselves can be carried out by tasks such as *specifying* temporal relationships and issuing dynamic *queries*. *Identifying* and *ranking* are the two low-level sub-tasks that work together to support the aforementioned tasks.

2.4.1.11 TimeRider

TimeRider (Rind, Aigner, Miksch, Wiltner, Pohl, Drexler, et al., 2011) offers an animated scatter plot to help users discover patterns in irregularly sampled patient data covering several time spans. As shown in Figure 11, time is represented by either traces or animation in TimeRider. Color, shape, and size of marks are used to encode up to three additional variables. Users can compare patient records of different time spans by synchronizing patients' age, calendar date, and the start and end of the treatment.



Y-Axis: HbA1c [%]

☑ Ranges

upper threshold 5.0

lower threshold 4.0

☐ color risk ranges

Color: Smoker

Shape: none

Size: none

Patient A
Body Mass Index: 35.5
HbA1c: 8.1 %
non smoker
data 27 days old

Patient B
Body Mass Index: 34.5
HbA1c: 7.3 %
non smoker
data from current date

☐ Ranges ● Axis: Body Mass Index

add filter

Cholesterol [mg/dl]

103.0 256.0

☐ invert filter

○ yes ● no

Currently visible time: 2009-02-11

Traces: none

Synch.: date of exa...

invisible after 53 days

Figure 2-11: TimeRider: Interactive visualization tool for pattern recognition in patient cohort data. Source: reprinted by permission from Springer Nature: Springer, Ergonomics and Health Aspects of Work with Computers, Visually Exploring Multivariate *Trends in Patient Cohorts Using Animated Scatter Plots*, Rind A, Aigner W, Miksch S, et al., copyright (2011).

This tool supports ***interpreting*** by allowing users to *detect* trends, clusters, and correlations and providing them with an *overview* to visually compare patient data in parallel. The sub-activities of *detecting* and *overviewing* can be carried out by tasks such as *specifying* temporal relationships, *clustering*, and *recognizing* patterns. *Identifying* and *aligning* are the sub-tasks that work together to support the aforementioned tasks.

2.4.1.12 VISITORS

VISITORS (Klimov et al., 2010a, 2010b) is an IVT that allows for exploration, analysis, and retrieval of raw temporal data. The tool uses raw numerical data (e.g., white blood cell counts) across time to derive temporal abstractions (e.g., durations of low, normal, or high blood-cell-count levels for patients). It then uses lower-level temporal abstractions in conjunction with raw data to generate higher-level abstractions. Finally, patient groups' values are aggregated and displayed. Figure 12 shows this tool's visualization environment, where raw numerical data is represented by line charts, whereas categorical data is displayed as tick marks or bars on a horizontal zoomable timeline.

In this tool, the activity of ***interpreting*** is supported by *exploring* patient data in different temporal granularities. The sub-activity of exploring can be carried out by tasks such as *specifying* relationships, *observing* the distribution of aggregated values of a group of patients, and *locating* records based on specific time and value constraints. VISITORS supports the activity of ***monitoring*** by sub-activities, such as *investigating* treatment effects, clinical trial results, and quality of clinical management processes. The latter sub-activity, *investigating*, can be carried out by the task of *recognizing* patterns as well as all the other tasks needed to support the activity of *interpreting*. Finally, *aggregating*, *classifying*, *aligning*, and *identifying* are the lowest-level sub-tasks that are supported by this tool.

2.4.1.13 Prima

Prima (Gresh et al., 2002) is a population-based IVT that allows users to explore the categorical and numerical data by constructing different linked views. This helps users to not only understand the large set of patient records but also discover patterns and trends in the dataset. The aggregated window provides an overview of the categorical variables

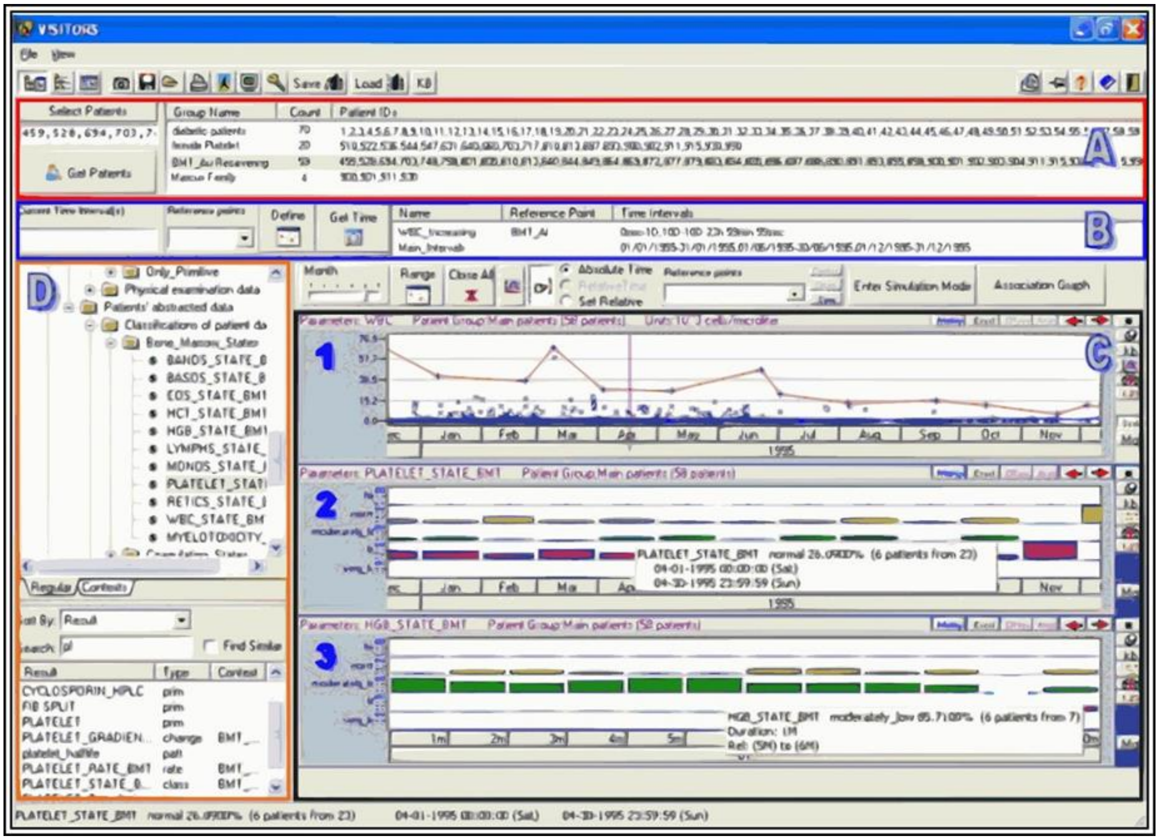


Figure 2-12: VISITORS: Interactive visualization tool for the exploration of multiple patient records. (A) displays lists of patients. (B) displays a list of time intervals. (C) displays the data for a group of 58 patients over the current time interval. Panel 1 shows the white blood cell raw counts for the patients, while Panels 2 and 3 display the states of monthly distribution of platelet and haemoglobin in higher abstraction, respectively. Abstractions are encoded in medical ontologies displayed in panels (D). Source: reprinted from Journal of Artificial Intelligence in Medicine, 49, Klimov D, Shahar Y, Taieb-Maimon M, *Intelligent visualization and exploration of time-oriented data of multiple patients*, 11-31., copyright (2010), with permission from Elsevier.

by showing the proportions of patients in each category for those variables using stacked bar charts. This window enables users to filter patients by applying a color "brush". It also displays correlations among different categorical variables through interactive coloring. Another view displays a histogram of numerical variables. The data can also be explored with a 2D scatter plot. Another view of the data is called multiple category tables. It shows the values of either a single variable or multiple categories. Finally, the tool incorporates the Kaplan–Meier curve to estimate the survival function from the patient data.

Prima supports the activity of ***interpreting*** by allowing users to *explore* patient data interactively, where this sub-activity itself can be accomplished by *recognizing* patterns and *specifying* temporal relationships. Finally, *aggregating* and *ranking* are the lowest-level sub-tasks that are supported by the tool.

2.4.1.14 WBIVS

WBIVS (Pieczkiewicz et al., 2007) is a web-based interactive tool that visualizes numerical and categorical variables for lung transplant home monitoring data. Numerical variables are displayed in line plots, while categorical variables are visualized in matrix plots. The tool visualizes ten variables in total. When a data point gets selected, all the other data points that belong to the same time period will get highlighted in the other charts. Moreover, users can find details about the last two chosen data points on the right part of the graph.

This tool supports the ***interpreting*** activity by allowing users to *explore* patient data interactively and *discover* patterns. ***Monitoring*** is supported by *investigating* treatment effects. The *exploring and discovering* sub-activities can be accomplished by tasks such as *specifying* temporal relationships among data points and *organizing* data for pattern recognition. These tasks can be composed of lowest-level sub-tasks, such as *identifying*, *classifying*, and *highlighting*.

2.4.2 Single-Patient Tools

Single-patient IVTs provide visualizations of one single-patient record at a time. These tools enable users to overview a given patient’s historical data, detect important events in the patient’s history, and recognize trends. In this section, we survey five single-patient IVTs.

2.4.2.1 Midgaard

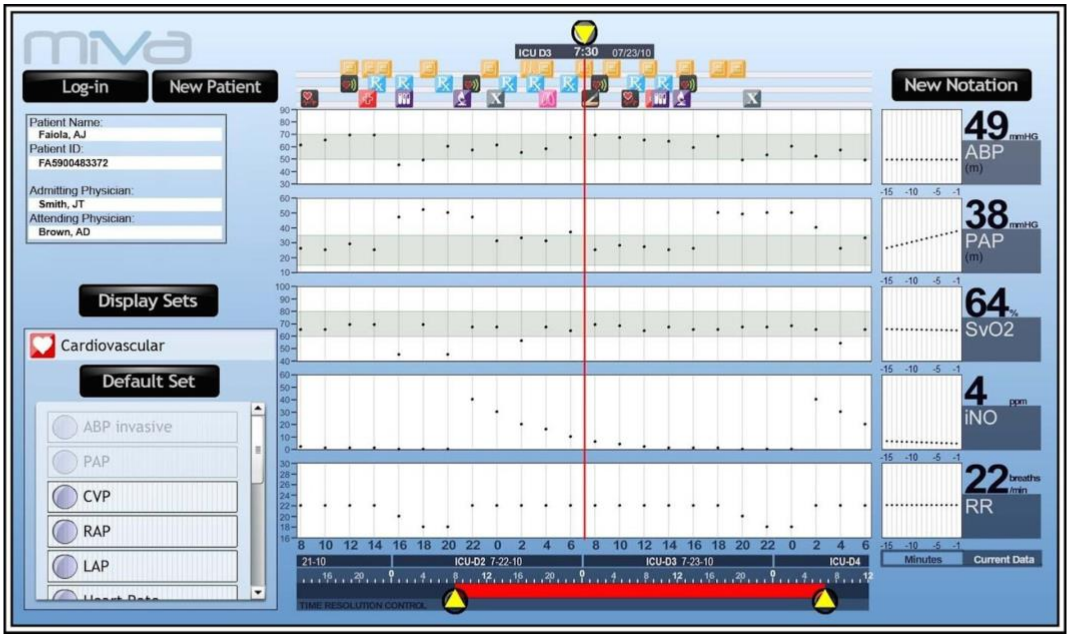
Midgaard (Bade et al., 2004) allows for exploration of the intensive care units' data at different levels of abstraction from overview to details. It uses visualizations to display numerical variables of treatment plans. It incorporates a complex semantic zoom method for numerical variables by calculating their categorical abstractions based on the available screen area and zoom level. Midgaard provides users with the ability to switch between different views such as a colored background, colored bars, area charts, or augmented line charts based on the level of details. The tool can progressively switches to a more detailed view to display all the individual data points when users zoom in or switch back to more compact graphical elements when they zoom out.

Midgaard can also visualize medical treatment plans using colored bars where each bar can contain further bars displaying sub-plans. It allows users to navigate and zoom by interacting with two time axes that are placed below the visualization area. The bottom axis displays a temporal overview of the patient record while the middle axis allows users to see specific time intervals in more detail.

The activity of ***interpreting*** is supported by *exploring* patient data at different levels of abstraction, where this sub-activity itself can be accomplished by tasks such as *recognizing* fluctuations in data. *Identifying* and *classifying* are the two sub-tasks that are supported by this tool.

2.4.2.2 MIVA

MIVA (Faiola & Newlon, 2011) (Medical information visualization assistant) is a tool that transforms and organizes biometric data into temporal resolutions to provide healthcare providers with contextual knowledge. It allows users to prioritize and customize visualizations based on specific clinical problems. It visualizes the data using point plots to display temporal changes in numerical values, where each variable is represented by a separate plot, as shown in Figure 13. MIVA enables users to detect changes in multiple physiological data points over time for faster and more accurate diagnosis. Users can control the data source, time resolutions, and time periods to narrow down the assessment of a patient's condition.



Patient Name:
Faiola, AJ
Patient ID:
FA5900483372

Admitting Physician:
Smith, JT
Attending Physician:
Brown, AD

Figure 2-13: MIVA: Interactive visualization tool to show the temporal change of numerical values where each variable is represented by an individual point plot. Source: image courtesy of Antony Faiola.

This tool supports the activity of ***interpreting*** by enabling users to carry out sub-activities such as *exploring* longitudinal relationships in patient data where this sub-activity can be accomplished by tasks such as *specifying* temporal relationships and *recognizing* patterns. At the level of sub-tasks, this tool supports *identifying* as well as *classifying*.

2.4.2.3 VIE–VISU

VIE–VISU (Horn et al., 2001) uses a set of glyphs to display changes in a patient's status over time in intensive care. Each glyph’s geometrical shape and color encodes categorical variables, while the numerical variables are represented by size of the glyph's elements. Every glyph can encode 15 variables that are classified by physiological systems. For instance, the respiratory parameters are mapped to a rectangle in the middle of the glyph; circulatory parameters are mapped to a triangle on top of the glyph, and the fluid balance parameters are shown by two smaller rectangles at the bottom of the glyph. By default, the tool displays 24 glyphs, one per hour.

The activity of ***interpreting*** can be accomplished by *overviewing* a patient's status, where this sub-activity is supported by tasks such as *recognizing* patterns. This tool supports ***monitoring*** by *evaluating* changes in patient's status over time. The task of *identifying* temporal relationships supports the sub-activity of *evaluating*. Finally, *aggregating* and *classifying* are two sub-tasks that can be carried out by the tool.

2.4.2.4 Lifelines

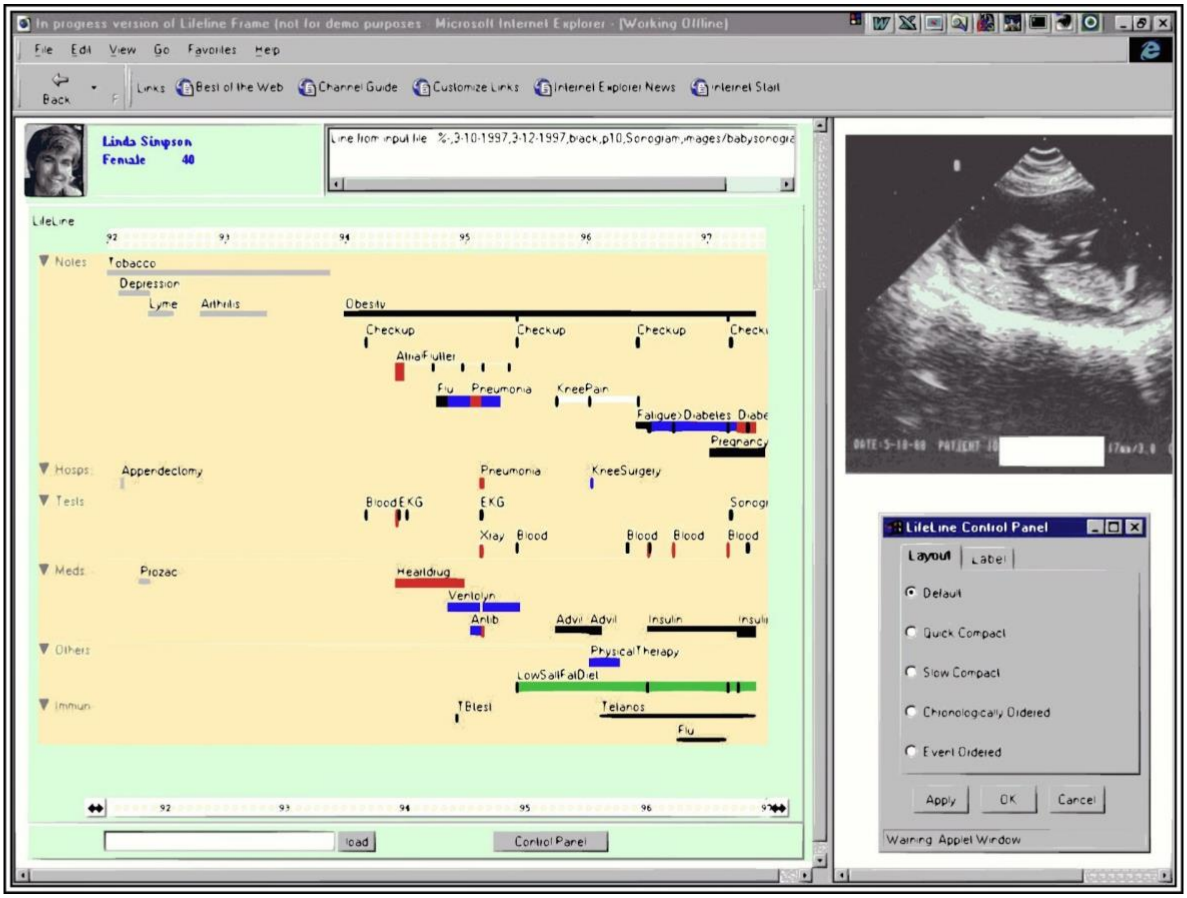
Lifelines (Plaisant et al., 1998) offers a visualization environment to show patient history on a zoomable timeline, where a patient's medical record is displayed by a set of events and lines. Episodes and events in a patient record are represented by a set of multiple line segments as shown in Figure 14. Color can be used to encode the states of categorical variables. This IVT provides an overview of a patient history to recognize trends, specify important events, and detect omissions in data.

The activity of ***interpreting*** is supported by *understanding* patient's status where this sub-activity itself can be carried out by tasks such as *recognizing* patterns and *specifying* temporal relationships. The tool supports ***monitoring*** by allowing users to carry out sub-activities such as *investigating* trends and anomalies in patient data. The *investigating* sub-activity is supported by *outlining* and *summarizing* the patient data. Finally, *aggregating*, *classifying*, and *identifying* are the sub-tasks that are supported by the tool.

2.4.2.5 VisuExplore

VisuExplore (Pohl et al., 2011; Rind, Aigner, Miksch, Wiltner, Pohl, Turic, et al., 2011) displays patient data in different views aligned with a horizontal timeline, where each view shows multiple variables. This IVT uses common visualization techniques that make it easy to use and learn. In this tool, numerical data are displayed using bar charts and line plots, whereas categorical data are represented using event charts and timeline charts, as shown in Figure 15.

In this tool, the activity of ***interpreting*** is supported by *exploring* temporal data of patients with chronic diseases, where this sub-activity can be carried out by tasks such as



Line from input file %;3-10-1997,3-12-1997,black,p10,Sonogram,images/babysonogra

● Default

Warning: Applet Window

Figure 2-14: Lifelines: interactive visualization tool that displays patient's medical histories on a timeline. Source: image courtesy of the University of Maryland Human–Computer Interaction Lab, http://hcil.umd.edu.

specifying temporal relationships. Finally, *aligning* and *identifying* are two sub-tasks that can be carried out by the tool.

2.5 Discussion and Limitations

In this paper, we have presented and proposed a framework to identify and analyze EHR-data-driven tasks and activities in the context of IVTs—that is, all the activities, sub-activities, tasks, and sub-tasks that are supported by EHR-based IVTs. Using a survey of 19 EHR-based IVTs, we demonstrate how these IVTs support activities by identifying the combination of sub-activities, tasks, and sub-tasks that work together to help users carry out the three higher-level activities as displayed in Table 3. ***Interpreting*** is supported by all IVTs surveyed in this paper. Eventflow, Similan, CoCo, Outflow, and

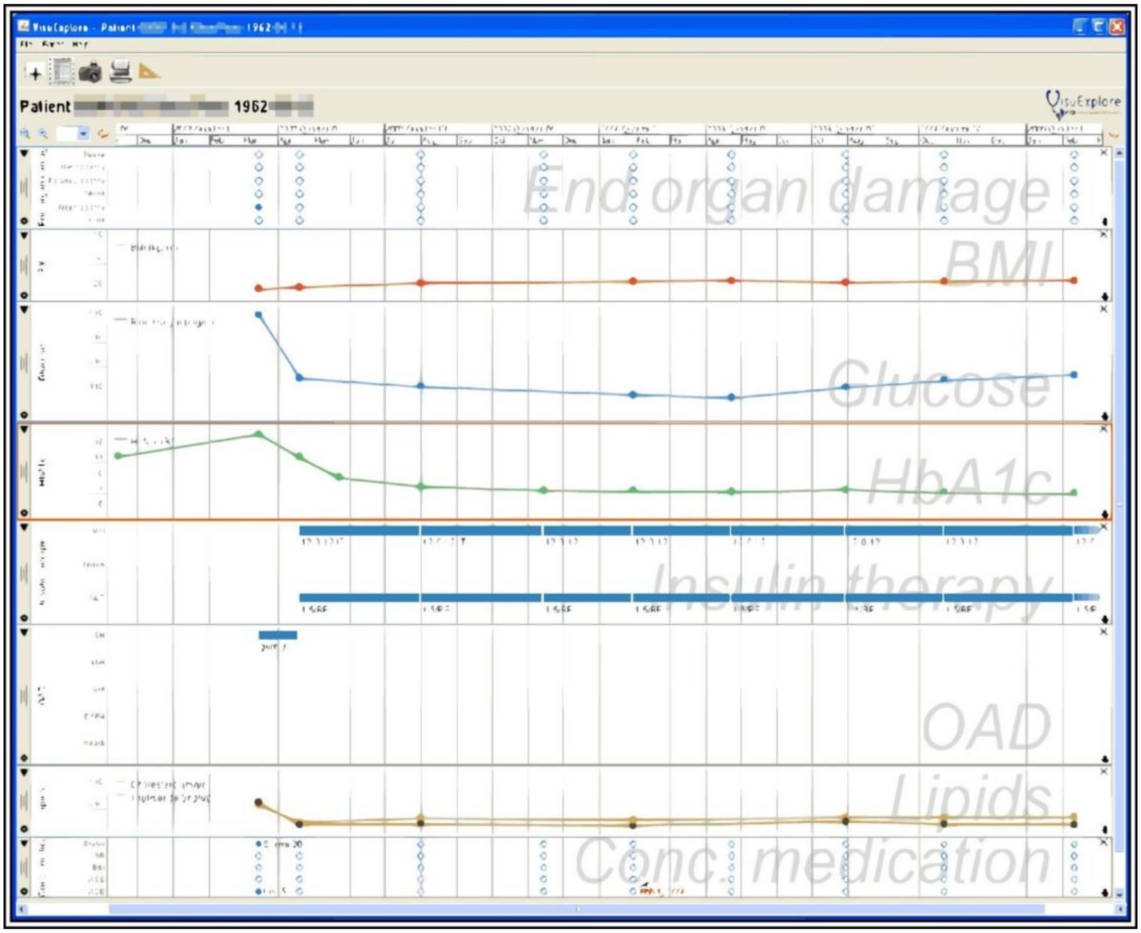


Figure 2-15: VisuExplore: interactive visualization tool that displays patient data in various views on a timeline. Source: reprinted by permission from Springer Nature: Springer, Human–Computer Interaction, Patient Development at a Glance: An Evaluation of *a Medical Data Visualization*, Pohl M, Wiltner S, Rind A, et al., copyright (2011).

Caregiver are the only IVTs that support ***predicting***, whereas Lifelines2, Lifeflow, Eventflow, Gravi++, IPBC, TimeRider, VISITORS, WBIVS, VIE-VISU, Lifelines, CoCo, and Visu-Explore are the tools that facilitate ***monitoring***. Going down from high-level activities, *recognizing* patterns and *specifying* temporal relationships are the most common sub-activities that help users with the activity of ***interpreting*** in most of the IVTs. The existing EHR-based IVTs support ***predicting*** by giving users the ability to perform sub-activities such as *learning* new hypotheses, *discovering* patients with similar symptoms to a target patient, and *detecting* early deterioration of a disease. Finally, the most common sub-activities that facilitate ***monitoring*** are *evaluating* the quality of care and *investigating* the development of a patient's status after treatment.

Our proposed framework can offer a number of benefits for designers, researchers, and evaluators of EHR-based IVTs. Firstly, the framework can help the designer to conceptualize activities, tasks, and sub-tasks of EHR-based IVTs systematically. Secondly, it can assist researchers in making sense of IVTs by providing them with all the activities that can be accomplished by carrying out different sets of sub-activities, tasks, and sub-tasks. Thirdly, this framework can be used by evaluators to identify the gaps in support of higher-level activities supported by existing IVTs. It appears that almost all existing IVTs focus on the activity of ***interpreting***, while only a few of them support ***predicting*** despite the importance of this activity in supporting users to find the patients that are at high risk and identify the risk factors of various diseases. Also, some of the EHR-based IVTs do not pay enough attention to ***monitoring***, even though this activity is beneficial in investigating the quality of clinical management processes. All these higher-level activities should be an integral part of a properly designed EHR-based IVT since healthcare providers use such tools to (1) better understand patients' condition, (2) anticipate the discourse of a specific disease, and (3) track patients' condition after treatment. Most of the tools surveyed in this paper can only satisfy a certain aspect of users' needs. According to a recent survey in the US, 40% of the clinicians are not satisfied with the existing EHR-based systems (EHRIntelligence, 2018). Therefore, a framework is needed to guide the designer of an IVT in choosing which activities, tasks, and sub-tasks the tool should support. Using questions such as, "What activities can users accomplish by executing a set of tasks?" or "What tasks should be supported to provide users with the ability to perform their activities?", we demonstrate how the proposed framework can be used by designers of EHR-based IVTs to systematically conceptualize and design the tasks and activities of such tools. Given the framework, all designers need to know is, which low-level sub-tasks, tasks, and sub-activities to select and how to blend and combine them to support higher-level activities and allow users to accomplish their overall goal. For instance, if a designer wants to design an IVT to monitor an infant's condition in the neonatal intensive care unit, they can choose different sets of sub-activities, such as *investigating* the effect of a specific treatment or *evaluating* changes in infant's status over time. Then, the designer selects a combination of tasks such as the temporal *ordering* of event sequences or displaying the distribution of temporal events to support the chosen sub-activities. Finally, a set of sub-tasks, such as *ranking*, *aggregating*, and *identifying*, are chosen to support the selected tasks.

We believe a successful EHR-based tool should be capable of doing more than just storing, retrieving, and exchanging patient data. It should support more complex activities, tasks, and sub-tasks to allow healthcare providers to accomplish their goals. Our proposed framework promises a new means for designers of EHR-based IVTs to understand the effectiveness of incorporating such activities, tasks, and sub-tasks in their tool. The use of our framework in EHR-based IVTs will also help physicians to make better treatment decisions and track changes in a patient's condition over time.

This paper has three key limitations. First, we do not investigate the completeness and accuracy of the data sources that IVTs are using as our survey relies on the descriptions of the IVTs found in publications and video tutorials. Second, as the main goal of this paper is the analysis of EHR-based IVTs, we exclude tools that are mainly dependent on statistical and machine learning methods. Finally, we do not consider commercial tools in this paper. This is because online descriptions of such tools do not systematically and thoroughly cover the features of these tools, i.e., their visualizations, interactions, and results.

The findings of this paper will lead to the development of best practices for creating similar frameworks in other domains. A possible area of future research involves developing frameworks for visual analytics tools that incorporate automated analysis techniques along with interactive visualizations to support the increasingly large and complex datasets in EHRs.

Table 2-3: Evaluation summary of the 19 existing tools based on the proposed framework.

	IVTs		Interpreting	Predicting	Monitoring

Population-based tools	Lifelines 2	Sub-activity	discovering, understanding,	no	investigating
		Tasks	locating, observing, ordering	n/a	locating, observing, ordering
		Sub-tasks	aggregating, identifying, ranking	n/a	aggregating, identifying, ranking
	Lifeflow	Sub-activity	exploring, overviewing	no	analyzing
		Tasks	ordering, recognizing	n/a	ordering, recognizing
		Sub-tasks	aggregating, classifying, identifying	n/a	aggregating, classifying, identifying
	Eventflow	Sub-activity	exploring, overviewing	learning	investigating
		Tasks	recognizing, summarizing	specifying, summarizing	detecting
		Sub-tasks	aggregating, classifying, identifying	aggregating, classifying, identifying	aggregating, classifying, identifying
	Similan	Sub-activity	discovering, exploring	discovering	no
		Tasks	detecting, recognizing	ordering, querying	n/a
		Sub-tasks	identifying, classifying, ranking	identifying, classifying, ranking	n/a
	CoCo	Sub-activity	exploring	learning	investigating
		Tasks	detecting	detecting	detecting

		Sub-tasks	classifying, identifying, ranking	identifying, classifying, ranking	identifying, classifying, ranking
	Outflow	Sub-activity	exploring, overviewing	discovering	no
		Tasks	detecting, specifying, summarizing	detecting, specifying, summarizing	n/a
		Sub-tasks	aggregating, classifying, identifying	aggregating, classifying, identifying	n/a
	Caregiver	Sub-activity	discovering	learning	n/a
		Tasks	specifying	clustering, specifying	n/a
		Sub-tasks	classifying, identifying, ranking	classifying, identifying, ranking	n/a
	Gravi++	Sub-activity	discovering, exploring	no	investigating
		Tasks	recognizing, specifying	n/a	recognizing, specifying
		Sub-tasks	classifying, identifying	n/a	classifying, identifying
	IPBC	Sub-activity	exploring	no	evaluating
		Tasks	recognizing, specifying	n/a	recognizing, specifying
		Sub-tasks	classifying, identifying, ranking	n/a	classifying, identifying, ranking
	Pattern Finder	Sub-activity	discovering, exploring	no	no

		Tasks	specifying, querying	n/a	n/a
		Sub-tasks	identifying, ranking	n/a	n/a
	Prima	Sub-activity	exploring	no	no
		Tasks	recognizing, specifying	n/a	n/a
		Sub-tasks	aggregating, ranking	n/a	n/a
	Timerider	Sub-activity	detecting, overviewing	no	investigating
		Tasks	clustering, recognizing, specifying	n/a	recognizing
		Sub-tasks	aligning, identifying	n/a	n/a
	VISITORS	Sub-activity	exploring	no	investigating
		Tasks	locating, observing, specifying	n/a	locating, observing, recognizing, specifying
		Sub-tasks	aggregating, aligning, classifying	n/a	aggregating, aligning, classifying, identifying
	WBIVS	Sub-activity	discovering, exploring	no	investigating
		Tasks	organizing, specifying	n/a	organizing, specifying
		Sub-tasks	classifying, highlighting, identifying	n/a	classifying, highlighting, identifying

Single-Patient Tools	Midgard	Sub-activity	exploring	no	no
		Tasks	recognizing	n/a	n/a
		Sub-tasks	classifying, identifying		
	MIVA	Sub-activity	exploring	no	no
		Tasks	recognizing, specifying	n/a	n/a
		Sub-tasks	classifying, identifying		
	VIE-Visu	Sub-activity	overviewing	no	evaluating
		Tasks	recognizing	n/a	specifying
		Sub-task	aggregating,classifying	n/a	aggregating, classifying
	Lifelines	Sub-activity	understanding	no	investigating
		Tasks	recognizing, specifying	n/a	outlining, summarizing
		Sub-tasks	aggregating, classifying, identifying	n/a	aggregating, classifying, identifying
	VisuExplore	Sub-activity	exploring	no	evaluating
		Tasks	specifying	n/a	recognizing
		Sub-tasks	aligning, identifying	n/a	identifying

Chapter 3

3 Visual Analytics for Electronic Health Records: a Review

This chapter has been submitted to Health Informatics journal.

Please note that the format has been changed to match the format of the dissertation. Figure, Section, and Table numbers mentioned herein are relative to the chapter number. For instance, “Figure 1” corresponds to Figure 3-1. Additionally, when the term “paper”, “work”, or “research” is used, it refers to this particular chapter.

3.1 Introduction

In recent years, medical organizations are increasingly deploying electronic health record (EHR)-based systems that generate, store, and manage their data. Therefore, the amount of data available to clinical researchers and clinicians continues to grow at an unprecedented rate, creating an untapped resource with the capacity to improve the healthcare system (Murdoch & Detsky, 2013). The EHR-based systems are used to detect hidden patterns and trends, monitor patient conditions (Doupi, 2012), reduce medical errors (A. Agrawal, 2009), detect adverse drug events (S. S. Abdullah, Rostamzadeh, Sedig, Lizotte, et al., 2020; Dey et al., 2018), and ultimately improve quality of care (Christensen & Grimsmo, 2008; Rostamzadeh et al., 2020; Tang & McDonald, 2006). However, despite the evidence showing the benefits of EHR-based systems, they rarely improve healthcare experts' ability to make better clinical decisions by having access to more comprehensive information (Heisey-Grove et al., 2014; Lau et al., 2012). Access to large volumes of clinical data has made some analytical and cognitive processes more difficult for healthcare experts. As the amount of data stored in EHRs continues to grow exponentially, and new EHR-based systems are implemented for those already overrun with too much data, there is a growing demand for computational systems that can handle the huge amount of clinical data.

Visual analytics (VA) systems have shown significant promise in addressing information overload challenges in EHRs by combining analytics techniques with interactive visualizations (D. A. Keim et al., 2010; Ola & Sedig, 2014). For a VA system to work well, there must be a strong coupling among all its components (Ribarsky et al., 2009; Sedig et al., 2012). Such components include but are not limited to tasks, interactive visual representations, and analytics techniques. Analytics has the potential to facilitate healthcare experts' clinical decision-making process by using techniques from various fields such as statistics, machine learning, and data mining. Completing analytics, interactive visualizations allow healthcare experts to explore the underlying data, alter the representations, and guide the analytics techniques to accomplish their tasks (Cortez & Embrechts, 2013; D. A. Keim et al., 2015; Vamathevan et al., 2019). VA systems fuse the strengths of both analytics techniques and interactive visualizations to support the execution of EHR-driven tasks. VA is needed to support the intuitive analysis of EHRs for healthcare experts while masking the data's underlying complexity. Clinical researchers can use VA to perform population-based analysis and gain insights from large volumes of patient data. Moreover, VA can also support physicians in tracking symptom evolution during disease progression and creating and visualizing detection models for disease surveillance (Goldsmith et al., 2010; Lo et al., 2013; Perer & Sun, 2012; Rajwan et al., 2013). The complex and diverse challenges and applications of VA in the analysis and exploration of EHRs have led to the development of several EHR-based VA systems, which aim to fulfill the computational and cognitive needs of healthcare experts. The design and development of such systems require collaboration with healthcare experts to assess their requirements and challenges and to better understand EHR-driven tasks from their perspective. This motivates us to systematically study and gather healthcare experts' needs and expectations and get an overview of the state-of-the-art EHR-based VA systems.

The purpose of this paper is to provide a comprehensive review of the state-of-the-art in EHR-based VA systems. We identify the primary dimensions of the EHR-based VA design space through the analysis of the literature. We then use these dimensions along with a characterization of different types of EHR-driven VA tasks to organize the existing systems. Furthermore, we identify the gaps and areas with little prior work, which remains a challenge for future research. To the best of our knowledge, no study has been conducted to review the existing VA systems that have been applied to EHRs. Thus, this review is equipped to help researchers identify which challenges remain insufficiently addressed and understand the primary dimensions that unify the existing work. Finally, the result can provide value to researchers and designers as an organized catalog of various approaches that are most appropriate for EHR-driven VA tasks.

The rest of the review is organized as follows. Section 2 presents the strategy for searching relevant literature and explains the selection criteria. Section 3 provides a brief overview of the EHR-based VA systems that met the selection criteria. In Section 4, we identify and explain the key dimensions of the EHR-based VA design space. Finally, in Section 5, we discuss how the selected EHR-based VA systems support these dimensions and identify the gaps.

3.2 Methods

3.2.1 Search Strategy

We conducted a systematic literature review to retain all the peer-reviewed studies published between 2010 to 2020. We collected all the studies that describe the design, development, and implementation of VA systems that have been applied to EHRs. Search keywords were grouped into three categories: visualization, analytics, and EHR (Table 1). An electronic literature search was conducted in August 2020 using PUBMED, IEEE XPLORE, WEB of SCIENCE, and GOOGLE SCHOLAR. We also utilized the related article function in PubMed on studies that were initially included to identify additional ones. This was supplemented using citation searching. Reference lists from highly relevant studies were also reviewed to find other relevant studies.

Table 3-1: Search terms used to identify studies related to EHR-based VA

KEYWORDS: (K1) AND (K2) AND (K3)	
within each group, the keywords are combined by the "OR" operator	
K1 (Visualization)	Visualization or visual
K2 (Analytics)	Analytics or analysis or data mining or machine learning

K3 (EHR)	EHR or electronic health record or electronic medical record or EMR or healthcare record or patient record or clinical data

3.2.2 Inclusion and exclusion criteria

Articles had to describe the development of VA systems that would be applied to EHRs. We included articles in our review if they met the following criteria: 1) articles must be published in a peer-reviewed journal or conference proceedings; 2) articles must be full papers with empirical evidence; and 3) articles must implement a VA system to support EHR-driven analytical tasks.

Articles were excluded if they were position papers explaining the need for VA, describe medical guidelines, or VA systems designed for administrative tasks with or in relation to patient data (e.g., scheduling and billing). We also excluded articles describing static visualizations because interaction is a key characteristic of VA systems. We also did not include articles on VA of syndromic surveillance, geospatial environmentally aware data, and genetics in our review because we were focused on clinical EHR data. Furthermore, we excluded articles that report the result of abstracts, surveys, feasibility studies, short reports, commentaries, letters, and studies not published in English.

3.2.3 Article selection and Analysis

We collected the authors, journal, title, year of publication, and abstract for each article in an Excel spreadsheet. In the first step, two reviewers screened the title and abstract for each article and eliminated those categorized with exclusion criteria or lacked inclusion criteria. If the reviewers could not assess the article's relevance based on the information provided by its title and abstract, they assessed the full article. In the next step, the full texts of articles that were deemed to be potentially relevant and/or the articles without enough information were reviewed by reviewers. The studies that were cited in eligible articles were also reviewed using a similar screening process. The articles identified for the review were examined by reviewers qualitatively, as described in Section 3.

3.2.4 Results

A total of 1037 references were retrieved from our initial search of electronic databases. A search of the gray literature and hand-searching references from articles resulted in an additional 32 papers. All titles and abstracts were reviewed, with duplicates removed (n = 256). We then excluded 781 articles based on the exclusion criteria. Then the full text of each of the remaining 32 articles was then read; 10 of these articles were excluded since they only described a visualization technique or an analysis technique with static visualization. The results of the screening process in the analysis are noted in the flow diagram in Figure 1. Finally, 22 articles were included in the review.

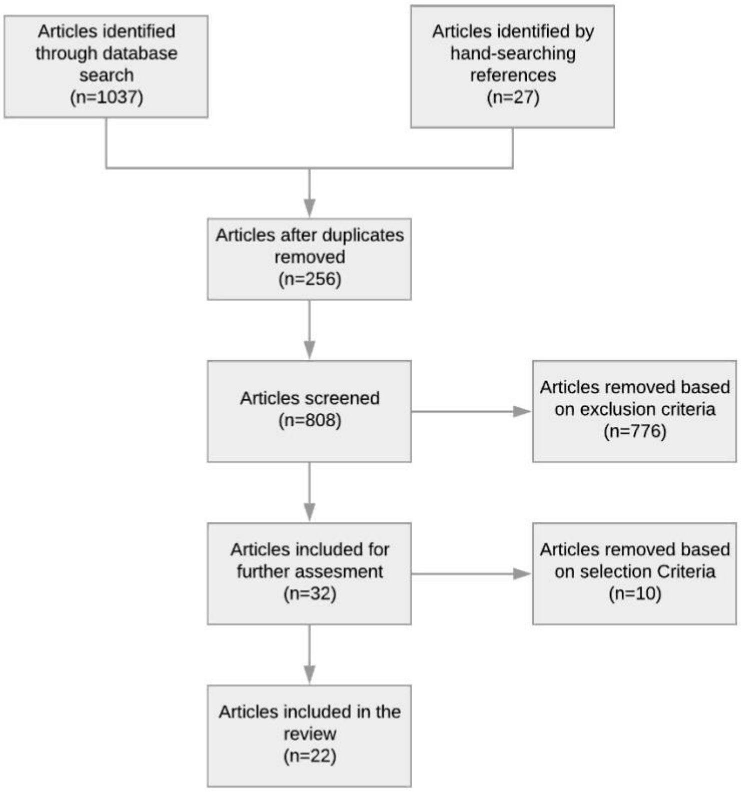

Articles identified through database search (n=1037)
Articles identified by hand-searching references (n=27)
Articles after duplicates removed (n=256)
Articles screened (n=808)
Articles removed based on exclusion criteria (n=776)
Articles included for further assesment (n=32)
Articles removed based on selection Criteria (n=10)
Articles included in the review (n=22)

Figure 3-1: Flow diagram of literature search results.

3.3 EHR-based Visual Analytics Systems

In this section, we provide an overview of the state-of-the-art VA systems that are applied to EHRs. We offer a brief summary of the system's overall goal and its analytics and visualization techniques. We then briefly describe how the system integrates analytical processes with interactive visualizations to help users accomplish their tasks.

3.3.1 Overview of Systems

DecisionFlow (D. Gotz & Stavropoulos, 2014) is a VA system that supports the analysis and exploration of temporal event sequences in high-dimensional datasets. It allows users to test different hypotheses regarding the factors that might affect the patient outcome and compare multiple complex patient event pathways by integrating on-demand statistical analysis techniques with interactive flow-based visualization. DecisionFlow helps users to specify a subsequence of interest with a milestone-based query interface. Then the matching data is aggregated to generate a DecisionFlow graph that contains a linear sequence of nodes (i.e., milestones) connected by directed edges. The system then analyzes the graph to extract multiple statistics (e.g., gender and age distributions and edge summary statistics). The system includes three main linked views-namely, the temporal flow view, edge overview view, and event statists view. The temporal flow view visualizes the DecisionFlow graph using a directed graph of nodes representing milestones where nodes are mapped to grey rectangles and are arranged in temporal order from left to right. The edges that connect these nodes are represented by two marks—namely, the time edges and the link edges, and they are color-coded to encode the average outcome. The edge overview panel summarizes the subsequence of interest that are returned from the query interface by showing multiple aggregate statistics. The event statistic view displays a color-coded bubble chart that represents different edge summary statistics.

