information

MDPI

Article

VERONICA: Visual Analytics for Identifying Feature Groups in Disease Classification

Neda Rostamzadeh [1], Sheikh S. Abdullah [1], Kamran Sedig [1,*], Amit X. Garg [2] and Eric McArthur [3]

[1] Insight Lab, Western University, London, ON N6A 3K7, Canada; nrostamz@uwo.ca (N.R.); sabdul9@uwo.ca (S.S.A.)
[2] Department of Medicine, Epidemiology, and Biostatistics, Western University, London, ON N6A 3K7, Canada; amit.garg@lhsc.on.ca
[3] ICES, London, ON N6A 3K7, Canada; eric.mcarthur@ices.on.ca
* Correspondence: sedig@uwo.ca; Tel.: +1-519-661-2111

Abstract: The use of data analysis techniques in electronic health records (EHRs) offers great promise in improving predictive risk modeling. Although useful, these analysis techniques often suffer from a lack of interpretability and transparency, especially when the data is high-dimensional. The emergence of a type of computational system known as visual analytics has the potential to address these issues by integrating data analysis techniques with interactive visualizations. This paper introduces a visual analytics system called VERONICA that utilizes the natural classification of features in EHRs to identify the group of features with the strongest predictive power. VERONICA incorporates a representative set of supervised machine learning techniques—namely, classification and regression tree, C5.0, random forest, support vector machines, and naive Bayes to support users in developing predictive models using EHRs. It then makes the analytics results accessible through an interactive visual interface. By integrating different sampling strategies, analytics algorithms, visualization techniques, and human-data interaction, VERONICA assists users in comparing prediction models in a systematic way. To demonstrate the usefulness and utility of our proposed system, we use the clinical dataset stored at ICES to identify the best representative feature groups in detecting patients who are at high risk of developing acute kidney injury.

Keywords: visual analytics; electronic health records; machine learning; data-driven healthcare; imbalanced data; prediction models; acute kidney injury; precision medicine

check for updates

Citation: Rostamzadeh, N.; Abdullah, S.S.; Sedig, K.; Garg, A.X.; McArthur, E. VERONICA: Visual Analytics for Identifying Feature Groups in Disease Classification. *Information* **2021**, *12*, 344. https://doi.org/10.3390/info12090344

Academic Editor: Barbara Pes

Received: 1 July 2021
Accepted: 23 August 2021
Published: 26 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

A key component of precision medicine is to determine a person's individualized estimates of different health outcomes, which then guides therapy to increase the chance of long-term good health. Identifying the group of features in electronic health records (EHRs) with the most substantial predictive power helps in the development of robust predictive models [1,2]. The data in EHRs has great promise for improving predictive risk modeling [3]. However, EHRs are often challenging to analyze due to their high dimensionality [4,5]. In recent years, several studies have incorporated various data mining and machine learning techniques to address this problem. Most of the existing studies use unsupervised learning techniques such as principal component analysis [6], K-means [7,8], and hierarchical clustering [9] to find the best representative group of features in high dimensional EHRs [10–18]. Although these unsupervised techniques have shown promise in managing high dimensional data, to our best knowledge, this problem has not been studied thoroughly using supervised techniques [19,20]. One of the main issues with both supervised and unsupervised techniques is that they suffer from a lack of interpretability and transparency [21–23]. In healthcare settings, it is essential to better understand how a given technique works. Therefore, increasing a technique's interpretability by involving humans in the analytics process can play a vital role in building trust with users [24–28].

The analytics results can be made accessible to users through visual analytics (VA) to address these issues.

Visual analytics (VA) integrates data analytics techniques with interactive visualizations to improve users' capabilities in performing data-driven tasks [21,29]. It enables users to achieve their goals through interactive exploration and manipulation of the data [30,31]. The design of a VA system is often not straightforward because it requires the designer to consider users' activities and tasks, the structure of the data, and human factors [20,32–34]. Thus, the designer needs to make several non-trivial decisions when developing such systems. For instance, one needs to consider which techniques to use, which features and samples to incorporate, and what level of granularity to look for when choosing a data analytics technique [29]. Similarly, it is important to determine how to map and classify data items and help users accomplish their tasks when developing interactive visualizations. Consequently, combining analytic techniques with interactive visualizations becomes a more complex challenge. Thus, it is important to involve stakeholders (e.g., clinical researchers and medical practitioners) in the design and development process of a VA system [35].

The purpose of this paper is to show how VA systems can be designed systematically to identify the best representative subset (i.e., a combination of groups) of high-dimensional EHRs. The proposed VA system, VERONICA (Visual analytics for idEntifying featuRe grOups iN dIsease ClAssification), takes advantage of the group structure of features stored in EHRs. EHRs are generally classified into different groups: comorbidities, medications, laboratory tests, hospital encounter codes, and demographics. It is possible to combine these groups to create multiple subsets of groups. For instance, one can create a subset by combining all features from both comorbidity and demographic groups. Depending on the predictive power of features within them, some groups or subsets (i.e., combinations of groups) are stronger predictors in identifying diseases in comparison to others. To identify the subset with the most substantial predictive power, VERONICA considers every possible subset of groups (i.e., groups of features) and applies several supervised learning techniques to each subset. It allows users to compare the results based on different performance measures through an interactive visual interface. VERONICA aims to assist healthcare providers at ICES-KDT, where ICES is a non-profit, world-leading research organization that utilizes population-based health data to produce knowledge on a broad range of healthcare issues, and KDT refers to the Kidney Dialysis and Transplantation Program located in London, Ontario, Canada. We utilize the clinical dataset housed at ICES to identify the best representative feature groups in detecting patients with a high risk of developing acute kidney injury to demonstrate VERONICA's utility and usefulness.

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

2. Background

In this section, we present the terminological and conceptual background to understand the design of VERONICA. We discuss different components of VA systems to provide a better understanding of the concept of VA. Finally, we provide a summary of the chosen machine learning techniques—namely, classification and regression tree (CART) [36], C5.0 [37], random forest [38], naïve Bayes (NB) [39], and support vector machine (SVM) [40].

2.1. Visual Analytics

Visual Analytics (VA) systems combine the strengths of data analysis and interactive visualizations to enable users to apply filters and explore and manipulate the data inter-

actively to accomplish their goals [41]. The processing load in VA is distributed between users and the key components of the system—namely, the analytics module and interactive visualization module [21,29,42–45].

2.1.1. Analytics Module

The analytics module is responsible for storing, pre-processing, transforming, and performing computerized analysis of the data. It involves three main stages: data pre-processing, data transformation, and data analysis [29]. The raw data retrieved from different sources gets processed in the pre-processing stage. This stage involves tasks such as fusion, integration, cleaning, and synthesis [46]. Then in the transformation stage, the pre-processed data is transformed into a form suitable for analysis. Common tasks in this stage include smoothing, aggregation, normalization, and feature generation [46]. Finally, the analysis stage involves the discovery of hidden patterns and relationships and allows for the extraction of useful and novel information from the data [47,48]. This can be carried out by applying various statistical and machine learning techniques (e.g., random forest, SVM, NB, and decision trees) to the transformed data. However, despite all the benefits, most of these computational techniques do not support proper exploration and manipulation of the computed results [21]. VA systems address this problem by allowing users to engage in a more involved discourse with the data through interactive visualizations.

2.1.2. Interactive Visualization Module

In VA systems, the interactive visualization module is composed of a mapping component that retrieves the analyzed data from the analytics module and generates interactive visual representations. It allows users to change the displayed information, modify the subset of the information displayed, and guide and control the intermediary steps of the analytical processes within the analytics module. This, in turn, incites a chain of reactions that leads to the execution of additional analysis processes. The interactive visualization module provides users with flexibility and supports their cognitive and perceptual needs as they engage in various complex tasks. However, despite the advantages of interactive visualizations in amplifying users' cognitive needs, they fell short when confronted with data-intensive tasks that require computational analysis [21]. Therefore, an approach that integrates analytical processes with interactive visualizations through VA is required to overcome these challenges [49–51].

