7-15-2020 10:30 AM

Visual Analytics of Electronic Health Records with a focus on Acute Kidney Injury

Sheikh S. Abdullah, *The University of Western Ontario*

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

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

Recommended Citation

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

Abstract

The increasing use of electronic platforms in healthcare has resulted in the generation of unprecedented amounts of data in recent years. The amount of data available to clinical researchers, physicians, and healthcare administrators continues to grow, which creates an untapped resource with the ability to improve the healthcare system drastically. Despite the enthusiasm for adopting electronic health records (EHRs), some recent studies have shown that EHR-based systems hardly improve the ability of healthcare providers to make better decisions. One reason for this inefficacy is that these systems do not allow for human-data interaction in a manner that fits and supports the needs of healthcare providers. Another reason is the information overload, which makes healthcare providers often misunderstand, misinterpret, ignore, or overlook vital data. The emergence of a type of computational system known as visual analytics (VA), has the potential to reduce the complexity of EHR data by combining advanced analytics techniques with interactive visualizations to analyze, synthesize, and facilitate high-level activities while allowing users to get more involved in a discourse with the data. The purpose of this research is to demonstrate the use of sophisticated visual analytics systems to solve various EHR-related research problems. This dissertation includes a framework by which we identify gaps in existing EHR-based systems and conceptualize the data-driven activities and tasks of our proposed systems. Two novel VA systems (VISA_M3R3 and VALENCIA) and two studies are designed to bridge the gaps. VISA_M3R3 incorporates multiple regression, frequent itemset mining, and interactive visualization to assist users in the identification of nephrotoxic medications. Another proposed system, VALENCIA, brings a wide range of dimension reduction and cluster analysis techniques to analyze high-dimensional EHRs, integrate them seamlessly, and make them accessible through interactive visualizations. The studies are conducted to develop prediction models to classify patients who are at risk of developing acute kidney injury (AKI) and identify AKI-associated medication and medication combinations using EHRs. Through healthcare administrative datasets stored at the ICES-KDT (Kidney Dialysis and Transplantation program), London, Ontario, we have demonstrated how our proposed systems and prediction models can be used to solve real-world problems.

Keywords

Visual Analytics; Visualization; Analytics; Interaction; Human-Data Interaction; Human-Computer Interaction; Electronic Health Records; Acute Kidney Injury; Machine Learning; Dimension Reduction, Cluster Analysis; Medication-Induced Acute Kidney Injury; Health-Related Tasks; Prediction Models

Summary for Lay Audience

Advances in healthcare technology have resulted in the generation of large amounts of electronic data in the form of electronic health records (EHRs). Adoption of EHR makes it easy to organize, access, and store medical records through computerized data management tools. Despite the potential benefits, healthcare professionals continue to report difficulty in adopting EHR-based systems. One of the main reasons for this problem is the complicated and improperly designed user interfaces in these systems, which often makes healthcare providers overlook vital information. The purpose of this research is to prove the use of visual analytics (VA) to solve various EHR-related problems. VA combines automated analysis with interactive visualizations for effective reasoning, understanding and decision making based on complex data. Through a literature survey and proposed framework, we first analyze the existing EHR-based systems and understand why they fail to fulfill the computational demand of EHRs. Two novel VA systems (VISA_M3R3 and VALENCIA) and two studies are designed to demonstrate how the VA approach can be used to overcome the challenges of EHRs. VISA_M3R3 is designed to assist healthcare providers in the identification of medications that may associate with a higher risk of developing acute kidney injury (AKI). VALENCIA provides users with the ability to explore high-dimensional EHRs using a number of dimension reduction and cluster analysis algorithms. The studies are conducted to identify AKI-associated medication and medication combinations and predict the risk of developing AKI using EHRs. Through healthcare administrative datasets stored at the ICES-KDT (Kidney Dialysis and Transplantation program), we have shown how our proposed approach can be used to solve real-world problems.

Co-Authorship Statement

Chapter 1 is my original work in explaining the motivation, identifying the problem, introducing the dissertation, and describing associations between different sections. Chapters 2, which focus on a survey and proposed framework, was a collaborative effort with my supervisor, Kamran Sedig and a colleague, Neda Rostamzadeh (another graduate student working under the supervision of Kamran Sedig).

Chapters 3, 4, 5, and 6, were a collaborative effort with my supervisor, Kamran Sedig, and three colleagues, namely—Neda Rostamzadeh, Amit Garg, and Eric McArthur. It is important to mention that all the systems and studies presented in this dissertation incorporate personal health-related datasets stored at ICES. It is a mandatory requirement that any publication that uses ICES data must include responsible ICES scientists as co-authors. Amit Garg is a senior scientist and program lead of the Kidney, Dialysis & Transplantation Research Program. Eric McArthur is a Local lead analyst at ICES Western. Amit Garg and Eric McArthur were responsible for providing us access to the data and ensuring the published results comply with the ICES' privacy guidelines. Daniel Lizotte, an assistant professor in the Department of Computer Science and the Department of Epidemiology & Biostatistics, helped me to conceptualize the study presented in Chapter 4.

I was primarily responsible for the design, analysis, implementation, and writing of the original draft for the visual analytics systems and studies presented in Chapters 3,4, 5, and 6. My supervisor and Neda Rostamzadeh helped me with the conceptualization and revision of the manuscripts. Finally, Chapter 8 is my original work, to summarize the chapters and outline future research areas.

Acknowledgments

First, I would like to say sincere thanks to my supervisor, Kamran Sedig. His guidance, mentorship, and insights in this research have made this a valuable experience for me. I am truly grateful to Amit Garg for his support and guidance. I would like to thank all the members of the Insight lab, especially Neda Rostamzadeh. Thank you for assisting me throughout this project. I would like to express my gratitude and appreciation for Rey Acedillo, Eric McArthur, Elaine Zibrowski, Flory Tsobo Muanda, and Stephanie Dixon, whose brainstorming sessions, suggestions, and reviews have been invaluable throughout this journey. I would like to thank Daniel Lizotte for his valuable advice. I would like to thank Janice Wiersma and Sean Leonard for being a constant support. I am also thankful to the Western Computer Science Department and all its member's staff.

And my biggest thanks to my family for all the love and support you have shown me. My parents are the reason I am here today, thanks for believing in me. I am extremely grateful to my wife for her constant support and patience. Finally, I would like to thank almighty God for providing me with the courage and strength to complete this research.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Introduction

1.1 Motivation

The increasing use of electronic platforms in healthcare has produced unprecedented amounts of data in recent years. In the healthcare industry, as a part of modernizing their operations, the medical organizations are adopting electronic health records (EHRs) and deploying new information technology systems that generate, collect, digitize, and analyze their data (Caban and Gotz, 2015). This data includes, but is not limited to, medical and demographic records of patients, hospital and emergency room records, and results of laboratory tests. The amount of information available to clinical researchers, physicians, healthcare administrators, and policymakers continues to grow, which creates an untapped resource with the ability to drastically improve the healthcare system (Kamal, 2014; Murdoch and Detsky, 2013). While initially created for archiving patient records and supporting healthcare administrative tasks such as billing, many researchers have observed the secondary use of EHRs for clinical research purposes (Shickel et al., 2018). Healthcare providers use modern systems to diagnose patients (Graber et al., 2017), detect hidden patterns and trends, study the effects of medications (Feng et al., 2019), determine the effectiveness of treatments (Cowie et al., 2017), monitor patient improvement (Doupi, 2012), reduce medical errors (Agrawal, 2009), and ultimately improve quality of care (Ali et al., 2007; Christensen and Grimsmo, 2008; Tang and McDonald, 2006). Despite the growing interest in adopting EHRs, some studies have shown that EHR-based systems hardly improve the ability of healthcare providers to make better decisions (Heisey-Grove et al., 2014; Lau et al., 2012). One of the main reasons for this inefficacy is that these systems do not allow for human-data interaction in a manner that supports and fits the needs of healthcare providers (Himmelstein et al., 2010; Rind, 2013). Another reason is the information overload that arises when the number of data items exceeds the limit of human cognition (Halford et al., 2005). The users often misunderstand, misinterpret, ignore, or overlook vital data because of information overload. In the healthcare domain, information overload often leads to an incorrect diagnosis, wrong interpretation of patient conditions, and erroneous treatment decisions (Caban and Gotz, 2015). For example, in a survey about the efficacy of electronic health records (EHRs) in the U.S., of over 500 primary-care physicians, only 66% of physicians were somewhat satisfied with existing EHR systems. Many physicians still continue to report problems. About 40% of physicians thought that there are more challenges with existing EHR-based systems than benefits. These physicians suggested that these systems' user interfaces were not designed well. They found this more important than the incorporation of analytics capabilities that support diagnosis, management, and prevention. Of those surveyed, 72% wanted improved user interfaces in these systems; whereas 43% believed that predictive analytics would improve the efficacy of EHR-based systems (EHRIntelligence, 2018). Thus, it seems that there is a growing demand for computational systems that integrates automated analysis techniques with user interfaces that facilitate interaction with visualizations of data (i.e., interactive visualizations).

Interactive visualizations can be defined as computational systems that store and process data and use visual representations to amplify human cognition (Proctor and Vu, 2012; Sedig and Parsons, 2016). They have the potential to boost the utilization of data in healthcare by providing a means to access the EHR data at various levels of granularity and abstraction. Interactive visualizations enable users to explore the underlying data, modify the representation, and change different visual elements to achieve their goals. For the last two decades, several EHR-based visualization systems have been developed to support healthcare providers to perform various data-driven activities (Rind et al., 2013). However, there are some gaps in support for certain types of higher-level activities and tasks supported by these systems for a number of reasons. Firstly, some of the visualizations are not capable of dealing with fast-paced data generated by different healthcare organizations (Cybulski et al., 2015; Zhang et al., 2012). Secondly, some improperly designed visualizations encode too much information at once, which often overwhelm the cognitive abilities of users and limit users’ ability to make time-sensitive decisions (Pike et al., 2009; Tominski, 2015). Finally, most of these systems can only represent a limited number of attributes and relationships within the data (Aimone et al., 2013; Faisal et al., 2013; Kosara and Miksch, 2002; Lavado et al., 2018). When working with high dimensional healthcare data, it is important to analyze hidden, non-explicit, and unknown relationships among the attributes. Thus, even the complex visualization systems are often inadequate to fulfill the computational demand of EHRs because they do not incorporate analytical processes, which is essential for recognizing hidden patterns and trends.

Data analytics is the process of investigating raw data to gain both deeper and novel insights on associations within the data (Koh and Tan, 2005). Data analytics includes algorithms, techniques, and methods from different fields, such as statistics, machine learning, and data mining, to assist users in informed decision-making (Han and Kamber, 2011). There are several systems developed in recent years that employ different analytics techniques to predict patient outcomes, enable disease diagnosis and prognosis, make treatment-related decisions, and discover relationships between risk factors (Yoo et al., 2012). Although these systems are designed to analyze large amounts of data, they often fail to build trust with healthcare providers. One of the main reasons lies in their lack of transparency and interpretability. The intermediary steps, adjustment of the configuration parameters, and theoretical assumptions are kept hidden from end-users, which limits their application in healthcare settings (Yoo et al., 2012). In addition, most of the analytics systems are not capable of efficiently managing ill-defined problems because they do not consider human judgment in the decision-making process (Ola and Sedig, 2014). In order to address these issues, analytical processes need to be made accessible through visualizations.

Despite the advantages, both interactive visualization systems, with compelling interaction and representation techniques and data analytics systems, with their powerful computational capabilities, fall short in fulfilling the computational and cognitive demands of EHRs. Thus, it seems that a combined approach may be needed—that is, combining analytical processes with interactive visualizations. Visual analytics (VA) has the potential to address the needs of EHRs by combining the strengths and alleviate the limitations of both types of systems mentioned above (Ola and Sedig, 2014). VA manages the complexity of EHRs and supports visuo-analytical reasoning in such a way that the initially overwhelming scale of data becomes a treasured asset (Kamal, 2014). It enables users to analyze, synthesize, and facilitate high-level cognitive activities while at the same time get more involved in the discourse with the data (D. Keim et al., 2010a; Thomas and Cook, 2006). Although the VA approach conceivably supports different EHR-driven activities (e.g., exploration of patient history and identification of patients at risk), to date, healthcare falls behind other sectors in the development of VA systems. The design of such systems is not straightforward, which requires designers to take into consideration users' activities and tasks, human factors, and the structure of the data. A number of complicated decisions need to be made by the designers. For instance, when choosing an analysis technique, it is important to consider which algorithm to use, which samples and features to incorporate, and what granularity to seek for a specific task. Similarly, when developing visualizations, one needs to determine how to encode and organize data elements and how to support users' tasks. Consequently, integrating analysis techniques with visualizations results in a more complicated challenge. Thus, there is a lack of direction and confusion over how to design effective VA systems for EHRs (Carroll et al., 2014; Folorunso and Shawn Ogunseye, 2008; Turner et al., 2008).

The goal of this dissertation is to demonstrate how VA systems can be designed for EHRs. To begin with, we conducted a systematic literature survey to examine the design of existing EHR-based systems. Since there were not too many VA systems that are designed for EHRs, we included the EHR-based interactive visualization systems as well in the survey. We then presented a framework to analyze and evaluate EHR-data-driven tasks and activities of these systems. The framework helped us to identify gaps in the existing systems and conceptualize the data-driven activities and tasks of EHR-based VA systems. In light of this, we designed and developed two novel VA systems

(VISA_M3R3 and VALENCIA) and conducted two independent studies. The systems and studies in this dissertation were mainly focused on acute kidney injury (AKI) because they were designed to assist the clinicians, epidemiologists, and analysts at the ICES-KDT program. ICES is 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. KDT refers to the Kidney Dialysis and Transplantation program located in London, Ontario, Canada. We demonstrated the usefulness of these systems by investigating the process of analyzing the health administrative datasets housed at ICES to gain novel insights into the data and fulfill the tasks at hand. The tasks included, but are not limited to, predicting AKI, identifying AKI-associated medication, examining the synergistic effects of AKI-associated medication combinations, and identifying risk-factors for AKI.

One of the main contributions of this dissertation is the conceptualization and design of human- and activity-centered computational systems for healthcare. There are several challenges that designers might face when developing a computational system for healthcare providers. These challenges include, but are not limited to, providing busy physicians timely information in the precise format, visualizing comparative-effectiveness and casual relationships, facilitating data-driven decision-making, and characterizing and understanding similarity among information items. This dissertation describes how these challenges can be addressed using a combination of statistical methods, data mining algorithms, machine learning techniques, and information visualization. This dissertation also demonstrates how VA systems can be designed in a systematic way. It describes different components of VA in a structured manner and explains the design decisions that need to be made while developing a VA system. This dissertation then illustrates how different design choices can lead to the development of an optimized VA system for healthcare. Finally, this dissertation demonstrates how healthcare providers' abilities to interact with data mining and machine learning processes can be improved by using well-designed VA systems. Through the development of two novel VA systems, this dissertation offers the healthcare domain with evidence of the efficacy of VA for analyzing EHRs. This research has implications for other domains that require their data to be made accessible and analyzable through VA.

1.2 Structure of this dissertation

The rest of this dissertation is divided into six chapters, as follows:

In **Chapter 2**, we present a framework to identify and analyze EHR-data-driven tasks and activities in the context of interactive visualization systems—that is, all the activities, sub-activities, tasks, and sub-tasks that are and can be supported by EHR-based systems. We conducted a systematic literature survey to analyze the researches that describe the design, implementation, and/or evaluation of these systems. The survey includes an overview of their goals, a short description of their visualizations, and an analysis of how sub-activities, tasks, and sub-tasks combine and blend to accomplish their higher-level activities. Our proposed framework reveals gaps in support of some higher-level activities supported by these systems. This chapter provides background for the dissertation.

In **Chapter 3**, we describe how VA systems can be designed to utilize the prescription data stored in EHRs. To achieve this, we propose and describe VISA_M3R3, a novel VA system designed to assist healthcare providers in identifying medications and medication combinations that associate with a higher risk of AKI. By integrating multiple logistic regression models, data visualization, frequent itemset mining, and human-data interaction mechanisms, VISA_M3R3 allows users to explore complex relationships between medications, medication combinations, and AKI in such a way that would be difficult without the aid of a VA system.

In **Chapter 4**, we present a population-based retrospective cohort study to test the hypotheses generated from the VISA_M3R3 and understand the synergistic effect of AKI-inducing medication combinations. By integrating multivariable logistic regression, frequent itemset mining, and stratified analysis, this study is designed to explore complex relationships between medications and AKI. We demonstrate that our results are consistent with previous studies through an electronic literature search and a consultation with a nephrologist in this chapter.

In **Chapter 5**, we present another novel VA system, called VALENCIA, to address the challenges of high-dimensional EHRs in a systematic way. VALENCIA brings together a wide range of cluster analysis and dimension reduction techniques, integrate them seamlessly, and make them accessible to users through interactive visualizations. It offers a balanced distribution of processing load between users and the system to facilitate the performance of high-level cognitive activities. Through a case study, we demonstrate how VALENCIA can be used to analyze the healthcare administrative dataset stored at ICES. During the cluster analysis of ICES datasets using VALENCIA, we identify several risk factors that may associate with AKI by investigating the characteristics of clusters where AKI is common. This motivated us to conduct a separate study on predicting AKI, which is described in chapter 6.

In **Chapter 6**, we employ a number of machine learning techniques to identify older patients who are at risk of developing AKI within 90 days after they are discharged from the hospital or emergency department. The records of one million patients are included in this study who visited the hospital or emergency department in Southwestern Ontario between 2014 and 2016. We developed sixteen prediction models based on combinations of four machine learning techniques and four ensemble-based methods along with a cost-sensitive logistic regression model. These models are evaluated through 10-fold cross-validation and compared based on the AUROC metric. We also validate features that are most relevant in predicting AKI with a healthcare expert through a participatory design process to improve the performance and reliability of the models.

In **Chapter 7**, we outline the conclusions drawn from the research presented in the preceding chapters, explain the contributions of this work to the broader scientific community, and discuss some areas of future research.

It is important to note that the chapters of this dissertation are self-sufficient and can be read individually or sequentially. Chapters 2,3, 4, and 5 have been published; Chapters 6 has been accepted for publication. This dissertation is written in an integrated article format, so Chapters 2 through 6 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. We changed the format to match the general format of the dissertation. Figure, Table and Section numbers specified herein are relative to the chapter number. For example, “Table 1” corresponds to Table 2-1; “Figure 1” corresponds to Figure 2-1; and “Section 1.1” corresponds to Section 2.1.1. Moreover, when the term “paper”, “research”, or "work" is used, it refers to this specific 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 and Grimsmo, 2008; Tang and 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 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 and Jacko, 2007; Sedig and 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 and 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 and 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 and Parsons, 2016, 2013). 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.

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 IVTs, 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. (Wang et al., 2008); Wongsuphasawat et al. (Wongsuphasawat et al., 2011); Wongsuphasawat and Gotz (Wongsuphasawat and Gotz, 2012); Malik et al. (Malik et al., 2014); Fails (Fails et al., 2006); Klimov et al. (Klimov et al., 2010); Wongsuphasawat (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 et al., 2011a); Plaisant et al. (Plaisant et al., 1998); Faiola and Newlon (Faiola and 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 et al., 2011b); and Ordonez et al. (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; Kumar et al., 2007; Låg et al., 2014), ***predicting*** (Amarasingham et al., 2014; Cohen et al., 2014; Kankanhalli et al., 2016; Raghupathi and Raghupathi, 2014; Allan F. Simpao et al., 2014; Wang et al., 2018), and ***monitoring*** (Anderson et al., 2015; Hauskrecht et al., 2013; Kho et al., 2007; Li and 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).

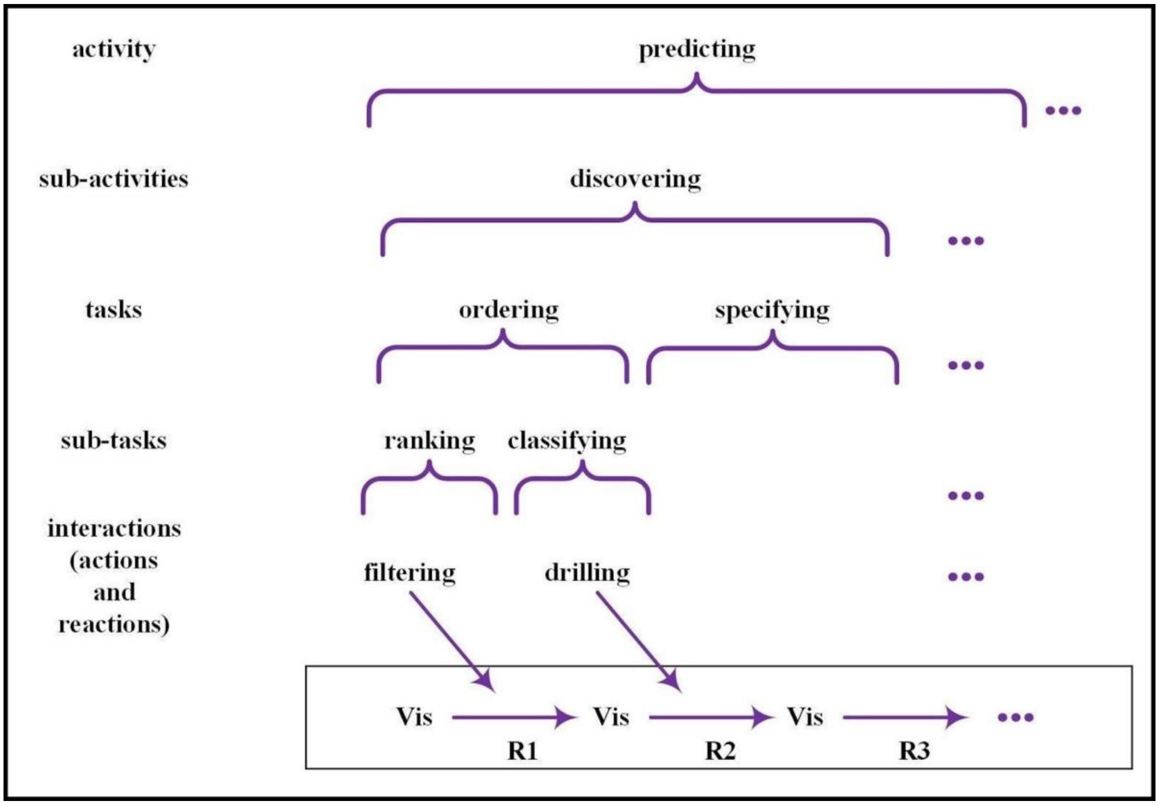


Figure 2-1: Relationships among activities, tasks, and interactions. Top-down view: activity is made up of sub-activities, tasks, sub-tasks, and interactions. Bottom-up view: activity emerges over time, through performance of tasks and interactions. Visualizations are depicted as Vis and reactions as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

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).

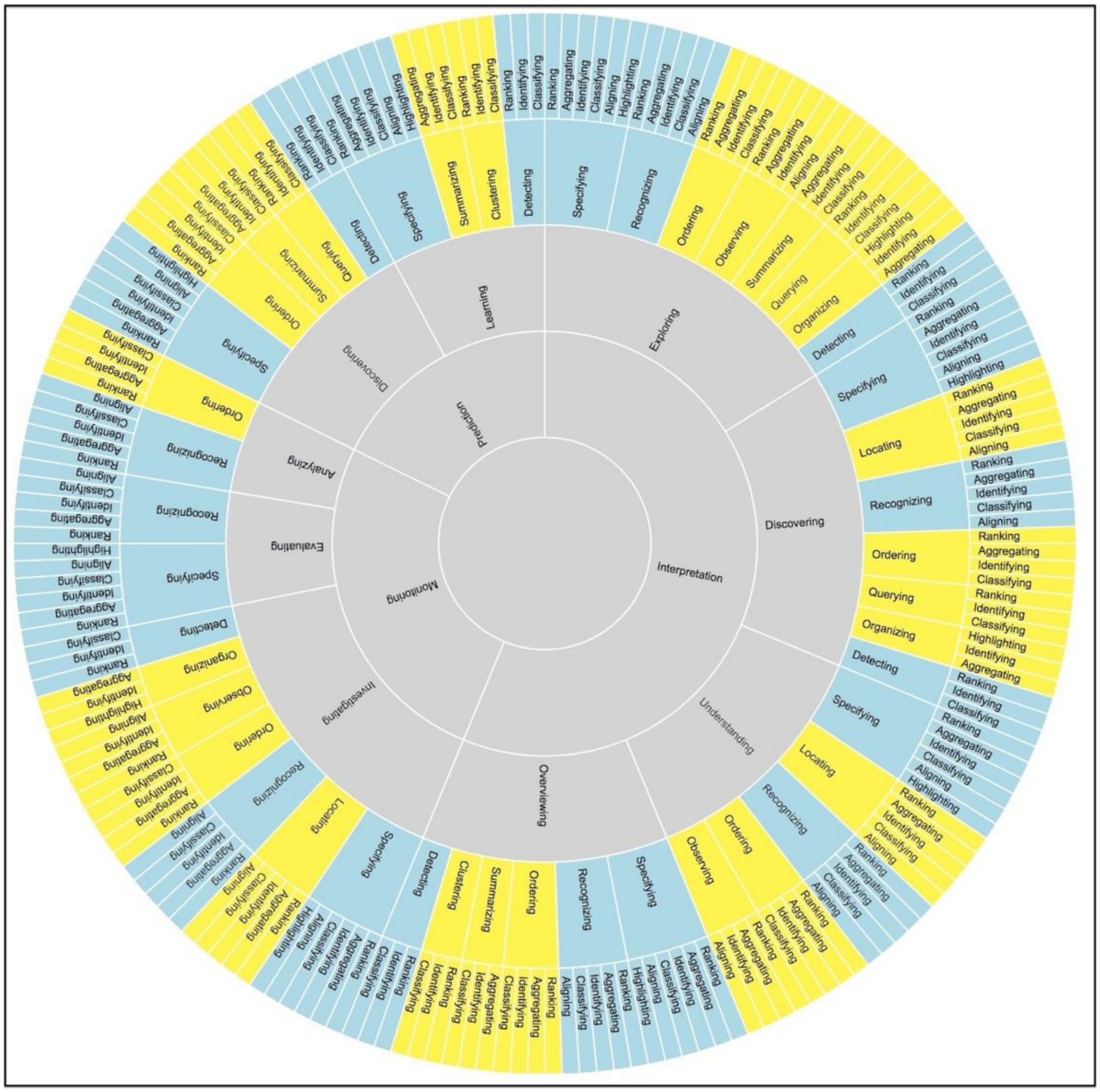


Figure 2-2: Overview of the proposed activity and task analysis framework. The visual tasks are represented as blue and interactive tasks are represented as yellow.

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*.

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

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 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 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.

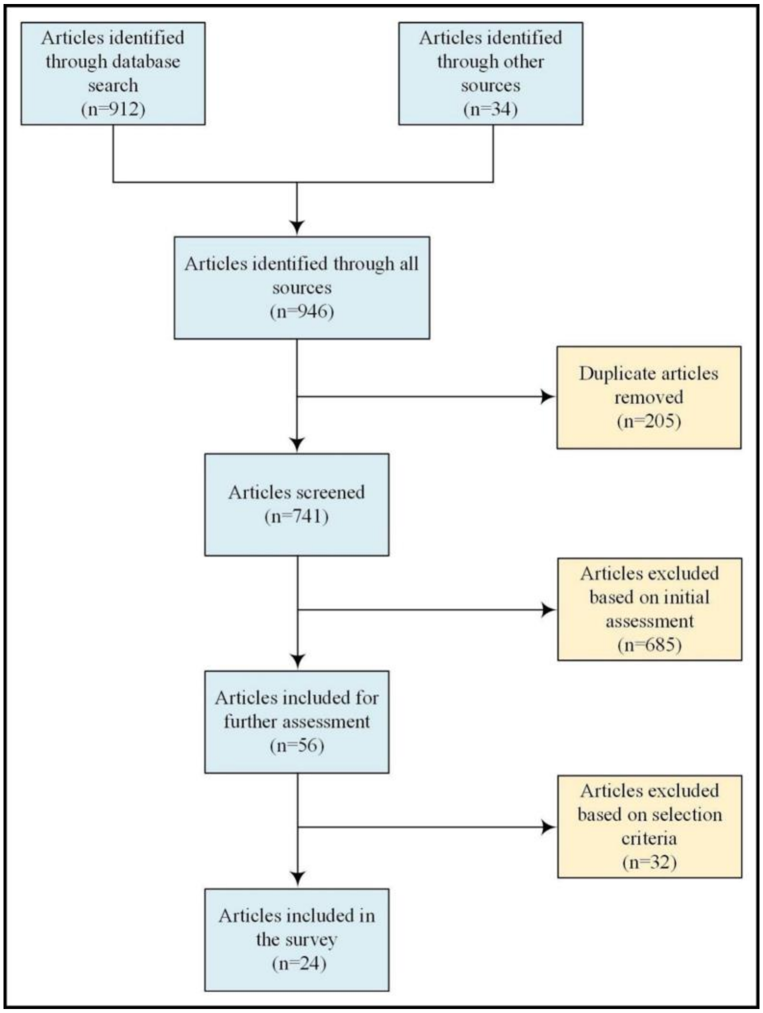

Articles identified through database search (n=912)
Articles identified through other sources (n=34)
Articles identified through all sources (n=946)
Duplicate articles removed (n=205)
Articles screened (n=741)
Articles excluded based on initial assessment (n=685)
Articles included for further assessment (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 aggregate form and simultaneously. 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 (Wang et al., 2009, 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.*

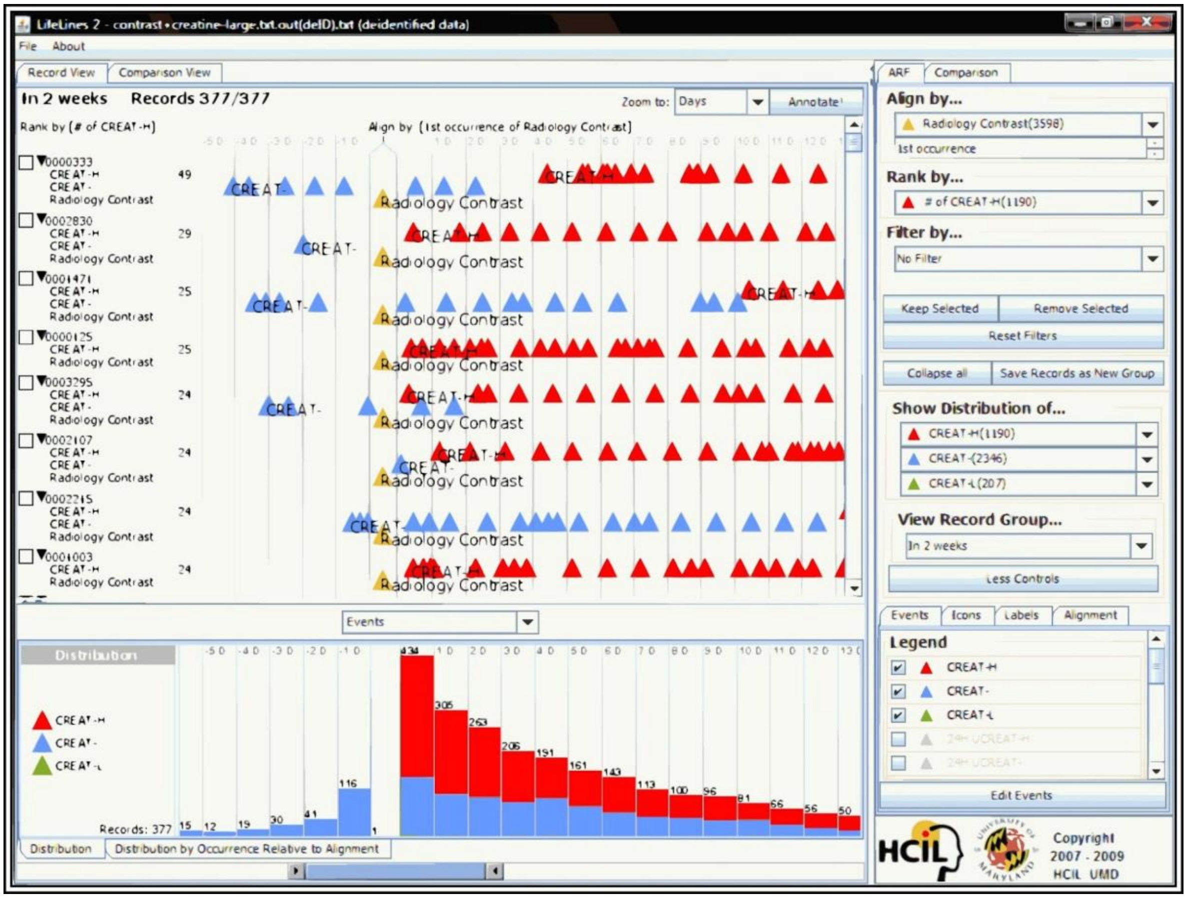


Figure 2-4: Lifelines2: Interactive visualization tool for temporal categorical data. Source: Image courtesy of the University of Maryland Human–Computer Interaction Lab, http://hcil.umd.edu.

2.4.1.2 Lifeflow

Lifeflow (Guerra Gómez et al., 2011; 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.

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.

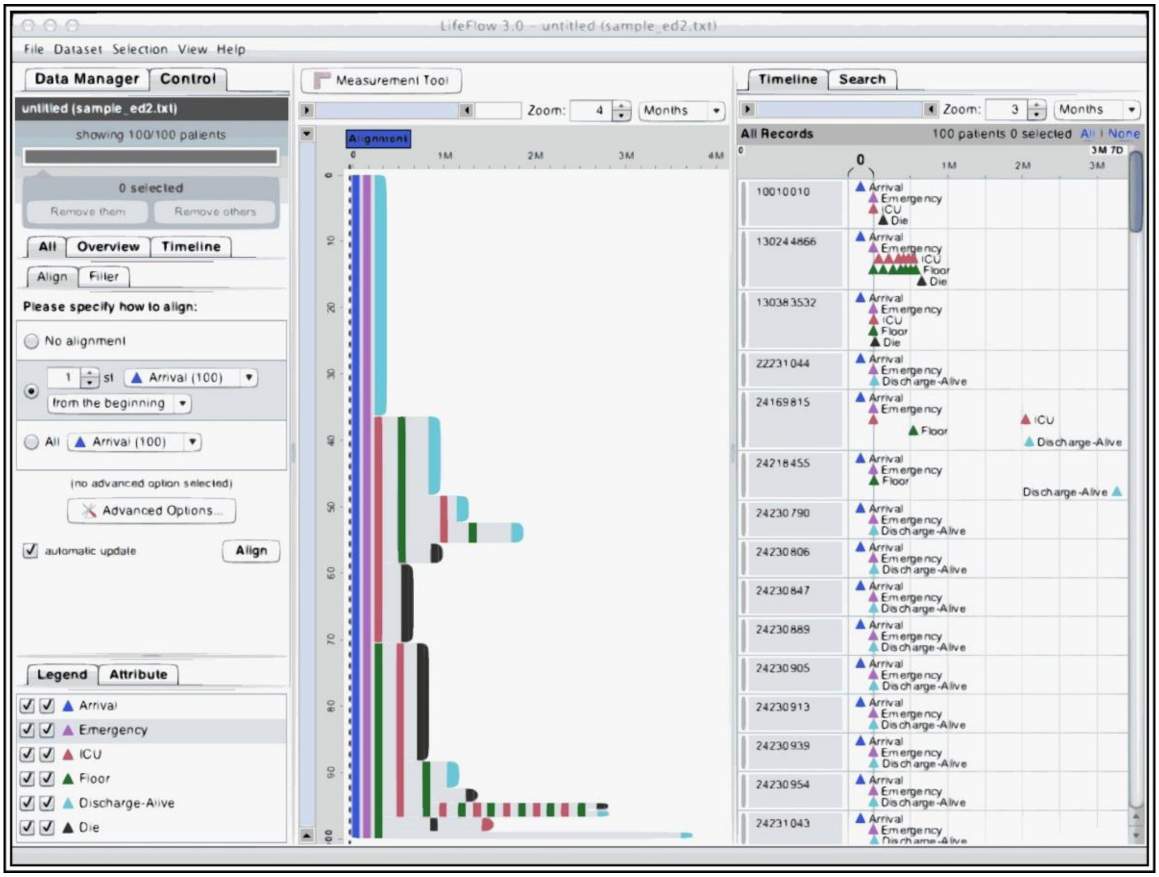


Figure 2-5: Lifeflow: Interactive visualization tool that provides an overview of event sequences. Source: Image courtesy of the University of Maryland Human–Computer Interaction Lab, http://hcil.umd.edu.

2.4.1.3 Eventflow

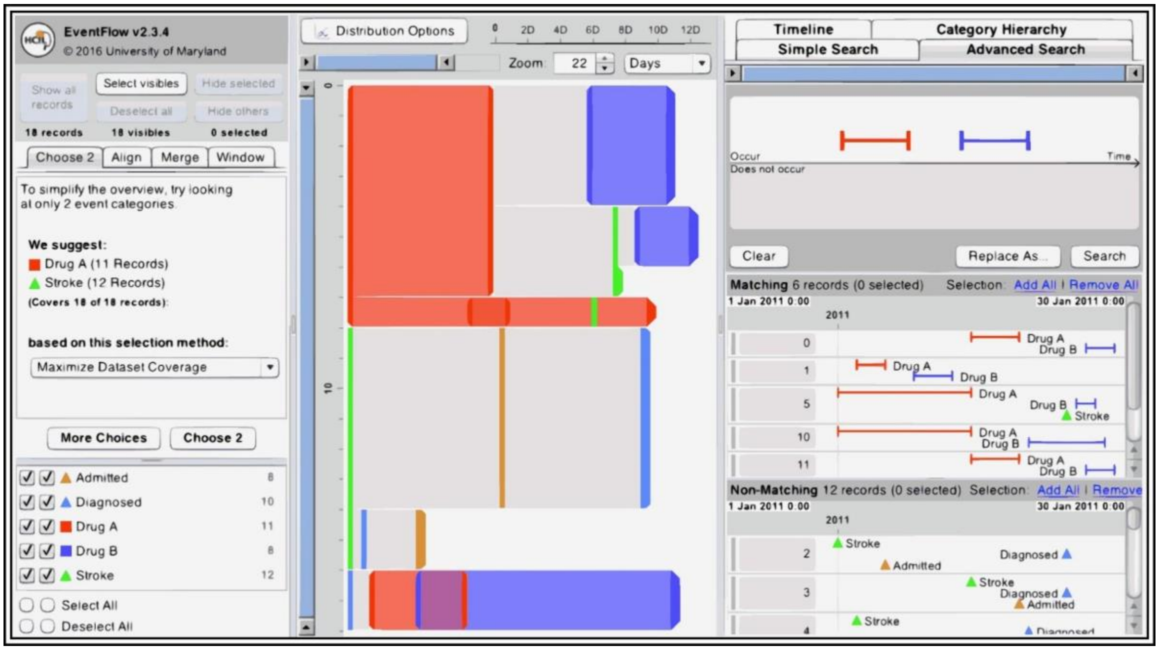
Eventflow (Monroe et al., 2013) provides users with the ability to query, explore, and visualize interval data interactively. It allows pattern recognition by visualizing events in both a timeline that displays all individual records and an aggregated overview that shows common and rare patterns. As 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.

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*.



EventFlow v2.3.4
© 2016 University of Maryland

18 records 18 visibles 0 selected

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

☑	☑	▲ Admitted	8
☑	☑	▲ Diagnosed	10
☑	☑	🟥 Drug A	11
☑	☑	🟦 Drug B	8
☑	☑	▲ Stroke	12

○ ○ Select All
○ ○ Deselect All

Zoom: 22 Days

Matching 6 records (0 selected) Selection: Add All | Remove All

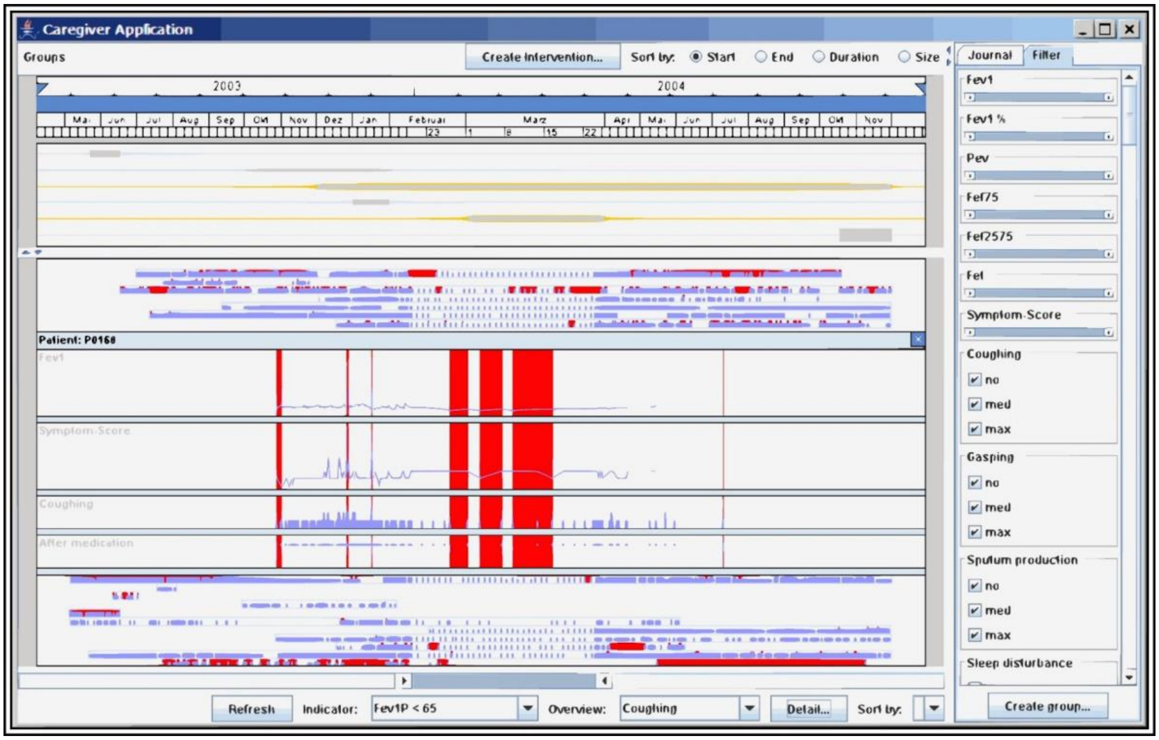
Non-Matching 12 records (0 selected) Selection: Add All | Remove

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.

2.4.1.5 CoCo

CoCo (Malik et al., 2015, 2014) 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.



Sort by: ● Start ○ End ○ Duration ○ Size

Coughing
☑ no
☑ med
☑ max

Gasping
☑ no
☑ med
☑ max

Sputum production
☑ no
☑ med
☑ max

Sleep disturbance

Indicator: Fev1P < 65

Overview: Coughing

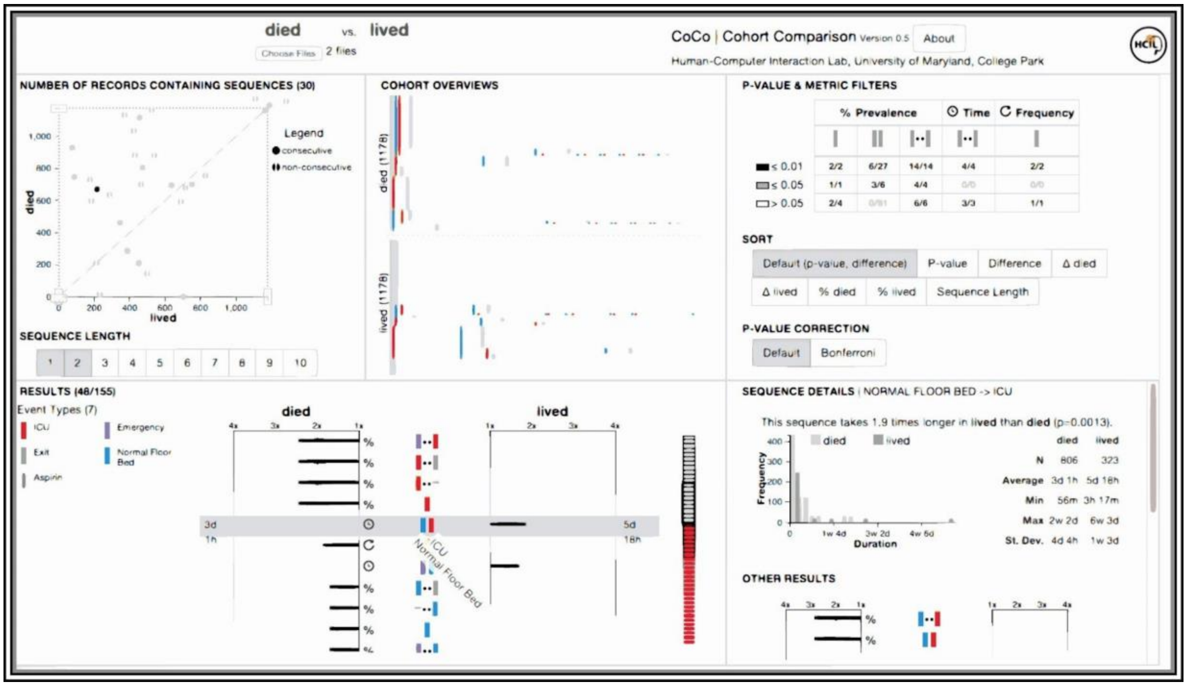
Sort by:

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

2.4.1.6 Similan

Similan (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.



	% Prevalence			⊙ Time	C 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	0/81	6/6	3/3	1/1

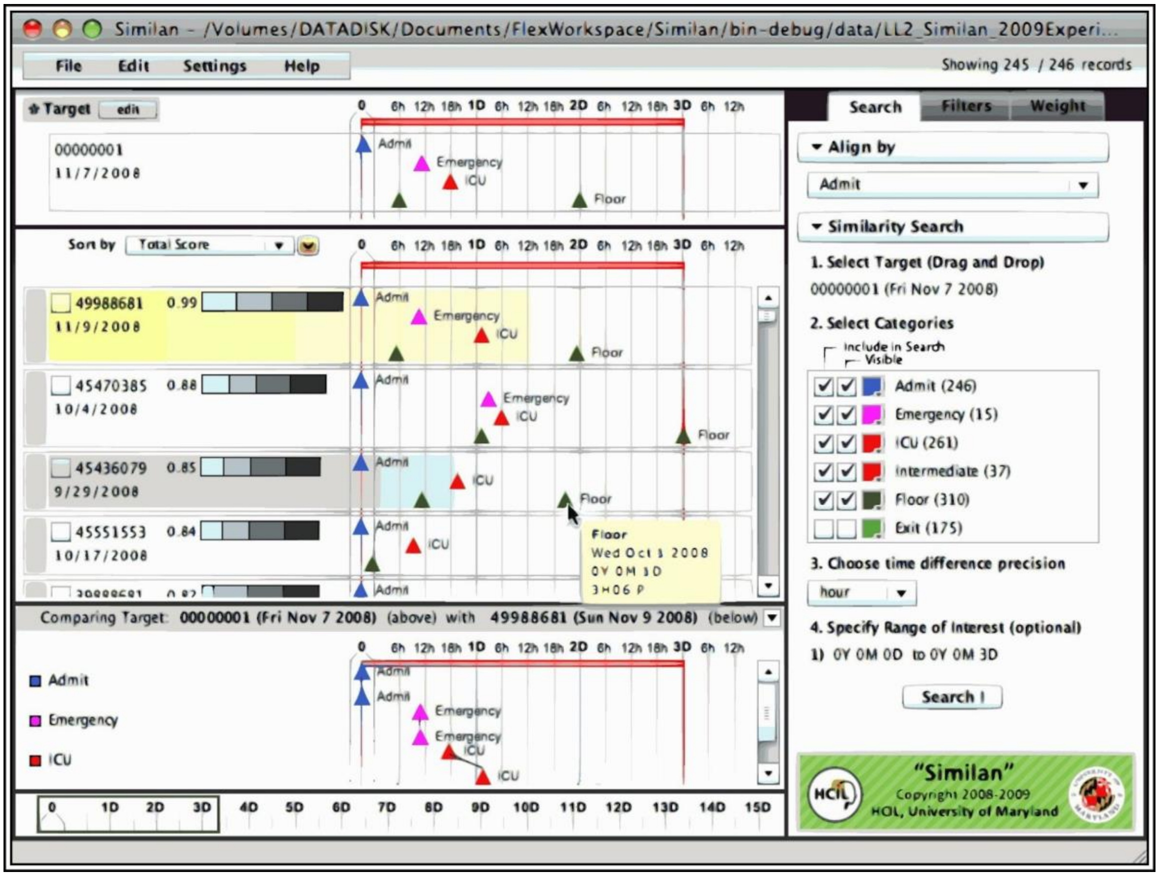
This sequence takes 1.9 times longer in **lived** than **died** (p=0.0013).

	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.

2.4.1.7 Outflow

Outflow (Wongsuphasawat and Gotz, 2012, 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 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.



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

Showing 245 / 246 records

Target: 00000001 11/7/2008

Sort by: Total 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

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)

Align by: Admit

Similarity Search

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

2. Select Categories

Include in Search	Visible	
☑	☑	🟦 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

Search !

"Similan"
Copyright 2008-2009
HCIL, University of Maryland

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.

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.

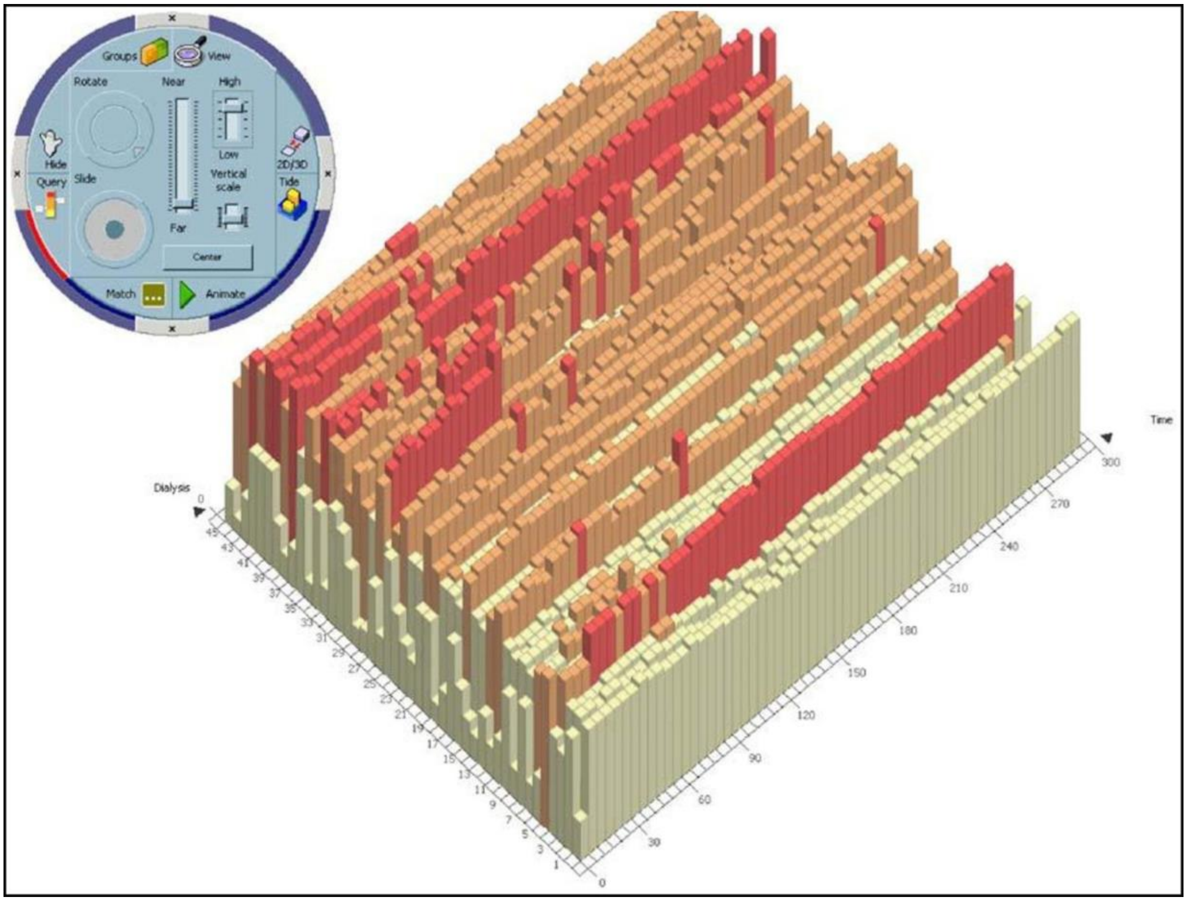


Figure 2-10: IPBC: 3D visualization tool for analysis of numerical data from multiple hemodialysis sessions. Source: reprinted from Journal of Visual Languages & Computing, 14, Chittaro L, Combi C, Trapasso G, Data mining on temporal data: a visual approach and its clinical application to hemodialysis, 591-620, Copyright (2003), with permission from Elsevier.