RetainVIS (Kwon et al., 2018) is a VA system that assists healthcare experts in the exploration of patient medical records in the context of risk prediction tasks. It provides users with the means to investigate common patterns in a patient's history to identify which medical codes or patient visits (i.e., sequence and timing) contribute to the prediction score. It can also help users to conduct different what-if analysis by testing hypothetical scenarios on patients (e.g., edit/add/remove medical code, alter visit intervals). Furthermore, RetainVIs allows users to provide feedback to the model based on their domain knowledge if the model acts in an undesirable manner. RetainVIS generates prediction scores based on the RetainEX technique, a bidirectional recurrent neural networks (RNN) model that harnesses the temporal information stored in patient records (e.g., time intervals between patient visits). It increases the interpretability and interactivity of models by calculating code-level and visit-level contribution scores.

This system integrates RetainEX with multiple interactive visualizations. The Overview summarizes patients regarding their contribution scores, medical codes, and predicted diagnosis risks using a scatter plot, multiple bar charts, an area chart, and a circle chart. Patient Summary shows a temporal summary of the selected patients. It contains a table, a code bar chart, and a contribution progress area chart. Patient Summary provides a summary description of the selected patients and represents aggregated contribution scores of medical codes over time and their mean contribution scores. Patient List shows selected patients in a row of rectangles. It allows users to compare and explore multiple patients and select a patient of interest to view their details in the Patient Details view. Patient Details view is composed of a line chart of prediction scores, a temporal code chart of contribution scores of medical codes, and a code bar chart representing the most contributing medical codes for each patient. Finally, Patient Editor represents each patient visit horizontally in a temporal order and lists medical codes for each visit downwards. It allows users to test hypothetical scenarios by changing the date of the visit or inserting new medical codes into a visit. Once the user changes are complete, the system generates the new model and returns the new predicted risk and contribution scores on top of the original records.

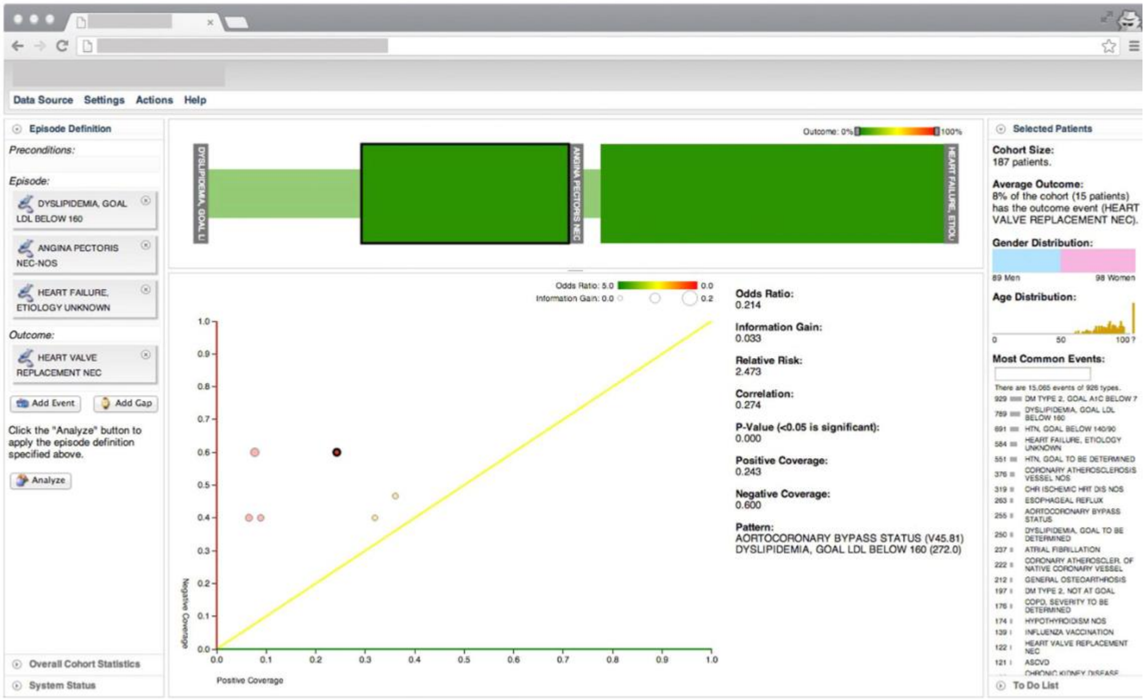
DPvis (Kwon et al., 2020) is a VA system that supports clinical researchers in interactively discovering and exploring disease progression patterns and studying interactions between such patterns and patient's characteristics. It also allows users to test and refine hypotheses for multiple clinically relevant subgroup cohorts in an ad hoc manner. DPVis models disease progression pathways by characterizing a patient's clinical course as a sequence of transitions between multiple states where each state describes a co-occurring pattern of observed symptoms and variables. Then, it uses a class of unsupervised models, namely- continuous-time hidden Markov models (CT-HMMs), to discover these hidden states and state transitions from large-scale longitudinal patient records. These models identify associations between disease progression patterns and various observed variables and predict a patient's future states. DPVis combines the outcome of HMM models with interactive visualizations to assist medical experts in interpreting these models and clinically make sense of the discovered patterns.

DPVis is composed of seven linked views. The Static Variable Distribution view contains a list of selected measures in a horizontal bar chart. The Observed Attributes view contains feature matrix, feature distribution, feature heatmap over time, and feature over time. It summarizes all the characteristics of disease states that are discovered by HMM. State Transitions view shows multiple representations of state-to-state transition patterns over a series of visits or over time. It includes four views-namely, Pathway over Observation, Pathway by Time Unit, Pathway Waterfall, and State Transition Chord Diagram. Frequently Occurring State Transition Pattern view shows a list of frequently occurring state sequential patterns. Subject Timeline represents an individual patient's observations over time. It contains Dual Kernel Densities view and Subject List view. State Sequence Query Builder allows users to create and refine cohorts based on state transitions. Cohort view enables users to load and save intermediate results. Once users create more than two cohorts in the Cohorts view, they can trigger the Comparison Mode between the selected views. This selection then turns all views into the Comparison Mode.

The VA system for pharmacovigilance (i.e., drug safety) in electronic medical records developed by **Ledieu et al.** (Ledieu et al., 2018) integrates a modified version of the Smith-Waterman (SW) sequence alignment algorithm with an interactive web interface to detect inappropriate drug administration and inadequate treatment decisions in patient sequences. The SW algorithm is used to compare a reference sequence (i.e., a sequence specified by the user) and a patient's sequence, where each sequence is considered a string of characters. Each character in the sequence represents a clinical event, such as a laboratory test result or a drug administration. The algorithm calculates a similarity score for each comparison. A high similarity score corresponds to a higher similarity between the reference and the patient sequences. This VA system allows users to create the reference sequence(s) in a query interface. It provides them with a visual dictionary of event types (e.g., the discretized numerical events are encoded by color-coded squares or the direction of arrows represents the trend of change) in a grey rectangular area. To form a pattern, users can drag and drop these icons down to a query line. The system also enables the user to indicate time-constraint events in the query. The adopted SW algorithm returns the search result, which is displayed as a list of patients and their corresponding sequences, sorted based on their similarity score to the reference sequence. Each sequence is aligned to the reference pattern or its closest match. The time interval between the time-constraint event and the aligned events is shown by a vertical line along with the time duration in days on top of it.

Gotz et al. (D. Gotz et al., 2014) develop a VA system to explore and query clinical event sequences stored in EHRs by combining on-demand analytics with visual queries and interactive visualizations (Figure 2). The visual query module provides an intuitive user interface that enables users to retrieve cohorts of patients that satisfy complex clinical episode specifications. Users can define a clinical episode by specifying milestones, time gaps, preconditions (i.e., a set of constraints that should be satisfied before the starting milestone), and outcome measures in the query interface. Upon submission of the query, the system returns a set of matching patient event sequences. The returned event sequence for each patient includes the specified milestones and several intermediate events that occur between milestones. Each episode is subdivided into a series of intermediate episodes at each milestone.

Frequent pattern mining (FPM) is then performed first on the overall episode as well as on each of the intermediate episodes that are retrieved by the visual query module. The FPM engine includes two main components-namely, the frequent pattern miner and the statistical pattern analyzer. The frequent pattern miner uses the bitmap-based Sequential PAttern Miner (SPAM) (Ayres et al., 2002) algorithm for pattern discovery. SPAM employs a search strategy that combines a depth-first traversal of the search space with an



Click the "Analyze" button to apply the episode definition specified above.

Odds Ratio:
0.214

Information Gain:
0.033

Relative Risk:
2.473

Correlation:
0.274

P-Value (<0.05 is significant):
0.000

Positive Coverage:
0.243

Negative Coverage:
0.600

Pattern:
AORTOCORONARY BYPASS STATUS (V45.81)
DYSLIPIDEMIA, GOAL LDL BELOW 160 (272.0)

Cohort Size:
187 patients.

Average Outcome:
8% of the cohort (15 patients) has the outcome event (HEART VALVE REPLACEMENT NEC).

Most Common Events:

There are 15,065 events of 926 types.

929	DM TYPE 2, GOAL A1C BELOW 7
789	DYSLIPIDEMIA, GOAL LDL BELOW 160
691	HTN, GOAL BELOW 140/90
584	HEART FAILURE, ETIOLOGY UNKNOWN
551	HTN, GOAL TO BE DETERMINED
376	CORONARY ATHEROSCLEROSIS VESSEL NOS
319	CHR ISCHEMIC HRT DIS NOS
263	ESOPHAGEAL REFLUX
255	AORTOCORONARY BYPASS STATUS
250	DYSLIPIDEMIA, GOAL TO BE DETERMINED
237	ATRIAL FIBRILLATION
222	CORONARY ATHEROSCLER. OF NATIVE CORONARY VESSEL
212	GENERAL OSTEOARTHROSIS
197	DM TYPE 2, NOT AT GOAL
176	COPD, SEVERITY TO BE DETERMINED
174	HYPOTHYROIDISM NOS
139	INFLUENZA VACCINATION
122	HEART VALVE REPLACEMENT NEC
121	ASCVD
	CHRONIC KIDNEY DISEASE

Figure 3-2: The screenshot of the VA system developed by Gotz et al. (D. Gotz et al., 2014) including, the visual query panel, the milestone timeline, the cohort overview, and the pattern diagram. Source: Reprinted from Journal of Biomedical Informatics, 48, Gotz D, Wang F, Peter A, A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data, 148-159., Copyright (2014), with permission from Elsevier.

efficient pruning mechanism. It takes a set of event sequences and a user-specified support as inputs and returns a set of frequent patterns as an output. Then the statistical pattern analyzer computes correlations (e.g., Pearson correlation, odds ratio, and information gain) between the mined patterns and the outcome measure. Finally, an interactive visualization allows users to explore the results and discover temporal patterns. The interactive visualization component is composed of three linked views. The cohort overview shows the age and gender distributions for patients that satisfy the query specifications. The milestone timeline represents the sequence of milestones using a series of ordered, vertical grey bars. The bars are connected by color-coded edges, where each edge has two parts-namely, the time edge, and the link edge. The time edge maps the mean duration between the milestones while the link edge connects the bars to show sequentially. The pattern diagram shows the set of patterns mined from the part of the episode that is selected in the milestone timeline in a scatter plot where the x and y axes encode the level of support for a specific pattern for patients with positive and negative outcomes, respectively.

The VA system developed by **Simpao et al.** (A. F. Simpao et al., 2014) facilitates the dynamic and continuous monitoring of medication alerts and care providers' responses through an automated, user-friendly dashboard. It allows pharmacists and care providers to examine and filter the alert data based on patient location and ordering provider type and to identify which specific orders triggered the drug-drug interaction alerts. This VA dashboard is an integral part of a hospital quality improvement initiative to improve medication safety and reduce alert fatigue by deactivating irrelevant alert rules. The system is developed in collaboration with a clinical decision support committee that is asked to perform three interventions to deactivate irrelevant drug-drug interaction alert rules. The impact of these interventions on pharmacists' alerts and override rates is analyzed using an interrupted time-series framework with piecewise regression. Baseline IQRs and median rates are compared to IQRs and median rates following three intervention phases of drug-drug interaction deactivations and are tested for statistical significance using the Wilcoxon rank-sum test. The user interface of this system includes a central display area with graphical and tabular data representations. Medication alert and override rates, different alert types, and various care providers, and patient characteristics are displayed and explored at a specific time point or across a user-defined time interval using multiple filters and limits.

The MOSAIC dashboard system (Arianna Dagliati et al., 2018) aims to support the prediction and diagnosis of type 2 diabetes mellitus (T2DM) by analyzing clinical and home monitoring data. The system integrates data querying and mining technique with an interactive user interface to assist caregivers in devising management strategies and therapeutic interventions for T2DM complications. The mining techniques are triggered by the query module that is responsible for retrieving the data from the i2b2 data warehouse, calling the proper data mining technique, and sending the results back to the user interface. The data mining module implements several temporal analytics models such as temporal abstractions, the care flow mining algorithm, drug exposure pattern detection, and risk prediction models for T2DM complications. Temporal abstractions are extracted using the Time Series Abstractor (JTSA) tool that provides a library of techniques that can be employed for time-series processing and abstraction (Sacchi et al., 2015). The care flow mining technique uses the temporal sequence of events to determine the most frequent clinical pathways patients experience during their care process, automatically generating groups of patients with similar care histories (A. Dagliati et al., 2017). The proportion of days covered is used to summarize the dug purchase patterns using the data gathered from administrative resources. Finally, several risk prediction models are generated to estimate the risk of T2DM complications (Arianna Dagliati et al., 2018).

The graphical user interface of MOSAIC has two primary components designed for 1) clinical decision support and 2) outcome assessment on populations of interest. The clinical decision support system dashboard is composed of three sections-namely, metabolic control, frequent temporal patterns, and drug purchase patterns. The metabolic control evaluation section is based on a "traffic light" metaphor to enable quick assessment of the control level of certain parameters. The frequent pattern mining section is composed of a scatter plot and a timeline plot. The drug purchase graph shows all the purchases made by a patient for each drug class using a scatter plot. The outcome assessment dashboard provides an overview of the treatments' outcomes on the population of patients with T2DM to clinical researchers. It includes summary charts that represent patient counts grouped by clinical and demographic variables. It also shows the most frequent temporal patterns of the patients selected in the summary chart using timeline graphs.

VisualDecisionLinc (Mane et al., 2012) is a VA system that helps clinicians to identify subpopulations of patients with similar clinical characteristics and to understand the risks and effectiveness of different treatment options for these subpopulations using psychiatric patients' data with major depressive disorder (MDD). The system aims to improve and simplify the decision-making process by reducing the number of available therapeutic options to those that have proven to be most effective with minimal side-effects. To define the MDD comparative population, VisualDecisionLinc uses a patient data-driven approach where the patient's medical profile is used as 'seed' data (i.e., patients with a primary diagnosis of MDD and their last prescribed medications) to identify a comparable group of patients with similar clinical characteristics. At the computational level, the system creates a bin for each medication and inserts patients into bins of their prescribed medication. At the same time, the system tags patients based on their treatment outcome response, which is reported in the database in the form of a clinical global impression (CGI) score. CGI score is a 7-point scale that offers a brief score of the clinician's assessment of the severity of the patient's illness prior to and after starting treatment. A lower CGI score indicates a better treatment outcome for the patient. After the binning process is done, the system uses additional computational processes to quantify the collective comparative MDD patient response into a '% Patient Improved' score.

VisualDecisionLinc is composed of five linked views. Data view of patient demographics shows patient demographic data such as age, gender, and race, to name a few. Data view of summarized medication response displays '% Patient Improved' score and the absolute number of patients that are used to compute this score. Color-coded dots placed next to the medication names encode the '% Patient Improved' score greater than 10 (green dots) and less (red dots). Data view of comorbidities shows a list of comorbid conditions among patients on a selected medication from the summarized medication view. Data view of contextual patient treatment outcome shows the CGI score of a patient over time. It also displays prescribed medications and their timespan using horizontal bars below the CGI temporal view. Finally, the data view of median-based historical response to medication shows the historic outcome response to the selected medication. Blue and red lines reflect the median-based historical trend in medication outcome from the comparative populations and patient's response to the selected medication in the past, respectively.

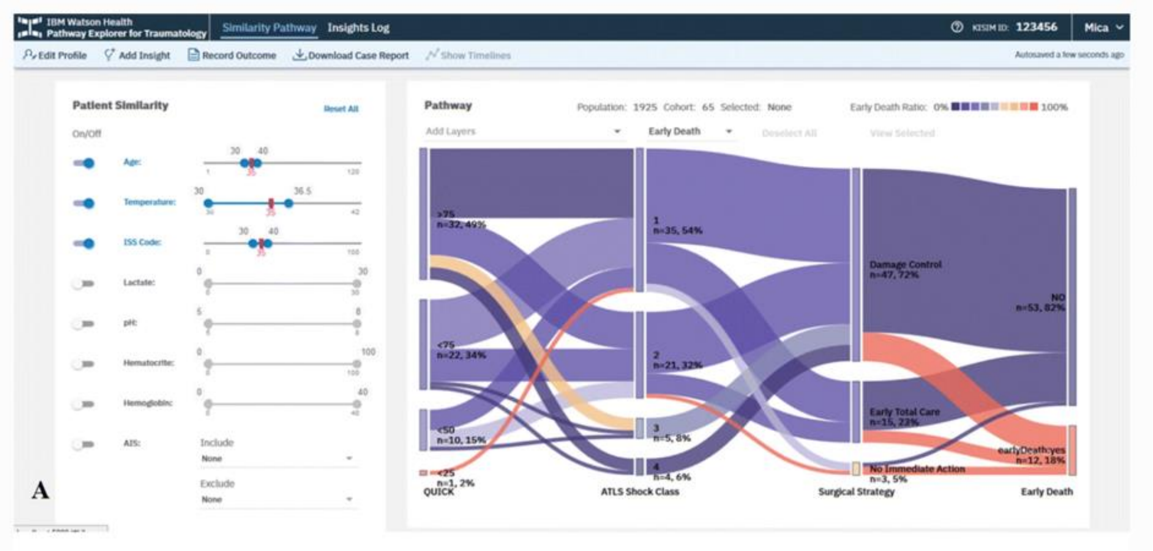
Care Pathway Explorer (Perer et al., 2015) is an interactive hierarchical information exploration system that can help physicians analyze patients’ longitudinal medical records. The system provides an overview of the frequent patterns that are mined from patient event sequences. The physician then studies these patterns and interactively selects patterns of interest for more details. The system computes the group of patients that match the physician's specified sub-traces. Then the event traces for those patients are extracted using a deeper level of the user-specified hierarchy. The system feeds these traces to the frequent pattern miner engine, which mines frequent patterns and analyzes how these patterns are associated with outcomes using a modified version of the SPAM algorithm (Ayres et al., 2002). The patterns are then visualized alongside meaningful statistics.

The visual interface of Care Pathway Explorer features two complementary views. The overview contains a bubble chart and represents events of the most frequent patterns mined by the frequent pattern miner engine. Each bubble encodes a medical event that occurs frequently among patients and is computationally positioned close to events with which it most frequently occurs to show an overview of clusters of patterns. The flow view shows how bubbles connect to each other using a visualization similar to the Sankey diagram. Events in the most frequent patterns are encoded by nodes, and event nodes belonging to the same pattern are connected by edges. Both bubbles and patterns are color-coded according to their association with the outcome, which is determined by the Pearson correlation.

RegressionExplorer (Dingen et al., 2019) is an interactive VA system that enables clinical researchers to quickly generate, compare, and evaluate many regression models. It also helps to formulate new hypotheses and steer the development of models by allowing the user to compare candidate models across several subpopulations. Upon loading the dataset and selecting the appropriate responder that captures the condition of interest, the system allows the researcher to analyze the one-to-one relationships between each covariate and the responder by performing a univariate analysis. The results are displayed as colored rectangles next to the variable names in the univariate analysis view. The significance level of an effect is determined using p-value, where a lower p-value results in a higher level of significance and a more saturated color. Red represents a positive effect, while blue represents a negative effect. Next, the system allows the user to perform stepwise multivariate analysis by dragging variables from the list of variables to the variable selection view. After each selection, the system generates a new model displayed as a single row of the multivariate model matrix. Columns in the matrix show the levels of significance for the included covariates following the same convention as for the univariate view. The system also displays histograms, along with some basic descriptive statistics for all the covariate distributions to provide basic checks and interpretation during analysis.

Another integral part of the RegressionExplorer is subgroup analysis that allows the user to gain more insight into the subpopulations throughout the univariate and multivariate analysis. To support subgroup analysis, the system enables the user to drag and drop a variable from the univariate analysis view to the population view, which leads to the partition of the population. If the user drops another variable into the population view, all the previously created subpopulations are partitioned recursively. The subpopulation tree is represented as an icicle plot. The system follows the same basic approach for both univariate and multivariate analysis when handling subpopulations. The primary difference is that the cells that used to show significant effects are now subdivided into sub-cells (i.e., icicle plots).

The VA system developed by **Mica et al.** (Mica et al., 2020) helps guide patient assessment and therapeutic decisions for physicians using severely injured patients' clinical data in a trauma center (Figure 3). The system allows the user to filter cohorts of patients based on multiple parameters, including age, body temperature, ISS, multiple lab results, and AIS score. With every change of the filtering criteria, a query is sent to the server to extract a group of patients that satisfy the query specifications using several algorithms such as statistical frequency grouping, time interval simplification, and consecutive event merging. The system enables the user to explore the results using a variation of the Sankey diagram. Each node in the graph encodes a medical state (e.g., treatment or outcome), and each link encodes transitions between consecutive states in the cohort of interest. The height of nodes and links represents the relative number of patients that share the state and transition, respectively. The color encodes the ratio of patients that develop the outcome of interest. Statistically, to justify the distribution of


KISIM ID: 123456

Population: 1925 Cohort: 65 Selected: None

Age: 30 40 (35)

Temperature: 30 36.5 (35)

ISS Code: 30 40 (35)

AIS: Include None Exclude None

Figure 3-3: The screenshot of the VA system developed by Mica et al. (Mica et al., 2020) shows the pathway of the early death outcome of a hypothetical patient with an age of 35 years, an ISS of 35, and a temperature at admission of 35 °C using a Sankey diagram. Source: Reprinted by permission from Springer Nature: [Springer] [World Journal of Surgery [Development of a Visual Analytics Tool for Polytrauma Patients: Proof of Concept for a New Assessment Tool Using a Multiple Layer Sankey Diagram in a Single-Center Database, Mica, L.; Niggli, C.; Bak, P.; Yaeli, A.; McClain, M.; Lawrie, C.M.; Pape, H.-C.], Copyright (2020).

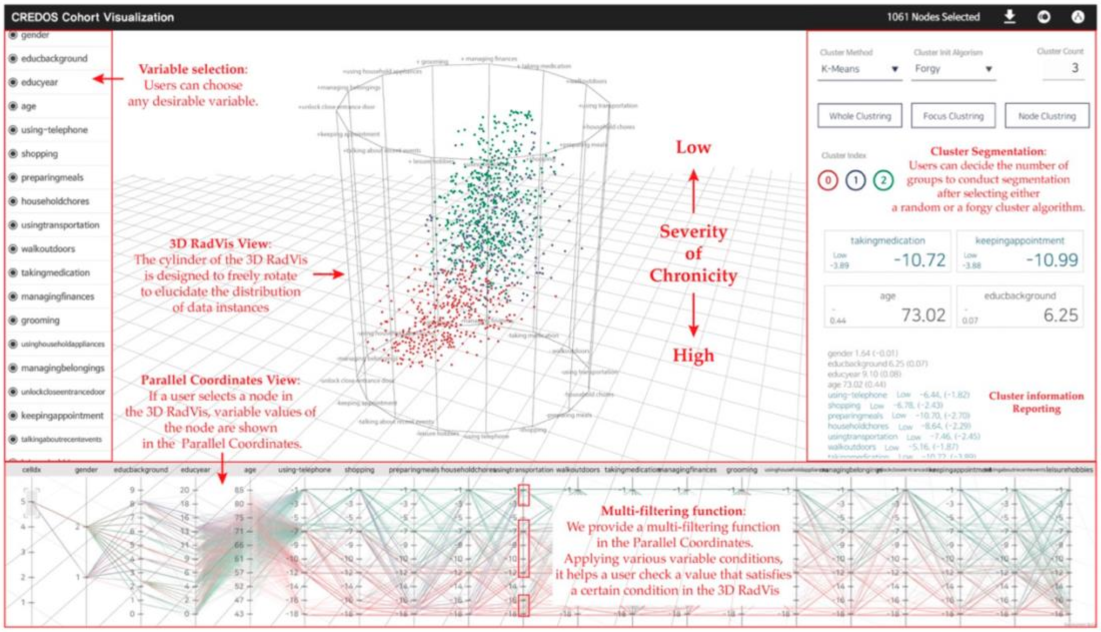
patients based on clinical scores, the system integrates binary logistic retrogression along with receiver operating characteristic (ROC).

Visual Temporal Analysis Laboratory (ViTA-Lab) (Klimov et al., 2015) integrates temporal data mining techniques with query-driven interactive visualizations to support a knowledge-based exploration of time-oriented clinical data and the discovery of interesting patterns within it. ViTA-Lab is composed of three main interfaces. The main visualization interface provides an overview of the longitudinal concepts and the distribution of derived temporal abstractions (TA) for individual and multiple patients at different temporal granularities. It provides the user with a knowledge-based browser and a graphical widget for selecting an individual patient or a group of patients. It uses a scatter diagram over time and a modified version of the bar chart visualization technique to show the distribution of TAs and help the user discover trends in these distributions.

The temporal association chart (TAC) allows visual exploration and discovery of probabilistic temporal associations among the distributions of various abstract concepts at different times. TAC's input is a group of patients and a set of concepts that are chosen within the same or a different time window panel. The system calculates the distributions of values for each concept within the chosen time. Each concept is represented by a rectangular bar. The corresponding data values between two consecutive concepts for each patient are linked. Multiple links, including the same pair of values for a group of patients, are aggregated into a temporal association rule. This rule indicates the probability of having the second concept's value, given the first concept's value, and the total frequency of that combination. Thus, a group of patients who have this specific combination of values from two concepts, simultaneously or at different times based on the user-specified time period, is represented by a temporal rule.

The pattern explorer supports the exploration of temporal patterns that are discovered by data-driven computational techniques. It works based on a version of the KarmaLego algorithm, which is used for the discovery of frequent temporal patterns (Moskovitch & Shahar, 2015a, 2015b). Components of the output’s temporal pattern (a pair of concept and value) are represented by horizontal lines that are ordered according to each component's start time, maintaining, in a proportional fashion, the mean duration of each component and of the time gaps among components. The color of the same type of component in all patterns stays the same. The pattern explorer allows the user to recognize the meaning of a temporal pattern, that is, which components make up the pattern, and what temporal associations such as overlaps, before, or after hold between them.

RadVis (Ha et al., 2019) is a VA system that supports psychiatrists in analyzing and exploring multidimensional medical datasets for patients who have dementia (Figure 4). It allows the user to get a better understanding of the characteristics of patient clusters and analyze the variable values of data comprising each cluster at the same time. The



Variable selection: Users can choose any desirable variable.

3D RadVis View: The cylinder of the 3D RadVis is designed to freely rotate to elucidate the distribution of data instances

Parallel Coordinates View: If a user selects a node in the 3D RadVis, variable values of the node are shown in the Parallel Coordinates.

Cluster Segmentation: Users can decide the number of groups to conduct segmentation after selecting either a random or a forgy cluster algorithm.

Multi-filtering function: We provide a multi-filtering function in the Parallel Coordinates. Applying various variable conditions, it helps a user check a value that satisfies a certain condition in the 3D RadVis

gender 1.64 (-0.01)
educbackground 6.25 (0.07)
educyear 9.10 (0.08)
age 73.02 (0.44)
using-telephone Low -6.44, (-1.82)
shopping Low -6.78, (-2.43)
preparingmeals Low -10.70, (-2.70)
householdchores Low -8.64, (-2.29)
usingtransportation Low -7.46, (-2.45)
walkoutdoors Low -5.16, (-1.87)

Figure 3-4: The screenshot of RadVis (Ha et al., 2019) combing 3D RadVis and parallel coordinates. Source: image used under CC-BY 4.0 License.

system enables the user to select variables of interest from "Variable Selection Menu" and select "Cluster Segmentation Menu" to segment clusters of patients based on their traits. The user can choose the number of clusters for segmentation after selecting either a forgy cluster or a random cluster algorithm. Following either of the clustering algorithms, the cluster's central value is calculated based on the number of clusters. After the Euclidean distance between the central value and each node is calculated, multiple nodes are included to obtain clusters of similar value. This process is repeated until the central value stays constant.

RadVis displays the distribution of data instances using 3-dimensional radial coordinate visualization (3D RadVis) that prevents node overlap. Furthermore, it facilitates the distribution of several nodes into optimum positions regardless of the number of dimensions. A patient with dementia is represented by a single node in this visualization. Nodes are color-coded according to the cluster they belong to. RadVis also supports a multi-filtering function through parallel coordinates plot to assign different conditions for a more comprehensive analysis. The parallel coordinates plot is used to display both categorical and numerical variables. It allows the user to check a value that satisfies a specific condition in the 3D RadVis. It also displays the variable values of a node that is selected in the 3D RadVis.

The predictive VA system developed by **Sun et al.** (J. Sun et al., 2014) aims to predict the risk and timing of deterioration in hypertension control using EHRs. The system is composed of three main modules. The feature engineering module converts clinical data into a feature matrix and a target label vector that can be used to build the predictive model. The target label is derived based on the physician's assessment of blood pressure control status as in-control (i.e., positive) versus out-of-control (i.e., negative). The positive and negative transition points (from an episode of positive (negative) assessment points into negative (positive) points) are considered as target labels for the prediction model. Next, to turn event sequences into feature variables, the system specifies an observation window for each feature concept (e.g., diagnosis concept). It then aggregates all the events of the same feature concept within the observation window into a single value. The system then applies a two-level feature selection process. In the first level, within the same concept, features are chosen based on the information gain. Then a greedy forward selection algorithm is used to choose which concepts to keep. In the next step, the system starts iteratively combining features from different concepts until the combination fails to improve the performance of the prediction. Finally, various techniques, such as naive Bayes, logistic regression, and random forests, are used to generate transition point models. The system allows the user to explore the prediction results and other events through interactive visualization. An individual patient's timeline is represented by a line, and each hypertension control assessment event is represented by a circle. Red and blue circles represent in-control and out-of-control blood pressure episodes, respectively.

The VA system developed by **Guo et al.** (Guo et al., 2020) helps clinicians to explore medical records from both multivariate and temporal perspectives and identify and analyze similar records (Figure 5). The system integrates an unsupervised learning-based technique with interactive linked views to support physicians in several tasks such as finding similar records based on a focal patient record, comparing patients' medical

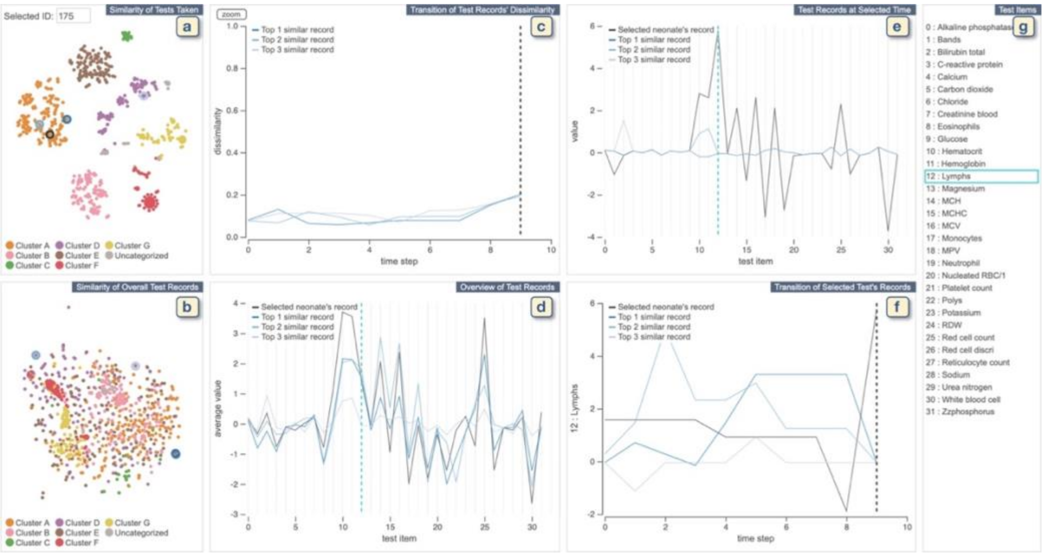


Figure 3-5: The screenshot of the VA system developed by Guo et al. (Guo et al., 2020). (a) and (b) display each neonate's similarities of tests taken and the records of test values, respectively. (c) shows changes of dissimilarities of test records over time between the neonate chosen from (a) or (b) and top-3 similar neonates. (d) displays a statistical overview of the chosen neonate and top-3 similar neonates. (e) provides all the test results at the selected time in (c) or (f). (f) displays the temporal changes of values of the chosen test in (d) or (e). (g) lists all medical test names. Source: Reprinted from Journal of Visual Informatics, 4, Guo, R.; Fujiwara, T.; Li, Y.; Lima, K.M.; Sen, S.; Tran, N.K.; Ma, K.-L., Comparative visual analytics for assessing medical records with sequence embedding, 72-85., Copyright (2020), with permission from Elsevier.

feature values at a specific time point or identifying (dis)similar time stamps among similar records. The system provides two overviews of all patients: One is for patients' similarities according to the combination of tests taken during the collected time period, and the other view shows patient's similarities according to the test values. To create the first overview, the system applies the Jaccard index (Gower & Warrens, 2014) to compute the similarity. Then it extracts clusters of similar patients by combing a dimensionality reduction (DR) technique (i.e., t-SNE) and a density-based clustering method (i.e., HDBSCAN). For the second overview, the system first calculates the similarity of each pair of the test records and then similar to the other overview; it applies t-SNE to visualize the similarity relationships. To visualize the clustering information, each point (i.e., each patient's record) is colored based on the assigned cluster-ID. The system allows the user to select a patient of interest from these overviews. It then automatically searches for the top-3 similar patients based on the pre-computed similarities. The system uses autoencoder-based event embedding (Kramer, 1991) and sequence to sequence learning (seq2seq) (Sutskever et al., 2014) technique to handle various event types and convert records with different lengths to vectors of the same length. Then, it computes the similarity of each pair of patients using a certain distance metric, such as the Euclidean distance. The system provides multiple line charts to show changes of dissimilarities of test records over time between the patient of interest and top-3 similar patients and to visualize a statistical overview of the focal and top-3 similar patients.

SubVIS (Hund et al., 2016) is a VA system to support medical experts in interpreting high-dimensional clinical data and exploring subspace clusters from different perspectives (Figure 6). It enables the user to analyze each subspace independent of its association to a certain clustering technique. It allows the use of every subspace clustering technique available at OpenSubspace Framework (Müller et al., 2009). SubVis allows a three-level exploration of data and clusters through its interface. The first level provides the user with a general overview of all the detected subspace clusters, their properties, and the distribution of dimensions within each subspace cluster using interactive bar charts. A matrix-based heatmap is also available to give more details on the association between the pair-wise distance. The second exploration level allows the user to choose a subset of relevant clusters in the multidimensional scaling (MDS) (Cox & Cox, 2008) plot to get an aggregated overview of the cluster members in an aggregation table. The distance between various clusters in the MDS plot shows their pair-wise similarity. SubVis contains various similarity measures, such as Overlapping, Jaccard Index, and Dice Coefficient. The system enables the user to inspect the distribution of the cluster members in every dimension for each cluster. In the last exploration level, a table-lens-like view (Rao & Card, 1994) supports the exploration of

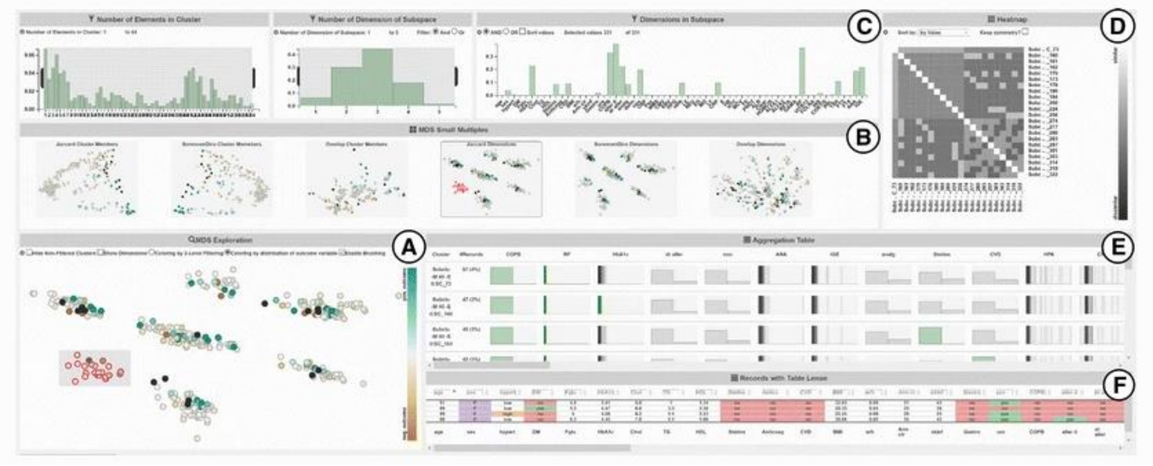
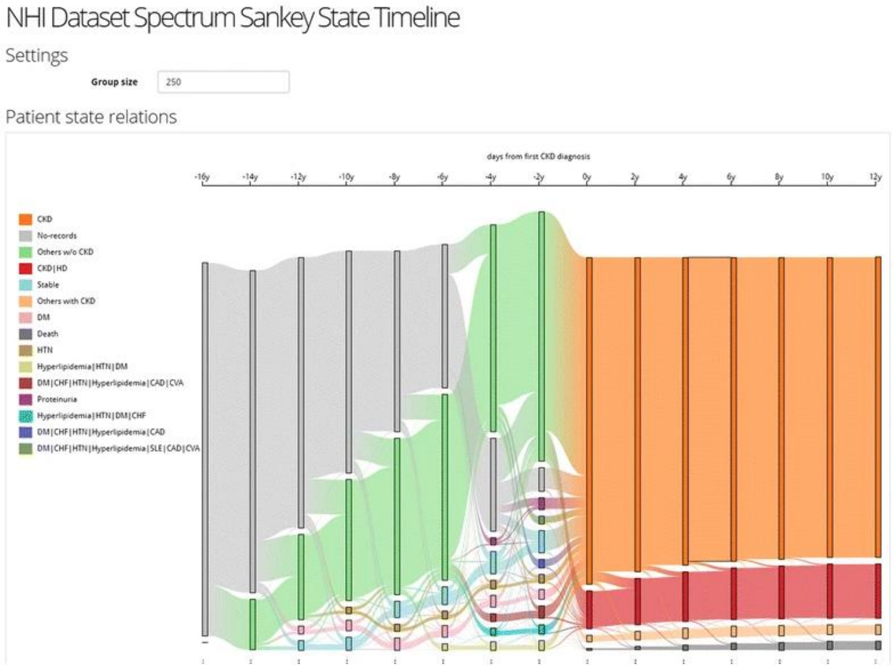


Figure 3-6: A screenshot of SubVIS (Hund et al., 2016) including (A) MDS projection plot, (B) MDS small multiples, (C) barcharts showing the distribution properties of the subspaces, (D) heatmap, (E) aggregation table, and (F) table lens. Source: image used under CC-BY License.

the actual data records and provides interactive coloring and sorting of the record and its dimension.

The VA system developed by **Huang et al.** (Huang et al., 2015) supports the interactive exploration of patient trajectories to assist physicians and clinical researchers in identifying chronic diseases and determining how a group of patients with chronic diseases might go on to develop other comorbidities over time (Figure 7). The system first aligns patient trajectories based on the time they are diagnosed with a specific chronic disease. Then once the user specifies the time windows, the patient trajectories are divided based on their timestamps, and patients within the same time window are aggregated into one. The system then clusters the patient records at each time window based on a similarity measure and create a set of cohorts. The system supports frequency-based cohort clustering and hierarchical cohort clustering techniques. A cohort of patient trajectory network is built based on the clustering result where each node represents a cohort at a time window, and each edge shows the relationship between two cohorts at consecutive time windows where their members overlap. The system allows the user to filter edges using the variance-based association filtering technique by adjusting the entropy threshold. When the threshold is zero, only associations between fully



NHI Dataset Spectrum Sankey State Timeline

Settings

Group size: 250

Patient state relations

days from first CKD diagnosis
-16y
-14y
-12y
-10y
-8y
-6y
-4y
-2y
0y
2y
4y
6y
8y
10y
12y
CKD
No-records
Others w/o CKD
CKD|HD
Stable
Others with CKD
DM
Death
HTN
Hyperlipidemia|HTN|DM
DM|CHF|HTN|Hyperlipidemia|CAD|CVA
Proteinuria
Hyperlipidemia|HTN|DM|CHF
DM|CHF|HTN|Hyperlipidemia|CAD
DM|CHF|HTN|Hyperlipidemia|SLE|CAD|CVA

Figure 3-7: The screenshot of the VA system developed by Huang et al. (Huang et al., 2015) shows the result of frequency-based cohort clustering using a Sankey diagram. Source: image used under CC-BY 4.0 License.

overlapped cohorts are shown; in the case when the threshold is high, all associations are visualized. A Sankey-like timeline then visualizes the output results. The nodes are color-coded based on the unique comorbidities, and the color of the edges is determined by the two nodes it connects (i.e., a gradient for smooth transitions). Each cohort has a label that shows its dominant features. In addition, the cardinality of both nodes and edges are represented by their height.

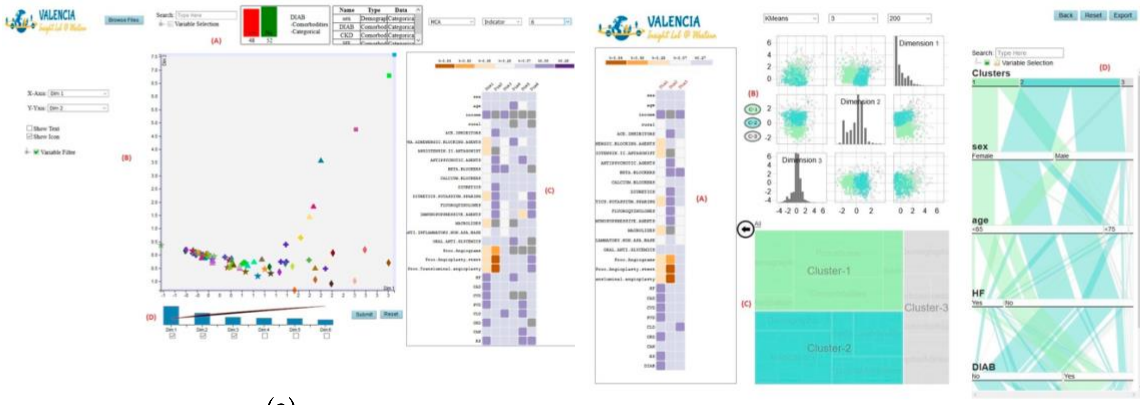
CarePre (Jin et al., 2020) is a clinical decision assistance system that supports the exploration and interpretation of deep learning prediction models that are developed to predict future diagnosis events for a focal patient based on their medical background. It assists physicians in making more informed decisions by letting them analyze contributing factors in prediction results and explore the outcomes of possible treatments through interactive visualizations. CarePre allows the physician to input potential diagnoses (based on the patient's symptoms and tests) for a focal patient into the system. The system then automatically estimates the risk of future diseases for the patient based on their medical history using a state-of-the-art deep learning technique and allows the physician to explore the results and the details of the historical medical records in the prediction view. The prediction view shows the patient's event sequence leading up to the time point of prediction, which is represented by rectangular nodes arranged horizontally in order of their occurrence. The predicted likelihood of each diagnosis is also displayed as a series of rectangular nodes where the color saturation for each node shows the prevalence of the predicted diagnosis across the records for a population of similar patients.