2.2. Machine Learning Techniques

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

2.2.1. Decision Tree

Decision trees are among the most popular and powerful classification techniques that can provide informative models [52]. They construct a set of predictive rules to solve the classification problems using the recursive partitioning process. In their simplest form (e.g., C4.5 [37]), each feature is tested and ranked based on its ability to split the remaining data. The training data is propagated through the decision tree branches until enough features are chosen to correctly classify them. The classifier has a tree-like structure where each of its leaf nodes corresponds to a subset of the data that belongs to one class. Two widely known methods for generating decision trees are Classification and Regression Trees (CART) and C5.0. CART is based on a tree-growing algorithm that uses the GINI index as its splitting criteria. The strategy is to choose the feature whose GINI Index is minimum after each split. On the other hand, C5.0 builds the tree by splitting based on the feature that yields the most considerable information gain (Entropy). These classifiers are robust in handling missing values since the tree-growing process is not affected by missing data [53]. However, despite all the benefits, they tend to over-fit the training

data [54]. Random forest addresses this problem by generating an ensemble of decision trees where each tree is built from a random arrangement of features [38,55]. A new object passes through every tree in the forest to get classified based on a vector of features. Each distinctive tree gives a classification and votes for the class. The final classification of the new object is based on the majority "vote" of all the trees in the forest.

2.2.2. Support Vector Machines

Support Vector Machines are among the most successful and robust classification techniques [40,56]. They aim to identify an optimal separating hyperplane that can distinctly divide the instances of multiple classes in a multi-dimensional space by maximizing the minimum distance from the hyperplane to the closest instance. Although models produced by SVM are often hard to interpret and understand, they work well on classification tasks involving a large number of features [57]. SVM is first outlined for the linearly separable classification problems, but a linear classifier might not be the most appropriate candidate for the binary classification. SVM can support non-linear decision surfaces using kernel functions. Due to its good generalization ability and its low sensitivity to the curse of high-dimensionality, SVM is often used in many classification problems.

2.2.3. Naive Bayes

Naive Bayes is a simple and powerful probabilistic classifier that often creates stable and accurate models [39]. The model is based on the probability of each class and the conditional probability of each class given each feature. These probabilities that are directly calculated from the data can be used for the classification of new data based on the Bayes theorem. Naive Bayes makes a simplistic assumption that all the features are independent of one another. Despite this assumption and its simplistic design, it can be very efficient, particularly when the data is high-dimensional.

2.3. Class Imbalance Problem

In EHRs, data are usually composed of "negative" samples with only a small percentage of "positive" ones, resulting in the imbalanced classification problem. The imbalance problem in the healthcare domain, where one class often has notably fewer samples than the other class, affects the performance of classification techniques. The former class is known as the minority class, and the latter is known as the majority class. Most standard classification techniques, such as support vector machines, assume that both classes are equally common and aim to maximize the overall classification accuracy without accounting for uneven distribution of the minority and majority classes. Thus, the impact of the imbalance problem in the performance of classification techniques could have adverse consequences. It often results in a learning bias to the majority class and poor sensitivity toward the minority class [58,59]. In EHRs with imbalance class distribution, accurately detecting samples from the minority class is of great importance as they often correspond to high-impact events. For instance, among patients with suspicious mole(s) pigmentation, the prevalence of patients with cancer (i.e., minority class) is significantly lower than patients who are likely not to have cancer (i.e., the majority class). In this example, the incorrect classification of a cancer patient as a patient without cancer will incur an unacceptably high cost, thus making the class imbalance into a problem of great importance in predictive learning, especially in the healthcare domain. A common strategy to address the imbalance problem is to rebalance the class distribution at the data level using sampling techniques [60–63]. In the next section, we discuss some of the widely used sampling techniques in more detail.

Sampling Techniques

In their simplest forms, random oversampling duplicates random samples from the minority class, while random undersampling selects random samples from the majority class [64]. One of the main issues of undersampling is the removal of valuable information

if a large number of samples are discarded from the majority class. A considerable deletion of samples from the majority class might also change the distribution of the majority class, resulting in a change in the distribution of the overall dataset. On the other hand, since oversampling increases the size of the training data, it will result in an increased training time. It has also been shown that oversampling approaches might cause overfitting since classification techniques tend to focus on replicated minority samples [65]. Overfitting occurs when a prediction model fits too closely to the training set and is then incapable of generalizing the new data. To avoid overfitting, oversampling approaches that create artificial minority samples are preferred [66]. The synthetic minority oversampling technique (SMOTE) is an oversampling approach that randomly selects samples from the minority class and generates artificial minority samples by random interpolation between the selected samples and their nearest neighbors [67]. The generation of new minority class samples will lead to a more balanced class distribution compared to the original minority to majority class ratio. One potential disadvantage of SMOTE is that it creates the same number of artificial samples for each original minority sample without taking the neighboring samples into consideration, which ultimately increases the occurrence of overlaps between classes [68]. Several modifications of SMOTE that improve its performance by adjusting minority sample selection procedures have been proposed in the literature. For instance, adaptive synthetic sampling adaptively alters the number of artificial samples from the minority class following the density of majority samples around the original samples from the minority class [69].

3. Related Work

The most common application of VA in EHRs is identifying and exploring patient cohorts [70]. Several EHR-based VA systems have been developed to facilitate the process of generating and comparing multiple patient cohorts and identifying risk factors associated with a specific disease. For instance, VisualDecisionLinc [71] is a VA system that supports clinicians in identifying groups of patients with similar characteristics to understand the effectiveness of different treatments for those patients by providing summaries of patient outcomes and treatment options in a dashboard. Similarly, PHENOTREE [72] is a hierarchical and interactive phenotyping EHR-based VA system that allows users to interactively explore patient groups and explore hierarchical phenotypes by integrating principal component analysis and a user interface. VALENCIA [19] is another EHR-based VA system that facilitates the exploration of high-dimensional data stored in EHRS by combining various dimensionality reduction and clustering techniques with interactive visualizations. It allows clinical researchers to identify which features are more important within each cluster of patients. RadVis [73] is a VA system that enables clinicians to better understand the characteristics of patient clusters. It allows the user to apply different clustering techniques and displays the result using a 3-dimensional radial coordinate visualization. Likewise, Guo et al. [74] developed another EHR-based VA system to assist clinical researchers in clustering similar patients, comparing values of medical features of patients, and finding similar time tamps among similar patients. To support the user in performing these tasks, the system integrates a dimensionality reduction technique and a density-based clustering method with multiple interactive linked views. SubVIS [75] is another VA system that assists clinical researchers in exploring and interpreting high-dimensional clinical data by integrating different subspace clustering techniques and an interactive visual interface. Similarly, Huang et al. [76] developed a VA system that supports the exploration of patient trajectories to help clinical researchers in identifying how a group of similar patients with a specific disease might develop other comorbidities over time. The system integrates frequency-based and hierarchical clustering techniques with a Sankey-like timeline to support clinical researchers in performing these tasks. Most of these existing EHR-based VA systems that have been developed to manage high dimensional data use unsupervised learning techniques such as dimensionality reduction, principal component analysis, and clustering techniques. Although these techniques have shown

great promise in addressing this issue, to our best knowledge, this problem has not been studied thoroughly using supervised techniques.

4. Materials and Methods

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

4.1. Design Process and Participants

The design and development of VA systems is an integrated process that requires various sets of skills and expertise. In light of this, we adopted a participatory design approach to obtain the needs and requirements of the healthcare providers and to understand the real-world EHR-driven tasks that they perform. Participatory design is a co-operative approach that places users at the center of the design process. It is an iterative group effort that requires all the stakeholders to work together to ensure the system meets their expectations [35]. An epidemiologist, a clinician-scientist, a statistician, data scientists, and computer scientists were involved in the conceptualization, design, and evaluation of this VA system. It is important to optimize the communication between all stakeholders involved in the process because they might experience a language gap due to their different backgrounds. For instance, it is critical to ensure that the medical terms are comprehensible to the team members with a technical background and the motivations of the analysis process and design decisions are well-addressed across the team. In light of this, we asked healthcare experts to provide us with their formative feedback on different design decisions and a list of tasks they perform on EHRs. Multiple participatory design approaches are used to obtain the healthcare providers' needs and identify opportunities that can significantly improve VERONICA's performance through more effective visualizations and analysis techniques.