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.

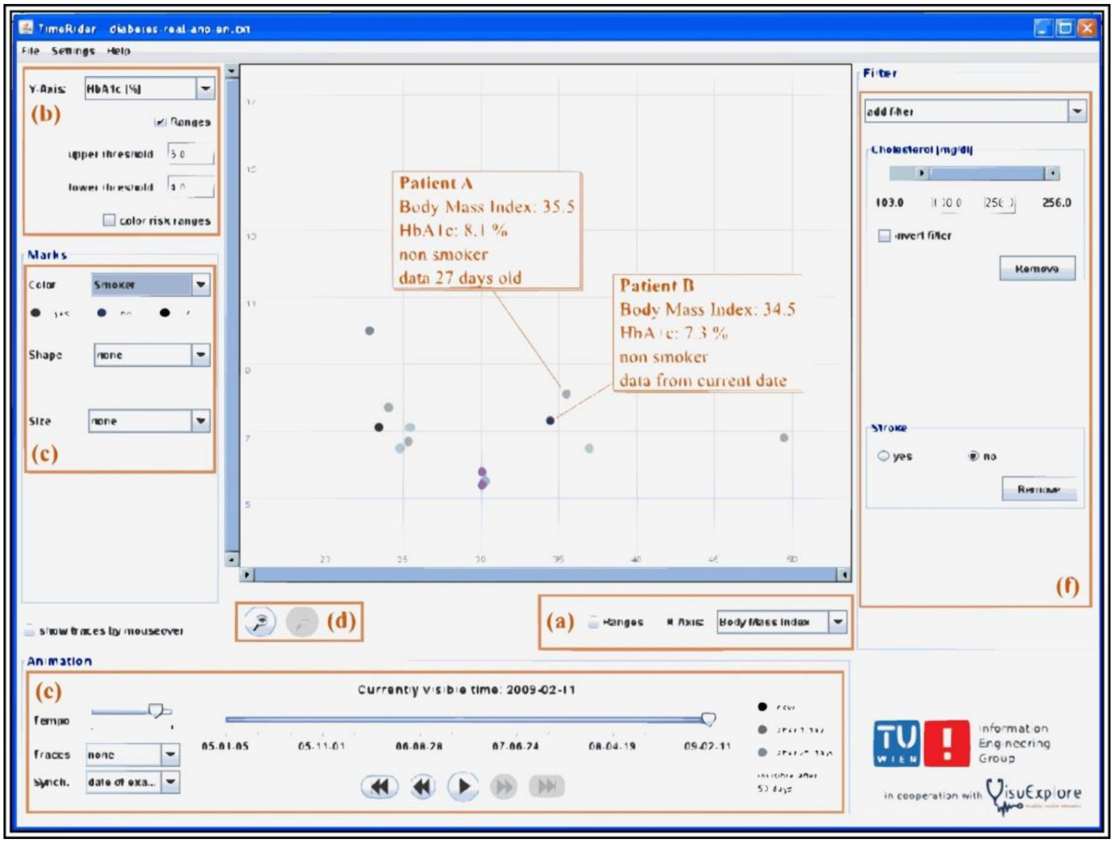
2.4.1.10 PatternFinder

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 et al., 2011a) 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.



TimeRider diabetes-real-anon.csv

Y-Axis: HbA1c [%]

☑ Ranges

upper threshold 5.0

lower threshold 1.0

□ color risk ranges

Color: Smoker

Shape: none

Size: none

□ show traces by mouseover

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]

103.0 100.0 256.0 256.0

□ invert filter

Stroke: ○ yes ● no

Currently visible time: 2009-02-11

Traces: none

Synch.: date of exa...

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., 2010) 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.

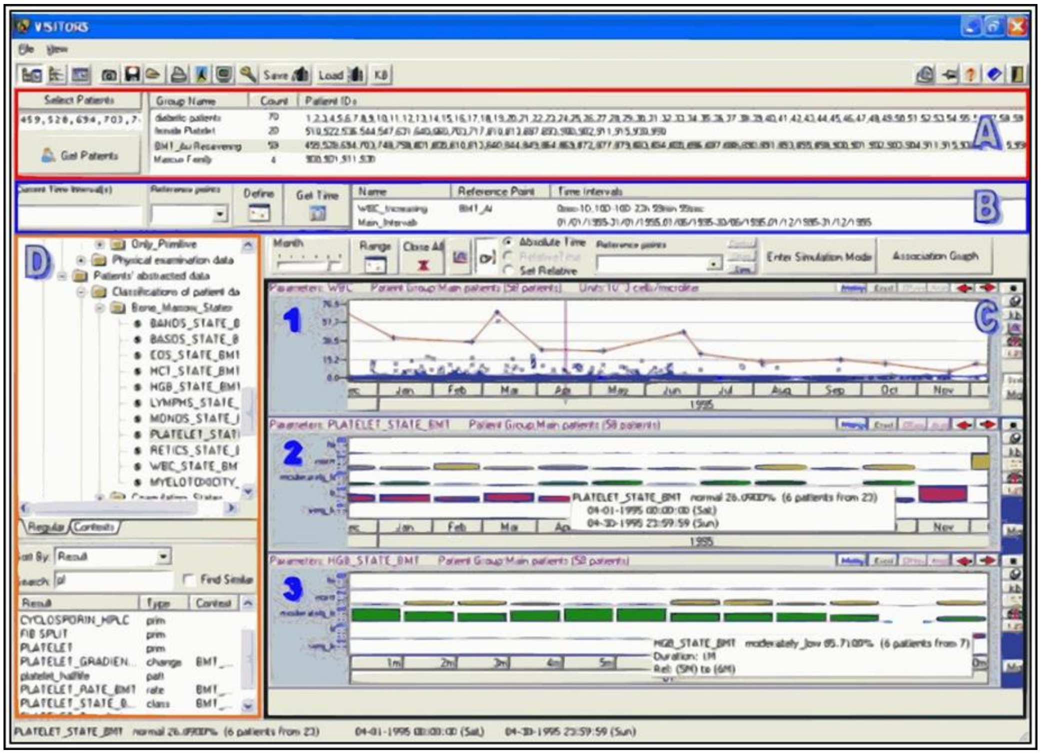


Figure 2-12: VISITORS: Interactive visualization tool for the exploration of multiple patient records. (A) displays lists of patients. (B) displays a list of time intervals. (C) displays the data for a group of 58 patients over the current time interval. Panel 1 shows the white blood cell raw counts 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.

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 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

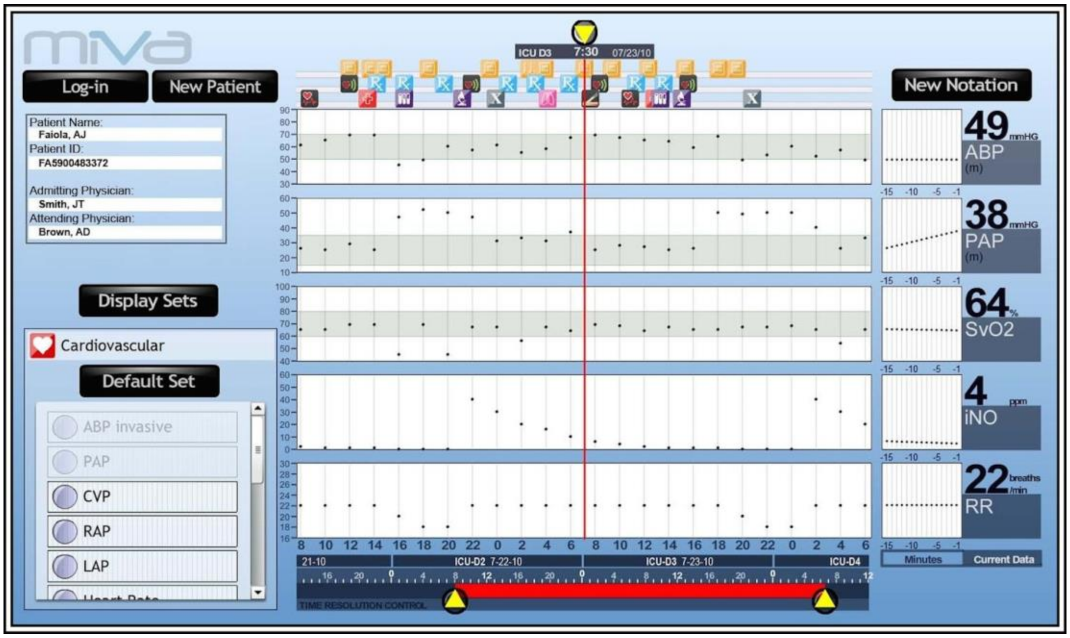
Midgaard (Bade et al., 2004) allows for exploration of the intensive care units’ data at different levels of abstraction from overview to details. It uses visualizations to display numerical variables of treatment plans. It incorporates a complex semantic zoom method for numerical variables by calculating 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 and 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.

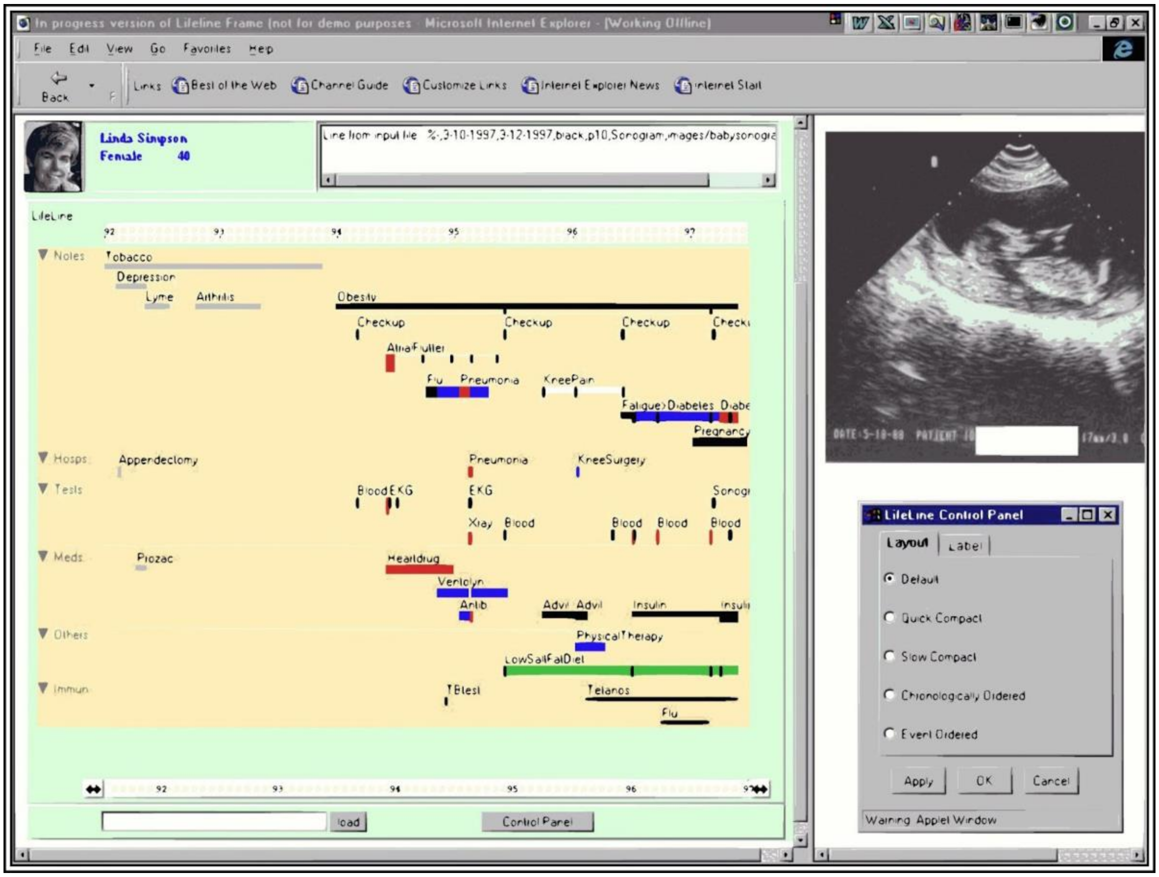


Figure 2-14: Lifelines: interactive visualization tool that displays patient's medical histories on a timeline. Source: image courtesy of the University of Maryland Human–Computer Interaction Lab, http://hcil.umd.edu.

2.4.2.5 VisuExplore

VisuExplore (Pohl et al., 2011; Rind et al., 2011b) 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.

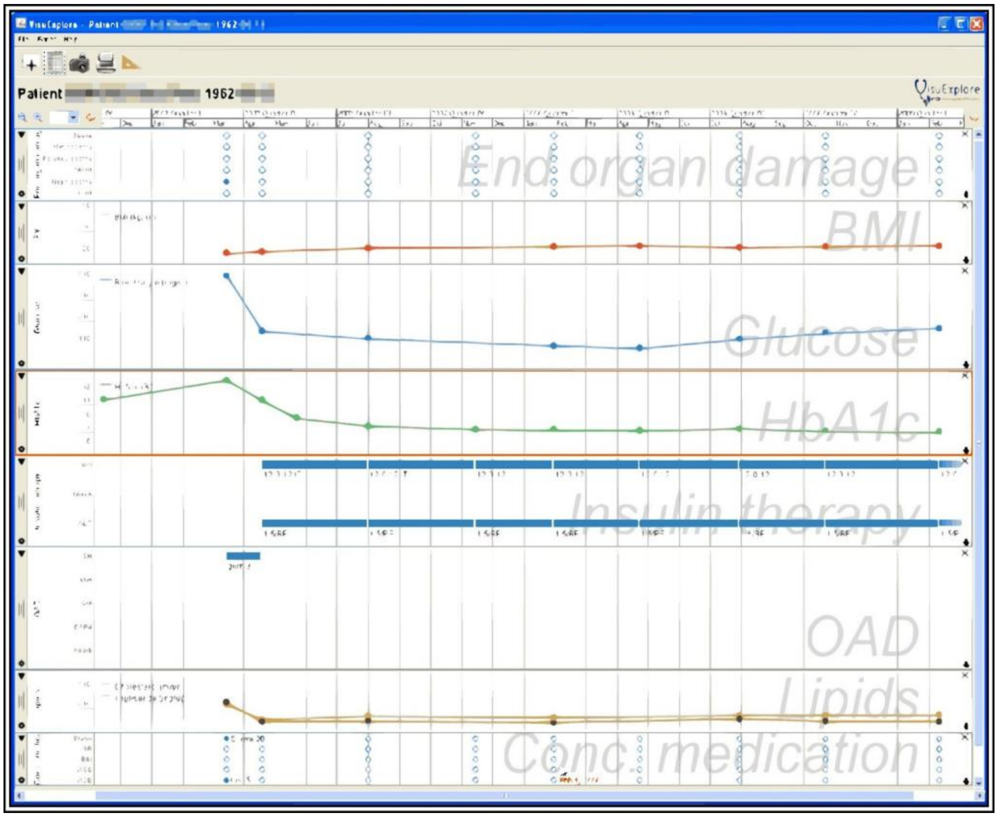


Figure 2-15: VisuExplore: interactive visualization tool that displays patient data in various views on a timeline. Source: reprinted by permission from Springer Nature: Springer, Human–Computer Interaction, Patient Development at a Glance: An Evaluation of a Medical Data Visualization, Pohl M, Wiltner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

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 *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 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 course 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 system (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 Multiple regression analysis and frequent itemset mining of electronic medical records: A visual analytics approach using VISA_M3R3

This chapter has been published as S.S. Abdullah, N. Rostamzadeh, K. Sedig, A.X. Garg, and E. McArthur, “Multiple regression analysis and frequent itemset mining of electronic medical records: A visual analytics approach using VISA_M3R3” in the Data Journal, *5*(2), 33; March 2020. We changed the format to match the general format of the dissertation. Figure, Table and Section numbers specified herein are relative to the chapter number. For example, “Table 1” corresponds to Table 3-1; “Figure 1” corresponds to Figure 3-1; and “Section 1.1” corresponds to Section 3.1.1. Moreover, when the term “paper”, “research”, or "work" is used, it refers to this specific chapter.

3.1 Introduction

As part of modernizing their operations, healthcare and medical organizations are adopting electronic medical records (EMRs) and deploying new information technology systems that generate, collect, digitize, and analyze their data (Caban and Gotz, 2015). With the development of EMRs and the extensive use of computerized provider order entry tools, patients’ medication profile data is now accessible and processable for secondary reuses (Abramson et al., 2011; Delamarre et al., 2015). The amount of prescription data available to clinical researchers, pharmaceutical scientists, and clinician-scientists continues to grow, creating an analyzable resource for generating insights that can help improve the healthcare system (Kamal, 2014; Murdoch and Detsky, 2013). Healthcare providers use modern EMR-based systems to identify adverse drug events (Hannan, 1999; Honigman et al., 2001), study medication-medication interactions (Rinner et al., 2015), investigate medication effects on particular medical conditions (Gruchalla, 2000; Tandon et al., 2015), and ultimately prevent medication errors (Agrawal, 2009; Gildon et al., 2019; Singer and Duarte Fernandez, 2015).

A common problem in clinical medicine which may lead to development of acute kidney injury (AKI) is medication-induced nephrotoxicity (Assadi and Ghane Shahrbaf, 2015; Fusco et al., 2016; Khan et al., 2017). AKI can be defined as a sudden loss of kidney function over a short period of time (Porter et al., 2014; Nicholas M. Selby et al., 2012). The rate of medication-induced AKI can be as high as 60 percent (Gandhi et al., 2000; Kaufman et al., 1991; Nash et al., 2002; Schetz et al., 2005). Many prior studies have assessed the impact of individual nephrotoxic medications on AKI (Alexander et al., 2017; Moffett and Goldstei, 2011; Ryan M. Rivosecchi et al., 2016). The combination of multiple medications can further increase the risk of AKI through synergistic or accumulative nephrotoxicity (Schetz et al., 2005). For each additional nephrotoxic medication, the chance of developing AKI may increase by 53 percent (Cartin-Ceba et al., 2012). Rivosecchi et al., through an exhaustive literature search, further emphasize the need for a comprehensive understanding of how medication combinations alter the risk of AKI (Ryan M. Rivosecchi et al., 2016). According to a Center for Disease Control report, as of 2017, there were more than 5,000 medications in the market and 1,000 adverse medication effects known in the literature. So, for drug-drug interactions there may be 125 billion possible adverse medication effects between all possible pairs of medications (Collins, 2018; Zitnik et al., 2018). An individual clinical study is often required to test the nephrotoxicity of each medication or medication combination. Therefore, it is impossible to comprehensively assess medication-induced AKI through this number of clinical studies.

Data analytics can offer a solution to this problem by employing algorithms, methods, and techniques from different fields, such as data mining, statistics, and machine learning (Han and Kamber, 2011). Data analytics is the investigation of raw data to gain both novel and deeper insights on associations within the data (Koh and Tan, 2005). There are several tools designed and developed in recent years that employ advanced machine learning techniques to improve drug-safety science, predict adverse drug reactions, and identify drug-drug interactions (Basile et al., 2019; Dey et al., 2018; Lysenko et al., 2018; Munsaka, 2017; Schmider et al., 2019; Vamathevan et al., 2019). While most clinical machine learning tools are designed to incorporate large amounts of data, they are not capable of efficiently managing ill-defined problems that need human judgment. The main challenge of using machine learning techniques lies with their lack of interpretability and transparency, hence limiting their application in healthcare settings (Vamathevan et al., 2019).

Interactive visualizations have the potential to address this challenge by providing a means to access the data at various levels of granularity and abstraction (Rind et al., 2011b). They can be defined as computational systems that store and process data and use visual representations to amplify human cognition (Sedig and Parsons, 2016; Wilson, 2014). Interactive visualizations allow users to explore the underlying data, modify representations, and change different visual elements to achieve their goals. In recent years, several EMR-based systems have been developed to interactively visualize patient prescription history (Ozturk et al., 2014), potential adverse medication events (Duke et al., 2010), and prescription behaviors (Van der Corput et al., 2014). Most of these systems only represent a limited number of attributes and relationships within the data (Faisal et al., 2013; Kosara and Miksch, 2002; Lavado et al., 2018; A Rind et al., 2011). When working with high-dimensional EMR data, it can be useful to analyze hidden, non-explicit, and unknown relationships among all the data attributes (Lee and Yoon, 2017; Perer et al., 2015). One of the main issues with traditional data visualization systems is that they do not incorporate analytical processes, which are essential for recognizing hidden patterns and trends in the data. Therefore, interactive data visualization systems, alone and without data analytics components, fall short of satisfying the computational needs and requirements of users.

While beneficial, both data analytics systems, with their advanced computational capabilities and interactive visualization systems, with powerful interaction and representation mechanisms, when used individually, prove inadequate in certain situations. The emergence of a type of computational system known as visual analytics (VA) has the potential to reduce the complexity of EMR data by combining the strengths and alleviate the limitations of both aforementioned systems (Parsons et al., 2015; Saffer et al., 2004; A. F. Simpao et al., 2014). VA can improve the capabilities of users to perform complex data-driven tasks by analyzing EMRs in such a way that would be difficult or sometimes even impossible to do otherwise. Even though VA is suitable for different healthcare activities (e.g., prediction of diseases, exploration of patient history, and identification of adverse medication events), to date, healthcare environments lag behind other sectors in the development of such systems (Amarasingham et al., 2014; Caban and Gotz, 2015; Feng et al., 2019).

The purpose of this study is to demonstrate how VA systems can be designed in a systematic way: 1) to examine the association between medications and AKI, in particular, and 2) to support other clinical investigations involving EMRs, in general. To this end, we present a novel system that we have developed, called VISA_M3R3—VISual Analytics, VISA for Multiple Regression analyses and fRequent itemset Mining of electronic Medical Records, M3R3. VISA_M3R3 is intended to assist clinicians and healthcare researchers at the ICES-KDT (Kidney Dialysis and Transplantation), located in London, Ontario, Canada. We demonstrate VISA_M3R3 by investigating the process of identifying medications and medication combinations that associate with a higher risk of AKI using ICES health administrative data. To our knowledge, no prior VA system has been designed to examine how different medications affect kidney function and increase the risk of developing AKI. While few VA systems have been developed for other areas in healthcare (Basole et al., 2015; Bernard et al., 2015; Gotz et al., 2012; Huang et al., 2015; Klimov et al., 2015; Mittelstädt et al., 2014; Ninkov and Sedig, 2019; Perer et al., 2015; A. F. Simpao et al., 2014), VISA_M3R3 is novel in that it integrates multiple regression models (i.e., multivariable logistic regression), frequent itemset mining (i.e., Eclat algorithm), data visualization, and human-data interaction mechanisms in an integrated fashion. As such, the design concept of VISA_M3R3 can be generalized for the development of other EMR-based VA systems that apply multivariable regression and frequent itemset mining to gain novel and deep insights into massive clinical data that exist for different health conditions (e.g., diabetes and heart failure, to name a few).

The rest of this paper is organized as follows. Section 2 provides an overview of the terminological and conceptual background to understand the design of VISA_M3R3. Section 3 describes the methodology employed for the design of the proposed VA system. Section 4 presents VISA_M3R3 by providing a description of its structure, components, and results. Finally, Section 5 discusses the usefulness and limitations of the proposed system and some future areas of application.

3.2 Background

This section presents the necessary background concepts and terminology for understanding the design of VISA_M3R3. VA systems fuse the strengths of automated analysis and interactive visualizations to allow users to explore data interactively, identify patterns, apply filters, and manipulate data to achieve their goals. This process is more complicated than an automated internal analysis coupled with an external visualization to show the results. It is both data-driven and user-driven and requires re-computation when users manipulate data through visual representations. VA not only relies on computational techniques and analytics but also supports human-in-the-loop mechanisms that allow users to employ human judgment to reach evidence-based conclusions. To understand the concepts of VA, we discuss the spatial structure and different modules of VA systems in this section.

3.2.1 Spatial Structure of Visual Analytics

To conceptualize the spatial structure of VA, Sedig et al. (Sedig et al., 2012; Sedig and Parsons, 2016) proposed its processing load to be divided into at least 5 spaces: information space, computing space, representation space, interaction space, and mental space. The information space represents bodies of data that come from different sources. Data may come from abstract spaces (e.g., treatment plans) or concrete spaces (e.g., prescriptions). Data is then processed in the computing space, which may include (1) pre-processing techniques such as data cleaning, filtering, fusion, integration, and normalization and (2) data processing and transformation techniques such as data mining, mathematical procedures, and statistical methods. Since the underlying processing is carried out in the computing space, users of the VA system ideally do not need to be concerned with any computational work of this space. Resulting data items are then encoded into perceptible visual forms in the representation space. In order to achieve their goals through a visually perceptible interface, users can choose actions from a set of available options (i.e., the interaction space) to act upon existing visualizations in the representation space. Finally, the mental space refers to users perceiving and processing changes in the interface through carrying out mental operations such as apprehension, induction, deduction, judgment, and memory encoding.

In healthcare settings, it is important for the designer to find a balanced distribution of the processing load among the above five spaces. VA systems can offer such a balanced distribution of processing load through a proper integration of advanced analytics techniques (i.e., data mining, statistics, and machine learning) with visual representations to facilitate high-level cognitive activities and tasks while at the same time allowing users to get more involved in interactive conversation with the data through its manipulation, analysis, and synthesis (D. Keim et al., 2010b; Ola and Sedig, 2018; Thomas and Cook, 2006).

3.2.2 Modules of Visual Analytics Systems

The information processing load in a VA system is distributed between the user and the main components of the VA system—namely, the analytics and the interactive visualization modules (Cui, 2019; Jeong et al., 2015; D. Keim et al., 2010b; Ola and Sedig, 2014; Parsons and Sedig, 2014; Sedig and Parsons, 2013). The data analytics module encompasses the computing space and deals with the analysis of data from the information space. The interactive visualization module encompasses representation and interaction spaces.

3.2.2.1 Data Analytics Module

Human cognition has limitations when engaged in data-intensive mental tasks, especially when the data is large and complex (Green and Maciejewski, 2013; Ola and Sedig, 2014).

The analytics module of the VA system supports user cognition by carrying out most of the computational load. It provides users with the ability to make time-critical decisions by placing the majority of the processing load in the computing space. In a VA system, data analytics should not be solely controlled by the system. Instead, users should be involved in controlling the parameters, settings, and intermediary steps of the processing stage. The primary responsibility of the analytics module is to store, prepare, analyze, transform, and perform computerized analysis of the raw data. In the context of VA, the analytics process can be divided into three main stages: data pre-processing, data transformation, and data analysis (Ola and Sedig, 2014).

The raw data from the information space gets processed in the pre-processing stage. Data often contains errors, exceptions, noise, and/or uncertainty. There are several possible reasons for having inaccurate data in EMRs. For instance, problems might arise from confusing data collection manual, faulty instruments, or incorrect data entry. The data analytics module might derive incorrect patterns if the data is noisy or erroneous. Therefore, it is very important to pre-process raw EMR data retrieved from a variety of sources. Data pre-processing includes cleaning, integration, and reduction (Han et al., 2011).

The pre-processed data is then transformed into forms appropriate for data analytics algorithms. The quality of information, knowledge, and insight extracted from a dataset can be improved by its transformation (Kusiak, 2001). Strategies for data transformation may include smoothing, attribute construction (i.e., feature generation), aggregation, normalization, and discretization (Han and Kamber, 2011).

Finally, data analysis is the stage to uncover previously undetected relationships among data items and extract the implicit, previously unknown, and possibly useful information from data (Agrawal et al., 1993; Sahu et al., 2008). The data analysis process includes, but is not limited to, frequent itemset mining, regression, classification, and clustering. Usually, these techniques allow analysis of limited types of variables and do not support heterogeneous data (D. Keim et al., 2010b). VA systems overcome this limitation by incorporating interactive visualizations and human reasoning in the decision-making loop.

3.2.2.2 Interactive Visualization Module

Interactive visualization is an integral part of VA for organizing data items in the information space and mapping them to visual structures. Interactive visual representations provide users with the ability to change and modify the displayed data and to guide the analysis process. This, in turn, will set off a chain of internal reactions that lead to the execution of additional data analysis processes. Interactive visualizations can potentially bridge the gap between the internal mental representation of the user and the external representations of the system by allowing the information processing load to be distributed between the user and the system.

Design of visualizations is straightforward when dealing with simple tasks. As tasks require completion of one or more subtasks, they become more complex. As tasks become more complex, design becomes less apparent, particularly when dealing with massive amounts of heterogeneous data (Heer and Kandel, 2012; Sedig and Parsons, 2013). To support complex, EMR-driven tasks, visualizations require some initial analysis (D. Keim et al., 2010b). For instance, the task of identifying high-risk medications for a certain medical condition includes sub-tasks such as finding association between the medical condition and medications (through data analysis), observing their relationships (through visual representations), and filtering medications that are associated with the medical condition (through analysis and visualization). Furthermore, because external structures of data affect how users perform tasks, another challenge involves determining how to organize a large number of data items in the visual representations. To support the performance of complex tasks, VA combines advanced, behind-the-scene analytics techniques with interactive external visualizations that organize data items (Kehrer and Hauser, 2013; Keim et al., 2008).

3.2.3 Visual Analytics and Analytical Reasoning

User-triggered actions, consequent reactions, and discourse with information are essential in a VA system whose function is to facilitate users' analytical reasoning activities—activities that refer to both rational and logical analysis of data as well as evaluation of results. Such activities also involve analogical, deductive, and inductive reasoning to reach conclusions (Sedig and Parsons, 2013), and emerge from a series of lower-level tasks (e.g., developing hypotheses or identifying relationships among data elements) (Heuer, 1999; Thomas and Cook, 2006). In order to reach a conclusion, some of these lower-level tasks take place in an iterative and non-linear manner depending on cognitive needs and overall goals of the user (Sedig and Parsons, 2013). Generally speaking, analytical reasoning can be viewed as transforming given data into information, knowledge, and insight (Gilhooly, 2004; Sedig and Parsons, 2013). This derived knowledge and insight serves as a foundation for other cognitive activities such as decision-making or problem-solving (Han et al., 2011; Leighton, 2004).

EMRs contain large bodies of complex data, and, oftentimes, EMR-driven tasks are ill-defined. Thus, users have to rely on their experience, knowledge, and judgement to perform complex activities (i.e., decision-making and problem-solving) in a healthcare setting (Varga and Varga, 2016). Human-in-the-loop mechanisms involving interaction with the visual and analytical modules of VA systems can thus help healthcare activities (Green and Maciejewski, 2013).

3.3 Materials and Methods

This section describes the methodology we have employed to design the proposed VA system, namely VISA_M3R3. For our EMR-based data, we use Ontario’s healthcare databases housed in the ICES facility to illustrate how VISA_M3R3 can be used to identify AKI-associated medications and medication combinations among older patients. In Section 3.1, we provide an overview of the design process and participants. We then describe data sources and cohort entry criteria in Sections 3.2 and 3.3, respectively. Section 3.4 explains the implementation details of our VA system. Finally, in Section 3.5, we introduce the components of VISA_M3R3 and briefly describe how the overall system works, which is also discussed more extensively in Section 4.

3.3.1 Design Process and Participants

Healthcare tasks usually include both well- and ill-defined problems. The well-defined tasks have specific goals, clear expected solutions, and, oftentimes, a single solution path. On the contrary, ill-defined tasks do not have clear goals, expected solutions, or solution paths (Arifin et al., 2017).

To help us understand how healthcare practitioners perform real-world tasks, and to help us conceptualize and design VISA_M3R3, we adopted a participatory design approach. Participatory design is a co-operative approach that involves all stakeholders (e.g., partners, end-users, or customers) in the design process to ensure the end product meets their needs (Muller, 2007). A clinician-scientist, a statistician, an epidemiologist, data scientists, and computer scientists were involved in the design and evaluation process of VISA_M3R3. During the initial stage in the participatory design process, we realized that healthcare experts solve ill-defined problems in many different ways. It is difficult and sometimes impossible to determine a single correct problem-solving strategy (i.e., analytics and/or visualization techniques) for ill-defined tasks. Different techniques have their strengths and weaknesses, and there are different criteria to find out which technique is more appropriate for a specific problem. As such, we asked experts to provide us with 1) a list of varying real-world, EMR-driven tasks that they perform, 2) analytics techniques they usually rely on to accomplish those tasks, 3) visualization techniques with which they are familiar, and 4) formative feedback on design decisions. In our collaboration with experts, we recognized two high-level tasks to consider in designing VISA_M3R3 system. 1) They would like to study the relationships between prescribed medications and AKI; 2) They would like to identify commonly prescribed medication combinations and understand the impact of different combinations on AKI. We were told that healthcare experts usually use different regression techniques to accomplish these types of tasks. Since the system has been designed to assist clinicians and healthcare researchers at the ICES-KDT program, we decided to incorporate the analytical and visualization techniques with which they are more familiar. This was essential to build trust between the proposed system and its end-users.

3.3.2 Data Sources

For the particular version of VISA_M3R3, we are primarily interested in analyzing medications prescribed to older hospitalized patients in Ontario. Accordingly, we obtained patient characteristics, prescriptions, and hospital admission data from 5 health administrative databases. We used the Ontario Drug Benefit Program database to get medication use data. We acquired patient characteristics data from the Registered Persons Database, which contains demographic data on all Ontario residents who have ever been issued a health card. We obtained hospital admissions and emergency department (ED) visit data from the Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System, respectively. International Classification of Diseases, ninth (pre-2002), and tenth revision (post-2002) codes was used to identify the baseline comorbidities and incidence of AKI from ED visit and hospital admission data.

3.3.3 Cohort Entry Criteria

We developed a cohort of individuals aged 65 years or older who were admitted to hospital or who visited the ED between April 1, 2014 and March 31, 2016. The ED visit date or hospital admission date served as the index (cohort entry date). If an individual had multiple ED visits or hospital admissions, we selected the first incident. Individuals with an invalid healthcare number, age, and/or sex were excluded from the cohort. A 120-day look-back window from the index date was used to capture the associated medication use data. We used a 5-year look-back window to identify relevant baseline comorbidities.

3.3.4 Implementation Details

The current VISA_M3R3 system is implemented in HTML, JavaScript library D3, standard PHP programming language, and R packages. R was used to develop the Analytics module. Html and D3 were used to create various external representations in the Visualization module. The communication between these two modules is implemented using PHP and JavaScript.

Most of the data analytics components were developed in R (version-3) because it 1) provides extensive support for carrying out data mining operations such as regression and frequent itemset mining, 2) is available in ICES workstations, 3) has a vast array of libraries, 4) is a platform-independent tool, 5) is an open-source tool, and 6) is constantly growing and providing updates whenever new features are available.

We used D3 to implement external representations of the Visualization module because of the following reasons. 1) D3 offers a data-driven approach to help users attach their data to the DOM (Document Object Model) element. 2) It allows users to get access to full capabilities of modern web-browsers. 3) D3 uses a functional style that enables users to reuse JavaScript code and add functionalities. 4) It is compatible with other programming languages and platforms that have been used in this system. And 5) D3 is free and open-source software.

3.3.5 Workflow

As shown in Figure 1, VISA_M3R3 has three modules: Analytics, Visualization, and Interaction. The Analytics module is composed of two components: 1) single-medication analyzer and 2) multiple-medications analyzer. The Visualization module is composed of five views: 1) single-medication view, 2) multiple-medications view, 3) frequent-itemsets view, 4) covariates view, and 5) medication-hierarchy view. The Interaction module provides users with six main actions: 1) arranging, 2) drilling, 3) filtering, 4) searching, 5) selecting, and 7) transforming. The basic workflow of the system is as follows.

First, an integrated dataset is created from different EMR databases stored at ICES. Next, the inclusion and exclusion criteria are applied to build the final cohort. The variables in the comorbidity and prescription data are then encoded and transformed into forms appropriate for analysis. After applying pre-processing techniques, we split the dataset into two groups. One contains the single medication data, and the other contains medication combination data; the latter is generated from the frequent itemset mining algorithm. We develop a number of multivariable regression models on both groups of data. The models are then validated through Bonferroni correction and mapped into respective visual representations. We developed five views to represent data items created from different analysis techniques. The output of the single-medication and multiple-medications analyzers are encoded into two scatter plots in the single-medication and multiple-medications views, respectively. The frequent-itemsets view represents the result of the frequent itemset mining algorithm using a chord diagram. The covariates view allows users to control the information presented in other views though sliders. The medication-hierarchy view includes a data table to display additional information about data elements from the original dataset. Users are allowed to perform a number of actions on the visual representations to manipulate data items. For instance, users can highlight and/or filter out certain items and drill down into the details of the selected data elements in different views.

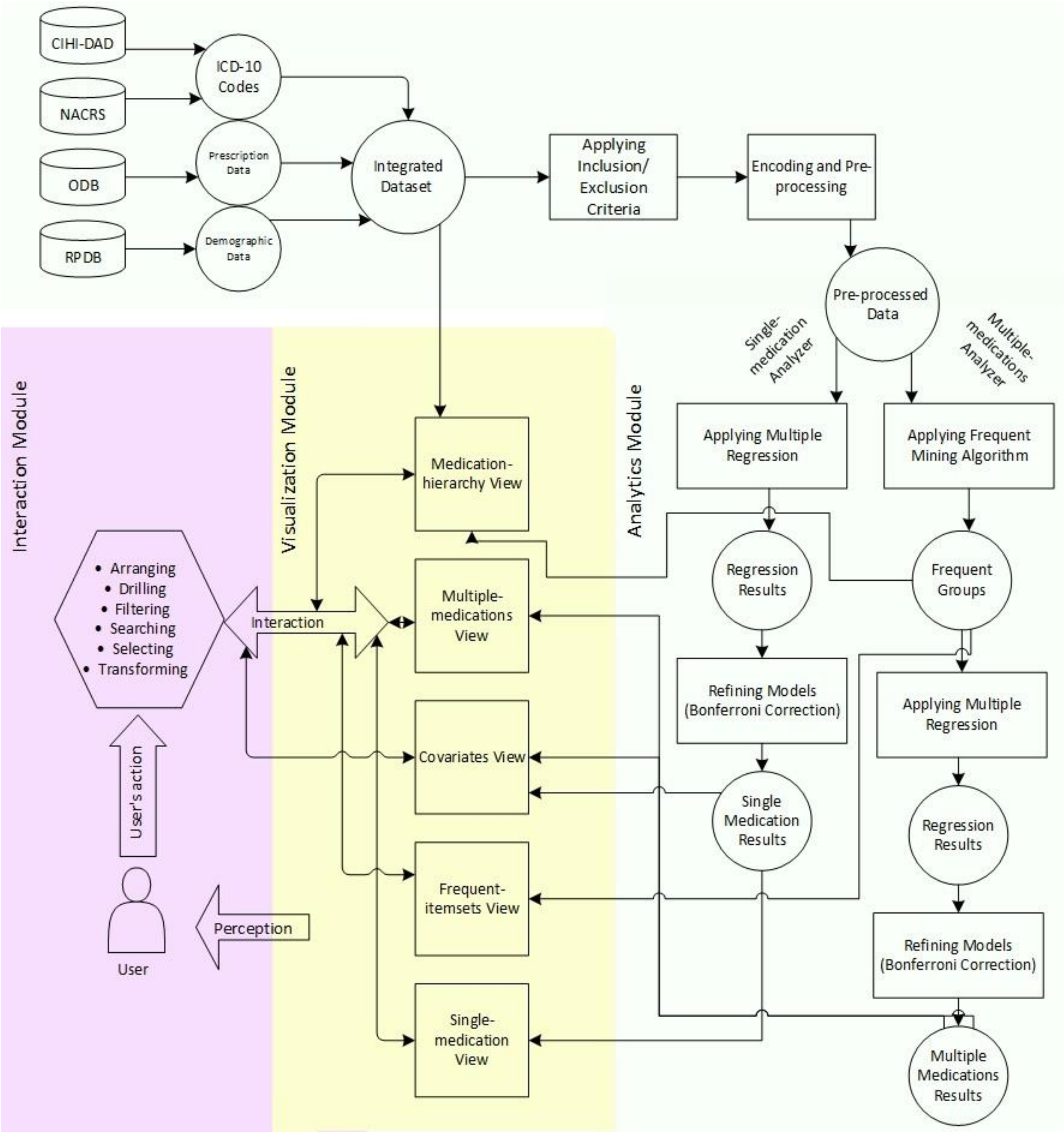

CIHI-DAD
NACRS
ODB
RPDB
ICD-10 Codes
Prescription Data
Demographic Data
Integrated Dataset
Applying Inclusion/ Exclusion Criteria
Encoding and Pre-processing
Pre-processed Data
Single-medication Analyzer
Multiple-medications Analyzer
Interaction Module
Visualization Module
Analytics Module
Applying Multiple Regression
Applying Frequent Mining Algorithm
Regression Results
Frequent Groups
Refining Models (Bonferroni Correction)
Applying Multiple Regression
Single Medication Results
Regression Results
Refining Models (Bonferroni Correction)
Multiple Medications Results
Medication-hierarchy View
Multiple-medications View
Covariates View
Frequent-itemsets View
Single-medication View
Arranging
Drilling
Filtering
Searching
Selecting
Transforming
Interaction
User's action
Perception
User

Figure 3-16: Workflow diagram of VISA_M3R3. Different colors are used to show the separation of the three main modules.

3.4 Design of VISA_M3R3 and Results

In this section, we describe the three main components of VISA_M3R3 as well as some results. Section 4.1 (Analytics module) explains how the data is processed and offers a summary of its results. Section 4.2 (Visualization module) describes VISA_M3R3's interfaces and discusses how the system helps users in interpreting results. Finally, Section 4.3 (Interaction module) illustrates how users can interact with the displayed data.

3.4.1 Analytics Module

We use VISA_M3R3 to analyze ICES' EMRs to identify individual medications and medication combinations that are associated with AKI. Our system aims to facilitate understanding of relationships among medications, medication combinations, and AKI. The Analytics module of VISA_M3R3 performs an individual and group analysis using logistic regression and frequent itemset mining to achieve this goal.

3.4.1.1 Single-Medication Analyzer

Single-medication analyzer includes the regression models created to identify the association between each medication and AKI. In order to capture an accurate association, we include the demographic and comorbidity variables as potential covariates in the models. For demographics (i.e., the study of a population based on certain non-medical factors), we include the following variables in the models: age, sex, income quintile, rural location, and long-term care. For comorbidity (commonly defined as any distinct additional disease or condition that has existed during the clinical course of a patient who has the first disease or condition under observation), we include the following variables in the models: diabetes mellitus, hypertension, heart failure, coronary artery disease, cerebrovascular disease, peripheral vascular disease, chronic liver disease, chronic kidney disease, major cancers, and kidney stones. We obtain the medication prescription data from the Ontario Drug Benefit Program database. This database includes medication name, medication dose, date filled, and route-of-administration of the prescriptions. We identify 595 different medications by analyzing prescriptions that have been filled 120 days before the index date. Thus, we create 595 binary variables to record the medication use data for each medication and each patient. We also gather the class and subclass information of these medications from the literature.

We combine data from different sources into a single dataset. The combined dataset contains 5 demographic, 10 comorbidity, and 595 medication variables for each patient included in the cohort. In total, there are 926,005 unique patients in the dataset. Next, we apply the necessary pre-processing and transformation techniques on the combined dataset to make it ready for the regression analysis. We use the “glm” function in R packages to develop separate multivariable logistic regression models (Williams et al., 1984) for each medication in the dataset. Thus, the regression formula includes AKI as the response variable and medication, demographics, and comorbidities as predictor variables. The “family” argument in the “glm” formula is set to “binomial”. We use the “summary” function to obtain the estimate, *p-value*, standard error, and z-score for each coefficient. In addition, the “confit” function is used to compute 95 percent confidence intervals and odds ratio.

VISA_M3R3 provides users with the ability to compare regression models based on their odds ratios, confidence intervals, *p-values*, and standard errors. Odds ratio measures the association between medication and AKI. A high odds ratio for a specific medication indicates a stronger positive association between that medication and AKI. A list of statistically significant medications is created by filtering models based on the *p-value* of the medication variable’s coefficient. A small *p-value* indicates that it is unlikely that an observed relationship between the predictor (i.e., medication) and response variable (i.e., AKI) is due to chance. Out of 595, we find 55 medications that are strongly associated with AKI. In order to avoid false positives when comparing multiple independent models, we make the alpha value lower based on the Bonferroni correction to account for the number of comparisons being done. A *p-value* less than 8.4e-5 (divide 0.05 by 595) is considered to be statistically significant in this context. Next, we calculate the frequency of each medication in the list. Data items produced through the single-medication analyzer include odds ratios, confidence intervals, *p-values*, standard errors, and usage frequencies of 55 medications. Users of VISA_M3R3 can explore and manipulate these data items to make sense of how an individual medication can affect AKI. Users' sensemaking tasks include, but are not limited to, identifying medications with high odds ratio and lower *p-value*, understanding the comparative risk of medications, assessing the behavior of medication class or subclass, and exploring data items at various levels of abstraction.

3.4.1.2 Multiple-Medications Analyzer

In order to identify the medication combinations that are associated with AKI, we first prepare a dataset of frequently prescribed medications. Since we have 595 individual medications, the total number of combinations is a large number. Therefore, we use the Eclat algorithm (Agrawal et al., 1993) to obtain frequent combinations with a support of 0.07%. Eclat is a frequent itemset mining algorithm that employs a depth-first search to discover groups of items that frequently occur in a transaction database. An itemset that appears in at least a pre-defined number of transactions is called a frequent itemset. At this stage, a total of 24,212 frequent itemsets (i.e., medication groups) are produced from 595 individual medications.

A number of binary variables are created to record the usage of the mediation groups. We set the value of a particular medication group for a patient when that patient has been dispensed all medications within the group within 120 days before the index date (at least once per medication). Next, we apply a multivariable logistic regression model on each medication group to identify potential accumulative nephrotoxicity. The formula includes group variables, individual medication variables that belong to the group, demographic variables, and comorbidities as predictors. Statistically significant medication groups are identified by filtering the models based on a Bonferroni-corrected alpha value (divide 0.05 by the number of medication groups). We also calculate the usage frequency of 78 medication groups that are found to be statistically significant.

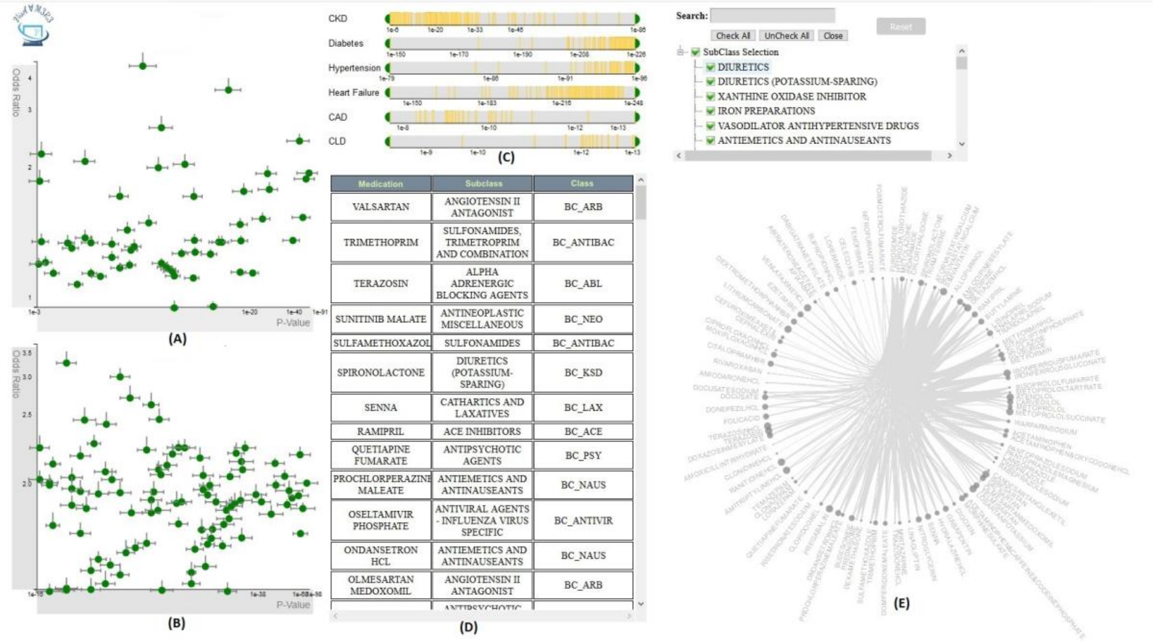
In the multiple-medications analyzer, we employ a combination of frequent itemset mining and logistic regression to generate data items such as frequent medication combinations, statistically significant medication groups, *p-values*, odds ratios, confidence intervals, and standard errors. These data items allow users to understand the synergistic effect of a combination of different medications on AKI. Users' sensemaking tasks include, but are not limited to, identifying medication groups with high impact on AKI, understanding the comparative risk of medications within a group, and exploring data items at various levels of abstraction. VISA_M3R3 organizes data items in different visual representations to allow users to perform these tasks.

3.4.2 Visualization Module

VISA_M3R3 (Figure 2) is composed of five main views: single-medication view, multiple-medications view, covariates view, medication-hierarchy view, and frequent-itemsets view. These views are supported by a number of selection controls, such as search bar and collapsible tree structure. Each of these visualizations represents an important aspect of the Analytics module. In this section, we discuss how data items generated in the Analytics module are encoded as visual representations to allow users perform the activities and tasks mentioned in the previous section.

3.4.2.1 Single-Medication View

Single-medication view uses a scatter plot to represent the results of individual regression models for all the medications, as displayed in Figure 3. The generated scatter plot displays each model in proximity to each other based on their p-value and odds ratio. A linear scale is used for the vertical axis (odds ratio), whereas a log scale is used for the horizontal axis (p-value) since the p-value is exponential. Medications that are plotted closer together affect the risk of developing AKI in a similar manner. The regression model for each medication is encoded as a glyph where horizontal lines on both sides of each circle represent the confidence interval, and the vertical line shows the standard error of the model. The single-medication view enables users to identify high-risk medications that are associated with AKI and understand the comparative risk of these medications. For instance, the glyph in the top-right corner with a p-value of 1e-45 and an odds ratio of 2.4 represents Metolazone. These values suggest that the odds of developing AKI for a patient using this medication are more than two times higher than a patient with similar conditions who is not using it.



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

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
	ANTIPSYCHOTIC	

Figure 3-17: The Visualization module of VISA_M3R3 is composed of five views: (A) single-medication view, (B) multiple-medications view, (C) covariates view, (D) medication-hierarchy view, and (E) frequent-itemsets view.