In the next step, the physician can specify a query to retrieve a group of similar patients to help interpret the prediction results. CarePre measures similarity between sequences by computing the similarity between each pair of events using the Euclidean distance of the corresponding event vectors. It then displays event sequence data for the focal patient as well as a group of similar patients. It also aggregates the event sequences for similar patients into a flow-based visualization to allow a one-to-many comparison between the focal patient and a group of similar patients and to show the overall evolution of treatments and diseases over time. Lastly, the physician can explore alternative treatment plans and identify the key factors that contribute to the prediction result through various interactions such as editing the focal patient's events (e.g., adding events, changing the order) in the prediction view and comparing the edited event sequence in the outcome analysis view.

Peekquence (Kwon et al., 2016) is a VA system that aims to make the frequent sequence mining results more interpretable by allowing the user to explore the patterns by ranking them based on their variability or correlation to the outcome. It can also integrate patterns with a patient timeline to help the user understand where the patterns occur in the actual data. Peekquence uses the SPAM (Ayres et al., 2002) frequent sequence mining algorithm to detect the most frequent sequences. The system uses four linked views to visualize the result of SPAM on the patient's medical records. All the views use an event glyph to visualize the event sequences. The event glyph represents each unique event type appearing in the mined patterns by a circle and is color-coded based on a categorical ontology. These event glyphs are labeled with an abbreviation of the name of the event type. The sequence network view displays the frequency of co-occurring events within patterns that are mined using SPAM. The event types are represented by the nodes, and the two co-occurring nodes within patterns are connected by an edge. The pattern list view displays all the mined patterns, aligned vertically. Each row represents a pattern that is visualized as a sequence of circular event glyphs. Furthermore, the association of the patterns with the outcome is represented by the stacked bar chart next to the sequence. The event co-occurrence histogram view shows the frequency of co-occurring events with a selected pattern from the pattern list view. Each event type is represented by a bar partitioned into three blocks to show events occurring before, within, and after the chosen pattern. Lastly, the timeline view displays the patient's event sequences aligned according to the selected pattern.

PHENOTREE (Baytas et al., 2016) is a hierarchical and interactive phenotyping VA system that allows physicians to participate in the phenotyping process of large scale patient records. It enables the user to explore patient cohorts, and to create, interpret, and evaluate phenotypes by generating and navigating a phenotype hierarchy. The system uses the sparse principal component analysis (SPCA) to identify key clinical features that describe the population given a cohort or sub-cohorts of patients. These key clinical features are used to build deeper phenotypes at finer granularities by expanding the phenotype hierarchy. Patients that are associated with each key feature are grouped into individual sub-cohorts. The system then iteratively applies the SPCS to each sub-cohort of patients created in the previous step. PHENOTREE assists physicians in identifying groups of phenotypes and their corresponding patient sub-cohorts at different granularities through this process. The system utilizes the radial Reingold-Tilford tree to visualize the results. Each node in the tree represents a structured phenotype and a sub-cohort characterized by this phenotype.

VALENCIA (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020b) is a VA system that aims to address the challenges of high-dimensional EHRs by integrating several dimensionality reduction (DR) and cluster analysis (CA) techniques with real-time analytics and interactive visualizations (Figure 8). VALENCIA’s analytics engine has two components—namely, DR and CA engines. The DR engine incorporates several DR techniques to transform EHRs from the high-dimensional space to one with lower dimensions. The CA engine then uses several clustering techniques to classify the data points in this low-dimensional space into meaningful groups with similar characteristics. VALENCIA allows the user to choose the most appropriate combination of DR and CA techniques and explore the results through two main views—namely, DR and CA views. The DR view has four subviews, including raw-data, projected-features, association, and variance subviews. These subviews allow the user to choose their features of interest, select the DR technique, adjust the configuration parameters, investigate how features are associated with transformed dimensions, and choose dimensions to be included in the CA engine. The CA view has three subviews—namely, hierarchical subview, frequency subview, and projected-observation subview. These subviews allow the user to examine the hierarchical structure of the CA results, choose the CA technique and configuration parameters, and observe the distribution of features in each subset of the data.

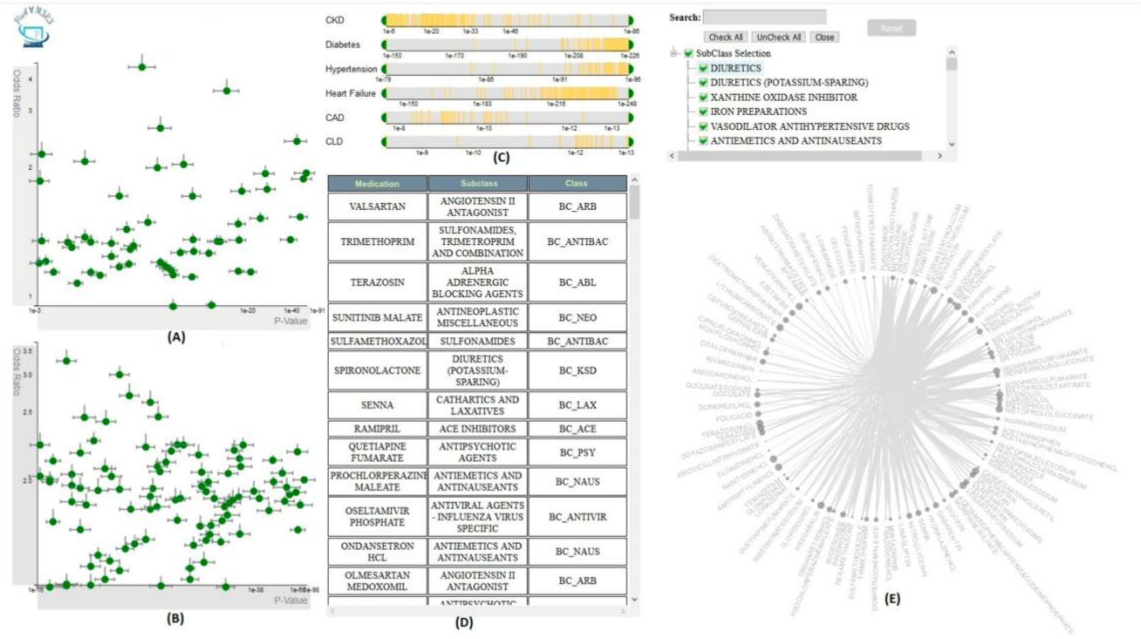


(a)

(b)

Figure 3-8: The screenshot of VALENCIA (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020b) showing (a) the DR view and (b) the CA view. Source: image by authors.

VISA_M3R3 (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020a) is a VA system that allows clinical researchers to identify medications or medication combinations that are associated with a higher risk of acute kidney injury (AKI) (Figure 9). The system incorporates regression, frequent itemset mining, and interactive visualization to help the user explore the relationship between medications and AKI. The analytics module of Visa_M3R3 is composed of two components. The first component is the single-medication analyzer that focuses on finding associations between individual medications and AKI using multivariate regression. The multiple-medications analyzer aims to identify associations between medication combinations and AKI using frequent itemset mining and regression. All models are validated through Bonferroni correction and represented in multiple interactive views. The regression models generated from single-medication and multiple-medications analyzers are represented in two scatter plots in the single-medication and multiple-medication views. The output of the frequent itemset mining is shown using a chord diagram in the frequent-itemset view. The user can filter and control the information presented in other views using sliders in the covariates view. Finally, the medication-hierarchy view displays additional information regarding data elements using a data table.



Medication	Subclass	Class
VALSARTAN	ANGIOTENSIN II ANTAGONIST	BC_ARB
TRIMETHOPRIM	SULFONAMIDES, TRIMETROPRIM AND COMBINATION	BC_ANTIBAC
TERAZOSIN	ALPHA ADRENERGIC BLOCKING AGENTS	BC_ABL
SUNITINIB MALATE	ANTINEOPLASTIC MISCELLANEOUS	BC_NEO
SULFAMETHOXAZOL	SULFONAMIDES	BC_ANTIBAC
SPIRONOLACTONE	DIURETICS (POTASSIUM-SPARING)	BC_KSD
SENNA	CATHARTICS AND LAXATIVES	BC_LAX
RAMIPRIL	ACE INHIBITORS	BC_ACE
QUETIAPINE FUMARATE	ANTIPSYCHOTIC AGENTS	BC_PSY
PROCHLORPERAZINE MALEATE	ANTIEMETICS AND ANTINAUSEANTS	BC_NAUS
OSELTAMIVIR PHOSPHATE	ANTIVIRAL AGENTS - INFLUENZA VIRUS SPECIFIC	BC_ANTIVIR
ONDANSETRON HCL	ANTIEMETICS AND ANTINAUSEANTS	BC_NAUS
OLMESARTAN MEDOXOMIL	ANGIOTENSIN II ANTAGONIST	BC_ARB

☑ SubClass Selection
- ☑ DIURETICS
- ☑ DIURETICS (POTASSIUM-SPARING)
- ☑ XANTHINE OXIDASE INHIBITOR
- ☑ IRON PREPARATIONS
- ☑ VASODILATOR ANTIHYPERTENSIVE DRUGS
- ☑ ANTIEMETICS AND ANTINAUSEANTS

Figure 3-9: The screenshot of VISA_M3R3 (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020a) showing (A) the single-medication view, (B) the multiple-medication view, (C) the covariates view, (D) the data table, and (E) the frequent-itemset view. Source: image by authors.

3.4 Design Space

In this section, we introduce the primary dimensions of the design space of EHR-based VA systems and highlight the key elements in each dimension that are frequently used for designing and developing these systems. For a VA system to work well, there must be harmonious functioning among all of its components. Such components include VA tasks, analytics and data models, and visual representations. One way in which we can investigate the strength of the coupling among components of VA systems is through the lens of interaction. Therefore, the four key dimensions that are used to evaluate the existing systems include VA tasks, analytics, visualizations, and interactions (for comparison see (Sedig et al., 2012; Sedig & Parsons, 2016)). To identify the main categories for each dimension in the design space, we have examined the existing EHR-based VA systems (see (Mane et al., 2012; D. Gotz et al., 2014; D. Gotz & Stavropoulos, 2014; A. F. Simpao et al., 2014; J. Sun et al., 2014; Huang et al., 2015; Klimov et al., 2015; Perer et al., 2015; Baytas et al., 2016; Hund et al., 2016; Kwon et al., 2016; Arianna Dagliati et al., 2018; Kwon et al., 2018; Ledieu et al., 2018; Dingen et al., 2019; Ha et al., 2019; S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020a; . S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020b; Guo et al., 2020; Jin et al., 2020; Kwon et al., 2020; Mica et al., 2020)).

3.4.1 VA Tasks

In this section, we summarize the EHR-based VA tasks that have gained attention from researchers over the past decade. We classify these tasks into four categories according to their objectives: 1) understanding the progression of diseases, 2) discovering and exploring cohorts of interest, 3) learning and understanding prediction models, and 4) discovering adverse events.

Understanding the Progression of Diseases: VA techniques can be used to model and visualize a patient's medical condition over time, which is known as the patient's medical trajectory. Research studies have shown that different patient trajectories can have different associated risks for the same outcome (Oh et al., 2016; Yadav et al., 2015). For instance, a patient may die due to cardiovascular complications, kidney complications, or peripheral complications. Although the outcome is the same, disease progression paths that lead to the outcome is typically different. Thus, studying such various trajectories can result in the development of more tailored treatment plans, the discovery of biomarkers, and the development of different risk estimation indices. Comorbidity analysis, which is the process of analyzing and exploring associations among diseases, is another key factor in improving the quality of care, especially for older patients who suffer from multiple diseases. Therefore, understanding the incidence, prevalence, and coincidence of diseases is the foundation for making important policy decisions. Thus, there have been many EHR-based VA systems developed to accomplish this task. For instance, Huang et al. (Huang et al., 2015) have developed a system that supports the exploration of patient trajectories to help clinical researchers detect chronic diseases and determine how a set of patients with multiple chronic diseases might go on to develop other comorbidities over time. DPvis (Kwon et al., 2020) supports the interactive exploration of disease progression patterns and the discovery of interactions between such patterns and patient's characteristics. Similarly, DecisionFlow (D. Gotz & Stavropoulos, 2014) and Peekquence (Kwon et al., 2016) allow comparison of multiple complex patient event pathways by combining statistical analysis processes with interactive flow-based visualizations.

Discovering and Exploring Cohorts of Interest: The identification of a cohort (group) of patients who meet predefined criteria from a large patient population has various use cases, including survival analysis, clinical trial recruitment, and other retrospective studies (Mathias et al., 2012; Strom et al., 2011). Cohort identification forms a platform for future clinical research studies in areas such as predicting complications, pharmacovigilance, and detecting adverse events. Traditionally, this process is carried out through chart reviews by primary care staff and research staff in individual practices to query the clinical systems for patients matching a specific set of criteria. However, manual cohort identification can be extremely challenging and time-consuming, depending on the complexity of the criteria. This is because the patient data satisfying these criteria is buried within large volumes of data stored in EHRs. Thus, there is a need for electronic phenotyping algorithms to replace the manual chart reviews for cohort identification.

A phenotype can be defined as a specification of an observable state of an organism. It can be applied to patient characteristics that are inferred from EHRs, such as clinical conditions, blood type, or physical traits. Phenotype algorithms that characterize or identify phenotypes can be used for the direct identification of cohorts based on clinical or medical characteristics, risk factors, and complications, thereby allowing clinical researchers to improve patient outcomes. These algorithms can be generated using various forms of machine learning techniques. However, an integrated approach that combines these analysis techniques with visualization is more likely to facilitate the process of creating and comparing different patient cohorts, determining risk factors associated with a particular disease, and discovering hidden structures in the patient data. As a result, several VA systems have been developed recently to address this issue. For instance, Mane et al. (Mane et al., 2012) developed VisualDecisionLinc to help clinical researchers identify subpopulations of patients with similar clinical characteristics to help them evaluate the risks and effectiveness of different treatment options. Similarly, PHENOTREE (Baytas et al., 2016) is another VA system that allows clinical researchers to explore patient cohorts and create and evaluate phenotypes by generating a phenotype hierarchy.

Learning and Understanding Prediction Models: The focus on creating prediction models is increasing in many areas of clinical research. These models aim to assist physicians in personalized decision making with regards to diagnosis, prognosis, and treatment. Examples of successful risk prediction models are the Apache system that estimates the risk of hospital mortality, the Framingham heart score that predicts cardiovascular mortality, and the Nottingham Prognostic Index that allocates patients with breast cancer to different risk groups (Chalmers et al., 2013; Galea et al., 1992; Gaziano et al., 2010; Knaus et al., 1991; Nashef et al., 1999; Timmerman et al., 2005). Despite the strong performance of these models, it is often challenging for physicians to understand how the prediction models arrive at an estimated risk. The black-box nature of most of these models can impede their wide adoption in clinical practice since there is little tolerance for errors in medical decision making.

Thus, providing interpretability and transparency in prediction models is critical in validating the resulting predictions. To address these needs, VA systems provide clinical researchers with accurate, fast, and trustworthy interpretation of prediction models by integrating effective visual representations with machine learning techniques (Munzner, 2014; Treisman, 1985; Ware, 2019). For instance, RetainVIS (Kwon et al., 2018) combines interpretable and interactive RNN-based models and interactive visualization to allow exploration of patient records in the context of prediction tasks.

Discovering adverse events: Adverse events can be defined as the harmful effects of medical care on a patient's medical condition. They are caused by medical management rather than the patient's underlying condition (Medicine & America, 2000). For instance, an infection developed during the treatment of a different condition is considered an adverse event. Adverse events are responsible for 2.9%-16-6% of all acute hospitalizations, and studies have shown that 30%-58% of all these events are preventable (Brennan et al., 1991; Leape et al., 1991; E. J. Thomas et al., 1999, 2000; Wilson et al., 1995). Adverse events can also often be linked to drugs. Adverse drug events cause 3.5 million physician visits, 125,000 hospitalizations, and 98,000 drug-related deaths each year (Torio et al., 2006). Even though drugs are tested for any potential adverse events and are cleared for marketing to the medical community, unsuspected adverse events are occasionally detected. This is due to the fact that clinical trials are usually limited to short time periods and include only a small test cohort. In addition, the frequency of these adverse events may be so low that they are hard to detect in clinical trials. Another issue in detecting adverse drug events is confounding by indication. For instance, insulin is prescribed for diabetes. Myocardial infarction is a common comorbid disease for patients who have diabetes and thus, detecting the adverse event myocardial infarction for insulin is a false positive ("a confounding effect"). There are several approaches to detect significant adverse drug events using automatic analysis techniques; however, most of these approaches overlook low-frequency events. Furthermore, the domain knowledge regarding the confounding effect should be included in these automatic analysis techniques. VA systems can address these issues by involving domain experts in the analysis process. For instance, Ledieu et al. (Ledieu et al., 2018) developed a VA system for pharmacovigilance in electronic medical records to detect inappropriate drug administration and inadequate treatment decisions in patient sequences. VISA_M3R3 (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020a) is another EHR-based VA system that helps healthcare providers to identify medications that are associated with a higher risk of acute kidney injury.

3.4.2 Analytics

There have been several analytics methods used for visual analysis of EHR data. These methods include 1) classification, 2) clustering, 3) frequent pattern mining, 4) association rule mining, 5) time series mining, 7) regression, 8) inference, and 9) dimensionality reduction.

Classification: Classification is used to classify data points into predefined categorical class labels. "Class" is the feature in a dataset in which users are most interested. In statistics, it can be defined as the dependent variable. In order to classy data points, a classification technique generates a model, including classification rules. Classification is a two-step process, including training and testing. In the training step, a classification model is built by analyzing training data that contains class labels. The accuracy of the classification model depends on the degree to which classifying rules are correct. In the testing step, the classifier's (i.e., classification model) ability to classify unknown data points for prediction is examined. Some of the most common classification techniques that are used in the analysis of EHRs are support vector machine (Cortes & Vapnik, 1995; Cristianini et al., 2000), decision tree (Breiman et al., 1984), naïve Bayes (Lewis, 1998), and neural network (Daniel, 2013). For instance, RetainVIS (Kwon et al., 2018) uses the RetainEX technique (i.e., a bidirectional recurrent neural networks (RNN) model) to generate prediction scores based on the temporal information stored in EHRs and help the user identify which medical codes or patient visits contribute to the prediction score.

Clustering: Clustering is an unsupervised learning technique that occurs by analyzing only independent variables. In other words, unlike classification techniques, clustering techniques do not use "class". Thus, clustering is best used for exploratory studies, especially if those studies include large volumes of data, but very little is known about the data. Clustering groups the data points into a certain number of clusters so that points within a cluster have high similarity and points from different clusters have low similarity. The similarities between the data points are measured using their feature values. Some of the most commonly used clustering techniques in exploring EHRs are k-means (Hartigan & Wong, 1979; Jain, 2010), hierarchical clustering (Nielsen, 2016), model-based clustering (Fraley & Raftery, 2002), and density-based clustering (Ester et al., 1996). For instance, the VA system developed by Guo et al. (Guo et al., 2020) uses HDBSCAN which is a density-based clustering technique to cluster similar patients according to the combination of tests taken during a specific time period.

Pattern discovery: Pattern discovery aims to identify statistically significant associations and frequently occurring patterns in the data. In the analysis of EHR data, pattern discovery can be further classified into frequent pattern mining and association rule mining techniques (R. Agrawal, Imielinski, et al., 1993). The purpose of the frequent pattern mining is to identify the inherent regularities in the EHR data. In other words, these techniques can be used to find common subsequences in the clinical event sequence dataset. Frequent pattern mining can be further extended to other problems such as sequential pattern mining and time-series mining that are very common when dealing with clinical event sequences. Association rules can be considered as a second-stage output of frequent pattern mining. Association rule mining is often used to discover relationships among data items. Association rule mining techniques can be employed to identify underlying relationships among health conditions, symptoms, and diseases in the healthcare field. For instance, the VA system developed by Gotz et al. (D. Gotz et al., 2014) helps the user to explore the clinical event sequences using a Frequent Pattern Mining (FPM) engine. The FPM engine has two main components, including the Frequent Pattern Miner and the Statistical Pattern Analyzer. The frequent pattern miner uses the SPAM (Ayres et al., 2002) algorithm for pattern discovery. Then the Statistical Pattern Analyzer computes correlations between the mined patterns and the outcome measure.

Regression: Regression techniques are often used to identify associations between features, such as the extent to which feature A affects feature B. Logistic regression is a special type of regression that is commonly used in the analysis of clinical data (Ismail & Anil, 2014). It draws a separating line among classes using the training data; then, it applies the line to classify the test data's unknown data points. Logistic regression is often used to analyze the relationship between a dependent feature (e.g., patient outcomes) and one or more independent features (e.g., patient comorbidities, symptoms, and laboratory test results). For instance, RegressionExplorer (Dingen et al., 2019) allows the user to formulate a new hypothesis and steer the development of models by creating, comparing, and evaluating multiple regression models.

Inference: Inference refers to the process of reaching conclusions based on the evidence found in the existing data. However, conclusions drawn from inference are only justifiable under some specific conditions and can be false when applied to unobserved data. One of the inference techniques used in the analysis of the clinical event sequences is graphical models. Graphical models show the conditional dependence between clinical events using an event correlation graph, such as the Markov chain (Stopar et al., 2019) and Bayesian Networks (Bhattacharjya et al., 2020). *For instance,* DPvis (Kwon et al., 2020) uses continuous-time hidden Markov models to learn how various diseases go through different states, discover biomarkers (i.e., observed variables) that can characterize the disease progression, and to use these biomarkers to identify diseases earlier in patients.

Dimensionality Reduction: Dimensionality reduction is the process of transforming a high-dimensional dataset into a dataset with reduced dimensionality without losing too much information (Siwek et al., 2013). Dimensionality reduction techniques help the user to get a better understanding of the underlying structure of the data by removing multicollinearity and creating a dataset with a smaller volume. In Clinical settings, dimensionality reduction is often required, as EHRs are often high dimensional. Thus, by reducing the dimensions, one can mitigate this issue and possibly decrease the computational time for analysis and visualization of the EHR data. For instance, PHENOTREE (Baytas et al., 2016) uses sparse principal component analysis (SPCA) to identify primary clinical features that describe the patient population to assist the user in building and navigating a phenotype hierarchy and exploring patient cohorts.

3.4.3 Visualizations

We identify four categories of visualizations that are commonly used in EHR-based VA systems: 1) Relation-based, 2) time-based, 3) hierarchy-based, and 4) flow-based visualizations.

Relation-based: Relation-based visualizations show connections and relationships between two or more attributes. They are inherent to the clustering and association tasks within VA. A variety of visualization techniques can be used to display relations, such as scatter plots, parallel coordinates plots, bubble charts, bar charts, and heatmaps. For instance, Gotz et al. (D. Gotz et al., 2014) use a scatter plot to show the distribution of the most frequent patterns with respect to their level of support for patients with positive and negative outcomes.

Time-based: Time-based visualizations show data or the sequence of clinical events over a time period. The main function of these visualizations is to assist the analysis and reasoning process of healthcare experts when investigating patients' clinical history. The primary time-based visualization technique is Timeline. Timeline displays a series of clinical events in a temporal order where each event is generally represented by an icon and is encoded by size, shape, or color to distinguish events with different characteristics. For instance, Peekquence (Kwon et al., 2016) displays each patient's entire event sequence in a timeline to assist users in discovering patterns in the patient’s event sequences.

Hierarchy-based: Hierarchy-based visualizations show how data items are ordered and ranked in a system. Several visualization techniques can be used to display the hierarchical structure of the data, such as tree diagrams, treemaps, and icicle plots. For instance, DecisionFlow (D. Gotz & Stavropoulos, 2014) aggregates clinical event sequences with the similar occurrence of milestone events into a tree of sequences, where each node encodes an event positioned according to its prefix in the sequence. VALENCIA (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020b) uses a treemap to display the distribution of patient characteristics in different clusters.

Flow-based: Flow-based visualizations show flows and their quantities with respect to one another. Sankey diagrams and parallel sets are two of the main flow-based visualization techniques that are used in EHR-based VA systems to provide an overview of transitions between different types of clinical events. For instance, Care Pathway Explorer (Perer et al., 2015) uses a Sankey diagram to show how clinical events in the most frequent patterns are connected to each other, where each event is represented by a node and nodes belonging to the same pattern are connected by edges.

3.4.4 Interactions

Interaction is an integral part of VA and plays a vital role in the success of EHR-based VA systems. We adapted the epistemic actions introduced as part of the framework proposed by Sedig et al. (Sedig & Parsons, 2013) to classify and evaluate interactions used in EHR-based VA systems. Epistemic actions are actions that are taken to alter the visualizations in a manner that supports the user's analytical and cognitive needs (mental processes). The subset of the actions identified in the framework commonly used in EHR-based VA systems (Sedig & Parsons, 2013) include: 1) arranging, 2) comparing, 3) drilling, 4) filtering, 5) searching, 6) selecting, 7) transforming, 8) translating, 9) animating/freezing, 10) collapsing/expanding, 11) inserting/removing, and 12) linking/unlinking.

Arranging: Arranging refers to acting upon visualizations to change their ordering, either temporally or spatially. Some variants of this epistemic action that are commonly used in EHR-based VA systems are sorting, ordering, organizing, and ranking. For instance, VALENCIA (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020b) supports arranging by allowing the user to sort the heatmap that represents the result of dimensionality reduction techniques based on either a dimension or a feature by clicking on the corresponding column or row.

Comparing: Comparing refers to acting upon visualizations to determine their degree of similarity or dissimilarity. The degree of similarity is often defined as the distance between or proximity of value or meaning in EHRs-based VA systems. For instance, RegressionExplorer (Dingen et al., 2019) allows the user to investigate a regression model's behavior on a specific subpopulation and compare it with its behavior on a different subpopulation. It supports this action by letting the user drag and drop a feature from the variable selection view to the population view, which results in the partition of the patient population.

Drilling: Drilling is acting upon visualizations to bring out deep information that is currently not displayed. Its main functionality is to make perceptually inaccessible information available for further investigation. Drilling is a fundamental action in EHR-based VA systems as it helps the user process and examine desired information more deeply when dealing with a large volume of data stored in EHRs. For instance, Visa_M3R3 (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020a) supports this action by allowing the user to hover their mouse over a glyph representing a regression model in the scatter plot to get additional information about the corresponding model.

Filtering: Filtering refers to acting upon visualizations to show a subset of their elements based on specific criteria. It allows the user to adjust the level of details, which is an essential feature of the process of abstraction in the exploration of complex high-dimensional EHRs. Thus, filtering is integral to many EHR-based VA tasks. For instance, the VA system developed by Mica et al. (Mica et al., 2020) allows the user to filter a cohort of patients according to various parameters (e.g., body temperature, age, and lab results).

Searching: Searching refers to acting upon visualizations to locate or seek out the existence of position of certain relationships, items, or structures. Some variants of this action are querying and seeking. Searching is commonly used in EHR-based VA systems. For instance, Visa_M3R3 (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020a) supports this action by allowing the user to enter the name of the medication of interest in a search bar, allowing that medication to get highlighted in other views.

Selecting: Selecting refers to acting upon visualizations to focus on or choose them either individually or as a group. This action is necessary for performing other actions in VA systems. By selecting an information item and making it visually distinctive, the user can keep track of it within a large volume of information, even when it is going through some changes. Most of the EHR-based VA systems support selecting. For instance, DecisionFlow (D. Gotz & Stavropoulos, 2014) allows the user to perform edge selection by clicking on time edges in the temporal flow view. The system then outlines the corresponding rectangular mark representing the edge and updates the overview and the edge statistics view to display information regarding the selected edge.

Transforming: Transforming refers to acting upon visualizations to modify their geometric form. This epistemic action can change the look, size, or orientation of visualizations by scaling, rotating, magnifying, and/or distorting them. Magnifying visualizations is the most common variant of this action in EHR-based VA systems. For instance, Visa_M3R3 (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020a) applies the cartesian fisheye distortion technique on both axes of the scatter plot representing regression models to help the user distinguish between models when the glyphs are densely clustered.

Translating: Translating is acting upon visualizations to convert them into alternative conceptually- or informationally-equivalent forms. This action has a high degree of utility for all EHR-based VA tasks as each alternative visualization form reveals different aspects of the data. For instance, SubVIS (Hund et al., 2016) supports translating by allowing the user to choose a more advanced representation of glyphs to show the subspace's underlying dimensions where each dimension is encoded by a small line around the border of the dot.

Animating/Freezing: Animating/Freezing refers to acting upon dynamic visualizations to create movement in constituent parts or oppositely to stop. Animating can be used to observe temporal trends and show complex relationships among clinical events in EHRs. For instance, the VA system developed by Gotz et al. (D. Gotz et al., 2014) uses a three-staged animation process to transition between different views in the Pattern Diagram view to highlight temporal patterns by allowing comparison between mining results at different parts of a clinical episode.

Collapsing/Expanding: Collapsing/Expanding is acting upon visualizations to make them compact or, oppositely, make them diffuse. These actions can facilitate investigating the associations among data items when dealing with complex high-dimensional EHRs. Collapsing enables the user to condense a set of data items into one, thus reducing complexity and facilitating the understanding of overall associations and patterns within EHRs while expanding allows the user to explore them in more detail. For instance, DPvis (Kwon et al., 2020) allows the user to convert Pathway over Observation diagram (i.e., a stacked Sankey diagram displaying pathways for subjects) into a bipartite Sankey diagram, which includes two stacked bars with paths between them. The converted view simplifies the transition by only displaying the changeover from a start to an end, thereby allowing the user to follow the state pathways between two consecutive patient visits.

Inserting/Removing: Inserting/Removing refers to acting upon visualizations to add new visualizations into them, or oppositely to take out unwanted or unnecessary parts. Such actions can facilitate the exploration of EHRs, allowing the user to create hypothetical scenarios by interjecting or getting rid of clinical events in patient trajectories and observing the effect. For instance, CarePre (Jin et al., 2020) enables the user to edit the focal patients' event sequence within the prediction view by inserting new events and removing existing ones. This allows the user to determine key factors that affect the prediction results and explore how changes in the patient's record (i.e., a novel treatment or the absence of a comorbidity) impact those results.

Linking/Unlinking: linking/Unlinking refers to acting upon visualizations to establish an association between them, or oppositely to disconnect their associations. In general, EHR-based VA systems with multiple coordinated views are assumed to support these actions. For instance, in VisualDecisionLinc (Mane et al., 2012), all the views are linked together to create a coordinated display, where filtering updates on one of the views prompts relevant updates to data items in other views. This aids the user in their decision-making process and evaluation of multiple treatment options by helping them to better understand the relationship between different data elements.

3.5 Discussion

The reviewed VA systems demonstrate a broad spectrum of VA tasks and analytics methods, visualizations, and interactions to deal with the challenges of complex data stored in EHRs. The EHR-based VA systems are getting increasingly popular in recent years. Figure 10 includes a timeline that shows most of these systems were developed between 2014 to 2020. We evaluate these systems by analyzing their strengths and weaknesses using the dimensions introduced in Section 4. Figure 11 provides an overview of the characteristics of the systems with respect to the four primary dimensions, including VA tasks, analytics, visualizations, and interactions.

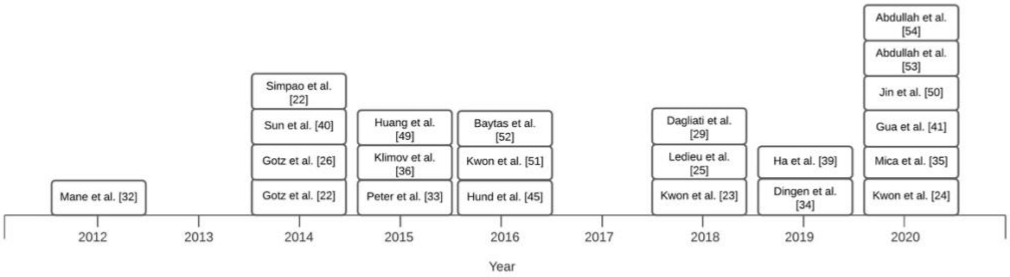
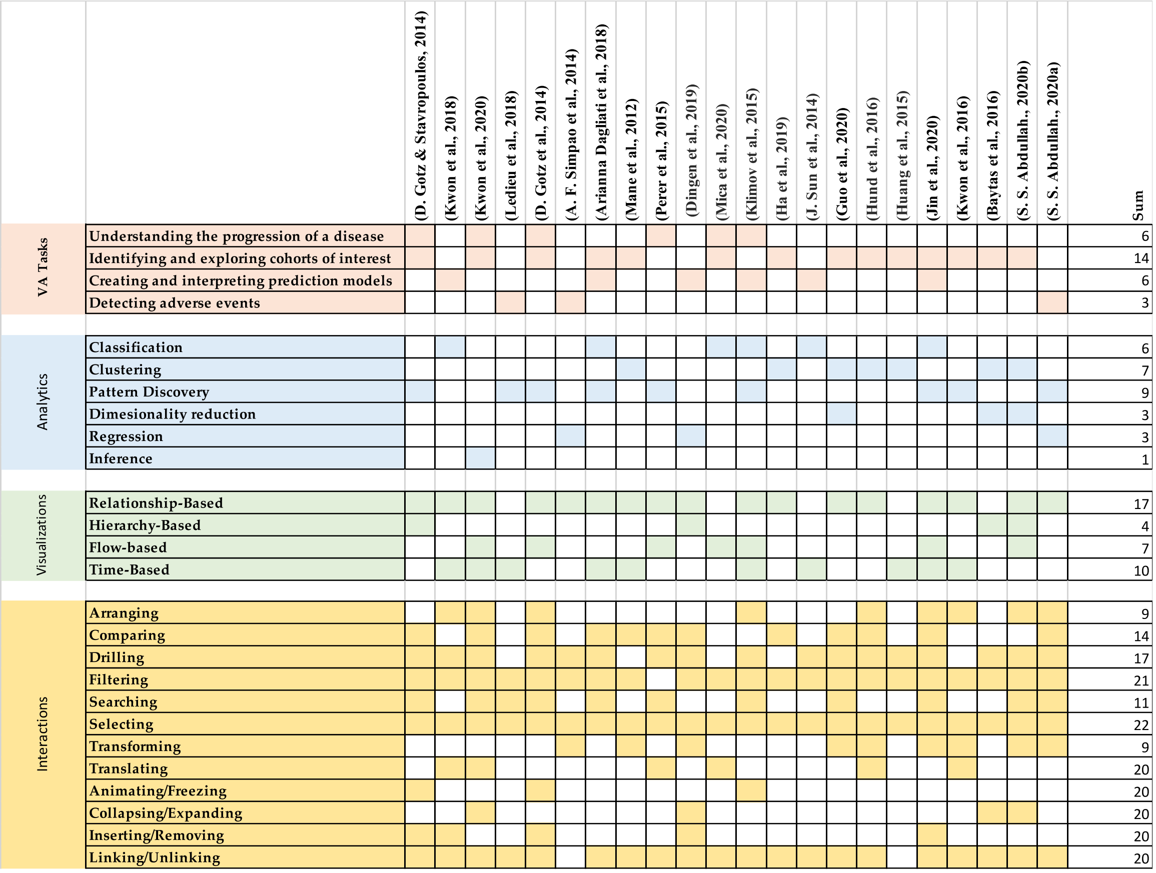


Figure 3-10: The figure shows the original order of the creation of the VA systems in a timeline.

The most common VA task that is supported by the systems is discovering and exploring patient cohorts. This task's popularity is mostly because of its numerous use cases, including survival analysis, clinical trial recruitment, and other kinds of retrospective studies (Mathias et al., 2012; Strom et al., 2011). Moreover, identifying patients who satisfy pre-defined criteria can form the platform for future studies in areas such as predicting patient outcomes, pharmacovigilance, and understanding patient trajectories. Thus, as seen in Figure 11, discovering and exploring patient cohorts is supported by most systems that support other VA tasks. Other VA tasks supported by several systems are understanding the progression of a disease and learning and exploring prediction models. Understanding the incidence, prevalence, and coincidence of diseases is the



		(D. Gotz & Stavropoulos, 2014)	(Kwon et al., 2018)	(Kwon et al., 2020)	(Ledieu et al., 2018)	(D. Gotz et al., 2014)	(A. F. Simpao et al., 2014)	(Arianna Dagliati et al., 2018)	(Mane et al., 2012)	(Perer et al., 2015)	(Dingen et al., 2019)	(Mica et al., 2020)	(Klimov et al., 2015)	(Ha et al., 2019)	(J. Sun et al., 2014)	(Guo et al., 2020)	(Hund et al., 2016)	(Huang et al., 2015)	(Jin et al., 2020)	(Kwon et al., 2016)	(Baytas et al., 2016)	(S. S. Abdullah., 2020b)	(S. S. Abdullah., 2020a)	Sum
VA Tasks	Understanding the progression of a disease	■		■		■				■		■	■											6
	Identifying and exploring cohorts of interest	■		■		■		■	■			■		■		■	■	■	■	■	■	■		14
	Creating and interpreting prediction models		■					■			■		■		■				■					6
	Detecting adverse events				■		■																■	3
Analytics	Classification		■					■				■	■		■				■					6
	Clustering								■					■		■	■	■			■	■		7
	Pattern Discovery	■			■	■		■		■			■						■	■			■	9
	Dimesionality reduction															■					■	■		3
	Regression						■				■												■	3
	Inference			■																				1
Visualizations	Relationship-Based	■	■	■		■	■	■	■	■	■		■	■		■	■		■	■		■	■	17
	Hierarchy-Based	■									■										■	■		4
	Flow-based			■		■				■		■	■						■			■		7
	Time-Based		■	■	■			■	■				■		■			■	■	■				10
Interactions	Arranging		■	■		■							■				■		■	■		■	■	9
	Comparing	■		■		■		■	■	■	■			■		■	■		■				■	14
	Drilling	■	■	■		■	■	■		■	■		■		■	■	■	■	■		■	■	■	17
	Filtering	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	21
	Searching	■		■	■	■		■		■			■			■			■			■	■	11
	Selecting	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	22
	Transforming						■		■		■					■	■		■	■		■	■	9
	Translating		■	■						■		■					■			■				20
	Animating/Freezing	■				■							■											20
	Collapsing/Expanding			■							■										■	■		20
	Inserting/Removing	■	■			■					■								■					20
	Linking/Unlinking	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	20

Figure 3-11: The reviewed EHR-based visual analytics system. Each system is labeled by the relevant element in the design space. The rows are grouped and colored by dimensions of the design space: VA tasks, analytics, visualizations, and interactions.

foundation on which important policy decisions are made, and thus researchers have spent considerable efforts on this task. Similarly, learning and exploring prediction models is the most natural and immediately impactful task. Conversely, discovering adverse events is not as popular as the other VA tasks. This could be due to the fact that this task requires extensive collaborations.

Pattern discovery is one of the most widely used analytics methods in EHR-based VA systems. It is often used to uncover common subsequences in the clinical event sequences and quantify the similarity between event sequences. This analytics method is commonly used in most of the systems that support understanding the progression of diseases. The other popular analytics method is clustering. For instance, the VA system developed by Gotz et al. (D. Gotz et al., 2014) and Care Pathway Explorer (Perer et al., 2015) use the SPAM (Ayres et al., 2002) frequent sequence mining algorithm to detect the most frequent patterns and examines how these patterns are associated with patient outcomes. Clustering often aims to organize patients into several groups, where patients within each group have similar characteristics. Most of the systems that support cohort discovery task use different clustering techniques. For instance, the VA system developed by Guo et al. (Guo et al., 2020), VALENCIA (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020b), and RadVis (Ha et al., 2019) apply several clustering techniques to assists clinical researchers in the identification of cohorts of patients with similar clinical characteristics. One of the other commonly used analytics methods is classification. It is often used in VA systems that support learning and exploring prediction models. For instance, RetainVIS (Kwon et al., 2018) and CarePre (Jin et al., 2020) support this VA task by creating deep learning prediction models and allowing the user to explore and interpret the results. Dimensionality reduction, regression, and inference are the other analytics methods that are used in EHR-based VA techniques. Dimensionality reduction techniques are usually used as a pre-processing step, followed by clustering techniques.

Most of the systems use multiple visualizations to allow the user to explore the data and the analytics results from different perspectives. The most common visualization used in the systems is relation-based visualizations, including scatter plots, bar charts, heatmaps, and parallel coordinates plots. Scatter plots are most suitable in representing clustering techniques, while bar charts are usually used to show the distribution of clinical features and their contributions to the prediction models. The other common visualization used in the EHR-based VA systems is time-based. The reviewed systems frequently adopt time-based visualizations such as timelines to display temporal distribution of clinical events in different time granularities and reveal temporal information among clinical event sequences. CarePre (Jin et al., 2020) and RetainVIS (Kwon et al., 2018) represent patient’s clinical events in a temporal order in a time-based visualization that allows the user to conduct what-if analyses by modifying these events and get the newly generated predicted risks for the patient. Thus, the systems can display the result of classification and pattern discovery techniques by adopting time-based visualizations. Flow-based visualizations such as Sankey diagrams and parallel sets have also been adopted by many EHR-based VA systems. They are mostly used to provide an overview of the progression pathways of clinical events within a cohort and thus, help the user to understand which clinical features, pathways, or other structures are more associated with the outcome of interest. Finally, a small number of EHR-based VA systems adopt hierarchy-based visualizations such as icicle plots and treemap to reveal the hierarchical organization of features or event sequences within EHRs. For instance, while DecisionFlow (D. Gotz & Stavropoulos, 2014) uses a hierarchy-based tree to display the aggregated progression patterns of interest, VALENCIA (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020b) allows the user to explore the hierarchical structure of clustering results using a tree map.

EHR-based VA studied in this review support a wide range of interactions. Selecting, filtering, likening/unlinking, and drilling are the most common epistemic actions supported by the systems. Selecting is supported by all the systems as this action is often regarded as a way for the user to perform additional manipulations on the selected data items. Filtering is also supported by most of the systems since these systems need to handle a large volume of data stored in EHRs. Likening/unlinking is another common action supported by the systems as most of the VA systems with multiple views support this action through brushing and linking technique. Drilling is the fourth epistemic action to play a leading part in the systems. Almost all of the reviewed systems provide a function to display additional details about data items, typically in a tooltip. Comparing is a common epistemic action in the reviewed systems, especially when investigating the similarities and differences between clinical event sequence data. Some of the reviewed systems support searching action through an intuitive visual query interface. It is surprising that the systems do not more widely support the translating action, given the wide range of possible visual encodings. Inserting/removing actions are mostly utilized in the systems that allow the user to test different hypotheses regarding the factors that might affect the patient outcome by adding and removing different event types to/from the patient’s event sequence. Collapsing/expanding action is not widely supported by the reviewed systems. These actions are mostly used in systems that adopt a hierarchy-based visualization. Finally, animating/freezing is only supported by three systems. It is interesting to note that the systems that support these actions are used to perform pattern discovery and understanding the progression of diseases.