4.2. Data Sources

We formed a derivation cohort using large linked administrative healthcare databases held at ICES. We ascertained hospital and patient characteristics, outcome, and drug use from five administrative databases (see Table A1). These datasets were linked using unique encoded identifiers that were derived from health card numbers of patients and were analyzed at ICES. The Ontario Drug Benefit Program database is used to identify prescription drug use. This database contains highly accurate patient records of all outpatient prescriptions administered to patients aged 65 years or older, with an error rate of less than 1% [77]. We acquired vital statistics from the Ontario Registered Persons Database, which includes demographic data on all Ontarians who have ever been issued a health card. We identified baseline comorbidity data, ED visits, and hospital admission codes from the National Ambulatory Care Reporting System and the Canadian Institute for Health Information Discharge Abstract Database (hospitalizations). We used ICD-10 (i.e., International Classification of Diseases, post-2002) codes to identify hospital encounter codes and baseline comorbidities. In addition, baseline comorbidity data and health claims for physician services were acquired from the Ontario Health Insurance Plan database. All the coding definitions for the comorbidities are provided in Tables A2 and A3.

4.3. Cohort Entry Criteria

We created a cohort of patients aged 65 years or older who were admitted to a hospital or visited an emergency department (ED) between 2014 and 2016. The discharge date from the hospital or ED served as the index date, also referred to as the cohort entry date. If a patient had multiple ED visits and hospital admissions, we chose the first incident. Individuals with invalid data regarding age, sex, and the health-card number were excluded. In addition, we excluded individuals who: (1) previously received a kidney transplant or dialysis treatment as the assessment of acute kidney injury is usually no longer relevant in patients with end-stage kidney disease; (2) left the hospital or ED against

medical advice or without being seen by a physician; and (3) had acute kidney injury recorded during their hospital admission or ED visit prior to hospital discharge, as acute kidney injury was already present prior to the follow-up period. The diagnosis codes for the exclusion criteria are shown in Table A4.

4.4. Response Variable

Acute Kidney Injury (AKI) is defined as a sudden deterioration of kidney function. It is associated with a lower chance of survival, prolonged hospital stays, subsequent morbidity after discharge, and incremental healthcare costs [78–81]. A system that detects early AKI or predicts its clinical manifestations with considerable lead-time allows healthcare experts to provide more effective treatments to prevent AKI. We build models to predict hospital admission with AKI within 90 days after being discharged from ED or hospital. The incidence of AKI is identified using the Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System based on ICD-10 diagnostic codes (i.e., "N17").

4.5. Input Features

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

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

Table 1. The features included in the GP visit Group.

Features
Minor assessment
General assessment
General re-assessment
Consultation
Repeat consultation
Intermediate assessment or well-baby care
Mini assessment
Complex house call assessment
House call assessment
Limited consultation
Special family and general practice consultation
Comprehensive family and general practice consultation
Care of the elderly FPA
Periodic health visit—adult 65 years of age and older
Chronic disease shared appointment-2 patients (per unit)
Chronic disease shared appointment—3 patients (per unit)
Chronic disease shared appointment—4 patients (per unit)
Chronic disease shared appointment—5 patients (per unit)
Chronic disease shared appointment—6 to 12 patients (per unit)
Nursing home or home for the aged—first 2 subsequent visits per patient per month (per visit)
Nursing home or home for the aged—additional subsequent visits (maximum 2 per patient per month) per visit
Additional visits due to intercurrent illness per visit

Table 2. The top ten features included in hospital encounter codes and medications groups.

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

5. Implementation Details

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

We use R to develop different components of the Analytics module because it (1) offers support in performing various sampling and machine learning techniques, (2) is an open-source and platform-independent tool, (3) has several libraries, (4) is available in the ICES environment, and (5) has a large community and user forums.

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

6. Workflow

As shown in Figure 1, VERONICA has two modules: Analytics and Interactive Visualization. The Analytics module utilizes the group structure of features stored in EHRs to identify the subset of feature groups that best represent the data in the prediction of AKI. The Interactive Visualization module maps the data items generated by the Analytics module into interactive visual representations to assist users in exploring the results. It supports six main interactions: (1) arranging, (2) drilling, (3) searching, (4) filtering, (5) transforming, and (6) selecting.

The basic workflow of VERONICA is as follows. First, we gather patient and hospital characteristics from five different databases stored at ICES. We then classify these features into five main groups—namely, hospital encounter codes, comorbidities, GP visits, medications, and demographics. The features included in these groups are pre-processed and transformed into forms appropriate for the analysis. We then create all possible subsets of groups (i.e., thirty-one groups), as shown in Figure 2. In the next step, we apply undersampling and SMOTE [67] to each subset to obtain two sampled datasets. Next, five machine learning techniques, namely CART, C5.0, random forest, naïve Bayes, and SVM, are applied to each sampled dataset, generating 310 prediction models. We use the area under the receiver operating characteristic curve (AUROC) to report the performance of these models. To help users compare and explore the analytic results, we make them accessible to users through interactive visualizations. The Interactive Visualization module uses an interactive visual interface to show the results of the Analytics module. It allows users to explore the prediction models and compare their performance. The interface is supported by several controls, such as a search bar, selection buttons, and drop-down menus. Finally, several interactions are built into the system to allow users to manipulate the results.

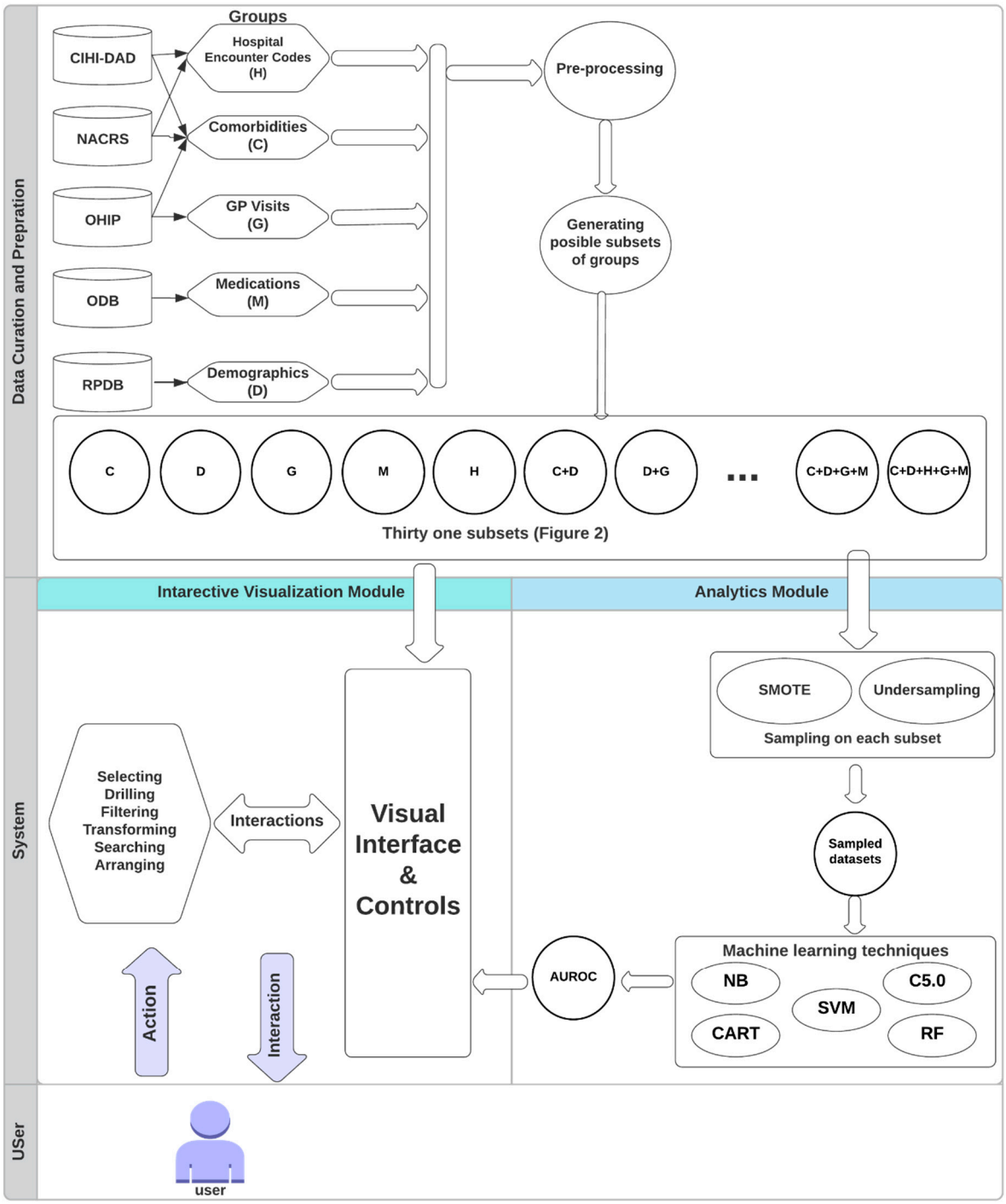

Figure 1. The workflow diagram of VERONICA.