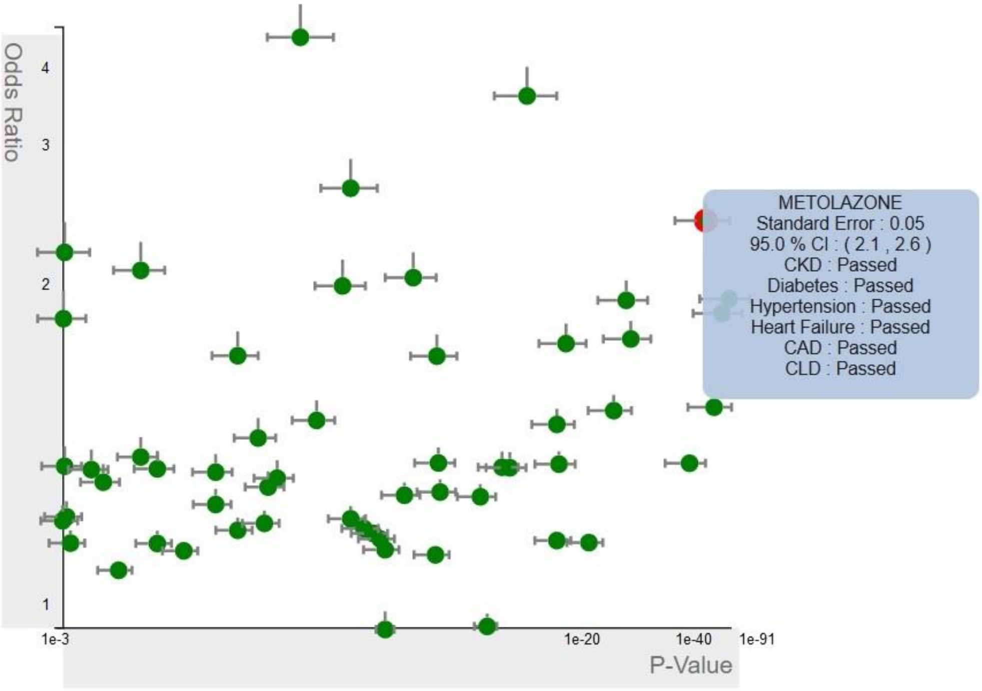


Figure 3-18: Scatter plot of single-medication view.

3.4.2.2 Multiple-Medications View

The multiple-medications view, displayed in Figure 4, uses another scatter plot to represent the results of the regression analysis of groups that are created by the frequent itemset mining algorithm. Each glyph in this scatter plot encodes a medication group model. Similar to the single-medication view, horizontal lines on both sides of each circle in the glyph represent the confidence interval, and the vertical line shows the standard error of the model. We map the p-value and odds ratio to the x- and y-axis, respectively. The multiple-medications view provides users with the ability to detect medication groups that are associated with AKI. For instance, through frequent itemset mining analysis, we find that the pair of Gabapentin and Furosemide medications are frequently prescribed together. As shown in Figure 4, this pair appears to be associated with AKI with a p-value of 1e-26.

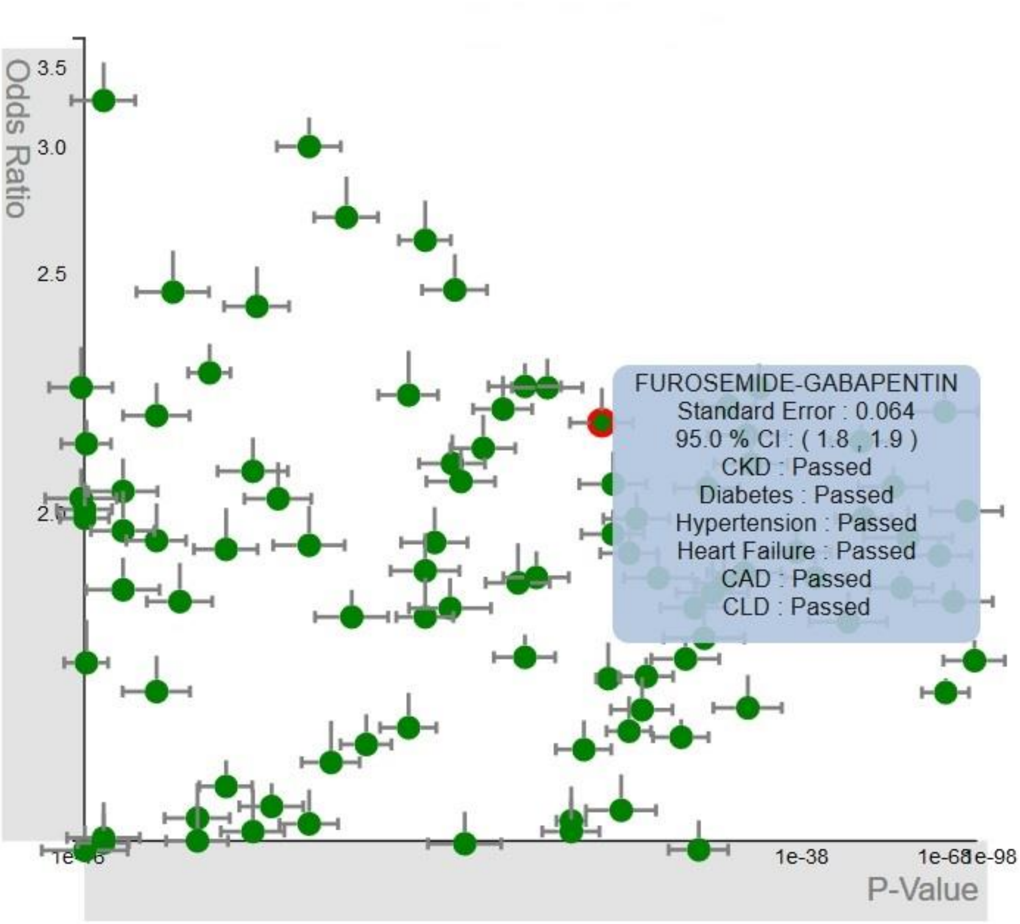


Figure 3-19: Scatterplot of multiple-medications view.

3.4.2.3 Frequent-Itemsets View

Frequent-itemsets view represents the result of the frequent itemset mining analysis by showing all possible combinations of the most frequent items using a chord diagram. As shown in Figure 5, medications are mapped to nodes along the circumference of the circle. Each node consists of an individual circle and a text field showing the name of the medication. Each chord (link) connects two nodes (medications) if they co-occur in the dataset within a certain timeframe. For instance, as shown in Figure 5, there are links between Moxifloxacin Hcl and three other medications (Furosemide, Allopurinol, and Amlodipine besylate) because these three medications have been prescribed with

Moxifloxacin Hcl more than a certain number of times (0.07 percent of the total population) within 120 days prior to the index date.

The size of the circle of each node displays the frequency of the medication in the dataset. Higher usage frequency of a certain medication results in a larger radius for the circle representing that medication. This allows users to visually compare medications based on their use frequency. For instance, a relatively large radius of the circle representing Ramipril indicates that it is one of the frequently prescribed medications in Figure 5-B.

The nodes that belong to the same subclass are placed close to each other separated by spaces. This enables users to visually identify the nodes that share common characteristics (i.e., belong to the same subclass). For instance, users can detect that Furosemide, Hydrochlorothiazide, Metolazone, Indapamide, and Chlorthalidone are all Diuretics; therefore, they are placed in the same group (Figure 5-A). The frequent-itemsets view also reveals subclasses that are composed of a higher number of AKI-associated medications. It can be observed from Figure 5 (C-1 and C-2) that there are two subclasses (Angiotensin and Beta-blockers) that contain six medications that are associated with AKI.

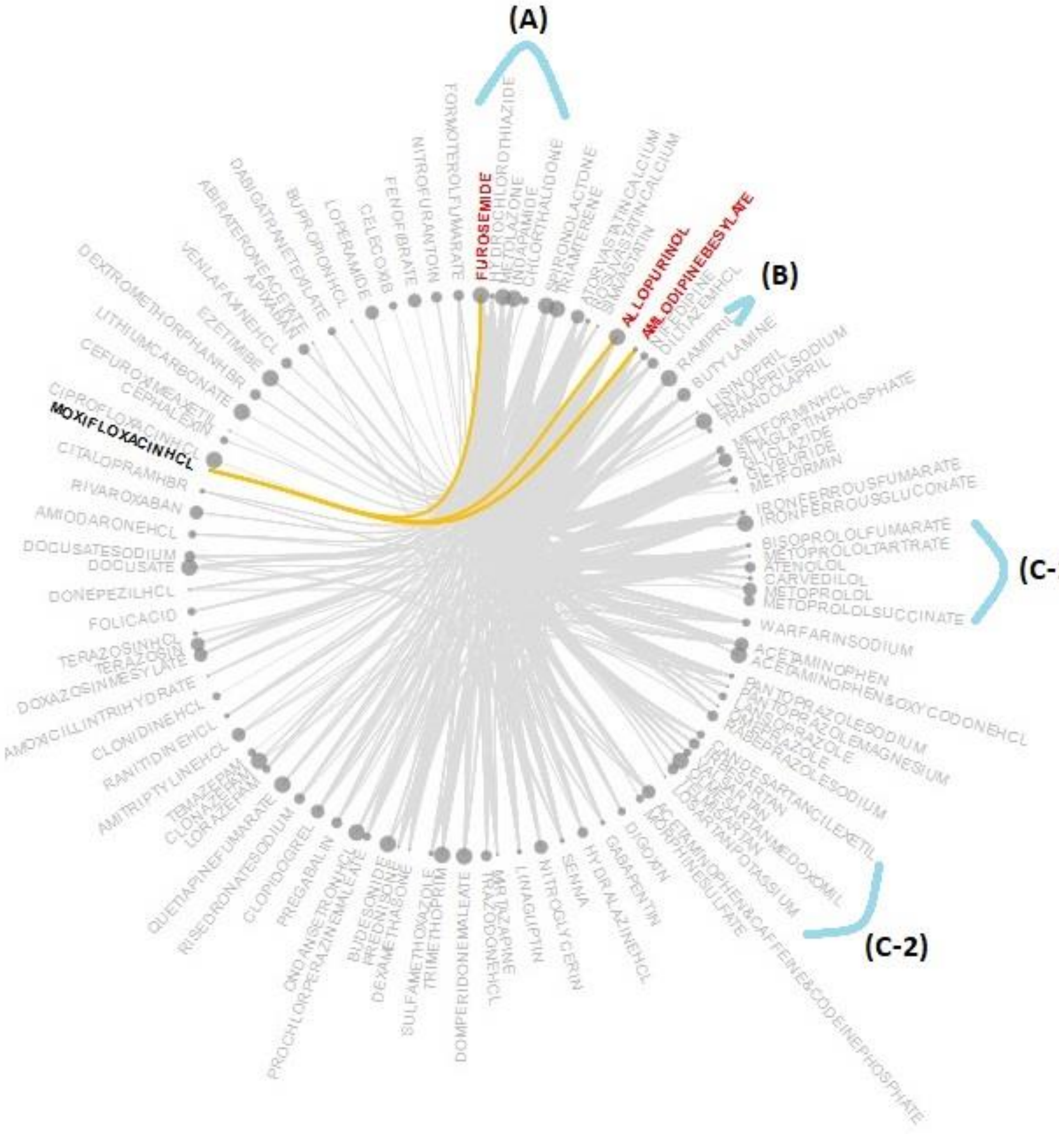


Figure 3-20: Chord diagram showing the results of the frequent itemset mining analysis in the frequent-itemsets view.

3.4.2.4 Covariates View

The covariates view is composed of several sliders that filter data items with respect to different covariates involved in the regression model. The number of sliders depends on the number of covariates that are found to be statistically significant based on the result of the regression analysis. As displayed in Figure 6, six sliders are generated to create control for cancer, diabetes, hypertension, heart failure, coronary artery disease, and coronary liver disease.

Each slider included in the covariates view has three components (a rectangle, vertical lines, and two arc-shaped handles). The rectangle contains the other two components in it. The length of the rectangle represents a linear or log scale, depending on the type of variable it is representing. A linear scale is used when the slider represents the odds ratio of a covariate. We use a log scale to represent the p-value of a covariate. All sliders are generated based on the p-value of the covariates. The vertical lines in the rectangles represent the regression models of both single-medication and multiple-medications analyzers. The placement of the line on the horizontal axis depends on the p-value or odds ratio of the covariate in the corresponding model. For instance, in the slider representing diabetes (second from the top in Figure 6), most of the models are densely clustered in the right corner. This indicates that diabetes has a high impact on the association between medications and AKI. Two arc-shaped handles are placed on both ends of the rectangle to allow users to choose a range of values on the horizontal axis.

3.4.2.5 Medication-Hierarchy View

The medication-hierarchy view contains a data table to provide a list of medications that have been selected through other views, as displayed in Figure 7. The table has three sortable columns for medications, subclasses, and higher-level classes. Each subclass contains a set of medications that share common chemical structures and mechanisms of action, and/or are used to treat similar diseases. A class contains medication subclasses that can be grouped together because of their similarity.

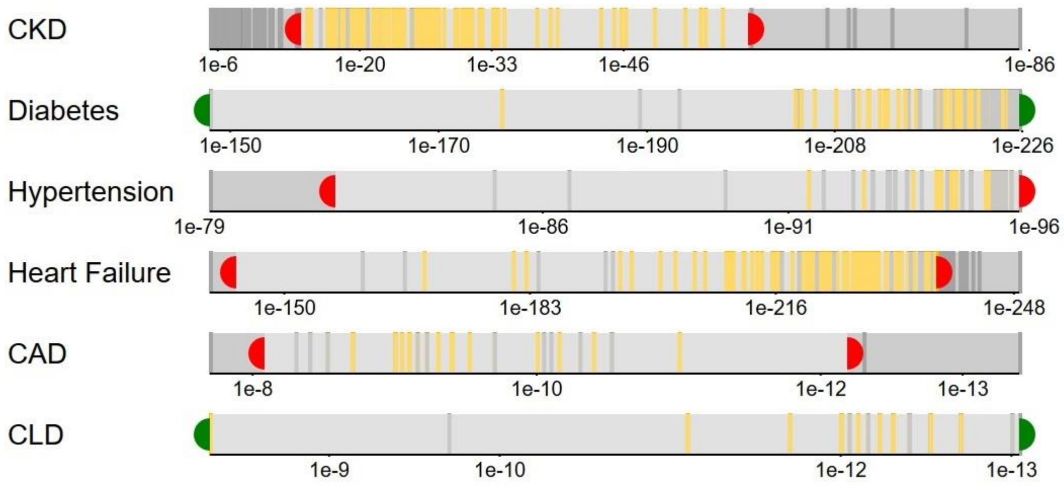
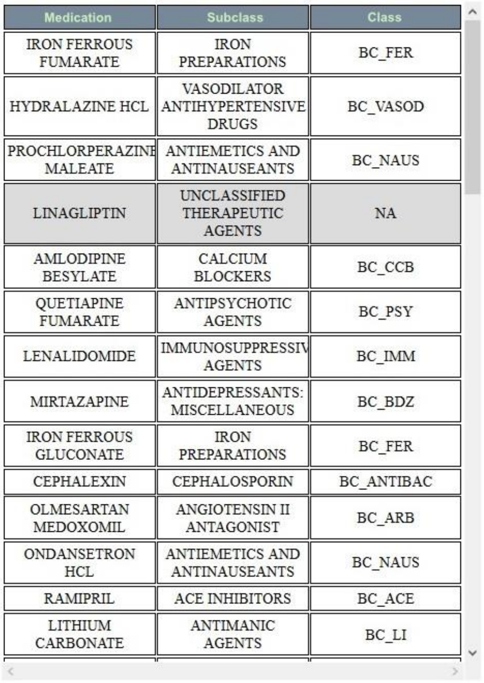


Figure 3-21: Six sliders representing different covariates in the covariates view.



Medication	Subclass	Class
IRON FERROUS FUMARATE	IRON PREPARATIONS	BC_FER
HYDRALAZINE HCL	VASODILATOR ANTIHYPERTENSIVE DRUGS	BC_VASOD
PROCHLORPERAZINE MALEATE	ANTIEMETICS AND ANTINAUSEANTS	BC_NAUS
LINAGLIPTIN	UNCLASSIFIED THERAPEUTIC AGENTS	NA
AMLODIPINE BESYLATE	CALCIUM BLOCKERS	BC_CCB
QUETIAPINE FUMARATE	ANTIPSYCHOTIC AGENTS	BC_PSY
LENALIDOMIDE	IMMUNOSUPPRESSIV AGENTS	BC_IMM
MIRTAZAPINE	ANTIDEPRESSANTS: MISCELLANEOUS	BC_BDZ
IRON FERROUS GLUCONATE	IRON PREPARATIONS	BC_FER
CEPHALEXIN	CEPHALOSPORIN	BC_ANTIBAC
OLMESARTAN MEDOXOMIL	ANGIOTENSIN II ANTAGONIST	BC_ARB
ONDANSETRON HCL	ANTIEMETICS AND ANTINAUSEANTS	BC_NAUS
RAMIPRIL	ACE INHIBITORS	BC_ACE
LITHIUM CARBONATE	ANTIMANIC AGENTS	BC_LI

Figure 3-22: The medication-hierarchy view shows the list of medications and their classes and subclasses.

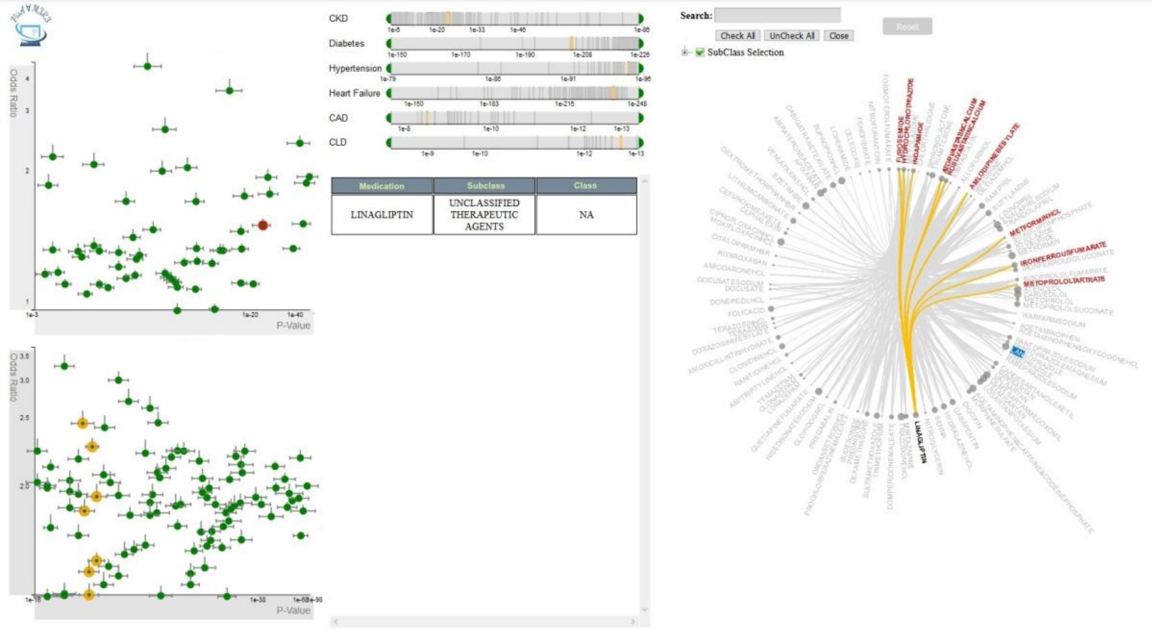
3.4.3 Interaction Module

The Interaction module of VISA_M3R3 is intended to support human-in-the-loop processes of VA. Using the many interactions provided by this module, users can gain insight into the data and manipulate the incorporated data analysis techniques. In this section, we will explore these interactions and discuss how they assist users in identifying high-risk medications and understanding the association between medication groups and AKI. We describe interactions that can be performed in each of the views discussed in the previous section. These interactions not only affect displayed data at the selected view but also change the representation of the data in other views.

3.4.3.1 Single-Medication View Interactions

As shown in Figure 8, the glyphs representing regression models of individual medications are placed very close to each other in the scatter plot. It is sometimes difficult for users to distinguish between models when the glyphs are densely clustered. In order to address this issue, we use the Cartesian fisheye distortion technique on both axes of the scatter plot. Fisheye distortion enables users to zoom in on small areas of the plot without losing sense of its overall structure. Users can apply fisheye distortion by moving their mouse pointer over the grey rectangular areas on both axes of the scatter plot. Fisheye distortion magnifies the local region around the mouse continuously. Users have the ability to enable and disable the fisheye distortion action by clicking on the grey rectangular areas. The color of the rectangular area gets lighter when the fisheye distortion action is disabled. As shown in Figure 8, fisheye on the top-left scatter plot is disabled (light grey rectangles) and bottom-left scatter plot is enabled (relatively dark grey rectangles).

The model selection interaction of the single-medication view affects all the other views. Using this interaction (Figure 8), users can highlight a single medication model throughout VISA_M3R3 in order to 1) determine positions of group models that include the selected medication in the multiple-medications view, 2) detect the position of the selected medication in the covariates view, 3) observe the class and subclass of the selected medication in the medication-hierarchy view, and 4) identify other medications that are frequently prescribed with the selected medication in the frequent-itemsets view. The selected medication is highlighted using the red color in the top-left scatter plot in Figure 8. The glyphs representing corresponding groups in the bottom-left scatter plot, vertical lines representing the medication in the covariates view, and links between selected medication and other frequently used medications in the frequent-itemsets view are all highlighted using the amber color. The utility of this interaction is when users are interested in learning more about a medication that is strongly associated with AKI. They would select a glyph at the top-right corner of the scatter plot, whereupon VISA_M3R3 would highlight and display the relevant information associated with that glyph. Another interaction supported by this view is hovered drilling. This interaction enables users to drill into scatter plot glyphs and get additional information about their corresponding model (Figure 3).

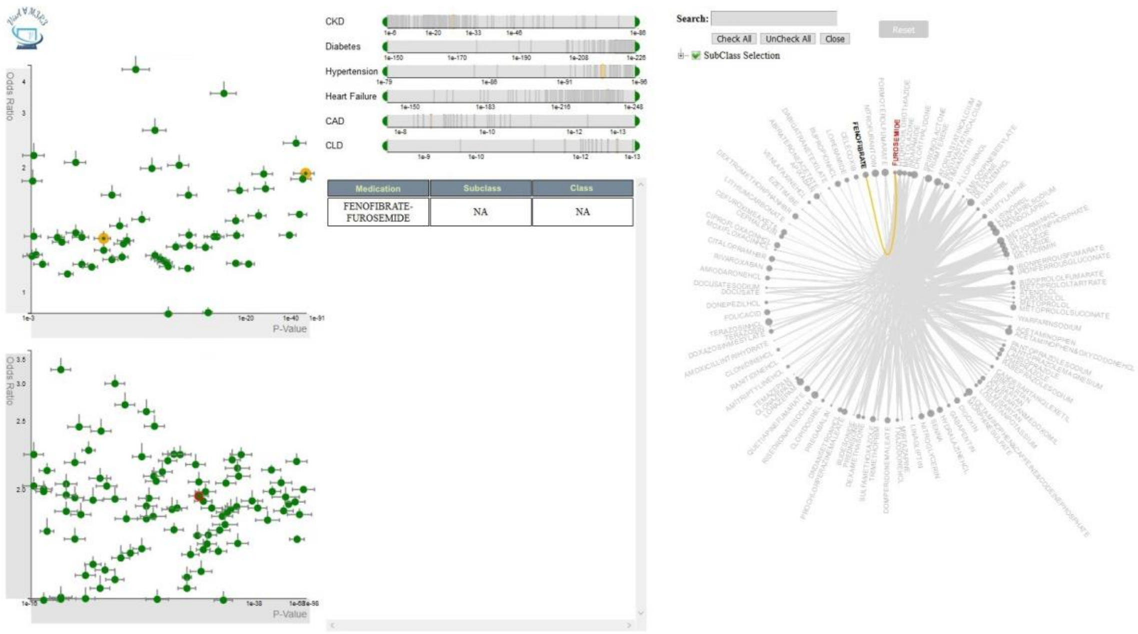


Medication	Subclass	Class
LINAGLIPTIN	UNCLASSIFIED THERAPEUTIC AGENTS	NA

Figure 3-23: Overview of interactions in the single-medication view.

3.4.3.2 Multiple-Medications View Interactions

We designed the interactions of the multiple-medications view in a similar manner to the interactions of the single-medication view. The only difference is how we have designed the selection interaction. The group model selection interaction affects all the other views. Using this interaction (Figure 9), users can highlight a group model throughout the system in order to 1) identify the position of single models included in the selected group in the single-medication view, 2) determine the position of the selected group in the covariates view, 3) observe the class and subclass of medications included in the selected group in the medication-hierarchy view, and 4) highlight the nodes and links representing the group in the frequent-itemsets view. To maintain consistency across all views, the color scheme of the multiple-medications view is similar to the single-medication view. This interaction can be used when users want additional information about a specific group model; they can select the corresponding glyph and observe whether medications included in the selected group are associated with AKI individually in the single-medication view.



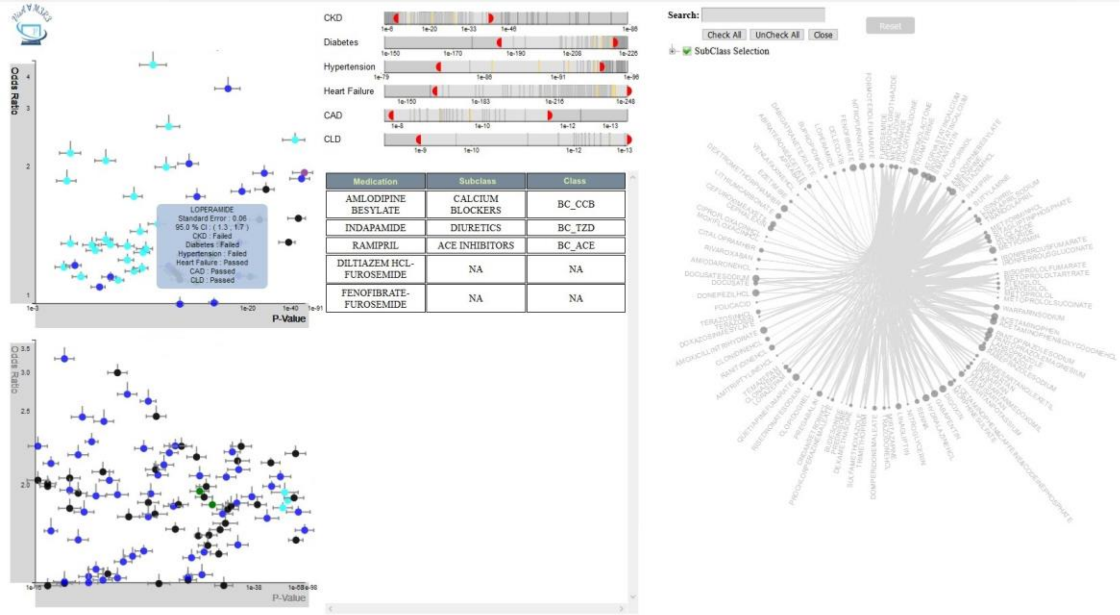
Medication	Subclass	Class
FENOFIBRATE-FUROSEMIDE	NA	NA

Figure 3-24: Overview of interactions in the multiple-medications view.

3.4.3.3 Covariates View Interactions

The single-medication and multiple-medications analyzers produce a set of regression models. These models can be described by a certain number of common attributes (e.g., *p-value* and odds ratio of each covariate) because all of them include the same set of demographic and comorbidity variables as their covariates. The value of an attribute changes based on how each covariate affects the model. It is essential to understand the impact of covariates on both single and group models.

Users can create complex queries composed of several simpler queries related to attributes of different covariates. In each simple query, users apply a filter to the models by selecting a specific range in each slider. Figure 10 shows an example of a complex query involving *p-value* of six covariates. Users can drag both ends of the given sliders to choose a certain range. The color of the range selector changes from green to red when a slider is active. The color of the vertical line representing the model changes from grey to amber when the corresponding model satisfies the criteria of the complex query. Also, the medication-hierarchy view displays the list of models that meet the criteria of the complex query.



LOPERAMIDE
Standard Error : 0.06
95.0 % CI : (1.3 ; 1.7)
CKD : Failed
Diabetes : Failed
Hypertension : Failed
Heart Failure : Passed
CAD : Passed
CLD : Passed

Medication	Subclass	Class
AMLODIPINE BESYLATE	CALCIUM BLOCKERS	BC_CCB
INDAPAMIDE	DIURETICS	BC_TZD
RAMIPRIL	ACE INHIBITORS	BC_ACE
DILTIAZEM HCL-FUROSEMIDE	NA	NA
FENOFIBRATE-FUROSEMIDE	NA	NA

Figure 3-25: Overview of interactions in the covariates view.

In many situations, users struggle to choose appropriate ranges for the sliders. As a result, the query might produce an empty or a limited result set. In order to address this issue, we implemented a sensitivity encoding mechanism in VISA_M3R3 (Spence, 2002). The sliders are set to their maximum and minimum ranges by default. In this case, the color of the glyphs in both scatter plots is set to green because all models satisfy the query. The color of the glyph in the scatter plots encodes the number of simple queries its corresponding model satisfies in the covariates view, as shown in Table 1 and Figure 10.

Table 3-4: Sensitivity encoding using color coding of glyphs.

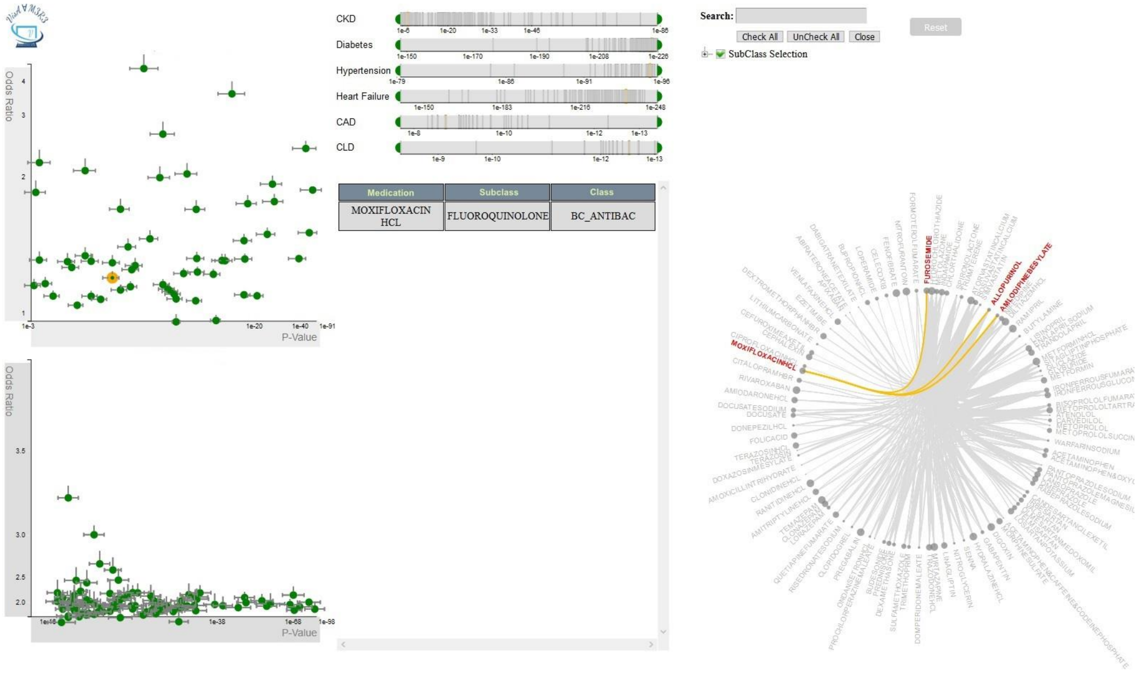
Number of satisfied filters	Color of the glyphs
6	Green
5	Black
4	Blue
3	Cyan
2	Purple
1	Grey
0	Yellow

3.4.3.4 Frequent-Itemsets View Interactions

The selection interaction of the frequent-itemsets view affects the single-medication view, covariates view, and medication-hierarchy view. Using this action (Figure 11), users can select a single medication from the chord diagram by clicking on its corresponding node in order to 1) identify other medications that are frequently prescribed with the selected medication in the frequent-itemsets view, 2) understand the association between the selected medication and AKI in the single-medication view, 3) determine the position of the selected medication in the covariates view, and 4) observe the class and subclass of the selected medication in the medication-hierarchy view. Figure 11 shows an example of this interaction. Selecting Moxifloxacin Hcl would highlight the links and the names of the other medications (i.e., Furosemide, Allopurinol, and Amlodipine besylate) that are frequently consumed with Moxifloxacin Hcl.

3.4.3.5 Medication-Hierarchy View Interactions

Medication-hierarchy view supports two interactions as shown in Figure 12. Users can sort the table based on medication name, subclass, or class by clicking on the corresponding column header. For instance, if they click on “Medication”, medication names in the table get sorted alphabetically. They can also sort in the opposite order by clicking on the same header again. In addition, users can click on any row in the table to select the corresponding medication or medication groups. Selected medications get highlighted in all other views.

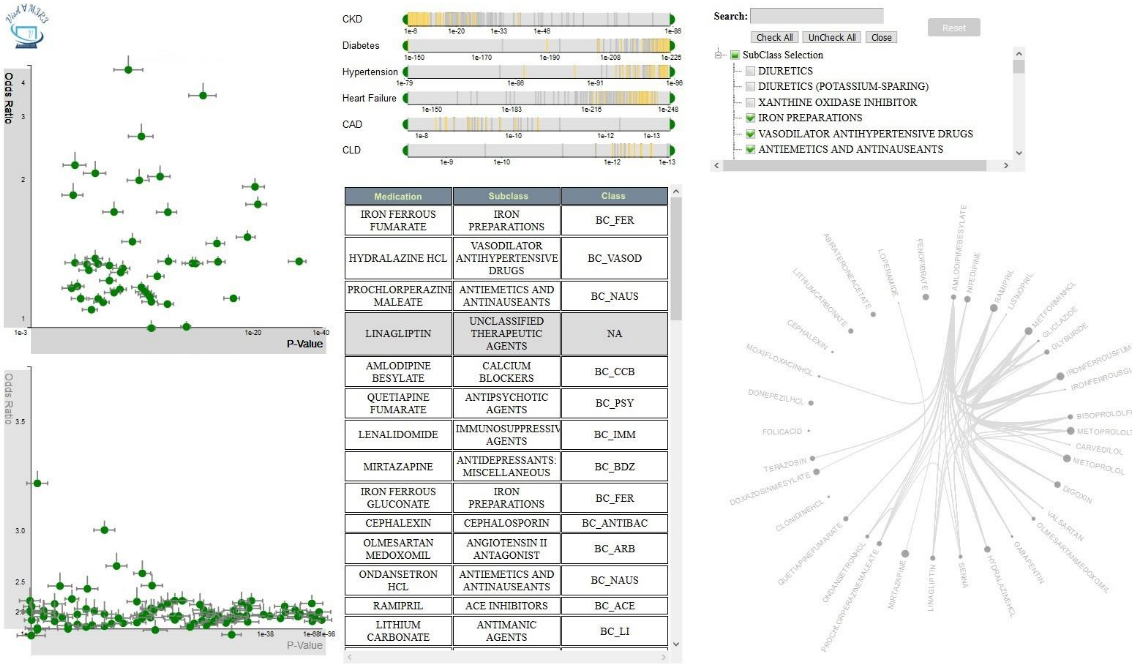


Medication	Subclass	Class
MOXIFLOXACIN HCL	FLUOROQUINOLONE	BC_ANTIBAC

Figure 3-26: Overview of interactions in the frequent-itemsets view.

3.4.3.6 Selection Controls

Selection controls include a search bar, a collapsible tree structure, and several buttons to control the information displayed in different views (top-right corner of Figure 12). If users are interested in learning about a specific medication, they can enter the name of that medication (or part of the name) in the search bar and the information related to that medication gets displayed in the medication-hierarchy view. Users can expand the tree structure by clicking on the “+” icon at the top-right corner to get a menu of medication subclasses. Each item in the menu is linked to a checkbox. It is possible to limit data items displayed in other views by selecting these checkboxes. For instance, as shown in Figure 12, users have selected a number of subclasses such as Iron preparations, Vasodilator antihypertensive, and Antiemetics & Antinauseants in the collapsible tree structure to limit the number of data items shown in the scatter plots, data table, and chord diagram.



Search:

Check All | UnCheck All | Close

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

Medication	Subclass	Class
IRON FERROUS FUMARATE	IRON PREPARATIONS	BC_FER
HYDRALAZINE HCL	VASODILATOR ANTIHYPERTENSIVE DRUGS	BC_VASOD
PROCHLORPERAZINE MALEATE	ANTIEMETICS AND ANTINAUSEANTS	BC_NAUS
LINAGLIPTIN	UNCLASSIFIED THERAPEUTIC AGENTS	NA
AMLODIPINE BESYLATE	CALCIUM BLOCKERS	BC_CCB
QUETIAPINE FUMARATE	ANTIPSYCHOTIC AGENTS	BC_PSY
LENALIDOMIDE	IMMUNOSUPPRESSIV AGENTS	BC_IMM
MIRTAZAPINE	ANTIDEPRESSANTS: MISCELLANEOUS	BC_BDZ
IRON FERROUS GLUCONATE	IRON PREPARATIONS	BC_FER
CEPHALEXIN	CEPHALOSPORIN	BC_ANTIBAC
OLMESARTAN MEDOXOMIL	ANGIOTENSIN II ANTAGONIST	BC_ARB
ONDANSETRON HCL	ANTIEMETICS AND ANTINAUSEANTS	BC_NAUS
RAMIPRIL	ACE INHIBITORS	BC_ACE
LITHIUM CARBONATE	ANTIMANIC AGENTS	BC_LI

Figure 3-27: Overview of interactions in the medication-hierarchy view and selection controls.

3.5 Discussion

In this paper, we have shown how VA systems can be designed to address the challenges of prescription data stored in EMRs in a systematic way. To achieve this, we have reported the development of VISA_M3R3, a VA system designed to assist medical researchers at ICES’ KDT program. VISA_M3R3 incorporates three main components: an Analytics module, made up of single-medication analyzer and multiple-medications analyzer; a Visualization module, made up of five views: single-medication view, multiple-medications view, covariates view, frequent-itemsets view, and medication-hierarchy view; and an Interaction module, made up of a set of different human-data interactions. VISA_M3R3 is unique in the manner in which it combines multivariable regression with Eclat to support underlying processing in the computing space and implements fisheye and sensitivity encoding to provide support for the representation and interaction spaces. It offers a balanced distribution of processing load through a proper integration of analytics techniques (i.e., regression and frequent itemset mining in the Analytics module) with visual representations (i.e., different interactive views in the Visualization module) to facilitate high-level cognitive tasks. Some of the main tasks commonly performed by researchers, and which VISA_M3R3 is designed to support, include: 1) compare multiple regression models, 2) understand the relationship between different predictors and a response variable, 3) identify the frequent itemsets from items of interest, and 4) interpret multivariable regression models. VISA_M3R3 is primarily designed as a research tool for the medical researchers at ICES' KDT program, and it is up to them to decide how this system will be applied within the healthcare system. A number of training materials have been prepared to assist new users who are not familiar with the analytics and visualization techniques incorporated in VISA_M3R3 to use the system effectively.

We have demonstrated how VISA_M3R3 can be used to detect AKI-associated medications among older patients who visited the hospital or emergency department in Ontario between 2014 to 2016 using ICES health administrative data. We have seen that VISA_M3R3 allows healthcare researchers to generate hypotheses, understand the relationships among data elements (e.g., medications and diseases), and recognize patterns and trends that would be otherwise difficult to identify. About 9% of all the medications that are prescribed to the older patients have been found to be associated with AKI. Using VISA_M3R3, we detect 55 medications (Furosemide, Allopurinol, Hydrochlorothiazide, Atorvastatin, Spironolactone, Olmesartan Medoxomil, to name a few) and 78 medication combinations (Furosemide & Oseltamivir Phosphate, Allopurinol

& Metolazone, Celecoxib & Quetiapine, and so on) that are associated with an increased risk of AKI. In general, medications belong to Angiotensin Receptor Blockers, Diuretics, Nonsteroidal Anti-inflammatory, and Xanthine Oxidase Inhibitors classes are found to be strongly associated with AKI. Moreover, some combinations of medication classes such as Anti-inflammatory & Antidepressants and Diuretics & Antiviral Agents have been identified with the evidence for increased risk of developing AKI. The lists of medications and medication combinations have been reviewed by a nephrologist to validate the results. Most of these medications are already known to be nephrotoxic in the existing literature, which confirms the accuracy of our findings through VISA_M3R3 (Chao et al., 2015; Kwok M. Ho and Power, 2010; Perez-Ruiz, 2017; Pierson-Marchandise et al., 2017; Verdoodt et al., 2018; Wu et al., 2014).

In terms of the extensibility and scalability of VISA_M3R3, we have designed it in a modular way so that it can easily accept new data sources, data types, and analysis techniques. VISA_M3R3 can be used to investigate many other clinical problems, such as identifying risk factors associated with hypertension, and understanding the relationship between dietary habits and diabetes. To test the applicability of the system in different healthcare areas, we have used VISA_M3R3 to detect hospital admission codes (i.e., reasons for hospitalization) that are associated with AKI using healthcare utilization database housed at ICES. We detected 8,543 itemsets by analyzing the hospital admission codes that co-occur frequently. Using VISA_M3R3 to analyze this data, 185 individual codes and 215 group codes are found to be statistically significant. The top few reasons for hospitalization (representing admission codes associated with AKI) include 1) Essential hypertension, 2) Malignant neoplasm of bladder, 3) Non-follicular (diffuse) lymphoma, 4) Mycosis fungoides, 5) Iron deficiency anemia, and 6) Chronic obstructive pulmonary disease. This result also aligns with what has already been known from the literature, which more generally and comprehensively proves the efficacy of VISA_M3R3's design (Anderson et al., 2010; Da’as et al., 2001; Kandler et al., 2014; Malbrain et al., 1994; Martines et al., 2013).

There are four key limitations to the development of VISA_M3R3. The first one is that it reports the regression analysis result of the group models but does not consider how individual items within the group are affecting the outcome. For instance, in the study with medications, VISA_M3R3 reveals that the combination of Furosemide and Metoprolol increases the risk of AKI. However, it does not explain the additive risk of using Metoprolol with or without Furosemide and vice versa. This issue can be resolved by incorporating a stratified analysis on each item available in at least one group. The second limitation is that, even though we have had a participatory design and medical experts have evaluated VISA_M3R3 and have found it very useful and usable, we have not conducted any formal experimental usability studies to evaluate its performance, nor the efficacy of its human-data discourse mechanisms. The third one is that VISA_M3R3 incorporates a limited number of analytics techniques. Although there are more advanced machine learning algorithms in the literature, we decided to design the system based on techniques that are more interpretable to our end-users (i.e., clinicians and healthcare researchers). Fourth, the preparation of the dataset for VISA_M3R3 could be labor-intensive in some situations, depending on the data source and problem at hand. However, there are a number of readily available libraries and packages available to assist users with the data cut and preparative work.

3.6 Conclusion

The purpose of this study is to demonstrate how VA systems can be designed in a systematic way to support EMR-driven tasks and investigation of different clinical problems. We report the development of a VA system (called VISA_M3R3) and demonstrate how it can be used to help medical practitioners and researchers identify medications and medication combinations that associate with a higher risk of AKI. VISA_M3R3's novelty stems from its design: it incorporates multivariable regression, frequent itemset mining, data visualization, and human-data interaction mechanisms in an integrated fashion to support ill-defined, complex EMR-driven tasks. Using VISA_M3R3, we analyzed ICES health administrative data. Through this analysis, 55 medications and 78 medication groups, strongly associated with AKI, were identified.

Although, through clinical studies, a number of these AKI-associated medications and medication groups are known by medical researchers, some of them have never been studied before. VISA_M3R3 can alert and raise physicians' awareness of such potentially AKI-associated medications. This, in turn, can prompt healthcare providers to conduct further clinical investigations to improve healthcare research outcomes. Finally, VISA_M3R3's design concepts are generalizable. They can be used to systematically develop any VA system whose goal is to support medical tasks involving analysis of EMR data using multiple regression models and frequent itemset mining. Applications of such VA systems can lead to the emergence of best practices for developing similar VA systems in other medical domains.

Chapter 4

4 Machine Learning for Identifying Medication-Associated Acute Kidney Injury

This chapter has been published as S.S. Abdullah, N. Rostamzadeh, K. Sedig, D. J. Lizotte, A.X. Garg, and E. McArthur, “Machine Learning for Identifying Medication-Associated Acute Kidney Injury” in the Health Section of the Informatics Journal, Volume 7; May 2020. We changed the format to match the general format of the dissertation. Figure, Table, and Section numbers specified herein are relative to the chapter number. For example, “Table 1” corresponds to Table 4-1; “Figure 1” corresponds to Figure 4-1; and “Section 1.1” corresponds to Section 4.1.1. Moreover, when the term “paper”, “research”, or "work" is used, it refers to this specific chapter.

4.1 Introduction

Acute kidney injury (AKI), defined as a sudden loss of kidney function over a short period of time, affects approximately 10% of patients admitted to hospitals worldwide (Porter et al., 2014; Selby et al., 2012). It is associated with increased mortality, morbidity, and estimated incremental health care costs of more than $200 million in Canada annually (Collister et al., 2017). Medication-induced nephrotoxicity is very common in clinical practice. It accounts for 19% of cases of AKI in a hospital setting (Collister et al., 2017; Gandhi et al., 2000; Kaufman et al., 1991; Miyahara, 1978; Nash et al., 2002; Uchino et al., 2005) and is associated with increased healthcare expenditure (Choudhury and Ahmed, 2006; Collister et al., 2017). For instance, using the medication utilization data in Canada for 2013, Morgan et al. (2016) have reported an estimated healthcare cost of $419 million due to inappropriate prescriptions (Morgan et al., 2016).

Over the last two decades, the incidence rate of AKI has increased in Canada (Liu et al., 2010; Mehrabadi et al., 2014), the United States (Nadkarni et al., 2016; Xue et al., 2006), and the United Kingdom (Kolhe et al., 2016). The increasing occurrence of AKI is related to the changing spectrum of diseases. There is a growing body of evidence showing that patients with multiple comorbidities and extrarenal complications are at a higher risk of developing AKI (Mehta et al., 2004; Siddiqui et al., 2012). For instance, Aikar et al. (Waikar et al., 2006) have shown that the high comorbidity rate, measured by the Deyo-Charlson comorbidity index, is associated with AKI. In a study of 681 AKI patients who are admitted to the intensive care unit, the occurrence of comorbid conditions is high: 37% have coronary artery disease, 30% have chronic kidney disease, 29% have diabetes mellitus, and 21% have chronic liver disease (Mehta et al., 2004). As a patient's number of comorbid conditions grow, there is a rise in associated hospitalizations, physician visits, prescriptions, and expenses (Zulman et al., 2014), ultimately leading to an increase in medication intake. Patients admitted to hospitals, particularly critically ill patients with multiple comorbidities, often take several medications, with up to 25% of these medications having nephrotoxic potential (Choudhury and Ahmed, 2006). A study in 2005 has revealed that out of 7 million adverse medication event reports, 2.7% include an incidence of AKI, of which 16% are known nephrotoxins, 18% are possible nephrotoxins, and the rest are new potential nephrotoxins (Uchino et al., 2005).

The use of nephrotoxic medications is associated with 16-25% of all AKI cases in the adult population (Pannu and Nadim, 2008; Uchino et al., 2005). Few studies have been conducted to identify medications that are commonly associated with AKI. Most of these studies have been limited in assessing the impact of known nephrotoxic medications (Alexander et al., 2017; Moffett and Goldstei, 2011; Rivosecchi et al., 2016). In addition, information on medication combinations that can cause AKI lacks in the literature. It is important to identify those combinations because a combination of multiple nephrotoxins may result in synergistic or accumulative nephrotoxicity, thus increasing the chance of renal failure (Schetz et al., 2005). For example, the risk of developing AKI increases by 53% for each additional nephrotoxic medication used by a patient (Cartin-Ceba et al., 2012). Hence, it is important to identify not only nephrotoxic medications but also medication combinations that affect the risk of AKI. Rivosecchi et al., through an exhaustive literature search, further emphasize the need for a comprehensive understanding of how medication combinations alter the risk of AKI (Rivosecchi et al., 2016). According to a CDC report in 2017, there are about 1,000 known adverse medication effects and 5,000 medications available in the pharmacies (FastStats - Therapeutic Drug Use), making for approximately 125 billion possible adverse medication effects between all possible pairs of medications (Zitnik et al., 2018). Thus, it is impossible to assess medication-induced AKI through this number of clinical trials comprehensively. Moreover, conducting a trial to determine whether to prescribe or not prescribe a potentially harmful combination would likely never receive research ethics board approval.

Data analysis has the potential to address this challenge by employing methods and techniques from different fields, such as data mining, statistics, and machine learning to accomplish various data-driven tasks (Han and Kamber, 2011). It can be used to investigate clinical data to gain both novel and deep insights to help healthcare providers examine medication-induced nephrotoxicity. Recently, several studies have been conducted to identify drug-drug interactions, improve drug-safety science, and predict adverse drug reactions using machine learning techniques (Vamathevan et al., 2019). For instance, Kandasamy et al. (2015) have developed a prediction model to identify drug-induced nephrotoxicity using human induced pluripotent stem cells and random forest (Kandasamy et al., 2015). In addition, Dey et al. (2018) have presented a deep learning framework to predict adverse drug reactions and detect molecular substructures associated with them (Dey et al., 2018). An automatic method of processing adverse event reports using artificial intelligence and robotics is presented in (Schmider et al., 2019). Lysenko et al. (2018) have incorporated Mashup (Cho et al., 2016) and a gradient-boosted tree to predict drug toxicity using biological network data (Lysenko et al., 2018). Although these studies are designed to deal with large bodies of data to solve different medication-related problems, the relationship between medications and AKI has not been studied before through automated data analysis. Automated data analysis techniques allow an incorporation of large quantities of data that creates an opportunity to include additional information to more comprehensively study individual medications and their combinations. It is essential to consider comorbidities while studying the effect of medications since it is not clear whether the underlying comorbidities or medications increase the risk of developing AKI. In addition to comorbidity data, demographics data such as age, sex, and region, are also considered as risk factors for AKI (K. D. Liu et al., 2019; Siew et al., 2016). Therefore, any complete study that investigates nephrotoxic medications or combinations should include demographic and comorbidity data in the analysis. Up until now, there is a lack of well-designed studies that consider demographic and comorbidity data while assessing the risk of developing AKI with the use of single or multiple medications. Even though the identification of nephrotoxic medications is crucial for improved patient care, it has not been studied thoroughly through machine learning techniques.

The purpose of this study is to identify individual medications associated with AKI in hospitalized patients using an automated machine learning approach. We also identify AKI-associated medication combinations and investigate whether the use of multiple medications results in multiplicative effects on the risk of developing AKI. Finally, we investigate how our findings are consistent with data in the existing literature. Our study differs from other studies in three ways: (1) we consider all the frequently used medications in the study, whether they have been known to be nephrotoxic or not; (2) we use a frequent itemset mining algorithm to identify frequent medication combinations and multivariable logistic regression to investigate the association between medication combinations and AKI; and (3) we incorporate the patient's demographic and comorbidity features as potential covariates in the regression models.

4.2 Materials and Methods

This section describes the methodology we have employed to conduct the study. We describe the design process, study setting, workflow, data sources, cohort entry criteria, input features, outcome, analysis processes, and tools.