As shown in Figure 11, this review enables researchers to identify the EHR-based VA research areas that requires more attention. First, it appears that most of the existing systems support a limited number of analytics methods, which is not appropriate for handling ill-defined tasks (i.e., tasks that do not have clear objectives or solution paths) in EHRs. Second, bipolar actions (e.g., animating/freezing and collapsing/expanding) are not commonly supported by the systems in comparison to unipolar action patterns (e.g., selecting and comparing). Lastly, most of the systems mainly allow for interactive exploration of the analytics results rather than illustrating the underlying working mechanisms of those techniques, which is essential in building trust with the user in healthcare settings. The findings of this paper can provide value to designers as an organized catalog of different approaches that are most suitable for EHR-driven tasks.

Chapter 4

4 Visual Analytics for Predicting Disease Outcomes Using Laboratory Test Results

This chapter has been submitted to Computers in Biology and Medicine.

Please note that the format has been changed to match the format of the dissertation. Figure, Section, and Table numbers mentioned herein are relative to the chapter number. For instance, “Figure 1” corresponds to Figure 4-1. Additionally, when the term “paper”, “work”, or “research” is used, it refers to this particular chapter.

4.1 Introduction

Accurate and early clinical diagnoses play an important role in the successful treatment of diseases. Every disease stems from or causes changes on a molecular and cellular level, and some of these changes can be detected through changes in urine and blood parameter values (Gunčar et al., 2018). Patterns within laboratory test results may contain additional information relevant to patient care that are not appreciated by even the most experienced physicians (Badrick, 2013; Cabitza & Banfi, 2018). Laboratories typically report test results as individual categorical and numerical values, but some individual results, particularly when studied in isolation, may have limited clinical value. Physicians often integrate several individual tests from a patient and interpret them in the context of medical knowledge and experience to use them for disease diagnosis and management. Furthermore, patients might have many individual tests, spanning years. There is a higher chance of overlooking important patterns in the increasing numbers of parameters that laboratories measure. While the manual approach to test interpretation is the routine procedure in most cases, data analytics offer the potential to improve the laboratory tests' diagnostic value (Louis et al., 2014). Several studies have been conducted to develop risk prediction models using laboratory test data, and some of these models were developed using data analytics techniques (Demirci et al., 2016; Diri & Albayrak, 2008; Goldstein et al., 2017; Y. Kumar & Sahoo, 2013; Lin et al., 2013; K. E. Liu et al., 2014; Lu & Ng, 2020; Nelson et al., 2012; Putin et al., 2016; Razavian et al., 2015; Richardson et al., 2016; Somnay et al., 2017; Surinova et al., 2015; Yang et al., 2020; Yuan et al., 2012).

These studies solely rely on performance metrics such as high accuracy scores to assess model performance. Furthermore, the analytics techniques used in these studies are often treated as black-boxes due to their unclear working mechanisms and their incomprehensible functions. Therefore, the question arises whether the user can trust these analytics techniques or not, especially in medical settings, where the model makes a critical decision about patients (Han et al., 2011; Krause, Perer, & Bertini, 2016). One way to increase the model's interpretability is by getting the user involved in the analytics process through an integrated approach called visual analytics (Kehrer & Hauser, 2013; D. A. Keim et al., 2010).

Visual Analytics (VA) is an emerging research discipline that integrates data analytics and interactive visualizations (D. Keim, Andrienko, et al., 2008; D. A. Keim et al., 2010; Ola & Sedig, 2018). It has the potential to enhance the user's confidence in the prediction results by improving their understanding of the modeling process and output (Munzner, 2014; Treisman, 1985; Ware, 2019). It is capable of illustrating the model's rationale, presenting the prediction result, and providing the user with the means to validate the model's response. VA allows the user to access, modify, and restructure the displayed data and guide the data analytics techniques. This, in turn, starts off internal reactions that result in execution of more analytical processes. VA aims to make the best possible use of the massive amount of data stored in EHRs by combining the strength of analytical processes and the user's visual perception and analysis capabilities (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020a, 2020b; Saffer et al., 2004; Allan F. Simpao et al., 2014). It enhances the user's ability to accomplish data-driven tasks by allowing them to examine and analyze EHRs in a way that would be very difficult to do otherwise (Ola & Sedig, 2014; Parsons et al., 2015).

The goal of this paper is to show how VA systems can be developed systematically to create disease prediction models using laboratory test result data. To this end, we present a novel proof of concept system called SUNRISE (viSUal aNalytics for exploring the association between laboRatory test results and a dIsease outcome using xgbooSt and Eclat). SUNRISE allows healthcare providers to examine associations between different groups of laboratory test results and a specific disease by letting them hypothesize new data by tweaking the test values and inspecting how the predictive model responds. It aims to support the user to go beyond judging predictive models based on their performance measures. Instead of just relying on the evaluation metrics, SUNRISE helps the user better understand how predictions are generated by illustrating their underlying working mechanisms. While several VA systems have been designed for other areas in healthcare (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020a, 2020b; Basole et al., 2015; Baytas et al., 2016; D. H. Gotz et al., 2012; Klimov et al., 2015; Mane et al., 2012; Mittelstädt et al., 2014; Ninkov & Sedig, 2019; Perer et al., 2015; Perer & Sun, 2012), SUNRISE is novel in that it integrates frequent itemset mining (i.e., Eclat algorithm), the extreme gradient boosting technique (i.e., XGBoost), visualization, and interaction in an integrated manner. We demonstrate the usefulness of SUNRISE through a case study of exploring associations between laboratory test results and acute kidney injury (AKI) using large provincial healthcare databases from Ontario, Canada stored at ICES (ICES – a not-for-profit, independent, world-leading research organization that utilizes population-based health data to produce knowledge on several healthcare issues).

The rest of this paper is organized as follows. In Section 2, we provide an overview of the conceptual background to understand the design of SUNRISE. We explain the methods used for the design of SUNRISE by providing a description of its structure and modules in Section 3. Section 4 presents a usage scenario of SUNRISE to demonstrate the potential utility of the system. Finally, in Section 5, we discuss the usefulness and limitations of the proposed VA system. Finally, Section 6 concludes the paper.

4.2 Background

This section introduces the necessary concepts and notions for understanding the design of SUNRISE. First, we describe the two components of visual analytics. Then, we briefly describe the concepts of machine learning techniques used in this paper.

4.2.1 Visual Analytics

Visual analytics (VA) can be defined as "the science of analytical reasoning facilitated by interactive visual interfaces" (J. J. Thomas & Cook, 2005). VA can help the user to gain insights from data via integration of analytics techniques and interactive visualization with human judgment. It can support the execution of data-driven cognitive tasks such as sense-making, knowledge discovery, and decision-making, to name a few (Endert et al., 2014; Parsons et al., 2015; Sedig & Parsons, 2016). The primary challenge that goes along with these tasks is that the user needs to rapidly analyze, interpret, compare, and contrast large amounts of information. VAs are capable of providing cognitive and computational possibilities to assist the user in performing these cognitive tasks by combining machine learning techniques, analytical processes, visualizations, and various interaction mechanisms (Angulo et al., 2016; K. Zhao et al., 2016). In summary, VA is composed of two integrated modules: an analytics module and an interactive visualization module (Sedig et al., 2012; Sedig & Parsons, 2016).

The analytics module combines machine learning with data processing techniques to reduce the cognitive load of the user when performing data-intensive tasks (Didandeh & Sedig, 2016; D. A. Keim et al., 2015; Ola & Sedig, 2014; Yu et al., 2016; K. Zhao et al., 2016). The analytics module is technology-independent, and it includes the use of processing techniques and data mining algorithms that fits best the needs of a domain. This module is composed of three primary steps: pre-processing, transformation, and analysis (Han et al., 2011; Jeong et al., 2015). In the pre-processing step, the raw data retrieved from multiple sources gets pre-processed. It includes tasks such as cleaning, integration, fusion, and synthesis (Han et al., 2011). Then the pre-processed data gets transformed into forms that are more appropriate for analysis. Examples of tasks that can be integrated into this stage are feature construction, normalization, aggregation, and discretization (Han et al., 2011). Finally, different data mining algorithms and machine learning techniques are applied to discover useful, unknown patterns from the data in the analysis stage. Data analysis includes tasks such as frequent itemset mining and classification. Despite all the benefits, most of these computational techniques are treated as black-box models and not developed with interpretability constraints. VA can provide the user with the underlying working mechanisms of these models to make them more trustworthy, informative, and easier to understand through interactive visualization.

Interactive visualization in VA involves mapping processed and derived data from the analytics module to visual structures (Sedig et al., 2012; Sedig & Parsons, 2016). It allows the user to interactively control and validate the analytical processes towards better interpretability and performance. It provides the user with new analytical possibilities that can be utilized in an iterative manner (Rostamzadeh et al., 2020). In the context of VA, these iterations can be regarded as discourses between the user and the VA. This back-and-forth communication supports the user by distributing the processing load between the user and the VA system during their analysis and exploration of the data (Z. Liu et al., 2008; Salomon, 1997; Sedig & Parsons, 2016).

4.2.2 Machine Learning Techniques

In this section, we briefly describe the overview of machine learning techniques used in this paper.

4.2.2.1 Frequent Itemset Mining (Eclat)

Frequent itemset mining that was first introduced by Agrawal and Srikant (R. Agrawal & Srikant, 1994) is a task of discovering features that frequently appear together in a database. Although frequent itemset mining was initially proposed to find groups of items that frequently co-occur in transactions made by customers, it now can be considered as a general mining task that can be used in many other domains such as image classification (Fernando et al., 2012), bioinformatics (Naulaerts et al., 2015), activity monitoring (Yunhao Liu et al., 2012), network traffic analysis (Brauckhoff et al., 2012; Glatz et al., 2014), analyzing customer reviews (Mukherjee et al., 2012), and disease prediction (Ilayaraja M. & Meyyappan T., 2015; C. Ordonez, 2006) to name just a few.

The problem of frequent itemset mining can be defined as follows. Let I be a set of items where I=$\{i_1, i_2, \ldots, i_m\}$. A transactional database T includes a set of transactions $\{t_1, t_2, \ldots, t_n\}$ where every transaction is a set of items ($t_i \subseteq I$) that can be identified by a unique transaction identifier (TID). An itemset x is a collection of items and can be characterized by a notion called support value ($\sup(x)$). Support is defined as the ratio of the number of transactions in T that contain x and the total number of transactions in T. It shows the frequency of appearance of an itemset in the database. An itemset is considered frequent if its support value is no less than a given minimum support threshold ($minsup$) that is defined by the user. The task of frequent itemset mining consists of extracting all frequent itemsets from database T given a minimum support threshold ($minsup$). Several techniques have been proposed to address this task. One of the most common frequent itemset mining techniques is the Eclat algorithm.

Eclat (Zaki, 2000) is a depth-first search approach that uses a vertical database format. A vertical database format represents the list of transactions where each item appears (i.e., $tid - list$ or $tid(x)$ for itemset x). The main benefit of this format is that it makes it possible to obtain the $tid - list$ of an itemset by simply intersecting the $tid - lists$ of its included items without requiring a full scan of the dataset. The main idea of Eclat is to utilize the $tid - lists$ intersections to obtain the support value of an itemset by using the property that $\sup (x)$=$|tid(x)|$. This algorithm first scans the dataset to obtain all frequent itemsets with k items, and then it generates all candidate itemsets that include $k + 1$ items from frequent k $-$itemsets. In the next step, it gets all frequent $(k + 1)$ $-$itemsets by leaving out all the non-frequent itemsets. It repeats these steps until no other candidate itemset can be generated.

4.2.2.2 Extreme Gradient Boosting

Extreme Gradient Boosting (i.e., XGBoost) belongs to a class of learning algorithms that aim to create a strong classifier by combining many "weak" classifiers — namely boosting techniques (Chen & Guestrin, 2016). XGBoost is chosen due to its scalability, excellent performance, and its efficient training speed (Möller et al., 2016; Pavlyshenko, 2016; Tamayo et al., 2016). This technique is an enhancement of the gradient boosting decision tree, and it is used for both regression and classification problems (Friedman, 2002).

The idea of XGBoost is to build decision trees sequentially such that each subsequent tree seeks to reduce the residuals of the previous trees. At each iteration, the tree that grows next in the sequence learns from its predecessors by fitting a new model to the last predicted residuals and then minimizing the loss when adding the latest prediction. XGBoost adds an additional custom regularization to the loss function to establish the objective function.

$$obj = L(\Theta) + \Omega(\Theta) \tag{1}$$

L is the loss function that measures how well the model fits on the training data and Ω represents the regularization term, which measures the model's complexity.

For a training data set with n samples, the model is given by a function of the sum of K tress:

$$\hat{y}_i = \sum_{i=1}^{k} f_k(x_i), f_k \in F \tag{2}$$

Where x_i is the independent variable, $\hat{y}_i$ is the predicted value corresponding to the dependent variable y_i, f_k represents the tree structure, and F is the collection of all possible trees. When the model is additive, we can write the prediction value at iteration t using the following equation:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3}$$

Then the objective function at iteration t can be defined as:

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k) \tag{4}$$

Where n represents the number of samples. Chen et.al (Chen & Guestrin, 2016) define the regularization term using the following equation:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5}$$

Where γ is the minimum loss reduction required to make a further split to a terminal node in the tree, T is the number of terminal nodes in the tree, λ is the regularization parameter, and w represents the vector of scores on terminal nodes.

We can re-write the objective function as:

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + constant \tag{6}$$

At each iteration, a tree that optimizes the objective function defined in equation (6) is created. In order to optimize this function, the second-order Taylor expansion of the loss function is taken.

$$obj^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{7}$$

Where g_i and h_i are the first and second derivatives of the loss function. By solving this equation, the optimal values for w_j (weights for a given tree structure) can be calculated as:

$$w_j = \frac{G_j}{H_j + \lambda} \tag{8}$$

Where G_j and H_j can be defined as:

$$G_j = \sum_{i \in I_j} g_j \tag{9}$$

$$H_j = \sum_{i \in I_j} h_j \tag{10}$$

Where I_j represents all the samples assigned to the j-th terminal node of the tree.

Now that we have a way to learn the weights for a given tree structure, the next step is to learn the structure of the tree. A set of candidate splits are proposed for each split, and the one that minimizes the loss function is selected. This is the criterion that we seek to minimize to find the optimal split in the tree and it can be defined as:

$$s = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{11}$$

Equivalently, we seek the split that maximizes the gain:

$$Gain = \frac{1}{2}[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}] \tag{12}$$

Given an input, each tree identifies a root-to-leaf path (i.e., decision path), which results in the prediction generated by each tree. If we assume the decision path is composed of M non-leaf nodes $d(p) = \{d_1, \dots, d_m, \dots, d_M\}$. The path can be represented as:

$$p = \{(x^{f_{d_1}} \otimes \tau_{d_1}), (x^{f_{d_2}} \otimes \tau_{d_2}), \dots, (x^{f_{d_M}} \otimes \tau_{d_M})\} \tag{13}$$

Where $f_{d_m}$is the feature at the node d_m,$\tau_{d_m}$is the corresponding threshold that is used to split node d_m into two child nodes, and $\otimes \in \{">", " \preccurlyeq "\}$ represents the boolean condition on each node d_m (Zhao et al., 2019). The final prediction is the weighted sum of the predictions for each individual tree.

4.3 Materials and Methods

This section describes the methods that have been used to design the proposed visual analytics system, namely SUNRISE. In Section 3.1, we give an overview of the design process and participants. Then, in Section 3.2, we briefly explain how the overall system works. Sections 3.3 and 3.4 then describe the analytics and interactive visualization modules of SUNRISE, respectively. Finally, we explain the implementation details of SUNRISE in Section 3.5.

4.3.1 Design Process and Participants

We adopted a participatory design approach in the development of SUNRISE to understand the EHR-driven tasks from the perspective of the healthcare providers, generate design solutions, collect feedback iteratively, and thus continuously improve our proposed VA system to meet the needs and expectations of the healthcare providers (Benyon, 2013; Hettinger et al., 2017; Sedig et al., 2017). Participatory design is an iterative group effort that requires all the stakeholders (e.g., end-users, partners, or customers) to work together to ensure the end product meets their needs and expectations (Leighton, 2004). Several computer scientists, data scientists, along with an epidemiologist and a clinician-scientist were involved in the conceptualization, design and evaluation of SUNRISE. It is critical to enhance the communication between all the members of the design team since healthcare experts might have a limited understanding of the technical background of the analytical processes, and medical terms might not be very comprehensible to the members of the team with a technical background. In light of this, we asked healthcare providers to provide us with their feedback on different design decisions. We performed formative evaluations continuously at every stage of the design process. This process was essential to build trust between the system and its end-users. In our collaboration with healthcare providers, we recognized two high-level tasks to consider in the design of SUNRISE: 1) They would like to examine the relationship between different groups of laboratory test results and the disease. 2) They would like to investigate the prediction result and track the decision path to determine how reliable the prediction is based on their domain knowledge.

4.3.2 Workflow

As shown in Figure 1, SUNRISE has two components, the Analytics module and Interactive visualization module. The Analytics module utilizes Eclat and XGBoost to generate prediction models. The Interactive Visualization module encodes the data items generated by the analytics module to four main sub-visualizations: 1) selection panel, 2) control panel, 3) probability meter, and 4) decision path panel. These sub-visualizations support multiple interactions to assist the user in achieving their tasks. These interactions include selecting, drilling, searching, measuring, and inserting/removing.

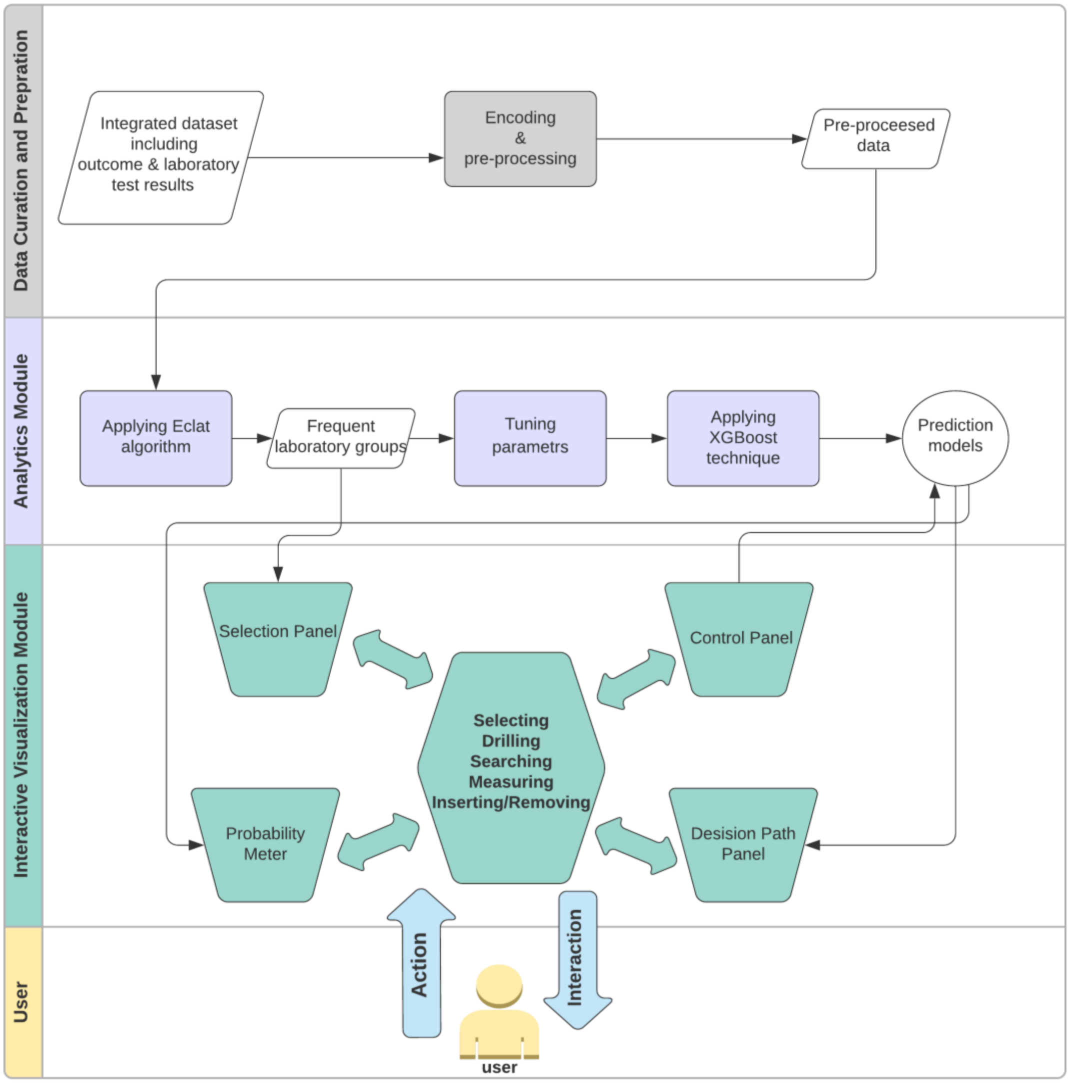


Figure 4-1: The workflow diagram of SUNRISE

The basic workflow of SUNRISE is as follows. First, we create an integrated dataset from different databases. Next, features in the laboratory test data are encoded and transformed into forms appropriate for analysis. In the next step, we apply the Eclat algorithm to the pre-processed dataset to obtain the frequent laboratory groups (i.e., frequent combinations of laboratory tests). For each laboratory group, we then create a subset of data with all the tests included in the group. In each subset, we only include rows where all the tests in the group are available. We then split each subset into train, validation, and test sets. We use the validation set to adjust the tuning parameters. Then, the XGBoost technique is applied to each subset with its corresponding tuning parameters. We develop four sub-visualizations in the Interactive Visualization module to allow the user to examine associations between laboratory groups and the outcome. The user can choose multiple laboratory tests using the selection control panel based on the result of the Eclat algorithm. When the user selects a test from the selection panel, the system then inserts a slider associated with the selected test in the input control panel. The user can probe the prediction models by creating input examples with their desired values using the input control panel. Upon clicking the "submit" button, the system passes the input (i.e., the chosen laboratory group and selected test values) to the Analytics module. The Analytics module uses the XGBoost model corresponding to the chosen laboratory group to predict the patient outcome and returns the results to the probability meter and the decision path panel. Finally, the user is able to observe the final prediction outcome and track the decision process that leads to the outcome to gain a deeper insight into the working mechanism of the prediction model.

4.3.3 Analytics Module

The Analytics module of SUNRISE generates prediction models using laboratory test data stored in EHRs by integrating the XGBoost techniques with Eclat. In this section, we describe how these techniques are combined to build the prediction models.

First, we create an integrated dataset from different databases. This dataset includes laboratory test data and the outcome for every patient. For a laboratory test, a patient might have multiple values from different times. Therefore, a sequence of laboratory test results can be formed. In order to represent this sequence for each patient, we use the average result. The outcome is considered positive if the patient develops the disease and it is considered negative otherwise. If there are many laboratory tests available, we cannot consider every possible combination of these tests because of limited memory and computational resources. Therefore, we use a frequent itemset mining technique to obtain the most frequent combinations to make the computations manageable. In order to generate more specialized prediction models, we use the Eclat algorithm to obtain frequent combinations of laboratory tests. Eclat is a fast frequent itemset mining technique that reduces memory requirements due to the use of depth-first search technique. We use the "arules" library to implement the Eclat algorithm with a specified minimum support to create several laboratory groups (i.e., frequent itemsets) from laboratory tests included in the dataset. Then for each group, we create a subset of data with all the laboratory tests that were included in the group. In order to get more accurate predictions, we only include rows where all the laboratory variables in the group are available in each subset. This approach allows us to deal with a more specialized model chosen based on the available laboratory tests in the prediction phase rather than a generalized model using the whole dataset.

In the next stage, we apply the XGBoost technique to each group. XGBoost has good model performance and can handle missing values and outliers very well. It provides various benefits, such as distributed computing, cache optimization, parallelization, and out-of-core computing. For each laboratory group, we split its corresponding subset into train, validation, and test sets to generate the prediction model. We use 80% of patients for training the model, 10% for validation and 10% for testing. The validation set is used to tune the hyperparameters when building the XGBoost model to avoid overfitting and control the "bias-variance" trade-off. We adjust the complexity of the model by changing values of tuning parameters minimum leaf weight (min_child_weight) and maximum tree depth (max_depth). Minimum leaf weight is the minimum weight that is required to generate a new node in the tree. A smaller value for this parameter allows for the generation of children that correspond to fewer samples, thus allowing for the creation of more complex trees that are more likely to overfit. Maximum tree depth is the maximum number of nodes allowed from the root of the tree to its farthest leaf. A large value for this parameter makes models more complex by letting the algorithm create more nodes. However, as we go deeper in the tree, splits become less relevant, thus causing the model to overfit. Another approach to avoid overfitting is to add randomness to make the model more robust to noise. Randomness is tuned by setting the sub-sampling rate (i.e., subsample parameter) at each sequential tree. Another parameter that can get adjusted is the model's learning rate (i.e., eta) that determines the contribution of each tree to the overall model. A low learning rate should result in better performance, but it will increase the computational cost. The final model is a linear combination of all trees in the sequence, along with their contributions weighted by the learning rate. In order to identify the best combination of parameters for each laboratory group, we use the random search approach, which is shown to have higher efficiency compared to a manual search and grid trials when given the same computation time. Another advantage of random search is that as opposed to the manual search, results obtained through random search are reproducible (Bergstra & Bengio, 2012). We use the combination of parameters with the best performance on the validation set to train the final model for each laboratory group.

We use the xgboost library in r to implement XGBoost and use the area under the receiver operating characteristic curve (i.e., AUROC) (Ferri et al., 2009; Garcıa et al., 2012) to measure the performance of all the models and choose the best combination of tuning parameters. A ROC curve shows the trade-off between specificity and sensitivity across different decision thresholds (i.e., threshold that is used for interpreting probabilities to class labels). Sensitivity measures how often a model classifies a patient as "at-risk" correctly. On the other hand, specificity is the capacity of a model to classify a patient as "risk-free" correctly (Parikh et al., 2008).

4.3.4 Interactive Visualization Module

The Interactive Visualization module is composed of four main sub-visualizations: the selection panel, control panel, probability meter, and decision path panel (Figure 2). In this section, we describe how data items that are generated in the analytics module are mapped into visual representations to allow healthcare providers to accomplish their tasks.

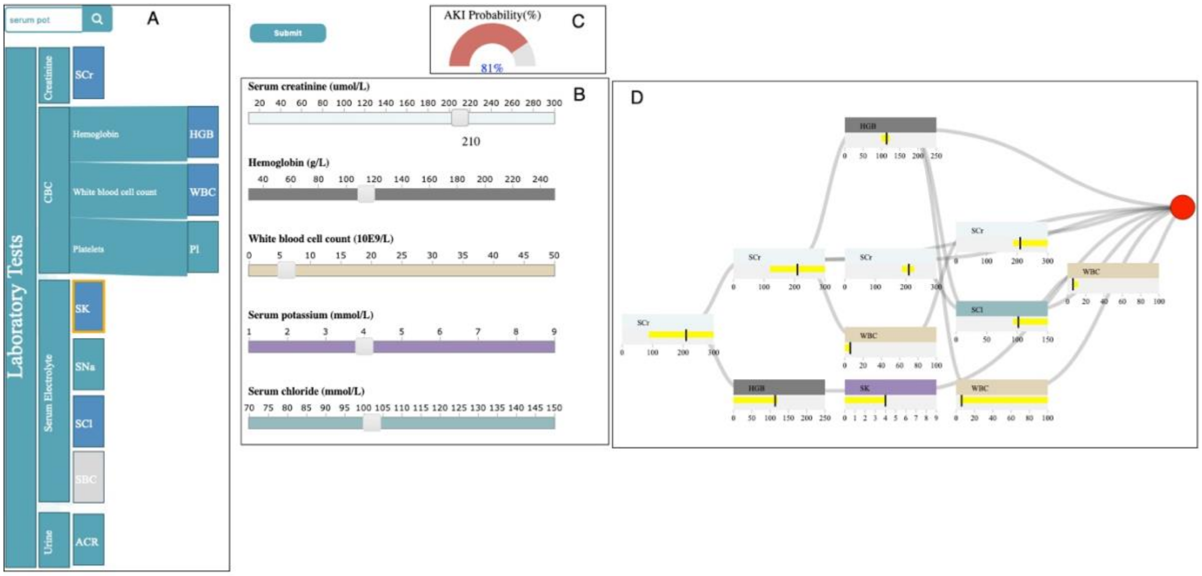


Figure 4-2: SUNRISE includes four main sub-visualizations: Selection Panel(A), Control Panel (B), Probability meter (C), and Decision Path Panel (D).

4.3.4.1 Selection Panel

The selection panel displays the hierarchical structure of the laboratory data using horizontally stacked rectangles ordered from left to right (Figure 2-A). The first rectangle to the left (i.e., root) that represents the laboratory tests takes the entire height. Each child node is placed to the right of its parent with the height proportional to the percentage it consumes relative to its siblings.

The selection panel utilizes the result of the Eclat algorithm from the Analytics module to allow the user to select their desired group of laboratory tests. The user can choose a test by clicking on its corresponding rectangle in the selection panel. This action changes the color of the selected rectangle from green to blue. When a test is selected, all the other rectangles corresponding to tests that are not in any laboratory group with the selected test become un-clickable and greyed out. The user can also insert/remove a slider corresponding to a test in the control panel by clicking/unclicking the rectangle corresponding to that test in the selection panel. We will describe the control panel in more detail in the next section.

The selection panel allows the user to observe the full name of laboratory tests that belong to a category by clicking on the rectangle corresponding to that category. In addition, when the user hovers the mouse over any of the rectangles, a tooltip with information regarding the test shows up. The selection panel is supported by a search bar. If the user enters the name of a specific laboratory test in the search bar, the border of the rectangle corresponding to the specified test becomes orange.

4.3.4.2 Control Panel

The control panel includes sliders corresponding to the laboratory tests that the user has chosen in the selection panel (Figure 2-B). It allows the user to probe the prediction models by creating input examples with their desired values and observe the output the model generates. When the user selects a test from the selection panel, the system inserts a slider associated with the selected test in the input control panel. Each slider is composed of a label including the full name and unit of measurement of its corresponding test, a horizontal axis with a linear scale representing the possible values of its associated test, and a rectangular handle that allows the user to change the values of the test. This panel allows the user to interactively tweak the values of the selected tests and see how the predictive model responds. The user can hover the mouse over the handle to observe the chosen value in any of the sliders.

When the user clicks on the "submit" button after selecting multiple tests from the selection panel and choosing the values of each test using their corresponding sliders, the system passes the information regarding the chosen laboratory group and selected test values to the Analytics module. The Analytics module uses the corresponding XGBoost prediction model that is associated with the selected group to predict the outcome and returns the results to the probability meter and the decision path panel.

4.3.4.3 Probability Meter

The probability meter is a radial gauge chart with a circular arc that shows the probability of developing the outcome (Figure 2-C). This probability is the outcome prediction after the system feeds the input (i.e., chosen laboratory group and laboratory test values) to its corresponding XGBoost model. The value inside the arc represents the probability. If the probability is less than 50 percent, then the shading of the arc is green; otherwise it is red.

4.3.4.4 Decision Path Panel

The Decision Path Panel allows the user to investigate the decision process of a prediction outcome to ensure its corresponding XGBoost model works appropriately when given an input (i.e., chosen laboratory group and laboratory test values) (Figure 2-D). The final prediction outcome in the XGBoost model is the additive sum of all the interim predictions from each individual tree, where each interim prediction has a unique decision path. Therefore, summarizing the structure of all the decision paths that lead to the final prediction can deepen the understanding of the model's working mechanism. Thus, this panel is designed to assist the user investigate the decision paths by summarizing the critical ranges of the laboratory tests involved in the chosen laboratory group and providing the detailed information of the decision paths layer by layer.

In order to reveal the structure of the decision paths that lead to the final prediction, we first summarize the features (i.e., laboratory tests included in the chosen group) at the layer level. A feature may appear multiple times at each layer of all the decision paths (Eq. 13) for an input data point of $x = \{x^1, x^2, \dots, x^H\}$ with H features of Q=$\{q^1, q^2, \dots, q^H\}$. In each layer l_i of all the decision paths, for each feature $q^h \in q_{l_i}$, we merge the ranges on q^h to $[\tau_{l_i}^{q^{h,l}}, \tau_{l_i}^{q^{h,u}}]$, where $\tau_{l_i}^{q^{h,l}} = Min\{\tau_{d_h} | q_{d_h} = q^h, \otimes = "\succ", d_h \in l_i\}$, and where $\tau_{l_i}^{q^{h,u}} = Max\{\tau_{d_h} | q_{d_h} = q^h, \otimes = " \preccurlyeq ", d_h \in l_i\}$.

We represent these summarized features using feature nodes. Each feature node summarizes the feature ranges for each laboratory test in each layer using a horizontal bar chart. The x-axis uses a linear scale to represent the possible values of the laboratory test associated with the feature node. The vertical bar represents the laboratory test value of the current input. The color of the feature node is identical to the color of its corresponding laboratory test slider in the input control panel. The user can hover the mouse over the feature node to observe the summarized ranges associated with the node.

We create a decision path flow by connecting the feature nodes from different layers using ribbons. The tooltip of a ribbon displays the pair of feature nodes that are connected by the hovered ribbon. This allows the user to examine the order of the features that appeared in the decision paths, which is very critical in measuring the importance of each feature. In the decision path panel, each column represents a layer where the layer depth increases from left to right. This supports the user in understanding how the ranges from each feature evolve from root to the terminal node (i.e., leaf). We append a circle to the decision path flow to encode the leaf that represents the final prediction outcome. If the probability of developing the outcome for the input data point is less than 50 percent, then the color of the circle is green; otherwise it is red (i.e., similar to the probability meter). The tooltip of the circle displays the probability of the outcome for the given input.

4.3.5 Implementation Details

SUNRISE is implemented using Java programming language, R packages, JavaScript library D3, Ajax, SAS, and standard HTML. D3 and HTML were used to implement the front end of the system, which includes the interactive visualization module. Several R libraries were used to develop the Analytics module. SAS was used to cut and integrate the data from multiple sources since the ICES data is stored in the SAS server. The communication between Analytics and Interactive Visualization modules is implemented using AJAX and Java.

We used R to develop the Analytics module since it 1) provides several packages to perform Eclat and XGBoost, 2) is open-source and platform-independent, and 3) is available in the ICES workstations.

We chose D3 to develop the Interactive Visualization module because it 1) provides a data-driven approach to attach data to the Document Object Model elements. 2) is an open-source library, 3) supports users in using the full capabilities of modern web-browsers, and 4) is compatible with other programming languages that have been used in this system.

4.4 Usage Scenario

In this section, we demonstrate how SUNRISE can assist healthcare experts in studying associations between laboratory test results and AKI using the data stored at ICES.

4.4.1 Data Description

We used a data cut that contained nine laboratory test results and the outcome of AKI for 229,620 patients, which were obtained from three health administrative databases (see Appendix A) from ICES. These datasets were linked using unique, encoded identifiers that were derived from patient health card numbers and were analyzed at ICES. We obtained outpatient albumin/creatinine ratio (ACr), serum creatinine (SCr) , serum sodium (SNa), serum potassium (SK), serum bicarbonate (SBC), serum chloride (SCl), hemoglobin (HGB), white blood cell count (WBC), and platelets (Pl) measurements from the Dynacare medical laboratories, which represents around one third of outpatient laboratory testing for Ontario residents. A 365 days lookback window was used to obtain the outpatient laboratory test data. Hospital admission codes and emergency department (ED) visits were identified from the National Ambulatory Care Reporting System and the Canadian Institute for Health Information Discharge Abstract Database (hospitalizations). ICD-10 (i.e., International Classification of Diseases, post-2002) codes were used to identify the incidence of AKI from ED visit and hospital admission data. The cohort included senior patients aged 65 years or older who visited ED or were admitted to hospital between April 1, 2014 and March 31, 2016. The hospital admission date or ED visit date served as the index date. If an individual had multiple hospital admissions or ED visits, the first incident was selected.

4.4.2 Outcome

AKI was the outcome variable for all the prediction models in this case study (Liangos et al., 2006a; Waikar et al., 2006). AKI is defined as a sudden deterioration of the kidney function in a short period of time (Liangos et al., 2006a; Waikar et al., 2007). The management and diagnosis of AKI can be a challenging task because of its complex etiology and pathophysiology. In the process of AKI diagnosis, the available information is complemented by additional data, which is obtained from patients' medical history and different diagnostic tests, including laboratory tests. Laboratory tests play a crucial role in the detection and diagnosis of AKI. The incidence of AKI was captured using the National Ambulatory Care Reporting System and Canadian Institute for Health Information Discharge Abstract Database based on the ICD-10 (International Classification of Diseases - Tenth Revision) diagnostic codes (i.e., "N17"). If an individual had multiple episodes of AKI, the first episode was selected. Positive cases were the ones in which AKI was acquired during the index date (i.e., 6,743), and negative cases were those when AKI was never developed (i.e., 222,877).

4.4.3 Case Study

First, the features in the laboratory test data are encoded and transformed into forms appropriate for analysis. For instance, if there is more than one result for a test on a patient, the average result is used. Thus, we created nine variables for each laboratory test reported in the past year prior to the index date for each patient. Then we apply Eclat with the minimum support of 0.05 to obtain the most frequent combinations of laboratory tests. At this stage, a total of 263 laboratory groups (i.e., frequent itemsets) were created from nine laboratory tests (see Appendix B). Next, we create a subset of data for each group only including rows where all tests in the group are available.

Generally, in most of the laboratory groups the prevalence of AKI is lower than 2.5 percent, which leads to an imbalanced class ratio. This issue can severely reduce the prediction performance, as most classifiers are developed to maximize the total number of correct predictions, and thus are more sensitive to the majority class. Therefore, if the imbalance issue is not addressed properly, then the classification result can be biased towards the majority class leading to poor performance on the prediction of AKI. The misclassification of AKI, including false positive and false negative cases affects the choice of treatment and prognosis, which consequently might increase the overuse of clinical resources and the risk of deterioration in the patient's condition. To address this issue, we set the weight of positive class (i.e., scale-pos_weigth) parameter in the XGboost models using following equation:

$$\text{Scale-pos-weight}=\sqrt{\text{number of non} - \text{AKI cases in each subset}/\text{number of AKI cases in each subset}} \qquad 14$$

After adjusting tuning parameters for each subset, we apply XGBoost with 100 trees and its corresponding tunning parameters and scale-pos-weight parameter to each subset (Table B1). Thus, we create 263 XGBoost prediction models in total.

As shown in Figure 3, the laboratory tests are classified into four categories: Creatinine, Complete Blood Count (CBC), Serum electrolyte, and Urine. Creatinine refers to SCr. CBC is composed of HGB, WBC, and Pl. Serum electrolyte contains SNa, SK, SBC, and SCl and Urine includes ACR. Now, let's assume the user is interested in exploring the relationship between SCr, SK, and SCl and AKI. The user can first select the rectangle corresponding to serum electrolytes to open it up and observe the full laboratory names included in that group and then select the rectangles corresponding to SCr, SK, and SCl in the selection panel. Upon selection, the system inserts a slider corresponding to the chosen test in the control panel. The system allows the user to probe the prediction model by generating input examples for their chosen tests using sliders in the control panel. As shown in Figure 3, the user has selected the SCr value of 70 umol/L, SK of 4 mmol/L, and SCl of 102 mmol/L through corresponding sliders. Upon submission, the Analytics module uses the XGBoost model generated with the subset of data including SCr, SK, and SCl to predict AKI with the input values and returns the result to the probability meter and the decision path panel. The probability meter in Figure 3 shows that the probability of developing AKI for a patient with the chosen values for SCr, SK, and SCl is 22 percent.

To ensure the prediction is reliable, the user examines the decision path panel to check the result (Figure 4). As shown in Figure 4, the user can observe that SCr is the only feature that appears in the first and second layers. Since we expect features near the root of the path to be more important than features near the leaves, SCr has higher importance than SCl, and SK when predicting AKI given the input. Also, if the user hovers the mouse over the SCr feature node in the root layer, they can see the split threshold for that

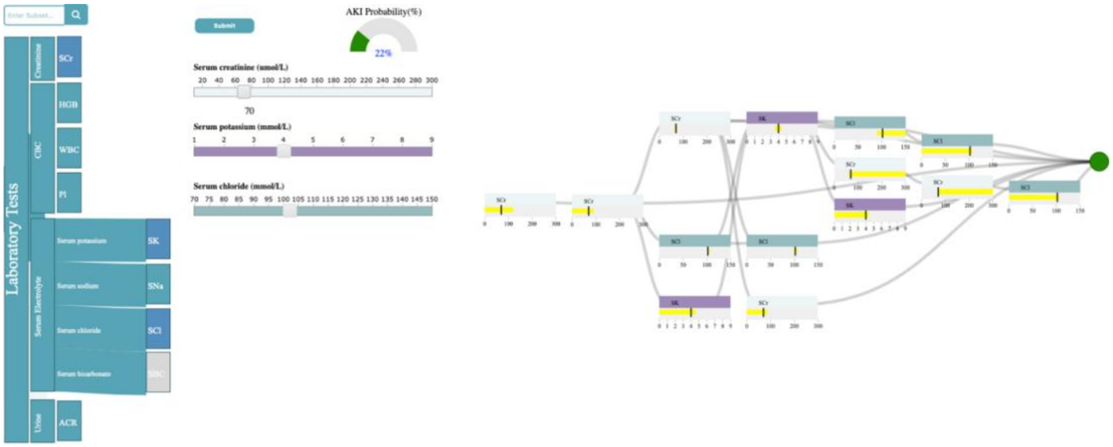


Figure 4-3: Shows how the probability meter and the decision path panel get updated upon submission with the user input of SCr value of 70 umol/L, SK of 4 mmol/L, and SCl of 102 mmol/L.

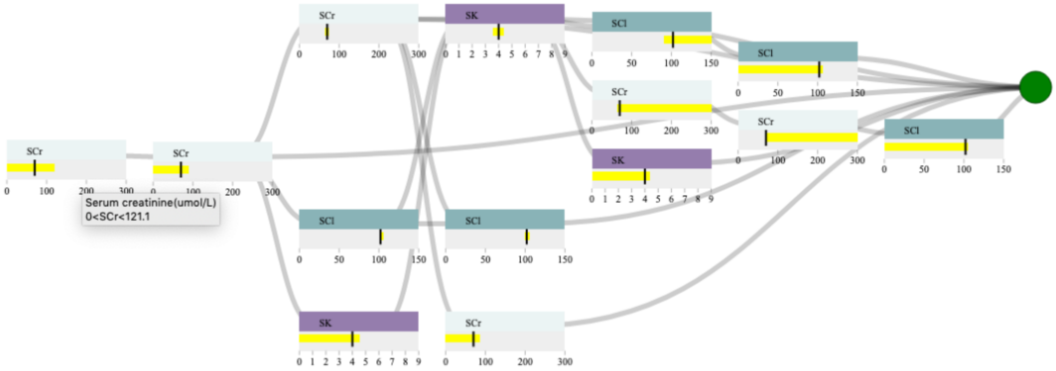


Figure 4-4: Shows how the decision path panel can help the user to ensure the reliability of the results.

specific node (i.e., SCr>121.1 umol/L). This information can guide the user to observe how the probability of developing AKI changes if they increase the SCr from 70 to 140 (i.e., a value greater than 121.1 umol/L for SCr). Figure 5 shows that the AKI probability is risen to 68 percent by increasing the SCr value. The user then might be curious to explore the association between SK, SCl and AKI. In this case, they can click on SCr rectangle in the selection panel to remove its corresponding slider from the control panel. Let's assume the user wants to observe how the changes in SK level would affect the probability of AKI. If the user increases SK to 6 mmol/L (high potassium level), the probability of AKI becomes 80 percent (Figure 6). The user can then observe the feature ranges and the path that led to this probability. For instance, they can observe that the split points for SK are around 4.7 to 5.1 mmol/L, which suggests that this range is critical when using SK in predicting AKI. They can also observe that SK and SCl have similar importance in AKI prediction based on the order they appear on the decision path.