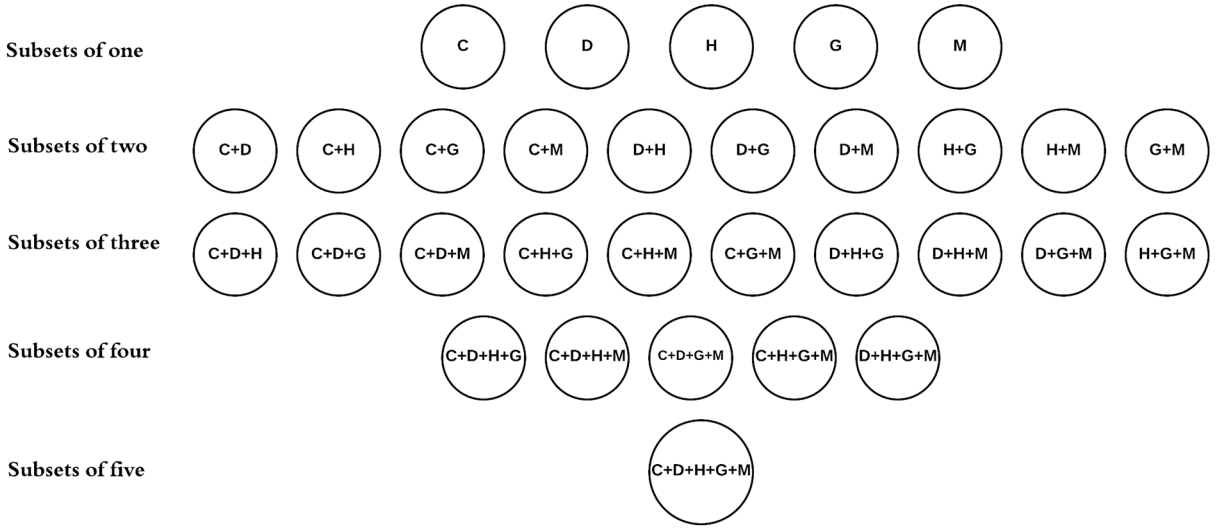

Figure 2. All possible subsets of 5 groups of features, including comorbidities (C), demographics (D), hospital encounter codes (H), GP visits (G), and medications (M).

7. The Design of VERONICA

We use VERONICA to identify the subset of groups that has the most substantial predictive power in the classification of AKI. VERONICA applies several machine learning techniques to each subset and allows exploration of the analysis results through interactive visualizations. In this section, we describe the two main components of the system. We explain how the data is processed and analyzed in the Analytics module. We then describe the Interactive Visualization module and how it assists users in the interpretation and exploration of the results.

7.1. Analytics Module

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

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

A total of 924,533 participants are included in the final cohort, of which 5993 experienced AKI after being discharged from the index encounter. This dataset has an imbalanced class distribution, where the negative class (i.e., non-AKI) is represented by a large number of patients (i.e., 899,449 patients) compared to the positive class (i.e., 5993). The proposed system supports a number of sampling techniques such as random oversampling, Borderline-SMOTE [82], and Adaptive Synthetic Sampling. In this paper, we

use undersampling and SMOTE. We configure these techniques so that the number of positive cases becomes equal compared to the negative cases. We use the DMwR package in R to implement the SMOTE algorithm. The "k" (i.e., nearest neighbors) and "perc.over" variables of the SMOTE algorithm are set to 5 and 100, respectively.

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

We compare the performance of all the generated models using AUROC [83,84]. A ROC curve shows the trade-off between sensitivity and specificity across different decision thresholds. Sensitivity measures how often a test classifies a patient as "at-risk" correctly. On the other hand, specificity is the capacity of a test to classify a patient as "risk-free" correctly [85]. The AUROC ranges from 0.51 to 0.89 for the classification of AKI among the generated models.

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

7.2. Interactive Visualization Module

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

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

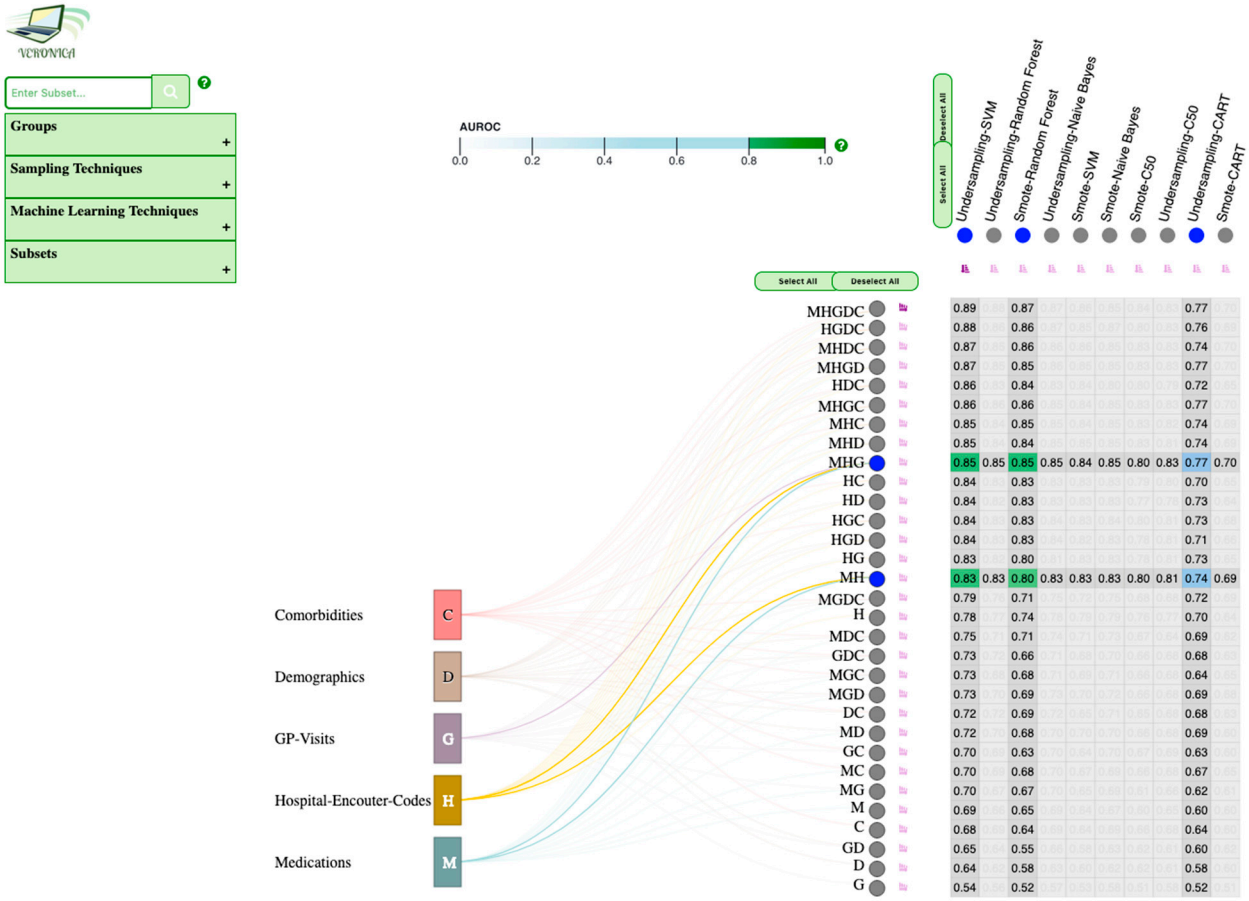

Figure 3. An overview of VERONICA.

Table 3. Groups and their representing codes.