4.2.1 Design Process and Participants

To help us understand how healthcare providers perform automated analysis, and to help us conceptualize and design our study, we adopted a participatory design method. It is a co-operative method that involves all stakeholders (e.g., designers, intermediary-users, and end-users) in the design process to ensure the output of the analysis meets their needs (Muller, 2007). A statistician, a clinician, an epidemiologist, and several computer scientists were involved in the design and evaluation process of this study. During the initial stage in the designing process, we realized that healthcare providers usually perform medication-safety related studies in many ways. It is difficult to determine a single correct analytics technique for these tasks because different techniques have their strengths and weaknesses. As such, we interviewed healthcare experts to identify the data-driven tasks and analytics techniques with which they are familiar. We identified four data-driven tasks to consider in designing this study through our collaboration with healthcare experts at the ICES-KDT (ICES - 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; KDT - *Kidney Dialysis and Transplantation program*), located in London, Ontario, Canada. 1) Studying the relationships between prescribed medications and AKI. 2) Identifying commonly prescribed medication combinations to older patients. 3) Examining the effect of a medication combination on AKI. 4) Investigating if a certain medication is associated with an increased risk of developing AKI when used with another medication. We came to know that healthcare experts usually rely on different regression techniques to accomplish such tasks. Thus, we decided to employ multivariable regression in this study. We also invited healthcare experts to provide us with formative feedback on design decisions and results.

4.2.2 Design and Setting

We performed a population-based retrospective cohort study in older adults from April 2014 to March 2016 in Ontario, Canada, using administrative health databases located at

ICES. These datasets were linked using unique encoded identifiers and analyzed at ICES. The use of data in this project was authorized under section 45 of Ontario’s Personal Health Information Protection Act, which does not require review by a Research Ethics Board.

Ontario has a population of approximately 13 million residents with universal access to hospital care and physician services, including 1.9 million people aged 65 years or older who have universal prescription drug coverage (14% of the population). We suppressed our results in cells with five or fewer patients to comply with privacy regulations and minimize the chance of re-identification of patients.

4.2.3 Workflow

Figure 1 illustrates the basic workflow of the study presented in this paper. In the first stage, we created an integrated dataset from different health administrative databases stored at ICES. The data sources are explained in Section 2.4. Next, we applied the inclusion and exclusion criteria presented in Section 2.5 to build the final cohort. The demographic and comorbidity features were then encoded and transformed into appropriate forms for analysis in Section 2.6. Section 2.7 describes the outcome (i.e., AKI) and how we identified the incidence of AKI. A brief description of the cohort is presented in Section 2.8. After that, we performed individual and combination analysis, which are discussed in Section 2.9 and 2.10, respectively. The results from both analyses were then validated and presented in Tables 2 and 3.

4.2.4 Data Sources

We ascertained patient characteristics, drug prescriptions, and outcome data from 5 health administrative databases (Appendix A). The datasets were linked using unique, encoded identifiers derived from health card numbers, and patient-level data were analyzed at ICES. We obtained vital statistics from the Ontario Registered Persons Database, which contains demographic data on all Ontario residents who have ever been issued a health card. We used the Ontario Drug Benefit Program database to identify prescription drug use. This database contains highly accurate records of all outpatient prescriptions dispensed to older patients, with an error rate of less than 1% (Levy et al., 2003). We identified hospital admissions, baseline comorbidity data, and emergency department visits from the National Ambulatory Care Reporting System (ED visits) and the Canadian Institute for Health Information Discharge Abstract Database (hospitalizations). We used the International Classification of Diseases, tenth revision (post-2002) codes to assess baseline comorbidities. Baseline comorbidity data were also obtained from the Ontario Health Insurance Plan database, which includes claims for physician services. Coding definitions for comorbidity data are represented in Appendix B.

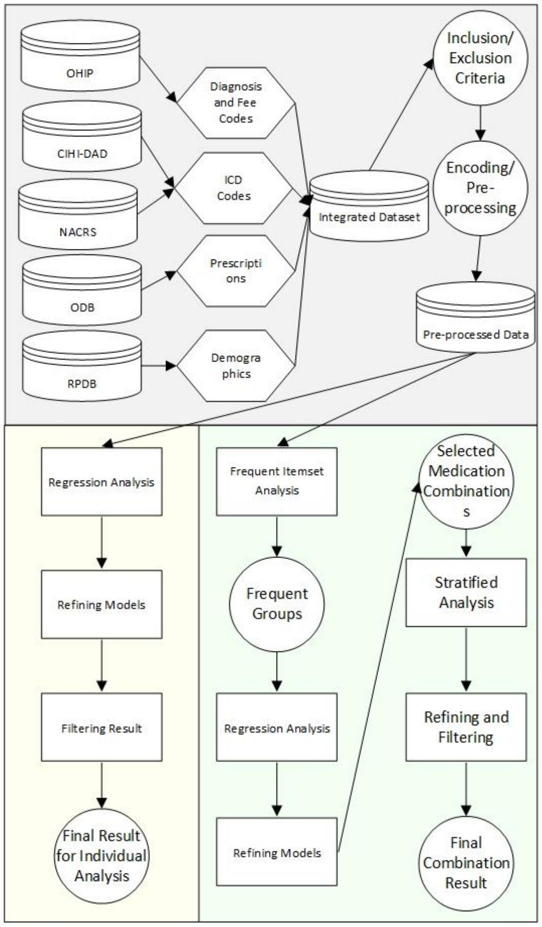


Figure 4-28: Workflow diagram of the study presented in this paper. Different colours are used to show the separation of three main parts (pre-processing, individual analysis and combination analysis).

4.2.5 Cohort Entry Criteria

We identified a cohort of individuals aged 65 years or older who were admitted to hospital or visited the emergency department (ED) between 1st April 2014 and 31st March 2016. The ED visit date or hospital admission date served as the index or cohort entry date. If an individual had multiple ED visits or hospital admissions, we selected the first incident. Individuals with invalid data regarding the health card number, age, and sex were excluded. We also exclude: (1) patients who previously received dialysis or a kidney transplant as AKI is often no longer relevant once a patient develops end-stage kidney disease (diagnosis codes for exclusion criteria are shown in Appendix C); and (2) patients who left the hospital against medical advice or without being seen by a physician.

4.2.6 Baseline Covariates

There were a total of 5 demographic, 10 comorbidity, and 595 medication features in the cohort, which serve as input for the analysis. Demographic information included age, sex, residency status (urban and rural), long term care, and socioeconomic status (income quintile according to Statistics Canada). We used a 5-year look-back window to identify relevant baseline comorbidities, including diabetes mellitus, hypertension, heart failure, coronary artery disease, cerebrovascular disease, peripheral vascular disease, chronic liver disease, chronic kidney disease, major cancers, and kidney stones.

All of the features in the cohort were categorical. We converted the comorbidity features into binary forms. For instance, if a patient had a particular comorbid condition, its corresponding value was taken as "1." We set the value for sex and residency status features if a patient was male and resided in urban areas. The income feature took an integer value ranged between 1 to 5 to represent the income quintile of a particular patient. All these features from different data sources were integrated using the encoded identifiers derived by ICES. Finally, the features in the cohort were transformed into a format and scale that were suitable for the analysis. For each feature in the cohort, we recorded the last value before the index date. Thus, we aggregated multiple values (rows) of a single feature into one by considering the latest values of that feature for each patient.

4.2.7 Outcome: Identification of AKI

AKI was the outcome variable for all the regression models in this study. We identified the incidence of AKI in the first visit to the ED or hospital admission between 1st April 2014 and 31st March 2016. The incidence of AKI was captured using the National Ambulatory Care Reporting System and the Canadian Institute for Health Information Discharge Abstract Database based on the International Classification of Diseases (ICD), Tenth Revision (ICD-10-CA) "N17" diagnostic codes. We considered the first incidence in case of multiple episodes of AKI for a patient. We set the value of the outcome variable if a patient was diagnosed with AKI. We recorded the first incidence of AKI in case there were multiple episodes.

4.2.8 Cohort Characteristics

A total of 924,533 participants were included in the derivation cohort, of which 25,084 (2.7%) had AKI during their hospital or ED encounter. Selected characteristics of this cohort are shown in Table 1. The mean age was 70 years, and 56% were women. Sixteen percent of the patients resided in rural areas, and 6% of them were in long term care. The pre-existing comorbidities were hypertension (88%), diabetes (38%), coronary artery disease (25%), major cancer (16%), heart failure (14%), cerebrovascular disease (3%), peripheral vascular disease (2%), chronic liver disease (4%), chronic kidney disease (9%), and kidney stones (1%).

Table 4-5: Baseline characteristics of patients admitted to the hospital or who visited the emergency department (ED).

Characteristics	**Patients admitted to hospital or visited ED**		
	Total Patients	**AKI**	**No AKI**
Cohort size	924,533	25084 (3%)	899449 (97%)
Age, yr, mean (SD)			
65 to <70	192,678	2522 (1.3%)	190156 (98.7%)
70 to <80	382,989	7946 (2.1%)	375043 (97.9%)
80 to <90	274,842	10370 (3.8%)	264472 (96.2%)

>=90	74,024	4246 (5.7%)	69778 (94.3%)
Women	516,175	12139 (2.4%)	504036 (97.6%)
Year of cohort entry (index date)			
2014-2015	605,244	16689 (2.8%)	588555 (97.2%)
2015-2016	319,289	8395 (2.6%)	310894 (97.4%)
Rural residence	151,323	2097 (1.4%)	149226 (98.6%)
Long-term care	43,351	3118 (7.2%)	40233 (92.8%)
Income Quintile			
1	180,227	5466 (3%)	5466 (3%)
2	192,686	5515 (2.9%)	5515 (2.9%)
3	182,957	4909 (2.7%)	4909 (2.7%)
4	186,407	4829 (2.6%)	4829 (2.6%)
5	182,256	4365 (2.4%)	4365 (2.4%)
Comorbid conditions			
Hypertension	814,604	24209 (3%)	790395 (97%)
Diabetes	358,472	13837 (3.9%)	344635 (96.1%)
Heart failure	125,136	7623 (6.1%)	117513 (93.9%)
Coronary artery disease	239,437	8392 (3.5%)	231045 (96.5%)
Chronic liver disease	33,359	1245 (3.7%)	32114 (96.3%)
Cancer	145,286	4253 (2.9%)	141033 (97.1%)
Chronic kidney disease	86,442	7759 (9%)	78683 (91%)
Kidney stones	12,457	391 (3.1%)	12066 (96.9%)
Peripheral vascular disease	13,197	660 (5%)	12537 (95%)
Cerebrovascular disease	25,835	1180 (4.6%)	24655 (95.4%)

4.2.9 Individual Medication Analysis

We identified a total of 595 unique medications prescribed to about 1 million patients in the Ontario Drug Benefit Program database. The database includes medication name, medication dose, date filled, and route-of-administration of the prescriptions. We generated 595 binary features to record the use data for each medication and each patient. We set the value of a specific medication feature for a patient when the medication was administered to that patient in the 120 days prior to hospital presentation. When patients take a drug, it affects them differently based on body composition and metabolism. However, most physicians are not able to consider all of these factors when prescribing a medication. Thus, to investigate the association between medications and AKI, we intended to identify signals that affect a large population. If a particular signal is common in a large number of people (i.e., a population of one million patients), then the possibility of the existence of an association is very high. Our goal was to identify potential interactions that are not yet understood or perhaps known. We considered this as an information retrieval problem, such that our models were designed to discover the possible relationships between each medication and AKI. We developed a multivariable logistic regression model to predict AKI based on the demographic, comorbidity, and medication data and observed the attribute representing medication to understand the relationships between a particular medication and AKI. Logistic regression is a special type of regression technique used to predict the outcome of a binary dependent feature from one or several predictors. We developed separate regression models for each individual medication (i.e., 595 models). For each model, the regression coefficient and *p-value* of the medication attribute were analyzed to identify potential associations. It is important to mention that formal clinical studies are required to confirm such interactions. The study was designed to assist healthcare experts at the ICES-KDT program in choosing potential candidates for their future drug-safety studies.

The "glm" function in R packages was employed to implement multivariable logistic regression models (Williams et al., 1984). Model covariates included demographic features and baseline comorbidities. Thus, the formula in R included AKI as the response and comorbidities, demographics, and medication as predictor variables. The value for the "family" argument in the "glm" function was set to "binomial." We used the "summary" function to get the estimate, *p-value,* z-score, and standard error for each coefficient in the model. In addition, the "confit" function was used to compute the confidence interval and odds ratio.

In order to avoid type I error in comparing multiple independent regression models, we lowered the alpha value based on the Bonferroni correction to account for the number of comparisons being performed. We considered a Bonferroni-corrected *p-value* less than 8.4e-5 (divided 0.05 by the number of individual medications) as statistically significant for regression models with each medication.

4.2.10 Medication Combination Analysis

In order to identify the medication combinations that are associated with AKI, we first prepared the medication combinations data. Since the number of individual medications is 595, the total number of combinations is a large number. Hence, we used a data mining technique named Eclat (Agrawal et al., 1993) to select the frequent combinations that included prescription data of at least 0.07% of the total number of prescriptions. Eclat is a frequent itemset mining algorithm that uses a depth-first search to discover groups of items that frequently occur in a transaction database. An itemset that appears in at least a pre-defined number of transactions is called a frequent itemset. Each frequent medication combination was annotated with its support. The support of a medication combination was how many times it appeared in the medication database

We only included combinations of two medications in this analysis and identified 7,748 unique medication combinations. Then, we created binary features to record the presence of these combinations. We set the value of a specific combination feature for a patient when that patient had been dispensed all medications within the combination in the 120-day period before the index date. Similar to the individual medication analysis, we applied a multivariable logistic regression on each medication combination. The baseline covariates, such as demographics and comorbidities, and medication combination features were included as potential covariates in the models. We developed separate regression models for each medication combinations identified using frequent itemset mining analysis (i.e., 7,748 models). The regression coefficient and *p-value* of the medication combination attribute were analyzed to identify combinations that are associated with AKI. We then performed a stratified analysis to examine potential medication-medications interactions further. We created a subset of medication combinations based on their significance in the regression models. Statistically significant combinations were detected by filtering the regression models based on a Bonferroni-corrected alpha value, 6.5e-6 (divided 0.05 by the number of medication combinations).

Stratified analyses were conducted on each medication available in one or more combinations in the above subset. To do this, we created a list of unique medications (i.e., base medications) from the chosen subset of medication combinations. Then for each medication in the list, we identified the other medication that holds a combination with the base medication. In the next stage, we prepared two sub-cohorts. The first one includes both medications in the combination (base and other), and the second one excludes the other medication in the combination. Finally, we applied multivariable logistic regression on each sub-cohort that included the combination and/or base medication feature along with the baseline covariates. The same process was followed for each medication available on the list.

In this analysis, for each unique medication combination, we obtained two models for the sub-cohorts. In order to help us to assess how the other medication affects the outcome of the base medication, we compared the odds ratio of the combination attribute in the first model with the odds ratio of the base medication attribute in the second model. We tested the significance of all models in the stratified analysis using a Bonferroni-corrected alpha value. We calculated the percentage change in odds ratios to report the result of this analysis.

4.2.11 Tools and Technologies

SAS was used to cut and prepare the dataset because ICES administrative databases were stored in the SAS server (“SAS Enterprise BI Server,” n.d.). In addition, we used R packages (“RStudio | Open source & professional software for data science teams,” n.d.) to conduct the necessary statistical and machine learning analyses in this study. R was chosen because it 1) provides widespread support for carrying out data mining operations such as frequent itemset mining and multivariable regression, 2) is available on the ICES workstations, 3) has a rich array of libraries, 4) is platform-independent and open-source, and 5) is continuously growing and providing updates with new features.

4.3 Results

This section describes the results of the study. The results are divided into two subsections. The results of the individual medication analysis and medication combination analysis are discussed in Subsection 3.1 and 3.2, respectively.

4.3.1 Individual Medication and AKI

Some of the commonly prescribed medications in the 120 days before the ED visit were Atorvastatin Calcium (24%), Rosuvastatin Calcium (22%), Hydrochlorothiazide (20%), Amlodipine Besylate (19%), and Metformin Hcl (16%). A binary logistic regression model was fit to each medication, where demographic and comorbidity features were included as potential risk factors in the model to test the research hypothesis regarding the relationship between the likelihood of developing AKI and specific medications. Table 2 shows the full list of medications with their *p-values*, odds ratios, confidence intervals, and standard errors. The medication classes are shown in brackets with medication names. We sorted medications based on the *odds ratio* of the medication feature in each model. Out of 595 medications, 55 of them were found to be strongly associated with AKI (i.e., statistically significant after Bonferroni correction). Among these 55 medications, six of them were Diuretics, four were Beta-blockers, three of them belonged to Oral Anti-Glycemic, three of them were Prostatic Hyperplasia medications, and the rest of them belonged to 33 other medication classes.

Among demographics, age, sex, residency status, and long-term care attributes have shown statistically significant relationships with the probability of AKI. The fitted models revealed that keeping all other attributes constant, the odds of getting diagnosed with AKI for males over females varied between 1.35 to 1.38. The odds for older age groups (i.e., 80 to <90 and >=90) was higher. The odds for rural residents were 24-28% lower than the odds for urban residents. Similarly, the odds for patients in long term care were 41-45% higher. By analyzing the comorbidity attributes in the models, we identified that AKI was more likely to be associated with chronic kidney disease, hypertension, diabetes, and heart failure, and chronic liver disease. Among these attributes, chronic kidney disease, hypertension, and diabetes have shown very strong associations. The average odds ratios for chronic kidney disease, hypertension, and diabetes patients were 1.81, 1.64 and,1.41, respectively.

4.3.2 Medication Combinations and AKI

A medication combination was chosen for this analysis if it has been used by at least 700 patients during the study period using the eclat algorithm. The most frequent medication combinations were Amlodipine Besylate-and-Atorvastatin Calcium (7%), Atorvastatin Calcium-and-Metformin Hcl (6%), Atorvastatin Calcium-and-Ramipril (5%), Amlodipine Besylate-and-Hydrochlorothiazide (5%), Atorvastatin Calcium-and-Hydrochlorothiazide (5%), Metformin Hcl-and-Rosuvastatin Calcium (5%), and Hydrochlorothiazide-and-Rosuvastatin Calcium (4%).

In the next stage, we applied multivariable logistic regression on each selected combination. We filtered the combinations based on the *p-value* of the medication feature in each model. We found 78 combinations that were found to be strongly associated with AKI among 7,748 combinations. Then, we performed a stratified analysis on the strongly associated combinations and reported the percentage change in the odds ratio. We identified 37 cases where a second medication is associated with increasing the risk of developing AKI when used with another medication. Table 3 contains a filtered list of combinations with a percentage change of more than 40%.

Table 3 shows the medication names with classes, odds ratios of models with and without the second medication, and percentage change in odds ratios. In the stratified analysis, we found 16 and 27 distinct classes representing the first (Base Medication column) and second (Other Medication in Combination column) medications, respectively. The percentage change in odds ratio had increased by 80% when Indapamide was used with Clavulanic acid potassium or Amoxicillin. The combination of Allopurinol with Venlafaxine Hcl or Morphine Sulfate was associated with a possible increase in the odds of 55%. The odds of getting diagnosed with AKI increases if Alprazolam, Trandolapril, Metformin, Clonidine Hcl, Acetaminophen & Oxycodone Hcl, or Cefuroxime Axetil is used in combination with Furosemide. When Celecoxib, Pregabalin, or Atenolol was used with one of the Antipsychotic medications (Quetiapine), the average change in odds ratio was about 65%. It is interesting to note that Celecoxib (Anti-Inflammatory) was not found to be associated with AKI (Table 2) when used individually but appeared to be AKI-associated when used with Mirtazapine (Antipsychotic) or Quetiapine Fumarate (Antidepressants).

The relationship between AKI and potential covariates (i.e., demographics and comorbidities) in the combination models resembled the relationship of individual models. By analyzing the regression coefficients of the combination models, we identified patients with AKI were more likely to be men, resided in urban areas, lived in long-term care, had chronic kidney disease, hypertension, diabetes, and heart failure. AKI was less likely to be associated with income quintile, peripheral vascular disease, chronic liver disease, and cerebrovascular disease.

Table 4-6: List of the individual medications sorted based on their odds ratios.

Medication	*P-value*	Odds ratio (OR)	Std. error	OR's 95% CI
Sunitinib Malate (Antineoplastic Miscellaneous)	1.6e-09	4.59	0.25	2.72 - 7.37
Lenalidomide (Immunosuppressive Agents)	9.4e-17	3.58	0.15	2.62 - 4.79
Abiraterone Acetate (Not Identified)	1.7e-10	2.61	0.15	1.92 - 3.48
Metolazone (Diuretics)	1.3e-60	2.38	0.05	2.14 - 2.63
Cyclosporine (Immunosuppressive Agents)	4.0e-06	2.18	0.17	1.54 - 3
Megestrol Acetate (Progesteron Analogues)	2.6e-07	2.08	0.14	1.56 - 2.72
Lithium Carbonate (Antimanic Agents)	4.7e-12	2.04	0.1	1.66 - 2.48
Atropine Sulfate & Diphenoxylate Hcl (Antidiarrhea)	3.4e-10	2	0.11	1.6 - 2.46
Furosemide (Diuretics)	2.6e-133	1.93	0.02	1.87 - 2
Prochlorperazine Maleate (Antiemetics And Antinauseants)	9.1e-26	1.93	0.06	1.7 - 2.17
Spironolactone (Diuretics (Potassium-Sparing))	2.6e-112	1.87	0.03	1.77 - 1.97
Methyldopa (Centrally Acting Antiadrenergic)	4.9e-06	1.84	0.13	1.4 - 2.37
Hydralazine Hcl (Vasodilator Antihypertensive Drugs)	1.5e-26	1.76	0.05	1.58 - 1.95
Dexamethasone (Corticosteroids, Plain)	2.4e-19	1.74	0.06	1.54 - 1.96
Ondansetron Hcl (Antiemetics And Antinauseants)	9.1e-13	1.69	0.07	1.46 - 1.94
Clonidine Hcl (Centrally Acting Antiadrenergic)	3.9e-06	1.69	0.09	1.4 - 2.02
Allopurinol (Xanthine Oxidase Inhibitor)	1.2e-81	1.51	0.02	1.45 - 1.57
Linagliptin (Unclassified Therapeutic Agents)	4.1e-24	1.5	0.04	1.38 - 1.62
Loperamide (Antidiarrhea)	1.4e-09	1.47	0.06	1.29 - 1.66
Glyburide (Oral Anti-Glycemic)	1.3e-12	1.46	0.04	1.34 - 1.58
Chlorthalidone (Diuretics)	1.2e-18	1.42	0.06	1.25 - 1.59
Atenolol (Beta Blockers)	1.23e-08	1.4	0.02	1.06 - 1.47
Acetylsalicylic Acid & Dipyridamole (Adenosine Diphosphate Inhibitors)	2.9e-07	1.36	0.06	1.21 - 1.53
Olmesartan Medoxomil (Angiotensin Ii Antagonist)	7.9e-13	1.35	0.04	1.24 - 1.46
Iron Ferrous Fumarate (Iron Preparations)	1.9e-39	1.34	0.02	1.29 - 1.4
Quetiapine Fumarate (Antipsychotic Agents)	4.4e-06	1.34	0.03	1.26 - 1.43
Nortriptyline Hcl (Tricyclic Antidepressant)	7.2e-19	1.34	0.06	1.18 - 1.51
Mirtazapine (Antidepressants: Miscellaneous)	2.1e-15	1.33	0.04	1.24 - 1.43
Iron Ferrous Gluconate (Iron Preparations)	8.2e-16	1.33	0.04	1.24 - 1.43
Terazosin (Alpha Adrenergic Blocking Agents)	1.5e-07	1.33	0.05	1.19 - 1.48
Olanzapine (Antipsychotic Agents)	8.5e-07	1.33	0.06	1.18 - 1.48
Fenofibrate (Antilipemic: Fibrates)	3.6e-08	1.32	0.05	1.19 - 1.46
Carvedilol (Beta-Blockers)	5.1e-09	1.31	0.05	1.19 - 1.43
Doxazosin Mesylate (Alpha Adrenergic Blocking Agents)	6.6e-07	1.3	0.05	1.17 - 1.43
Folic Acid (Vitamin B Complex)	6.9e-09	1.28	0.04	1.18 - 1.39
Trimethoprim (Sulfonamides, Trimetroprim And Combination)	8.5e-12	1.27	0.03	1.19 - 1.36
Indapamide (Diuretics)	3.9e-08	1.26	0.03	1.19 - 1.33
Sulfamethoxazole (Anti-Bacterial Sulfonamide)	2.1e-14	1.26	0.03	1.18 - 1.35
Moxifloxacin Hcl (Fluoroquinolones)	1.8e-10	1.24	0.04	1.15 - 1.34

Nifedipine (Calcium Blockers)	3.3e-06	1.21	0.03	1.14 - 1.28
Lisinopril (Ace Inhibitors)	5.1e-06	1.21	0.04	1.12 - 1.31
Gabapentin (Gamma-Aminobutyric Acid (Gaba) Derivatives)	7.6e-09	1.2	0.03	1.13 - 1.27
Oseltamivir Phosphate (Antiviral Agents - Influenza Virus Specific)	1.1e-10	1.2	0.04	1.11 - 1.3
Metoprolol (Beta-Blockers)	6.4e-11	1.19	0.03	1.12 - 1.25
Donepezil Hcl (Cholinesterase Inhibitors)	1.9e-08	1.18	0.03	1.11 - 1.25
Gliclazide (Oral Anti-Glycemic)	3.7e-11	1.17	0.03	1.12 - 1.23
Hydrochlorothiazide (Diuretics)	1.9e-18	1.16	0.02	1.12 - 1.2
Metoprolol Tartrate (Beta-Blockers)	1.7e-21	1.16	0.02	1.11 - 1.21
Amlodipine Besylate (Calcium Blockers)	2.4e-06	1.15	0.02	1.12 - 1.19
Valsartan (Angiotensin Ii Antagonist)	3.1e-11	1.15	0.03	1.09 - 1.21
Digoxin (Digitalis Preparations)	1.85e-06	1.15	0.03	1.09 - 1.22
Bisoprolol Fumarate (Beta-Blockers)	9.9e-08	1.14	0.02	1.1 - 1.18
Senna (Cathartics and Laxatives)	1.7e-09	1.14	0.02	1.08 - 1.19
Ramipril (Ace Inhibitors)	9.7e-15	1.13	0.02	1.09 - 1.17
Metformin Hcl (Oral Anti-Glycemic)	1.8e-11	1.1	0.02	1.06 - 1.14

Table 4-7: List of the medication combinations sorted based on their percentage change in odds ratios.

Base Medication	Other Medication in Comb.	Base Odds Ratio	Comb Odds Ratio	%Chg in Odds Ratio
Indapamide (Diuretics)	Clavulanic Acid Potassium (Penicillins)	1.24	2.22	79.00
Indapamide (Diuretics)	Amoxicillin (Penicillins)	1.24	2.21	78.27
Furosemide (Diuretics)	Alprazolam (Benzodiazepine Derivatives)	1.86	3.27	75.86
Donepezil Hcl (Cholinesterase Inhibitors)	Indapamide (Diuretics)	1.16	2.00	72.77
Mirtazapine (Antidepressants: Miscellaneous)	Celecoxib (Non-Steroidal Anti-Inflammatory: Non-Asa Base)	1.31	2.27	72.41
Quetiapine Fumarate (Antipsychotic Agents)	Celecoxib (Non-Steroidal Anti-Inflammatory: Non-Asa Base)	1.32	2.26	70.79
Nortriptyline Hcl (Tricyclic Antidepressant)	Acetaminophen & Oxycodone Hcl (Analgesics And Antipyretics: Miscellaneous)	1.27	2.13	67.49
Doxazosin Mesylate (Alpha Adrenergic Blocking Agents)	Perindopril Tert.Butylamine (Ace Inhibitors)	1.22	2.02	66.03
Metoprolol Tartrate (Beta-Blockers)	Amitriptyline Hcl (Tricyclic Antidepressant)	1.15	1.90	65.96
Iron Ferrous Fumarate (Iron Preparations)	Bupropion Hcl (Antidepressants)	1.33	2.21	65.63
Nortriptyline Hcl (Tricyclic Antidepressant)	Lorazepam (Benzodiazepine Derivatives)	1.25	2.06	64.85
Furosemide (Diuretics)	Trandolapril (Ace Inhibitors)	1.86	3.01	61.77
Indapamide (Diuretics)	Donepezil Hcl (Cholinesterase Inhibitors)	1.24	2.00	61.77
Allopurinol (Xanthine Oxidase Inhibitor)	Venlafaxine Hcl (Selective Serotonin Reuptake Inhibitors - Other)	1.49	2.41	61.65
Terazosin (Alpha Adrenergic Blocking Agents)	Irbesartan (Angiotensin Ii Antagonist)	1.27	2.03	59.84
Fenofibrate (Antilipemic: Fibrates)	Candesartan Cilexetil (Angiotensin Ii Antagonist)	1.27	2.02	58.17
Terazosin (Alpha Adrenergic Blocking Agents)	Pantoprazole Sodium (Proton Pump Inhibitors)	1.25	1.97	57.23
Lithium Carbonate (Antimanic Agents)	Atorvastatin Calcium (Antilipemic: Statins)	1.84	2.86	55.79
Spironolactone (Diuretics (Potassium-Sparing))	Clonazepam (Benzodiazepine Derivatives)	1.85	2.81	52.26
Allopurinol (Xanthine Oxidase Inhibitor)	Morphine Sulfate (Narcotics: Opiate Agonists)	1.50	2.26	50.77
Iron Ferrous Fumarate (Iron Preparations)	Meloxicam (Non-Steroidal Anti-Inflammatory: Non-Asa Base)	1.33	2.01	50.75
Folic Acid (Vitamin B Complex)	Hydrochlorothiazide (Diuretics)	1.22	1.82	49.59
Dexamethasone (Corticosteroids, Plain)	Gabapentin (Gamma-Aminobutyric Acid (Gaba) Derivatives)	1.67	2.49	49.42
Quetiapine Fumarate (Antipsychotic Agents)	Pregabalin (Anticonvulsants: Miscellaneous)	1.32	1.95	47.67
Dexamethasone (Corticosteroids, Plain)	Ramipril (Ace Inhibitors)	1.69	2.48	46.97
Metoprolol (Beta-Blockers)	Omeprazole (Proton Pump Inhibitors)	1.17	1.71	46.70
Ondansetron Hcl (Antiemetics And Antinauseants)	Ranitidine Hcl (Histamine H2 Receptor Antagonist)	1.62	2.37	46.49
Furosemide (Diuretics)	Metformin (Oral Anti-Glycemics)	1.86	2.69	44.56
Quetiapine Fumarate (Antipsychotic Agents)	Atenolol (Beta-Blockers)	1.32	1.90	43.70

Iron Ferrous Fumarate (Iron Preparations)	Metformin (Oral Anti-Glycemics)	1.34	1.91	43.05
Spironolactone (Diuretics (Potassium-Sparing))	Candesartan Cilexetil (Angiotensin Ii Antagonist)	1.84	2.62	42.47
Spironolactone (Diuretics (Potassium-Sparing))	Enalapril Sodium (Ace Inhibitors)	1.86	2.61	40.83
Furosemide (Diuretics)	Acetaminophen & Oxycodone Hcl (Analgesics and Antipyretics: Miscellaneous)	1.85	2.60	40.35
Spironolactone (Diuretics (Potassium-Sparing))	Dabigatran Etexilate (Anticoagulants Miscellaneous)	1.85	2.60	40.22
Furosemide (Diuretics)	Clonidine Hcl (Centrally Acting Antiadrenergic)	1.86	2.61	40.11
Furosemide (Diuretics)	Cefuroxime Axetil (Cephalosporin)	1.86	2.61	40.02

4.4 Discussion

In this study, we demonstrated how machine learning techniques could help with the identification of potentially nephrotoxic medications using administrative health databases housed in ICES. Nephrotoxic medications are responsible for about 20% of episodes of AKI, and the rate of medication-induced nephrotoxicity leading to AKI among older patients is approximately 66% (Kohli et al., 2000; Peres and da Cunha, 2013). We have presented methods for identifying medications and medication combinations that are associated with AKI using regression and frequent itemset mining algorithms. We found that 9% of all the prescribed medications were possibly associated with AKI by analyzing the medication data of one million older patients included in our study. Our analysis identified Angiotensin II Receptor Blockers, Antibacterial Agents, Diuretics, Iron Preparations, Nonsteroidal Anti-inflammatory Drugs, and Xanthine Oxidase Inhibitors as medication classes that were significantly associated with AKI. In a recent study of the French national pharmacovigilance database, Pierson-Marchandise et al. (2017) found that the majority of cases of medication-induced AKI were related to Antibacterial Agents, Antineoplastic Agents, Diuretics, Anti-inflammatory Drugs, and agents acting on the Renin-angiotensin system (Pierson-Marchandise et al., 2017). A similar conclusion was reached by a study of nursing home residents where Ace Inhibitors, Angiotensin II receptor Blockers, Antibiotics, and Diuretics were identified as the primary medication classes responsible for developing AKI.

Our study aimed to investigate how individual medication analysis results were consistent with what has been found in previous studies. We first reviewed the results with a nephrologist and learned that most of the statistically significant medications

(Table 2) were already known to be associated with AKI, which confirmed the accuracy of our findings. We also conducted an electronic literature search to find the research papers that studied the relationships between these medications and AKI. To ensure that relevant papers were not missed in our search, we used a relatively large set of keywords. We used two sets of keywords. The first set represented the medication, and the second was concerned with AKI. For the second set, we used the following terms: "AKI", "acute kidney injury", " acute renal failure", "acute phosphate nephropathy", "acute prerenal failure", and "anuria". All the studies included in this literature search were published from 1995 till 2019. We found evidence through the literature search that confirmed the association between 38 individual medications (among 55 statistically significant medications) and AKI, which more comprehensively proved the efficacy of our study.

To explain the results of individual medication analysis, we divided the identified medications into two main groups— known and likely-confounded. The medications that belong to the first group were already known to be associated with AKI. The relationships between AKI and these medications have previously been studied in the literature. The likely-confounded group contained medications that were not yet proven to be AKI-inducing. They were used to treat conditions that are associated with AKI, included in studies with kidney function, or not studied before. There is a lack of evidence regarding the association between AKI and some of these medications, such as Prochlorperazine Maleate and Terazosin. The complete list of medications that are divided into these groups is shown in Table 4. Out of 55 medications, there were 38 medications in the known group and 17 medications in the likely-confounded group. The key finding of the individual medication analysis was the list of medications included in the likely-confounded group. These medications can be suitable candidates for clinical drug-safety studies to investigate this potential association.

Table 4-8: The list of statistically significant medications from individual analysis divided into three groups.

Known	Likely-confounded
Furosemide(Bove et al., 2018; K. M. Ho and Power, 2010)	Hydralazine Hcl(Sari, 2019)
Allopurinol (Alirezaei et al., 2017; Perez-Ruiz, 2017)	Ondansetron Hcl(Aamdal, 1992)
Amlodipine(Pierson-Marchandise et al., 2017; Saruta et al., 1995)	Lithium Carbonate(Sari, 2019)
Hydrochlorothiazide (Pierson-Marchandise et al., 2017).	Bisoprolol Fumarate(J. Liu et al., 2019)
Iron Ferrous Fumarate (Leaf and Swinkels, 2016)	Abiraterone Acetate(Neyra et al., 2015)
Spironolactone(Juncos and Juncos, 2016)	Sunitinib Malate(Jha et al., 2013)
Bisoprolol(Pierson-Marchandise et al., 2017)	Carvedilol(Dupont, 1992)
Atenolol(Fleet et al., 2014)	Donepezil Hcl(Erbayraktar et al., 2017)
Metoprolol (Fleet et al., 2014)	Acetylsalicylic Acid (Sari, 2019)
Valsartan(Lopau et al., 2001)	Mirtazapine(Dev et al., 2014)
Indapamide(Pierson-Marchandise et al., 2017).	Loperamide(Mackowski et al., 2015)
Nifedipine(Mishima et al., 2017)	Doxazosin Mesylate(Mori et al., 2001)
Iron Ferrous Gluconate(Leaf and Swinkels, 2016)	Senna(Vanderperren et al., 2005)
Quetiapine(Yamada et al., 2018)	Megestrol Acetate(Boccanfuso et al., 2000; Rammohan et al., 2005)
Gabapentin (Miller and Price, 2009)	Nortriptyline(Dawlilng et al., 1981)
Linagliptin(Nandikanti et al., 2016)	Terazosin
Glyburide (McCoy et al., 2010)	Prochlorperazine Maleate
Lenalidomide(Lipson et al., 2010)	
Trimethoprim(Pierson-Marchandise et al., 2017).	
Olmesartan Medoxomil(Georgaki-Angelaki et al., 2009)	
Ramipril(Pierson-Marchandise et al., 2017).	
Gliclazide(Pierson-Marchandise et al., 2017).	
Atropine Sulfate (Pierson-Marchandise et al., 2017).	
Folic Acid(Gupta et al., 2012)	
Chlorthalidone(Peskoe et al., 1978)	
Clonidine Hcl(Allison, 2015)	
Fenofibrate(Pierson-Marchandise et al., 2017).	
Dipyridamole(Puri et al., 2016)	
Olanzapine(Hwang et al., 2014)	
Digoxin(Pierson-Marchandise et al., 2017).	
Lisinopril(Pierson-Marchandise et al., 2017).	
Methyldopa(Perazella, 2015)	
Oseltamivir Phosphate(Watanabe et al., 2014)	
Metolazone(Sean M. Bagshaw et al., 2007; Rp, 2019; Shulenberger et al., 2016)	
Cyclosporine(Bennett, 2013)	
Dexamethasone(Jacob et al., 2015; Kumar et al., 2009)	
Moxifloxacin Hcl(Bird et al., 2013)	
Sulfamethoxazole(Pierson-Marchandise et al., 2017)	

Through the medication combination analysis, we found that out of 25 thousand patients with AKI in our dataset, about 85% were prescribed multiple medications within 120 days prior to the index date. The incidence rate of developing AKI is usually higher among patients who are prescribed multiple medications. For instance, in a study of 38,782 adverse drug reaction events, out of 1,254 reported AKI cases, about 66%

included two or more concomitantly prescribed medications (Pierson-Marchandise et al., 2017). Another study suggested that there were statistically significant associations between the duration of simultaneous medication use and the development of AKI (Chang et al., 2012). Similarly, a study of Taiwan's National Health Insurance system showed that the concurrent use of certain medication classes (such as Diuretics, Beta Blockers, Calcium Channel Blockers, Alpha Blockers, Ace Inhibitors, Digoxin, and Platelet Aggregation Inhibitors) was strongly associated with the development of AKI (Chao et al., 2015). In order to compare our findings with the existing literature, we discussed the results of medication combination analysis using medication classes since most of the previous studies presented their results this way. As shown in Table 3, some of the AKI-associated combinations are Alpha Adrenergic Blocking Agents-and-Ace Inhibitors, Corticosteroids-and-Ace Inhibitors, Diuretics-and-Ace Inhibitors, Potassium-Sparing Diuretics -and-Ace Inhibitors, Diuretics-and-Analgesics and Antipyretics, Tricyclic Antidepressant-and-Analgesics and Antipyretics, Alpha Adrenergic Blocking Agents-and-Angiotensin II Antagonist, and Antilipemic: Fibrates-and-Angiotensin II Antagonist. We have identified that using a combination of Diuretics with some specific medication classes are associated with increasing the risk of developing AKI. In line with our findings, the effect of using Diuretics with Renin Angiotensin Aldosterone System Agents, Ace inhibitors, or Penicillin on AKI has been investigated in several research studies (Adhiyaman et al., 2001; Audia et al., 2008; Fournier et al., 2014, 2012; Loboz and Shenfield, 2005; Steinhäuslin et al., 1993; Wu et al., 2014).

In order to verify the results of the medication combination analysis, we compared our findings with a recent study (Rivosecchi et al., 2016). In their study, Rivosecchi et al. identified 76 unique combinations of medication classes that were associated with AKI by assessing 2,139 citations. Overall, 73.7% of selected medication classes were categorized as very low quality, 15.8% were of low quality, and 10.5% were considered medium quality. We found that our results are consistent with the studies included in this literature review. It is important to note that there were 19 medications in our study that were not statistically significant individually but were found to be strongly associated with AKI when used with another medication (Table 2 and Table 3). There are also a few combinations of medication classes in our study, such as Antipsychotic Agents and Anti-inflammatory, Diuretics and Xanthine Oxidase Inhibitor, to name a few, which have been studied individually but there is a lack of evidence in the literature on how these combinations are associated with AKI (Dixit et al., 2010; Gois et al., 2016; Jiang et al., 2017; Karajala et al., 2009; Zhang et al., 2017). Clinical drug-safety studies need to be conducted to confirm these medication-medication interactions.

The main strength of the study presented in this paper was its exhaustive analysis of medication usage patterns of the one million hospitalized patients within a 120-day look-back window. It is noteworthy that all the patients were elderly (65 years or older), suffering from multiple diseases and taking several potentially nephrotoxic medications. We included most of the frequently prescribed medications and investigated all possible combinations among these medications in our study. Next, to assess the true impact of medications on AKI, we incorporated the patients' demographic and comorbidity features as covariates in the regression analysis. In addition, we performed a stratified analysis to investigate the synergistic effect of medication combinations on AKI, which made our study more comprehensive and unique in comparison to other studies. To our knowledge, this study introduced a novel analysis technique by integrating frequent itemset mining, regression, and stratification to identify medications and combinations that can potentially be associated with AKI.

This research also demonstrates how machine learning can be used to address a well-known problem in the medical domain. It highlights what needs to be considered when designing studies that are intended to incorporate machine learning techniques to support data-driven tasks using health administrative datasets.

4.5 Limitations

Our study has some limitations. First, our results can only be generalized to the elderly, as we only had complete medication data on those aged 65 and older. Second, our study population might have included clinically unstable patients who were admitted to the hospital or emergency department. This could be a confounding factor as clinically unstable patients are more likely to take multiple concomitant medications, increasing their chances of developing AKI. Third, our findings can only be generalized to the population of Ontario since the models were derived and validated in cohorts from hospitals in Ontario. Lastly, there could be multiple reasons for which a patient is prescribed with medication, and these reasons may lead to the development of AKI rather than the medication itself. The study was designed to assist healthcare researchers at the ICES-KDT program in identifying potential candidates for their future medication-safety studies. This is not a confirmatory analysis, and proper clinical studies are required to confirm the findings.

4.6 Conclusion

Medication-induced nephrotoxicity is one of the major causes of AKI worldwide. In the present study of the ICES database, we identify the individual medications and medication combinations that are potentially associated with AKI by applying a combination of regression and frequent itemset mining techniques to this field for the first time. We have shown that our results are consistent with previous studies throughout this paper. Although most of the medications that we identify are already known to be associated with AKI, some of them have not been thoroughly studied yet. Our findings would raise awareness to conduct clinical research on these potentially nephrotoxic medications. Attention should be directed at avoiding nephrotoxic treatments when an at-risk situation is identified to reduce the chance of patients developing AKI. This requires not only careful monitoring by prescribers but also comprehensive studies on these medications. Ongoing research in this field might provide us with more reliable methods in the detection of potentially nephrotoxic medications and their combinations, thus allowing timely intervention to prevent AKI. This research will help machine learning researchers to understand what needs to be considered when designing studies that are intended to incorporate machine learning methods to accomplish various data-driven tasks using healthcare datasets.

Chapter 5

5 Visual Analytics for Dimension Reduction and Cluster Analysis of High Dimensional Electronic Health Records

This chapter has been published as S.S. Abdullah, N. Rostamzadeh, K. Sedig, A.X. Garg, and E. McArthur, “Visual Analytics for Dimension Reduction and Cluster Analysis of High Dimensional Electronic Health Records” in the Health Section of the Informatics Journal, Volume 7; May 2020. We changed the format to match the general format of the dissertation. Figure, Table, and Section numbers specified herein are relative to the chapter number. For example, “Table 1” corresponds to Table 5-1; “Figure 1” corresponds to Figure 5-1; and “Section 1.1” corresponds to Section 5.1.1. Moreover, when the term “paper”, “research”, or "work" is used, it refers to this specific chapter.

5.1 Introduction

The increasing use of EHR-based (Electronic Health Record) systems in healthcare has resulted in generating data at an unprecedented rate in recent years (Caban and Gotz, 2015; Murdoch and Detsky, 2013). EHR data includes, but is not limited to, medical and demographic records, healthcare administrative records, and results of laboratory tests (Cowie et al., 2017). The complex, diverse, and growing information available in EHRs creates promising opportunities for the healthcare providers to drastically improve the healthcare system (Kamal, 2014; Murdoch and Detsky, 2013). It is often challenging for healthcare providers to keep pace with the large volumes of heterogeneous data stored in EHRs (Rind et al., 2019). Automated data analysis techniques based on data mining and machine learning hold great promise to fulfill the computational requirements of EHRs (Marlin et al., 2012; Wetzel, 2001). There are currently a variety of efforts underway to organize, analyze, and interpret EHRs using unsupervised machine learning techniques such as clustering (Estiri et al., 2019; Foguet-Boreu et al., 2015; Haraty et al., 2015; Khalid et al., 2018; Liao et al., 2016; Marlin et al., 2012).

Cluster analysis (CA) can be used to discover hidden patterns in EHRs by grouping entities (e.g., patients, medications) with similar features into homogenous groups (i.e., clusters) while increasing heterogeneity across different groups (Dilts et al., 1995; McLachlan, 1992). It divides data into meaningful, useful, and natural groups without prior knowledge of the labels or nature of the groupings. With the large amount of unlabeled data stored in EHRs, CA has the potential to characterize medical records into meaningful groupings. Several studies have been conducted that employ different clustering techniques to identify multimorbidity patterns (Foguet-Boreu et al., 2015), implausible clinical observations (Estiri et al., 2019), and risk factors for a disease (Doust and Walsh, 2011). Despite the effectiveness of using CA in analyzing EHRs, it suffers from a problem which has been referred to as the "curse of dimensionality". This problem arises when the dataset is high-dimensional (i.e., has more than 1000 features), a very common occurance in EHRs (Ruan et al., 2019). In such situations, the output of CA is not reproducible and meaningful since variances among data elements become sparse and large (Adachi, 2017; Ronan et al., 2016). One solution is to employ dimensionality reduction (DR) techniques that can potentially reduce the number of features to a manageable size before using CA (Mitsuhiro and Yadohisa, 2015).

DR refers to the transformation of the original high-dimensional dataset into a new dataset with reduced dimensionality without loss of much information (Siwek et al., 2013). DR techniques are developed based on the idea that most high-dimensional datasets contain overlapping information (Wilke, 2019). DR techniques can be used to improve the performance of CA by removing multicollinearity and creating a small-volume dataset. Many recent studies have combined techniques from both CA and DR to find similarity among data elements and form meaningful groups (Wenskovitch et al., 2018). Despite the fact that a combination of DR and CA can result in efficient processing time and better interpretability, a number of complicated decisions need to be made when using techniques from both families (i.e., CA and DR) (Sembiring et al., 2011). For instance, when applying CA, it is important to consider which technique and distance measure to use, which features and samples to include, and what granularity to seek (Demiralp, 2017). Similarly, one needs to determine the optimal values for the configuration parameters when using a DR technique (Halpern et al., 2012). Consequently, combining these techniques results in more complicated problems. Given a wide range of techniques for DR and CA, determining which combination of techniques of CA and DR techniques leads to the desired results is not straightforward (Wenskovitch et al., 2018). Moreover, the intermediary steps of the analysis processes of CA and DR are often hidden from users, making it difficult to choose optimal values for the configuration parameters (Yoo et al., 2012). Therefore, one of the challenges of using these techniques lies with their lack of transparency and interpretability, hence limiting their application in EHR-based systems.

In order to address this issue, analysis processes can be made accessible to users through interactive visualizations. Interactive visualizations provide users with an overview of the data while at the same time enabling them to access, restructure, and modify the amount and form of displayed information (D. A. Keim et al., 2010; Thomas and Cook, 2005). They allow exploration of the visualized data to answer user-initiated queries (Sedig and Parsons, 2013). In recent years, several EHR-based visualization systems have been developed to support healthcare providers in performing various user-driven activities (Rind et al., 2011b). Although users are often good at visually perceiving the overall structure of the data, it is difficult for them to extract meaningful patterns from visualization systems when the data is large and high-dimensional. Most of these systems can only represent a limited number of features within the data due to the limited real estate space on display devices (Aimone et al., 2013; Faisal et al., 2013; Kosara and Miksch, 2002; Lavado et al., 2018). Another issue with visualization systems is that they do not incorporate analytical processes, hence falling short in fulfilling the computational demands of EHRs. Thus, an integrated approach may be needed in which automated analysis techniques (i.e., DR and CA) and user interfaces that facilitate interaction with visualizations of data (i.e., interactive visualizations) are coupled together.

Visual analytics fuses the strengths of analysis techniques and interactive visualizations to allow users to explore data interactively, identify patterns, apply filters, and manipulate data as required to achieve their goals (Parsons et al., 2015; Saffer et al., 2004; A. F. Simpao et al., 2014). This process is more complicated than an automated internal analysis coupled with an external visual representation to show the results of the analysis. It is both data-driven and user-driven and requires re-computation when users manipulate the data through the visual interface (Ola and Sedig, 2014).

The purpose of this study is to demonstrate how visual analytics systems can be designed in a systematic way to analyze the large-scale high-dimensional data in EHRs. To this end, we present a novel system that we have developed, called VALENCIA—Visual Analytics for Cluster Analysis and Dimension Reduction of High Dimensional Electronic Health Records. VALENCIA is intended to assist healthcare providers at ICES-KDT (ICES - 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; KDT - *Kidney Dialysis and Transplantation program*), located in London, Ontario, Canada. This visual analytics system allows users to choose from multiple DR and CA techniques with different configuration parameters, combine these techniques, and compare analysis results through interactive visualizations. We demonstrate the usefulness of this system by investigating the process of analyzing the health administrative data housed at ICES to gain novel and deep insights into the data and tasks at hand while at the same time identifying the most appropriate combination of analysis techniques. While few visual analytics systems have been developed for different areas in healthcare (Choo et al., 2013; Demiralp, 2017; Klimov et al., 2015; Ninkov and Sedig, 2019; Stasko et al., 2008; Wise, 1999), VALENCIA is novel in that it integrates a number of DR and CA techniques, real-time analytics, data visualization, and human-data interaction mechanisms in a systematic way. As such, the design concepts of VALENCIA can be generalized for the development of other visual analytics systems that deal with high-dimensional datasets in other domains (e.g., insurance, finance, and bioinformatics, to name a few).

The rest of this paper is organized as follows. Section 2 provides an overview of the conceptual and terminological background to understand the design of VALENCIA. Section 3 briefly describes other visual analytics systems that are related to VALENCIA. Section 4 explains the methodology employed for the design of the proposed system by describing its structure and components. Section 5 presents a usage scenario of VALENCIA to illustrate the usefulness of the system. Finally, Section 6 discusses conclusions and some future areas of application.

5.2 Background

This section presents the necessary terminology and concepts for understanding the design of VALENCIA. First, we describe the components of visual analytics. Afterwards, we briefly describe the processes of DR and CA. Finally, the healthcare stakeholders subsection introduces intended users of the system.

5.2.1 Visual Analytics

Visual analytics combines advanced analytics techniques with visual representations to analyze, synthesize, and facilitate high-level cognitive activities while allowing users to get more involved in discourse with the data (D. A. Keim et al., 2010; Thomas and Cook, 2006). The information processing load of visual analytics is distributed between users and the main components of the system—namely, the analytics and interactive visualization engines (Cui, 2019; Jeong et al., 2015; D. Keim et al., 2010b; Ola and Sedig, 2014; Parsons and Sedig, 2014; Sedig and Parsons, 2013). The analytics engine deals with the analysis of the data and carries out most of the computational load. The interactive visualization engine incorporates visual representations to amplify human cognition when working with the data (Sears and Jacko, 2007; Sedig and Parsons, 2016).