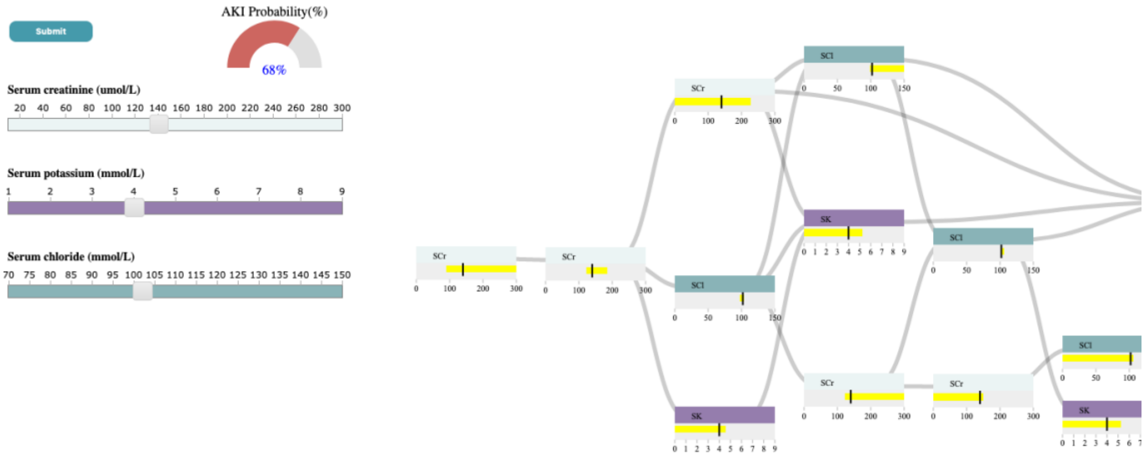


Figure 4-5: Shows how the system gets updated when the user increases the value of SCr from 70 to 140 (umol/L).

4.5 Discussion and Limitations

The purpose of this paper is to: 1) show how VA systems can be designed to examine relationships between laboratory test results and a specific disease outcome and 2) study the structure and the working mechanisms of the risk prediction models. To accomplish these tasks, we have reported the development of SUNRISE, a VA system designed to support healthcare providers. SUNRISE incorporates two main components: the Analytics module and Interactive visualization module. The Analytics module integrates a frequent itemset mining technique (i.e., Eclat) with XGBoost to develop risk prediction models. The Interactive Visualization module then maps the data items generated by the analytics module to four main sub-visualizations—namely, the selection panel, control panel, probability meter, and decision path panel. SUNRISE is unique in how it integrates XGBoost with Eclat to develop prediction models, and it allows the user to interact with

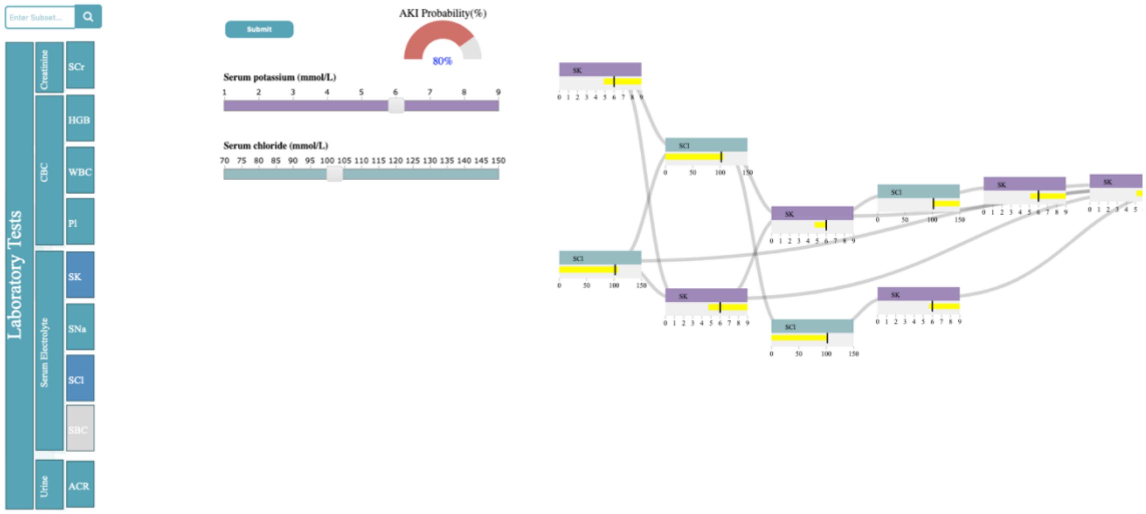


Figure 4-6: Shows how the system gets updated when the user removes SCr from the control panel and increase the value of SK to 6 (mmol/L).

the model and audit the decision process through multiple interactive sub-visualizations. SUNRISE provides a balanced distribution of processing load through the seamless integration of computational techniques (i.e., frequent itemset mining and XGBoost in the Analytics module) with interactive visual representations (i.e., sub-visualizations in the Interactive Visualization module) to support the user's cognitive tasks. It provides the user with the means to probe the prediction model by creating input instances and observing the model's output. Furthermore, it allows the user to examine how a particular input example's risk might change if it had different values. Finally, SUNRISE helps the user gain deeper insight into the underlying working mechanism of the model, increasing their confidence in the generated predictions.

In terms of generalizability, SUNRISE is designed in a modular way so that it can easily accept new data sources and data types. SUNRISE can be used to investigate other clinical problems, such as exploring the association between medication dosage and diabetes. Although SUNRISE focuses on making XGBoost interpretable, a similar approach can be applied to other tree-based ensemble techniques such as Random forest. Random forest uses several decision trees and generates final predictions by summarizing the output of all internal trees. Unlike XGBoost, decision trees in Random forest are trained independently. One potential enhancement to support Random forest is to summarize the paths based on whether they generate positive predictions or negative ones and let the user compare them in the same view.

One of the primary considerations in the design of SUNRISE is scalability. To make the control and decision path panels less cluttered because of the user's limited visual capacity when the number of laboratory tests increases, we restrict the maximum number of tests that can get inserted into the control panel by adjusting the minimum support parameter of Eclat.

This research has several limitations. The first limitation is that, although we used a participatory design approach and medical researchers have assessed SUNRISE and found it useful, we have not conducted any formal usability studies to assess SUNRISE's performance and the efficiency of its interaction mechanisms. Second, the decision path panel may not function properly if the number of layers in the XGBoost trees gets higher due to screen space limitations and computational resources. Third, as we use curves to link the feature cells from different layers in the decision path panel, these curves might have overlapping problems. Another limitation is that the prediction models might be prone to overfitting because a small validation set might lead to an unstable model at a particular hyperparameter set. This will result in validation error measurements that are overoptimistic. Finally, we aggregated laboratory test results for a patient by taking the average of their test results in the past 365 days before the index date. As such, we might have lost vital information regarding laboratory tests. To address this issue, in future versions, we plan to offer the user different aggregation functions such as the trend of change (i.e., increase/decrease) in tests over a certain period of time.

4.6 Conclusion

The overall goal of this paper is to show how VA systems can be designed systematically to support the investigation of different clinical problems. We report the development of a VA system—namely, SUNRISE and demonstrate how it can be employed to assist healthcare providers in exploring associations between laboratory test results and a disease outcome. SUNRISE's novelty stems from its design: it incorporates XGBoost, frequent itemset mining, visualization, and human-data interaction mechanisms in an integrated manner to support complex EHR-driven tasks. We illustrate SUNRISE's value and usefulness through a usage scenario of investigating and exploring relationships between laboratory test results and AKI using the data stored at ICES. We demonstrate how it can help clinicians and researchers at ICES probe the AKI risk prediction models by hypothesizing input examples and observing the model's output. They can also audit the decision process to verify the reliability of the prediction models. Finally, the design concepts employed in SUNRISE are generalizable. They can be utilized to systematically develop any VA system whose purpose is to support clinical tasks involving investigation and analysis of EHR data using XGBoost and frequent itemset mining. Applications of such VA systems result in the emergence of best practices for developing similar VA systems in other medical domains.

Chapter 5

5 VERONICA: Visual Analytics for Identifying Feature Groups in Disease Classification

This chapter has been submitted to Computers in Biology and Medicine.

Please note that the format has been changed to match the format of the dissertation. Figure, Section, and Table numbers mentioned herein are relative to the chapter number. For instance, “Figure 1” corresponds to Figure 5-1. Additionally, when the term “paper”, “work”, or “research” is used, it refers to this particular chapter.

5.1 Introduction

A key component of precision medicine is to determine a person’s individualized estimates of different health outcomes, which then guides therapy to increase the chance of long-term good health. Identifying the group of features in electronic health records (EHRs) with the most substantial predictive power helps in the development of robust predictive models (Bengio et al., 2013; Jordan & Mitchell, 2015). The data in EHRs has great promise for improving predictive risk modeling (Hersh, 2007). However, EHRs are often challenging to analyze due to their high dimensionality (Jensen et al., 2012; Weiskopf & Weng, 2013). In recent years, several studies have incorporated various data mining and machine learning techniques to address this problem. Most of the existing studies use unsupervised learning techniques such as principal component analysis (Hotelling, 1933), K-means (Hartigan & Wong, 1979; Jain, 2010), and hierarchical clustering (Nielsen, 2016) to find the best representative group of features in high dimensional EHRs (Alexander et al., 2020; Lütz, 2019; Khalid et al., 2018; Miotto et al., 2016, 2016; Lasko et al., 2013; Marlin et al., 2012; L. Wang et al., 2020; Panahiazar et al., 2015; Langavant et al., 2018). Although these unsupervised techniques have shown promise in managing high dimensional data, to our best knowledge, this problem has not been studied thoroughly using supervised techniques (S. Abdullah, 2020; S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020b). One of the main issues with both supervised and unsupervised techniques is that they suffer from a lack of interpretability and transparency (Caruana et al., 2008; Johnstone & Titterington, 2009; D. A. Keim et al., 2010). In healthcare settings, it is essential to better understand how a given technique works. Therefore, increasing a technique's interpretability by involving humans in the analytic process can play a vital role in building trust with users (Krause, Perer, & Bertini, 2016; Krause, Perer, & Ng, 2016; S. Liu et al., 2017; Spinner et al., 2020; X. Zhao et al., 2019). The analytic results can be made accessible to users through visual analytics (VA) to address these issues.

Visual analytics (VA) integrates data analytics techniques with interactive visualizations to improve users' capabilities in performing data-driven tasks (D. A. Keim et al., 2010; Ola & Sedig, 2014). It enables users to achieve their goals through interactive exploration and manipulation of the data (Parsons et al., 2015; Allan F. Simpao et al., 2014). The design of a VA system is often not straightforward because it requires the designer to consider users’ activities and tasks, the structure of the data, and human factors (Sedig et al., 2012; S. Abdullah, 2020). Thus, the designer needs to make several non-trivial decisions when developing such systems. For instance, one needs to consider which techniques to use, which features and samples to incorporate, and what level of granularity to look for when choosing a data analytics technique (Ola & Sedig, 2014). Similarly, it is important to determine how to map and classify data items and help users accomplish their tasks when developing interactive visualizations. Consequently, combining analytic techniques with interactive visualizations becomes a more complex challenge. Thus, it is important to involve stakeholders (e.g., clinical researchers and medical practitioners) in the design and development process of a VA system (Leighton, 2004).

The purpose of this paper is to show how VA systems can be designed systematically to identify the best representative subset (i.e., a combination of groups) of high-dimensional EHRs. The proposed VA system, VERONICA (Visual analytics for idEntifying featuRe grOups iN dIsease ClAssification), takes advantage of the group structure of features stored in EHRs. EHRs are generally classified into different groups: comorbidities, medications, laboratory tests, hospital encounter codes, and demographics. It is possible to combine these groups to create multiple subsets of groups. For instance, one can create a subset by combining all features from both comorbidity and demographic groups.

Depending on the predictive power of features within them, some groups or subsets (i.e., combinations of groups) are stronger predictors in identifying diseases in comparison to others. To identify the subset with the most substantial predictive power, VERONICA considers every possible subset of groups (i.e., groups of features) and applies several supervised learning techniques to each subset. It allows users to compare the results based on different performance measures through an interactive visual interface. VERONICA aims to assist healthcare providers at ICES-KDT—where ICES is a non-profit, world-leading research organization that utilizes population-based health data to produce knowledge on a broad range of healthcare issues, and KDT refers to the Kidney Dialysis and Transplantation Program located in London, Ontario, Canada. We utilize the clinical dataset housed at ICES to identify the best representative feature groups in detecting patients with high risk of developing acute kidney injury to demonstrate VERONICA's utility and usefulness.

The rest of the paper is organized as follows. Section 2 gives an overview of the conceptual and terminological background to understand the design of VERONICA. Section 3 briefly describes existing EHR-based VA systems. Section 4 explains the methodology used for the design of VERONICA. Section 5 presents VERONICA by describing its structure and components. We address the limitations of the system in Section 6. Finally, Section 7 discusses the conclusions and future areas of application.

5.2 Background

In this section, we present the terminological and conceptual background to understand the design of VERONICA. We discuss different components of VA systems to provide a better understanding of the concept of VA. Finally, we provide a summary of the chosen machine learning techniques—namely, classification and regression tree (CART) (Wilkinson, 2004), C5.0 (Quinlan, 2014), random forest (Breiman, 2001), naïve Bayes (NB) (Lewis, 1998), and support vector machine (SVM) (Cristianini et al., 2000).

5.2.1 Visual Analytics

Visual Analytics (VA) systems combine the strengths of data analysis and interactive visualizations to enable users to apply filters, and explore and manipulate the data interactively to accomplish their goals (J. J. Thomas & Cook, 2005). The processing load in VA is distributed between users and the key components of the system—namely, the analytics module and interactive visualization module (Cui, 2019; Jeong et al., 2015; D. A. Keim et al., 2010; Ola & Sedig, 2014; Parsons & Sedig, 2014; Sedig & Parsons, 2013).

5.2.1.1 Analytics Module

The analytics module is responsible for storing, pre-processing, transforming, and performing computerized analysis on the data. It involves three main stages: data pre-processing, data transformation, and data analysis (Ola & Sedig, 2014). The raw data retrieved from different sources gets processed in the pre-processing stage. This stage involves tasks such as fusion, integration, cleaning, and synthesis (Han et al., 2011). Then in the transformation stage, the pre-processed data is transformed into a form suitable for analysis. Common tasks in this stage include smoothing, aggregation, normalization, and feature generation (Han et al., 2011). Finally, the analysis stage involves the discovery of hidden patterns and relationships and allows for the extraction of useful and novel information from the data (R. Agrawal, Swami, et al., 1993; Sahu et al., 2008). This can be done by applying various statistical and machine learning techniques (e.g., random forest, SVM, NB, and decision trees) to the transformed data. However, despite all the benefits, most of these computational techniques do not support proper exploration and manipulation of the computed results (D. A. Keim et al., 2010). VA systems address this problem by allowing users to engage in a more involved discourse with the data through interactive visualizations.

5.2.1.2 Interactive Visualization Module

In VA systems, the interactive visualization module is composed of a mapping component that retrieves the analyzed data from the analytics module and generates interactive visual representations. It allows users to change the displayed information, modify the subset of the information displayed, and guide and control the intermediary steps of the analytical processes within the analytics module. This, in turn, incites a chain of reactions that leads to the execution of additional analysis processes. The interactive visualization module provides users with flexibility and supports their cognitive and perceptual needs as they engage in various complex tasks. However, despite the advantages of interactive visualizations in amplifying users' cognitive needs, they fell short when confronted with data-intensive tasks that require computational analysis (D. A. Keim et al., 2010). Therefore, an approach that integrates analytical processes with interactive visualizations through VA is required to overcome these challenges (Kehrer & Hauser, 2013; D. Keim, Mansmann, et al., 2008).

5.2.2 Machine Learning Techniques

In this section, we give a brief overview of all machine learning techniques used in VERONICA.

5.2.2.1 Decision Tree

Decision trees are among the most popular and powerful classification techniques that can provide informative models (Breiman et al., 1984). They construct a set of predictive rules to solve the classification problems using the recursive partitioning process. In their simplest form (e.g., C4.5 (Quinlan, 2014)), each feature is tested and ranked based on its ability to split the remaining data. The training data is propagated through the decision tree branches until enough features are chosen to correctly classify them. The classifier has a tree-like structure where each of its leaf nodes corresponds to a subset of the data that belongs to one class. Two widely known methods for generating decision trees are Classification and Regression Trees (CART) and C5.0. CART is based on a tree-growing algorithm that uses the GINI index as its splitting criteria. The strategy is to choose the feature whose GINI Index is minimum after each split. On the other hand, C5.0 builds the tree by splitting based on the feature that yields the most considerable information gain (Entropy). These classifiers are robust in handling missing values since the tree-growing process is not affected by missing data (Ismail & Anil, 2014). However, despite all the benefits, they tend to over-fit the training data (Deng et al., 2011). Random forest addresses this problem by generating an ensemble of decision trees where each tree is built from a random arrangement of features (Breiman, 2001; Liaw & Wiener, 2002). A new object passes through every tree in the forest to get classified based on a vector of features. Each distinctive tree gives a classification and votes for the class. The final classification of the new object is based on the majority "vote" of all the trees in the forest.

5.2.2.2 Support Vector Machines

Support Vector Machines are among the most successful and robust classification techniques (Cortes & Vapnik, 1995; Cristianini et al., 2000). They aim to identify an optimal separating hyperplane that can distinctly divide the instances of multiple classes in a multi-dimensional space by maximizing the minimum distance from the hyperplane to the closest instance. Although models produced by SVM are often hard to interpret and understand, they work well on classification tasks involving a large number of features (Ghaddar & Naoum-Sawaya, 2018). SVM is first outlined for the linearly separable classification problems, but a linear classifier might not be the most appropriate candidate for the binary classification. SVM can support non-linear decision surfaces using kernel functions. Due to its good generalization ability and its low sensitivity to the curse of high-dimensionality, SVM is often used in many classification problems.

5.2.2.3 Naive Bayes

Naive Bayes is a simple and powerful probabilistic classifier that often creates stable and accurate models (Lewis, 1998). The model is based on the probability of each class and the conditional probability of each class given each feature. These probabilities that are directly calculated from the data can be used for the classification of new data based on the Bayes theorem. Naive Bayes makes a simplistic assumption that all the features are independent of one another. Despite this assumption and its simplistic design, it can be very efficient, particularly when the data is high-dimensional.

5.3 Related work

In this section, we describe some of the existing EHR-based VA systems. MatrixFlow (Perer & Sun, 2012) is a VA system that helps users discover complex temporal patterns across patient groups in EHRs. Likewise, Simpao et al. (A. F. Simpao et al., 2014) developed a VA system to support the management of adverse drug alerts in EHRs to enhance drug safety. It helps clinicians to explore not only alert types and patient characteristics but also medication alerts. VisualDecisionLinc (Mane et al., 2012) is another VA system that facilitates the interpretation of EHRs by providing summaries of patient outcomes and treatment options in a dashboard. It allows users to identify patient subpopulations with similar clinical characteristics to assist them in the decision-making process. Visual Temporal Analysis Laboratory (ViTA-Lab) (Klimov et al., 2015) is a VA-based framework designed to investigate temporal medical data. It combines longitudinal data mining techniques with query-driven visualizations to help users discover temporal patterns within time-oriented medical data. Care Pathway Explorer (Perer et al., 2015) enables clinical researchers to identify common event sequences and examine how they are connected with patient outcomes. It combines a sequence mining procedure with an interactive visual interface to achieve this. PHENOTREE (Baytas et al., 2016) is another VA system that enables interactive exploration of patient groups and the interpretation of hierarchical phenotypes by combining principal component analysis with an interactive user interface. VISA_M3R3 (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020a; S. S. Abdullah, Rostamzadeh, Sedig, Lizotte, et al., 2020) is a VA system that assists clinical researchers in identifying medications that are associated with a higher risk of acute kidney injury. It integrates frequent itemset mining, multivariable regression, and interactive visualizations to support investigation of high-risk medications. Finally, VALENCIA (S. S. Abdullah, Rostamzadeh, Sedig, Garg, et al., 2020b) is another VA system that incorporates various clustering and dimensionality reduction techniques with interactive visualizations to allow exploration of high dimensional data stored in EHRs.

5.4 Materials and Methods

In this section, we describe the methods we used to design the proposed VA system—namely, VERONICA.

5.4.1 Design process and Participants

The design and development of VA systems is an integrated process that requires various sets of skills and expertise. In light of this, we adopted a participatory design approach to obtain the needs and requirements of the healthcare providers and to understand the real-world EHR-driven tasks that they perform. Participatory design is a co-operative approach that places users at the center of the design process. It is an iterative group effort that requires all the stakeholders to work together to ensure the system meets their expectations (Leighton, 2004). An epidemiologist, a clinician-scientist, a statistician, data scientists, and computer scientists were involved in the conceptualization, design, and evaluation of this VA system. It is important to optimize the communication between all stakeholders involved in the process because they might experience a language gap due to their different backgrounds. For instance, it is critical to ensure that the medical terms are comprehensible to the team members with a technical background, and the motivations of the analysis process and design decisions are well-addressed across the team. In light of this, we asked healthcare experts to provide us with their formative feedback on different design decisions and a list of tasks they perform on EHRs. Multiple participatory design approaches are used to obtain the healthcare providers' needs and to identify opportunities that can significantly improve the VERONICA’s performance through more effective visualizations and analysis techniques.

5.4.2 Data Sources

We formed a derivation cohort using large linked administrative healthcare databases held at ICES. We ascertained hospital and patient characteristics, outcome, and drug use from five administrative databases (see Appendix C). These datasets were linked using unique encoded identifiers that were derived from health card numbers of patients and were analyzed at ICES. The Ontario Drug Benefit Program database is used to identify prescription drug use. This database contains highly accurate patient records of all outpatient prescriptions administered to patients aged 65 years or older, with an error rate of less than 1% (Levy et al., 2003). We acquired vital statistics from the Ontario Registered Persons Database, which includes demographic data on all Ontarians who have ever been issued a health card. We identified baseline comorbidity data, ED visits, and hospital admission codes from the National Ambulatory Care Reporting System and the Canadian Institute for Health Information Discharge Abstract Database (hospitalizations). We used ICD-10 (i.e., International Classification of Diseases, post-

2002) codes to identify hospital encounter codes and baseline comorbidities. In addition, baseline comorbidity data and health claims for physician services were acquired from the Ontario Health Insurance Plan database. All the coding definitions for the comorbidities are provided in Appendix D and E.

5.4.3 Cohort Entry Criteria

We created a cohort of patients aged 65 years or older who were admitted to a hospital or visited an emergency department (ED) between 2014 and 2016. The discharge date from the hospital or ED served as the index date, also referred to as the cohort entry date. If a patient had multiple ED visits and hospital admissions, we chose the first incident. Individuals with invalid data regarding age, sex, and the health-card number were excluded. In addition, we excluded individuals who: (1) previously received a kidney transplant or dialysis treatment as the assessment of acute kidney injury is usually no longer relevant in patients with end-stage kidney disease; (2) left the hospital or ED against medical advice or without being seen by a physician; and (3) had acute kidney injury recorded during their hospital admission or ED visit prior to hospital discharge, as acute kidney injury was already present prior to the follow-up period. The diagnosis codes for the exclusion criteria are shown in Appendix F.

5.4.4 Response Variable

Acute Kidney Injury (AKI) is defined as a sudden deterioration of the kidney function. It is associated with a lower chance of survival, prolonged hospital stays, subsequent morbidity after discharge, and incremental healthcare costs (Collister et al., 2017; Liangos et al., 2006a). A system that detects early AKI or predicts its clinical manifestations with considerable lead-time allows healthcare experts to provide more effective treatments to prevent AKI. We build models to predict hospital admission with AKI within 90 days after being discharged from ED or hospital. The incidence of AKI is identified using the Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System based on ICD-10 diagnostic codes (i.e., "N17").

5.4.5 Input Features

The final cohort includes 162 unique features. These features can be classified into five groups—namely, demographics, comorbidities, hospital encounter codes, general practitioner (GP) visits, and medications. The demographic group includes four features—namely, age, sex, region, and income quintile. The comorbidity group contains ten known risk factors of AKI, including diabetes mellitus, chronic kidney disease, chronic liver disease, cerebrovascular disease, coronary artery disease, hypertension, major cancers, peripheral vascular disease, heart failure, and kidney stones. These comorbidity features are detected prior to index hospital admission or ED visit. We applied a 5-year look-back window to identify these features. The GP visit group contains twenty-three features that are identified based on the billing codes from the Ontario Health Insurance Plan database (Table 1).

Table 5-1: The features included in the GP visits Group

Features
Minor assessment
General assessment
General re-assessment
Consultation
Repeat consultation
Intermediate assessment or well-baby care
Mini assessment
Complex house call assessment
House call assessment
Limited consultation
Special family and general practice consultation

Comprehensive family and general practice consultation
Care of the elderly FPA
Periodic health visit - adult 65 years of age and older
Chronic disease shared appointment-2 patients (per unit)
Chronic disease shared appointment - 3 patients (per unit)
Chronic disease shared appointment - 4 patients (per unit)
Chronic disease shared appointment - 5 patients (per unit)
Chronic disease shared appointment - 6 to 12 patients (per unit)
Nursing home or home for the aged - first 2 subsequent visits per patient per month (per visit)
Nursing home or home for the aged - additional subsequent visits (maximum 2 per patient per month) per visit
Additional visits due to intercurrent illness (see General Preamble GP33) per visit

The hospital encounter code group includes 1878 diagnostic codes that were detected during the index hospital admission and ED visit. The medication group consists of 595 medications prescribed to the patients within 120 before the index date. We apply the Chi-Square test for feature selection on the hospital encounter code and medication groups and then filter the chosen features with a healthcare expert. We select seventy and fifty-five most significant features for hospital encounter code and medication groups, respectively based on the result of the chi square test. The ten most important features in the hospital encounter code and medication groups are shown in Table 2.

Table 5-2: The top ten features included in hospital encounter codes and medications groups

Hospital encounter codes	**Medications**
Acute myeloid leukemia	Sunitinib Malate

Diffuse non-Hodgkin's lymphoma	Lenalidomide
Chronic kidney disease	Abiraterone Acetate
Congestive heart failure	Metolazone
Cholecystitis	Cyclosporine
Lymphoid leukemia	Megestrol Acetate
Malignant neoplasm of bladder	Lithium Carbonate
Decubitus ulcer	Atropine Sulfate & Diphenoxylate Hcl
Abnormal serum enzyme levels	Furosemide
Secondary and unspecified malignant neoplasm of lymph nodes	Prochlorperazine Maleate

5.4.6 Implementation Details

VERONICA is implemented in HTML, JavaScript library D3, and R packages. R is used to develop the Analytics module. Html and D3 are used to build the interface and controls in the Interactive Visualization module. We implement the communication between these two modules using PHP and JavaScript.

We use R to develop different components of the Analytics module because it (1) offers support in performing various sampling and machine learning techniques, (2) is an open-source and platform-independent tool, (3) has several libraries, and (4) is available in the ICES environment.

D3 (Data Driven Documents) will be used to implement the interactive visualizations, and the Java programming language will be used to integrate data analytics with the visualizations. D3 (1) is an open-source Javascript library that works with web standards, (2) provides users with the full capabilities of the modern web browsers, (3) enables them to reuse JavaScript code and add different functionalities, and (4) is compatible with multiple platforms and other programming languages that are used in the implementation of VERONICA.

5.4.7 Workflow

As shown in Figure 1, VERONICA has two modules: Analytics and Interactive Visualization. The Analytics module utilizes the group structure of features stored in EHRs to identify the subset of feature groups that best represent the data in the prediction of AKI. The Interactive Visualization module maps the data items generated by the Analytics module into interactive visual representations to assist users in exploring the results. It supports six main interactions: 1) arranging, 2) drilling, 3) searching, 4) filtering, 5) transforming, and 7) selecting.

The basic workflow of VERONICA is as follows. First, we gather patient and hospital characteristics from five different databases stored at ICES. We then classify these features into five main groups—namely, hospital encounter codes, comorbidities, GP visits, medications, and demographics. The features included in these groups are pre-processed and transformed into forms appropriate for the analysis. We then create all possible subsets of groups (i.e., thirty-one groups), as shown in Figure 2. In the next step, we apply undersampling and SMOTE (Chawla et al., 2002) to each subset to obtain two sampled datasets. Next, five machine learning techniques—namely, CART, C5.0, random forest, naïve Bayes, and SVM are applied to each sampled dataset, generating 310 prediction models. We use the area under the receiver operating characteristic curve (AUROC) to report the performance of these models. To help users compare and explore the analytic results, we make them accessible to users through interactive visualizations. The Interactive Visualization module uses an interactive visual interface to show the results of the Analytics module. It allows users to explore the prediction models and compare their performance. The interface is supported by several controls, such as a search bar, selection buttons, and drop-down menus. Finally, several interactions are built into the system to allow users to manipulate the results.

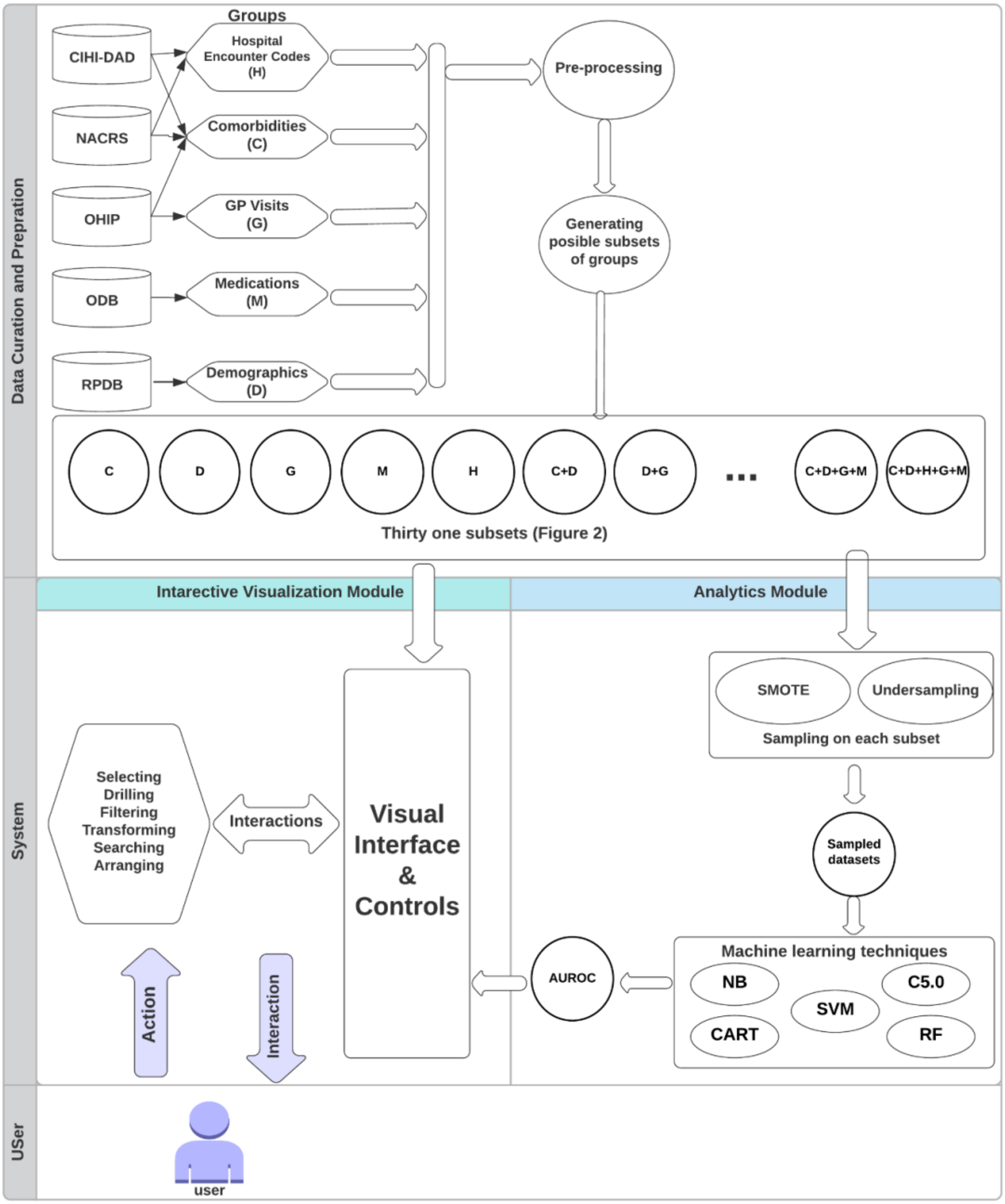


Figure 5-1: The workflow diagram of VERONICA.

5.5 The Design of VERONICA

We use VERONICA to identify the subset of groups that has the most substantial predictive power in the classification of AKI. VERONICA applies several machine learning techniques to each subset and allows exploration of the analysis results through

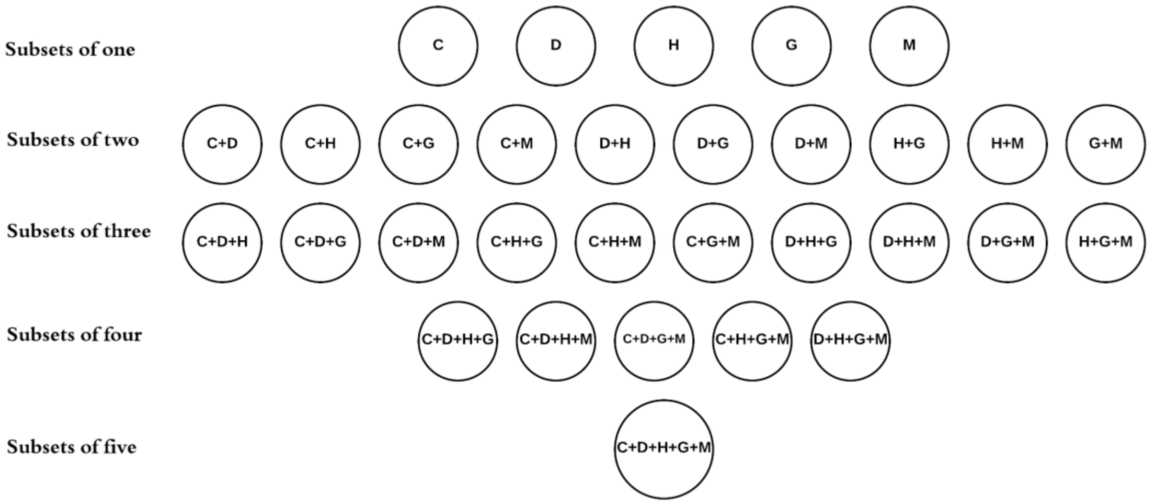


Figure 5-2: All possible subsets of 5 groups of features, including Comorbidities (C), Demographics (D), Hospital encounter codes (I), GP visits (G), and Medications (M).

interactive visualizations. In this section, we describe the two main components of the system. We explain how the data is processed and analyzed in the Analytics module. We then describe the Interactive Visualization module and how it assists users in the interpretation and exploration of the results.

5.5.1 Analytics Module

The Analytics module utilizes a representative set of machine learning and sampling techniques to identify the subset that best represents the data in identifying AKI. Three tree-based classifiers (CART, C5.0, and random forest), one kernel-based classifier (SVM), and one probabilistic classifier (naive Bayes) are used in this analysis. In this section, we explain how these techniques can be employed to analyze the data.

We classify features stored in our clinical dataset into five main groups based on the domain knowledge—namely, demographics, comorbidities, medications, hospital encounter codes, and GP visits. For each feature included in these groups, the last recorded value before the index date is chosen. The features in comorbidity, medication, hospital encounter code, and GP visit groups are set to either "Y" or "N". If an individual is prescribed a medication or has a comorbid condition, then its corresponding value is set to "Y". If there is evidence of a particular hospital encounter code to be present for a patient, we set its corresponding value to “Y”. We create multiple dummy variables for the age feature where each variable represents a specific age range. If a patient's age lays within a specified range, then the corresponding variable is set to "1". The region feature takes either "R" or "U" representing rural or urban, respectively. The sex feature takes either "M" or "F" for males and females. The income feature takes an integer value that lies within 1 to 5 to represent the income quintile. All features included in the cohort are transformed into a scale and format suitable for further analysis by machine learning techniques.

A total of 924,533 participants are included in the final cohort, of which 5,993 experienced AKI after being discharged from the index encounter. This dataset has an imbalanced class distribution, where the negative class (i.e., non-AKI) is represented by a large number of patients (i.e., 899,449 patients) compared to the positive class (i.e., 5993). In such an imbalanced dataset, classification techniques generally do not perform well since most of them assume that the ratios of each class are equal (Gu et al., 2009; Laza et al., 2011; Yang Liu et al., 2011; Zhang et al., 2010). Thus, the cost gets skewed in favor of the negative class in imbalanced datasets (Gu et al., 2008; Longadge & Dongre, 2013; Maloof, 2003). This makes the classifiers often biased toward the majority cla**ss.** One approach to improve the classification process is to use sampling strategies (Blagus & Lusa, 2013; Drummond & Holte, n.d.; Nguyen et al., 2012; Rahman & Davis, 2013). We apply two sampling approaches to solve this issue: (1) undersampling and (2) synthetic minority oversampling technique (SMOTE). Undersampling removes samples from the negative class randomly until the dataset is balanced. Although it reduces the strain on storage and improves run time, it might result in the loss of useful information. On the other hand, SMOTE is a sampling approach that involves oversampling minority class by creating samples synthetically using the k-nearest-neighbours algorithm. We configure undersampling and SMOTE so that the number of positive cases becomes equal compared to the negative cases. We use the DMwR package in R to implement the

SMOTE algorithm. The “k” (i.e., nearest neighbors) and “perc.over” variables of the SMOTE algorithm are set to 5 and 100, respectively.

To develop the prediction models, we first split the dataset into training and test sets. The training and test set includes 903,442 and 2000 cases, respectively. In the next step, we create every possible subset of groups, as shown in Figure 2. The total number of subsets is $2^5 - 1 = 31$ where 5 is the number of groups. We then apply both undersampling and SMOTE to each subset to obtain two sampled datasets. We develop ten prediction models for each subset by applying five machine learning techniques—namely, CART, C5.0, random forest, naive Bayes, and SVM to the sampled datasets. We created a total of $31 * 2 * 5 = 310$ models where 31, 2, and 5 are the number of subsets, sampling approaches, and machine learning techniques, respectively. In each model, AKI is the response variable and all features included in the subset are predictor variables. The CART and C5.0 classifiers are implemented using the "rpart" and "C50" packages in R, respectively. We use the "e1071" package in R to implement naive Bayes and SVM with a radial kernel (kernel=" radial"). Random forest is implemented using the "randomForest" package in R with fifty trees (i.e., ntree=50).

We compare the performance of all the generated models using AUROC (Ferri et al., 2009; Garcıa et al., 2012). A ROC curve shows the trade-off between sensitivity and specificity across different decision thresholds. Sensitivity measures how often a test classifies a patient as "at-risk" correctly. On the other hand, specificity is the capacity of a test to classify a patient as "risk-free" correctly (Parikh et al., 2008). The AUROC ranges from 0.51 to 0.89 for the classification of AKI among the generated models.

In total, VERONICA generates 310 models that are built by applying five machine learning techniques mentioned above on two sampled datasets (i.e., undersampled and SMOTE) for each subset. As a result, a large number of models and subsets are generated, which makes it difficult for users to understand the results. To overcome this issue, the data items generated by the Analytics module are made available to users through an interactive visual interface.

5.5.2 Interactive Visualization Module

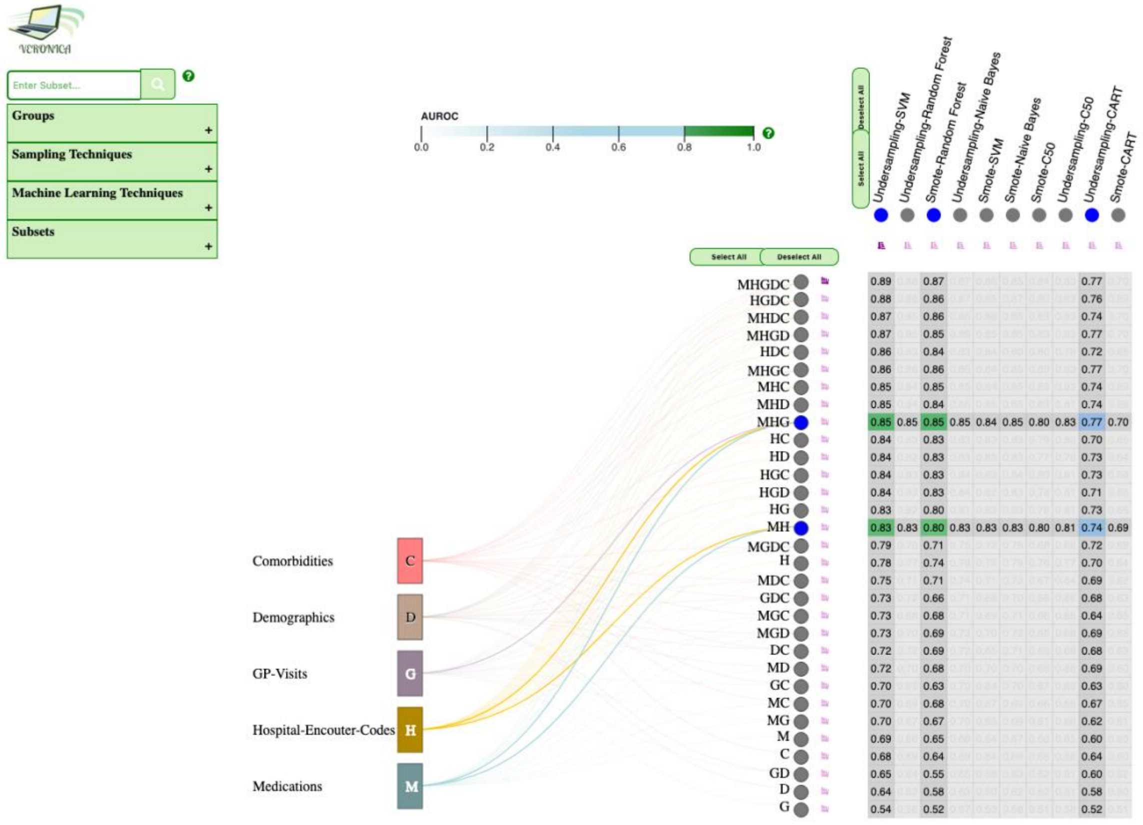
VERONICA is composed of an interactive visual interface and several selection controls, such as a search bar, drop-down menus, and selection buttons. In this section, we explain how data items produced by the Analytics module and subsets of groups are mapped into visual representation to allow users to accomplish various tasks.