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

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

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

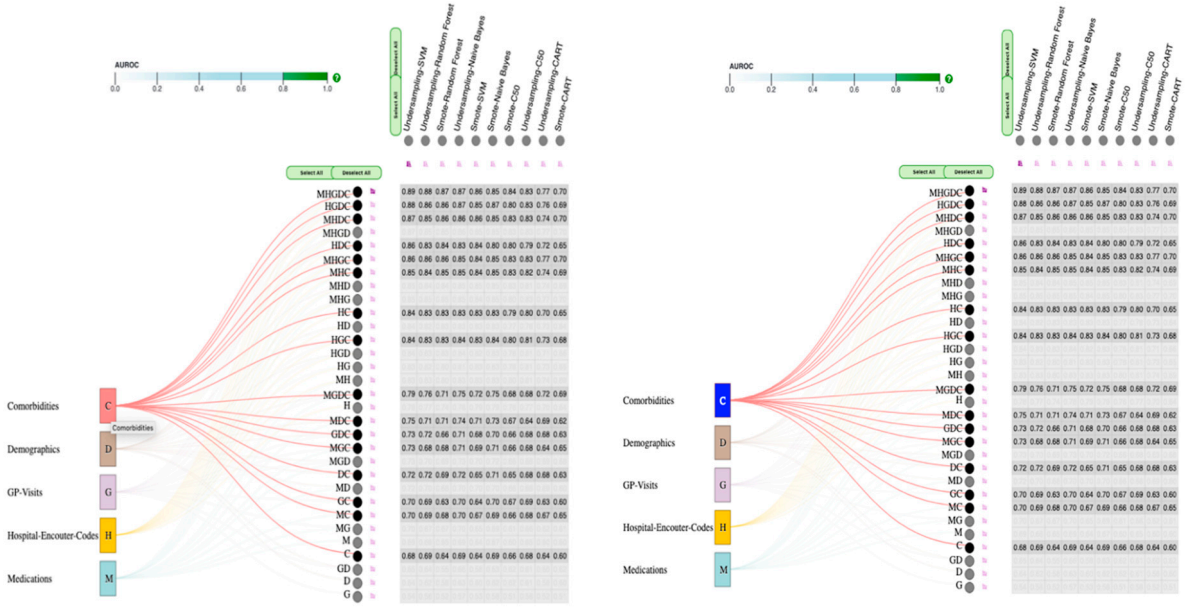

Figure 4. How the system gets updated when users hover over (**A**) or select (**B**) a rectangle representing groups.

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

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

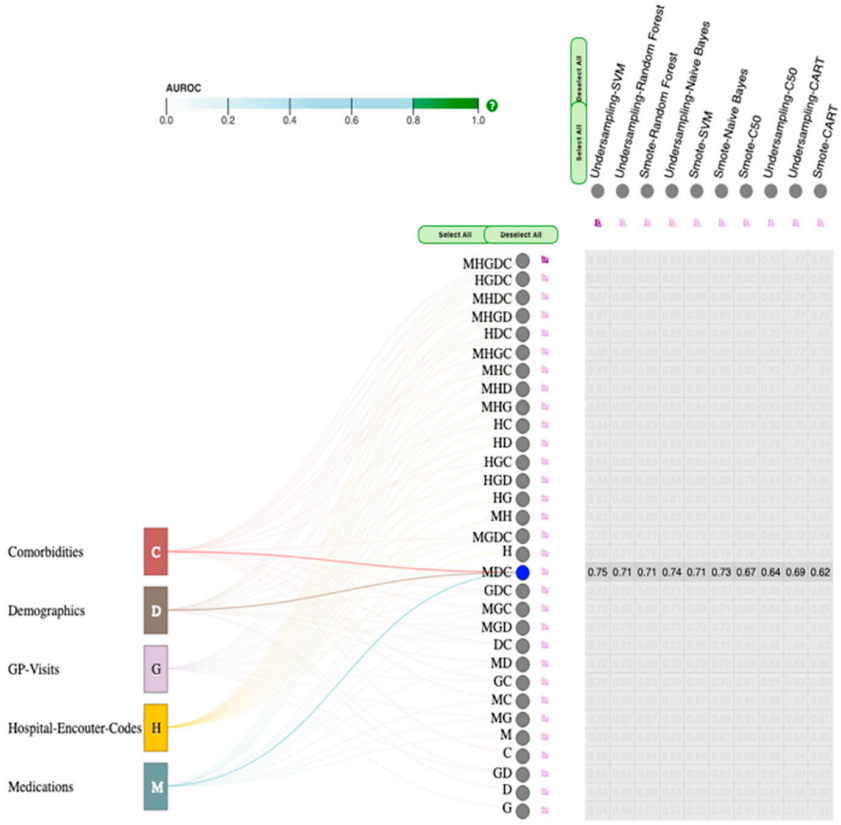

Figure 5. How the system gets updated when users select a circle representing subsets.

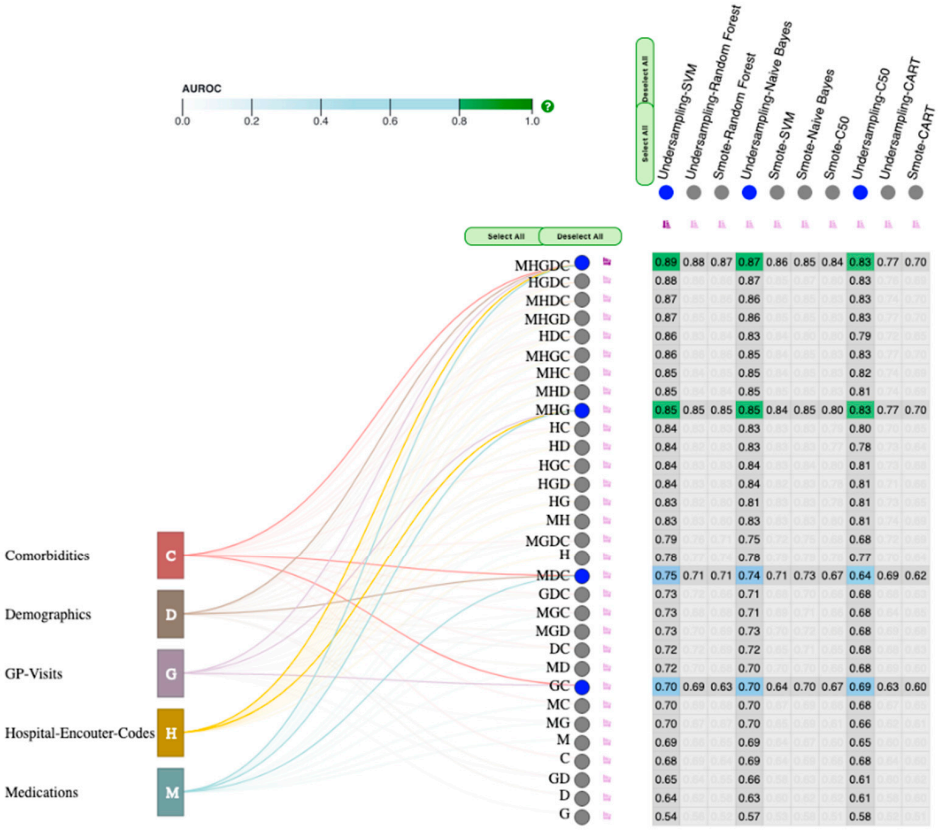

Figure 6. How the system gets updated when users select multiple subset nodes and techniques.

Users can also hover the mouse over the cells in the heatmap to highlight the labels and circles representing the hovered cell. Additionally, this interaction changes the cell's color based on its corresponding AUROC value. The system enables users to sort the cells by rows and columns based on their corresponding AUROC values by clicking on the pink sort icons. For instance, cells in the heatmap are sorted by the "MHGDC" subset and "undersampling-SVM" technique in Figure 6.

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

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

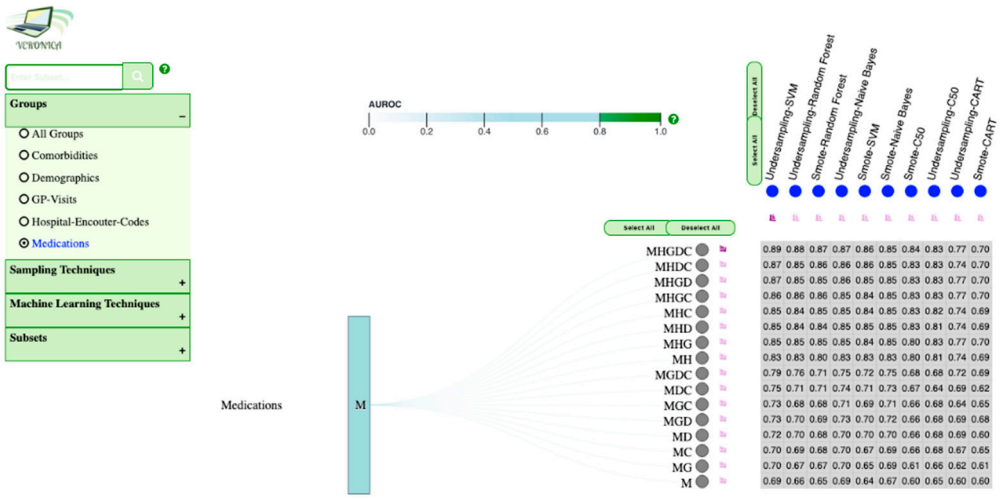

Figure 7. How the system gets updated when users select "Medications" from the "Groups" drop-down menu.