Human cognition is limited when confronted with data-intensive tasks, especially when the data is high-dimensional and complex (Green and Maciejewski, 2013; Ola and Sedig, 2014). The analytics engine of the system incorporates techniques from different fields such as statistics, machine learning, and data mining to support human cognition in such situations. Although the analytics engine carries out the majority of the computational load of the system, users are responsible for controlling the configuration parameters and internal steps of the analysis. The main responsibility of the analytics engine is to store, pre-process, transform, and analyze the data. This process can be divided into three stages: data pre-processing, data transformation, and data analysis (Ola and Sedig, 2014). The pre-processing stage is responsible for preparing raw data from different sources, which includes procedures such as cleaning, integration, and reduction (Han et al., 2011). Next, in the transformation stage, the pre-processed data is transformed into forms suitable for analysis (Kusiak, 2001). The transformation stage includes procedures such as smoothing, aggregation, feature generation, discretization, and normalization (Han and Kamber, 2011). Finally, in the data analysis stage, various statistical and machine learning techniques are applied to the transformed data to discover hidden patterns among data items and extract implicit, novel, and useful information (Agrawal et al., 1993; Sahu et al., 2008). Most of these techniques are intended for users with significant experience and do not allow proper exploration of the intermediary steps and computed results. Visual analytics addresses these issues by incorporating interactive visualization in the human-in-the-loop process.

The interactive visualization engine provides users with the ability to change the displayed data, filter the subset of the information displayed, tune the configuration parameters of the analysis techniques, and control the intermediary steps of the analytics engine. This, in turn, sets off a chain of reactions that will result in the execution of additional data analysis processes. Despite the benefits of interactive visualizations in enhancing the cognitive needs of users, they prove inadequate when faced with problems requiring heavy computations (Ola and Sedig, 2014). Another challenge is to determine how to organize a large number of data items in visual representations, especially when the data is high-dimensional. Therefore, an integrated approach that combines data analysis with interactive visualizations through visual analytics is more suitable for a comprehensive exploration of high-dimensional EHR data (Kehrer and Hauser, 2013; Keim et al., 2008).

5.2.2 Dimension Reduction

Most of the high-dimensional EHR datasets consist of multiple correlated features that offer overlapping data (e.g., most of the diabetes patients use similar medications). DR techniques can be used on such datasets to reduce dimensions without losing much information. This has long been one of the leading research topics in statistics, data mining, and machine learning (Sorzano et al., 2014). In addition to data analysis, DR techniques have been widely used in visualization research due to their ability to represent high-dimensional datasets in a low-dimensional space (Cook et al., 2007, p. 1; Fujiwara et al., 2020; Hege et al., 2009; Xin Geng et al., 2005). For instance, it is possible to transform a high-dimensional comorbidity dataset into a dataset with reduced dimensions to represent it in a scatter plot where relative positions among coordinates indicate the pairwise relationships among the transformed dimensions.

There are many DR techniques in the literature. Each DR technique has its own set of parameters, optimization criteria, and behaviours, which in turn affects data types and tasks that the technique supports. Different DR techniques should be represented using different types of visual representations because the internal mechanisms of these techniques are dissimilar. DR techniques can be broadly categorized into two groups: supervised and unsupervised (Cunningham, 2008). Most of the unsupervised DR techniques only consider the pairwise relationships among data items. Thus, the generated lower-dimensional projection can be represented in a cartesian-coordinate-based visualization. On the other hand, supervised techniques take into account additional information about the cluster structure of the data items. Therefore, supervised DR techniques require the class labels associated with cluster structure to obtain a low-dimensional projection of the original data.

In many existing visual analytics systems, DR techniques have been used as a preprocessing step to prepare the data for traditional machine learning methods that work well with a lower number of features (Mitsuhiro and Yadohisa, 2015; Obaid et al., 2019; Yan et al., 2006). A number of DR techniques are incorporated in our proposed system to help users understand the high-dimensional EHR data better and prepare the data for CA. Since the cluster structure and/or class labels are not available at the initial stage, we only incorporate unsupervised DR techniques in VALENCIA.

5.2.3 Cluster Analysis

CA can be instrumental in retrieving the cluster structure information from the transformed dimensions. It is a machine learning method that partitions data items with similar characteristics into groups called clusters. When CA is applied on a dataset containing comorbidities data, it creates different patient groups/clusters each having similar comorbid conditions. The groups formed by CA offer valuable insights into the data. In the above example, if a patient with an unknown comorbidity profile belongs to a cluster where diabetes and hypertension are common, there is a high chance for that patient to have those conditions. Moreover, CA results can be used to create an additional categorical feature to improve the performance of the data mining methods. Furthermore, CA has the potential to add significant value to visual analytics systems by offering a visual understanding of natural groupings of data items in the dataset (Choo et al., 2013; Demiralp, 2017).

The overall goal of CA is to determine the similarity between data items. There are different ways to measure similarity. Accordingly, CA techniques can be divided into four categories: connectivity, centroid, distribution, and density techniques (Kameshwaran and Malarvizhi, 2014). When data items are placed in a data space, connectivity techniques assume that items closer to each other exhibit more similarity than items that are farther away. Centroid techniques determine the similarity of data items by measuring closeness to the centroids using an iterative approach. Distribution techniques are based on the assumption that all data items in the same cluster share a common distribution (e.g., normal, gaussian, to name a few). Finally, density-based techniques analyze the density of the data items in a data space and group different density regions into clusters.

Each CA technique has its own set of configuration parameters, optimization criteria, and behaviours, which affects its performance for different datasets. Our goal in the design of VALENCIA is to assist users explore high-dimensional EHR data from different perspectives and identify the best CA technique that fits their needs. Thus, we incorporate at least one CA technique from each category (i.e., connectivity, centroid, distribution, and density) in our proposed system.

5.2.4 Healthcare Stakeholders

For the purposes of this study, we characterize stakeholders as those people who are integrally involved in the healthcare system to provide different services, such as medical practitioners, clinical researchers, and so on. With the growth of healthcare organizations, the interrelationship among healthcare stakeholders is getting complex (Davis, 2019). Irrespective of their field of expertise, stakeholders interact with EHRs at some level to perform numerous tasks to achieve novel healthcare solutions. For instance, medical practitioners use the historic treatment plan data to forecast the progress of treatments (Soyiri and Reidpath, 2013), or clinical researchers develop frameworks to discover temporal knowledge from healthcare administrative data (Klimov et al., 2015). To support complex, data-driven tasks, EHR data require some initial analysis to allow healthcare stakeholders to get insight into the distribution of the data and understand relationships among data items. The initial analysis may include preprocessing and compression of high-dimensional data to make it ready for other machine learning and statistical methods. Because of their lack of support for interactive visualizations, particularly when dealing with high-dimensional data, it is often difficult to do the above-mentioned task with conventional data analysis systems (i.e., R, SAS, Weka, to name a few). VALENCIA is designed to assist healthcare stakeholders at the ICES-KDT program (i.e., clinicians, scientists, epidemiologists, and analysts) to be able to explore and analyze healthcare administrative data housed at ICES.

5.3 Related work

In this section, we discuss some of the available visual analytics systems. There are not too many EHR-based systems that adopt DR and/or CA techniques. Thus, we include any visual analytics systems that incorporate DR and/or CA techniques in this section. In addition, we provide a brief overview of visual analytics systems that are designed for EHRs, whether they are tied to DR/CA or not. This section is divided into four parts: ones using DR, CA, both DR and CA, and EHR.

5.3.1 DR-Based Visual Analytics Systems

GGobi19 (Cook et al., 2007) is a visual analytics system that uses a DR technique called grand tour (Asimov, 1985) to represent encoded high-dimensional data. The advantage of this technique in comparison with other DR techniques is that it supports exploration of the high-dimensional space by allowing users to continuously modify the basis vectors into which data items are mapped. However, the grand tour technique can only be used when the data is not very high-dimensional. Because of this limitation, the application of GGobi19 is restricted when dealing with a very large number of dimensions which is often the case in EHRs. Another visual analytics system that uses a DR technique (specifically, PCA) to represent high-dimensional data is iPCA (Hege et al., 2009). The use of DR on high-dimensional data often results in significant information loss. iPCA offers a solution to this problem by introducing the idea of reducing the dimensions to an intermediary size and visualizing them using parallel coordinates plot. Thus, iPCA allows exploration of reduced dimensional data without loss of much information from the original dataset. It can also help users get a better understanding of the role of the reduced dimensions by visualizing the PCA basis vectors. Praxis (Cavallo and Demiralp, 2018) is another system that allows users to change the input and output of DR techniques dynamically and observe these changes through interactive visualizations. Praxis implements PCA and a number of autoencoder-based DR techniques. TimeCluster (Ali et al., 2019) is another system that incorporates DR, deep convolutional auto-encoder, scatter plot, and time-series graph to analyze large time-series data. It allows users to compare the results of multiple DR techniques visually. Although most of these systems are designed to assist users in exploring high-dimensional data using DR, they only include a limited number of DR techniques. Moreover, some of these systems, such as GGobi19 and iPCA do not support exploration of very high-dimensional data because they visualize the features of the original data along with the results of DR.

5.3.2 CA-Based Visual Analytics Systems

The Hierarchical Clustering Explorer (HCE) (Seo and Shneiderman, 2003) allows users to explore the results of CA of gene expression data using dendrograms and heatmaps. Although it enables users to visually compare the results of CA, it only supports hierarchical clustering techniques. Similar to the HCE, Matchmaker (Lex et al., 2010) allows users to arrange and compare multiple clusters simultaneously using heatmaps and parallel coordinates. It shows raw data along with the clustering results. ClusterSculptor (Nam et al., 2007) is a visual analytics system that uses k-means as the clustering engine to aid users in the derivation of classification hierarchies. Although it allows users to tune the configuration parameters through an interactive visual interface, it does not support any other clustering techniques. iGPSe (Ding et al., 2014) is another system that is designed to visually compare the results of clustering of different expression data types using parallel sets. It allows users to investigate which features are shared between multiple clusters from two different CA techniques. Both iGPSe and HCE have interpretability problems for large datasets because of having too many crossing lines. CComViz (Zhou et al., 2009) resolves this issue by rearranging clusters and their items to minimize visual clutter between features. XCluSim (L’Yi et al., 2015) also supports the comparison of several CA results of gene expression datasets using a force-directed layout, dendrogram, and parallel sets. XCluSim offers a better understanding of the characteristics of each CA technique and its parameters along with results. Although most of the abovementioned visual analytics systems are designed to compare multiple CA results, they often suffer from lack of interpretability when dealing with large datasets. A combination of the CA with DR can resolve this issue, especially when the data is high-dimensional.

5.3.3 DR and CA-Based Visual Analytics Systems

IN-SPIRE (Wise, 1999) is a visual analytics system for processing text documents; it incorporates both CA, DR, and interactive visualizations. It uses a bag-of-words model to encode the documents as high-dimensional vectors and then applies k-means with a specific number of clusters. Although it is equipped to deal with a large amount of data, it offers only a limited number of interactions to alter the analysis techniques and their configurations. Another system that utilizes both CA and DR for analyzing documents and their entities is Jigsaw (Stasko et al., 2008). To reduce the number of keywords in the vocabulary, Jigsaw implements an automatic named-entity extraction technique. It then uses k-means to display related documents and their keywords through visualization. Similar to IN-SPIRE, Jigsaw supports a limited number of interactions and does not allow users to change the CA technique. Testbed (Choo et al., 2013) is another system that addresses these limitations by incorporating seventeen DR and four CA techniques to analyze large-scale high-dimensional datasets. It allows users to apply any combinations of these techniques to visually compare their results. Another system for interactive exploration of high-dimensional data is Clustrophile (Demiralp, 2017); this system incorporates six DR and two CA techniques. It allows users to tune different configuration parameters and observe the changes through several interactive visualizations such as a heatmap and a scatter plot. Despite the advantages, both Testbed and Clustrophile allow users to apply clustering on the original dataset, which can be very high-dimensional. Some CA and visualization techniques may not perform well in those situations due to the “curse of dimensionality”.

5.3.4 EHR-Based Visual Analytics Systems

MatrixFlow (Perer and Sun, 2012) is a visual analytics system that assists users in discovering subtle temporal patterns across patient cohorts stored in EHRs. It integrates an advanced network modeling framework (i.e., Orion (Heer and Perer, 2014)) with interactive visualizations to represent networks of clinical events as a temporal flow of matrices. Another visual analytics system is VisualDecisionLinc (Mane et al., 2012) that facilitates the interpretation of large amounts of clinical data by providing overviews of treatment options and patient outcomes in an interactive dashboard. It enables clinicians to identify patient subpopulations that share similar medical characteristics to help them in the decision-making process. Simpao et al. (Simpao et al., 2015) developed a dashboard to facilitate the monitoring of medication alerts in EHRs to reduce irrelevant alerts and improve medication safety. It assists clinicians in exploring not only medication alerts but also alert types and patient characteristics. Visual Temporal Analysis Laboratory (ViTA-Lab) (Klimov et al., 2015) is an interactive and data-driven framework that is designed for the investigation of temporal clinical data. It combines query-driven visualizations with longitudinal data mining techniques to assist users in discovering temporal patterns within time-oriented clinical data. Another visual analytics system is Care Pathway Explorer (Perer et al., 2015) that enables users to discover common clinical event sequences and helps them to study how these event sequences are associated with patient outcomes. In order to achieve this, it integrates a frequent sequence mining technique with an interactive user interface. PHENOTREE (Baytas et al., 2016) allows interactive exploration of patient cohorts and interpretation of hierarchical phenotypes by integrating sparse principal component analysis with an interactive visual interface. VISA_M3R3 (Abdullah et al., 2020) is a recent visual analytics system that incorporates multiple regression, frequent itemset mining, and interactive visualization to assist users in the identification of nephrotoxic medications using EHRs. Although most of these systems incorporate complex visualization and enable users to interactively explore EHR data, they only include a limited number of analytics techniques. Moreover, some of these systems do not allow users to access and modify the analytics engine through visualization, which is an essential aspect of visual analytics.

5.4 Methods

This section describes the methodology we have employed to design the proposed visual analytics system, namely VALENCIA. In Section 4.1, we provide an overview of the design process and participants. We then describe task analysis and design criteria in

Sections 4.2. Then, in Section 4.3, we introduce the components of VALENCIA and briefly describe how the overall system works, also discussed more extensively in Section 4.4, 4.5, and 4.6. Finally, Section 4.7 outlines the implementation details of VALENCIA.

5.4.1 Design Process and Participants

Healthcare stakeholders usually deal with both well- and ill-defined tasks to solve various research problems. The well-defined tasks have clear expected solutions, specific goals, and, oftentimes, a single solution path. Unlike well-defined tasks, ill-defined tasks do not have a solution path (Varga and Varga, 2016). To help us understand how healthcare stakeholders perform real-world tasks, and to help us conceptualize and design VALENCIA, we adopted a participatory design approach. It is a co-operative approach that involves all stakeholders in the design process to ensure the output meets their requirements (Leighton, 2004). The system was primarily designed to assist the healthcare experts at the ICES-KDT program located in London, Ontario, Canada. A clinician-scientist, an epidemiologist, a data scientist, and two computer scientists were involved in the conceptualization, design, and evaluation process. They were from the computer science and epidemiology department of Western University. Participants were identified and contacted through the ICES-KDT. During the primary stage of the design process, we discerned that exploring EHR through DR, CA, and interactive visualization is not a straightforward task. It is often difficult to understand which analytics technique produces the desired result for a given dataset, which visualization technique is more suitable for the analysis results, or which interaction techniques are more appropriate to meet the requirements of the user. It becomes an ill-defined problem when analytics and interactive visualizations are combined in a VA system. In order to make appropriate design decisions, we interviewed healthcare experts in our team (i.e., a clinician-scientist and epidemiologist) to understand 1) data-driven tasks they perform with EHRs 2) analytics techniques they rely on to accomplish those tasks, and 3) visualizations with which they are familiar. We negotiated with healthcare experts the possibility of using several semi-structured interviews, which allowed new concepts to be brought up during the process. We conducted these interviews in person at the ICES-KDT center. Typical stakeholders of the system are involved in assessing and suggesting features towards similar systems regularly. In our collaboration with experts, we first finalized the analytics techniques that could allow them to accomplish data-driven tasks they would like to perform with the system. We then created several horizontal prototypes to narrow down the visualization design possibilities and selected appropriate visualization techniques for the data, analytics, and users. We performed formative evaluations continuously at every stage of the design and development process. This process was essential to build trust between the proposed system and its end-users.

5.4.2 Task Analysis and Design Criteria

In our collaboration with the healthcare stakeholders, we recognized four high-level tasks to consider in designing VALENCIA.

5.4.2.1 Displaying an Overview of the Data

Users would like to explore the features of the dataset so that they can decide which features to incorporate in the analysis. For instance, they would like to see frequencies of distinct categories for the categorical variables. Since some analysis techniques work best with specific data types, it is important to understand the characteristics of the features and their distributions.

5.4.2.2 Allowing Iteration Over DR Techniques

Choosing the appropriate DR technique is not a straightforward task. Users have to make several decisions such as which technique to use, which values for the configuration parameters are appropriate, and how many transformed dimensions to retain, to name a few. After the initial selection, users would like to refine their decisions in an iterative manner.

5.4.2.3 Allowing Iteration Over CA Techniques

Users would like to explore the data using different CA techniques with various parameter settings. They want to investigate how the clusters are formed and verify the results. Users would like to refine their decisions by going through the CA process iteratively.

5.4.2.4 Facilitating Reasoning about DR and CA

Users often would like to understand which features of the dataset are affecting the transformed dimensions, which dimensions are essential in identifying a given cluster, and how different selections of features, dimensions, techniques, and/or parameters influence the results. Since clustering is performed on the transformed data, users would like to know the summary statistics of different features and identify which feature groups or features are more important within each cluster.

5.4.3 Workflow

As shown in Figure 1, VALENCIA has two modules: the analytics engine and the interactive visualization engine. The analytics engine is composed of two components: 1) DR engine and 2) CA engine. The interactive visualization engine is composed of two views: 1) DR view, and 2) CA view. The DR view has four subviews: 1) raw-data subview, 2) projected-features subview, 3) association subview, and 4) variance subview; it supports eight interactions: selecting, drilling, filtering, annotating, arranging, searching, and transforming. The CA view is composed of three subviews: 1) hierarchical subview, 2) frequency subview, and 3) projected-observation subview; it supports six interactions: selecting, drilling, filtering, arranging, searching, and transforming.

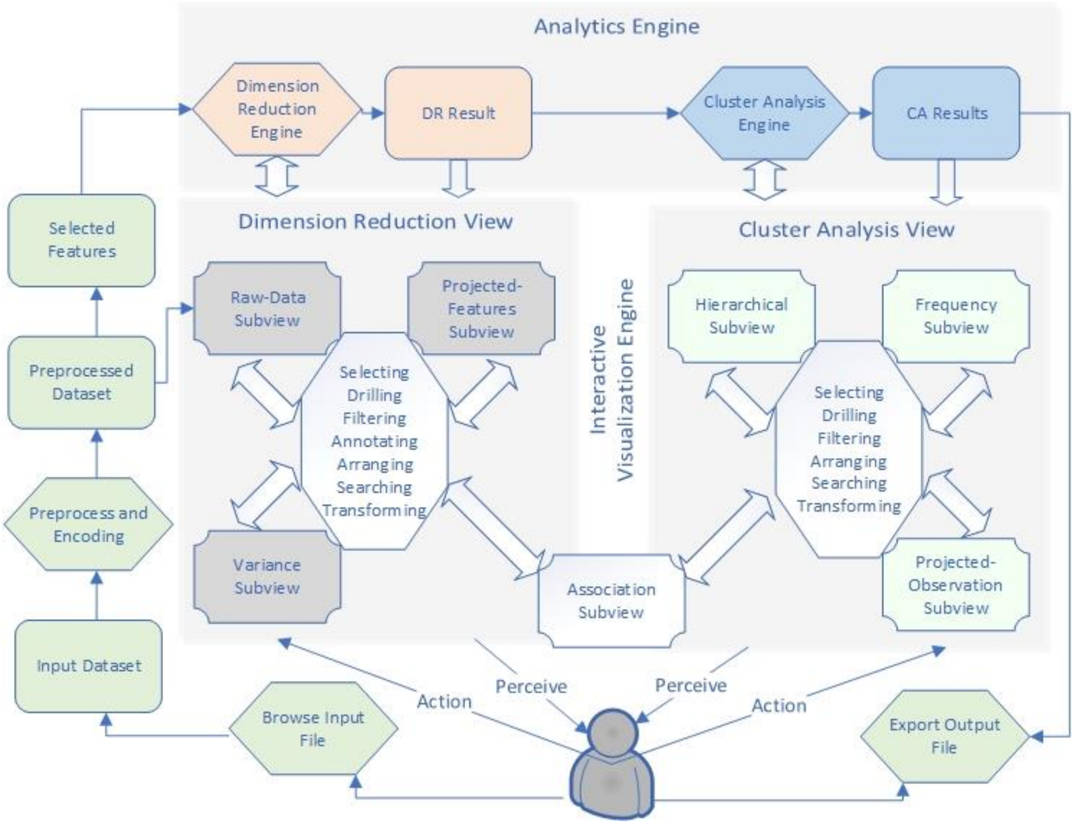


Figure 5-29: Basic workflow of VALENCIA. The backgrounds of the components are color-coded to show the similarity between processes.

The basic workflow of VALENCIA is as follows. Once the data is loaded, it gets preprocessed and encoded via the default encoding scheme. Users can then interactively explore the dataset through the raw-data subview to choose their features of interest. Next, upon selection of the DR technique and configuration parameters, the subset of the data containing the chosen features is analyzed in the DR engine. The system updates the projected-features, association, and variance subviews when the data items are generated in the DR engine. Users can observe representation of the categories of different features in proximity to each other based on their values in the projected dimensions through the projected-features subview. The association subview allows users to understand which features are most significantly associated with different dimensions. Users can observe the amount of variation retained by each projected dimension from the variance subview. This subview also allows users to select the dimensions to be analyzed through the CA engine. Users can observe the hierarchical structure of the CA result through the hierarchical subview. After selecting the dimensions, when users click the submit button, they get to the CA view. Upon selection of the CA technique and configuration parameters, the CA engine generates data items to be represented in the hierarchical, frequency, and projected-observation subviews. The frequency subview displays the distribution of features in each subset of the data selected through the hierarchical subview. The projected observation subview allows users to explore the positions of the observations in the dataset with respect to the projected dimensions. The association subview is shared between both the DR and CA views; however, the data in this subview gets filtered in the CA view based on the selection through the variance subview. Finally, users can export the output of the analysis using the export button in the CA view.

5.4.4 Encoding and Preprocessing

VALENCIA accepts input files in the JSON (JavaScript Object Notation) format and enables output to be exported in the same format. It has a built-in preprocessing procedure to encode the categorical features. The system enables users to select multiple features within a group (e.g., diabetes and hypertension in comorbidities group), all features of a group (e.g., all comorbidities or medications) or all features in all groups. A collapsible tree structure is implemented to support this operation in VALENCIA. The subset of the data containing selected features are then transferred to the analytics engine for further processing.

5.4.5 Analytics Engine

The analytics engine of VALENCIA has two main components: 1) the DR engine (a sub-engine of the analytics engine) that transforms the EHR data from the high-dimensional space to a space of lower dimensions, and 2) the CA engine (a sub-engine of the analytics engine) that organizes objects in low-dimensional space into meaningful groups whose members share similar characteristics in some way. Several techniques belonging to both families are incorporated in VALENCIA. Users are able to analyze the inputted dataset using DR, CA, or a combination of both techniques. Some studies in the literature have identified several limitations of combining some specific DR and CA techniques (e.g., (Arabie, 1994; De Soete and Carroll, 1994; Mitsuhiro and Yadohisa, 2015; Rocci et al., 2011; Timmerman et al., 2010; Vichi and Kiers, 2001)). For instance, DR techniques that rely on probability distribution (e.g., t-distributed stochastic neighbour embeddings) are not suitable for distance or density-based CA techniques. VALENCIA overcomes these limitations by providing users with the ability to choose a combination from a number of DR and CA techniques and verify the results of the analysis with both original and low-dimensional data through interactive visualizations. This analysis process is iterative, which allows users to go through any number of combinations until an optimal solution is found.

5.4.5.1 DR Engine

In analytical activities, it is often challenging for users to choose a DR technique among an abundance of available algorithms. There is no single solution to the problem of recognizing which technique is appropriate for a particular dataset. The choice of a DR technique primarily depends on the nature of the data. It also depends on the domain knowledge of users and the problem at hand. Linear DR techniques such as correspondence analysis (Hirschfeld, 1935), classical multidimensional scaling (CMDS) (Torgerson, 1958), principal component analysis (PCA) (F.R.S, 1901; Hotelling, 1933), multiple correspondence analysis (MCA) (Greenacre and Blasius, 2006), and multiple factor analysis (MFA) (Escofier and Pagès, 1994) are better at representing the global structure of the data. On the other hand, nonlinear techniques such as t-Stochastic neighbour embedding techniques (t-SNE) (Maaten and Hinton, 2008) and nonmetric multidimensional scaling (NMDS) (Kruskal, 1964; Shepard, 1962) are better at representing and preserving local interactions. VALENCIA incorporates eight linear and nonlinear DR techniques to allow users to analyze high-dimensional EHR data. Some of the well-known DR techniques that are implemented in VALENCIA include PCA, MCA, MFA, and t-SNE.

PCA uses variance to obtain principal components (i.e., orthogonal vectors) in the feature space that accounts for the maximum variance in the data. Although PCA is originally designed for continuous features, a special version of PCA, categorical principal component analysis (princals), can be used for categorical features (De Leeuw, 2005; Gifi, 1990). VALENCIA uses R libraries "PCA" and "princals" to implement PCA. On the other hand, MCA is a correspondence analysis technique for compressing and visualizing datasets with multiple categorical features. It is a generalization of PCA when the features to be analyzed are categorical instead of continuous (Abdi and Williams, 2010). To implement MCA, VALENCIA uses the "MCA" function from the "FactoMineR" package in R. In addition, MFA is a multivariate analysis technique to summarize or visualize complex datasets where observations are described by multiple sets of features structured into different groups. The distance between observations is defined based on the contribution of all active groups. To implement this technique, we use the "MFA" function in the "FactoMineR" package in R.

Unlike PCA, t-SNE is a nonlinear dimensionality reduction technique that can deal with more complex patterns in multidimensional space (Maaten and Hinton, 2008). It relies on the probability distribution of observations in the high-dimensional space to calculate the probability in the corresponding low-dimensional space. This technique is implemented using the "Rtsne" package in VALENCIA. NMDS is another nonlinear dimensionality reduction technique that uses rank-orders to collapse data from high-dimensional space into a limited number of dimensions. VALENCIA uses the "vegan" package to implement NMDS.

Determining the suitable number of new dimensions in the lower-dimensional space is a challenging task. The optimal number of dimensions to keep for CA mainly depends on the dataset. Users are often interested in particular signals in the dataset, and the choice of dimensions also depends on whether the signal of interest is captured within the dimensions in the reduced space. Thus, choosing the appropriate dimensions is crucial in VALENCIA because the DR engine is used to prepare the data for CA. It is also important to reduce the number of dimensions to an appropriate size because of the limitation of the screen space, especially when users want to visually explore the high-dimensional data. For instance, in the case of principal component analysis with a high-dimensional dataset, the first two or three principal components may describe a small fraction of variance of the dataset and/or may not capture the variation of interest (i.e., the signal of interest can be a confounding factor). In those situations, users may need to explore higher-order components through visualization and select a combination of low- and higher-order components to preserve the desired variance. VALENCIA allows users to explore the projected dimensions produced through different DR techniques using interactive visualizations. Users have the ability to adjust not only the number of dimensions but also different configuration parameters of a particular DR technique. It is important for users to find the optimal values of configuration parameters to get their desired results from the DR engine. Some arguments are adjusted automatically by the system based on the type of features in the dataset. The data items for the visual representations are produced based on the values of different arguments in the DR engine.

5.4.5.2 CA Engine

It is often difficult to interpret and visualize the results of CA when the data is high-dimensional. To address this issue, VALENCIA employs DR techniques to lower the dimension from possibly thousands to a manageable size, making it possible not only to apply different CA techniques on the projected data but also to incorporate different visualization techniques. It also offers the flexibility of analyzing a dataset containing mixed features because some CA techniques might not work well in such situations (Mitsuhiro and Yadohisa, 2015). Similar to DR, there is no single CA technique that suits every dataset and/or problem. Moreover, the configuration parameters of CA techniques need to be adjusted for different problems to find an optimal solution. There are several CA algorithms in the literature, and new algorithms are often introduced to solve different problems. Many of these algorithms are problem- and data-specific. Since there is no globally optimal CA technique, VALENCIA incorporates a number of CA techniques from different methods (i.e., connectivity, centroid, distribution, and density) that work together with a number of DR techniques. Through the integration of DR and CA, it allows users to identify patterns and groups in low-dimensional space and discover knowledge in multidimensional data.

One of the widely used CA techniques is k-means (Hartigan and Wong, 1979; Jain, 2010), a centroid-based method that partitions the data into clusters. It defines clusters in such a way that the total within-cluster (i.e., intra-cluster) variation is minimized. In general, this algorithm first selects k observations as initial centers or centroids from the dataset. Then, all remaining observations are assigned to their closest centroid using a distance function. Next, the new mean value of each cluster and its centroid are calculated. All the observations are reassigned based on the updated cluster means. These steps are repeated until convergence is achieved. To implement this technique in VALENCIA, we use the "kmeans" function in the "stats" package in R.

Unlike k-means, hierarchical clustering (Nielsen, 2016) does not require users to specify the number of clusters initially. It comes in two forms: agglomerative and divisive (Rokach and Maimon, 2005). Agglomerative clustering works in a "bottom-up" manner. Observations are initially considered as single clusters, and similar clusters are then combined to create new clusters with multiple observations. This process is repeated until all observations are grouped in a single cluster. On the contrary, divisive clustering works in a "top-down" manner where observations are combined or divided based on a similarity measure. We use the "dist()" function in R to compute distances between observations. Agglomerative and divisive techniques are implemented using "hclust()” in “stats" and “diana()” in “cluster” packages, respectively, to generate hierarchical trees in VALENCIA.

The density-based clustering (Ester et al., 1996) can be used to identify clusters of different sizes and shapes from the data. Each cluster must contain a minimum number of observations. It seeks the regions in the data space that have a high density of observations, which are separated by low-density regions. VALENCIA uses the "dbscan" function in the "fpc" package in R to provide support for density-based clustering. Users can define the radius of the neighbourhood around an observation by choosing "eps" argument and the minimum number of observations within a specified radius using "MinPts" argument.

Model-based clustering (Fraley and Raftery, 2002) assumes that the data is generated by an original model and tries to recover that model based on certain criteria. The recovered model is then used to define the clusters. Unlike other techniques mentioned above, model-based techniques implement a soft assignment, where each observation is assigned with a probability of belonging to a cluster. One of the well-known criteria to determine the model parameters is maximum likelihood. VALENCIA uses the "mclust" package in R to provide support for model-based clustering. This package uses maximum likelihood to fit different models, which can be compared based on their Bayesian information criterion score.

There are several ways to assess the quality of CA, each of which has limitations relating to the subjective quality of individual evaluations (Feldman and Sanger, 2007). VALENCIA allows users to develop a feedback loop with the system through a series of interactions. Users adjust different configuration parameters to observe their effects on features of interest to evaluate the performance of a particular CA technique. In order to find the optimal CA technique for a dataset, users can try several configuration settings. For example, when working with k-means, the "centers" argument can be modified to control the number of initial cluster centroids, and "iter.max" can be tuned to regulate the maximum number of iterations. While users have the flexibility to adjust some arguments, many arguments are adjusted automatically by the system. Despite the ubiquitous use of DR and CA techniques in the literature, their combination can be difficult to interpret, especially in relation to the features of the original dataset. To overcome this issue, the data items produced through the CA engine are made available to users through a number of visualizations. These visualizations represent the distribution of clustered observations in both high- and low-dimensional space, allowing users to verify the results and avoid misinterpretation.

5.4.6 Interactive Visualization Engine

VALENCIA is composed of two main views: DR and CA. The DR view is composed of 4 subviews: raw-data, projected-features, association, and variance. The CA view is composed of 3 subviews: hierarchical, frequency, and projected-observations. These views are supported by several selection controls, such as collapsible tree structure, drop-down menu, search bar, and checkbox. Each of these views represents an important aspect of the analytics engine. In this section, we describe how data items generated in the analytics engine are mapped onto visual representations to allow healthcare stakeholders to achieve the tasks mentioned in Section 4.2.

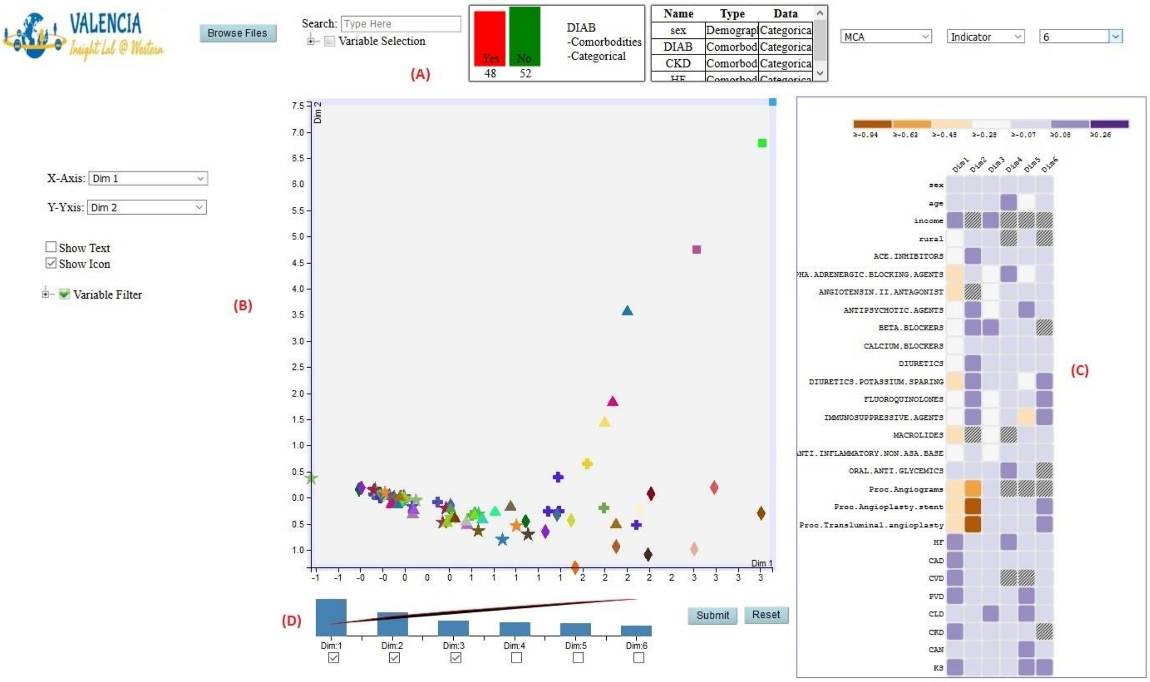
5.4.6.1 DR View

The components in the DR view allow healthcare stakeholders to import raw data, explore features, select features of interest, apply DR techniques, adjust configuration parameters, analyze DR results, and generate data items for the CA engine. This section describes four main subviews of the DR view (Figure 2).

5.4.6.1.1 Raw-data Subview

The raw-data subview is composed of a collapsible tree structure, bar chart, and data table. Upon selection of an input file, VALENCIA maps the hierarchical features of the preprocessed data into a collapsible tree structure. Users can expand the tree structure multiple times by clicking on the “+” icon in each level; this reveals groupings of the features in that level. The lowest level of the tree contains the actual feature names.

The grouping or feature name at each level of the tree structure has a checkbox, allowing users to select not only a specific feature but also a group of features. The list of selected features and relevant information is shown in a data table. Moreover, users can hover the mouse over any feature in the tree structure to see the distribution of that feature through a bar chart. The data table and bar chart are on the right side of the tree structure, as seen in the top-middle section of Figure 2.



Name	Type	Data
sex	Demograp	Categorica
DIAB	Comorbod	Categorica
CKD	Comorbod	Categorica
HF	Comorbod	Categorica

Figure 5-30: The DR view containing (A) raw-data subview, (B) projected-features subview, (C) association subview, and (D) variance subview.

5.4.6.1.2 Projected-Features Subview

The projected-features subview includes a scatter plot, collapsible tree structure, search bar, and several drop-down menus. Initially, users select a DR technique, relevant configuration parameters, and the number of projected dimensions to engage with the DR engine. Upon these selections, the coordinates of the chosen features (selected through the raw-data subview) are mapped onto a scatter plot. The scatter plot displays glyphs representing categories of each feature in proximity to each other based on their values in the projected dimensions. All the categories of a specific feature are encoded with the same color and all the features belonging to the same group are represented by a specific shape (e.g., triangle, rectangle, star, to name a few). Each category can also be represented by its corresponding label. Both the glyph and label can be turned on/off via two separate checkboxes. In the scatter plot, a linear scale is used for both horizontal and vertical axes to represent the selected dimensions. Users can interactively adjust the dimensions corresponding to the axes via two drop-down menus. The displayed information in the scatter plot can be filtered using a collapsible tree structure. This tree structure shows the list of chosen features through the raw-data subview. It allows users to select features of interest to observe their positions in the scatter plot. The tree structure is accompanied by a search bar that enables users to look for a specific group and/or feature.

Users can click on a glyph representing a category of a specific feature to observe the position of other glyphs and labels belonging to that feature. This interaction filters out all other glyphs to make it easy for users to investigate the feature of interest. Users can drill the glyphs for additional information by hovering the mouse over them. It is sometimes difficult for users to distinguish between glyphs when they are densely clustered in the scatter plot. In order to address this issue, VALENCIA provides scrolling to allow users to zoom in/out on the scatter plot. While zooming, users may wish to see glyphs that are not visible in the visual representation of the scatter plot. In such situations, users can navigate through the scatter plot by selecting any region within the representation (with the mouse) and dragging it to the desired location. These interactions are useful for exploring high-dimensional and heavily-categorized datasets.

5.4.6.1.3 Association Subview

Once the DR technique is applied, the correlation coefficient between each feature and projected dimension is shown in a heatmap in the association subview. The heatmap visualizes the magnitude and direction of the correlations through variations in coloring. It allows users to cross-examine multivariate data, through placing features in the columns and projected dimensions in the rows. Users can identify patterns by examining variance across multiple features and dimensions through this subview. They can also detect similarities between both features and dimensions and observe if any correlations exist between them. Only the significantly correlated (i.e., filtered by p-value) coefficients are included in the heatmap, leaving the unassociated cells empty. Each cell in the heatmap contains a color-coded numerical value representing the relationship between the feature and dimension in the connecting row and column. The color-coding is based on a color scale that blends from one particular color to another, to show the difference between low and high values. In order to assist users in interpreting the heatmap, a legend is included in the association subview. The legend contains a gradient scale, which is created by blending dark brown and navy blue.

Users can sort the heatmap based on either a feature or dimension by clicking on the corresponding row or column header. This allows users to observe which dimensions best represent each feature and how different features affect each dimension. Users can drill to obtain the actual value of the coefficient by hovering the mouse over the corresponding cell. Users may face difficulty while exploring this subview because of the limited screen space, especially when the dataset is high-dimensional. To address this issue, VALENCIA supports selecting any region of the subview with the mouse (left-click) and dragging it to the desired position. It also allows users to zoom in/out on the heatmap by scrolling the mouse within the region specified for this subview. These interactions make it possible for users to observe all the elements of the heatmap and investigate features of interest more closely.

5.4.6.1.4 Variance Subview

The variance subview includes a line-column chart and checkboxes that correspond to each projected dimension. The line-column chart combines a line graph and column chart by using a common x-axis. The column chart encodes each projected dimension in a vertical bar, allowing users to compare the proportion of variance retained by that dimension using the eigenvalues measure. The line chart encodes the cumulative percentage, obtained by adding the successive variances to calculate the running total. This subview supports drilling (mouse over) by displaying both actual and cumulative variance. Users can select a dimension by clicking on its corresponding checkbox. This allows users to choose a subset of projected dimensions so that it can be analyzed with the CA engine. In practice, users tend to look for a minimum number of projected dimensions that cover maximum variance in the dataset.

5.4.6.2 CA View

The components in this view allow users to apply different CA techniques, adjust configuration parameters, analyze the output, and export the final result (Figure 3). This view shares a common subview (i.e., association subview) with the DR view. The three main subviews of the CA view are described in this section.

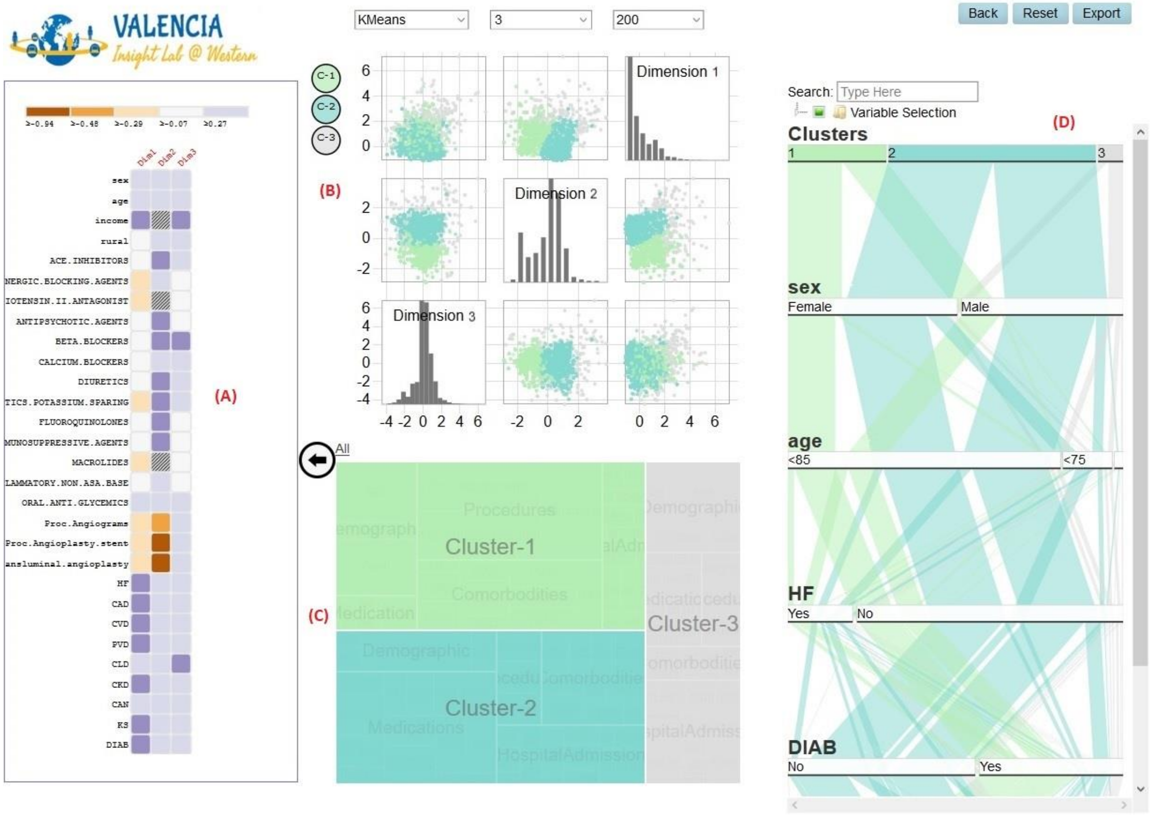


Figure 5-31: The CA view containing (A) association subview, (B) projected-observations subview (C) hierarchical subview, and (D) frequency subview.

5.4.6.2.1 Hierarchical Subview

Upon selection of a CA technique and relevant configuration parameters, the hierarchical structure of the clustered data is displayed in a zoomable treemap in the hierarchical subview. The space in the visual representation of the treemap is divided into nested rectangles. The set of rectangles in the first, second, and third levels represents clusters, groups within a particular cluster, and features within a particular group, respectively. There are several algorithms in the literature that can be used to determine the size of the rectangles in a treemap. VALENCIA determines the size of the rectangles based on the impact of each feature on a particular cluster. The algorithm to compute the size is presented in Procedure 1. For hierarchies, the size of a rectangle that contains other rectangles is determined by the sum of areas of the contained rectangles. All the rectangles representing groups and features within a cluster are encoded with the same color. VALENCIA automatically assigns colors to different clusters. The sets of rectangles in the first and second levels are transparent, showing the contained rectangles in the background. The varying sizes, colors, and nested structures of the rectangles allow users to identify patterns that would be difficult to detect otherwise.

Procedure 5-1: Compute the size of the rectangles

```
Require: Raw dataset with cluster labels
compute the number of features in each group in number_of_groupfeatures []
compute max_groupfeatures = maximum value in number_of_groupfeatures []
compute frequency of each feature in the dataset
divide the dataset based on each cluster
for each cluster C in the dataset
  for each feature F in the dataset
      compute relative frequencies of feature F in cluster C
      feature_weight = (relative frequencies / frequency [F]) * 100
      adjusted_feature_weight [C,F] = (max_groupfeatures / number_of_groupfeatures [F]) * feature_weight
return adjusted_feature_weight [][]
```

Initially, the set of rectangles belonging to the first level (i.e., clusters) is visible in the representation of the treemap. Users can navigate through the rectangles in different levels by clicking on a rectangle representing a cluster or group. The top-left corner of the treemap contains a button and navigation links. The button allows users to get back to the previous level from particular levels (i.e., second or third). The navigation links get updated dynamically as users navigate through the treemap. These links allow users to jump into any level by clicking on them. Users can hover the mouse over a rectangle to bring out the label of the corresponding rectangle. When a rectangle is hovered, it becomes highlighted (black) to help users understand which rectangle will be selected if they click on it.

5.4.6.2.2 Frequency Subview

The frequency subview includes a parallel set, collapsible tree structure, search bar, and checkbox. Parallel Sets (Kosara, 2010) is a visualization technique that is developed mainly for interpreting categorical data. For each feature or cluster, horizontal bars are displayed for possible categories in the frequency subview. The width of the bar encodes the frequency (i.e., number of matches) of that category. Starting with the first feature, each of its corresponding categories is connected to the categories of the next feature, which reveals how that category is subdivided. This subdivision process gets repeated recursively, which creates a tree of “ribbons”. The relationship between horizontal bars and ribbons helps users understand the distribution of combinations of categories. The horizontal bars and ribbons are color-coded based on the categories of the first feature. VALENCIA assigns colors to different categories automatically to make sure they are visually distinguishable.

The data items displayed in the parallel sets can be controlled through a collapsible tree structure and an interaction with the treemap in the hierarchical subview. Users can select checkboxes of features in the collapsible tree structure to include them in the parallel sets. Features are organized into groups to make them easy to find. A search bar is also included to find a specific feature. The interaction with the tree structure helps users to investigate the distribution of features of interest in different clusters. The displayed information in the parallel sets can also be controlled by selecting a rectangle representing a cluster, group, or feature in the treemap. Initially, a common set of features along with clusters are shown in the parallel sets for the entire dataset. As users interact with the treemap, the subset of data belonging to the contained rectangles in the treemap is shown in the parallel sets. For instance, if users click on a rectangle representing a cluster in the treemap, only the data items belonging to that cluster are displayed in the parallel sets. This process continues until users reach the last level in the treemap. Whenever users interact with either the tree structure or the treemap, the parallel sets gets updated based on the latest interaction.

To get additional information, users can move their mouse over the components of the parallel sets to highlight them and bring out tooltips. The tooltip of each horizontal bar displays the frequency and percentage (as a fraction of the entire dataset) of its corresponding category. When users move their mouse over a horizontal bar, all the bars and ribbons connected to that particular bar get highlighted. The tooltip of a ribbon displays the combination of criteria (categories) that the ribbon represents along with the frequency and relative percentage. When users hover over a ribbon, all other connected ribbons get highlighted. Users can drag any features and categories to reorder them. The mouse pointer changes to help users understand which components are draggable. The features and categories can be dragged vertically and horizontally, respectively. This helps users to rearrange components of the parallel set and choose which feature should be used to color the ribbons.

5.4.6.2.3 Projected-Observations Subview

Projected-observations subview includes a scatter plot matrix and histograms. The scatterplot matrix is used to show the projected observation from the DR and CA analyses. It can be seen as a collection of scatterplots organized into a matrix where each scatterplot displays the relationship between a pair of projected dimensions. While each off-diagonal cell in the matrix maps a pair of distinct dimensions, there is no logical mapping for the diagonal cells. Therefore, VALENCIA incorporates histograms in the diagonal cells of the matrix. Histograms plot the frequency of observations in each projected dimension. The observations are color-coded based on their corresponding cluster. The same color scheme is used for both the treemap, scatter plot matrix, and parallel sets. The scatter plot matrix helps users determine the linear correlation between multiple dimensions and detect patterns in the distribution of the clustered observations using projected dimensions. Users can observe each histogram to visually detect the median, outlier, and distribution (e.g., normal, skewed, to name a few) of the observations.

When users apply brushing to select a region in any scatter plot, all observations outside the brushed region get grayed out in the scatter plot matrix. This interaction helps users investigate a set of observations in the region of interest. The mouse pointer changes when users move the mouse over any region that can be brushed. Several buttons are generated to filter observations displayed in the scatter plots and histograms. The number of buttons depends on the number of clusters. Each button and its corresponding cluster share the same color to help users understand the mapping. These buttons can be turned on/off by clicking on them. Each button can be used to filter observations of its corresponding cluster.

5.4.7 Implementation Details

The VALENCIA system is implemented using standard PHP programming language, R packages, JavaScript library D3, Ajax, JavaScript library jQuery, SAS, and standard HTML. D3, jQuery and HTML were used to develop the front end of the system, which includes all the external representations (i.e., interactive visualization engine). A number of packages in R were used to develop the analytics engine of the system. Since ICES data is stored in the SAS server, we used SAS to cut the data and integrate data from different sources. The communication between analytics and visualization engines is implemented using AJAX and PHP.

We used R to develop the components of the analytics engine because it 1) offers various packages to perform DR and CA, 2) is a platform-independent open-source tool, and 3) is available in the ICES working environment.

We chose D3 to develop various external representations mainly because it 1) offers a data-driven approach to attach data to the Document Object Model elements. 2) provides users with the ability to get access to the full capabilities of modern web-browsers, 3) is an open-source library, and 4) is compatible with other programming languages that have been used in our system.

5.5 Usage Scenario

In this section, we demonstrate how VALENCIA can assist healthcare stakeholders at the ICES-KDT program in the investigation and exploration of high-dimensional EHR data. The datasets include demographics, comorbidities, hospital admission codes, medication profiles, and procedures, all linked using unique identifiers derived from health card numbers. We describe multiple scenarios to demonstrate how intended users perform numerous tasks to achieve their goals in finding appropriate DR and/or CA techniques and optimal configuration settings. Throughout this process, users get an overall understanding of relationships among data items in the EHRs.

5.5.1 Data Sources

We ascertained patient characteristics, drug prescription, and healthcare utilization data from 5 health administrative databases housed at ICES. We obtained vital statistics from the Ontario Registered Persons Database that contains demographic data on all residents of the Province of Ontario who have a valid health card. We used the Ontario Drug Benefit program database to get the prescription drug use data. This database records all outpatient prescriptions dispensed to patients aged 65 years or older, with a very low error rate (Levy et al., 2003). We ascertained hospital admission, procedure, baseline comorbidity, and emergency department visit data from the National Ambulatory Care Reporting System (i.e., ED visits) and the Canadian Institute for Health Information

Discharge Abstract Database (i.e., hospitalizations). Baseline comorbidity data were also obtained from the Ontario Health Insurance Plan database, containing claims data for physician services.

5.5.2 Cohort Creation

For this analysis, we created a cohort of patients who visited an ED or hospital between April 1st, 2014 and March 31st, 2016. The hospital admission date or ED visit date served as the cohort entry date (i.e., index date). If a patient had multiple hospital admissions or ED visits, we chose the first incident. Patient records with invalid data regarding age, sex, and health-care number were excluded from the cohort. We captured the hospital admission diagnosis and procedural information on the index date. We applied a 5-year look-back window to obtain relevant baseline comorbidity data and 120 days look-back window to obtain prescription data. We used the International Classification of Diseases, tenth revision (post-2002) codes to identify baseline comorbidities.