As shown in Figure 3, groups of features (i.e., comorbidities, demographics, medications, hospital encounter codes, and GP visits) and their subsets are represented by a two-layer graph structure. In the first layer, the group nodes are mapped by color-coded rectangles, where each rectangle is labeled with a code representing the first letter of its corresponding group's name (Table 3). For instance, the rectangle representing the comorbidity group is color-coded in pink and is labeled with "C". The second layer includes all the nodes representing subsets of groups where each node includes a grey circle and a combination code in the text format. The combination code for each subset contains the first letters of all the groups that are included in the subset. For instance, as shown in Figure 3, the first grey circle from the top represents the subset of all groups, and it is labeled with "MHDGC". The connections between the nodes in the first and second layers are shown by color-coded links where the link's color is identical to its corresponding group node's color. Two nodes from the first and second layers are connected if the node in the first layer (i.e., group node) is included as one of the groups that make up the node in the second layer (i.e., subset node).

Table 5-3: Groups and their representing codes

Groups	**Codes**
Comorbidities	"C"
Demographics	"D"
GP visits	"G"
Hospital encounter codes	"H"
Medications	"M"

VERONICA uses a sortable heatmap to show the result of the Analytics module, as shown in Figure 3. It enables users to compare the performance of the generated models by placing the analysis techniques in the columns and subsets of groups in the rows. Each cell in the heatmap includes a color-coded numerical value representing the AUROC achieved by applying an analysis technique to a subset in the connecting column and row. The color of the cells of the heatmap is light grey by default. However, through different interactions, users can observe the cell's color based on the value of test AUROC corresponding to that cell. This color-coding is based on two gradient scales. The first gradient scale is created by blending different shades of green. It represents all the cells corresponding to models where AUROC is greater than 0.8. It is interesting to note that most of the models are densely clustered between 0.8 and 0.9. Thus, the second scale is built by blending different shades of blue to represent all the cells corresponding to models where AUROC is less than 0.8. We included a legend to assist users in interpreting the heatmap based on these gradient scales. There is also a help button ("?")

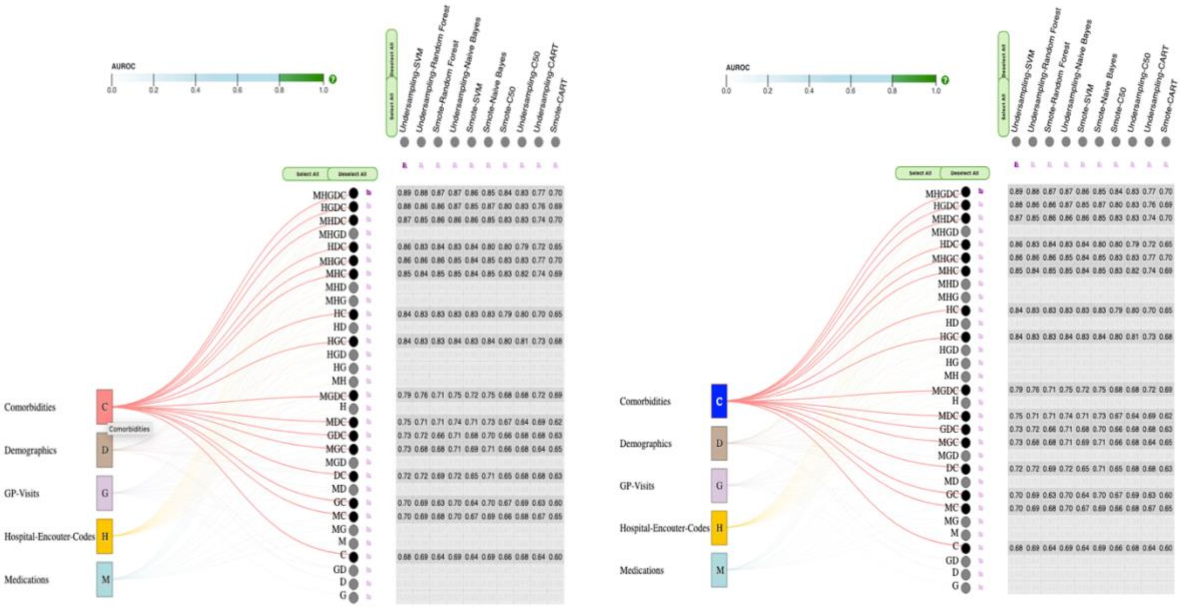


	Undersampling-SVM	Undersampling-Random Forest	Smote-Random Forest	Undersampling-Naive Bayes	Smote-SVM	Smote-Naive Bayes	Smote-C50	Undersampling-C50	Undersampling-CART	Smote-CART
MHGDC	0.89	[illegible]	0.87	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.77	[illegible]
HGDC	0.88	[illegible]	0.86	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.76	[illegible]
MHDC	0.87	[illegible]	0.86	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.74	[illegible]
MHGD	0.87	[illegible]	0.85	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.77	[illegible]
HDC	0.86	[illegible]	0.84	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.72	[illegible]
MHGC	0.86	[illegible]	0.86	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.77	[illegible]
MHC	0.85	[illegible]	0.85	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.74	[illegible]
MHD	0.85	[illegible]	0.84	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.74	[illegible]
MHG ●	0.85	0.85	0.85	0.85	0.84	0.85	0.80	0.83	0.77	0.70
HC	0.84	[illegible]	0.83	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.70	[illegible]
HD	0.84	[illegible]	0.83	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.73	[illegible]
HGC	0.84	[illegible]	0.83	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.73	[illegible]
HGD	0.84	[illegible]	0.83	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.71	[illegible]
HG	0.83	[illegible]	0.80	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.73	[illegible]
MH ●	0.83	0.83	0.80	0.83	0.83	0.83	0.80	0.81	0.74	0.69
MGDC	0.79	[illegible]	0.71	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.72	[illegible]
H	0.78	[illegible]	0.74	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.70	[illegible]
MDC	0.75	[illegible]	0.71	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.69	[illegible]
GDC	0.73	[illegible]	0.66	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.68	[illegible]
MGC	0.73	[illegible]	0.68	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.64	[illegible]
MGD	0.73	[illegible]	0.69	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.69	[illegible]
DC	0.72	[illegible]	0.69	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.68	[illegible]
MD	0.72	[illegible]	0.68	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.69	[illegible]
GC	0.70	[illegible]	0.63	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.63	[illegible]
MC	0.70	[illegible]	0.68	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.67	[illegible]
MG	0.70	[illegible]	0.67	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.62	[illegible]
M	0.69	[illegible]	0.65	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.60	[illegible]
C	0.68	[illegible]	0.64	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.64	[illegible]
GD	0.65	[illegible]	0.55	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.60	[illegible]
D	0.64	[illegible]	0.58	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.58	[illegible]
G	0.54	[illegible]	0.52	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.52	[illegible]

Figure 5-3: Overview of VERONICA

located to the right of the legend to provides users with additional information on how to interact with the heatmap.

Users can hover the mouse over any rectangle representing a group to highlight all the subset nodes that include the hovered group, links connecting the hovered group and highlighted subsets, and cells corresponding to the highlighted subsets (Figure 4-A). In addition, VERONICA allows users to select group nodes by clicking on their corresponding rectangles (Figure 4-B). The system then highlights all the subset nodes that contain all the groups corresponding to the selected rectangles, links connecting the selected groups and highlighted subsets, and rows of cells corresponding to the highlighted circles. In addition, to get additional information, users can move their mouse over the circles representing subsets to bring out tooltips. Furthermore, the system enables users to select any number of subsets by clicking on their corresponding circles.



(A) (B)

Figure 5-4: Shows how the system gets updated when users hover over (A) or select (B) a rectangle representing groups.

This interaction highlights all the cells corresponding to the selected subsets, group nodes that contain the selected subset, and links connecting the selected subset node and highlighted group nodes (Figure 5).

Users can observe the performance of different analysis techniques by clicking on circles representing the combinations. This interaction highlights all the cells in the heatmap representing the selected column. When a circle gets selected, its color changes to dark blue. As shown in Figure 6, when several subset nodes (or group nodes) and circles representing analysis techniques are selected simultaneously, the color of all the cells that both their rows and columns are selected changes based on the gradient scales mentioned above (i.e., shades of green or blue based on the value of the cell’s AUROC).

Users can also hover the mouse over the cells in the heatmap to highlight the labels and circles representing the hovered cell. Also, this interaction changes the cell's color based on its corresponding AUROC value. The system enables users to sort the cells by rows

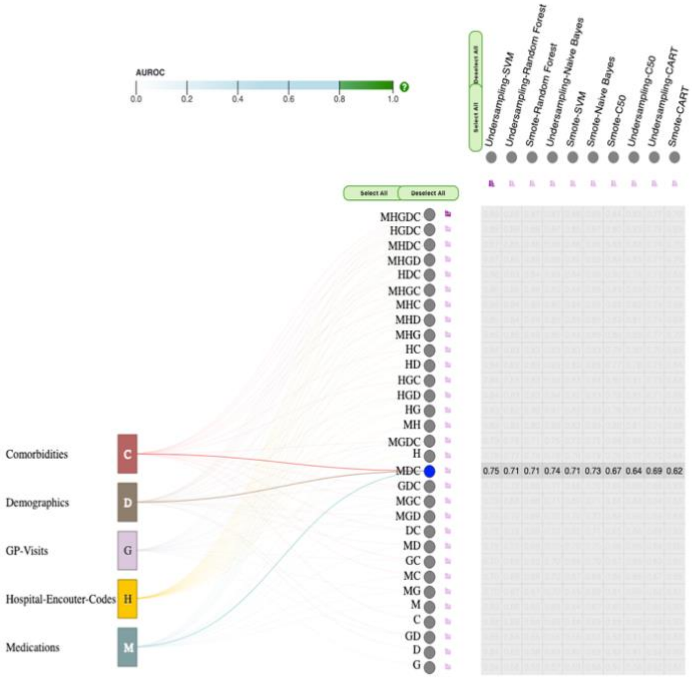
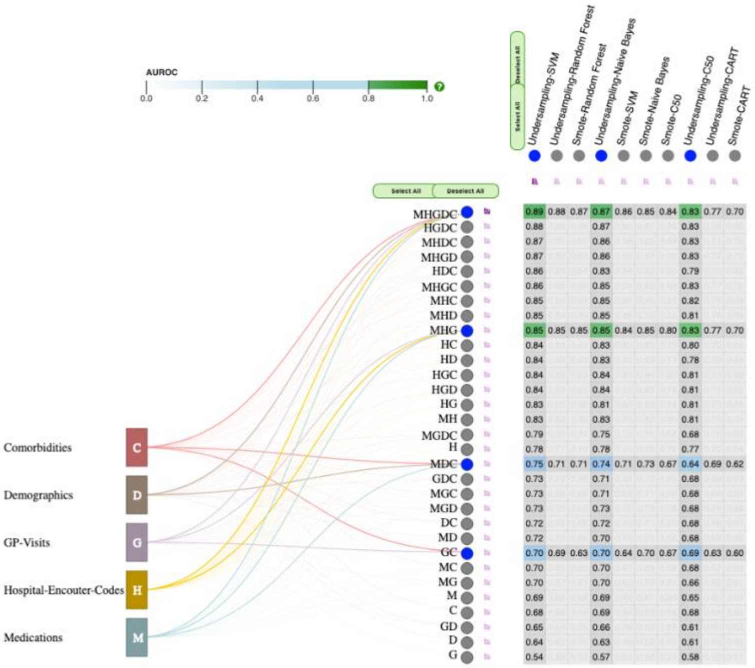


Figure 5-5: Shows how the system gets updated when users select a circle representing subsets.

and columns based on their corresponding AUROC values by clicking on the pink sort icons. For instance, cells in the heatmap are sorted by the "MHGDC" subset and "undersampling-SVM" technique in Figure 6.

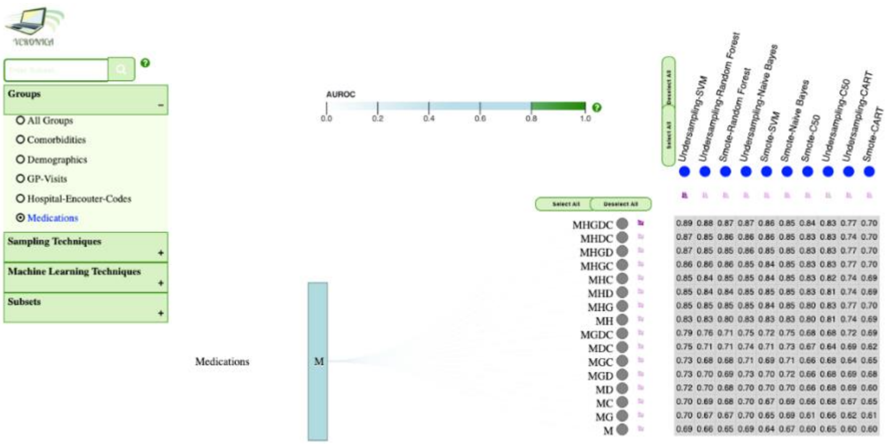
The horizontal and vertical groups of "Select All" and "Deselect All" buttons on the top left corner of the heatmap allow users to select/deselect all the subsets and techniques. These buttons help users easily get an overview of all the performances without selecting all the circles individually. VERONICA provides users with a search bar and four drop-down menus on the top left corner of the screen. Suppose users are interested in learning about a specific subset. In that case, they can enter the combination code corresponding to that subset in the search bar to change its color from black to green in the interface. In addition, when users hover their mouse over the help button placed besides the search bar, a tooltip appears with information on how to use the search bar.



	Undersampling-SVM	Undersampling-Random Forest	Smote-Random Forest	Undersampling-Naive Bayes	Smote-SVM	Smote-Naive Bayes	Smote-C50	Undersampling-C50	Undersampling-CART	Smote-CART
MHGDC	0.89	0.88	0.87	0.87	0.86	0.85	0.84	0.83	0.77	0.70
HGDC	0.88			0.87				0.83		
MHDC	0.87			0.86				0.83		
MHGD	0.87			0.86				0.83		
HDC	0.86			0.83				0.79		
MHGC	0.86			0.85				0.83		
MHC	0.85			0.85				0.82		
MHD	0.85			0.85				0.81		
MHG	0.85	0.85	0.85	0.85	0.84	0.85	0.80	0.83	0.77	0.70
HC	0.84			0.83				0.80		
HD	0.84			0.83				0.78		
HGC	0.84			0.84				0.81		
HGD	0.84			0.84				0.81		
HG	0.83			0.81				0.81		
MH	0.83			0.83				0.81		
MGDC	0.79			0.75				0.68		
H	0.78			0.78				0.77		
MDC	0.75	0.71	0.71	0.74	0.71	0.73	0.67	0.64	0.69	0.62
GDC	0.73			0.71				0.68		
MGC	0.73			0.71				0.68		
MGD	0.73			0.73				0.68		
DC	0.72			0.72				0.68		
MD	0.72			0.70				0.68		
GC	0.70	0.69	0.63	0.70	0.64	0.70	0.67	0.69	0.63	0.60
MC	0.70			0.70				0.68		
MG	0.70			0.70				0.66		
M	0.69			0.69				0.65		
C	0.68			0.69				0.68		
GD	0.65			0.66				0.61		
D	0.64			0.63				0.61		
G	0.54			0.57				0.58		

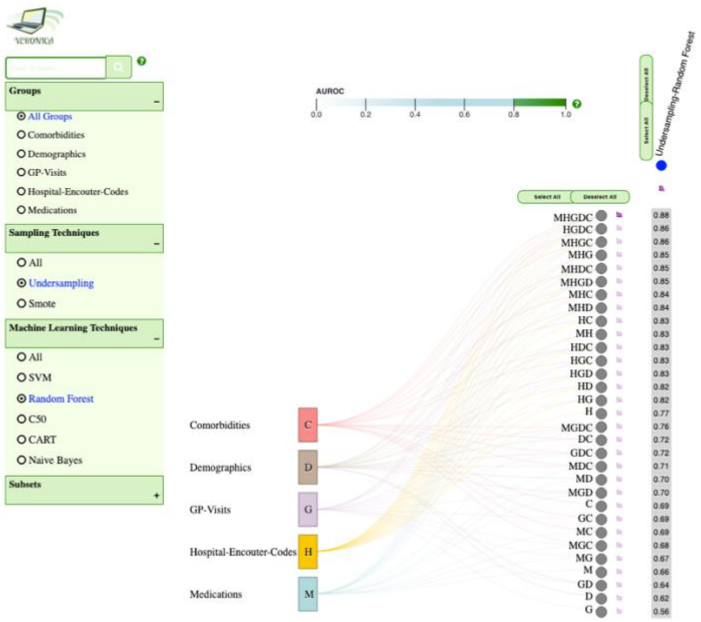
Figure 5-6: Shows how the system gets updated when users select multiple subset nodes and techniques.

The drop-down menus allow users to interactively filter subsets and techniques based on different criteria. This gives users great flexibility to focus on the data points of interest. The drop-down menus provide filtering based on groups, sampling techniques, machine learning techniques, and subsets from top to bottom, respectively. Each drop-down menu provides users with several options to choose from using radio buttons. The "Groups" menu allows users to focus on a specific group of features. If users select a group, the system only displays all subsets that contain the chosen group. For instance, Figure 7 shows how the system updates the interface if the "Medications" option is chosen from the menu. The "Sampling Techniques" and "Machine Learning Techniques" menus allow users to filter the columns of the heatmap based on sampling and machine learning techniques, respectively. For instance, if users are interested to learn how a specific combination of sampling and machine learning techniques such as SMOTE and random forest performs, they can select them in the second and third drop-down menus, respectively, as shown in Figure 8. The "Subsets" menu provides users with an option to compare all models that only include a specific number of groups. For instance, if users are interested in comparing the performance of all the techniques on subsets that only include two groups, they can choose "Subsets of Two" from the last menu (Figure 9). Users can filter data points based on different criteria by choosing an option from each menu (Figure 10). All these menus give users an option to reset the interface based on all groups, subsets, and techniques. Also, if users select any groups, subsets, or techniques, the system restores all the selections when it gets updated using any of the drop-down menus.



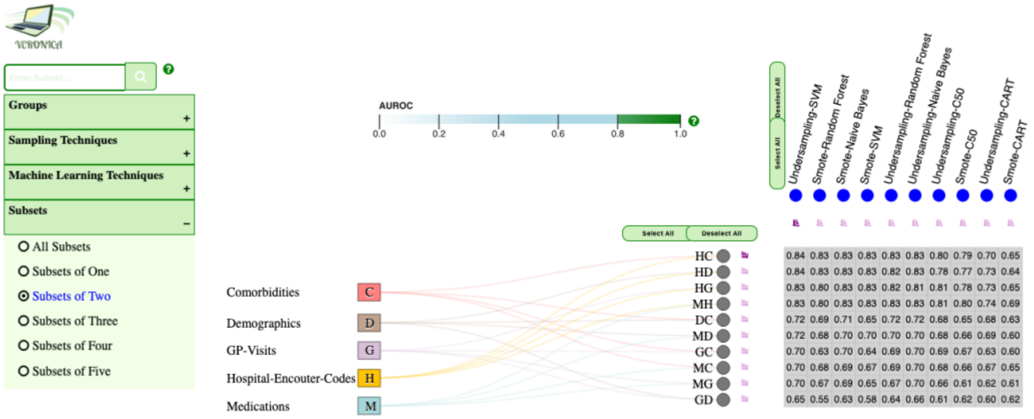
	Undersampling-SVM	Undersampling-Random Forest	Smote-Random Forest	Undersampling-Naive Bayes	Smote-SVM	Smote-Naive Bayes	Smote-C50	Undersampling-C50	Undersampling-CART	Smote-CART
MHGDC	0.89	0.88	0.87	0.87	0.86	0.85	0.84	0.83	0.77	0.70
MHDC	0.87	0.85	0.86	0.86	0.86	0.85	0.83	0.83	0.74	0.70
MHGD	0.87	0.85	0.85	0.86	0.85	0.85	0.83	0.83	0.77	0.70
MHGC	0.86	0.86	0.86	0.85	0.84	0.85	0.83	0.83	0.77	0.70
MHC	0.85	0.84	0.85	0.85	0.84	0.85	0.83	0.82	0.74	0.69
MHD	0.85	0.84	0.84	0.85	0.85	0.85	0.83	0.81	0.74	0.69
MHG	0.85	0.85	0.85	0.85	0.84	0.85	0.80	0.83	0.77	0.70
MH	0.83	0.83	0.80	0.83	0.83	0.83	0.80	0.81	0.74	0.69
MGDC	0.79	0.76	0.71	0.75	0.72	0.75	0.68	0.68	0.72	0.69
MDC	0.75	0.71	0.71	0.74	0.71	0.73	0.67	0.64	0.69	0.62
MGC	0.73	0.68	0.68	0.71	0.69	0.71	0.66	0.68	0.64	0.65
MGD	0.73	0.70	0.69	0.73	0.70	0.72	0.66	0.68	0.69	0.68
MD	0.72	0.70	0.68	0.70	0.70	0.70	0.66	0.68	0.69	0.60
MC	0.70	0.69	0.68	0.70	0.67	0.69	0.66	0.68	0.67	0.65
MG	0.70	0.67	0.67	0.70	0.65	0.69	0.61	0.66	0.62	0.61
M	0.69	0.66	0.65	0.69	0.64	0.67	0.60	0.65	0.60	0.60

Figure 5-7: Shows how the system gets updated when users select "Medications" from the "Groups" drop-down menu.



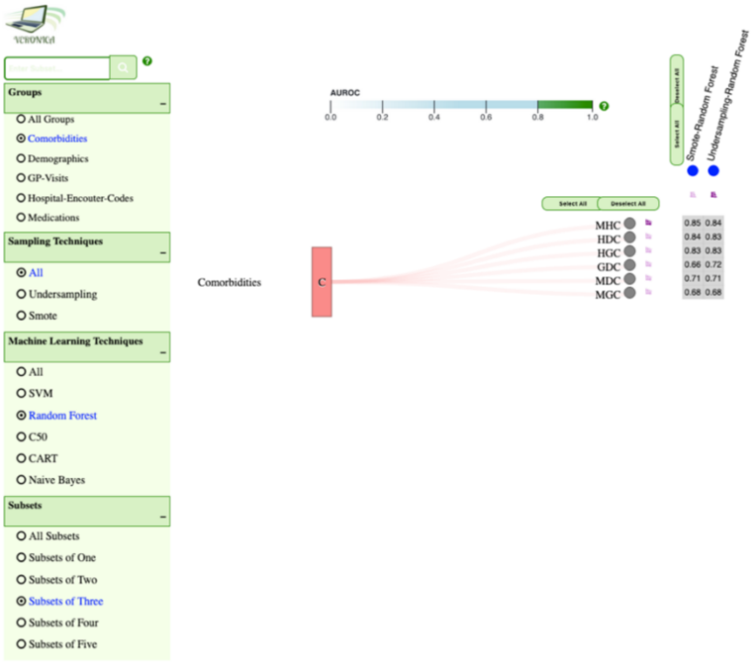
	Undersampling-Random Forest
MHGDC	0.88
HGDC	0.86
MHGC	0.86
MHG	0.85
MHDC	0.85
MHGD	0.85
MHC	0.84
MHD	0.84
HC	0.83
MH	0.83
HDC	0.83
HGC	0.83
HGD	0.83
HD	0.82
HG	0.82
H	0.77
MGDC	0.76
DC	0.72
GDC	0.72
MDC	0.71
MD	0.70
MGD	0.70
C	0.69
GC	0.69
MC	0.69
MGC	0.68
MG	0.67
M	0.66
GD	0.64
D	0.62
G	0.56

Figure 5-8: Shows how the system gets updated when users select "Undersampling" and "Random Forest" from the "Sampling Techniques" and "Machine Learning Techniques" drop-down menus.



	Undersampling-SVM	Smote-Random Forest	Smote-Naive Bayes	Smote-SVM	Undersampling-Random Forest	Undersampling-Naive Bayes	Undersampling-C50	Smote-C50	Undersampling-CART	Smote-CART
HC	0.84	0.83	0.83	0.83	0.83	0.83	0.80	0.79	0.70	0.65
HD	0.84	0.83	0.83	0.83	0.82	0.83	0.78	0.77	0.73	0.64
HG	0.83	0.80	0.83	0.83	0.82	0.81	0.81	0.78	0.73	0.65
MH	0.83	0.80	0.83	0.83	0.83	0.83	0.81	0.80	0.74	0.69
DC	0.72	0.69	0.71	0.65	0.72	0.72	0.68	0.65	0.68	0.63
MD	0.72	0.68	0.70	0.70	0.70	0.70	0.68	0.66	0.69	0.60
GC	0.70	0.63	0.70	0.64	0.69	0.70	0.69	0.67	0.63	0.60
MC	0.70	0.68	0.69	0.67	0.69	0.70	0.68	0.66	0.67	0.65
MG	0.70	0.67	0.69	0.65	0.67	0.70	0.66	0.61	0.62	0.61
GD	0.65	0.55	0.63	0.58	0.64	0.66	0.61	0.62	0.60	0.62

Figure 5-9: Shows how the system gets updated when users select "Subsets of Two" from the "Subsets" drop-down menu.



	Smote-Random Forest	Undersampling-Random Forest
MHC	0.85	0.84
HDC	0.84	0.83
HGC	0.83	0.83
GDC	0.66	0.72
MDC	0.71	0.71
MGC	0.68	0.68

Figure 5-10: Shows how the system gets updated when users select "Comorbidities", "Random Forest", and "Subsets of Three" from "Groups", "Machine Learning Techniques", and "Subsets" drop-down menus.

5.6 Limitations

This tool should be evaluated with respect to four limitations. The first limitation relates to the problem of using undersampling. The main issue with this sampling approach is that it results in the loss of potentially useful data that could be essential for the induction process. The second limitation is that the system only supports a limited number of data mining and sampling techniques. Third, the system is designed for imbalanced datasets. The sampling techniques are unnecessary if the dataset is balanced. Forth, most of the guidelines for AKI diagnosis rely on an increase in serum creatinine as a gold standard. However, these guidelines need a premorbid serum creatinine value to be used as a baseline creatinine, which was not available for all patients in this research. Therefore, the episode of AKI was identified using the ICD-10 code. The fifth limitation is that although the healthcare experts at ICES have found VERONICA helpful and usable through the participatory design process, we have not conducted a formal study to evaluate the system's performance or the efficiency of its user-information discourse mechanism.

5.7 Conclusion

In this paper, we demonstrate how VA systems can be designed to address the challenges stemming from the high dimensional EHRs to identify the subset of feature groups with the most predictive power in the classification of AKI systematically. To accomplish this, we have reported the development of VERONICA, a VA system designed to assist healthcare providers at ICES' KDT program. VERONICA incorporates two components: Analytics and Interactive Visualization modules. The Analytics module identifies the best representative subset of data in detecting the patients at high risk of developing AKI using different sampling and machine learning techniques. It incorporates two sampling techniques—namely, undersampling and SMOTE. It also uses a representative set of machine learning techniques, including CART, C5.0, random forest, SVM, and naive Bayes. Our clinical dataset includes comorbidities, demographics, hospital encounter codes, GP visits, and medications. The system generates a large number of prediction models by applying sampling and machine learning techniques mentioned above to each subset. The performance of all the generated models is reported using AUROC. The system enables users to access, explore, and compare these models through interactive visualizations. The Interactive Visualization module is composed of an interactive visual interface and several selection controls, such as a search bar, drop-down menus, and selection buttons. The interactive visual interface assists users in the exploration of the analytic results by providing them several interactions such as arranging, drilling, searching, filtering, transforming, and selecting.

In terms of VERONICA's scalability and extensibility, we design it in a modular way so that it can accept new data sources and sampling and machine learning techniques. VERONICA can be used to analyze high-dimensional datasets in many other domains, such as insurance, bioinformatics, and finance, where the features included in the dataset have a group structure.

Chapter 6

6 Conclusion

This dissertation has touched upon several aspects regarding the design and development of VA systems to accomplish complex EHR-driven tasks. Chapter 2 presents a framework for analyzing activities and tasks supported by EHR-based interactive visualization tools to provide a systematic approach to examine and evaluate these tools. To do so, we provide a brief survey of the existing EHR-based interactive visualization tools and examine patterns in these tools. In chapter 3, we conduct a comprehensive systematic review of existing EHR-based VA systems. We explain the review criteria, including visual analytics tasks and analytics, visualizations, and interactions supported by these VA systems. The EHR-based VA systems are evaluated using these criteria to identify the gaps and challenges of the use of VA in EHRs that remain insufficiently addressed. These challenges motivate us to design two novel EHR-based VA systems. Chapter 4 presented SUNRISE (viSUal aNalytics for exploring the association between laboRatory test results and a dIsease outcome using xgbooSt and Eclat), a VA system that allows interactive exploration of associations between laboratory test results and a disease outcome. Chapter 5 presents VERONICA (Visual analytics for idEntifying featuRe grOups iN dIsease ClAssification), another VA system that uses the classification of features in EHRs to identify the group of features with the strongest predictive power in detecting patients at high risk of developing a disease. Both of these VA systems are designed to assist the healthcare researchers at the ICES-KDT program. We demonstrate the utility of the proposed VA systems using large provincial healthcare databases from Ontario, Canada, stored at ICES to solve different AKI-related problems.

This chapter concludes the dissertation and is broken into three sections: 1) a brief overview of each chapter and some of its contributions, 2) general contributions of this research to the scientific literature, 3) area and future research and limitations.

6.1 Dissertation Summary

In Chapter 2, we conduct a systematic literature survey of 19 population-based and single-patient interactive visualization tools that involve EHRs as their primary source of data. We then present an activity and task analysis framework to examine EHR-driven tasks and activities supported by these tools. We provide an analysis of how sub-activities, tasks, and sub-tasks blend to accomplish the tool's main higher-level activities using the framework. The proposed framework can help designers to conceptualize the functionalities of new EHR-based tools. It can also be used to evaluate the existing tools and reveal the gaps in support of higher-level activities supported by these tools.

In Chapter 3, we provide an overview of the state-of-the-art in EHR-based VA systems. To do so, we conduct a systematic literature review to collect all the research papers that describe the design of VA systems that work with EHRs. We then identify the primary dimensions of the EHR-based VA design space, including VA tasks, analytics, visualizations, and interactions. This chapter gives an overview of the characteristics of the reviewed VA systems with respect to the selected dimensions. The findings of this chapter can provide value to designers and researchers as an organized catalog of various techniques that are most suitable for EHR-driven tasks. The evaluation of these systems is used to identify areas with little prior work that remains a challenge for future research.

In Chapter 4, we demonstrate how VA systems can be designed to facilitate the creation and interpretation of risk prediction models. We report the development of a novel proof of concept VA system called SUNRISE and show how it can be utilized to help healthcare providers to explore associations between laboratory test results and a disease outcome. SUNRISE integrates extreme gradient boosting (XGBoost) and a frequent itemset mining technique (i.e., Eclat) with multiple interactive sub-visualizations to develop and explore risk prediction models. It has shown to assist healthcare providers in 1) exploring relationships between various groups of laboratory test results and disease, 2) investigating the prediction results, 3) tracking the decision path by studying the working mechanisms of the models to verify their reliability, and 4) conducting what-if analysis by allowing them to probe the prediction model and observing its output. We show the utility of this system through a case study with AKI using healthcare data from Ontario, Canada, housed at ICES.

In Chapter 5, we explain how to design VA systems in a systematic manner to address the challenges stemming from high dimensional EHRs to improve predictive risk modeling. We introduce VERONICA, a novel proof of concept EHR-based VA system that utilizes the natural classification of features in EHRs to identify the group of features with the most substantial predictive power. It incorporates a representative set of supervised learning techniques, including classification and regression tree, C5.0, random forest, support vector machines, and naive Bayes to allow the analysis of the data from different perspectives. VERONICA helps healthcare providers in comparing and exploring the analytics results by making them accessible through an interactive visual interface. We demonstrate the usefulness of this VA system by identifying representative feature groups in detecting patients with a high risk of developing AKI using a clinical dataset stored at ICES.

6.2 General Contributions

As discussed in the introduction, the broad concern of this research lies in the design of VA systems that facilitate the performance of complex data-intensive tasks with EHRs. There is currently a lack of research in this area, and thus, we aim to bridge the gap through this work. This research describes the individual components of VA systems and presents the process of integrating these components systematically to design EHR-based VA systems that support the computational and cognitive demands of healthcare providers. Furthermore, this research explains how various EHR-driven tasks can be accomplished by combining machine learning techniques, data mining algorithms, visualization, and human-data interaction through VA. Using this research, designers can approach the development of EHR-based VA systems with a deeper understanding of how the process unfolds.

In Chapters 2 and 3, we have highlighted the need for coherent, unifying, and comprehensive frameworks that support the conceptualization of functionalities of EHR-based systems in an organized manner. In light of this, one of the main contributions of this research is the activity and tasks analysis framework that helps researchers to get a deeper understanding of interactive visualization tools by providing them with all the higher-level activities that can be achieved by performing different sets of tasks and sub-tasks. Furthermore, it can result in the development of best practices for designing similar frameworks in other fields. Through a systematic literature review of existing EHR-based VA systems, this research identifies and described the primary dimensions of EHR-based VA design space that unify the existing work in Chapter 3. The proposed design space can be used to evaluate the EHR-based VA systems from different perspectives. Moreover, the comprehensive review of the state-of-the-art in EHR-based VA systems can help researchers and evaluators to assess healthcare providers' demands and expectations and to better understand the EHR-driven tasks from their perspective. Also, the findings of this review can give value to designers as an organized archive of different analytics, visualization, and interaction approaches that are most appropriate for various EHR-driven VA tasks.

Another Contribution of this research is the development of EHR-based VA systems, known as SUNRISE and VERONICA, which are discussed in Chapters 4 and 5. To help us understand the real-world EHR-driven tasks performed by healthcare providers, we adopt a participatory design approach. We asked the stakeholders at the ICES-KDT program to provide us with formative feedback on design decisions in every step of the design and development process of the proposed VA systems. Through these evaluations**,** users at the ICES-KDT program find the systems helpful and sophisticated. These VA systems are scalable and can be reconfigured to work with other forms of data.

Finally, this research provides the healthcare domain with evidence of the efficacy of VA for handling EHRs. This research has suggestions for other areas that require their data to be made analyzable and accessible through VA. Designers in the following areas will likely find this research useful: data and information visualization, statistics, data science, and educational and learning technologies.

6.3 Limitations and Future Work

The systems presented in this dissertation lay a foundation for the usefulness of VA systems in healthcare. We identified a small number of interactive visualization tools that support higher-level activities, “predicting” and “monitoring” through the activity and task analysis framework. Thus, more research is required to explore how EHR-based VA systems can be extended to support these activities.

This research reports the development of two EHR-based VA systems—namely, SUNRISE and VERONICA. In Chapters 4 and 5, we describe the data sources, stakeholders, interface, input, output, and design criteria of these systems. Although, we describe these systems’ workflow to demonstrate their design, their architectures are not described thoroughly. We mainly focus on the conceptual challenges rather than practical difficulties while describing the proposed systems. For instance, we do not investigate the VA systems’ ability to interact with other systems or the number of users that can be supported simultaneously. Another limitation of the proposed systems lies in their capability to increase their capacity and functionalities based on users’ needs. The other limitation is that we cannot assess the scalability of these systems as they are both developed in an access-restricted virtual machine. Also, we are not required to develop web services or implement any adapters for the same reason.

We use various programming languages, platforms, and servers for implementing these systems. These systems are implemented using JavaScript library D3, HTML, Java programming language, Ajax, SAS, and R packages. The datasets are stored in the SAS server. R server is employed to perform all the underlying processing. Since different technologies are used to develop these systems, it is easy to incorporate new analytics techniques (e.g., using a new package in R), additional features in the user interface (e.g., changing D3 functions), and additional data (e.g., incorporating new datasets).

It should be noted that we have not conducted any formal studies to evaluate the performance of the proposed systems nor the effectiveness of their interaction mechanisms. However, during the design process, healthcare providers have evaluated these systems and found them useful. Thus, formal studies can help assess the efficiency of these systems for both expert and non-expert users. Both systems presented in this research are developed and tested using healthcare databases stored at ICES. Studies that ascertain the effectiveness of these systems with different datasets and settings will help us better understand the efficacy of these systems in the future.

References

Abdullah, S. (2020). Visual Analytics of Electronic Health Records with a focus on Acute Kidney Injury. Electronic Thesis and Dissertation Repository. https://ir.lib.uwo.ca/etd/7086

Abdullah, S. S., Rostamzadeh, N., Sedig, K., Garg, A. X., & McArthur, E. (2020a). Multiple Regression Analysis and Frequent Itemset Mining of Electronic Medical Records: A Visual Analytics Approach Using VISA_M3R3. Data, 5(2), 33. https://doi.org/10.3390/data5020033

Abdullah, S. S., Rostamzadeh, N., Sedig, K., Garg, A. X., & McArthur, E. (2020b). Visual Analytics for Dimension Reduction and Cluster Analysis of High Dimensional Electronic Health Records. Informatics, 7(2), 17. https://doi.org/10.3390/informatics7020017

Abdullah, S. S., Rostamzadeh, N., Sedig, K., Lizotte, D. J., Garg, A. X., & McArthur, E. (2020). Machine Learning for Identifying Medication-Associated Acute Kidney Injury. Informatics, 7(2), 18. https://doi.org/10.3390/informatics7020018

Agrawal, A. (2009). Medication errors: Prevention using information technology systems. British Journal of Clinical Pharmacology, 67(6), 681–686. https://doi.org/10.1111/j.1365-2125.2009.03427.x

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases.

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. Proceedings of the 20th International Conference on Very Large Data Bases, 487–499.

Agrawal, R., Swami, A., & Imielinski, T. (1993). Database Mining: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering, 5(6), 914–925. https://doi.org/10.1109/69.250074

Aigner, W., Kaiser, K., & Miksch, S. (2008). Visualization techniques to support authoring, execution, and maintenance of clinical guidelines. Computer-Based Medical Guidelines and Protocols: A Primer and Current Trends, 139, 140–159.

Alexander, N., Alexander, D. C., Barkhof, F., & Denaxas, S. (2020). Using Unsupervised Learning to Identify Clinical Subtypes of Alzheimer’s Disease in Electronic Health Records. Studies in Health Technology and Informatics, 270, 499–503. https://doi.org/10.3233/SHTI200210

Ali, T., Khan, I., Simpson, W., Prescott, G., Townend, J., Smith, W., & Macleod, A. (2007). Incidence and outcomes in acute kidney injury: A comprehensive population-based study. Journal of the American Society of Nephrology : JASN, 18(4), 1292–1298. https://doi.org/10.1681/ASN.2006070756

Amarasingham, R., Patzer, R. E., Huesch, M., Nguyen, N. Q., & Xie, B. (2014). Implementing electronic health care predictive analytics: Considerations and challenges. Health Affairs, 33(7), 1148–1154. https://doi.org/10.1377/hlthaff.2014.0352

Anderson, H. D., Pace, W. D., Brandt, E., Nielsen, R. D., Allen, R. R., Libby, A. M., West, D. R., & Valuck, R. J. (2015). Monitoring suicidal patients in primary care using electronic health records. The Journal of the American Board of Family Medicine, 28(1), 65–71. https://doi.org/10.3122/jabfm.2015.01.140181

Angulo, D. A., Schneider, C., Oliver, J. H., Charpak, N., & Hernandez, J. T. (2016). A Multi-facetted Visual Analytics Tool for Exploratory Analysis of Human Brain and Function Datasets. Frontiers in Neuroinformatics, 10. https://doi.org/10.3389/fninf.2016.00036

Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., Bernal-Delgado, E., Blomberg, N., Bock, C., Conesa, A., Del Signore, S., Delogne, C., Devilee, P., Di Meglio, A., Eijkemans, M., Flicek, P., Graf, N., Grimm, V., Guchelaar, H.-J., … Zanetti, G. (2016). Making sense of big data in health research: Towards an EU action plan. Genome Medicine, 8(1), 71. https://doi.org/10.1186/s13073-016-0323-y

Ayres, J., Flannick, J., Gehrke, J., & Yiu, T. (2002). Sequential pattern mining using a bitmap representation. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 429–435.

Bade, R., Schlechtweg, S., & Miksch, S. (2004). Connecting time-oriented data and information to a coherent interactive visualization. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 105–112.

Badrick, T. (2013). Evidence-Based Laboratory Medicine. The Clinical Biochemist Reviews, 34(2), 43.

Basole, R. C., Braunstein, M. L., Kumar, V., Park, H., Kahng, M., Chau, D. H., Tamersoy, A., Hirsh, D. A., Serban, N., Bost, J., Lesnick, B., Schissel, B. L., & Thompson, M. (2015). Understanding variations in pediatric asthma care processes in the emergency department using visual analytics. Journal of the American Medical Informatics Association, 22(2), 318–323. https://doi.org/10.1093/jamia/ocu016

Baytas, I. M., Lin, K., Wang, F., Jain, A. K., & Zhou, J. (2016). PhenoTree: Interactive Visual Analytics for Hierarchical Phenotyping From Large-Scale Electronic Health Records. IEEE Transactions on Multimedia, 18(11), 2257–2270. https://doi.org/10.1109/TMM.2016.2614225

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Benyon, D. (2013). Designing Interactive Systems: A comprehensive guide to HCI, UX and interaction design (3 edition). Pearson Canada.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. The Journal of Machine Learning Research, 13(1), 281–305.

Bhattacharjya, D., Shanmugam, K., Gao, T., Mattei, N., Varshney, K., & Subramanian, D. (2020). Event-driven continuous time bayesian networks. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), 3259–3266.

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics, 14(1), 106. https://doi.org/10.1186/1471-2105-14-106

Boonstra, A., Versluis, A., & Vos, J. F. (2014). Implementing electronic health records in hospitals: A systematic literature review. BMC Health Services Research, 14(1), 370.

Brauckhoff, D., Dimitropoulos, X., Wagner, A., & Salamatian, K. (2012). Anomaly Extraction in Backbone Networks Using Association Rules. IEEE/ACM Transactions on Networking, 20(6), 1788–1799. https://doi.org/10.1109/TNET.2012.2187306

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. https://zbmath.org/?q=an%3A0541.62042

Brennan, T. A., Leape, L. L., Laird, N. M., Hebert, L., Localio, A. R., Lawthers, A. G., Newhouse, J. P., Weiler, P. C., & Hiatt, H. H. (1991). Incidence of Adverse Events and Negligence in Hospitalized Patients. New England Journal of Medicine, 324(6), 370–376. https://doi.org/10.1056/NEJM199102073240604

Brodbeck, D., Gasser, R., Degen, M., Reichlin, S., & Luthiger, J. (2005). Enabling large-scale telemedical disease management through interactive visualization. European Notes in Medical Informatics, 1(1), 1172–1177.