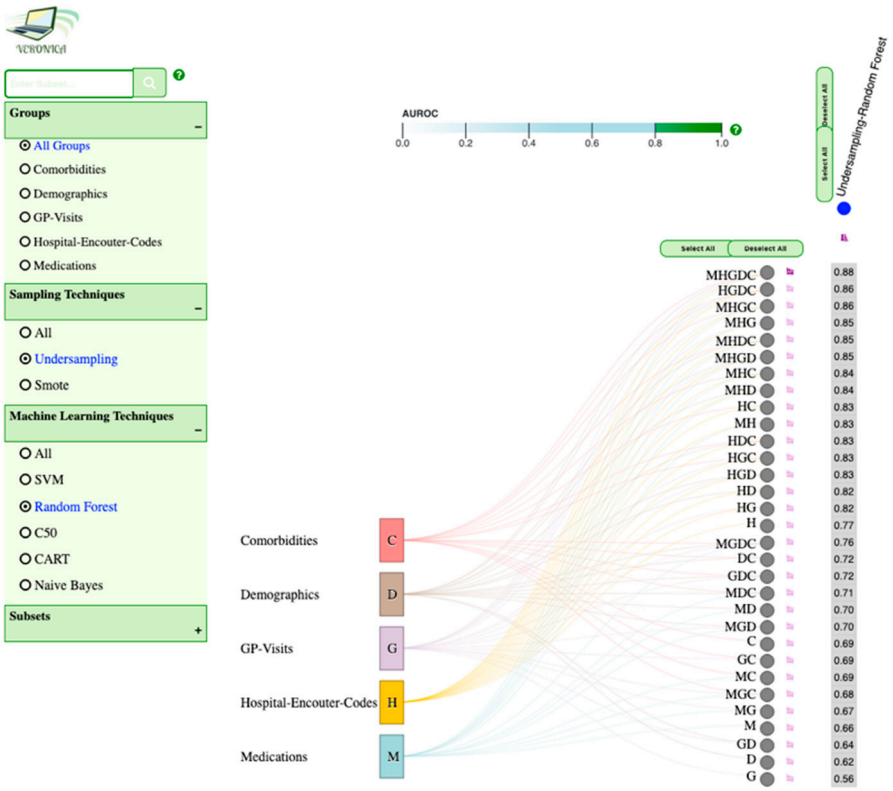

Figure 8. How the system gets updated when users select "Undersampling" and "Random Forest" from the "Sampling Techniques" and "Machine Learning Techniques" drop-down menus.

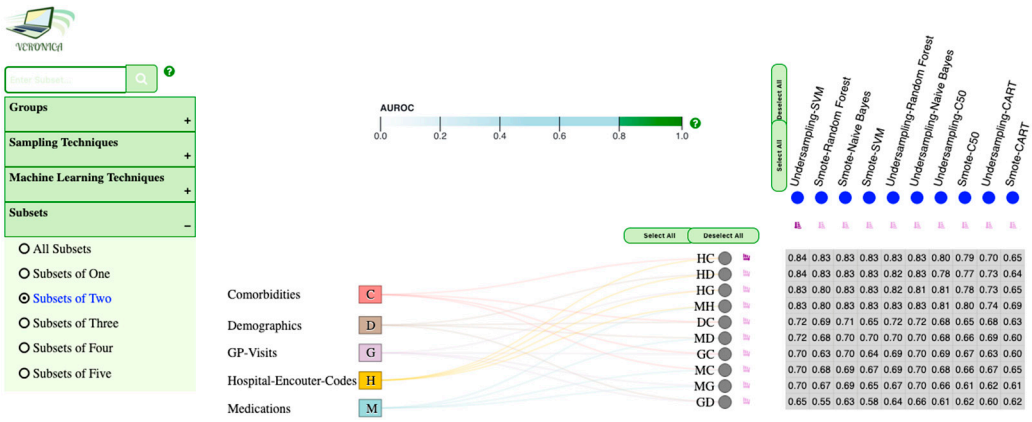

Figure 9. How the system gets updated when users select "Subsets of Two" from the "Subsets" drop-down menu.

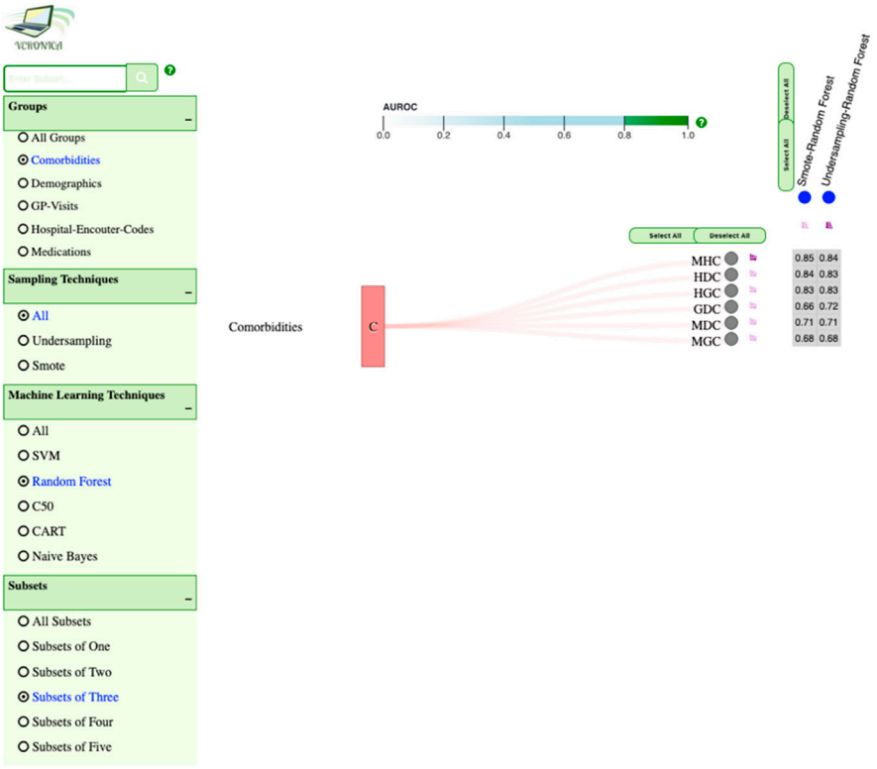

Figure 10. How the system gets updated when users select "Comorbidities", "Random Forest", and "Subsets of Three" from "Groups", "Machine Learning Techniques", and "Subsets" drop-down menus.

8. Limitations

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

9. Conclusions and Future Work

In this paper, we demonstrate how VA systems can be designed to address the challenges stemming from the high dimensional EHRs to identify the subset of feature groups with the most predictive power in the classification of AKI systematically. To accomplish this, we have reported the development of VERONICA, a VA system designed to assist healthcare providers at ICES' KDT program. VERONICA incorporates two components: Analytics and Interactive Visualization modules. The Analytics module identifies the best representative subset of data in detecting the patients at high risk of developing AKI using different sampling and machine learning techniques. It incorporates two sampling techniques—undersampling and SMOTE. It also uses a representative set of machine learn-

ing techniques, including CART, C5.0, random forest, SVM, and naive Bayes. Our clinical dataset includes comorbidities, demographics, hospital encounter codes, GP visits, and medications. The system generates a large number of prediction models by applying sampling and machine learning techniques mentioned above to each subset. The performance of all the generated models is reported using AUROC. The system enables users to access, explore, and compare these models through interactive visualizations. The Interactive Visualization module is composed of an interactive visual interface and several selection controls, such as a search bar, drop-down menus, and selection buttons. The interactive visual interface assists users in the exploration of the analytic results by providing them several interactions such as arranging, drilling, searching, filtering, transforming, and selecting.