5.5.3 Cohort Description

There were a total of 47 unique features and about 1 million patients in the cohort. The results of the analysis are suppressed to comply with the privacy regulations for reducing the possibility of patient reidentification. Therefore, the data points shown in the projected-observation subview are suppressed in cells with five or fewer patients. The cohort includes eleven comorbidities—namely, acute kidney injury, cerebrovascular disease, chronic kidney disease, chronic liver disease, coronary artery disease, diabetes mellitus, heart failure, hypertension, kidney stones, major cancers, and peripheral vascular disease. It contains four demographics features, including age, sex, income quintile, and location. There are thirteen features representing drug classes of ACE-inhibitors, alpha-adrenergic blocking agents, angiotensin II receptor blockers, beta-blockers, calcium blockers, potassium-sparing diuretics, other diuretics, antipsychotic agents, fluoroquinolones, macrolides, immunosuppressive agents, nonsteroidal anti-inflammatory agents, and oral anti-glymetics. The cohort contains three features to represent the procedures—namely, angiograms, angioplasty stent, and transluminal angioplasty. Finally, it contains sixteen hospital admission diagnosis codes, including fluid disorders, delirium, atrial fibrillation, mycoplasma, anemia, valve disorders, femur fracture, chronic ischemia, volume depletion, paralytic ileus, chronic pulmonary, septicemia, abnormal function, hyperplasia of prostate, dementia, and glomerular disorders.

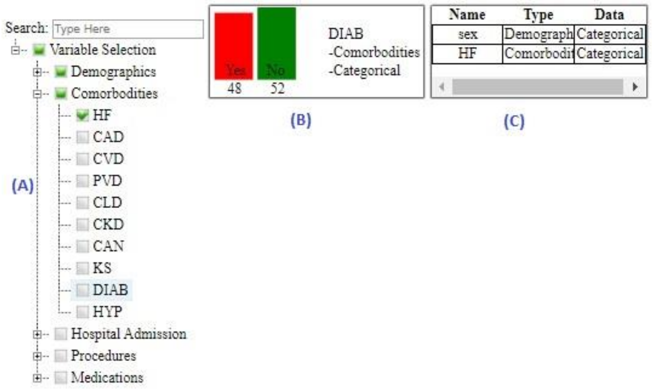
All the patients in the cohort are aged over 64 years, and the mean age is 70 years. About 56% of the patients are female, and 16% are from rural locations. The pre-existing comorbidities are hypertension (88%), diabetes (38%), coronary artery disease (25%), heart failure (14%), major cancer (16%), chronic kidney disease (9%), cerebrovascular disease (3%), peripheral vascular disease (2%), and kidney stones (1%). Some of the commonly prescribed drug classes are ace-inhibitors or angiotensin II receptor blockers (60%) and diuretics (57%). The most frequent diagnosis codes associated with AKI were chronic pulmonary (3%), atrial fibrillation (3%), chronic anaemia (2%), and ischaemic (2%).

5.5.4 Case Study

VALENCIA can be used in an iterative manner. This allows users to move freely among different stages, skipping some stages if needed, especially after going through the process of choosing a DR or CA technique once. In this study, we explain the process of using the system in a sequential manner to make it easier for readers to follow.

First, users import the data file by clicking on the “Browse Files” button in the DR view. The data file gets preprocessed by the system automatically.

Intended users can be interested in selecting a number of features from different feature groups. The imported dataset has five feature groups (i.e., demographics, comorbidities, hospital admission codes, procedures, and medications). Let us assume that a user analyzes the features using the raw-data subview and chooses fifteen features from hospital admission codes, twelve features from medications, and all features from procedures, demographics, and comorbidities through the collapsible tree structure (Figure 4-A). As shown in Figure 4-B, the user has the option to observe the description of each feature while choosing them. The selected features are displayed in a scrollable data table for verification as shown in Figure 4-C.

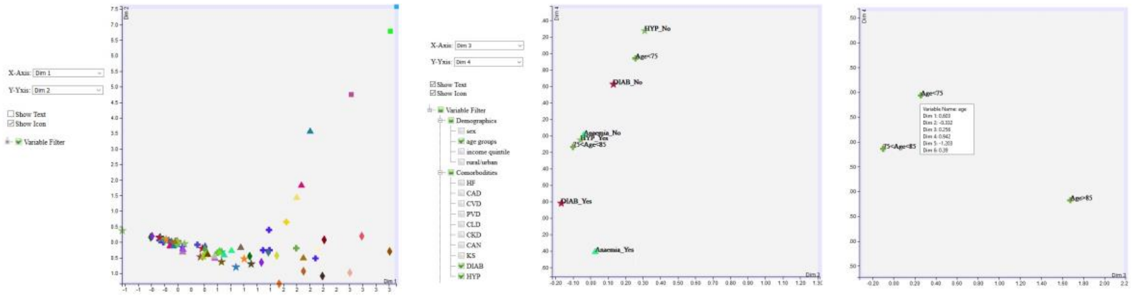


DIAB
-Comorbodities
-Categorical

Name	Type	Data
sex	Demograph	Categorical
HF	Comorbodit	Categorical

Figure 5-32: The raw-data subview containing (A) collapsible tree structure, (B) bar chart, and (C) data table.

The user has the option to choose the DR technique and set the configuration parameters for that technique. Let us assume that the user selects “MCA” as the DR technique and sets the method and number of dimensions to “indicator” and “6”, respectively. The DR engine then automatically sets indices for quantitative and categorical supplementary features. Upon these selections, the DR engine applies the selected technique with the specified configurations on the chosen features.


X-Axis: Dim 1
Y-Yxis: Dim 2
□ Show Text
☑ Show Icon

X-Axis: Dim 3
Y-Yxis: Dim 4
☑ Show Text
☑ Show Icon

Variable Name: age
Dim 1: 0.603
Dim 2: -0.332
Dim 3: 0.256
Dim 4: 0.942
Dim 5: -1.203
Dim 6: 0.39

(A) (B) (C)

Figure 5-33: Showing an overview of the projected-features subview, which includes (A) all glyphs with respect to dimensions one and two, (B) some selected glyphs and labels with regard to dimensions three and four, and (C) all the glyphs and labels representing age upon drilling.

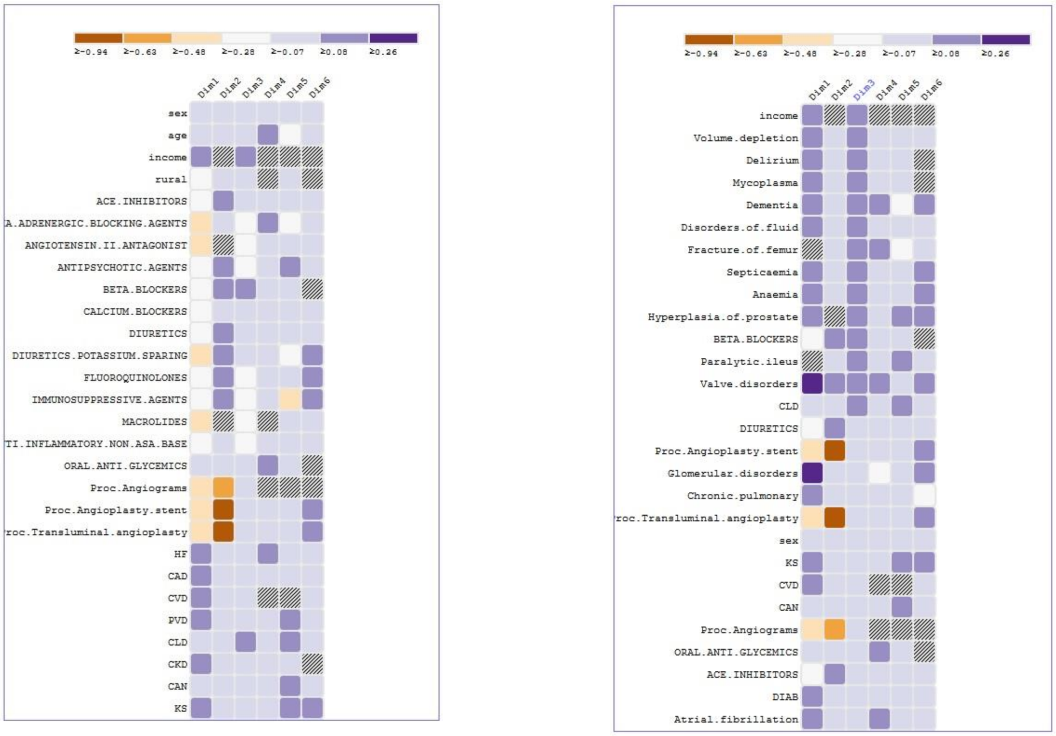
Then, VALENCIA updates the projected-features, association, and variance subviews when the data items are generated. As shown in Figure 5-A, the projected-features subview displays the coordinates of features relative to the dimensions. The first two dimensions (i.e., dimension one and two) are shown by default as axes of the scatter plot. In Figure 5-B, the user can change the X- and Y-axes from default to dimensions three and four. Initially, the scatter plot displays all the glyphs corresponding to all feature categories. As shown in Table 1, the shapes of the glyphs are chosen automatically by the system based on different groups of features such as comorbidities, demographics, and so on. The user is interested in investigating a few specific features, and thus they select age from demographics, diabetes mellitus and hypertension from comorbidities, and anemia from the hospital admission codes using the collapsible tree structure in the projected-features subview (Figure 5-B). Since the glyphs displayed in the scatter plot belong to different groups (i.e., demographics, comorbidities, and admission codes), they are encoded by different shapes and colors. However, all the categories belonging to a feature (e.g., male and female categories for feature sex) are represented by the same shape and color. The user selects the checkbox to observe the label of each glyph in Figure 5-B. They click on the glyph representing age to observe the position of other glyphs and labels (i.e., different categories of age) belonging to that feature (Figure 5-C).

Table 5-9: Showing the shapes of the glyphs based on different groups of features.

Group	**Shape**
Demographics	+ (Plus)
Comorbidities	★ (Star)
Hospital admission codes	▲ (Triangle)
Procedures	■ (Rectangle)
Medications	♦ (Diamond)

Although the chosen features in Figure 5-B contribute to the definition of dimension four, they are not well represented in dimension three. This makes the user interested in investigating which features contribute most to dimension three using the heatmap in the association subview. As shown in Figure 6-A, the heatmap displays the correlation between features and dimensions. There are six columns to represent six dimensions and 44 rows to represent the features. The positive relationships between features and dimensions are encoded with colors ranging from light blue to dark blue, whereas negative ones range from light brown to dark brown. The cells are empty when the correlation between a specific row and column is not significant (e.g., between income and dimension two). In order to find the features that are related to dimension three, the user can click on "Dim3" column header once to sort the features in a descending order. This reveals that "income", "volume depletion", "delirium", "mycoplasma", and "dementia" are positively correlated to dimension three (Figure 6-B). Then the user can select these features in the projected-features subview to investigate these correlations more closely.

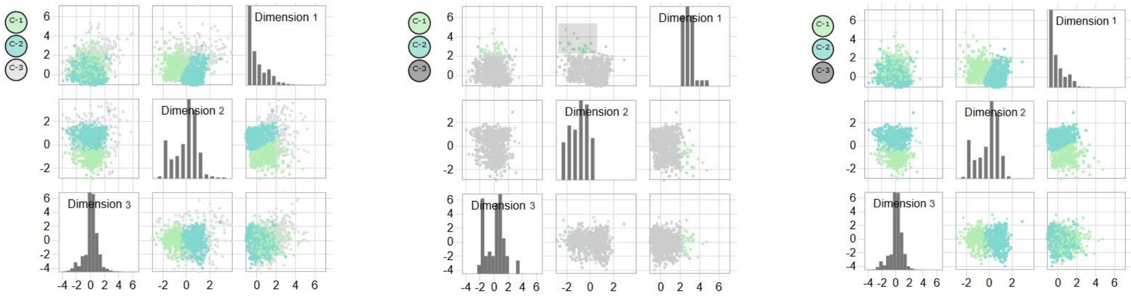
After going through the above-mentioned process iteratively, the user finalizes the number of features, DR technique, and configuration parameters. At the final stage of the DR view, the user chooses the dimensions to be included in the CA engine by observing the line-column chart in the variance subview. This helps the user to understand the amount of variation retained by each dimension. Let us assume that the user selects checkboxes for dimension one, two, and three after analyzing them thoroughly, as shown in the bottom-left corner of Figure 2. Upon clicking the "submit" button, the system takes the user to the CA view.



(A) (B)

Figure 5-34: Showing an overview of the association subview, which includes (A) a heatmap representing the association between six dimensions and 44 features and (B) a heatmap where all the features are sorted in a descending order based on dimension three.

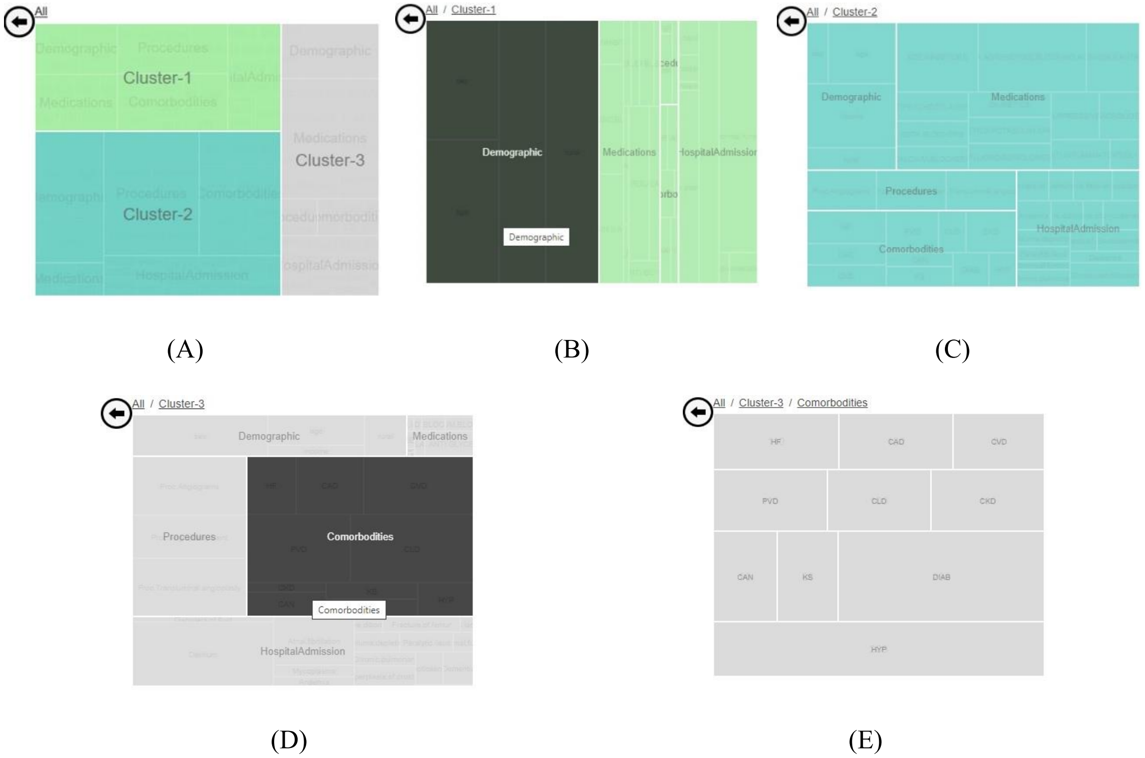
Once the CA view is loaded, the user chooses a CA technique and relevant configuration parameters to activate the CA engine. Let us assume that the user selects “kmeans” as their desired CA technique and sets the number of clusters and maximum number of iterations to “3” and “100”, respectively. Upon these selections, when the data items are generated based on the results of CA, VALENCIA updates the hierarchical, frequency, and projected-observations subviews. As shown in Figure 7-A, the projected-observations subview displays the clustered observations in the low-dimensional space. The user can verify the output of the chosen CA technique by observing the distribution of observations that are color-coded based on different clusters. For instance, the user can observe that the clusters are more distinguishable from each other in the scatter plots between dimensions one and two (Figure 7-A). In order to understand the distribution of the observations better and detect outliers, the user applies brushing on a region in a scatter plot (between dimensions one & two). This helps the user to investigate how the observations in the selected region are distributed in other scatter plots (Figure 7-B). As shown in Figure 7-C, when the user clicks on button “C-3”, the system removes all the observations belonging to cluster three.



(A) (B) (C)

Figure 5-35: Showing the overview of the projected-observation subview, which displays (A) all the observations color-coded based on clusters, (B) the brushing interaction, and (C) observations in cluster-1 and cluster-2 because cluster-3 is filtered out.

This allows the user to compare the remaining clusters more easily. If the user becomes interested in getting additional information about the dimensions (e.g., which features are associated with these dimensions), they can use the association subview. Although this subview is also available in the DR view, it is included in the CA view to allow the user to retrieve such information without switching between views. As shown in the left corner of Figure 3, the association subview within the CA view contains information of the first three dimensions based on the user’s selection in the DR view. Next, if the user is interested in exploring the hierarchical structure of the clustered data in high-dimensional space (raw data), they can refer to the hierarchical subview.



(A) (B) (C)

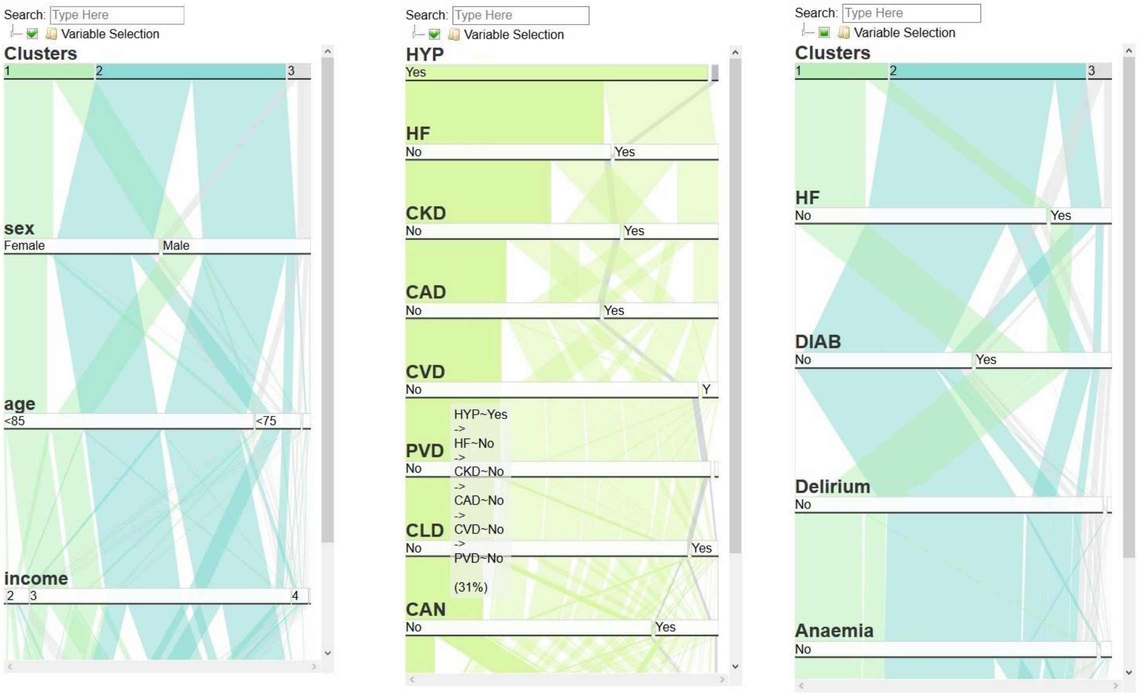
(D) (E)

Figure 5-36: Showing the overview of the hierarchical subview, which displays (A) all the clusters, (B) feature groups within cluster-1, (C) feature groups within cluster-2, (D) feature groups within cluster-3, and (E) features within the comorbidity group in cluster-3.

The hierarchical subview allows the user to detect which clusters cover the maximum amount of variation of the data (Figure 8-A). The user clicks on a rectangle representing a cluster (i.e., cluster-1, cluster-2, and cluster-3) in this subview to observe how different feature groups contribute to the variance of a particular cluster. For instance, demographics, medications, and comorbidities have the highest contributions to cluster-1, cluster-2, and cluster-3, respectively (as shown in Figure 8-B, 8-C, and 8-D). Then, the user can click on the rectangle representing comorbidities within cluster-3, which reveals that diabetes mellitus (DIAB) and hypertension (HYP) are the dominating features in this group (Figure 8-E).

The user can consult the frequency subview to get the frequency distribution of clusters and features. As shown in Figure 9-A, 32%, 59%, and 9% of the patients are assigned to cluster one, two, and three, respectively. About 82% of these patients are aged between 65 and 85, and most of them are assigned to the first two clusters. Upon interacting with the hierarchical subview, the frequency subview gets updated dynamically to allow the user to get additional information at every level. For instance, Figure 9-B displays the frequencies of the comorbidities when the user selects the “cluster-3”->“comorbidities” rectangles (Figure 8-E) in the hierarchical subview. The user can observe that 93% of the patients in cluster-3 have hypertension. In order to change the color of the ribbons based on the outcome of the heart failure feature (HF), they can reorder the horizontal bars by dragging hypertension (HYP) to the top.

The user can also observe the distribution of all other comorbidity features within cluster-3. Next, let us assume the user becomes interested in checking how the patients who have heart failure, diabetes mellitus, anemia, and delirium are subdivided into different clusters. The user can activate the collapsible tree structure by clicking on particular checkboxes corresponding to these features to filter the displayed information. Figure 9-C shows how patients in different clusters are subdivided into these features and vice versa. It is possible to explore the interrelationship between not only the clusters and features but also different features in this manner. For example, the user can observe that most of the patients belonging to cluster-1 have diabetes. In order to investigate this relationship more closely, the user can change (i.e., from clusters to feature) the ordering and color-coding by moving diabetes mellitus (DIAB) to the top. The color scheme for clusters in the frequency, hierarchical, and projected-observations subviews are identical; this makes it easy for the user to visually perceive the connection between these subviews (Figure 3).



(A) (B) (C)

Figure 5-37: Showing the overview of the frequency subview, which shows the distribution of (A) different clusters and demographics, (B) all the comorbidities within a particular cluster, and (C) clusters and some user-selected features.

At any stage of the analysis in the CA view, the user can click the "Back" button to navigate back to the DR view. They can switch between the DR and CA views as many times as is required. After going through this iterative process of applying different CA techniques, tuning configuration parameters, and analyzing results with different subviews, the user exports the resulting dataset by clicking on the "Export" button. The output dataset contains all the data elements along with cluster labels for each patient.

5.6 Conclusion

In this study, we have shown how visual analytics systems can be designed to address the challenges of high-dimensional data stored in EHRs in a systematic way. To achieve this, we have reported the development of VALENCIA, a visual analytics system designed to assist healthcare stakeholders at the ICES-KDT program. VALENCIA incorporates two main components: an analytics engine, made up of two sub-engines: the DR engine and the CA engine; and an interactive visualization engine, made up of the DR view and the CA view. The main contribution of VALENCIA is to bring a wide range of state-of-the-art and traditional analysis techniques, integrate them seamlessly, and make them accessible through interactive visualizations. VALENCIA offers a balanced distribution of processing load between users and the system through a proper integration of analytics techniques (i.e., the DR and CA engines) with visual representations (i.e., different interactive views in the interactive visualization engine) to facilitate the performance of high-level cognitive tasks. Through a real case study, we have demonstrated how VALENCIA can be used to analyze the healthcare administrative dataset of older patients who visited the hospital or emergency department in Ontario between 2014 to 2016. Through the formative evaluations conducted during the participatory design process, we have seen that VALENCIA assists healthcare experts in 1) exploring datasets using different DR and CA techniques, 2) generating hypotheses, 3) identifying relationships among data items, 4) evaluating results of the analysis, and 5) recognizing patterns and trends that would be otherwise difficult to identify without such a system. A number of training materials have been prepared to assist new users in getting familiar with the system. Users at the ICES-KDT program were able to identify suitable analysis techniques and configuration settings for their health administrative datasets. They got familiar with different analytics techniques quickly while exploring them through VALENCIA, although they never worked with those techniques before. They also have reported that the interactive visual interface makes it easy for them to explore the analysis results.

In terms of the scalability and extensibility of VALENCIA, we designed it in a modular way so that it can easily accept new data sources and analysis techniques (both DR and CA). VALENCIA can be used to analyze high-dimensional datasets in many other domains, such as insurance, biotechnology, finance, and image processing.

The study should be evaluated with respect to four limitations. The first one is that, as the size of the dataset grows, its computational time for the DR and CA techniques increases; this limits the real-time functionality of the interactive visualizations. The second limitation is that, even though we have had a participatory design and healthcare experts have evaluated VALENCIA and have found it helpful and usable, we have not conducted any formal studies to assess its performance, nor the efficiency of its human-data discourse mechanisms. Third, since the system has been designed for a healthcare organization, we have not tested the performance of the system on any other domain except healthcare. Fourth, some subviews of the system may not function properly if the number of features in the dataset gets too large due to limitations of screen space and computational resources.

Chapter 6

6 Predicting Acute Kidney Injury: A Machine Learning Approach using Electronic Health Records

This chapter is accepted for publication as S.S. Abdullah, N. Rostamzadeh, K. Sedig, A.X. Garg, and E. McArthur, “Predicting Acute Kidney Injury: A Machine Learning Approach using Electronic Health Records” in the Information Journal, July 2020. We changed the format to match the general format of the dissertation. The Figure, Table, and Section numbers specified herein are relative to the chapter number. For example, “Table 1” corresponds to Table 6-1; “Figure 1” corresponds to Figure 6-1; and “Section 1.1” corresponds to Section 6.1.1. Moreover, when the term “paper”, “research”, or "work" is used, it refers to this specific chapter.

6.1 Introduction

Acute kidney injury (AKI) is common among patients admitted to hospitals, affecting approximately 10% of hospitalized patients and more than 25% of patients in the intensive care unit (Porter et al., 2014; Selby et al., 2012). AKI is defined as an abrupt loss of kidney function over a short period of time [2]. AKI may lead to prolonged hospital stays, lower chance of survival, and a higher risk of developing chronic kidney disease. Over the last 10-15 years, the incidence rate of AKI has increased in the United States (Nadkarni et al., 2016; Wu et al., 2014), the United Kingdom (Kolhe et al., 2016), and Canada (Liu et al., 2010; Mehrabadi et al., 2014). The growing incidence rate of AKI is associated with the changing spectrum of diseases. There is an increasing body of evidence proving that patients with extrarenal complications and multiple comorbidities are at a greater risk of developing AKI (Mehta et al., 2004; Siddiqui et al., 2012). Aikar et al. (Waikar et al., 2006) have shown that the high comorbidity rate, measured by the Deyo-Charlson comorbidity index, is associated with AKI. As a patient's number of comorbid conditions grows, there is a rise in associated physician visits, healthcare utilization, medication intake, and hospitalizations (Zulman et al., 2014), ultimately leading to an increase in healthcare expenditure. Given the associated risk and expense, a promising strategy is required to improve the care for AKI patients. However, a UK-based report published in 2009 demonstrated significant under-recognition of AKI, leading to delayed recognition, inadequate treatment, and ineffective monitoring (Ali et al., 2007; Bagshaw et al., 2007).

Thus, there is a rising demand for techniques that can be used for the detection of AKI. However, the complex pathophysiology and etiology of AKI make the diagnosis and management of this disease challenging. There are different guidelines such as RIFLE (Eriksen et al., 2003), AKIN (Palevsky et al., 2013), WRF (Gottlieb et al., 2002) and KDIGO (Clinical Practice Guideline, 2012) for AKI diagnosis. Most of these guidelines rely on a rise in serum creatinine (i.e., a laboratory test) alone as the gold standard. However, serum creatinine-based guidelines are often not ideal for the diagnosis of AKI among older patients because the age-related deteriorations in glomerular filtration rates affect the baseline measure (Kate et al., 2016). Another limitation of this measurement is due to the fact that serum creatinine may vary with muscle mass since it is a product of muscle catabolism (Delanaye et al., 2017). In addition, serum creatinine-based guidelines require a premorbid serum creatinine value to be used as a baseline creatinine, which may not be available for all patients (Mohamadlou et al., 2018). Although some guidelines also rely on urine output to diagnose AKI, it is only monitored for patients with reduced kidney function (Kate et al., 2016). Despite these challenges, even if AKI can be diagnosed properly, the clinicians often fail to intervene due to a lack of time and treatment options. The treatments of AKI are primarily focused on avoiding nephrotoxic medications and administering supportive care (Clinical Practice Guideline, 2012). Although more advanced treatments are identified in recent years, their effectiveness has not been proven in clinical trials yet (Pozzoli et al., 2018). Thus, interventions often have poor performance if a patient has developed AKI already (Lieske et al., 2014; Mehta, 2011). So, it is more effective to predict AKI prior to its diagnosis. A number of recent studies have shown that AKI is predictable and avoidable if early risk factors can be identified using Electronic Health Records (EHRs). For instance, Kate et al. (2016) have revealed that it is possible to predict up to 30 percent of AKI cases in the hospital settings using the patient data stored in EHRs (Kate et al., 2016).

EHR contains patient medical records, such as comorbid conditions, medications, laboratory test results, diagnosis codes, demographics, and discharge summaries, which can be used for the risk profiling of patients (Mohamadlou et al., 2018; Rostamzadeh et al., 2020). With the evolution of EHRs and the widespread use of information technology systems, these medical records are available nowadays for subsequent reuses (Abdullah et al., 2020a, 2020b; Abramson et al., 2011; Delamarre et al., 2015). EHRs offer an opportunity to employ machine learning techniques to recognize risk factors associated with AKI and identify patients at risk of developing AKI. Several clinical decision support systems have been developed in recent years for earlier detection of AKI using machine learning techniques (Abdullah et al., 2020c; Cheng et al., 2017; Davis et al., 2017; Gameiro et al., 2020; Ibrahim et al., 2019; Rashidi et al., 2020; Tran et al., 2019). However, many of these systems suffer from various performance and design related issues such as lack of predictive power, substantial trade-offs between sensitivity and specificity, a limited number of machine learning techniques, small population size, lack of predictors, and limited patient populations (Gameiro et al., 2020; Mohamadlou et al., 2018).

This study is designed to predict AKI among hospitalized and emergency department patients using machine learning techniques. We incorporate ICES' healthcare administrative datasets containing one million older patients' medical records who visited the hospital or emergency department between 2014 and 2016. We developed 31 prediction models based on different combinations of two sampling techniques, three ensemble methods, and eight classifiers. Our study differs from other studies in several ways: (1) we developed prediction models for patients who are at risk of developing AKI within 90 days timeframe after being discharged from hospital or emergency department; (2) we included a large number of predictors to train the models; and (3) we validated the important features of each model with healthcare experts through formative evaluations to improve the performance and reliability of the models. The rest of this study is organized as follows. Section 2 describes the methodology employed for the design of the study. Section 3 presents the experimental results. Finally, Section 4 includes the discussion, and Section 5 describes the limitations of the study.

6.2 Materials and Methods

We discuss the data sources and methodology in this section that includes the design process, settings, design flow, data integration, cohort entry criteria, input features, outcomes, and proposed machine learning techniques.

6.2.1 Design Process and Participants

As a part of continuing clinical research, medical experts usually conduct clinical trials and case studies in their areas of expertise. In many situations, the result of these clinical studies is not reproducible due to limited and specific population size. Machine learning can help healthcare experts evaluate the relevance of such studies and explore more complicated relationships among data elements. Despite the advantages, one significant drawback of the machine learning approach is a general lack of interpretability. Thus, it is underexplored in clinical studies as most of the healthcare experts often find it difficult to understand these models and results (Spasic and Nenadic, 2020). On the other hand, although computer science experts are more experienced in working with machine learning techniques, they are not familiar with clinical terms. It becomes difficult for them to interpret and validate the analysis results without the help of domain experts. To address this issue, we adopted a participatory design approach to conceptualize and design our study. It is a co-operative approach that includes all stakeholders (e.g., users, designers, and evaluators) in the process to make sure the result of the analysis meets their needs (Muller, 2007). A clinician, a statistician, an epidemiologist, and several data scientists participated in the conceptualization, design and evaluation process of this study. During the primary stage of the design process, we came to know that healthcare experts perform studies to predict diseases in many different ways. There is no single correct analysis technique because different techniques have their strengths and weaknesses, and the selection of an appropriate technique for a task is not straightforward. As such, we invited healthcare experts in our team to provide us with a list of analysis techniques, which they usually practice. We decided to employ both traditional and state-of-art analysis techniques to build trust with end-users and, at the same time, allow them to explore complex relationships in the dataset.

6.2.2 Study Design and Setting

We conducted a population-based retrospective cohort study in older patients who visited a hospital or emergency department between April 1st, 2014 and March 31st, 2016, using health administrative databases stored at ICES. These datasets were connected using unique encoded identifiers and analyzed at ICES.

Ontario has a population of about 13 million residents with universal access to physician services and hospital care, which includes 1.9 million people aged 65 years or older. We suppressed the results of this study in cells with five or fewer patients to comply with ICES privacy regulations and minimize the possibility of reidentification of patients.

6.2.3 Workflow

Figure 1 shows the basic workflow of the study described in this paper. In the first step, we created an integrated dataset from five different health administrative databases. The data sources are discussed in Section 2.4. Next, we describe the inclusion and exclusion criteria in Section 2.5. The features in the comorbidity, prescription, demographic, and hospital admission codes data were encoded and transformed into suitable forms for the analysis in the preprocessing stage, which is discussed in Section 2.8. The analysis techniques and results are presented in Section 2.9 and 3, respectively.

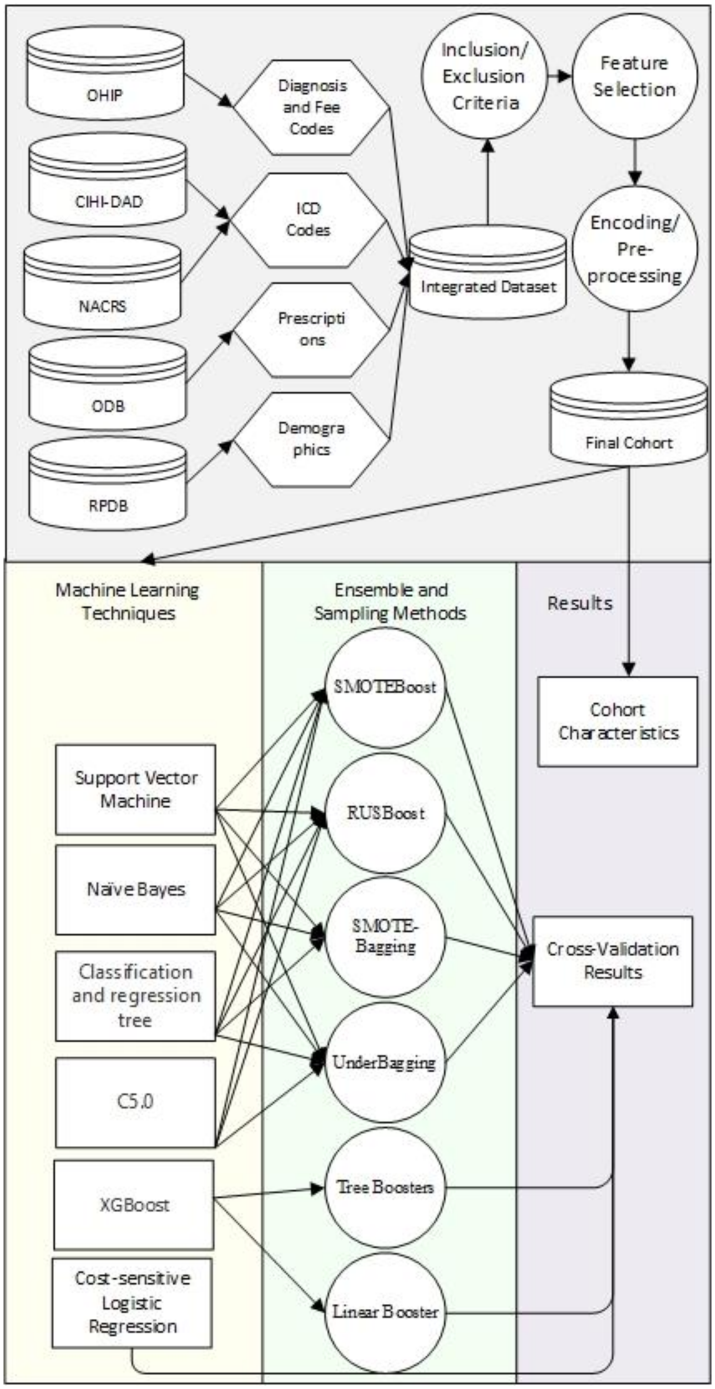


Figure 6-38: Workflow diagram of the presented study where different colours are used to represent three main parts (data integration and preprocessing, analysis and validation). The figure shows how different combinations are formed using two

sampling techniques (i.e., under sampling and SMOTE), three ensemble methods (i.e., boosting, bagging, and XGBoost), and eight machine learning classifiers.

6.2.4 Data Sources

We ascertained patient characteristics, drug prescriptions, outcome and medical history data from 5 administrative databases (as shown in Appendix A). These datasets are linked using a unique identifier, which is derived from health card numbers. We collected vital statistics from the Ontario Registered Persons Database, which includes demographic data of all residents in Ontario who have a valid health card. We utilized the Ontario Drug Benefit Program database to get prescription medication use data. Ontario Drug Benefit Program holds all the outpatient prescription records dispensed to older patients, which has an error rate of less than 1% (Levy et al., 2003). We ascertained baseline comorbidity, emergency department visit, and hospital admission data from the National Ambulatory Care Reporting System (i.e., for the emergency department) and the Canadian Institute for Health Information Discharge Abstract Database (i.e., for hospital admissions). We applied the ICD-10 (i.e., International Classification of Diseases, post-2002) codes to identify baseline comorbidities within the look-back window. In addition, Baseline comorbidity data were acquired from the Ontario Health Insurance Plan database, which holds claim records for physician services. ICES Physician Database was used to obtain the demographic, education, practice, and specialty information on all physicians. All the coding definitions for the comorbidity databases are provided in Appendix B.

6.2.5 Cohort Entry Criteria

We identified a cohort of individuals aged 65 years or older who visited the emergency department or were admitted to hospital between 2014 and 2016 (Figure 2). The hospital admission or emergency department discharge dates were taken as the cohort entry or index date. If a patient had multiple hospital admissions and emergency department visits, we chose the first incident. We excluded patients with invalid or missing age, sex, and health card number. In addition, we excluded patients who: (1) previously underwent a kidney transplant or dialysis treatment as AKI is usually no longer relevant once patients develop end-stage kidney disease; (2) left the emergency department or hospital without being seen by a physician or against medical advice; and (3) developed AKI during emergency department visit or hospital admission as they are already under observation. The diagnosis codes for the exclusion criteria are presented in Appendix C.

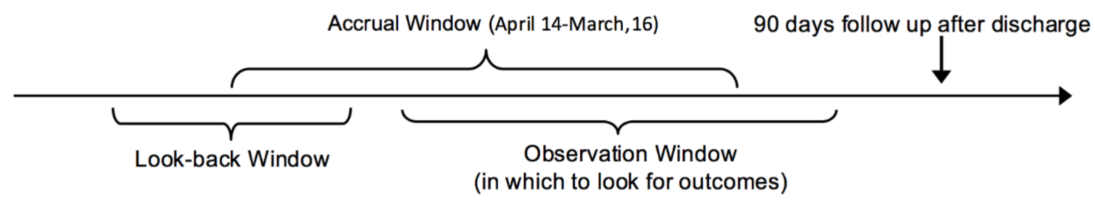

Accrual Window (April 14-March,16)
90 days follow up after discharge
Look-back Window
Observation Window
(in which to look for outcomes)

Figure 6-39: Provides an overview of data creation plan and how we prepared the final cohort.

6.2.6 Input Features

We used the Chi-Square test for feature selection and then filtered the selected features with a healthcare expert. The final cohort included about one million patients and a total of 86 unique features. The cohort contained eleven comorbidity features—namely, chronic kidney disease, diabetes mellitus, cerebrovascular disease, coronary artery disease, hypertension, chronic liver disease, major cancers, peripheral vascular disease, heart failure, and kidney stones. We applied a 5-year look-back window to detect these baseline comorbidities. There were four demographics features—namely, sex, age, region, and income quintile. We included 55 medications that were prescribed to the patients within 120 days before the first hospital admission or emergency department visit. These medications belonged to thirteen distinct drug classes—namely, ACE-inhibitors (blood pressure and heart failure), beta-blockers (blood pressure), alpha-adrenergic blocking agents (blood pressure), angiotensin-receptor blockers (blood pressure), calcium blockers (blood pressure), macrolides (antibiotics), fluoroquinolones (antibiotics), potassium-sparing diuretics (weak diuretic), other diuretics, nonsteroidal anti-inflammatory agents (pain relievers), oral hypoglycemic (diabetes mellitus), and immunosuppressive agents (immune system activity).

The cohort also included sixteen ICD-10 diagnosis codes that were identified during the index hospitalization or emergency department visit. The codes were related to delirium, mycoplasma pneumoniae, disorders of fluid, electrolyte and acid-base balance (e.g., hyperosmolality and hypernatraemia, hypo-osmolality and hyponatraemia, acidosis, alkalosis, mixed disorder of acid-base balance, hyperkalaemia, hypokalaemia, fluid overload, and other disorders of electrolyte and fluid balance), atrial fibrillation, anemia, femur fracture, valve disorders, atherosclerotic cardiovascular disease, diseases of the digestive system (e.g., paralytic ileus, intussusception, volvulus, gallstone ileus, other impaction of intestine, intestinal adhesions with obstruction, and other and unspecified intestinal obstruction ileus), Certain infectious and parasitic diseases (e.g., sepsis due to Staphylococcus aureus, other specified Staphylococcus, Haemophilus influenzae, Escherichia coli, Pseudomonas, Serratia marcescens, other gram-negative organisms, gram-negative Septicaemia, and Enterococcus), dehydration and other volume depletion, abnormal function (e.g., abnormal results of function tests of central nervous system, peripheral nervous system and special senses, pulmonary function tests, cardiovascular function tests, kidney function tests, liver function tests, thyroid function tests, other endocrine function tests, and electrocardiogram suggestive of ST-segment elevation myocardial infarction , abnormal cardiovascular function tests, and other abnormal results of cardiovascular function tests), chronic pulmonary (e.g., chronic obstructive pulmonary disease with acute lower respiratory infection and acute exacerbation and other specified chronic obstructive pulmonary disease), dementia, glomerular disorders (e.g., glomerular disorders in infectious and parasitic diseases, neoplastic diseases, blood diseases and disorders involving the immune mechanism, diabetes mellitus, other endocrine, nutritional and metabolic diseases, and systemic connective tissue disorders), and hyperplasia of prostate.

6.2.7 Outcome: Identification of AKI

Machine learning models were built to predict AKI within 90 days after being discharged from the hospital or emergency department. Positive cases were those in which patients revisited hospital or emergency department with AKI within 90 days after being discharged, and negative cases were the ones when hospitalizations or emergency department visits with AKI never took place. There was a total of 899,449 negative and 5,993 positive cases in the dataset. There were no recurrent AKI examples (i.e., excluded 25,084 patients) in the data because we excluded the cases where AKI or dialysis was acquired during the index hospital stay or emergency department visit.

The incidence of AKI was detected using the Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System based on the ICD-10 (International Classification of Diseases - Tenth Revision) diagnostic codes (i.e., ICD-10 code of AKI is "N17").

6.2.8 Data Preprocessing

For each feature described in Section 2.5, the last recorded value before the first hospital admission or emergency department visit was captured. Medication, diagnosis code, and comorbidity features were set to either "Y" or "N." If a patient had a certain comorbid condition or was prescribed a medication, then its corresponding value was taken as "Y." Instead of reporting individual ages, we calculated age group features for the patients. If a patient's age laid within the specified range of an age group, we set the value to "1" for that corresponding feature. The sex feature took either "M" or "F" if the information is available in the dataset. Patients with invalid age or sex were removed from the cohort. The region feature took either "R" or "U" to represent rural and urban, respectively. The income feature took an integer value ranged between 1 to 5 to represent the income quintile of a particular patient.

All these features from different data sources were integrated using the encoded identifiers derived by ICES using patient health card numbers. The features in the cohort were transformed into a format and scale that was suitable for the machine learning techniques. For each patient, we aggregated multiple values (rows) of a single feature into one by considering the latest values of that feature.

6.2.9 Analysis using Machine Learning Techniques

We employed both traditional and state-of-art analysis techniques to build trust with end-users and, at the same time, allow them to explore complex relationships in the dataset. We developed 31 AKI prediction models based on combinations of eight classifiers—namely, classification and regression tree (CART) (Wilkinson, 2015), C5.0 (Quinlan, 2014), naïve Bayes (NB) (Lewis, 1998), logistic regression (Bahnsen et al., 2014), and support vector machine (SVM) with four different kernels (linear, polynomial, sigmoid, and radial) (Cristianini and Shawe-Taylor, 2000), two sampling techniques—namely, under sampling and SMOTE, and three ensemble methods—namely, Boosting, Bagging, and XGBoost. These techniques are chosen for several reasons: 1) They each represent different types of machine learning methods. For example, the decision tree is a rule-based, regression is a statistical, and naïve Bayes is a probability-based method. 2) Each of these methods has its own set of advantages and limitations. For instance, decision tree models are more human-interpretable but often fail to represent complex relationships among data elements. On the contrary, SVM is equipped to model complex non-linear relationships using different kernels but difficult to interpret. 3) Medical experts are more familiar with regression than other machine learning algorithms, which convinced us to include regression in this analysis.

6.2.9.1 Ensemble-Based Methods

Since the number of negative cases was significantly higher than the number of positive cases, we considered the dataset as highly imbalanced. Traditional machine learning techniques that are designed to optimize the overall accuracy tend to achieve poor performance in this class imbalanced learning scenario. An ensemble method offers a solution to this problem by combining several classification models to obtain better performance than the base classifiers (Dietterich, 2000). To deal with the class imbalance issue in this study, we incorporated four different combinations of ensemble and sampling methods—namely, SMOTEBoost, SMOTE-Bagging, UnderBagging, and RUSBoost that are available in the "embc" package of R (Barandela et al., 2003; Freund and Schapire, 1997; Wang and Yao, 2009). The RUSBoost was implemented using the "rus" function in the "ebmc" package. The weak learners in RUSBoost are trained on random under-sampled datasets (Seiffert et al., 2010). Those learners are then combined to generate the final ensemble model. We used the "sbo" function to implement SMOTEBoost. SMOTE (Synthetic Minority Oversampling Technique) is a sampling technique that synthesizes new instances for the minority class using the k-nearest-neighbours algorithm (Chawla et al., 2002). SMOTEBoost returns several weak learners that are trained on SMOTE-generated datasets along with their error estimations (Galar et al., 2012). The "sbag" function was used to implement SMOTEBagging, which combines SMOTE and random over-sampling to rebalance the dataset [44]. We used the "ub" function to implement the UnderBagging method. Unlike other ensemble methods discussed above, UnderBagging only incorporates random under-sampling to reduce the instances of the majority class in each bag to rebalance the class distribution. We configured this function in such a way that the amount of majority instances became equal compared to the minority instances (i.e., imbalance ratio = 1). We used NB, SVM, CART, and C50 as weak learners for the ensemble methods, which are discussed in the following subsections.

6.2.9.1.1 Support Vector Machine

The objective of the SVM is to find an optimal separating hyperplane in a multi-dimensional space (i.e., depending on the number of features) that distinctly divides the instances of different classes. Although SVM models are often not human-interpretable, it has been proven to work well on prediction tasks involving a large number of features [18]. It has become popular in healthcare research recently because it is more effective in analyzing high dimensional EHRs. In addition, the regularisation parameters of SVM kernels help users avoid over-fitting. Since the performance of the models widely varies depending on the selection of the kernel (Tomar and Agarwal, 2013) and kernels are quite sensitive to over-fitting (Cawley and Talbot, 2010), one of the main challenges is to select an appropriate kernel. Thus, we tested the performance of four well-known kernel functions in this study—namely, linear, polynomial, sigmoid, and radial.

6.2.9.1.2 Decision Tree

A decision tree is the representation of possible outcomes of a decision depending on certain conditions (Quinlan, 2014). It is similar to a flowchart where every non-leaf node represents a test for a specific feature, and the leaf node represents a particular outcome. Decision tree reduces the ambiguity of complicated clinical decisions and requires reduced effort for data preparation compared to other techniques. It can be an effective technique to analyze datasets with missing values because the tree-building process is not affected by the missing data (Niuniu and Yuxun, 2010). We choose the decision tree mainly because it is easy to interpret and understand. Despite the advantages, decision tree models are often volatile, meaning that a minor alteration in the training data may cause a massive change in the structure of the tree. To overcome this issue, we included other types of base classifiers along with decision tree and verified the structure of the generated tree with a healthcare expert. We incorporated two different algorithms to develop decision tree models in this study. The classification and regression tree (CART) were implemented using “rpart” package (Wilkinson, 2015), and the C5.0 classifier was implemented using the “C50” package in R (Quinlan, 2014).

6.2.9.1.3 Naïve Bayes

NB is a simple probabilistic classifier established on Bayes theorem (Lewis, 1998), which is exceptionally fast to train compared to other complex techniques (Tomar and Agarwal, 2013). Classification of the new data using this technique only requires mathematical operations based on the feature probability. We choose NB mainly because it is less sensitive to missing data. However, since this technique is designed based on the assumption of feature independence, the performance may deteriorate when features in the training data are related. We used the "naive Bayes" package to implement the NB algorithm in this study (McCallum and Nigam, 1998).

6.2.9.2 Logistic Regression

Logistic regression draws a separating line among the classes using the training dataset and then applies that line to classify the unknown data points. It is used to analyze the relationships between one dependent feature and one or more independent features. Logistic regression models are informative as they reveal the association among features in terms of odds ratios. Over the last decades, logistic regression techniques have become very popular in healthcare studies (Ismail and Anil, 2014). Although logistic regression models are not designed to support imbalanced classification directly, they can be modified to work with skewed distributions. In order to adjust the regression coefficients while training with the imbalance data, we implemented a cost-sensitive regression model. We adjusted the weight of the minority class based on the cost of its misclassification compared to the cost of misclassifying the majority class. We used internal 10-fold cross-validation during training to determine the appropriate weight for the minority class.

6.2.9.3 XGBoost

XGBoost (i.e., eXtreme Gradient Boosting) is an advanced implementation of gradient boosted decision trees that can be used for ranking, regression, and classification problems (Chen and Guestrin, 2016). One of the main advantages of XGBoost is that it supports parallel computation, which makes it faster than other implementations of gradient boosting. Because of its time complexity and performance superiority, it has been widely used in healthcare research, such as analysis of EHRs (C. Wang et al., 2018) and cancer diagnosis (C.-W. Wang et al., 2018). We used the "xgboost" package to implement XGBoost in R. Since this implementation of XGBoost only works with numeric data, we converted the categorical features in our dataset into numerical vectors. The "xgboost" package includes both a tree learning algorithm and linear model solver. We implemented both algorithms to compare their performance. This package has a built-in mechanism to control the balance of positive and negative weights as well. To train the models with unbalanced data, we adjusted the "scale_pos_weight" parameter based on the ratio of the negative class to the positive class (Wang et al., 2019). We performed a grid search on the parameters of XGBoost and tuned the regularization parameters using the best parameters from the grid search.