Caban, J. J., & Gotz, D. (2015). Visual analytics in healthcare—Opportunities and research challenges. Journal of the American Medical Informatics Association, 22(2), 260–262. https://doi.org/10.1093/jamia/ocv006

Cabitza, F., & Banfi, G. (2018). Machine learning in laboratory medicine: Waiting for the flood? Clinical Chemistry and Laboratory Medicine (CCLM), 56(4), 516–524. https://doi.org/10.1515/cclm-2017-0287

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. Proceedings of the 25th International Conference on Machine Learning, 96–103. https://doi.org/10.1145/1390156.1390169

Chalmers, J., Pullan, M., Fabri, B., McShane, J., Shaw, M., Mediratta, N., & Poullis, M. (2013). Validation of EuroSCORE II in a modern cohort of patients undergoing cardiac surgery. European Journal of Cardio-Thoracic Surgery, 43(4), 688–694. https://doi.org/10.1093/ejcts/ezs406

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chittaro, L., Combi, C., & Trapasso, G. (2003). Data mining on temporal data: A visual approach and its clinical application to hemodialysis. Journal of Visual Languages & Computing, 14(6), 591–620.

Christensen, T., & Grimsmo, A. (2008). Instant availability of patient records, but diminished availability of patient information: A multi-method study of GP's use of electronic patient records. BMC Medical Informatics and Decision Making, 8(1), 12.

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Affairs, 33(7), 1139–1147. https://doi.org/10.1377/hlthaff.2014.0048

Collister, D., Pannu, N., Ye, F., James, M., Hemmelgarn, B., Chui, B., Manns, B., & Klarenbach, S. (2017). Health Care Costs Associated with AKI. Clinical Journal of the American Society of Nephrology : CJASN, 12(11), 1733–1743. https://doi.org/10.2215/CJN.00950117

Combi, C., Keravnou-Papailiou, E., & Shahar, Y. (2010). Temporal Information Systems in Medicine. Springer Science & Business Media.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1–17. https://doi.org/10.1016/j.ins.2012.10.039

Cowie, M. R., Blomster, J. I., Curtis, L. H., Duclaux, S., Ford, I., Fritz, F., Goldman, S., Janmohamed, S., Kreuzer, J., Leenay, M., Michel, A., Ong, S., Pell, J. P., Southworth, M. R., Stough, W. G., Thoenes, M., Zannad, F., & Zalewski, A. (2017). Electronic health records to facilitate clinical research. Clinical Research in Cardiology, 106(1), 1–9. https://doi.org/10.1007/s00392-016-1025-6

Cox, M. A. A., & Cox, T. F. (2008). Multidimensional Scaling. In C. Chen, W. Härdle, & A. Unwin (Eds.), Handbook of Data Visualization (pp. 315–347). Springer. https://doi.org/10.1007/978-3-540-33037-0_14

Cristianini, N., Shawe-Taylor, J., & Shawe-Taylor, D. of C. S. R. H. J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

Cui, W. (2019). Visual Analytics: A Comprehensive Overview. IEEE Access, 7, 81555–81573. https://doi.org/10.1109/ACCESS.2019.2923736

Dagliati, A., Sacchi, L., Zambelli, A., Tibollo, V., Pavesi, L., Holmes, J. H., & Bellazzi, R. (2017). Temporal electronic phenotyping by mining careflows of breast cancer patients. Journal of Biomedical Informatics, 66, 136–147. https://doi.org/10.1016/j.jbi.2016.12.012

Dagliati, Arianna, Sacchi, L., Tibollo, V., Cogni, G., Teliti, M., Martinez-Millana, A., Traver, V., Segagni, D., Posada, J., Ottaviano, M., Fico, G., Arredondo, M. T., De Cata, P., Chiovato, L., & Bellazzi, R. (2018). A dashboard-based system for supporting diabetes care. Journal of the American Medical Informatics Association, 25(5), 538–547. https://doi.org/10.1093/jamia/ocx159

Daniel, G. G. (2013). Artificial Neural Network. In A. L. C. Runehov & L. Oviedo (Eds.), Encyclopedia of Sciences and Religions (pp. 143–143). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8265-8_200980

Demirci, F., Akan, P., Kume, T., Sisman, A. R., Erbayraktar, Z., & Sevinc, S. (2016). Artificial neural network approach in laboratory test reporting: Learning algorithms. American Journal of Clinical Pathology, 146(2), 227–237.

Deng, H., Runger, G., & Tuv, E. (2011). Bias of Importance Measures for Multi-valued Attributes and Solutions. In T. Honkela, W. Duch, M. Girolami, & S. Kaski (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2011 (pp. 293–300). Springer. https://doi.org/10.1007/978-3-642-21738-8_38

Dey, S., Luo, H., Fokoue, A., Hu, J., & Zhang, P. (2018). Predicting adverse drug reactions through interpretable deep learning framework. BMC Bioinformatics, 19(S21), 476–476. https://doi.org/10.1186/s12859-018-2544-0

Didandeh, A., & Sedig, K. (2016). Externalization of Data Analytics Models: In S. Yamamoto (Ed.), Human Interface and the Management of Information: Information, Design and Interaction (pp. 103–114). Springer International Publishing. https://doi.org/10.1007/978-3-319-40349-6_11

Dingen, D., van't Veer, M., Houthuizen, P., Mestrom, E. H. J., Korsten, E. H. H. M., Bouwman, A. R. A., & van Wijk, J. (2019). RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis. IEEE Transactions on Visualization and Computer Graphics, 25(1), 246–255. https://doi.org/10.1109/TVCG.2018.2865043

Diri, B., & Albayrak, S. (2008). Visualization and analysis of classifiers performance in multi-class medical data. Expert Systems with Applications, 34(1), 628–634. https://doi.org/10.1016/j.eswa.2006.10.016

Doupi, P. (2012). Using EHR data for monitoring and promoting patient safety: Reviewing the evidence on trigger tools. Studies in Health Technology and Informatics, 180, 786–790.

Drummond, C., & Holte, R. C. (n.d.). C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. 8.

EHRIntelligence. (2018, June 5). *40% of Physicians See More EHR Challenges than Benefits*. EHRIntelligence. https://ehrintelligence.com/news/40-of-physicians-see-more-ehr-challenges-than-benefits

Endert, A., Hossain, M. S., Ramakrishnan, N., North, C., Fiaux, P., & Andrews, C. (2014). The human is the loop: New directions for visual analytics. Journal of Intelligent Information Systems, 43(3), 411–435. https://doi.org/10.1007/s10844-014-0304-9

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd, 96(34), 226–231.

Fails, J. A., Karlson, A., Shahamat, L., & Shneiderman, B. (2006). A visual interface for multivariate temporal data: Finding patterns of events across multiple histories. 2006 IEEE Symposium On Visual Analytics Science And Technology, 167–174.

Faiola, A., & Newlon, C. (2011). Advancing critical care in the ICU: A human-centered biomedical data visualization systems. International Conference on Ergonomics and Health Aspects of Work with Computers, 119–128.

Feng, C., Le, D., & Mccoy, A. B. (2019). Using Electronic Health Records to Identify Adverse Drug Events in Ambulatory Care: A Systematic Review Background and Significance. Appl Clin Inform, 10, 123–128. https://doi.org/10.1055/s-0039-1677738

Fernando, B., Fromont, E., & Tuytelaars, T. (2012). Effective Use of Frequent Itemset Mining for Image Classification. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), Computer Vision – ECCV 2012 (pp. 214–227). Springer. https://doi.org/10.1007/978-3-642-33718-5_16

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. Pattern Recognition Letters, 30(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

Fraley, C., & Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association, 97(458), 611–631. https://doi.org/10.1198/016214502760047131

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Galea, M. H., Blamey, R. W., Elston, C. E., & Ellis, I. O. (1992). The Nottingham prognostic index in primary breast cancer. Breast Cancer Research and Treatment, 22(3), 207–219. https://doi.org/10.1007/BF01840834

Garcıa, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the suitability of numerical performance measures for class imbalance problems. International Conference in Pattern Recognition Applications and Methods, 310–313.

Gaziano, T. A., Bitton, A., Anand, S., Abrahams-Gessel, S., & Murphy, A. (2010). Growing epidemic of coronary heart disease in low- and middle-income countries. Current Problems in Cardiology, 35(2), 72–115. https://doi.org/10.1016/j.cpcardiol.2009.10.002

Ghaddar, B., & Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. European Journal of Operational Research, 265(3), 993–1004. https://doi.org/10.1016/j.ejor.2017.08.040

Glasziou, P., Irwig, L., & Mant, D. (2005). Monitoring in chronic disease: A rational approach. BMJ, 330(7492), 644–648. https://doi.org/10.1136/bmj.330.7492.644

Glatz, E., Mavromatidis, S., Ager, B., & Dimitropoulos, X. (2014). Visualizing big network traffic data using frequent pattern mining and hypergraphs. Computing, 96(1), 27–38. https://doi.org/10.1007/s00607-013-0282-8

Goldsmith, M.-R., Transue, T. R., Chang, D. T., Tornero-Velez, R., Breen, M. S., & Dary, C. C. (2010). PAVA: Physiological and anatomical visual analytics for mapping of tissue-specific concentration and time-course data. Journal of Pharmacokinetics and Pharmacodynamics, 37(3), 277–287. https://doi.org/10.1007/s10928-010-9160-6

Goldstein, B. A., Navar, A. M., & Carter, R. E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. European Heart Journal, 38(23), 1805–1814.

Gotz, D. H., Sun, J., & Cao, N. (2012). Multifaceted visual analytics for healthcare applications. IBM Journal of Research and Development, 56(5), 6:1-6:12. https://doi.org/10.1147/jrd.2012.2199170

Gotz, D., & Stavropoulos, H. (2014). Decisionflow: Visual analytics for high-dimensional temporal event sequence data. IEEE Transactions on Visualization and Computer Graphics, 20(12), 1783–1792.

Gotz, D., Wang, F., & Perer, A. (2014). A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. Journal of Biomedical Informatics, 48, 148–159. https://doi.org/10.1016/j.jbi.2014.01.007

Gower, J. C., & Warrens, M. J. (2014). Similarity, dissimilarity, and distance, measures of. Wiley StatsRef: Statistics Reference Online, 1–11.

Gresh, D. L., Rabenhorst, D. A., Shabo, A., & Slavin, S. (2002). Prima: A case study of using information visualization techniques for patient record analysis. IEEE Visualization, 2002. VIS 2002., 509–512.

Groves, M., O’rourke, P., & Alexander, H. (2003). Clinical reasoning: The relative contribution of identification, interpretation and hypothesis errors to misdiagnosis. Medical Teacher, 25(6), 621–625. https://doi.org/10.1080/0142159031000160568

Gu, Q., Cai, Z., Zhu, L., & Huang, B. (2008). Data Mining on Imbalanced Data Sets. 2008 International Conference on Advanced Computer Theory and Engineering, 1020–1024. https://doi.org/10.1109/ICACTE.2008.26

Gu, Q., Zhu, L., & Cai, Z. (2009). Evaluation Measures of the Classification Performance of Imbalanced Data Sets. In Z. Cai, Z. Li, Z. Kang, & Y. Liu (Eds.), Computational Intelligence and Intelligent Systems (pp. 461–471). Springer. https://doi.org/10.1007/978-3-642-04962-0_53

Guerra Gómez, J., Wongsuphasawat, K., Wang, T. D., Pack, M., & Plaisant, C. (2011). Analyzing incident management event sequences with interactive visualization. Transportation Research Board 90th Annual Meeting Compendium of Papers.

Gunčar, G., Kukar, M., Notar, M., Brvar, M., Černelč, P., Notar, M., & Notar, M. (2018). An application of machine learning to haematological diagnosis. Scientific Reports, 8. https://doi.org/10.1038/s41598-017-18564-8

Guo, R., Fujiwara, T., Li, Y., Lima, K. M., Sen, S., Tran, N. K., & Ma, K.-L. (2020). Comparative Visual Analytics for Assessing Medical Records with Sequence Embedding. ArXiv:2002.08356 [Physics, Stat]. http://arxiv.org/abs/2002.08356

Ha, H., Lee, J., Han, H., Bae, S., Son, S., Hong, C., Shin, H., & Lee, K. (2019). Dementia Patient Segmentation Using EMR Data Visualization: A Design Study. International Journal of Environmental Research and Public Health, 16(18). https://doi.org/10.3390/ijerph16183438

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? Psychological Science, 16(1), 70–76. https://doi.org/10.1111/j.0956-7976.2005.00782.x

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 83–124.

Harerimana, G., Kim, J. W., Yoo, H., & Jang, B. (2019). Deep Learning for Electronic Health Records Analytics. IEEE Access, 7, 101245–101259. https://doi.org/10.1109/ACCESS.2019.2928363

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100–108. JSTOR. https://doi.org/10.2307/2346830

Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G. F., & Clermont, G. (2013). Outlier detection for patient monitoring and alerting. Journal of Biomedical Informatics, 46(1), 47–55. https://doi.org/10.1016/j.jbi.2012.08.004

Heisey-Grove, D., Danehy, L. N., Consolazio, M., Lynch, K., & Mostashari, F. (2014). A national study of challenges to electronic health record adoption and meaningful use. Medical Care, 52(2), 144–148. https://doi.org/10.1097/MLR.0000000000000038

Hersh, W. R. (2007). Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. The American Journal of Managed Care, 13(6 Part 1), 277–278.

Hettinger, A. Z., Roth, E. M., & Bisantz, A. M. (2017). Cognitive engineering and health informatics: Applications and intersections. Journal of Biomedical Informatics, 67, 21–33. https://doi.org/10.1016/j.jbi.2017.01.010

Himmelstein, D. U., Wright, A., & Woolhandler, S. (2010). Hospital computing and the costs and quality of care: A national study. The American Journal of Medicine, 123(1), 40–46. https://doi.org/10.1016/j.amjmed.2009.09.004

Hinum, K., Miksch, S., Aigner, W., Ohmann, S., Popow, C., Pohl, M., & Rester, M. (2005). Gravi++: Interactive Information Visualization of Highly Structured Temporal Data. J. UCS 11, 1792–1805. https://doi.org/10.3217/jucs-011-11-1792

Horn, W., Popow, C., & Unterasinger, L. (2001). Support for fast comprehension of ICU data: Visualization using metaphor graphics. Methods of Information in Medicine, 40(05), 421–424.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6), 417–441. https://doi.org/10.1037/h0071325

Huang, C.-W., Lu, R., Iqbal, U., Lin, S.-H., Nguyen, P. A. (Alex), Yang, H.-C., Wang, C.-F., Li, J., Ma, K.-L., Li, Y.-C. (Jack), & Jian, W.-S. (2015). A richly interactive exploratory data analysis and visualization tool using electronic medical records. BMC Medical Informatics and Decision Making, 15(1), 92. https://doi.org/10.1186/s12911-015-0218-7

Hund, M., Böhm, D., Sturm, W., Sedlmair, M., Schreck, T., Ullrich, T., Keim, D. A., Majnaric, L., & Holzinger, A. (2016). Visual analytics for concept exploration in subspaces of patient groups. Brain Informatics, 3(4), 233–247. https://doi.org/10.1007/s40708-016-0043-5

Ilayaraja M., & Meyyappan T. (2015). Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets. Procedia Computer Science, 70, 586–592. https://doi.org/10.1016/j.procs.2015.10.040

Ismail, B., & Anil, M. (2014). Regression methods for analyzing the risk factors for a life style disease among the young population of India. Indian Heart Journal, 66(6), 587–592. https://doi.org/10.1016/j.ihj.2014.05.027

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. Nature Reviews. Genetics, 13(6), 395–405. https://doi.org/10.1038/nrg3208

Jeong, D. H., Ji, S. Y., Suma, E. A., Yu, B., & Chang, R. (2015). Designing a collaborative visual analytics system to support users' continuous analytical processes. Human-Centric Computing and Information Sciences, 5(1). https://doi.org/10.1186/s13673-015-0023-4

Jin, Z., Cui, S., Guo, S., Gotz, D., Sun, J., & Cao, N. (2020). CarePre: An Intelligent Clinical Decision Assistance System. ACM Transactions on Computing for Healthcare, 1(1), 6:1–6:20. https://doi.org/10.1145/3344258

Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1906), 4237–4253. https://doi.org/10.1098/rsta.2009.0159

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science (New York, N.Y.), 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kamal, N. (2014). Big Data and Visual Analytics in Health and Medicine: From Pipe Dream to Reality. Journal of Health & Medical Informatics, 05(05). https://doi.org/10.4172/2157-7420.1000e125

Kankanhalli, A., Hahn, J., Tan, S., & Gao, G. (2016). Big data and analytics in healthcare: Introduction to the special section. Information Systems Frontiers, 18(2), 233–235. https://doi.org/10.1007/s10796-016-9641-2

Kehrer, J., & Hauser, H. (2013). Visualization and visual analysis of multifaceted scientific data: A survey. IEEE Transactions on Visualization and Computer Graphics, 19(3), 495–513. https://doi.org/10.1109/TVCG.2012.110

Keim, D. A., Mansmann, F., & Thomas, J. (2010). Visual analytics: How much visualization and how much analytics? ACM SIGKDD Explorations Newsletter, 11(2), 5. https://doi.org/10.1145/1809400.1809403

Keim, D. A., Munzner, T., Rossi, F., & Verleysen, M. (2015). Bridging Information Visualization with Machine Learning (Dagstuhl Seminar 15101). Dagstuhl Reports, 5(3), 1–27. https://doi.org/10.4230/DagRep.5.3.1

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), Information Visualization: Human-Centered Issues and Perspectives (pp. 154–175). Springer. https://doi.org/10.1007/978-3-540-70956-5_7

Keim, D., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). Mastering the Information Age Solving Problems with Visual Analytics. Eurographics Association. http://diglib.eg.org/handle/10.2312/14803

Keim, D., Mansmann, F., Schneidewind, J., Thomas, J., & Ziegler, H. (2008). Visual analytics: Scope and challenges. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4404 LNCS, 76–90. https://doi.org/10.1007/978-3-540-71080-6_6

Khalid, S., Judge, A., & Pinedo-Villanueva, R. (2018). An Unsupervised Learning Model for Pattern Recognition in Routinely Collected Healthcare Data: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies, 266–273. https://doi.org/10.5220/0006535602660273

Kho, A., Rotz, D., Alrahi, K., Cárdenas, W., Ramsey, K., Liebovitz, D., Noskin, G., & Watts, C. (2007). Utility of commonly captured data from an EHR to identify hospitalized patients at risk for clinical deterioration. AMIA Annual Symposium Proceedings, 2007, 404–408.

Klimov, D., Shahar, Y., & Taieb-Maimon, M. (2010a). Intelligent selection and retrieval of multiple time-oriented records. Journal of Intelligent Information Systems, 35(2), 261–300.

Klimov, D., Shahar, Y., & Taieb-Maimon, M. (2010b). Intelligent visualization and exploration of time-oriented data of multiple patients. Artificial Intelligence in Medicine, 49(1), 11–31. https://doi.org/10.1016/j.artmed.2010.02.001

Klimov, D., Shknevsky, A., & Shahar, Y. (2015). Exploration of patterns predicting renal damage in patients with diabetes type II using a visual temporal analysis laboratory. Journal of the American Medical Informatics Association: JAMIA, 22(2), 275–289. https://doi.org/10.1136/amiajnl-2014-002927

Knaus, W. A., Wagner, D. P., Draper, E. A., Zimmerman, J. E., Bergner, M., Bastos, P. G., Sirio, C. A., Murphy, D. J., Lotring, T., Damiano, A., & Harrell, F. E. (1991). The APACHE III Prognostic System: Risk Prediction of Hospital Mortality for Critically III Hospitalized Adults. Chest, 100(6), 1619–1636. https://doi.org/10.1378/chest.100.6.1619

Komaroff, A. L. (1979). The variability and inaccuracy of medical data. Proceedings of the IEEE, 67(9), 1196–1207. https://doi.org/10.1109/PROC.1979.11435

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal, 37(2), 233–243.

Krause, J., Perer, A., & Bertini, E. (2016). Using Visual Analytics to Interpret Predictive Machine Learning Models. ArXiv:1606.05685 [Cs, Stat]. http://arxiv.org/abs/1606.05685

Krause, J., Perer, A., & Ng, K. (2016). Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 5686–5697. https://doi.org/10.1145/2858036.2858529

Kumar, M., Stoll, N., Kaber, D., Thurow, K., & Stoll, R. (2007). Fuzzy filtering for an intelligent interpretation of medical data. 2007 IEEE International Conference on Automation Science and Engineering, 225–230. https://doi.org/10.1109/COASE.2007.4341714

Kumar, Y., & Sahoo, G. (2013). Prediction of different types of liver diseases using rule based classification model. Technology and Health Care, 21(5), 417–432.

Kwon, B. C., Anand, V., Severson, K. A., Ghosh, S., Sun, Z., Frohnert, B. I., Lundgren, M., & Ng, K. (2020). DPVis: Visual Analytics with Hidden Markov Models for Disease Progression Pathways. IEEE Transactions on Visualization and Computer Graphics.

Kwon, B. C., Choi, M.-J., Kim, J. T., Choi, E., Kim, Y. B., Kwon, S., Sun, J., & Choo, J. (2018). Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. IEEE Transactions on Visualization and Computer Graphics, 25(1), 299–309.

Kwon, B. C., Verma, J., & Perer, A. (2016). Peekquence: Visual analytics for event sequence data. ACM SIGKDD 2016 Workshop on Interactive Data Exploration and Analytics, 1.

Låg, T., Bauger, L., Lindberg, M., & Friborg, O. (2014). The role of numeracy and intelligence in health-risk estimation and medical data interpretation. Journal of Behavioral Decision Making, 27(2), 95–108. https://doi.org/10.1002/bdm.1788

Langavant, L. C. de, Bayen, E., & Yaffe, K. (2018). Unsupervised Machine Learning to Identify High Likelihood of Dementia in Population-Based Surveys: Development and Validation Study. Journal of Medical Internet Research, 20(7), e10493. https://doi.org/10.2196/10493

Lasko, T. A., Denny, J. C., & Levy, M. A. (2013). Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. PLOS ONE, 8(6), e66341. https://doi.org/10.1371/journal.pone.0066341

Lau, F., Price, M., Boyd, J., Partridge, C., Bell, H., & Raworth, R. (2012). Impact of electronic medical record on physician practice in office settings: A systematic review. BMC Medical Informatics and Decision Making, 12(1), 10–10. https://doi.org/10.1186/1472-6947-12-10

Laza, R., Pavón, R., Reboiro-Jato, M., & Fdez-Riverola, F. (2011). Evaluating the effect of unbalanced data in biomedical document classification. Journal of Integrative Bioinformatics, 8(3), 105–117.

Leape, L. L., Brennan, T. A., Laird, N., Lawthers, A. G., Localio, A. R., Barnes, B. A., Hebert, L., Newhouse, J. P., Weiler, P. C., & Hiatt, H. (1991). The Nature of Adverse Events in Hospitalized Patients. New England Journal of Medicine, 324(6), 377–384. https://doi.org/10.1056/NEJM199102073240605

Ledieu, T., Bouzille, G., Plaisant, C., Thiessard, F., Polard, E., & Cuggia, M. (2018). Mining clinical big data for drug safety: Detecting inadequate treatment with a DNA sequence alignment algorithm. AMIA Annual Symposium Proceedings, 2018, 1368–1376.

Lee, C., Luo, Z., Ngiam, K. Y., Zhang, M., Zheng, K., Chen, G., Ooi, B. C., & Yip, W. L. J. (2017). Big Healthcare Data Analytics: Challenges and Applications. In S. U. Khan, A. Y. Zomaya, & A. Abbas (Eds.), Handbook of Large-Scale Distributed Computing in Smart Healthcare (pp. 11–41). Springer International Publishing. https://doi.org/10.1007/978-3-319-58280-1_2

Leighton, J. P. (2004). Defining and Describing Reason. In The nature of reasoning. (pp. 3–11). Cambridge University Press.

Lesselroth, B. J., & Pieczkiewicz, D. S. (2011). Data visualization strategies for the electronic health record. Nova Science Publishers, Inc. https://experts.umn.edu/en/publications/data-visualization-strategies-for-the-electronic-health-record

Levy, A. R., O’Brien, B. J., Sellors, C., Grootendorst, P., & Willison, D. (2003). Coding accuracy of administrative drug claims in the Ontario Drug Benefit database. The Canadian Journal of Clinical Pharmacology = Journal Canadien De Pharmacologie Clinique, 10(2), 67–71.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. European Conference on Machine Learning, 4–15.

Li, X., & Wang, Y. (2016). Adaptive online monitoring for ICU patients by combining just-in-time learning and principal component analysis. Journal of Clinical

Monitoring and Computing, 30(6), 807–820. https://doi.org/10.1007/s10877-015-9778-4

Liangos, O., Wald, R., O'Bell, J. W., Price, L., Pereira, B. J., & Jaber, B. L. (2006a). Epidemiology and outcomes of acute renal failure in hospitalized patients: A national survey. Clinical Journal of the American Society of Nephrology : CJASN, 1(1), 43–51. https://doi.org/10.2215/CJN.00220605

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. 2, 6.

Lin, C., Karlson, E. W., Canhao, H., Miller, T. A., Dligach, D., Chen, P. J., Perez, R. N. G., Shen, Y., Weinblatt, M. E., & Shadick, N. A. (2013). Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. PloS One, 8(8), e69932.

Liu, K. E., Lo, C.-L., & Hu, Y.-H. (2014). Improvement of adequate use of warfarin for the elderly using decision tree-based approaches. Methods of Information in Medicine, 53(01), 47–53.

Liu, S., Wang, X., Liu, M., & Zhu, J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics, 1(1), 48–56. https://doi.org/10.1016/j.visinf.2017.01.006

Liu, Yang, Yu, X., Huang, J. X., & An, A. (2011). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Information Processing & Management, 47(4), 617–631. https://doi.org/10.1016/j.ipm.2010.11.007

Liu, Yunhao, Zhao, Y., Chen, L., Pei, J., & Han, J. (2012). Mining Frequent Trajectory Patterns for Activity Monitoring Using Radio Frequency Tag Arrays. IEEE Transactions on Parallel and Distributed Systems, 23(11), 2138–2149. https://doi.org/10.1109/TPDS.2011.307

Liu, Z., Nersessian, N., & Stasko, J. (2008). Distributed Cognition as a Theoretical Framework for Information Visualization. IEEE Transactions on Visualization and Computer Graphics, 14(6), 1173–1180. https://doi.org/10.1109/TVCG.2008.121

Lo, Y.-S., Lee, W.-S., & Liu, C.-T. (2013). Utilization of Electronic Medical Records to Build a Detection Model for Surveillance of Healthcare-Associated Urinary Tract Infections. Journal of Medical Systems, 37(2), 9923. https://doi.org/10.1007/s10916-012-9923-2

Longadge, R., & Dongre, S. (2013). Class Imbalance Problem in Data Mining Review. ArXiv:1305.1707 [Cs]. http://arxiv.org/abs/1305.1707

Louis, D. N., Gerber, G. K., Baron, J. M., Bry, L., Dighe, A. S., Getz, G., Higgins, J. M., Kuo, F. C., Lane, W. J., Michaelson, J. S., Le, L. P., Mermel, C. H., Gilbertson, J. R., & Golden, J. A. (2014). Computational pathology: An emerging definition. Archives of Pathology & Laboratory Medicine, 138(9), 1133–1138. https://doi.org/10.5858/arpa.2014-0034-ED

Lu, W., & Ng, R. (2020). Automated Analysis of Public Health Laboratory Test Results. AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science, 2020, 393–402.

Lütz, E. (2019). Unsupervised machine learning to detect patient subgroups in electronic health records. /paper/Unsupervised-machine-learning-to-detect-patient-in-L%C3%9CTZ/e11f5b060947f22ae7d80d053564546487dbc0bf

Malik, S., Du, F., Monroe, M., Onukwugha, E., Plaisant, C., & Shneiderman, B. (2014). An evaluation of visual analytics approaches to comparing cohorts of event sequences. EHRVis Workshop on Visualizing Electronic Health Record Data at VIS, 14.

Malik, S., Du, F., Monroe, M., Onukwugha, E., Plaisant, C., & Shneiderman, B. (2015). Cohort comparison of event sequences with balanced integration of visual analytics and statistics. Proceedings of the 20th International Conference on Intelligent User Interfaces, 38–49. https://doi.org/10.1145/2678025.2701407

Maloof, M. A. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. ICML-2003 Workshop on Learning from Imbalanced Data Sets II, 2, 2–1.

Mane, K. K., Bizon, C., Schmitt, C., Owen, P., Burchett, B., Pietrobon, R., & Gersing, K. (2012). VisualDecisionLinc: A visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry. Journal of Biomedical Informatics, 45(1), 101–106. https://doi.org/10.1016/j.jbi.2011.09.003

Marlin, B. M., Kale, D. C., Khemani, R. G., & Wetzel, R. C. (2012). Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics - IHI ’12, 389. https://doi.org/10.1145/2110363.2110408

Mathias, J. S., Gossett, D., & Baker, D. W. (2012). Use of electronic health record data to evaluate overuse of cervical cancer screening. Journal of the American Medical Informatics Association, 19(e1), e96–e101. https://doi.org/10.1136/amiajnl-2011-000536

Medicine, I. of, & America, C. on Q. of H. C. in. (2000). To Err Is Human: Building a Safer Health System. National Academies Press.

Mica, L., Niggli, C., Bak, P., Yaeli, A., McClain, M., Lawrie, C. M., & Pape, H.-C. (2020). Development of a Visual Analytics Tool for Polytrauma Patients: Proof of Concept for a New Assessment Tool Using a Multiple Layer Sankey Diagram in a Single-Center Database. World Journal of Surgery, 44(3), 764–772. https://doi.org/10.1007/s00268-019-05267-6

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports, 6(1), 26094. https://doi.org/10.1038/srep26094

Mittelstädt, S., Hao, M. C., Dayal, U., Hsu, M. C., Terdiman, J., & Keim, D. A. (2014). Advanced visual analytics interfaces for adverse drug event detection.

Proceedings of the Workshop on Advanced Visual Interfaces AVI, 237–244. https://doi.org/10.1145/2598153.2598156

Möller, A., Ruhlmann-Kleider, V., Leloup, C., Neveu, J., Palanque-Delabrouille, N., Rich, J., Carlberg, R., Lidman, C., & Pritchet, C. (2016). Photometric classification of type Ia supernovae in the SuperNova Legacy Survey with supervised learning. Journal of Cosmology and Astroparticle Physics, 2016(12), 008–008. https://doi.org/10.1088/1475-7516/2016/12/008

Monroe, M., Lan, R., Lee, H., Plaisant, C., & Shneiderman, B. (2013). Temporal event sequence simplification. IEEE Transactions on Visualization and Computer Graphics, 19(12), 2227–2236.

Moskovitch, R., & Shahar, Y. (2015a). Classification of multivariate time series via temporal abstraction and time intervals mining. Knowledge and Information Systems, 45(1), 35–74.

Moskovitch, R., & Shahar, Y. (2015b). Fast time intervals mining using the transitivity of temporal relations. Knowledge and Information Systems, 42(1), 21–48.

Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. Proceedings of the 21st International Conference on World Wide Web, 191–200. https://doi.org/10.1145/2187836.2187863

Müller, E., Günnemann, S., Assent, I., & Seidl, T. (2009). Evaluating clustering in subspace projections of high dimensional data. Proceedings of the VLDB Endowment, 2(1), 1270–1281. https://doi.org/10.14778/1687627.1687770

Munzner, T. (2014). Visualization Analysis and Design. CRC Press.

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. JAMA - Journal of the American Medical Association, 309(13), 1351–1352. https://doi.org/10.1001/jama.2013.393

Murphy, G., Hanken, M. A., & Waters, K. (1999). Electronic health records: Changing the vision.

Nashef, S. a. M., Roques, F., Michel, P., Gauducheau, E., Lemeshow, S., Salamon, R., & Group, the E. study. (1999). European system for cardiac operative risk evaluation (EuroSCORE). European Journal of Cardio-Thoracic Surgery, 16(1), 9–13. https://doi.org/10.1016/S1010-7940(99)00134-7

Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T. N., Vanden Berghe, W., Goethals, B., & Laukens, K. (2015). A primer to frequent itemset mining for bioinformatics. Briefings in Bioinformatics, 16(2), 216–231. https://doi.org/10.1093/bib/bbt074

Nelson, D. W., Rudehill, A., MacCallum, R. M., Holst, A., Wanecek, M., Weitzberg, E., & Bellander, B.-M. (2012). Multivariate outcome prediction in traumatic brain injury with focus on laboratory values. Journal of Neurotrauma, 29(17), 2613–2624.

Nguyen, H. M., Cooper, E. W., & Kamei, K. (2012). A comparative study on sampling techniques for handling class imbalance in streaming data. The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th

International Symposium on Advanced Intelligence Systems, 1762–1767. https://doi.org/10.1109/SCIS-ISIS.2012.6505291

Nielsen, F. (2016). Hierarchical Clustering. In F. Nielsen (Ed.), Introduction to HPC with MPI for Data Science (pp. 195–211). Springer International Publishing. https://doi.org/10.1007/978-3-319-21903-5_8

Ninkov, A., & Sedig, K. (2019). VINCENT: A visual analytics system for investigating the online vaccine debate. Online Journal of Public Health Informatics, 11(2).

Oh, W., Kim, E., Castro, M. R., Caraballo, P. J., Kumar, V., Steinbach, M. S., & Simon, G. J. (2016). Type 2 diabetes mellitus trajectories and associated risks. Big Data, 4(1), 25–30.

Ola, O., & Sedig, K. (2014). The challenge of big data in public health: An opportunity for visual analytics. Online Journal of Public Health Informatics, 5(3), 223. https://doi.org/10.5210/ojphi.v5i3.4933

Ola, O., & Sedig, K. (2018). Discourse with visual health data: Design of human-data interaction. Multimodal Technologies and Interaction, 2(1), 10. https://doi.org/10.3390/mti2010010

Ordonez, C. (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334–343. https://doi.org/10.1109/TITB.2006.864475

Ordonez, P., Oates, T., Lombardi, M. E., Hernandez, G., Holmes, K. W., Fackler, J., & Lehmann, C. U. (2012). Visualization of multivariate time-series data in a neonatal ICU. IBM Journal of Research and Development, 56(5), 7–1.

Panahiazar, M., Taslimitehrani, V., Pereira, N. L., & Pathak, J. (2015). Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics. Studies in Health Technology and Informatics, 210, 369–373.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. Indian Journal of Ophthalmology, 56(1), 45–50.

Parsons, P., & Sedig, K. (2014). Distribution of Information Processing While Performing Complex Cognitive Activities with Visualization Tools. In W. Huang (Ed.), Handbook of Human Centric Visualization (pp. 693–715). Springer. https://doi.org/10.1007/978-1-4614-7485-2_28

Parsons, P., Sedig, K., Mercer, R., Khordad, M., Knoll, J., & Rogan, P. (2015, October 25). Visual Analytics for Supporting Evidence-Based Interpretation of Molecular Cytogenomic Findings. https://doi.org/10.1145/2836034.2836036

Pavlyshenko, B. M. (2016). Linear, machine learning and probabilistic approaches for time series analysis. 2016 IEEE First International Conference on Data Stream Mining Processing (DSMP), 377–381. https://doi.org/10.1109/DSMP.2016.7583582

Perer, A., & Sun, J. (2012). MatrixFlow: Temporal Network Visual Analytics to Track Symptom Evolution during Disease Progression. AMIA Annual Symposium Proceedings, 2012, 716–725.

Perer, A., Wang, F., & Hu, J. (2015). Mining and exploring care pathways from electronic medical records with visual analytics. Journal of Biomedical Informatics, 56, 369–378. https://doi.org/10.1016/j.jbi.2015.06.020

Pieczkiewicz, D. S., Finkelstein, S. M., & Hertz, M. I. (2007). Design and evaluation of a web-based interactive visualization system for lung transplant home monitoring data. AMIA Annual Symposium Proceedings, 2007, 598.

Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D., & Shneiderman, B. (1998). LifeLines: Using visualization to enhance navigation and analysis. Of Patient Records", Proceedings of the American Medical Informatic Association Annual Fall Symposium.

Pohl, M., Wiltner, S., Rind, A., Aigner, W., Miksch, S., Turic, T., & Drexler, F. (2011). Patient development at a glance: An evaluation of a medical data visualization. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), Human-Computer Interaction – INTERACT 2011 (Vol. 6949, pp. 292–299). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23768-3_24

Putin, E., Mamoshina, P., Aliper, A., Korzinkin, M., Moskalev, A., Kolosov, A., Ostrovskiy, A., Cantor, C., Vijg, J., & Zhavoronkov, A. (2016). Deep biomarkers of human aging: Application of deep neural networks to biomarker development. Aging (Albany NY), 8(5), 1021–1030. https://doi.org/10.18632/aging.100968

Quinlan, J. R. (2014). C4. 5: Programs for machine learning. Elsevier.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. Health Information Science and Systems, 2(1), 3. https://doi.org/10.1186/2047-2501-2-3

Rahman, M. M., & Davis, D. N. (2013). Cluster Based Under-Sampling for Unbalanced Cardiovascular Data. 6.

Rajwan, Y. G., Barclay, P. W., Lee, T., Sun, I.-F., Passaretti, C., & Lehmann, H. (2013). Visualizing Central Line –Associated Blood Stream Infection (CLABSI) Outcome Data for Decision Making by Health Care Consumers and Practitioners—An Evaluation Study. Online Journal of Public Health Informatics, 5(2), 218. https://doi.org/10.5210/ojphi.v5i2.4364

Rao, R., & Card, S. K. (1994). The table lens: Merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 318–322.

Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. (2015). Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data, 3(4), 277–287.

Reisman, M. (2017). EHRs: The Challenge of Making Electronic Data Usable and Interoperable. Pharmacy and Therapeutics, 42(9), 572–575.

Ribarsky, W., Fisher, B., & Pottenger, W. M. (2009). Science of Analytical Reasoning: Information Visualization. https://doi.org/10.1057/ivs.2009.28

Richardson, A., Signor, B. M., Lidbury, B. A., & Badrick, T. (2016). Clinical chemistry in higher dimensions: Machine-learning and enhanced prediction from routine clinical chemistry data. Clinical Biochemistry, 49(16), 1213–1220. https://doi.org/10.1016/j.clinbiochem.2016.07.013

Rind, A., Aigner, W., Miksch, S., Wiltner, S., Pohl, M., Drexler, F., Neubauer, B., & Suchy, N. (2011). Visually exploring multivariate trends in patient cohorts using animated scatter plots. In M. M. Robertson (Ed.), Ergonomics and Health Aspects of Work with Computers (pp. 139–148). Springer Berlin Heidelberg.

Rind, A., Aigner, W., Miksch, S., Wiltner, S., Pohl, M., Turic, T., & Drexler, F. (2011). Visual exploration of time-oriented patient data for chronic diseases: Design study and evaluation. Symposium of the Austrian HCI and Usability Engineering Group, 301–320.

Rind, A., Wang, T. D., Aigner, W., Miksch, S., Wongsuphasawat, K., Plaisant, C., & Shneiderman, B. (2013). Interactive information visualization to explore and query electronic health records. Foundations and Trends® in Human–Computer Interaction, 5(3), 207–298. https://doi.org/10.1561/1100000039

Rostamzadeh, N., Abdullah, S. S., & Sedig, K. (2020). Data-Driven Activities Involving Electronic Health Records: An Activity and Task Analysis Framework for Interactive Visualization Tools. Multimodal Technologies and Interaction, 4(1), 7. https://doi.org/10.3390/mti4010007

Sacchi, L., Capozzi, D., Bellazzi, R., & Larizza, C. (2015). JTSA: An open source framework for time series abstractions. Computer Methods and Programs in Biomedicine, 121(3), 175–188. https://doi.org/10.1016/j.cmpb.2015.05.006

Saeed, M., Lieu, C., Raber, G., & Mark, R. G. (2002). MIMIC II: A massive temporal ICU patient database to support research in intelligent patient monitoring. Computers in Cardiology, 641–644. https://doi.org/10.1109/CIC.2002.1166854

Saffer, J. D., Burnett, V. L., Chen, G., & van der Spek, P. (2004). Visual analytics in the pharmaceutical industry. IEEE Computer Graphics and Applications, 24(5), 10–15. https://doi.org/10.1109/MCG.2004.40

Sahu, H., Shrma, S., & Gondhalakar, S. (2008). A Brief Overview on Data Mining Survey. Ijctee, 1(3), 114–121.

Salomon, G. (1997). Distributed Cognitions: Psychological and Educational Considerations. Cambridge University Press.

Sears, A., & Jacko, J. A. (2007). The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications, Second Edition. CRC Press.

Sedig, K., Naimi, A., & Haggerty, N. (2017). ALIGNING INFORMATION TECHNOLOGIES WITH EVIDENCEBASED HEALTH-CARE ACTIVITIES: A DESIGN AND EVALUATION FRAMEWORK. Human Technology, 13(2).

Sedig, K., & Parsons, P. (2013). Interaction design for complex cognitive activities with visual representations: A pattern-based approach. AIS Transactions on Human-Computer Interaction, 5(2), 84–133.

Sedig, K., & Parsons, P. (2016). Design of visualizations for human-information interaction: A pattern-based framework. Synthesis Lectures on Visualization, 4(1), 1–185. https://doi.org/10.2200/S00685ED1V01Y201512VIS005

Sedig, K., Parsons, P., & Babanski, A. (2012). Towards a characterization of interactivity in visual analytics. Journal of Multimedia Processing Technologies, 3, 12–28.

Shah, J. R. (2014). Electronic Health Records: Challenges and Opportunities. 189–204.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

Siegel, E. (2013). Predictive analytics: The power to predict who will click, buy, lie, or die. John Wiley & Sons.

Silow-Carroll, S., Edwards, J. N., & Rodin, D. (2012). Using electronic health records to improve quality and efficiency: The experiences of leading hospitals. Issue Brief (Commonwealth Fund), 17, 1–40.

Simpao, A. F., Ahumada, L. M., Desai, B. R., Bonafide, C. P., Galvez, J. A., Rehman, M. A., Jawad, A. F., Palma, K. L., & Shelov, E. D. (2014). Optimization of drug-drug interaction alert rules in a pediatric hospital's electronic health record system using a visual analytics dashboard. Journal of the American Medical Informatics Association, amiajnl-2013-002538. https://doi.org/10.1136/amiajnl-2013-002538

Simpao, Allan F., Ahumada, L. M., Gálvez, J. A., & Rehman, M. A. (2014). A review of analytics and clinical informatics in health care. Journal of Medical Systems, 38(4), 45. https://doi.org/10.1007/s10916-014-0045-x

Siwek, K., Osowski, S., Markiewicz, T., & Korytkowski, J. (2013). Analysis of medical data using dimensionality reduction techniques. Przegląd Elektrotechniczny, 89(2a), 279–281.