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

Future research directions include (but are not limited to) the following. Further research is needed to effectively evaluate the performance of the system by comparing it with other standard feature selection techniques. In addition, we plan to measure the effectiveness of the system for different datasets that support natural groupings. Furthermore, since the proposed system is developed in an access-restricted virtual machine [20,86], we could not evaluate the systems' scalability. Thus, further efforts are needed to access VERONICA more comprehensively by conducting formal studies.

Author Contributions: Conceptualization, N.R., S.S.A., K.S., A.X.G. and E.M.; methodology, N.R., S.S.A. and K.S.; software, N.R., S.S.A.; validation, N.R., S.S.A., K.S., A.X.G. and E.M.; data curation, N.R., S.S.A. and E.M.; writing—original draft preparation, N.R. and S.S.A.; writing—review and editing, N.R., S.S.A., K.S., A.X.G. and E.M.; visualization, N.R., S.S.A. and K.S.; supervision, K.S. and A.X.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Projects that use data collected by ICES under section 45 of Ontario's Personal Health Information Protection Act (PHIPA), and use no other data, are exempt from research ethics board review. The use of the data in this project is authorised under section 45 and approved by ICES' Privacy and Legal Office.

Informed Consent Statement: ICES is a prescribed entity under PHIPA. Section 45 of PHIPA authorises ICES to collect personal health information, without consent, for the purpose of analysis or compiling statistical information with respect to the management of, evaluation or monitoring of, the allocation of resources to or planning for all or part of the health system.

Data Availability Statement: The study dataset is held securely in coded form at ICES. While legal data sharing agreements between ICES and data providers (eg, healthcare organisations and government) prohibit ICES from making the dataset publicly available, access might be granted to those who meet prespecified criteria for confidential access, available at www.ices.on.ca/DAS (email das@ices.on.ca). The full dataset creation plan and underlying analytic code are available from the authors upon request, understanding that the computer programs might rely upon coding templates or macros that are unique to ICES and are therefore either inaccessible or require modification.

Acknowledgments: This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). Parts of this material are based on data and information compiled and provided by: (CIHI). The analyses, conclusions, opinions and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred.

Conflicts of Interest: The authors declare that there is no conflict of interest. Amit Garg is supported by Adam Linton, Chair in Kidney Health Analytics and a Clinician Investigator Award from the Canadian Institutes of Health Research (CIHR).

Appendix A

Table A1. List of databases held at 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).

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 the National Ambulatory Care Reporting System collect diagnostic and procedural variables for inpatient stays and ED visits, respectively. Diagnostic and inpatient procedural coding uses 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

Table A2. Coding definitions for comorbid conditions.

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

Table A2. *Cont.*

Variable	Database	Code	Set Code
Coronary artery disease (excluding angina)	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	4801, 4802, 4803, 4804, 4805, 481, 482, 483
		Canadian Classification of Health Interventions	1IJ50, 1IJ76
		International Classification of Diseases 9th Revision	412, 410, 411
		International Classification of Diseases 10th Revision	I21, I22, Z955, T822
	Ontario Health Insurance Plan	Diagnosis	410, 412
		Fee code	R741, R742, R743, G298, E646, E651, E652, E654, E655, Z434, Z448
Diabetes	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	250
		International Classification of Diseases 10th Revision	E10, E11, E13, E14
	Ontario Health Insurance Plan	Diagnosis	250
		Fee code	Q040, K029, K030, K045, K046
Heart failure	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	4961, 4962, 4963, 4964
		Canadian Classification of Health Interventions	1HP53, 1HP55, 1HZ53GRFR, 1HZ53LAFR, 1HZ53SYFR
		International Classification of Diseases 9th Revision	I500, I501, I509, I255, J81
		International Classification of Diseases 10th Revision	I21, I22, Z955, T822
	Ontario Health Insurance Plan	Diagnosis	428
		Fee code	R701, R702, Z429
Hypertension	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	401, 402, 403, 404, 405
		International Classification of Diseases 10th Revision	I10, I11, I12, I13, I15
	Ontario Health Insurance Plan	Diagnosis	401, 402, 403

Table A2. *Cont.*

Variable	Database	Code	Set Code
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

Table A3. Diagnostic codes for health care utilization characteristics.

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

Table A4. Diagnostic codes for exclusion criteria.

Variable	Database	Code Set	Code
Dialysis	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	5127, 5142, 5143, 5195, 6698
		Canadian Classification of Health Interventions	1PZ21, 1OT53DATS, 1OT53HATS, 1OT53LATS, 1SY55LAFT, 7SC59QD, 1KY76, 1KG76MZXXA, 1KG76MZXXN, 1JM76NC, 1JM76NCXXN
		International Classification of Diseases 9th Revision	V451, V560, V568, 99673
		International Classification of Diseases 10th Revision	T824, Y602, Y612, Y622, Y841, Z49, Z992
	Ontario Health Insurance Plan	Fee code	R850, G324, G336, G327, G862, G865, G099, R825, R826, R827, R833, R840, R841, R843, R848, R851, R946, R943, R944, R945, R941, R942, Z450, Z451, Z452, G864, R852, R853, R854, R885, G333, H540, H740, R849, G323, G325, G326, G860, G863, G866, G330, G331, G332, G861, G082, G083, G085, G090, G091, G092, G093, G094, G095, G096, G294, G295
Kidney transplant	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Health Interventions	1PC85
	Ontario Health Insurance Plan	Fee code	S435, S434

References

1. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
2. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
3. Hersh, W.R. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *Am. J. Manag. Care* **2007**, *13*, 277–278.
4. Jensen, P.B.; Jensen, L.J.; Brunak, S. Mining electronic health records: Towards better research applications and clinical care. *Nat. Rev. Genet.* **2012**, *13*, 395–405. [CrossRef]
5. Weiskopf, N.G.; Weng, C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 144–151. [CrossRef]
6. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417–441. [CrossRef]
7. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C* **1979**, *28*, 100–108. [CrossRef]
8. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
9. Nielsen, F. Hierarchical Clustering. Introduction to HPC with MPI for Data Science. In *Undergraduate Topics in Computer Science*; Nielsen, F., Ed.; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 195–211. ISBN 978-3-319-21903-5.
10. Alexander, N.; Alexander, D.C.; Barkhof, F.; Denaxas, S. Using Unsupervised Learning to Identify Clinical Subtypes of Alzheimer's Disease in Electronic Health Records. *Stud. Health Technol. Inform.* **2020**, *270*, 499–503. [CrossRef] [PubMed]
11. Lütz, E. Unsupervised Machine Learning to Detect Patient Subgroups in Electronic Health Records. Available online: /paper/Unsupervised-machine-learning-to-detect-patient-in-L%C3%9CTZ/e11f5b060947f22ae7d80d053564546487dbc0bf (accessed on 11 November 2020).
12. Khalid, S.; Judge, A.; Pinedo-Villanueva, R. 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, Funchal, Madeira, Portugal, 19–21 January 2018; SCITEPRESS—Science and Technology Publications: Funchal, Portugal, 2018; pp. 266–273.
13. Miotto, R.; Li, L.; Kidd, B.A.; Dudley, J.T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.* **2016**, *6*, 26094. [CrossRef] [PubMed]
14. Lasko, T.A.; Denny, J.C.; Levy, M.A. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLoS ONE* **2013**, *8*, e66341. [CrossRef]