6.2.10 Tools and Technologies

We primarily used two different data analysis software: SAS and R. SAS was used to cut and process the cohort because ICES health administrative databases were stored in a SAS server ("SAS Enterprise BI Server," n.d.). We used SAS programming, SQL, and predefined macros to prepare data for analysis. Then we loaded the preprocessed dataset in R packages ("RStudio | Open source & professional software for data science teams," n.d.) for additional analysis using machine learning techniques. We chose R mainly because it 1) is installed on the ICES workstations already, 2) has a rich array of machine learning libraries, 3) is open-source and platform-independent, and 4) is continuously providing updates with new libraries.

6.3 Results

This section presents the results of this study. We divided the results into two subsections. First, we provide an overview of the dataset in Subsection 3.1. The results of predictive models are presented in Subsection 3.2.

6.3.1 Cohort Characteristics

A total of 905,442 participants were included in the derivation cohort, of which 5,993 had AKI during their hospital admission or emergency department visit after being discharged from the index encounter. We excluded 25,084 patients who developed AKI during the index hospitalization or emergency department visit. Selected characteristics of the derivation cohort are presented in Table 1.

Table 6-10: Baseline characteristics of patients in the cohort who were admitted to the hospital or visited the emergency department between 2014 and 2016.

Characteristics	Patients admitted to hospital or visited ED		
	Total Patients	**AKI**	**No AKI**
Cohort size	905,442	5993	899,449
Age, yr, mean (SD)			
65 to <70	181,088 (20%)	589	180,499
70 to <80	371,231 (41%)	1911	369,320
80 to <90	269,147 (30%)	2485	269,147
>=90	81,489 (9%)	1008	80,481
Sex			
Women	507,047 (56%)	2901	504,146
Year of cohort entry (index date)			
2014-2015	588,537 (65%)	3987	584,550
2015-2016	316,904 (34%)	2006	314,898
Location			
Rural residence	144,870 (16%)	501	144,369
LTC			
Long-term care	36,217 (4%)	745	35,472
Income Quintile			
1 (lowest)	172,035 (19%)	1,306	170,729
2	189,143 (21%)	1,318	187,825
3	182,588 (20%)	1,173	181,415
4	181,086 (20%)	1,154	179,932
5 (highest)	180,590 (20%)	1,043	179,547
Comorbid conditions (by codes)			
Hypertension	814,604 (88%)	5784	808,820
Diabetes	358,472 (38%)	3306	355,166
Heart failure	125,136 (14%)	1821	123,315
Coronary artery disease	239,437 (26%)	2005	237,432
Chronic liver disease	33,359 (4%)	297	33,062
Cancer	145,286 (16%)	1016	144,270
Chronic kidney disease	86,442 (9%)	1854	84,588
Kidney stones	12,457 (1%)	93	12,364
Peripheral vascular disease	13,197 (2%)	158	13,039
Cerebrovascular disease	25,835 (3%)	282	25,553
Hospital Diagnosis Codes			
Disorders of fluid, electrolyte and acid-base balance (E87)	13563 (1%)	962	12601
Delirium (F05)	4996 (1%)	342	4654
Atrial fibrillation (I48.91)	34120 (4%)	1978	32142
Mycoplasma pneumoniae (B96)	6197 (1%)	434	5763
Anaemia (D64.9)	11814 (1%)	791	11023
Valve disorders (I35)	1261 (1%)	186	1075
Fracture of femur (S72)	7263 (1%)	231	7032
Atherosclerotic cardiovascular disease (I25.10)	21472 (2%)	1256	20216
Volume depletion (E86.9)	3739 (1%)	240	3499
Diseases of the digestive system (K00-K95)	4552 (1%)	264	4288
Abnormal functions of organs and systems (R94.8)	11348 (2%)	725	10623
Chronic pulmonary (J81.1)	24217 (3%)	971	23246

Hyperplasia of prostate (N40.1)	5047 (1%)	153	4894
Certain infectious and parasitic diseases (A00-B99)	1191 (1%)	105	1086
Dementia (F03. 90)	8714 (1%)	390	8324
Glomerular disorders (N08)	3988 (1%)	569	3419

All the patients in the cohort were aged 65 years or older, where the mean age was 70 years. Among the participants, about 56% were women. About six percent of patients were in long term care, and sixteen percent were from rural areas. The pre-existing comorbidities were diabetes (38%), hypertension (88%), major cancer (16%), coronary artery disease (25%), cerebrovascular disease (3%), heart failure (14%), chronic kidney disease (9%), kidney stones (1%), and peripheral vascular disease (2%). Some of the commonly prescribed medications were rosuvastatin calcium (22%), atorvastatin calcium (24%), amlodipine besylate (19%), metformin hcl (16%), and hydrochlorothiazide (20%).

6.3.2 Classification Results

We evaluated all of the machine learning models using 10-fold cross-validation (Japkowicz and Shah, 2011). The cohort was divided into ten equal groups, where nine groups were used for training, and the tenth group was used for testing. We repeated this process ten times, using different parts for training and testing, and assessed the performance of the models for each fold. We then combined the results of these folds to calculate the evaluation scores. We measured the validity of the tests in terms of sensitivity and specificity. Sensitivity is the capacity of a test to classify an individual as “at-risk” correctly. It represents the probability of a test being positive when “AKI” is present. On the contrary, specificity refers to the ability to classify an individual as “risk-free” correctly. Since predicting AKI was a binary classification problem (i.e., AKI or Non-AKI), all of the machine learning techniques were capable of providing a confidence score along with the output. The trade-off between sensitivity and 1-specificity was achieved by altering the threshold on the confidence scores, generating the receiver operating characteristic (ROC) curve. We used the ROC space to compare the performances of alternative tests in terms of 1-specificity and sensitivity. Thus, we computed and reported sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC). The AUROC ranged from 0.61 to 0.88 for predicting AKI among 31 machine learning models. The average AUROC values of ensemble methods were higher than the cost-sensitive logistic regression model. Among the sampling-based ensemble methods, the performances of the UnderBagging and RUSBoost methods were better than the SMOTE. We achieved the best result of AUROC 0.88 with 1) a combination of RUSBoost and SVM using a sigmoid kernel and 2) XGBoost using a tree learning algorithm. The AUROC of the linear boosting algorithm (XGBoost) was 0.84, which was higher than the cost-sensitive logistic regression but lower than the tree learning algorithm (XGBoost). Since it is a disease prediction problem, high sensitivity was more useful than specificity. The highest sensitivity was 0.90, which was achieved using SVM-sigmoid and SVM-radial kernels with RUSBoost and SMOTE-Bagging, respectively. The complete list of performance measures is presented in Table 2.

Table 6-11: Performances of the machine learning techniques grouped by four ensemble-based methods and results of XGBoost and cost-sensitive regression analysis.

Ensemble-Based Methods	Machine Learning Techniques	Sensitivity	Specificity	AUROC
NA	**Logistic Regression**	0.79	0.72	0.77
SMOTEBoost	CART	0.77	0.69	0.74
	C5.0	0.84	0.78	0.83
	NB	0.61	0.89	0.75
	SVM (linear)	0.84	0.74	0.79
	SVM (polynomial)	0.78	0.82	0.81
	SVM (sigmoid)	0.76	0.85	0.84
	SVM (radial)	0.70	0.83	0.82
SMOTE-Bagging	CART	0.60	0.71	0.68
	C5.0	0.62	0.84	0.79
	NB	0.69	0.73	0.72
	SVM (linear)	0.76	0.84	0.81
	SVM (polynomial)	0.82	0.73	0.80
	SVM (sigmoid)	0.84	0.71	0.81
	SVM (radial)	0.90	0.74	0.86
UnderBagging	CART	0.71	0.83	0.79
	C5.0	0.88	0.76	0.85
	NB	0.58	0.72	0.61
	SVM (linear)	0.77	0.84	0.83
	SVM (polynomial)	0.85	0.71	0.84
	SVM (sigmoid)	0.89	0.71	0.85
	SVM (radial)	0.79	0.90	0.86
RUSBoost	CART	0.78	0.74	0.76
	C5.0	0.84	0.77	0.82
	NB	0.68	0.72	0.71

	SVM (linear)	0.84	0.78	0.83
	SVM (polynomial)	0.74	0.85	0.82
	SVM (sigmoid)	0.90	0.79	0.88
	SVM (radial)	0.71	0.87	0.85
XGBoost	Tree boosting	0.89	0.81	0.88
	Linear boosting	0.86	0.77	0.84

6.4 Discussion

In this study, we demonstrated how machine learning techniques could help with the prediction of AKI using administrative health databases stored at ICES. Several machine learning-based models have been developed in recent studies to predict AKI among patients in ICU and post-operative (Abdullah et al., 2020c; Cheng et al., 2017; Davis et al., 2017; Gameiro et al., 2020; Ibrahim et al., 2019; Rashidi et al., 2020; Tran et al., 2019). However, most of these models only focus on a specific medical condition and consider the risk factors associated with that condition. For instance, Go et al. (2010) examined how AKI affects the risk of chronic kidney disease, cardiovascular events, and other patient-related outcomes in hospital settings (Go et al., 2010). The earlier AKI can be predicted, the better the chances are to prevent AKI and its associated cost. The features that have been used in most of the existing studies work better in predicting AKI if their values are recorded closer to the timing of AKI onset. However, it may not be beneficial to detect AKI close to its onset because clinicians will not have enough time to intervene. Thus, there is a trade-off between accuracy and usefulness, which can be optimized using information available in EHRs. Although some studies have developed risk stratification models for AKI using EHRs (Kane-Gill et al., 2015; Matheny et al., 2010), they can only predict hospital-acquired AKI and do not consider patients who are at risk of developing AKI after being discharged. To our best knowledge, there are no previous studies in the literature that predict the risk of AKI after being discharged from the hospital using both the historical and healthcare utilization data. Thus, this study is not only novel but also clinically relevant because it provides clinicians with the ability to intervene and treat patients before AKI cause irreversible damage.

We analyzed all AKI events that took place within 90 days after being discharged from the hospital or emergency department and developed prediction models to identify high-risk patients. We decided to choose 90 days timeframe for following up because 1) out of all AKI cases within six months after discharge, about 85 percent were acquired within this timeframe; and 2) it was a reasonable timeframe considering the trade-off between the models' usefulness (from a clinical point of view) and predictive power (from a machine learning point of view). Table 3 shows how many AKI acquired cases were identified within different time intervals. The machine learning models presented in this study can be adapted to make predictions at any other timeframes if needed.

We incorporated eight different machine learning classifiers and three ensemble methods, and two sampling techniques to develop 31 prediction models. Although each combination of machine learning techniques and ensemble-based methods performed reasonably well, the performance of SVM with sigmoid kernel and tree-based XGBoost produced better results than other techniques in general. The performance of all of the ensemble-based methods were consistent and produced similar results for different base classifiers. The results shown in Table 2 indicate that the models agreed with each other.

Table 6-12: The number of AKI cases are grouped into six time periods.

Intervals	Readmission with AKI
1-3 days	415
4-7 days	534
8-14 days	888
15-30 days	1517
31-60 days	3579
61-90 days	1499

To understand the models better, we explored the features that are important in each prediction model. We analyzed this information with a nephrologist to confirm the correctness of the models. We observed the odds ratio and p-value of the features in the regression model, feature importance in decision tree and XGBoost models, and coefficients in the SVM-linear models to understand the association between different features and AKI. The features included in this study can be divided into four categories—namely, demographics, comorbidities, medications, and diagnosis codes.

In general, features from comorbidities and hospital diagnosis codes were more associated with AKI. Although the importance of the features varied based on the machine learning techniques, most of the features that stood out were common among these models. For instance, diabetes mellitus, hypertension, coronary artery disease, heart failure, major cancers, chronic liver disease, peripheral vascular disease, and chronic kidney disease were the comorbidity features that were important in most of the prediction models. These comorbid conditions are already known to be associated with AKI in the literature (Dylewska et al., 2019; Girman et al., 2012; Hsu and Hsu, 2016; Olsson et al., 2013; Rydén et al., 2014). The medication features that contributed to the higher risk of AKI include furosemide, allopurinol, hydrochlorothiazide, atorvastatin, metolazone, sunitinib malate, spironolactone, dexamethasone, chlorthalidone, atenolol, dexamethasone and oseltamivir phosphate. These medications are known to be nephrotoxic (Chao et al., 2015; Ho and Power, 2010; Perez-Ruiz, 2017; Pierson-Marchandise et al., 2017; Verdoodt et al., 2018; Wu et al., 2014). Delirium, anaemia, mycoplasma, fluid disorders, atrial fibrillation, atherosclerotic cardiovascular disease, mycoplasma pneumoniae, hyperplasia of prostate, glomerular disorders, and valve disorders were the features belonging to the diagnosis codes that were associated with increasing the risk of AKI in the prediction models. Several studies in the literature associate these medical conditions with AKI (Carrara et al., 2017; Godin et al., 2013; Ng et al., 2016; Siew et al., 2017; Zaleska-Kociecka et al., 2019). Among the demographic features, age, sex, location (i.e., urban or rural residence), and long-term care were found to be associated with AKI in most of the prediction models. Similar to comorbidity, medication, and diagnosis code, these demographic features are already known to be associated with AKI (Evans et al., 2017; Neugarten and Golestaneh, 2018; Yokota et al., 2018) in the literature, which more conclusively proves the correctness of the prediction models. Through a comprehensive analysis of ICES' healthcare administrative datasets, this study shows that AKI is predictable using EHRs. Successful implementation of these prediction models in a healthcare setting can potentially reduce the risk of AKI among older patients.

6.5 Limitations and Future Work

The study should be evaluated with respect to several limitations. First, our models were trained and tested on a cohort of older patients (65 years or older), which limits the generalizability of the models. Second, we excluded patients with missing or invalid demographics information. This may affect the performance of the models if the excluded data includes any interesting or rare cases. Third, the models are based on a cohort containing Ontario patients only, which limits this study to a specific geographic location. Fourth, the proposed prediction models are trained and tested on a specific patient cohort. It is essential to test the models' performance with real-time medical data before applying them in a clinical setting. Fifth, since we developed 31 prediction models, and many of them have different mechanisms to identify feature importance, the interrelationships produced by these models are very complex. This study only identifies the most significant predictors but does not incorporate any rankling system for predictors. Finally, we identified the episode of AKI using ICD-10 codes, which may not include undetected cases in hospital settings. Moreover, since AKI was identified using the diagnosis code, this study does not consider the severity of AKI. Our future work concerns a deeper analysis of severe AKI that requires dialysis.

6.6 Conclusion

AKI is characterized by a sharp decline in renal function and associated with increased health-related costs and mortality. AKI is avoidable and may be preventable through an earlier prediction using risk factors available in EHRs. This study is designed to identify older patients who are discharged from the hospital or emergency department and at risk of developing AKI within 90 days after discharge. We employ eight traditional and state-of-art machine learning classifiers along with two sampling techniques, and three ensemble methods to build AKI prediction models. The performances of these models were consistent, and a maximum AUROC of 0.88 was achieved through 10-fold cross-validation. We analyzed the models with a healthcare expert and identified features that are most relevant in predicting AKI. Most of these features are already known to be AKI-associated, which proves the correctness and feasibility of the prediction models. This study predicts the risk of AKI for a patient after being discharged from the hospital or emergency department, which provides healthcare providers enough time to intervene, monitor them more carefully, and avoid prescribing nephrotoxic medications for such patients.

Chapter 7

7 Conclusion

This dissertation has discussed several aspects relating to the design of VA for EHRs. First, a systematic literature survey has been conducted to analyze the design and implementation of existing EHR-based systems. We then presented a framework to evaluate EHR-data-driven tasks and activities in the existing systems. The gaps that we identified during the analysis with the framework motivated us to design new EHR-based VA systems. Therefore, we designed and developed two novel VA systems—namely, VISA_M3R3 (VISual Analytics, VISA for Multiple Regression analyses and fRequent itemset Mining of electronic Medical Records, M3R3) and VALENCIA (Visual Analytics for Cluster Analysis and Dimension Reduction of High Dimensional Electronic Health Records). We also conducted two independent population-based retrospective cohort study to test the hypothesis and ideas generated from the VA systems. These systems and studies are designed to assist the healthcare researchers at the ICES-KDT program. We demonstrated the effectiveness of the proposed systems by investigating the process of analyzing the health administrative data housed at ICES to solve different AKI-related problems.

This chapter, which serves as a conclusion of the dissertation, is divided into three sections: 1) a review of the chapters and some of their contributions, 2) general contribution of this dissertation to the scientific literature, and 3) some future research areas.

7.1 Dissertation Summary

In **Chapter 2**, we have presented a framework to examine EHR-data-driven activities and tasks in the context of interactive visualizations. Using a literature survey of 19 EHR-based existing systems, we demonstrated how different combinations of sub-tasks, tasks, and sub-activities work together to help users achieve their overall goals in the system. The proposed framework can help 1) designers to conceptualize activities, sub-activities, tasks, and sub-tasks of new systems, 2) researchers to understand the design concepts in a systematic way, and 3) evaluators to assess existing EHR-based interactive visualization systems.

In **Chapter 3**, we described how VA systems could be designed systematically to support EHR-driven tasks and investigate complex clinical problems. We developed VISA_M3R3 that integrates multiple statistical and machine learning techniques with interactive visualization to identify potentially nephrotoxic medications. VISA_M3R3 has shown to assist healthcare researchers in 1) comparing multiple logistic regression models, 2) understanding the relationships among predictors and response variable, 3) identifying frequent itemsets from items of interest, and 4) interpreting regression results. VISA_M3R3 can also be used to develop an alert system to raise physicians' awareness of AKI-associated medications. This, in turn, will prompt healthcare providers to carry out additional clinical investigations on these high-risk medications.

VISA_M3R3 only visualizes regression models of medication combinations but do not investigate how individual medications within combinations are affecting AKI. In **Chapter 4**, we presented a population-based retrospective cohort to overcome this limitation and understand the synergistic effect of AKI-inducing medication combinations more comprehensively. Through an investigation of prescription records of one million adult patients stored in the ICES datasets, we identified 55 AKI-inducing medications among a total of 595 medications and 78 AKI-inducing medication combinations among a total of 7,748 frequent medication groups. We also identified 37 cases where a medication is associated with increased risk of developing AKI when combined with another medication. Finally, we performed an electronic literature search and consulted with a nephrologist to verify the findings of this study. Although many of the medications and medication combinations that we detected are already known to be nephrotoxic, some of them have not been investigated before. This study will assist healthcare researchers in identifying candidates for future drug-safety studies.

In **Chapter 5**, we explained how VA systems could be designed to address the challenges of high-dimensional EHRs. We introduced VALENCIA that integrates a wide range of traditional and state-of-the-art analysis techniques with several interactive visualizations to provide a deeper understanding of the structure of data, results, control parameters, and analytical processes. We have demonstrated the utility of VALENCIA using a case study. Through a number of formative evaluations, we have found that VALENCIA assists healthcare providers in 1) exploring EHRs using different dimension reduction and clustering techniques, 2) identifying relationships among different features, 3) generating hypotheses, and 4) comparing results of different analysis techniques.

We identified multiple risk factors of AKI while performing a cluster analysis of the ICES dataset using VALENCIA, which motivated us to design another study for predicting AKI. In **Chapter 6**, we employed a number of machine learning techniques to develop prediction models to identify patients who were at risk of developing AKI within 90 days after they were discharged from the emergency department or hospital. We included demographics, comorbid conditions, medications, and hospital diagnosis codes as predictor variables. A total of sixteen prediction models based on combinations of four machine learning techniques and four ensemble-based methods, along with a cost-sensitive logistic regression model, were developed for this study. The performances of these models were consistent, and we achieved an AUROC of 0.88 through ten-fold cross-validation.

7.2 General Contributions

As described in Chapter 1, the broad concern of this research surrounds the design of visual analytics for EHRs. Currently, there is a scarcity of research in this field, and therefore, we intended to bridge the gap through this work. This dissertation presents the design process of human- and activity-centered computational systems for healthcare. It is often challenging to fulfill the computational and cognitive demands of healthcare providers when designing such systems. This dissertation also describes how different EHR-related challenges can be addressed by combining statistical methods, data mining algorithms, machine learning techniques, and information visualization. It demonstrates how VA systems can be designed systematically and how healthcare providers' capabilities to interact with machine learning processes can be improved by VA. In addition, this dissertation offers the healthcare domain with evidence of the effectiveness of VA for managing EHRs. This research has suggestions for other domains that require their data to be made accessible and analyzable through VA.

Through a systematic survey in Chapter 2, this dissertation provides a detailed analysis of the EHR-data-driven tasks and activities supported by the existing interactive visualization systems, which was lacking in the literature. Moreover, the proposed activity and task analysis framework is helpful for the designer and evaluators of any EHR-based visualization system. This framework will lead to the development of best practices for designing related frameworks in other domains.

Another contribution of this dissertation is the VA systems, VISA_M3R3 and VALENCIA, that are discussed in Chapters 3 and 5. To help us learn how healthcare providers perform real-world tasks, and to help us design and develop these systems, we adopted a participatory design approach. We performed formative evaluations at every step of the design and development process with the stakeholders at the ICES-KDT program. Through these evaluations, the healthcare experts found the systems useful and sophisticated. The systems would benefit not only healthcare researchers across the globe but also designers of EHR-based VA systems. These systems are also scalable and can be reconfigured to work with other forms of data.

Other contributions emerge from the population-based studies presented in Chapters 4 and 6. Chapter 4 presented an automated approach to identify AKI-associated medication and medication combinations. The drug-safety studies in clinical settings are usually costly and time-consuming. The proposed approach can help healthcare researchers to prepare a short list of suitable candidates for clinical studies, which not only saves money and time but also supports the identification of potentially unknown nephrotoxic medication combinations. In addition, the study in Chapter 6 demonstrates how machine learning can be used to predict AKI using EHRs. Most of the existing AKI prediction models are designed for a specific medical condition and predict AKI close to its onset. To the best of our knowledge, Chapter 6 presents the first study that predicts the risk of AKI for patients who are discharged from the hospital or emergency department, which provides enough time for clinicians to intervene.

7.3 Limitations and Future Work

The systems and studies presented in this dissertation lay a foundation for the usefulness of VA tools in healthcare. Through the activity and task analysis framework, we identified that there are a limited number of interactive visualization systems that support higher-level activity, “predicting” and “monitoring.” The proposed systems are primarily designed to support “interpreting” and “predicting.” More research is needed to explore how EHR-based VA systems can be extended to support the activity, “monitoring.”

This dissertation reports the development of two VA systems—namely, VISA_M3R3 and VALENCIA. We described the data sources, input, output, stakeholders, interface, design criteria, levels of abstraction and different subsystems of these systems in Chapters 3 and 5. Although we included workflow diagrams to describe the systems' design, the architectures of these systems are not described thoroughly. The description was mainly focused on the conceptual challenges but did not consider the practical difficulties of these complex systems. For instance, we did not investigate how many users can be supported by the systems simultaneously or the systems’ capability of interacting with other systems. Since both VISA_M3R3 and VALENCIA were developed to process sensitive patient records in an access-restricted offline setting, it was not required to create web services or implement any adapter (i.e., software that connects two systems and reconciles the distinctions between them).

Another limitation of these systems lies in their ability to increase their functionalities and capacity based on users’ demand. The description of these systems does not describe the challenges regarding scalability and extensibility. Scalability refers to methods that guarantee that the functionality and quality of a system are maintained as the number of users goes up or the complexity of the data increases. On the other hand, extensibility refers to a system's ability to adapt to new interfaces, functionalities, data, and input. We employed multiple programming languages, platforms, and servers for developing these systems. The current VISA_M3R3 and VALENCIA systems were implemented using HTML, JavaScript library D3, standard PHP programming language, SAS, and R packages. The datasets were stored in the SAS server. R server was incorporated to perform all the underlying processing. Web server (PHP) was used to host the HTML files to maintain communication with the R server. Since different technologies are combined to develop these systems, it is easy to incorporate new analysis algorithms (e.g., using a new library or package in R), additional features in the interface (e.g., modifying D3 functions), and supplementary data (e.g., incorporating new datasets). However, we were not able to determine the scalability of these systems. Since we developed them in an access-restricted virtual machine, we did not get a chance to access the systems' scalability. We will evaluate these systems more comprehensively in the future.

Although healthcare experts have evaluated VISA_M3R3 and VALENCIA during the design process and found it useful, we have not conducted formal studies to assess their performances, nor the efficiency of their human-information discourse mechanisms. Thus, additional studies can help ascertain the effectiveness of these systems for both expert and non-expert users.

Both systems presented in this dissertation are implemented and tested using ICES-KDT datasets. Studies that assess the effectiveness of these systems with different datasets and settings will provide a better understanding of the efficacy of the systems.

References

Aamdal, S., 1992. Can ondansetron hydrochloride (Zofran) enhance the nephrotoxic potential of other drugs? Annals of oncology : official journal of the European Society for Medical Oncology 3, 774. https://doi.org/10.1093/oxfordjournals.annonc.a058342

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley interdisciplinary reviews: computational statistics 2, 433–459.

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

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

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

Abramson, E.L., Barrón, Y., Quaresimo, J., Kaushal, R., 2011. Electronic prescribing within an electronic health record reduces ambulatory prescribing errors. Joint Commission Journal on Quality and Patient Safety 37, 470–478. https://doi.org/10.1016/S1553-7250(11)37060-2

Adachi, S., 2017. Rigid geometry solves “curse of dimensionality” effects in clustering methods: An application to omics data. PLoS One 12. https://doi.org/10.1371/journal.pone.0179180

Adhiyaman, V., Asghar, M., Oke, A., White, A.D., Shah, I.U., 2001. Nephrotoxicity in the elderly due to co-prescription of angiotensin converting enzyme inhibitors and nonsteroidal anti-inflammatory drugs. Journal of the Royal Society of Medicine 94, 512–4. https://doi.org/10.1177/014107680109401005

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

Agrawal, R., Swami, A., Imielinski, T., 1993. Database Mining: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering 5, 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.

Aimone, A.M., Perumal, N., Cole, D.C., 2013. A systematic review of the application and utility of geographical information systems for exploring disease-disease relationships in paediatric global health research: The case of anaemia and malaria. International Journal of Health Geographics 12. https://doi.org/10.1186/1476-072X-12-1

Alexander, T., McArthur, E., Jandoc, R., Welk, B., Hayward, J.S., Jain, A.K., Braam, B., Flockerzi, V., Garg, A.X., Quinn, R.R., 2017. Antihypertensive medications and the risk of kidney stones in older adults: A retrospective cohort study. Hypertension Research 40, 837–842. https://doi.org/10.1038/hr.2017.42

Ali, M., Jones, M.W., Xie, X., Williams, M., 2019. TimeCluster: dimension reduction applied to temporal data for visual analytics. Vis Comput 35, 1013–1026. https://doi.org/10.1007/s00371-019-01673-y

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, 1292–8. https://doi.org/10.1681/ASN.2006070756

Alirezaei, A., Argani, H., Asgharpour, M., Bahadorimonfared, A., Bakhtiyari, M., 2017. An update on allopurinol and kidney failure; new trend for an old drug. https://doi.org/10.15171/jrip.2017.57

Allison, S.J., 2015. Effect of perioperative aspirin and clonidine on AKI. Nature Reviews Nephrology 11.

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, 1148–1154. https://doi.org/10.1377/hlthaff.2014.0352

Anderson, F.A., Wyman, A., Varon, J., McCullough, P.A., Devlin, J.W., Weir, M.R., Katz, J.N., Szczech, L.A., Granger, C.B., Dasta, J.F., Amin, A., Frank, W., Frank Peacock, W., 2010. Circulation Treatment of Acute Hypertension Investigators Hypertension Acute Kidney Injury and Cardiovascular Outcomes in Acute Severe Hypertension Acute Kidney Injury and Cardiovascular Outcomes in Acute Severe

Hypertension 121, 2183–2191. https://doi.org/10.1161/CIRCULATIONAHA.109.896597

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. J Am Board Fam Med 28, 65–71. https://doi.org/10.3122/jabfm.2015.01.140181

Arabie, P., 1994. Cluster analysis in marketing research. Advanced methods of marketing research 160–189.

Arifin, S., Zulkardi, Z., Indra Putri, R., Hartono, Y., Susanti, E., 2017. Developing Ill-defined problem-solving for the context of “South Sumatera.” Journal of Physics: Conference Series 943, 12038. https://doi.org/10.1088/1742-6596/943/1/012038

Asimov, D., 1985. The grand tour: a tool for viewing multidimensional data. SIAM journal on scientific and statistical computing 6, 128–143.

Assadi, F., Ghane Shahrbaf, F., 2015. Ghane Shahrbaf F, Assadi F. Drug-induced renal disorders. Journal of Renal Injury Prevention J Renal Inj Prev 4, 57–60. https://doi.org/10.12861/jrip.2015.12

Audia, P., Feinfeld, D.A., Dubrow, A., Winchester, J.F., 2008. Metformin-induced lactic acidosis and acute pancreatitis precipitated by diuretic, celecoxib, and candesartan-associated acute kidney dysfunction. Clinical toxicology (Philadelphia, Pa.) 46, 164–6. https://doi.org/10.1080/15563650701355314

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., Guo, Y.-K., Gut, I.G., Hanbury, A., Hanif, S., Hilgers, R.-D., Honrado, Á., Hose, D.R., Houwing-Duistermaat, J., Hubbard, T., Janacek, S.H., Karanikas, H., Kievits, T., Kohler, M., Kremer, A., Lanfear, J., Lengauer, T., Maes, E., Meert, T., Müller, W., Nickel, D., Oledzki, P., Pedersen, B., Petkovic, M., Pliakos, K., Rattray, M., i Màs, J.R., Schneider, R., Sengstag, T., Serra-Picamal, X., Spek, W., Vaas, L.A.I., van Batenburg, O., Vandelaer, M., Varnai, P., Villoslada, P., Vizcaíno, J.A., Wubbe, J.P.M., Zanetti, G., 2016. Making sense of big data in health research: Towards an EU action plan. Genome Medicine 8, 71. https://doi.org/10.1186/s13073-016-0323-y

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

Bagshaw, Sean M., Delaney, A., Jones, D., Ronco, C., Bellomo, R., 2007. Diuretics in the Management of Acute Kidney Injury: A Multinational Survey. Acute Kidney Injury 156, 236–249. https://doi.org/10.1159/000102089

Bagshaw, Sean M, George, C., Bellomo, R., ANZICS Database Management Committee, 2007. Changes in the incidence and outcome for early acute kidney injury in a cohort of Australian intensive care units. Critical care (London, England) 11, R68–R68. https://doi.org/10.1186/cc5949

Bahnsen, A.C., Aouada, D., Ottersten, B., 2014. Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring, in: 2014 13th International Conference on Machine Learning and Applications. Presented at the 2014 13th International Conference on Machine Learning and Applications, pp. 263–269. https://doi.org/10.1109/ICMLA.2014.48

Barandela, R., Valdovinos, R.M., Sánchez, J.S., 2003. New Applications of Ensembles of Classifiers. Patt. Analy. App. 6, 245–256. https://doi.org/10.1007/s10044-003-0192-z

Basile, A.O., Yahi, A., Tatonetti, N.P., 2019. Artificial Intelligence for Drug Toxicity and Safety. Trends in Pharmacological Sciences. https://doi.org/10.1016/j.tips.2019.07.005

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, 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, 2257–2270. https://doi.org/10.1109/TMM.2016.2614225

Bennett, W.M., 2013. Cyclosporine and tacrolimus nephrotoxicity. Utd 1–18.

Bernard, J., Sessler, D., Bannach, A., May, T., Kohlhammer, J., 2015. A visual active learning system for the assessment of patient well-being in prostate cancer research, in: ACM International Conference Proceeding Series. Association for Computing Machinery. https://doi.org/10.1145/2836034.2836035

Bird, S.T., Etminan, M., Brophy, J.M., Hartzema, A.G., Delaney, J.A., 2013. Risk of acute kidney injury associated with the use of fluoroquinolones. Cmaj 185, E475–E482.

Boccanfuso, J.A., Hutton, M., McAllister, B., 2000. The effects of megestrol acetate on nutritional parameters in a dialysis population. Journal of Renal Nutrition 10, 36–43. https://doi.org/10.1016/S1051-2276(00)90021-9

Boonstra, A., Versluis, A., Vos, J.F., 2014. Implementing electronic health records in hospitals: a systematic literature review. BMC health services research 14, 370.

Bove, T., Belletti, A., Putzu, A., Pappacena, S., Denaro, G., Landoni, G., Bagshaw, S.M., Zangrillo, A., 2018. Intermittent furosemide administration in patients with or at risk for acute kidney injury: Meta-analysis of randomized trials. PLoS ONE 13, e0196088. https://doi.org/10.1371/journal.pone.0196088

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, 1172–1177.

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

Carrara, C., Abbate, M., Sabadini, E., Remuzzi, G., 2017. Acute Kidney Injury and Hemolytic Anemia Secondary to Mycoplasma pneumoniae Infection. Nephron 137, 148–154. https://doi.org/10.1159/000478991

Carroll, L.N., Au, A.P., Detwiler, L.T., Fu, T.-C., Painter, I.S., Abernethy, N.F., 2014. Visualization and analytics tools for infectious disease epidemiology: a systematic review. J Biomed Inform 51, 287–298. https://doi.org/10.1016/j.jbi.2014.04.006

Cartin-Ceba, R., Kashiouris, M., Plataki, M., Kor, D.J., Gajic, O., Casey, E.T., 2012. Risk factors for development of acute kidney injury in critically ill patients: a systematic review and meta-analysis of observational studies. Critical care research and practice 2012, 691013. https://doi.org/10.1155/2012/691013

Cavallo, M., Demiralp, Ç., 2018. A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI ’18. Association for Computing Machinery, Montreal QC, Canada, pp. 1–13. https://doi.org/10.1145/3173574.3174209

Cawley, G.C., Talbot, N.L.C., 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation 29.

Chang, Y.-P., Huang, S.-K., Tao, P., Chien, C.-W., 2012. A population-based study on the association between acute renal failure (ARF) and the duration of polypharmacy. BMC nephrology 13, 96. https://doi.org/10.1186/1471-2369-13-96

Chao, C.-T., Tsai, H.-B., Wu, C.-Y., Lin, Y.-F., Hsu, N.-C., Chen, J.-S., Hung, K.-Y., 2015. Cumulative Cardiovascular Polypharmacy Is Associated With the Risk of Acute Kidney Injury in Elderly Patients. Medicine 94, e1251. https://doi.org/10.1097/MD.0000000000001251

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, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, San Francisco, California, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, P., Waitman, L.R., Hu, Y., Liu, M., 2017. Predicting Inpatient Acute Kidney Injury over Different Time Horizons: How Early and Accurate? AMIA Annu Symp Proc 2017, 565–574.

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, 591–620.

Cho, H., Berger, B., Peng, J., 2016. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. Cell Syst 3, 540-548.e5. https://doi.org/10.1016/j.cels.2016.10.017

Choo, J., Lee, H., Liu, Z., Stasko, J., Park, H., 2013. An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data, in: Visualization and Data Analysis 2013. International Society for Optics and Photonics, p. 865402.

Choudhury, D., Ahmed, Z., 2006. Drug-associated renal dysfunction and injury. Nature Clinical Practice Nephrology. https://doi.org/10.1038/ncpneph0076

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, 12.

Clinical Practice Guideline, 2012. KDIGO Clinical Practice Guideline for Acute Kidney Injury 2.

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, 1139–1147. https://doi.org/10.1377/hlthaff.2014.0048

Collins, N., 2018. AI predicts drug pair side effects | Stanford News [WWW Document]. URL https://news.stanford.edu/2018/07/10/ai-predicts-drug-pair-side-effects/ (accessed 1.5.20).

Collister, D., Pannu, N., Ye, F., James, M., Hemmelgarn, B., Chui, B., Manns, B., Klarenbach, S., Alberta Kidney Disease Network, 2017. Health Care Costs Associated with AKI. Clinical journal of the American Society of Nephrology : CJASN 12, 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.

Cook, D., Swayne, D.F., Buja, A., 2007. Interactive and dynamic graphics for data analysis: with R and GGobi. Springer Science & Business Media.

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. Clin Res Cardiol 106, 1–9. https://doi.org/10.1007/s00392-016-1025-6

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods [WWW Document]. Cambridge Core. https://doi.org/10.1017/CBO9780511801389

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

Cunningham, P., 2008. Dimension Reduction, in: Cord, M., Cunningham, P. (Eds.), Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval, Cognitive Technologies. Springer, Berlin, Heidelberg, pp. 91–112. https://doi.org/10.1007/978-3-540-75171-7_4

Cybulski, J.L., Keller, S., Nguyen, L., Saundage, D., 2015. Creative problem solving in digital space using visual analytics. Computers in Human Behavior, Digital Creativity: New Frontier for Research and Practice 42, 20–35. https://doi.org/10.1016/j.chb.2013.10.061

Da'as, N., Polliack, A., Cohen, Y., Amir, G., Darmon, D., Kleinman, Y., Goldfarb, A.W., Ben-Yehuda, D., 2001. Kidney involvement and renal manifestations in non-Hodgkin's lymphoma and lymphocytic leukemia: a retrospective study in 700 patients. European Journal of Haematology 67, 158–164. https://doi.org/10.1034/j.1600-0609.2001.5790493.x

Davis, E., 2019. What is a health care contract? Health values 4, 82–86, 89.

Davis, S.E., Lasko, T.A., Chen, G., Siew, E.D., Matheny, M.E., 2017. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc 24, 1052–1061. https://doi.org/10.1093/jamia/ocx030

Dawlilng, S., Lynn, K., Rosser, R., Braithwaite, R., 1981. The pharmacokinetics of nortriptyline in patients with chronic renal failure. Br J Clin Pharmacol 12, 39–45.

De Leeuw, J., 2005. Multivariate analysis with optimal scaling.

De Soete, G., Carroll, J.D., 1994. K-means clustering in a low-dimensional Euclidean space, in: New Approaches in Classification and Data Analysis. Springer, pp. 212–219.

Delamarre, D., Bouzille, G., Dalleau, K., Courtel, D., Cuggia, M., 2015. Semantic integration of medication data into the EHOP Clinical Data Warehouse. Studies in health technology and informatics 210, 702–6.

Demiralp, Ç., 2017. Clustrophile: A tool for visual clustering analysis. arXiv preprint arXiv:1710.02173.

Dev, V., Dixon, S.N., Fleet, J.L., Gandhi, S., Gomes, T., Harel, Z., Jain, A.K., Shariff, S.Z., Tawadrous, D., Weir, M.A., Garg, A.X., 2014. Higher anti-depressant dose and major adverse outcomes in moderate chronic kidney disease: a retrospective population-based study. BMC Nephrol 15, 79. https://doi.org/10.1186/1471-2369-15-79

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

Dietterich, T.G., 2000. Ensemble Methods in Machine Learning, in: Multiple Classifier Systems, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1

Dilts, D., Khamalah, J., Plotkin, A., 1995. Using cluster analysis for medical resource decision making. Medical Decision Making 15, 333–346.

Ding, H., Wang, C., Huang, K., Machiraju, R., 2014. iGPSe: a visual analytic system for integrative genomic based cancer patient stratification. BMC Bioinformatics 15, 203. https://doi.org/10.1186/1471-2105-15-203

Dixit, M., Doan, T., Kirschner, R., Dixit, N., 2010. Significant Acute Kidney Injury Due to Non-steroidal Anti-inflammatory Drugs: Inpatient Setting. Pharmaceuticals (Basel) 3, 1279–1285. https://doi.org/10.3390/ph3041279

Doupi, P., 2012. Using EHR data for monitoring and promoting patient safety: reviewing the evidence on trigger tools. Stud Health Technol Inform 180, 786–790.

Doust, D., Walsh, Z., 2011. Mediterranean Conference on Information Systems (MCIS) 2011 DATA MINING CLUSTERING : A HEALTHCARE APPLICATION.

Duke, J.D., Li, X., Grannis, S.J., 2010. Data visualization speeds review of potential adverse drug events in patients on multiple medications. Journal of Biomedical Informatics 43, 326–331. https://doi.org/10.1016/j.jbi.2009.12.001

Dupont, A.G., 1992. Carvedilol and the kidney. The clinical investigator 70, S127–S131.

Dylewska, M., Chomicka, I., Małyszko, J., 2019. Hypertension in patients with acute kidney injury. Wiad. Lek. 72, 2199–2201.

EHRIntelligence, 2018. 40% of Physicians See More EHR Challenges than Benefits [WWW Document]. EHRIntelligence. URL https://ehrintelligence.com/news/40-of-physicians-see-more-ehr-challenges-than-benefits (accessed 5.14.20).

Erbayraktar, Z., Evlice, A., Yener, G., Ulusu, N.N., 2017. Effects of donepezil on liver and kidney functions for the treatment of Alzheimer’s disease. Journal of integrative neuroscience 16, 335–346.

Eriksen, B.O., Hoff, K.R.S., Solberg, S., 2003. Prediction of acute renal failure after cardiac surgery: retrospective cross-validation of a clinical algorithm. Nephrol. Dial. Transplant. 18, 77–81. https://doi.org/10.1093/ndt/18.1.77

Escofier, B., Pagès, J., 1994. Multiple factor analysis (AFMULT package). Computational Statistics & Data Analysis 18, 121–140. https://doi.org/10.1016/0167-9473(94)90135-X

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

Estiri, H., Klann, J.G., Murphy, S.N., 2019. A clustering approach for detecting implausible observation values in electronic health records data. BMC Medical Informatics and Decision Making 19, 142. https://doi.org/10.1186/s12911-019-0852-6

Evans, R.D.R., Hemmilä, U., Craik, A., Mtekateka, M., Hamilton, F., Kawale, Z., Kirwan, C.J., Dobbie, H., Dreyer, G., 2017. Incidence, aetiology and outcome of community-acquired acute kidney injury in medical admissions in Malawi. BMC Nephrology 18, 21. https://doi.org/10.1186/s12882-017-0446-4

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

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

Faisal, S., Blandford, A., Potts, H.W., 2013. Making sense of personal health information: Challenges for information visualization. Health Informatics Journal 19, 198–217. https://doi.org/10.1177/1460458212465213

Feldman, R., Sanger, J., 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

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

Fleet, J.L., Weir, M.A., McArthur, E., Ozair, S., Devereaux, P.J., Roberts, M.A., Jain, A.K., Garg, A.X., 2014. Kidney function and population-based outcomes of initiating oral atenolol versus metoprolol tartrate in older adults. American journal of kidney diseases 64, 883–891.

Foguet-Boreu, Q., Violán, C., Rodriguez-Blanco, T., Roso-Llorach, A., Pons-Vigués, M., Pujol-Ribera, E., Cossio Gil, Y., Valderas, J.M., 2015. Multimorbidity Patterns in Elderly Primary Health Care Patients in a South Mediterranean European Region: A Cluster Analysis. PLoS One 10. https://doi.org/10.1371/journal.pone.0141155

Folorunso, O., Shawn Ogunseye, O., 2008. Challenges in the adoption of visualization system: a survey. Kybernetes 37, 1530–1541. https://doi.org/10.1108/03684920810907841

Fournier, J.-P., Lapeyre-Mestre, M., Sommet, A., Dupouy, J., Poutrain, J.-C., Montastruc, J.-L., 2012. Laboratory monitoring of patients treated with antihypertensive drugs and newly exposed to non steroidal anti-inflammatory drugs: a cohort study. PloS one 7, e34187. https://doi.org/10.1371/journal.pone.0034187

Fournier, J.-P., Sommet, A., Durrieu, G., Poutrain, J.-C., Lapeyre-Mestre, M., Montastruc, J.-L., French Network of Regional Pharmacovigilance Centres, 2014. Drug interactions between antihypertensive drugs and non-steroidal anti-inflammatory agents: a descriptive study using the French Pharmacovigilance database. Fundamental & clinical pharmacology 28, 230–5. https://doi.org/10.1111/fcp.12014

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

Freund, Y., Schapire, R.E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 55, 119–139. https://doi.org/10.1006/jcss.1997.1504

F.R.S, K.P., 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2, 559–572. https://doi.org/10.1080/14786440109462720

Fujiwara, T., Chou, J.-K., Shilpika, Xu, P., Ren, L., Ma, K.-L., 2020. An Incremental Dimensionality Reduction Method for Visualizing Streaming Multidimensional Data. IEEE Trans. Visual. Comput. Graphics 26, 418–428. https://doi.org/10.1109/TVCG.2019.2934433

Fusco, S., Garasto, S., Corsonello, A., Vena, S., Mari, V., Gareri, P., Ruotolo, G., Luciani, F., Roncone, A., Maggio, M., Lattanzio, F., 2016. Medication-Induced Nephrotoxicity in Older Patients. Current Drug Metabolism 17, 608–625. https://doi.org/10.2174/1389200217666160406115959

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Trans. Syst., Man, Cybern. C 42, 463–484. https://doi.org/10.1109/TSMCC.2011.2161285

Gameiro, J., Branco, T., Lopes, J.A., 2020. Artificial Intelligence in Acute Kidney Injury Risk Prediction. J Clin Med 9. https://doi.org/10.3390/jcm9030678

Gandhi, T.K., Burstin, H.R., Cook, E.F., Puopolo, A.L., Haas, J.S., Brennan, T.A., Bates, D.W., 2000. Drug complications in outpatients. Journal of General Internal Medicine 15, 149–154. https://doi.org/10.1046/j.1525-1497.2000.04199.x

Georgaki-Angelaki, E., Stergiou, N., Naoum, E., Papassotiriou, I., Anagnostakou, M., 2009. Olmesartan medoxomil-induced acute renal failure in a premature newborn following maternal exposure during pregnancy: a case report and review of the literature. NDT plus 2, 295–297.

Gifi, A., 1990. Nonlinear multivariate analysis. Wiley.

Gildon, B., Condren, M., Hughes, C., 2019. Impact of Electronic Health Record Systems on Prescribing Errors in Pediatric Clinics. Healthcare 7, 57. https://doi.org/10.3390/healthcare7020057

Gilhooly, K.J., 2004. Working Memory and Reasoning., in: The Nature of Reasoning. Cambridge University Press, Gilhooly, Kenneth J.: Psychology Group, Human Sciences Dept., Brunei University, West London, Uxbridge, United Kingdom, UBS 3PH, ken.gilhooly@brunel.ac.uk, pp. 49–77.

Girman, C.J., Kou, T.D., Brodovicz, K., Alexander, C.M., O'Neill, E.A., Engel, S., Williams-Herman, D.E., Katz, L., 2012. Risk of acute renal failure in patients with Type 2 diabetes mellitus. Diabet. Med. 29, 614–621. https://doi.org/10.1111/j.1464-5491.2011.03498.x

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

Go, A.S., Parikh, C.R., Ikizler, T.A., Coca, S., Siew, E.D., Chinchilli, V.M., Hsu, C.-Y., Garg, A.X., Zappitelli, M., Liu, K.D., Reeves, W.B., Ghahramani, N., Devarajan, P., Faulkner, G.B., Tan, T.C., Kimmel, P.L., Eggers, P., Stokes, J.B., Assessment Serial Evaluation, and Subsequent Sequelae of Acute Kidney Injury Study Investigators, 2010. The assessment, serial evaluation, and subsequent sequelae of acute kidney injury (ASSESS-AKI) study: design and methods. BMC Nephrol 11, 22. https://doi.org/10.1186/1471-2369-11-22

Godin, Mélanie, Godin, M., Bouchard, J., Mehta, R.L., 2013. Fluid Balance in Patients with Acute Kidney Injury: Emerging Concepts. NEC 123, 238–245. https://doi.org/10.1159/000354713

Gois, P.H.F., Canale, D., Volpini, R.A., Ferreira, D., Veras, M.M., Andrade-Oliveira, V., Câmara, N.O.S., Shimizu, M.H.M., Seguro, A.C., 2016. Allopurinol attenuates rhabdomyolysis-associated acute kidney injury: Renal and muscular protection. Free Radical Biology and Medicine 101, 176–189. https://doi.org/10.1016/j.freeradbiomed.2016.10.012

Gottlieb, S.S., Abraham, W., Butler, J., Forman, D.E., Loh, E., Massie, B.M., O'connor, C.M., Rich, M.W., Stevenson, L.W., Young, J., Krumholz, H.M., 2002. The prognostic importance of different definitions of worsening renal function in congestive heart failure. Journal of cardiac failure 8, 136–41.

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

Graber, M.L., Byrne, C., Johnston, D., 2017. The impact of electronic health records on diagnosis. Diagnosis (Berl) 4, 211–223. https://doi.org/10.1515/dx-2017-0012

Green, T.M., Maciejewski, R., 2013. A role for reasoning in visual analytics, in: Proceedings of the Annual Hawaii International Conference on System Sciences. pp. 1495–1504. https://doi.org/10.1109/HICSS.2013.58

Greenacre, M., Blasius, J., 2006. Multiple correspondence analysis and related methods. CRC press.

Gresh, D.L., Rabenhorst, D.A., Shabo, A., Slavin, S., 2002. Prima: A case study of using information visualization techniques for patient record analysis, in: IEEE Visualization, 2002. VIS 2002. IEEE, pp. 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, 621–625. https://doi.org/10.1080/01421590310001605688

Gruchalla, R.S., 2000. Clinical assessment of drug-induced disease. Lancet 356, 1505–1511. https://doi.org/10.1016/S0140-6736(00)02885-3

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

Gupta, A., Puri, V., Sharma, R., Puri, S., 2012. Folic acid induces acute renal failure (ARF) by enhancing renal prooxidant state. Experimental and toxicologic pathology 64, 225–232.

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

Halpern, Y., Horng, S., Nathanson, L.A., Shapiro, N.I., Sontag, D., 2012. A comparison of dimensionality reduction techniques for unstructured clinical text, in: Icml 2012 Workshop on Clinical Data Analysis.

Han, J., Kamber, M., 2011. Data Mining: Concepts and Techniques.

Han, J., Kamber, M., Pei, J., 2011. Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems).

Hannan, T.J., 1999. Detecting adverse drug reactions to improve patient outcomes, in: International Journal of Medical Informatics. Elsevier Science Ireland Ltd, pp. 61–64. https://doi.org/10.1016/S1386-5056(99)00020-9

Haraty, R.A., Dimishkieh, M., Masud, M., 2015. An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data. International Journal of Distributed Sensor Networks 11, 615740. https://doi.org/10.1155/2015/615740

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, 100–108. 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, 47–55. https://doi.org/10.1016/j.jbi.2012.08.004

Heer, J., Kandel, S., 2012. Interactive analysis of big data. XRDS: Crossroads, The ACM Magazine for Students 19, 50. https://doi.org/10.1145/2331042.2331058

Heer, J., Perer, A., 2014. Orion: a system for modeling, transformation and visualization of multidimensional heterogeneous networks. Information Visualization 13, 111–133. https://doi.org/10.1177/1473871612462152

Hege, H.C., Hotz, I., Muntzner, T., 2009. iPCA: An Interactive System for PCA-based Visual Analytics.

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, 144–148. https://doi.org/10.1097/MLR.0000000000000038

Heuer, R.J., 1999. Psychology of intelligence analysis. Center for the Study of Intelligence, Central Intelligence Agency.

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, 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 to Explore Highly Structured Temporal Data. J. UCS 11, 1792–1805. https://doi.org/10.3217/jucs-011-11-1792

Hirschfeld, H.O., 1935. A Connection between Correlation and Contingency. Mathematical Proceedings of the Cambridge Philosophical Society 31, 520–524. https://doi.org/10.1017/S0305004100013517

Ho, Kwok M., Power, B.M., 2010. Benefits and risks of furosemide in acute kidney injury. Anaesthesia. https://doi.org/10.1111/j.1365-2044.2009.06228.x

Ho, K. M., Power, B.M., 2010. Benefits and risks of furosemide in acute kidney injury. Anaesthesia 65, 283–293. https://doi.org/10.1111/j.1365-2044.2009.06228.x

Honigman, B., Light, P., Pulling, R.M., Bates, D.W., 2001. A computerized method for identifying incidents associated with adverse drug events in outpatients. International Journal of Medical Informatics 61, 21–32. https://doi.org/10.1016/S1386-5056(00)00131-3

Horn, W., Popow, C., Unterasinger, L., 2001. Support for fast comprehension of ICU data: Visualization using metaphor graphics. Methods of information in medicine 40, 421–424.