Somnay, Y. R., Craven, M., McCoy, K. L., Carty, S. E., Wang, T. S., Greenberg, C. C., & Schneider, D. F. (2017). Improving diagnostic recognition of primary hyperparathyroidism with machine learning. Surgery, 161(4), 1113–1121. https://doi.org/10.1016/j.surg.2016.09.044

Spinner, T., Schlegel, U., Schäfer, H., & El-Assady, M. (2020). explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. IEEE Transactions on Visualization and Computer Graphics, 26(1), 1064–1074. https://doi.org/10.1109/TVCG.2019.2934629

Stead, W. W., & Lin, H. S. (2009). *Computational Technology for Effective Health Care*. National Academies Press. https://doi.org/10.17226/12572

Stopar, L., Skraba, P., Grobelnik, M., & Mladenic, D. (2019). StreamStory: Exploring Multivariate Time Series on Multiple Scales. IEEE Transactions on Visualization and Computer Graphics, 25(4), 1788–1802. https://doi.org/10.1109/TVCG.2018.2825424

Strom, B. L., Schinnar, R., Jones, J., Bilker, W. B., Weiner, M. G., Hennessy, S., Leonard, C. E., Cronholm, P. F., & Pifer, E. (2011). Detecting pregnancy use of non-hormonal category X medications in electronic medical records. Journal of the American Medical Informatics Association, 18(Supplement_1), i81–i86. https://doi.org/10.1136/amiajnl-2010-000057

Sun, J., McNaughton, C. D., Zhang, P., Perer, A., Gkoulalas-Divanis, A., Denny, J. C., Kirby, J., Lasko, T., Saip, A., & Malin, B. A. (2014). Predicting changes in hypertension control using electronic health records from a chronic disease management program. Journal of the American Medical Informatics Association: JAMIA, 21(2), 337–344. https://doi.org/10.1136/amiajnl-2013-002033

Sun, W., Cai, Z., Liu, F., Fang, S., & Wang, G. (2017). A survey of data mining technology on electronic medical records. 2017 IEEE 19th International Conference on E-Health Networking, Applications and Services (Healthcom), 1–6. https://doi.org/10.1109/HealthCom.2017.8210774

Surinova, S., Choi, M., Tao, S., Schüffler, P. J., Chang, C.-Y., Clough, T., Vysloužil, K., Khoylou, M., Srovnal, J., & Liu, Y. (2015). Prediction of colorectal cancer diagnosis based on circulating plasma proteins. EMBO Molecular Medicine, 7(9), 1166–1178.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27, 3104–3112.

Tamayo, D., Silburt, A., Valencia, D., Menou, K., Ali-Dib, M., Petrovich, C., Huang, C. X., Rein, H., Laerhoven, C. van, Paradise, A., Obertas, A., & Murray, N. (2016). A MACHINE LEARNS TO PREDICT THE STABILITY OF TIGHTLY PACKED PLANETARY SYSTEMS. The Astrophysical Journal, 832(2), L22. https://doi.org/10.3847/2041-8205/832/2/L22

Tang, P. C., & McDonald, C. J. (2006). Electronic health record systems. In E. H. Shortliffe & J. J. Cimino (Eds.), Biomedical Informatics: Computer Applications in Health Care and Biomedicine (pp. 447–475). Springer New York. https://doi.org/10.1007/0-387-36278-9_12

Thomas, E. J., Studdert, D. M., Burstin, H. R., Orav, E. J., Zeena, T., Williams, E. J., Howard, K. M., Weiler, P. C., & Brennan, T. A. (2000). Incidence and Types of Adverse Events and Negligent Care in Utah and Colorado. Medical Care, 38(3), 261–271.

Thomas, E. J., Studdert, D. M., Newhouse, J. P., Zbar, B. I. W., Howard, K. M., Williams, E. J., & Brennan, T. A. (1999). Costs of Medical Injuries in Utah and Colorado. Inquiry, 36(3), 255–264.

Thomas, J. J., & Cook, K. A. (2005). Illuminating the Path: The Research and Development Agenda for Visual Analytics. https://www.hsdl.org/?abstract&did=

Tia Gao, Greenspan, D., Welsh, M., Juang, R. R., & Alm, A. (2005). Vital signs monitoring and patient tracking over a wireless network. 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, 102–105. https://doi.org/10.1109/IEMBS.2005.1616352

Timmerman, D., Testa, A. C., Bourne, T., Ferrazzi, E., Ameye, L., Konstantinovic, M. L., Van Calster, B., Collins, W. P., Vergote, I., Van Huffel, S., & Valentin, L. (2005). Logistic Regression Model to Distinguish Between the Benign and Malignant Adnexal Mass Before Surgery: A Multicenter Study by the International Ovarian Tumor Analysis Group. Journal of Clinical Oncology, 23(34), 8794–8801. https://doi.org/10.1200/JCO.2005.01.7632

Torio, C. M., Elixhauser, A., & Andrews, R. M. (2006). Trends in Potentially Preventable Hospital Admissions among Adults and Children, 2005–2010: Statistical Brief #151. In Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Agency for Healthcare Research and Quality (US). http://www.ncbi.nlm.nih.gov/books/NBK137748/

Treisman, A. (1985). Preattentive processing in vision. Computer Vision, Graphics, and Image Processing, 31(2), 156–177. https://doi.org/10.1016/S0734-189X(85)80004-9

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. In Nature Reviews Drug Discovery (Vol. 18, Issue 6, pp. 463–477). Nature Publishing Group. https://doi.org/10.1038/s41573-019-0024-5

Waikar, S. S., Curhan, G. C., Ayanian, J. Z., & Chertow, G. M. (2007). Race and Mortality after Acute Renal Failure. Journal of the American Society of Nephrology : JASN, 18(10), 2740–2748. https://doi.org/10.1681/ASN.2006091060

Waikar, S. S., Curhan, G. C., Wald, R., McCarthy, E. P., & Chertow, G. M. (2006). Declining Mortality in Patients with Acute Renal Failure, 1988 to 2002. Journal of the American Society of Nephrology, 17(4), 1143–1150. https://doi.org/10.1681/ASN.2005091017

Wang, L., Tong, L., Davis, D., Arnold, T., & Esposito, T. (2020). The application of unsupervised deep learning in predictive models using electronic health records. BMC Medical Research Methodology, 20(1), 37. https://doi.org/10.1186/s12874-020-00923-1

Wang, T. D., Plaisant, C., Shneiderman, B., Spring, N., Roseman, D., Marchand, G., Mukherjee, V., & Smith, M. (2009). Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. IEEE Transactions on Visualization and Computer Graphics, 15(6), 1049–1056. https://doi.org/10.1109/TVCG.2009.187

Wang, Taowei David, Plaisant, C., Quinn, A. J., Stanchak, R., Murphy, S., & Shneiderman, B. (2008). Aligning temporal data by sentinel events: Discovering patterns in electronic health records. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 457–466. https://doi.org/10.1145/1357054.1357129

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3–13. https://doi.org/10.1016/j.techfore.2015.12.019

Ware, C. (2019). Information Visualization: Perception for Design. Morgan Kaufmann.

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. Journal of the American Medical Informatics Association: JAMIA, 20(1), 144–151. https://doi.org/10.1136/amiajnl-2011-000681

Wilkinson, L. (2004). Classification and regression trees. Systat, 11, 35–56.

Wilson, R. M., Runciman, W. B., Gibberd, R. W., Harrison, B. T., Newby, L., & Hamilton, J. D. (1995). The Quality in Australian Health Care Study. Medical Journal of Australia, 163(9), 458–471. https://doi.org/10.5694/j.1326-5377.1995.tb124691.x

Wongsuphasawat, K., & Gotz, D. (2012). Exploring flow, factors, and outcomes of temporal event sequences with the Outflow visualization. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2659–2668. https://doi.org/10.1109/TVCG.2012.225

Wongsuphasawat, Krist. (2009). Finding comparable patient histories: A temporal categorical similarity measure with an interactive visualization. IEEE Symposium on Visual Analytics Science and Technology (VAST).

Wongsuphasawat, Krist, & Gotz, D. (2011). Outflow: Visualizing patient flow by symptoms and outcome. IEEE VisWeek Workshop on Visual Analytics in Healthcare, Providence, Rhode Island, USA, 25–28.

Wongsuphasawat, Krist, Guerra Gómez, J. A., Plaisant, C., Wang, T. D., Taieb-Maimon, M., & Shneiderman, B. (2011). LifeFlow: Visualizing an overview of event sequences. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1747–1756. https://doi.org/10.1145/1978942.1979196

Yadav, P., Pruinelli, L., Hangsleben, A., Dey, S., Hauwiller, K., Westra, B. L., Delaney, C. W., Kumar, V., Steinbach, M. S., & Simon, G. J. (2015). Modelling trajectories for diabetes complications. Proceedings of the 4th Workshop on Data Mining for Medicine and Healthcare. 2015 SIAM International Conference on Data Mining.

Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2018). Mining Electronic Health Records (EHRs): A Survey. ACM Computing Surveys, 50(6), 85:1–85:40. https://doi.org/10.1145/3127881

Yang, H. S., Hou, Y., Vasovic, L. V., Steel, P., Chadburn, A., Racine-Brzostek, S. E., Velu, P., Cushing, M. M., Loda, M., Kaushal, R., Zhao, Z., & Wang, F. (2020). Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. Clinical Chemistry. https://doi.org/10.1093/clinchem/hvaa200

Yazhini, K., & Loganathan, D. (2019). A State of Art Approaches on Deep Learning Models in Healthcare: An Application Perspective. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 195–200. https://doi.org/10.1109/ICOEI.2019.8862730

Yu, Y., Long, J., Liu, F., & Cai, Z. (2016). Machine Learning Combining with Visualization for Intrusion Detection: A Survey. In V. Torra, Y. Narukawa, G. Navarro-Arribas, & C. Yañez (Eds.), Modeling Decisions for Artificial Intelligence (pp. 239–249). Springer International Publishing. https://doi.org/10.1007/978-3-319-45656-0_20

Yuan, C., Ming, C., & Chengjin, H. (2012). UrineCART, a machine learning method for establishment of review rules based on UF-1000i flow cytometry and dipstick or reflectance photometer. Clinical Chemistry and Laboratory Medicine (CCLM), 50(12), 2155–2161.

Zaki, M. J. (2000). Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering, 12(3), 372–390. https://doi.org/10.1109/69.846291

Zhang, Y.-P., Zhang, L.-N., & Wang, Y.-C. (2010). Cluster-based majority under-sampling approaches for class imbalance learning. 2010 2nd IEEE International Conference on Information and Financial Engineering, 400–404.

Zhao, K., Ward, M., Rundensteiner, E., & Higgins, H. (2016). MaVis: Machine Learning Aided Multi-Model Framework for Time Series Visual Analytics. Electronic Imaging, 2016(1), 1–10.

Zhao, X., Wu, Y., Lee, D. L., & Cui, W. (2019). iForest: Interpreting Random Forests via Visual Analytics. IEEE Transactions on Visualization and Computer Graphics, 25(1), 407–416. https://doi.org/10.1109/TVCG.2018.2864475

Appendices

Appendix A: List of databases held at ICES used for the development of SUNRISE (an independent, non-profit, world-leading research organization that uses population-based health and social data to produce knowledge on a broad range of healthcare issues).

Data Source	Description	Study Purpose
Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System	The Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System collect diagnostic and procedural variables for inpatient stays and ED visits, respectively. Diagnostic and inpatient procedural coding use the 10th version of the Canadian Modified International Classification of Disease system 10th Revision (after 2002).	Cohort description, exposure, and outcome estimation
Dynacare (formerly known as Gamma-Dynacare Medical Laboratories)	Database that contains all outpatient laboratory test results from all Dynacare laboratory locations across Ontario since 2002. Dynacare is one of the three largest laboratory providers in Ontario and contains records on over 59 million tests each year.	Outpatient laboratory tests

Appendix B: Laboratory groups created by Eclat algorithm and their corresponding XGBoost tunning parameters and AUROC

Group	Laboratory Tests in the Group	AUROC	max_depth	eta	subsample	min_child_weigth	gamma
1	SBC,SCr,SK,SNa	0.78	9	0.084	0.843	0	3
2	SBC,SCr,SNa	0.77	8	0.031	0.963	10	2
3	SBC,SK,SNa	0.66	3	0.178	0.724	7	4
4	SBC,SCr,SK	0.76	7	0.068	0.972	10	2
5	SBC,SCr	0.76	9	0.047	0.912	5	5
6	SBC,SK	0.61	3	0.286	0.96	0	5
7	SBC,SNa	0.66	6	0.178	0.77	0	3

8	ACr,HGB,Pl,SCl,SCr,SK,SNa,WBC	0.81	7	0.16	0.989	5	3
9	ACr,Pl,SCl,SCr,SK,SNa,WBC	0.82	8	0.123	0.753	6	0
10	ACr,HGB,Pl,SCl,SK,SNa,WBC	0.76	3	0.268	0.721	4	3
11	ACr,HGB,Pl,SCl,SCr,SK,SNa	0.81	10	0.285	0.818	4	0
12	ACr,Pl,SCl,SCr,SK,SNa	0.81	8	0.196	0.785	0	2
13	ACr,HGB,Pl,SCl,SK,SNa	0.75	3	0.245	0.948	8	2
14	ACr,Pl,SCl,SK,SNa,WBC	0.75	4	0.295	0.741	1	5
15	ACr,HGB,SCl,SCr,SK,SNa,WBC	0.81	7	0.268	0.758	4	0
16	ACr,SCl,SCr,SK,SNa,WBC	0.81	6	0.209	0.819	9	0
17	ACr,HGB,SCl,SK,SNa,WBC	0.77	3	0.275	0.834	3	2
18	ACr,HGB,SCl,SCr,SK,SNa	0.81	9	0.279	0.948	0	2
19	ACr,SCl,SCr,SK,SNa	0.8	7	0.212	0.861	7	0
20	ACr,HGB,SCl,SK,SNa	0.75	3	0.286	0.96	0	5
21	ACr,SCl,SK,SNa,WBC	0.76	10	0.217	0.938	5	3
22	ACr,Pl,SCl,SK,SNa	0.74	7	0.231	0.766	2	5
23	ACr,HGB,Pl,SCl,SCr,SNa,WBC	0.81	8	0.274	0.856	3	4
24	ACr,Pl,SCl,SCr,SNa,WBC	0.81	10	0.269	0.718	0	1
25	ACr,HGB,Pl,SCl,SNa,WBC	0.76	9	0.188	0.925	7	4
26	ACr,HGB,Pl,SCl,SCr,SNa	0.81	7	0.268	0.95	1	2
27	ACr,Pl,SCl,SCr,SNa	0.8	7	0.284	0.728	5	4
28	ACr,HGB,Pl,SCl,SNa	0.74	3	0.279	0.714	7	3

29	ACr,Pl,SCl,SNa,WBC	0.74	10	0.084	0.921	5	3
30	ACr,HGB,SCl,SCr,SNa,WBC	0.81	10	0.267	0.91	6	5
31	ACr,SCl,SCr,SNa,WBC	0.81	6	0.169	0.752	9	2
32	ACr,HGB,SCl,SNa,WBC	0.77	4	0.264	0.772	6	3
33	ACr,HGB,SCl,SCr,SNa	0.8	9	0.274	0.981	5	2
34	ACr,SCl,SCr,SNa	0.8	9	0.027	0.714	2	0
35	ACr,HGB,SCl,SNa	0.75	4	0.295	0.741	1	5
36	ACr,SCl,SNa,WBC	0.74	6	0.22	0.81	3	4
37	ACr,Pl,SCl,SNa	0.72	3	0.264	0.702	0	2
38	ACr,SCl,SK,SNa	0.73	5	0.282	0.818	9	5
39	ACr,HGB,Pl,SCl,SCr,SK,WBC	0.81	9	0.112	0.977	3	5
40	ACr,Pl,SCl,SCr,SK,WBC	0.81	6	0.23	0.7	9	4
41	ACr,HGB,Pl,SCl,SK,WBC	0.78	6	0.289	0.761	3	4
42	ACr,HGB,Pl,SCl,SCr,SK	0.81	10	0.192	0.729	0	3
43	ACr,Pl,SCl,SCr,SK	0.8	5	0.292	0.754	3	0
44	ACr,HGB,Pl,SCl,SK	0.75	3	0.279	0.714	7	3
45	ACr,Pl,SCl,SK,WBC	0.75	3	0.269	0.915	0	1
46	ACr,HGB,SCl,SCr,SK,WBC	0.8	10	0.284	0.93	7	4
47	ACr,SCl,SCr,SK,WBC	0.8	7	0.29	0.834	3	0
48	ACr,HGB,SCl,SK,WBC	0.76	3	0.291	0.998	6	0
49	ACr,HGB,SCl,SCr,SK	0.81	9	0.267	0.941	9	1

50	ACr,SCl,SCr,SK	0.79	5	0.292	0.754	3	0
51	ACr,HGB,SCl,SK	0.75	3	0.271	0.863	7	1
52	ACr,SCl,SK,WBC	0.75	8	0.23	0.701	8	1
53	ACr,Pl,SCl,SK	0.74	4	0.295	0.741	1	5
54	ACr,HGB,Pl,SCl,SCr,WBC	0.81	10	0.248	0.991	9	0
55	ACr,Pl,SCl,SCr,WBC	0.8	8	0.234	0.831	1	4
56	ACr,HGB,Pl,SCl,WBC	0.77	4	0.29	0.926	4	3
57	ACr,HGB,Pl,SCl,SCr	0.8	9	0.208	0.929	3	1
58	ACr,Pl,SCl,SCr	0.79	9	0.277	0.718	1	2
59	ACr,HGB,Pl,SCl	0.75	3	0.284	0.764	8	1
60	ACr,Pl,SCl,WBC	0.74	5	0.28	0.956	5	2
61	ACr,HGB,SCl,SCr,WBC	0.8	7	0.27	0.765	4	3
62	ACr,SCl,SCr,WBC	0.8	5	0.257	0.805	2	4
63	ACr,HGB,SCl,WBC	0.76	3	0.265	0.898	6	1
64	ACr,HGB,SCl,SCr	0.8	10	0.191	0.728	0	3
65	ACr,SCl,SCr	0.8	8	0.298	0.987	4	3
66	ACr,HGB,SCl	0.74	3	0.221	0.833	8	2
67	ACr,SCl,WBC	0.73	5	0.293	0.73	4	2
68	ACr,Pl,SCl	0.73	6	0.215	0.888	8	5
69	ACr,SCl,SK	0.73	3	0.284	0.764	8	1
70	ACr,SCl,SNa	0.72	7	0.291	0.723	2	5

71	ACr,HGB,Pl,SCr,SK,SNa,WBC	0.83	6	0.294	0.912	2	0
72	ACr,Pl,SCr,SK,SNa,WBC	0.82	5	0.291	0.952	1	4
73	ACr,HGB,Pl,SK,SNa,WBC	0.77	4	0.287	0.771	6	0
74	ACr,HGB,Pl,SCr,SK,SNa	0.82	10	0.191	0.945	1	5
75	ACr,Pl,SCr,SK,SNa	0.82	6	0.25	0.763	5	5
76	ACr,HGB,Pl,SK,SNa	0.74	5	0.146	0.844	3	1
77	ACr,Pl,SK,SNa,WBC	0.75	5	0.238	0.92	6	4
78	ACr,HGB,SCr,SK,SNa,WBC	0.83	9	0.17	0.718	5	2
79	ACr,SCr,SK,SNa,WBC	0.83	10	0.175	0.76	5	5
80	ACr,HGB,SK,SNa,WBC	0.76	8	0.077	0.733	5	5
81	ACr,HGB,SCr,SK,SNa	0.81	4	0.276	0.923	2	2
82	ACr,SCr,SK,SNa	0.8	4	0.27	0.726	1	3
83	ACr,HGB,SK,SNa	0.75	4	0.228	0.945	7	0
84	ACr,SK,SNa,WBC	0.75	5	0.166	0.955	9	0
85	ACr,Pl,SK,SNa	0.72	4	0.201	0.784	0	4
86	ACr,HGB,Pl,SCr,SNa,WBC	0.82	7	0.278	0.927	2	4
87	ACr,Pl,SCr,SNa,WBC	0.83	7	0.27	0.765	4	3
88	ACr,HGB,Pl,SNa,WBC	0.76	3	0.278	0.988	0	1
89	ACr,HGB,Pl,SCr,SNa	0.82	10	0.187	0.891	4	5
90	ACr,Pl,SCr,SNa	0.81	3	0.219	0.823	8	0
91	ACr,HGB,Pl,SNa	0.74	3	0.095	0.855	3	4

92	ACr,Pl,SNa,WBC	0.74	6	0.289	0.79	9	5
93	ACr,HGB,SCr,SNa,WBC	0.82	9	0.277	0.948	10	4
94	ACr,SCr,SNa,WBC	0.83	8	0.141	0.714	10	1
95	ACr,HGB,SNa,WBC	0.76	4	0.173	0.773	6	4
96	ACr,HGB,SCr,SNa	0.82	7	0.278	0.927	2	4
97	ACr,SCr,SNa	0.81	3	0.297	0.988	9	5
98	ACr,HGB,SNa	0.74	4	0.043	0.725	1	4
99	ACr,SNa,WBC	0.74	5	0.216	0.799	1	3
100	ACr,Pl,SNa	0.71	3	0.265	0.898	6	1
101	ACr,SK,SNa	0.71	5	0.288	0.888	10	5
102	ACr,HGB,Pl,SCr,SK,WBC	0.83	10	0.206	0.828	9	2
103	ACr,Pl,SCr,SK,WBC	0.83	9	0.186	0.889	3	5
104	ACr,HGB,Pl,SK,WBC	0.77	7	0.231	0.766	2	5
105	ACr,HGB,Pl,SCr,SK	0.82	5	0.176	0.999	2	4
106	ACr,Pl,SCr,SK	0.82	5	0.213	0.719	6	3
107	ACr,HGB,Pl,SK	0.75	5	0.085	0.929	5	4
108	ACr,Pl,SK,WBC	0.75	3	0.278	0.988	0	1
109	ACr,HGB,SCr,SK,WBC	0.83	6	0.205	0.799	1	4
110	ACr,SCr,SK,WBC	0.83	5	0.236	0.755	1	4
111	ACr,HGB,SK,WBC	0.77	3	0.284	0.764	8	1
112	ACr,HGB,SCr,SK	0.82	10	0.191	0.945	1	5

113	ACr,SCr,SK	0.81	3	0.275	0.929	8	1
114	ACr,HGB,SK	0.74	4	0.028	0.828	9	5
115	ACr,SK,WBC	0.75	9	0.181	0.83	10	4
116	ACr,Pl,SK	0.72	5	0.238	0.92	6	4
117	ACr,HGB,Pl,SCr,WBC	0.82	8	0.249	0.812	6	3
118	ACr,Pl,SCr,WBC	0.82	10	0.284	0.94	6	4
119	ACr,HGB,Pl,WBC	0.76	7	0.284	0.728	5	4
120	ACr,HGB,Pl,SCr	0.82	10	0.193	0.81	4	4
121	ACr,Pl,SCr	0.81	3	0.275	0.834	3	2
122	ACr,HGB,Pl	0.74	4	0.201	0.717	10	4
123	ACr,Pl,WBC	0.73	5	0.248	0.888	0	2
124	ACr,HGB,SCr,WBC	0.81	4	0.276	0.923	2	2
125	ACr,SCr,WBC	0.82	7	0.291	0.723	2	5
126	ACr,HGB,WBC	0.75	4	0.145	0.746	9	5
127	ACr,HGB,SCr	0.82	5	0.281	0.715	10	1
128	ACr,SCr	0.81	4	0.264	0.772	6	3
129	ACr,HGB	0.74	8	0.134	0.717	0	5
130	ACr,WBC	0.73	7	0.269	0.871	8	4
131	ACr,Pl	0.71	3	0.291	0.812	5	2
132	ACr,SK	0.7	3	0.286	0.96	0	5
133	ACr,SNa	0.7	8	0.224	0.757	7	3

134	ACr,SCl	0.72	7	0.148	0.823	5	5
135	HGB,Pl,SCl,SCr,SK,SNa,WBC	0.8	6	0.289	0.761	3	4
136	Pl,SCl,SCr,SK,SNa,WBC	0.79	3	0.293	0.736	8	2
137	HGB,Pl,SCl,SK,SNa,WBC	0.74	9	0.28	0.788	5	5
138	HGB,Pl,SCl,SCr,SK,SNa	0.8	5	0.282	0.818	9	5
139	Pl,SCl,SCr,SK,SNa	0.79	5	0.229	0.717	7	5
140	HGB,Pl,SCl,SK,SNa	0.72	8	0.134	0.717	0	5
141	Pl,SCl,SK,SNa,WBC	0.66	8	0.265	0.953	1	5
142	HGB,SCl,SCr,SK,SNa,WBC	0.8	5	0.282	0.818	9	5
143	SCl,SCr,SK,SNa,WBC	0.78	5	0.168	0.731	0	3
144	HGB,SCl,SK,SNa,WBC	0.72	3	0.264	0.702	0	2
145	HGB,SCl,SCr,SK,SNa	0.8	7	0.203	0.775	7	4
146	SCl,SCr,SK,SNa	0.78	5	0.282	0.818	9	5
147	HGB,SCl,SK,SNa	0.71	7	0.269	0.814	3	5
148	SCl,SK,SNa,WBC	0.65	8	0.26	0.776	1	5
149	Pl,SCl,SK,SNa	0.65	9	0.223	0.949	8	3
150	HGB,Pl,SCl,SCr,SNa,WBC	0.8	9	0.153	0.754	10	5
151	Pl,SCl,SCr,SNa,WBC	0.78	6	0.197	0.925	4	3
152	HGB,Pl,SCl,SNa,WBC	0.73	5	0.271	0.878	3	0
153	HGB,Pl,SCl,SCr,SNa	0.79	4	0.26	0.721	0	3
154	Pl,SCl,SCr,SNa	0.78	4	0.253	0.806	4	4

155	HGB,Pl,SCl,SNa	0.71	7	0.127	0.748	5	2
156	Pl,SCl,SNa,WBC	0.65	7	0.255	0.84	0	3
157	HGB,SCl,SCr,SNa,WBC	0.8	5	0.134	0.916	3	1
158	SCl,SCr,SNa,WBC	0.78	4	0.267	0.825	10	4
159	HGB,SCl,SNa,WBC	0.72	5	0.288	0.888	10	5
160	HGB,SCl,SCr,SNa	0.79	5	0.135	0.742	8	4
161	SCl,SCr,SNa	0.78	9	0.28	0.788	5	5
162	HGB,SCl,SNa	0.7	10	0.233	0.784	5	5
163	SCl,SNa,WBC	0.65	8	0.26	0.776	1	5
164	Pl,SCl,SNa	0.63	10	0.191	0.945	1	5
165	SCl,SK,SNa	0.64	6	0.267	0.868	1	1
166	HGB,Pl,SCl,SCr,SK,WBC	0.81	9	0.28	0.788	5	5
167	Pl,SCl,SCr,SK,WBC	0.79	6	0.289	0.79	9	5
168	HGB,Pl,SCl,SK,WBC	0.73	4	0.228	0.945	7	0
169	HGB,Pl,SCl,SCr,SK	0.8	7	0.278	0.927	2	4
170	Pl,SCl,SCr,SK	0.78	5	0.245	0.968	6	0
171	HGB,Pl,SCl,SK	0.71	6	0.133	0.806	7	4
172	Pl,SCl,SK,WBC	0.66	6	0.289	0.761	3	4
173	HGB,SCl,SCr,SK,WBC	0.8	5	0.194	0.79	6	4
174	SCl,SCr,SK,WBC	0.79	4	0.245	0.767	7	4
175	HGB,SCl,SK,WBC	0.72	3	0.241	0.714	9	2

176	HGB,SCl,SCr,SK	0.8	5	0.176	0.763	3	2
177	SCl,SCr,SK	0.78	7	0.164	0.743	6	5
178	HGB,SCl,SK	0.71	9	0.285	0.846	10	3
179	SCl,SK,WBC	0.64	7	0.261	0.732	0	2
180	Pl,SCl,SK	0.63	5	0.283	0.923	4	3
181	HGB,Pl,SCl,SCr,WBC	0.8	7	0.048	0.745	2	5
182	Pl,SCl,SCr,WBC	0.78	5	0.227	0.771	8	2
183	HGB,Pl,SCl,WBC	0.73	3	0.271	0.849	5	0
184	HGB,Pl,SCl,SCr	0.79	5	0.271	0.878	3	0
185	Pl,SCl,SCr	0.78	6	0.149	0.756	7	0
186	HGB,Pl,SCl	0.7	7	0.184	0.876	7	5
187	Pl,SCl,WBC	0.64	9	0.22	0.805	1	4
188	HGB,SCl,SCr,WBC	0.8	3	0.279	0.714	7	3
189	SCl,SCr,WBC	0.78	6	0.076	0.715	7	5
190	HGB,SCl,WBC	0.71	3	0.28	0.936	6	0
191	HGB,SCl,SCr	0.8	4	0.245	0.767	7	4
192	SCl,SCr	0.78	3	0.232	0.837	0	0
193	HGB,SCl	0.7	5	0.292	0.754	3	0
194	SCl,WBC	0.62	9	0.281	0.841	9	5
195	Pl,SCl	0.6	8	0.295	0.823	4	0
196	SCl,SK	0.61	9	0.277	0.948	10	4

197	SCl,SNa	0.64	9	0.232	0.873	8	1
198	HGB,Pl,SCr,SK,SNa,WBC	0.8	7	0.219	0.969	1	5
199	Pl,SCr,SK,SNa,WBC	0.8	4	0.287	0.771	6	0
200	HGB,Pl,SK,SNa,WBC	0.73	8	0.134	0.717	0	5
201	HGB,Pl,SCr,SK,SNa	0.8	5	0.273	0.885	5	5
202	Pl,SCr,SK,SNa	0.79	8	0.058	0.713	0	4
203	HGB,Pl,SK,SNa	0.69	7	0.291	0.723	2	5
204	Pl,SK,SNa,WBC	0.66	5	0.27	0.847	0	5
205	HGB,SCr,SK,SNa,WBC	0.8	5	0.201	0.744	3	4
206	SCr,SK,SNa,WBC	0.79	4	0.201	0.784	0	4
207	HGB,SK,SNa,WBC	0.71	7	0.291	0.723	2	5
208	HGB,SCr,SK,SNa	0.8	5	0.288	0.888	10	5
209	SCr,SK,SNa	0.79	6	0.22	0.81	3	4
210	HGB,SK,SNa	0.69	9	0.257	0.768	7	5
211	SK,SNa,WBC	0.64	5	0.283	0.923	4	3
212	Pl,SK,SNa	0.64	6	0.285	0.994	2	3
213	HGB,Pl,SCr,SNa,WBC	0.8	3	0.291	0.812	5	2
214	Pl,SCr,SNa,WBC	0.79	3	0.291	0.812	5	2
215	HGB,Pl,SNa,WBC	0.71	5	0.258	0.846	6	2
216	HGB,Pl,SCr,SNa	0.8	4	0.26	0.721	0	3
217	Pl,SCr,SNa	0.79	5	0.194	0.79	6	4

218	HGB,Pl,SNa	0.69	7	0.206	0.732	4	4
219	Pl,SNa,WBC	0.64	4	0.287	0.771	6	0
220	HGB,SCr,SNa,WBC	0.8	6	0.164	0.878	3	5
221	SCr,SNa,WBC	0.79	4	0.297	0.774	5	2
222	HGB,SNa,WBC	0.7	7	0.164	0.743	6	5
223	HGB,SCr,SNa	0.79	4	0.198	0.856	3	5
224	SCr,SNa	0.79	3	0.165	0.726	9	3
225	HGB,SNa	0.68	8	0.26	0.776	1	5
226	SNa,WBC	0.61	9	0.299	0.746	10	3
227	Pl,SNa	0.61	9	0.194	0.951	7	4
228	SK,SNa	0.63	5	0.283	0.923	4	3
229	HGB,Pl,SCr,SK,WBC	0.8	9	0.103	0.817	4	5
230	Pl,SCr,SK,WBC	0.8	4	0.201	0.717	10	4
231	HGB,Pl,SK,WBC	0.72	7	0.255	0.84	0	3
232	HGB,Pl,SCr,SK	0.8	5	0.245	0.968	6	0
233	Pl,SCr,SK	0.79	8	0.208	0.955	6	5
234	HGB,Pl,SK	0.69	9	0.159	0.953	10	5
235	Pl,SK,WBC	0.66	5	0.166	0.782	3	4
236	HGB,SCr,SK,WBC	0.8	7	0.106	0.776	0	2
237	SCr,SK,WBC	0.79	5	0.248	0.888	0	2
238	HGB,SK,WBC	0.7	5	0.291	0.952	1	4

239	HGB,SCr,SK	0.8	6	0.289	0.761	3	4
240	SCr,SK	0.79	6	0.081	0.787	8	3
241	HGB,SK	0.68	8	0.277	0.921	4	0
242	SK,WBC	0.64	9	0.11	0.738	10	5
243	Pl,SK	0.63	9	0.197	0.889	2	5
244	HGB,Pl,SCr,WBC	0.8	5	0.227	0.771	8	2
245	Pl,SCr,WBC	0.79	4	0.179	0.708	8	5
246	HGB,Pl,WBC	0.71	4	0.198	0.856	3	5
247	HGB,Pl,SCr	0.79	3	0.019	0.702	2	2
248	Pl,SCr	0.79	5	0.177	0.988	4	5
249	HGB,Pl	0.68	7	0.189	0.723	8	2
250	Pl,WBC	0.63	3	0.196	0.748	1	0
251	HGB,SCr,WBC	0.8	4	0.228	0.945	7	0
252	SCr,WBC	0.79	4	0.264	0.772	6	3
253	HGB,WBC	0.7	5	0.05	0.888	10	1
254	HGB,SCr	0.79	7	0.026	0.989	8	5
255	SCr	0.79	10	0.205	0.967	8	4
256	HGB	0.67	8	0.272	0.765	0	0
257	WBC	0.6	8	0.23	0.701	8	1
258	Pl	0.56	8	0.286	0.967	3	2
259	SK	0.62	9	0.259	0.704	1	0

260	SNa	0.59	6	0.133	0.855	0	0
261	SCl	0.6	10	0.048	0.797	10	4
262	ACr	0.7	5	0.163	0.872	3	3
263	SBC	0.62	3	0.144	0.91	10	4

Appendix C: List of databases held at ICES used for the development of VERONICA

Data Source	Description	Study Purpose
Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System	The Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System collect diagnostic and procedural variables for inpatient stays and ED visits, respectively. Diagnostic and inpatient procedural coding use the 10th version of the Canadian Modified International Classification of Disease system 10th Revision (after 2002).	Cohort creation, description, exposure, and outcome estimation
Ontario Drug Benefits	The Ontario Drug Benefits database includes a wide range of outpatient prescription medications available to all Ontario citizens over the age of 65. The error rate in the Ontario Drug Benefits database is less than 1%.	Medication prescriptions, description, and exposure
Registered Persons Database	The Registered Persons Database captures demographic (sex, date of birth, postal code) and vital status information on all Ontario residents. Relative to the Canadian Institute for Health Information Discharge Abstract Database in-hospital death flag, the Registered Persons Database has a sensitivity of 94% and a positive predictive value of 100%.	Cohort creation, description, and exposure
Ontario Health Insurance Plan	The Ontario Health Insurance Plan database contains information on Ontario physician billing claims for medical services using fee and diagnosis codes outlined in the Ontario Health Insurance Plan Schedule of Benefits. These codes	Cohort creation, stratification, description,

	capture information on outpatient, inpatient, and laboratory services rendered to a patient.	exposure, and outcome

Appendix D: Coding definitions for comorbid conditions.

Variable	Database	Code	Set Code
Major cancer	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	150, 154, 155, 157, 162, 174, 175, 185, 203, 204, 205, 206, 207, 208, 2303, 2304, 2307, 2330, 2312, 2334
		International Classification of Diseases 10th Revision	971, 980, 982, 984, 985, 986, 987, 988, 989, 990, 991, 993, C15, C18, C19, C20, C22, C25, C34, C50, C56, C61, C82, C83, C85, C91, C92, C93, C94, C95, D00, D010, D011, D012, D022, D075, D05
	Ontario Health Insurance Plan	Diagnosis	203, 204, 205, 206, 207, 208, 150, 154, 155, 157, 162, 174, 175, 183, 185
Chronic liver disease	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	4561, 4562, 070, 5722, 5723, 5724, 5728, 573, 7824, V026, 571, 2750, 2751, 7891, 7895
		International Classification of Diseases 10th Revision	B16, B17, B18, B19, I85, R17, R18, R160, R162, B942, Z225, E831, E830, K70, K713, K714, K715, K717, K721, K729, K73, K74, K753, K754, K758, K759, K76, K77
	Ontario Health Insurance Plan	Diagnosis	571, 573, 070
		Fee code	Z551, Z554

Coronary artery disease (excluding angina)	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	4801, 4802, 4803, 4804, 4805, 481, 482, 483
		Canadian Classification of Health Interventions	1IJ50, 1IJ76
		International Classification of Diseases 9th Revision	412, 410, 411
		International Classification of Diseases 10th Revision	I21, I22, Z955, T822
	Ontario Health Insurance Plan	Diagnosis	410, 412
		Fee code	R741, R742, R743, G298, E646, E651, E652, E654, E655, Z434, Z448
Diabetes	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	250
		International Classification of Diseases 10th Revision	E10, E11, E13, E14
	Ontario Health Insurance Plan	Diagnosis	250
		Fee code	Q040, K029, K030, K045, K046

Heart failure	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	4961, 4962, 4963, 4964
		Canadian Classification of Health Interventions	1HP53, 1HP55, 1HZ53GRFR, 1HZ53LAFR, 1HZ53SYFR
		International Classification of Diseases 9th Revision	I500, I501, I509, I255, J81
		International Classification of Diseases 10th Revision	I21, I22, Z955, T822
	Ontario Health Insurance Plan	Diagnosis	428
		Fee code	R701, R702, Z429
Hypertension	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	401, 402, 403, 404, 405
		International Classification of Diseases 10th Revision	I10, I11, I12, I13, I15
	Ontario Health Insurance Plan	Diagnosis	401, 402, 403

Kidney stones	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	5920, 5921, 5929, 5940, 5941, 5942, 5948, 5949, 27411
		International Classification of Diseases 10th Revision	N200, N201, N202, N209, N210, N211, N218, N219, N220, N228
Peripheral vascular disease	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	5125, 5129, 5014, 5016, 5018, 5028, 5038, 5126, 5159
		Canadian Classification of Health Interventions	1KA76, 1KA50, 1KE76, 1KG50, 1KG57, 1KG76MI, 1KG87, 1IA87LA, 1IB87LA, 1IC87LA, 1ID87LA, 1KA87LA, 1KE57
		International Classification of Diseases 9th Revision	4402, 4408, 4409, 5571, 4439, 444
		International Classification of Diseases 10th Revision	I700, I702, I708, I709, I731, I738, I739, K551
	Ontario Health Insurance Plan	Fee code	R787, R780, R797, R804, R809, R875, R815, R936, R783, R784, R785, E626, R814, R786, R937, R860, R861, R855, R856, R933, R934, R791, E672, R794, R813, R867, E649

Cerebrovascular disease (stroke or transient ischemic attack)	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	430, 431, 432, 4340, 4341, 4349, 435, 436, 3623
		International Classification of Diseases 10th Revision	I62, I630, I631, I632, I633, I634, I635, I638, I639, I64, H341, I600, I601, I602, I603, I604, I605, I606, I607, I609, I61, G450, G451, G452, G453, G458, G459, H340
Chronic kidney disease	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	4030, 4031, 4039, 4040, 4041, 4049, 585, 586, 5888, 5889, 2504
		International Classification of Diseases 10th Revision	E102, E112, E132, E142, I12, I13, N08, N18, N19
	Ontario Health Insurance Plan	Diagnosis	403, 585

Appendix E: Diagnostic codes for health care utilization characteristics.

Variable	Database	Code	Set Code
Family physician visit	Ontario Health Insurance Plan	Fee code	A001, A003, A004, A005, A006, A007, A008, A900, A901, A905, A911, A912, A967, K131, K132, K140, K141, K142, K143, K144, W003, W008, W121

Appendix F: Diagnostic codes for exclusion criteria.

Variable	Database	Code Set	Code

Dialysis	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	5127, 5142, 5143, 5195, 6698
		Canadian Classification of Health Interventions	1PZ21, 1OT53DATS, 1OT53HATS, 1OT53LATS, 1SY55LAFT, 7SC59QD, 1KY76, 1KG76MZXXA, 1KG76MZXXN, 1JM76NC, 1JM76NCXXN
		International Classification of Diseases 9th Revision	V451, V560, V568, 99673
		International Classification of Diseases 10th Revision	T824, Y602, Y612, Y622, Y841, Z49, Z992
	Ontario Health Insurance Plan	Fee code	R850, G324, G336, G327, G862, G865, G099, R825, R826, R827, R833, R840, R841, R843, R848, R851, R946, R943, R944, R945, R941, R942, Z450, Z451, Z452, G864, R852, R853, R854, R885, G333, H540, H740, R849, G323, G325, G326, G860, G863, G866, G330, G331, G332, G861, G082, G083, G085, G090, G091, G092, G093, G094, G095, G096, G294, G295
Kidney transplant	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Health Interventions	1PC85

	Ontario Health Insurance Plan	Fee code	S435, S434

Curriculum Vitae

Name:	Neda Rostamzadeh
Post-secondary Education and Degrees:	The University of Western Ontario London, Ontario, Canada Ph.D. (2014-2021)
	Linkoping University Norrkoping, Sweden M.Sc. (2010-2014)
	Shahid Beheshti University Tehran, Iran B.Sc. (2005-2009)
Honours and Awards:	Western Graduate Research Scholarship 2014-2018
Related Work Experience	Teaching Assistant The University of Western Ontario 2014-2020

Publications:

Rostamzadeh, N.; Abdullah, S.S.; Sedig, K. Data-Driven Activities Involving Electronic Health Records: An Activity and Task Analysis Framework for Interactive Visualization Tools. *Multimodal Technol. Interact.* 2020, *4*, 7.

Abdullah, S.S.; Rostamzadeh, N.; Sedig, K; Garg, A; McArthur E. Multiple regression analysis and frequent itemset mining of electronic medical records: A visual analytics approach using VISA_M3R3. *Multimodal Technol. Data.* 2020*, 3,5.*

Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Garg, A.X.; McArthur, E. Visual Analytics for Dimension Reduction and Cluster Analysis of High Dimensional Electronic Health Records. *Informatics* 2020, *7*, 17.

Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Lizotte, D.J.; Garg, A.X.; McArthur, E. Machine Learning for Identifying Medication-Associated Acute Kidney Injury. *Informatics* 2020, *7*, 18.

Abdullah, S.S.; Rostamzadeh, N.; Sedig, K; Garg, A; McArthur E. Predicting Acute Kidney Injury: A Machine Learning Approach using Electronic Health Records. *Accepted for publication in Information, July, 2020*.

Rostamzadeh, N. A Comparison of Volumetric Illumination Methods by Considering their Underlying Mathematical Models. In Proceedings of SIGRAD 2013; Visual Computing; June 13-14; 2013; Norrköping; Sweden; Linköping University Electronic Press, 2013; pp. 35–40.

Rostamzadeh, N.; Abdullah, S.S.; Sedig, K.Visual Analytics for Electronic Health Records: a Review. December 2020. Manuscript submitted for publication.

Rostamzadeh, N.; Abdullah, S.S.; Sedig, K; Garg, A; McArthur E. Visual Analytics for Predicting Disease Outcomes Using Laboratory Test Results. December 2020. Manuscript submitted for publication.

Rostamzadeh, N.; Abdullah, S.S.; Sedig, K; Garg, A; McArthur E. VERONICA: Visual Analytics for Identifying Feature Groups in Disease Classification. December 2020. Manuscript submitted for publication.