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

Hsu, R.K., Hsu, C., 2016. THE ROLE OF ACUTE KIDNEY INJURY IN CHRONIC KIDNEY DISEASE. Semin Nephrol 36, 283–292. https://doi.org/10.1016/j.semnephrol.2016.05.005

Huang, C.W., Syed-Abdul, S., Jian, W.S., Iqbal, U., Nguyen, P.A., Lee, P., Lin, S.H., Hsu, W.D., Wu, M.S., Wang, C.F., Ma, K.L., Li, Y.C., 2015. A novel tool for visualizing chronic kidney disease associated polymorbidity: A 13-year cohort study in Taiwan. Journal of the American Medical Informatics Association 22, 290–298. https://doi.org/10.1093/jamia/ocu044

Hwang, Y.J., Dixon, S.N., Reiss, J.P., Wald, R., Parikh, C.R., Gandhi, S., Shariff, S.Z., Pannu, N., Nash, D.M., Rehman, F., 2014. Atypical antipsychotic drugs and the risk for acute kidney injury and other adverse outcomes in older adults: a population-based cohort study. Annals of internal medicine 161, 242–248.

Ibrahim, N.E., McCarthy, C.P., Shrestha, S., Gaggin, H.K., Mukai, R., Magaret, C.A., Rhyne, R.F., Januzzi, J.L., 2019. A clinical, proteomics, and artificial intelligence-driven model to predict acute kidney injury in patients undergoing coronary angiography. Clin Cardiol 42, 292–298. https://doi.org/10.1002/clc.23143

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 J 66, 587–592. https://doi.org/10.1016/j.ihj.2014.05.027

Jacob, K.A., Leaf, D.E., Dieleman, J.M., Dijk, D. van, Nierich, A.P., Rosseel, P.M., Maaten, J.M. van der, Hofland, J., Diephuis, J.C., Lange, F. de, Boer, C., Kluin, J., Waikar, S.S., 2015. Intraoperative High-Dose Dexamethasone and Severe AKI after Cardiac Surgery. JASN 26, 2947–2951. https://doi.org/10.1681/ASN.2014080840

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, Award winning papers from the 19th International Conference on Pattern Recognition (ICPR) 31, 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Japkowicz, N., Shah, M., 2011. Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press.

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. https://doi.org/10.1186/s13673-015-0023-4

Jha, P.K., Vankalakunti, M., Siddini, V., Bonu, R., Prakash, G.K., Babu, K., Ballal, H.S., 2013. Sunitinib induced nephrotic syndrome and thrombotic microangiopathy. Indian journal of nephrology 23, 67.

Jiang, Y., McCombs, J.S., Park, S.H., 2017. A Retrospective Cohort Study of Acute Kidney Injury Risk Associated with Antipsychotics. CNS Drugs 31, 319–326. https://doi.org/10.1007/s40263-017-0421-4

Juncos, L.A., Juncos, L.I., 2016. Mineralocorticoid receptor antagonism in AKI: a new hope? Am Soc Nephrol.

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

Kameshwaran, K., Malarvizhi, K., 2014. Survey on Clustering Techniques in Data Mining 5, 5.

Kandasamy, K., Chuah, J.K.C., Su, R., Huang, P., Eng, K.G., Xiong, S., Li, Y., Chia, C.S., Loo, L.-H., Zink, D., 2015. Prediction of drug-induced nephrotoxicity and injury mechanisms with human induced pluripotent stem cell-derived cells and machine learning methods. Sci Rep 5. https://doi.org/10.1038/srep12337

Kandler, K., Jensen, M.E., Nilsson, J.C., Møller, C.H., Steinbrüchel, D.A., 2014. Acute kidney injury is independently associated with higher mortality after cardiac surgery. Journal of Cardiothoracic and Vascular Anesthesia 28, 1448–1452. https://doi.org/10.1053/j.jvca.2014.04.019

Kane-Gill, S.L., Sileanu, F.E., Murugan, R., Trietley, G.S., Handler, S.M., Kellum, J.A., 2015. Risk factors for acute kidney injury in older adults with critical illness: a retrospective cohort study. Am. J. Kidney Dis. 65, 860–869. https://doi.org/10.1053/j.ajkd.2014.10.018

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

Karajala, V., Mansour, W., Kellum, J.A., 2009. Diuretics in acute kidney injury. Minerva Anestesiol 75, 251–257.

Kate, R.J., Perez, R.M., Mazumdar, D., Pasupathy, K.S., Nilakantan, V., 2016. Prediction and detection models for acute kidney injury in hospitalized older adults. BMC Med Inform Decis Mak 16. https://doi.org/10.1186/s12911-016-0277-4

Kaufman, J., Dhakal, M., Patel, B., Hamburger, R., 1991. Community-Acquired Acute Renal Failure. American Journal of Kidney Diseases 17, 191–198. https://doi.org/10.1016/S0272-6386(12)81128-0

Kehrer, J., Hauser, H., 2013. Visualization and visual analysis of multifaceted scientific data: A survey. IEEE Transactions on Visualization and Computer Graphics. https://doi.org/10.1109/TVCG.2012.110

Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F., 2010a. 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, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 76–90. https://doi.org/10.1007/978-3-540-71080-6_6

Keim, D., Mansmann, F., Thomas, J., 2010b. Visual Analytics : How Much Visualization and How Much Analytics ? ACM SIGKDD Explorations Newsletter 11, 5–8. https://doi.org/10.1145/1809400.1809403

Khalid, S., Judge, A., Pinedo-Villanueva, R., 2018. An Unsupervised Learning Model for Pattern Recognition in Routinely Collected Healthcare Data:, in: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies. Presented at the 11th International Conference on Health Informatics, SCITEPRESS - Science and Technology Publications, Funchal, Madeira, Portugal, pp. 266–273. https://doi.org/10.5220/0006535602660273

Khan, S., Loi, V., Rosner, M.H., 2017. Drug-Induced Kidney Injury in the Elderly. Drugs and Aging. https://doi.org/10.1007/s40266-017-0484-4

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 Annu Symp Proc 2007, 404–408.

Klimov, D., Shahar, Y., Taieb-Maimon, M., 2010. Intelligent visualization and exploration of time-oriented data of multiple patients. Artificial Intelligence in Medicine 49, 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 22, 275–289. https://doi.org/10.1136/amiajnl-2014-002927

Koh, H.C., Tan, G., 2005. Data mining applications in healthcare. Journal of healthcare information management : JHIM 19, 64–72. https://doi.org/10.4314/ijonas.v5i1.49926

Kohli, H.S., Bhaskaran, M.C., Muthukumar, T., Thennarasu, K., Sud, K., Jha, V., Gupta, K.L., Sakhuja, V., 2000. Treatment-related acute renal failure in the elderly: a hospital-based prospective study. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association 15, 212–7. https://doi.org/10.1093/ndt/15.2.212

Kolhe, N. V, Muirhead, A.W., Wilkes, S.R., Fluck, R.J., Taal, M.W., 2016. The epidemiology of hospitalised acute kidney injury not requiring dialysis in England from 1998 to 2013: retrospective analysis of hospital episode statistics. International journal of clinical practice 70, 330–9. https://doi.org/10.1111/ijcp.12774

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

Kosara, R., 2010. Turning a table into a tree: Growing parallel sets into a purposeful project. Beautiful Visualization: Looking at Data through the Eyes of Experts, Steele J., Iliinsky N.,(Eds.). O’Reilly 193–204.

Kosara, R., Miksch, S., 2002. Visualization methods for data analysis and planning in medical applications, in: International Journal of Medical Informatics. pp. 141–153. https://doi.org/10.1016/S1386-5056(02)00072-2

Kruskal, J.B., 1964. Nonmetric multidimensional scaling: A numerical method. Psychometrika 29, 115–129. https://doi.org/10.1007/BF02289694

Kumar, M., Stoll, N., Kaber, D., Thurow, K., Stoll, R., 2007. Fuzzy filtering for an intelligent interpretation of medical data, in: 2007 IEEE International Conference on Automation Science and Engineering. Presented at the 2007 IEEE

International Conference on Automation Science and Engineering, pp. 225–230. https://doi.org/10.1109/COASE.2007.4341714

Kumar, S., Allen, D.A., Kieswich, J.E., Patel, N.S.A., Harwood, S., Mazzon, E., Cuzzocrea, S., Raftery, M.J., Thiemermann, C., Yaqoob, M.M., 2009. Dexamethasone Ameliorates Renal Ischemia-Reperfusion Injury. JASN 20, 2412–2425. https://doi.org/10.1681/ASN.2008080868

Kusiak, A., 2001. Feature transformation methods in data mining. IEEE Transactions on Electronics Packaging Manufacturing 24, 214–221. https://doi.org/10.1109/6104.956807

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, 95–108. https://doi.org/10.1002/bdm.1788

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, 10–10. https://doi.org/10.1186/1472-6947-12-10

Lavado, R., Hayrapetyan, S., Kharazyan, S., 2018. Expansion of the Benifits Package: The Experience of Armenia, The World Bank Group. https://doi.org/10.1145/958491.958516

Leaf, D.E., Swinkels, D.W., 2016. Catalytic iron and acute kidney injury. American Journal of Physiology-Renal Physiology 311, F871–F876.

Lee, C.H., Yoon, H.J., 2017. Medical big data: Promise and challenges. Kidney Research and Clinical Practice 36, 3–11. https://doi.org/10.23876/j.krcp.2017.36.1.3

Leighton, J.P., 2004. Defining and Describing Reason., in: The Nature of Reasoning. Cambridge University Press, Leighton, Jacqueline P.: Centre for Research in Applied Measurement and Evaluation (CRAME), Department of Educational Psychology, University of Alberta, 6-110 Education North, Edmonton, AB, Canada, T6G 2G5, Jacqueline.Leighton@ualberta.ca, pp. 3–11.

Lesselroth, B.J., Pieczkiewicz, D.S., 2011. Data visualization strategies for the electronic health record. Nova Science Publishers, Inc.

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. Canadian Journal of Clinical Pharmacology 10, 67–71.

Lewis, D.D., 1998. Naive (Bayes) at forty: The independence assumption in information retrieval, in: Nédellec, C., Rouveirol, C. (Eds.), Machine Learning: ECML-98, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 4–15. https://doi.org/10.1007/BFb0026666

Lex, A., Streit, M., Partl, C., Kashofer, K., Schmalstieg, D., 2010. Comparative Analysis of Multidimensional, Quantitative Data. IEEE Trans Vis Comput Graph 16, 1027–1035. https://doi.org/10.1109/TVCG.2010.138

Li, X., Wang, Y., 2016. Adaptive online monitoring for ICU patients by combining just-in-time learning and principal component analysis. J Clin Monit Comput 30, 807–820. https://doi.org/10.1007/s10877-015-9778-4

Liao, M., Li, Y., Kianifard, F., Obi, E., Arcona, S., 2016. Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. BMC Nephrol 17. https://doi.org/10.1186/s12882-016-0238-2

Lieske, J.C., Chawla, L., Kashani, K., Kellum, J.A., Koyner, J.L., Mehta, R.L., 2014. Biomarkers for acute kidney injury: where are we today? Where should we go? Clin. Chem. 60, 294–300. https://doi.org/10.1373/clinchem.2012.201988

Lipson, E.J., Huff, C.A., Holanda, D.G., McDevitt, M.A., Fine, D.M., 2010. Lenalidomide-induced acute interstitial nephritis. The oncologist 15, 961.

Liu, J., Sun, G., He, Y., Song, F., Chen, S., Guo, Z., Liu, B., Lei, L., He, L., Chen, J., 2019. Early β-blockers administration might be associated with a reduced risk of contrast-induced acute kidney injury in patients with acute myocardial infarction. Journal of thoracic disease 11, 1589.

Liu, K.D., Yang, J., Tan, T.C., Glidden, D. V, Zheng, S., Pravoverov, L., Hsu, C.-Y., Go, A.S., 2019. Risk Factors for Recurrent Acute Kidney Injury in a Large Population-Based Cohort. American journal of kidney diseases : the official journal of the National Kidney Foundation 73, 163–173. https://doi.org/10.1053/j.ajkd.2018.08.008

Liu, S., Joseph, K.S., Bartholomew, S., Fahey, J., Lee, L., Allen, A.C., Kramer, M.S., Sauve, R., Young, D.C., Liston, R.M., Kirby, R., León, J.A., 2010. Temporal Trends and Regional Variations in Severe Maternal Morbidity in Canada, 2003 to 2007. Journal of Obstetrics and Gynaecology Canada 32, 847–855. https://doi.org/10.1016/S1701-2163(16)34656-4

Loboz, K.K., Shenfield, G.M., 2005. Drug combinations and impaired renal function - The “triple whammy.” British Journal of Clinical Pharmacology 59, 239–243. https://doi.org/10.1111/j.0306-5251.2004.2188.x

Lopau, K., Hefner, L., Bender, G., Heidbreder, E., Wanner, C., 2001. Haemodynamic effects of valsartan in acute renal ischaemia/reperfusion injury. Nephrology Dialysis Transplantation 16, 1592–1597.

L’Yi, S., Ko, B., Shin, D., Cho, Y.-J., Lee, J., Kim, B., Seo, J., 2015. XCluSim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. BMC bioinformatics 16, S5.

Lysenko, A., Sharma, A., Boroevich, K.A., Tsunoda, T., 2018. An integrative machine learning approach for prediction of toxicity-related drug safety. Life Science Alliance 1. https://doi.org/10.26508/lsa.201800098

Maaten, L. van der, Hinton, G., 2008. Visualizing Data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

Mackowski, A., Chen, H.-K., Levitt, M., 2015. Successful management of chronic high-output ileostomy with high dose loperamide. BMJ Case Rep 2015. https://doi.org/10.1136/bcr-2015-209411

Malbrain, M.L.N.G., Lambrecht, G.L.Y., Daelemans, R., Lins, R.L., Hermans, P., Zachee, P., 1994. Acute renal failure due to bilateral lymphomatous infiltrates - Primary extranodal non-Hodgkin’s lymphoma (p-EN-NHL) of the kidneys: Does it really exist? Clinical Nephrology 42, 163–169.

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, in: Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI ’15. ACM, New York, NY, USA, pp. 38–49. https://doi.org/10.1145/2678025.2701407

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, in: EHRVis Workshop on Visualizing Electronic Health Record Data at VIS.

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, 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, in: Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics - IHI ’12. Presented at the the 2nd ACM SIGHIT symposium, ACM Press, Miami, Florida, USA, p. 389. https://doi.org/10.1145/2110363.2110408

Martines, A.M.F., Masereeuw, R., Tjalsma, H., Hoenderop, J.G., Wetzels, J.F.M., Swinkels, D.W., 2013. Iron metabolism in the pathogenesis of iron-induced kidney injury. Nature Reviews Nephrology. https://doi.org/10.1038/nrneph.2013.98

Matheny, M.E., Miller, R.A., Ikizler, T.A., Waitman, L.R., Denny, J.C., Schildcrout, J.S., Dittus, R.S., Peterson, J.F., 2010. Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. Med Decis Making 30, 639–650. https://doi.org/10.1177/0272989X10364246

McCallum, A., Nigam, K., 1998. A Comparison of Event Models for Naive Bayes Text Classification 8.

McCoy, A.B., Waitman, L.R., Gadd, C.S., Danciu, I., Smith, J.P., Lewis, J.B., Schildcrout, J.S., Peterson, J.F., 2010. A computerized provider order entry intervention for medication safety during acute kidney injury: a quality improvement report. American journal of kidney diseases 56, 832–841.

McLachlan, G.J., 1992. Cluster analysis and related techniques in medical research. Statistical Methods in Medical Research 1, 27–48.

Mehrabadi, A., Liu, S., Bartholomew, S., Hutcheon, J.A., Magee, L.A., Kramer, M.S., Liston, R.M., Joseph, K.S., Canadian Perinatal Surveillance System Public Health Agency of Canada, 2014. Hypertensive disorders of pregnancy and the recent increase in obstetric acute renal failure in Canada: population based retrospective cohort study. BMJ (Clinical research ed.) 349, g4731. https://doi.org/10.1136/bmj.g4731

Mehta, R.L., 2011. Management of Acute Kidney Injury: It’s the Squeaky Wheel That Gets the Oil! CJASN 6, 2102–2104. https://doi.org/10.2215/CJN.07720811

Mehta, R.L., Pascual, M.T., Soroko, S., Savage, B.R., Himmelfarb, J., Ikizler, T.A., Paganini, E.P., Chertow, G.M., 2004. Spectrum of acute renal failure in the intensive care unit: The PICARD experience. Kidney International 66, 1613–1621. https://doi.org/10.1111/j.1523-1755.2004.00927.x

Miller, A., Price, G., 2009. Gabapentin toxicity in renal failure: the importance of dose adjustment. Pain medicine 10, 190–192.

Mishima, E., Maruyama, K., Nakazawa, T., Abe, T., Ito, S., 2017. Acute kidney injury from excessive potentiation of calcium-channel blocker via synergistic CYP3A4 inhibition by clarithromycin plus voriconazole. Internal Medicine 56, 1687–1690.

Mitsuhiro, M., Yadohisa, H., 2015. Reduced k-means clustering with MCA in a low-dimensional space. Computational Statistics 30, 463–475. https://doi.org/10.1007/s00180-014-0544-8

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, in: Proceedings of the Workshop on Advanced Visual Interfaces AVI. Association for Computing Machinery, pp. 237–244. https://doi.org/10.1145/2598153.2598156

Miyahara, T., 1978. Drug-induced renal disorders. Nippon rinsho. Japanese journal of clinical medicine Suppl, 2320–2321. https://doi.org/10.12861/jrip.2015.12

Moffett, B.S., Goldstei, S.L., 2011. Acute kidney injury and increasing nephrotoxic-medication exposure in noncritically-Ill children. Clinical Journal of the American Society of Nephrology 6, 856–863. https://doi.org/10.2215/CJN.08110910

Mohamadlou, H., Lynn-Palevsky, A., Barton, C., Chettipally, U., Shieh, L., Calvert, J., Saber, N.R., Das, R., 2018. Prediction of Acute Kidney Injury With a Machine Learning Algorithm Using Electronic Health Record Data. Can J Kidney Health Dis 5. https://doi.org/10.1177/2054358118776326

Monroe, M., Lan, R., Lee, H., Plaisant, C., Shneiderman, B., 2013. Temporal event sequence simplification. IEEE transactions on visualization and computer graphics 19, 2227–2236.

Morgan, S.G., Hunt, J., Rioux, J., Proulx, J., Weymann, D., Tannenbaum, C., 2016. Frequency and cost of potentially inappropriate prescribing for older adults: a cross-sectional study. CMAJ Open 4, E346–E351. https://doi.org/10.9778/cmajo.20150131

Mori, Y., Matsubara, H., Nose, A., Shibasaki, Y., Masaki, H., Kosaki, A., Okigaki, M., Fujiyama, S., Tanaka-Uchiyama, Y., Hasegawa, T., Iba, O., Tateishi, E., Amano, K., Iwasaka, T., 2001. Safety and availability of doxazosin in treating hypertensive patients with chronic renal failure. Hypertens. Res. 24, 359–363. https://doi.org/10.1291/hypres.24.359

Muller, M., 2007. Participatory Design. pp. 1061–1081. https://doi.org/10.1201/9781410615862.ch54

Munsaka, M.S., 2017. Leveraging Machine Learning in the Analysis of Safety Data in Drug Research and Healthcare Informatics.

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

Nadkarni, G.N., Patel, A.A., Ahuja, Y., Annapureddy, N., Agarwal, S.K., Simoes, P.K., Konstantinidis, I., Kamat, S., Archdeacon, M., Thakar, C. V, 2016. Incidence, Risk Factors, and Outcome Trends of Acute Kidney Injury in Elective Total Hip and Knee Arthroplasty. American journal of orthopedics (Belle Mead, N.J.) 45, E12–E19.

Nam, E.J., Han, Y., Mueller, K., Zelenyuk, A., Imre, D., 2007. ClusterSculptor: A Visual Analytics Tool for High-Dimensional Data, in: 2007 IEEE Symposium on Visual Analytics Science and Technology. Presented at the 2007 IEEE Symposium on Visual Analytics Science and Technology, pp. 75–82. https://doi.org/10.1109/VAST.2007.4388999

Nandikanti, D.K., Gosmanova, E.O., Gosmanov, A.R., 2016. Acute kidney injury associated with linagliptin. Case reports in endocrinology 2016.

Nash, K., Hafeez, A., Hou, S., 2002. Hospital-acquired renal insufficiency. American Journal of Kidney Diseases 39, 930–936. https://doi.org/10.1053/ajkd.2002.32766

Neugarten, J., Golestaneh, L., 2018. Female sex reduces the risk of hospital-associated acute kidney injury: a meta-analysis. BMC Nephrology 19, 314. https://doi.org/10.1186/s12882-018-1122-z

Neyra, J.A., Rocha, N.A., Bhargava, R., Vaidya, O.U., Hendricks, A.R., Rodan, A.R., 2015. Rhabdomyolysis-induced acute kidney injury in a cancer patient exposed to denosumab and abiraterone: a case report. BMC nephrology 16, 118.

Ng, R.R.G., Tan, G.H.J., Liu, W., Ti, L.K., Chew, S.T.H., 2016. The Association of Acute Kidney Injury and Atrial Fibrillation after Cardiac Surgery in an Asian Prospective Cohort Study. Medicine (Baltimore) 95, e3005. https://doi.org/10.1097/MD.0000000000003005

Nielsen, F., 2016. Hierarchical Clustering, in: Nielsen, F. (Ed.), Introduction to HPC with MPI for Data Science, Undergraduate Topics in Computer Science. Springer International Publishing, Cham, pp. 195–211. 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. https://doi.org/10.5210/ojphi.v11i2.10114

Niuniu, X., Yuxun, L., 2010. Review of decision trees. pp. 105–109. https://doi.org/10.1109/ICCSIT.2010.5564437

Obaid, H.S., Dheyab, S.A., Sabry, S.S., 2019. The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning, in: 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON). IEEE, pp. 279–283.

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

Ola, O., Sedig, K., 2014. The Challenge of Big Data in Public Helth: An Opportunity for Visual Analytics. Online Journal of Public Health Informatics 5. https://doi.org/10.5210/ojphi.v5i3.4933

Olsson, D., Sartipy, U., Braunschweig, F., Holzmann, M.J., 2013. Acute kidney injury following coronary artery bypass surgery and long-term risk of heart failure. Circ Heart Fail 6, 83–90. https://doi.org/10.1161/CIRCHEARTFAILURE.112.971705

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, 7–1.

Ozturk, S., Kayaalp, M., McDonald, C.J., 2014. Visualization of patient prescription history data in emergency care. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium 2014, 963–968.

Palevsky, P.M., Liu, K.D., Brophy, P.D., Chawla, L.S., Parikh, C.R., Thakar, C.V., Tolwani, A.J., Waikar, S.S., Weisbord, S.D., 2013. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for acute kidney injury. Am. J. Kidney Dis. 61, 649–672. https://doi.org/10.1053/j.ajkd.2013.02.349

Pannu, N., Nadim, M.K., 2008. An overview of drug-induced acute kidney injury. Critical care medicine 36, S216-23. https://doi.org/10.1097/CCM.0b013e318168e375

Parsons, P., Sedig, K., 2014. Distribution of information processing while performing complex cognitive activities with visualization tools, in: Handbook of Human Centric Visualization. Springer New York, pp. 693–715. https://doi.org/10.1007/978-1-4614-7485-2_28

Parsons, P., Sedig, K., Mercer, R.E., Khordad, M., Knoll, J., Rogan, P., 2015. Visual Analytics for supporting evidence-based interpretation of molecular cytogenomic findings, in: ACM International Conference Proceeding Series. Association for

Computing Machinery, New York, New York, USA, pp. 1–8. https://doi.org/10.1145/2836034.2836036

Perazella, M.A., 2015. The Urine Sediment as a Biomarker of Kidney Disease. American Journal of Kidney Diseases 66, 748–755. https://doi.org/10.1053/j.ajkd.2015.02.342

Perer, A., Sun, J., 2012. MatrixFlow: temporal network visual analytics to track symptom evolution during disease progression. AMIA Annu Symp Proc 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

Peres, L.A.B., da Cunha, A.D., 2013. Acute nephrotoxicity of cisplatin: molecular mechanisms. Jornal brasileiro de nefrologia : ′orgão oficial de Sociedades Brasileira e Latino-Americana de Nefrologia. https://doi.org/10.5935/0101-2800.20130052

Perez-Ruiz, F., 2017. Treatment with Allopurinol is Associated with Lower Risk of Acute Kidney Injury in Patients with Gout: A Retrospective Analysis of a Nested Cohort. Rheumatology and Therapy 4, 419–425. https://doi.org/10.1007/s40744-017-0082-2

Peskoe, S.T., McMillan, J.H., Lorch, A., Sussman, H., Ozawa, T., 1978. Reversible acute renal failure associated with chlorthalidone therapy: possible drug-induced interstitial nephritis. J Med Assoc Ga 67, 17–18.

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, in: AMIA Annual Symposium Proceedings. American Medical Informatics Association, p. 598.

Pierson-Marchandise, M., Gras, V., Moragny, J., Micallef, J., Gaboriau, L., Picard, S., Choukroun, G., Masmoudi, K., Liabeuf, S., 2017. The drugs that mostly frequently induce acute kidney injury: a case − noncase study of a pharmacovigilance database. British Journal of Clinical Pharmacology 83, 1341–1349. https://doi.org/10.1111/bcp.13216

Pike, W.A., Stasko, J., Chang, R., O’Connell, T.A., 2009. The science of interaction. Information Visualization 8, 263–274. https://doi.org/10.1057/ivs.2009.22

Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D., Shneiderman, B., 1998. LifeLines: using visualization to enhance navigation and analysis, in: 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: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (Eds.), Human-Computer Interaction – INTERACT 2011. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 292–299. https://doi.org/10.1007/978-3-642-23768-3_24

Porter, C.J., Juurlink, I., Bisset, L.H., Bavakunji, R., Mehta, R.L., Devonald, M.A.J., 2014. A real-time electronic alert to improve detection of acute kidney injury in a large teaching hospital. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association 29, 1888–1893. https://doi.org/10.1093/ndt/gfu082

Pozzoli, S., Simonini, M., Manunta, P., 2018. Predicting acute kidney injury: current status and future challenges. J Nephrol 31, 209–223. https://doi.org/10.1007/s40620-017-0416-8

Proctor, R., Vu, K., 2012. Human-Computer Interaction Fundamentals Human Factors and Ergonomics Handbook of Human Factors in Web Design Handbook of Standards and Guidelines in Ergonomics and Human Factors.

Puri, N., Mohey, V., Singh, M., Kaur, T., Pathak, D., Buttar, H.S., Singh, A.P., 2016. Dipyridamole attenuates ischemia reperfusion induced acute kidney injury through adenosinergic A1 and A2A receptor agonism in rats. Naunyn-Schmiedeberg's archives of pharmacology 389, 361–368.

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 Inf Sci Syst 2, 3. https://doi.org/10.1186/2047-2501-2-3

Rammohan, M., Kalantar-Zadeh, K., Liang, A., Ghossein, C., 2005. Megestrol Acetate in a Moderate Dose for the Treatment of Malnutrition-Inflammation Complex in Maintenance Dialysis Patients. Journal of Renal Nutrition 15, 345–355. https://doi.org/10.1016/j.jrn.2004.10.006

Rashidi, H.H., Sen, S., Palmieri, T.L., Blackmon, T., Wajda, J., Tran, N.K., 2020. Early Recognition of Burn- and Trauma-Related Acute Kidney Injury: A Pilot Comparison of Machine Learning Techniques. Sci Rep 10, 205. https://doi.org/10.1038/s41598-019-57083-6

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, 7. https://doi.org/10.3390/mti4010007

Rind, A., 2013. Interactive Information Visualization to Explore and Query Electronic Health Records. Foundations and Trends® in Human–Computer Interaction 5, 207–298. https://doi.org/10.1561/1100000039

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

Rind, Alexander, Aigner, W., Miksch, S., Wiltner, S., Pohl, M., Turic, T., Drexler, F., 2011b. Visual exploration of time-oriented patient data for chronic diseases: design study and evaluation. Information Quality in e- … 301–320. https://doi.org/10.1007/978-3-642-25364-5_22

Rind, A., Wagner, M., Aigner, W., 2019. Towards a Structural Framework for Explicit Domain Knowledge in Visual Analytics. 2019 IEEE Workshop on Visual Analytics in Healthcare (VAHC) 33–40. https://doi.org/10.1109/VAHC47919.2019.8945032

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. Found. Trends Hum.-Comput. Interact. 5, 207–298. https://doi.org/10.1561/1100000039

Rind, A, Wang, T.D., Aigner, W, Miksch, S, Wongsuphasawat, K, Plaisant, C, Shneiderman, B, B., David Wang, T., Aigner, Wolfgang, Miksch, Silvia, Wongsuphasawat, Krist, Plaisant, Catherine, Shneiderman, Ben, 2011. Interactive Information Visualization to Explore and Query Electronic Health Records. Foundations and Trends R in Human-Computer Interaction 5, 207–298. https://doi.org/10.1561/1100000039

Rinner, C., Grossmann, W., Sauter, S.K., Wolzt, M., Gall, W., 2015. Effects of Shared Electronic Health Record Systems on Drug-Drug Interaction and Duplication Warning Detection. BioMed Research International 2015. https://doi.org/10.1155/2015/380497

Rivosecchi, Ryan M., Kellum, J.A., Dasta, J.F., Armahizer, M.J., Bolesta, S., Buckley, M.S., Dzierba, A.L., Frazee, E.N., Johnson, H.J., Kim, C., Murugan, R., Smithburger, P.L., Wong, A., Kane Gill, S.L., 2016. Drug Class Combination–Associated Acute Kidney Injury. Annals of Pharmacotherapy. https://doi.org/10.1177/1060028016657839

Rivosecchi, Ryan M, Kellum, J.A., Dasta, J.F., Armahizer, M.J., Bolesta, S., Buckley, M.S., Dzierba, A.L., Frazee, E.N., Johnson, H.J., Kim, C., Murugan, R., Smithburger, P.L., Wong, A., Kane Gill, S.L., 2016. Drug Class Combination-Associated Acute Kidney Injury. The Annals of pharmacotherapy 50, 953–972. https://doi.org/10.1177/1060028016657839

Rocci, R., Gattone, S.A., Vichi, M., 2011. A new dimension reduction method: Factor discriminant k-means. Journal of classification 28, 210–226.

Rokach, L., Maimon, O., 2005. Clustering Methods, in: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, Boston, MA, pp. 321–352. https://doi.org/10.1007/0-387-25465-X_15

Ronan, T., Qi, Z., Naegle, K.M., 2016. Avoiding common pitfalls when clustering biological data. Sci Signal 9, re6. https://doi.org/10.1126/scisignal.aad1932

Rp, B., 2019. Metolazone and Furosemide Combination in Cardiorenal Syndrome: Short-Term Safety and Efficacy Among Admitted Patients in a Tertiary Hospital. https://doi.org/10.19080/jojun.2018.06.555686

RStudio | Open source & professional software for data science teams [WWW Document], n.d. URL https://rstudio.com/ (accessed 2.19.20).

Ruan, T., Lei, L., Zhou, Y., Zhai, J., Zhang, L., He, P., Gao, J., 2019. Representation learning for clinical time series prediction tasks in electronic health records. BMC Medical Informatics and Decision Making 19, 259. https://doi.org/10.1186/s12911-019-0985-7

Rydén, L., Sartipy, U., Evans, M., Holzmann, M.J., 2014. Acute kidney injury after coronary artery bypass grafting and long-term risk of end-stage renal disease. Circulation 130, 2005–2011. https://doi.org/10.1161/CIRCULATIONAHA.114.010622

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, in: Computers in Cardiology. Presented at the Computers in Cardiology, pp. 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, 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, 114–121.

Sari, A., 2019. Nephrotoxic Effects of Drugs. Poisoning in the Modern World - New Tricks for an Old Dog? https://doi.org/10.5772/intechopen.83644

Saruta, T., Kanno, Y., Hayashi, K., Suzuki, H., 1995. Renal effects of amlodipine. Journal of human hypertension 9, S11–6.

SAS Enterprise BI Server [WWW Document], n.d. URL https://www.sas.com/en_ca/software/enterprise-bi-server.html (accessed 2.19.20).

Schetz, M., Dasta, J., Goldstein, S., Golper, T., 2005. Drug-induced acute kidney injury. Current Opinion in Critical Care. https://doi.org/10.1097/01.ccx.0000184300.68383.95

Schmider, J., Kumar, K., LaForest, C., Swankoski, B., Naim, K., Caubel, P.M., 2019. Innovation in Pharmacovigilance: Use of Artificial Intelligence in Adverse Event Case Processing. Clinical Pharmacology & Therapeutics 105, 954–961. https://doi.org/10.1002/cpt.1255

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

Sedig, K., Parsons, P., 2016. Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework. Synthesis Lectures on Visualization 4, 1–185. https://doi.org/10.2200/s00685ed1v01y201512vis005

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, 84–133. https://doi.org/10.17705/1thci.00055

Sedig, K., Parsons, P., Babanski, A., 2012. Towards a Characterization of Interactivity in Visual Analytics. JMPT 3, 12–28.

Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A., 2010. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE Trans. Syst., Man, Cybern. A 40, 185–197. https://doi.org/10.1109/TSMCA.2009.2029559

Selby, Nicholas M., Crowley, L., Fluck, R.J., McIntyre, C.W., Monaghan, J., Lawson, N., Kolhe, N. V., 2012. Use of electronic results reporting to diagnose and monitor AKI in hospitalized patients. Clinical Journal of the American Society of Nephrology 7, 533–540. https://doi.org/10.2215/CJN.08970911

Selby, N. M., Crowley, L., Fluck, R.J., McIntyre, C.W., Monaghan, J., Lawson, N., Kolhe, N. V., 2012. Use of Electronic Results Reporting to Diagnose and Monitor AKI in Hospitalized Patients. Clinical Journal of the American Society of Nephrology 7, 533–540. https://doi.org/10.2215/CJN.08970911

Sembiring, R.W., Zain, J.M., Embong, A., 2011. Dimension Reduction of Health Data Clustering. arXiv:1110.3569 [cs].

Seo, J., Shneiderman, B., 2003. Interactively Exploring Hierarchical Clustering Results, in: The Craft of Information Visualization. Elsevier, pp. 334–340. https://doi.org/10.1016/B978-155860915-0/50042-1

Shepard, R.N., 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. Psychometrika 27, 219–246. https://doi.org/10.1007/BF02289621

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 J Biomed Health Inform 22, 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

Shulenberger, C.E., Jiang, A., Devabhakthuni, S., Ivaturi, V., Liu, T., Reed, B.N., 2016. Efficacy and Safety of Intravenous Chlorothiazide versus Oral Metolazone in Patients with Acute Decompensated Heart Failure and Loop Diuretic Resistance. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy 36, 852–860. https://doi.org/10.1002/phar.1798

Siddiqui, N.F., Coca, S.G., Devereaux, P.J., Jain, A.K., Li, L., Luo, J., Parikh, C.R., Paterson, M., Thiessen Philbrook, H., Wald, R., Walsh, M., Whitlock, R., Garg, A.X., 2012. Secular trends in acute dialysis after elective major surgery - 1995 to 2009. CMAJ 184, 1237–1245. https://doi.org/10.1503/cmaj.110895

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

Siew, E.D., Fissell, W.H., Tripp, C.M., Blume, J.D., Wilson, M.D., Clark, A.J., Vincz, A.J., Ely, E.W., Pandharipande, P.P., Girard, T.D., 2017. Acute Kidney Injury as a Risk Factor for Delirium and Coma during Critical Illness. Am J Respir Crit Care Med 195, 1597–1607. https://doi.org/10.1164/rccm.201603-0476OC

Siew, E.D., Parr, S.K., Abdel-Kader, K., Eden, S.K., Peterson, J.F., Bansal, N., Hung, A.M., Fly, J., Speroff, T., Ikizler, T.A., Matheny, M.E., 2016. Predictors of Recurrent AKI. Journal of the American Society of Nephrology : JASN 27, 1190–200. https://doi.org/10.1681/ASN.2014121218

Simpao, A.F., Ahumada, L.M., Desai, B.R., Bonafide, C.P., Gálvez, J.A., Rehman, M.A., Jawad, A.F., Palma, K.L., Shelov, E.D., 2015. Optimization of drug-drug interaction alert rules in a pediatric hospital’s electronic health record system using a visual analytics dashboard. J Am Med Inform Assoc 22, 361–369. https://doi.org/10.1136/amiajnl-2013-002538

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. 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. J Med Syst 38, 45. https://doi.org/10.1007/s10916-014-0045-x

Singer, A., Duarte Fernandez, R., 2015. The effect of electronic medical record system use on communication between pharmacists and prescribers. BMC Family Practice 16. https://doi.org/10.1186/s12875-015-0378-7

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

Sorzano, C.O.S., Vargas, J., Montano, A.P., 2014. A survey of dimensionality reduction techniques. arXiv:1403.2877 [cs, q-bio, stat].

Soyiri, I.N., Reidpath, D.D., 2013. An overview of health forecasting. Environ Health Prev Med 18, 1–9. https://doi.org/10.1007/s12199-012-0294-6

Spasic, I., Nenadic, G., 2020. Clinical Text Data in Machine Learning: Systematic Review. JMIR Medical Informatics 8, e17984. https://doi.org/10.2196/17984

Spence, R., 2002. Sensitivity encoding to support information space navigation: A design guideline. Information Visualization 1, 120–129. https://doi.org/10.1057/palgrave.ivs.9500019

Stasko, J., Görg, C., Liu, Z., 2008. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. Information Visualization 7, 118–132. https://doi.org/10.1057/palgrave.ivs.9500180

Steinhäuslin, F., Munafo, A., Buclin, T., Macciocchi, A., Biollaz, J., 1993. Renal effects of nimesulide in furosemide-treated subjects. Drugs 46 Suppl 1, 257–62. https://doi.org/10.2165/00003495-199300461-00066

Tandon, V.R., Khajuria, V., Mahajan, V., Sharma, A., Gillani, Z., Mahajan, A., 2015. Drug-induced diseases (DIDs): An experience of a tertiary care teaching hospital from India. Indian Journal of Medical Research 142, 33–39. https://doi.org/10.4103/0971-5916.162093

Tang, P.C., McDonald, C.J., 2006. Electronic health record systems, in: Shortliffe, E.H., Cimino, J.J. (Eds.), Biomedical Informatics: Computer Applications in Health

Care and Biomedicine, Health Informatics. Springer New York, New York, NY, pp. 447–475. https://doi.org/10.1007/0-387-36278-9_12

Thomas, J.J., Cook, K.A., 2006. A visual analytics agenda. IEEE Computer Graphics and Applications 26, 10–13. https://doi.org/10.1109/MCG.2006.5

Thomas, J.J., Cook, K.A., 2005. Illuminating the Path: The Research and Development Agenda for Visual Analytics.

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

Timmerman, M.E., Ceulemans, E., Kiers, H.A., Vichi, M., 2010. Factorial and reduced K-means reconsidered. Computational Statistics & Data Analysis 54, 1858–1871.

Tomar, D., Agarwal, S., 2013. A survey on data mining approaches for healthcare. International Journal of Bio-Science and Bio-Technology 5, 241–266. https://doi.org/10.14257/ijbsbt.2013.5.5.25

Tominski, C., 2015. Interaction for Visualization. Synthesis Lectures on Visualization 3, 1–107. https://doi.org/10.2200/S00651ED1V01Y201506VIS003

Torgerson, W.S., 1958. Theory and methods of scaling, Theory and methods of scaling. Wiley, Oxford, England.

Tran, N.K., Sen, S., Palmieri, T.L., Lima, K., Falwell, S., Wajda, J., Rashidi, H.H., 2019. Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: A proof of concept. Burns 45, 1350–1358. https://doi.org/10.1016/j.burns.2019.03.021

Turner, A.M., Stavri, Z., Revere, D., Altamore, R., 2008. From the ground up: information needs of nurses in a rural public health department in Oregon. J Med Libr Assoc 96, 335–342. https://doi.org/10.3163/1536-5050.96.4.008

Uchino, S., Kellum, J.A., Bellomo, R., Doig, G.S., Morimatsu, H., Morgera, S., Schetz, M., Tan, I., Bouman, C., Macedo, E., Gibney, N., Tolwani, A., Ronco, C., Beginning and Ending Supportive Therapy for the Kidney (BEST Kidney) Investigators, 2005. Acute renal failure in critically ill patients: a multinational, multicenter study. JAMA 294, 813–8. https://doi.org/10.1001/jama.294.7.813

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. Nature Reviews Drug Discovery. https://doi.org/10.1038/s41573-019-0024-5

Van der Corput, P., Arends, J., Van Wijk, J.J., 2014. Visualization of Medicine Prescription Behavior. Computer Graphics Forum 33, 161–170. https://doi.org/10.1111/cgf.12372

Vanderperren, B., Rizzo, M., Angenot, L., Haufroid, V., Jadoul, M., Hantson, P., 2005. Acute Liver Failure with Renal Impairment Related to the Abuse of Senna Anthraquinone Glycosides. Ann Pharmacother 39, 1353–1357. https://doi.org/10.1345/aph.1E670

Varga, M., Varga, C., 2016. Visual Analytics: Data, Analytical and Reasoning Provenance. pp. 141–150. https://doi.org/10.1007/978-3-319-40226-0_9

Verdoodt, A., Honore, P.M., Jacobs, R., Waele, E. De, Gorp, V. Van, Regt, J. De, Spapen, H.D., 2018. Do statins induce or protect from acute kidney injury and chronic kidney disease: An update review in 2018. Journal of Translational Internal Medicine 6, 21–25. https://doi.org/10.2478/jtim-2018-0005

Vichi, M., Kiers, H.A., 2001. Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis 37, 49–64.

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 : JASN 17, 1143–50. https://doi.org/10.1681/ASN.2005091017

Wang, C., Deng, C., Wang, S., 2019. Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Binary Label-Imbalanced Classification with XGBoost. arXiv:1908.01672 [cs, stat].

Wang, C., Wang, S., Shi, F., Wang, Z., 2018. Robust Propensity Score Computation Method based on Machine Learning with Label-corrupted Data. arXiv:1801.03132 [cs, stat].

Wang, C.-W., Lee, Y.-C., Calista, E., Zhou, F., Zhu, H., Suzuki, R., Komura, D., Ishikawa, S., Cheng, S.-P., 2018. A benchmark for comparing precision medicine methods in thyroid cancer diagnosis using tissue microarrays. Bioinformatics 34, 1767–1773. https://doi.org/10.1093/bioinformatics/btx838

Wang, S., Yao, X., 2009. Diversity analysis on imbalanced data sets by using ensemble models, in: 2009 IEEE Symposium on Computational Intelligence and Data Mining. Presented at the 2009 IEEE Symposium on Computational Intelligence and Data Mining, pp. 324–331. https://doi.org/10.1109/CIDM.2009.4938667

Wang, T.D., Plaisant, C., Quinn, A.J., Stanchak, R., Murphy, S., Shneiderman, B., 2008. Aligning temporal data by sentinel events: Discovering patterns in electronic health records, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’08. ACM, New York, NY, USA, pp. 457–466. https://doi.org/10.1145/1357054.1357129

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, 1049–1056. https://doi.org/10.1109/TVCG.2009.187

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

Watanabe, R., Takahashi, K., Tada, K., Ishimura, A., 2014. A case of acute kidney injury associated with the use of oseltamivir and clarithromycin that was treated by hemodialysis. Nihon Toseki Igakkai Zasshi 47, 755–759. https://doi.org/10.4009/jsdt.47.755

Wenskovitch, J., Crandell, I., Ramakrishnan, N., House, L., Leman, S., North, C., 2018. Towards a Systematic Combination of Dimension Reduction and Clustering in Visual Analytics. IEEE Trans. Visual. Comput. Graphics 24, 131–141. https://doi.org/10.1109/TVCG.2017.2745258

Wetzel, R.C., 2001. The virtual pediatric intensive care unit: Practice in the new millennium. Pediatric Clinics 48, 795–814.

Wilke, C.O., 2019. Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures, 1 edition. ed. O’Reilly Media, Sebastopol, CA.

Wilkinson, L., 2015. Classification and Regression Trees 23.

Williams, D.A., McCullagh, P., Nelder, J.A., 1984. Generalized Linear Models. Biometrics 40, 566. https://doi.org/10.2307/2531415

Wilson, J.R., 2014. Fundamentals of systems ergonomics/human factors. Applied Ergonomics 45, 5–13. https://doi.org/10.1016/j.apergo.2013.03.021

Wise, J.A., 1999. The ecological approach to text visualization. Journal of the American Society for Information Science 50, 1224–1233.

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

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, 2659–2668. https://doi.org/10.1109/TVCG.2012.225

Wongsuphasawat, K., Gotz, D., 2011. Outflow: Visualizing patient flow by symptoms and outcome, in: IEEE VisWeek Workshop on Visual Analytics in Healthcare, Providence, Rhode Island, USA. American Medical Informatics Association, pp. 25–28.

Wongsuphasawat, K., Guerra Gómez, J.A., Plaisant, C., Wang, T.D., Taieb-Maimon, M., Shneiderman, B., 2011. LifeFlow: visualizing an overview of event sequences, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’11. ACM, New York, NY, USA, pp. 1747–1756. https://doi.org/10.1145/1978942.1979196

Wu, X., Zhang, W., Ren, H., Chen, X., Xie, J., Chen, N., 2014. Diuretics associated acute kidney injury: clinical and pathological analysis. Renal failure 36, 1051–5. https://doi.org/10.3109/0886022X.2014.917560

Xin Geng, De-Chuan Zhan, Zhi-Hua Zhou, 2005. Supervised nonlinear dimensionality reduction for visualization and classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 35, 1098–1107. https://doi.org/10.1109/TSMCB.2005.850151

Xue, J.L., Daniels, F., Star, R.A., Kimmel, P.L., Eggers, P.W., Molitoris, B.A., Himmelfarb, J., Collins, A.J., 2006. Incidence and mortality of acute renal failure in Medicare beneficiaries, 1992 to 2001. Journal of the American Society of Nephrology : JASN 17, 1135–42. https://doi.org/10.1681/ASN.2005060668

Yamada, H., Katsumori, Y., Kawano, M., Mori, S., Takeshige, R., Mukai, J., Imada, H., Shimoura, H., Takahashi, H., Horai, T., 2018. Quetiapine-related Acute Kidney Injury Requiring Transient Continuous Hemodiafiltration. Internal Medicine 57, 1763–1767.

Yan, J., Zhang, B., Liu, N., Yan, S., Cheng, Q., Fan, W., Yang, Q., Xi, W., Chen, Z., 2006. Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. IEEE transactions on Knowledge and Data Engineering 18, 320–333.

Yokota, L.G., Sampaio, B.M., Rocha, E.P., Balbi, A.L., Sousa Prado, I.R., Ponce, D., 2018. Acute kidney injury in elderly patients: narrative review on incidence, risk factors, and mortality. Int J Nephrol Renovasc Dis 11, 217–224. https://doi.org/10.2147/IJNRD.S170203

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.F., Hua, L., 2012. Data mining in healthcare and biomedicine: A survey of the literature. Journal of Medical Systems 36, 2431–2448. https://doi.org/10.1007/s10916-011-9710-5

Zaleska-Kociecka, M., Dabrowski, M., Stepinska, J., 2019. Acute kidney injury after transcatheter aortic valve replacement in the elderly: outcomes and risk management. Clin Interv Aging 14, 195–201. https://doi.org/10.2147/CIA.S149916

Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., Keim, D., 2012. Visual analytics for the big data era — A comparative review of state-of-the-art commercial systems, in: 2012 IEEE Conference on Visual Analytics Science and Technology (VAST). Presented at the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173–182. https://doi.org/10.1109/VAST.2012.6400554

Zhang, X., Donnan, P.T., Bell, S., Guthrie, B., 2017. Non-steroidal anti-inflammatory drug induced acute kidney injury in the community dwelling general population and people with chronic kidney disease: systematic review and meta-analysis. BMC Nephrology 18, 256. https://doi.org/10.1186/s12882-017-0673-8

Zhou, J., Konecni, S., Grinstein, G., 2009. Visually comparing multiple partitions of data with applications to clustering, in: Visualization and Data Analysis 2009. International Society for Optics and Photonics, p. 72430J.

Zitnik, M., Agrawal, M., Leskovec, J., 2018. Modeling polypharmacy side effects with graph convolutional networks, in: Bioinformatics. Oxford University Press, pp. i457–i466. https://doi.org/10.1093/bioinformatics/bty294

Zulman, D.M., Asch, S.M., Martins, S.B., Kerr, E.A., Hoffman, B.B., Goldstein, M.K., 2014. Quality of care for patients with multiple chronic conditions: The role of comorbidity interrelatedness. Journal of General Internal Medicine 29, 529–537. https://doi.org/10.1007/s11606-013-2616-9

Appendices

Appendix A: List of databases held at ICES.

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

Appendix B: Coding definitions for co-morbid 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	International Classification of Diseases 9th Revision	401, 402, 403, 404, 405

	Discharge Abstract Database	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 C: Diagnostic codes for exclusion criteria.

Variable	Database	Code Set	Code
Dialysis	Canadian Institute for Health Information	Canadian Classification of Diagnostic,	5127, 5142, 5143, 5195, 6698

	Discharge Abstract Database	Therapeutic and Surgical Procedures	
		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:	Sheikh S. Abdullah
Post-secondary Education and Degrees:	The University of Western Ontario London, Ontario, Canada Ph.D. (2014-2020)
	American International University-Bangladesh Dhaka, Bangladesh M.Sc. (2011-2013)
	American International University-Bangladesh Dhaka, Bangladesh B.Sc. (2007-2010)
Honours and Awards:	Summa Cum Laude B.Sc. (2011), M.Sc. (2013)
	Best Project Award B.Sc. (2011)
	Best Thesis Award M.Sc. (2013)
	Best Robotics Project National Electronic Project Competition (2009)
	Western Graduate Research Scholarship 2014-2018
Related Work Experience	Teaching Assistant, Department of Computer Science The University of Western Ontario 2014-2020
	Faculty, Department of Computer Science American International University-Bangladesh 2010-2014

Publications:

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*.

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.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.

TabinHasan, K.; Shaugat Abdullah, S.; Ahmed, R.; Giunchiglia, F. The History of Temporal Data Visualization and a Proposed Event Centric Timeline Visualization Model. *Int. J. Comput. Appl.* 2013, *70*, 27–33.

Giunchiglia, F.; Tabin Hasan, K.; Ahmed, R.; Shaugat Abdullah, S. Context Enabled Query and Minimalist Metadata Visualization: A Context Bound Approach for User and Content. *Int. J. Comput. Appl.* 2013, *69*, 34–40.

Giunchiglia, F.; Khandaker, T.H.; Sheikh Shaugat, A.; Rezwan, A. Minimalist Metadata Visualization: The Minimal Set of Context Dependent Attributes for Entity Identification. 2013.

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.*

Conferences:

Abdullah, S.S.; Rahaman, M.S.; Rahman, M.S. Analysis of stock market using text mining and natural language processing. In Proceedings of the 2013 International Conference on Informatics, Electronics and Vision, ICIEV 2013.

Abdullah, S.S.; Rahaman, M.S. Stock market prediction model using TPWS and association rules mining. In Proceedings of the Proceeding of the 15th International Conference on Computer and Information Technology, ICCIT 2012; pp. 390–395.

Book Chapters:

Sheikh Shaugat Abdullah and Saiful Azad (2014), Classical Cryptographic Algorithms. In S Azad & ASK Pathan, Practical Cryptography: Algorithms and Implementations Using C++ (pp. 11-34). Boca Raton, Florida. CRC Press.

Sheikh Shaugat Abdullah and Saiful Azad (2014), Rotor Machine. In S Azad & ASK Pathan, Practical Cryptography: Algorithms and Implementations Using C++ (pp. 35-44). Boca Raton, Florida. CRC Press.