15. Marlin, B.M.; Kale, D.C.; Khemani, R.G.; Wetzel, R.C. 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*; ACM Press: Miami, FL, USA, 2012; p. 389.
16. Wang, L.; Tong, L.; Davis, D.; Arnold, T.; Esposito, T. The application of unsupervised deep learning in predictive models using electronic health records. *BMC Med. Res. Methodol.* **2020**, *20*, 37. [CrossRef]
17. Panahiazar, M.; Taslimitehrani, V.; Pereira, N.L.; Pathak, J. Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics. *Stud. Health Technol. Inform.* **2015**, *210*, 369–373.
18. Langavant, L.C.D.; Bayen, E.; Yaffe, K. Unsupervised Machine Learning to Identify High Likelihood of Dementia in Population-Based Surveys: Development and Validation Study. *J. Med. Internet Res.* **2018**, *20*, e10493. [CrossRef] [PubMed]
19. 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. [CrossRef]
20. Abdullah, S.S. Visual Analytics of Electronic Health Records with a Focus on Acute Kidney Injury. Ph.D. Thesis, The University of Western Ontario, London, ON, Canada, 2020.
21. Keim, D.A.; Mansmann, F.; Thomas, J. Visual analytics: How much visualization and how much analytics? *ACM SIGKDD Explor. Newsl.* **2010**, *11*, 5. [CrossRef]
22. Caruana, R.; Karampatziakis, N.; Yessenalina, A. An empirical evaluation of supervised learning in high dimensions. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; Association for Computing Machinery: Helsinki, Finland, 2008; pp. 96–103.
23. Johnstone, I.M.; Titterington, D.M. Statistical challenges of high-dimensional data. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **2009**, *367*, 4237–4253. [CrossRef] [PubMed]
24. Krause, J.; Perer, A.; Bertini, E. Using Visual Analytics to Interpret Predictive Machine Learning Models. *arXiv* **2016**, arXiv:160605685.
25. Liu, S.; Wang, X.; Liu, M.; Zhu, J. Towards better analysis of machine learning models: A visual analytics perspective. *Vis. Inform.* **2017**, *1*, 48–56. [CrossRef]
26. Krause, J.; Perer, A.; Ng, K. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; Association for Computing Machinery: San Jose, CA, USA, 2016; pp. 5686–5697.
27. Zhao, X.; Wu, Y.; Lee, D.L.; Cui, W. iForest: Interpreting Random Forests via Visual Analytics. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 407–416. [CrossRef]
28. Spinner, T.; Schlegel, U.; Schäfer, H.; El-Assady, M. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 1064–1074. [CrossRef]
29. Ola, O.; Sedig, K. The challenge of big data in public health: An opportunity for visual analytics. *Online J. Public Health Inform.* **2014**, *5*, 223. [CrossRef]
30. Parsons, P.; Sedig, K.; Mercer, R.; Khordad, M.; Knoll, J.; Rogan, P. Visual Analytics for Supporting Evidence-Based Interpretation of Molecular Cytogenomic Findings. In Proceedings of the 2015 Workshop on Visual Analytics in Healthcare, Chicago, IL, USA, 25 October 2015.
31. Simpao, A.F.; Ahumada, L.M.; Gálvez, J.A.; Rehman, M.A. A review of analytics and clinical informatics in health care. *J. Med. Syst.* **2014**, *38*, 45. [CrossRef] [PubMed]
32. Sedig, K.; Parsons, P.; Babanski, A. Towards a characterization of interactivity in visual analytics. *J. Multimed. Process. Technol.* **2012**, *3*, 12–28.
33. Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Garg, A.X.; McArthur, E. Multiple Regression Analysis and Frequent Itemset Mining of Electronic Medical Records: A Visual Analytics Approach Using VISA_M3R3. *Data* **2020**, *5*, 33. [CrossRef]
34. 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. [CrossRef]
35. Leighton, J.P. Defining and Describing Reason. In *The Nature of Reasoning*; Leighton, J.P., Sternberg, R.J., Eds.; Cambridge University Press: Cambridge, UK, 2004; pp. 3–11. ISBN 0-521-81090-6.
36. Wilkinson, L. Classification and regression trees. *Systat* **2004**, *11*, 35–56.
37. Quinlan, J.R. *C4. 5: Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.
38. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Lewis, D.D. Naive (Bayes) at forty: The independence assumption in information retrieval. In Proceedings of the European Conference on Machine Learning, Chemnitz, Germany, 21 April 1998; Springer: Berlin/Heidelberg, Germany, 1998; pp. 4–15.
40. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000; ISBN 978-0-521-78019-3.
41. Thomas, J.J.; Cook, K.A. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*; IEEE Computer Society: Washington, DC, USA, 2005.
42. Sedig, K.; Parsons, P. Interaction design for complex cognitive activities with visual representations: A pattern-based approach. *AIS Trans. Hum.-Comput. Interact.* **2013**, *5*, 84–133. [CrossRef]
43. Cui, W. Visual Analytics: A Comprehensive Overview. *IEEE Access* **2019**, *7*, 81555–81573. [CrossRef]

44. Jeong, D.H.; Ji, S.Y.; Suma, E.A.; Yu, B.; Chang, R. Designing a collaborative visual analytics system to support users' continuous analytical processes. *Hum.-Cent. Comput. Inf. Sci.* **2015**, *5*, 5. [CrossRef]
45. Parsons, P.; Sedig, K. Distribution of Information Processing While Performing Complex Cognitive Activities with Visualization Tools. In *Handbook of Human Centric Visualization*; Huang, W., Ed.; Springer: New York, NY, USA, 2014; pp. 693–715. ISBN 978-1-4614-7485-2.
46. Han, J.; Kamber, M.; Pei, J. Data mining concepts and techniques third edition. In *The Morgan Kaufmann Series in Data Management Systems*; Elsevier: Amsterdam, The Netherlands, 2011; pp. 83–124.
47. Agrawal, R.; Swami, A.; Imielinski, T. Database Mining: A Performance Perspective. *IEEE Trans. Knowl. Data Eng.* **1993**, *5*, 914–925. [CrossRef]
48. Sahu, H.; Shrma, S.; Gondhalakar, S. A Brief Overview on Data Mining Survey. *IJCTEE* **2008**, *1*, 114–121.
49. Keim, D.; Mansmann, F.; Schneidewind, J.; Thomas, J.; Ziegler, H. Visual analytics: Scope and challenges. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 4404, pp. 76–90.
50. Kehrer, J.; Hauser, H. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 495–513. [CrossRef] [PubMed]
51. 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. [CrossRef]
52. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: London, UK, 1984.
53. Ismail, B.; Anil, M. Regression methods for analyzing the risk factors for a life style disease among the young population of India. *Indian Heart J.* **2014**, *66*, 587–592. [CrossRef] [PubMed]
54. Deng, H.; Runger, G.; Tuv, E. Bias of Importance Measures for Multi-valued Attributes and Solutions. In Proceedings of the Artificial Neural Networks and Machine Learning—ICANN 2011, Espoo, Finland, 14–17 June 2011; Honkela, T., Duch, W., Girolami, M., Kaski, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 293–300.
55. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 6.
56. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
57. Ghaddar, B.; Naoum-Sawaya, J. High dimensional data classification and feature selection using support vector machines. *Eur. J. Oper. Res.* **2018**, *265*, 993–1004. [CrossRef]
58. Holte, R.C.; Acker, L.E. Concept Learning and the Problem of Small Disjuncts. *IJCAI* **1989**, *89*, 813–818.
59. Weiss, G.M. Mining with rarity: A unifying framework. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 7–19. [CrossRef]
60. Blagus, R.; Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* **2013**, *14*, 106. [CrossRef] [PubMed]
61. Rahman, M.M.; Davis, D.N. Cluster Based Under-Sampling for Unbalanced Cardiovascular Data. *Proc. World Congr. Eng.* **2013**, *3*, 3–5.
62. Drummond, C.; Holte, R.C. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. In Proceedings of the Workshop on Learning from Imbalanced Datasets II, Washington, DC, USA, 21 August 2003; Volume 11, pp. 1–8.
63. Nguyen, H.M.; Cooper, E.W.; Kamei, K. A comparative study on sampling techniques for handling class imbalance in streaming data. In Proceedings of the The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems, Kobe, Japan, 20–24 November 2012; pp. 1762–1767.
64. Van Hulse, J.; Khoshgoftaar, T.M.; Napolitano, A. Experimental perspectives on learning from imbalanced data. In Proceedings of the 24th International Conference on Machine Learning, New York, NY, USA, 20–24 June 2007; Association for Computing Machinery: New York, NY, USA, 2007; pp. 935–942.
65. Chawla, N.V.; Japkowicz, N.; Kotcz, A. Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 1–6. [CrossRef]
66. Fernández, A.; del Río, S.; Chawla, N.V.; Herrera, F. An insight into imbalanced Big Data classification: Outcomes and challenges. *Complex Intell. Syst.* **2017**, *3*, 105–120. [CrossRef]
67. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
68. He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
69. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
70. Rostamzadeh, N.; Abdullah, S.S.; Sedig, K. Visual Analytics for Electronic Health Records: A Review. *Informatics* **2021**, *8*, 12. [CrossRef]
71. Mane, K.K.; Bizon, C.; Schmitt, C.; Owen, P.; Burchett, B.; Pietrobon, R.; Gersing, K. VisualDecisionLinc: A visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry. *J. Biomed. Inform.* **2012**, *45*, 101–106. [CrossRef]
72. Baytas, I.M.; Lin, K.; Wang, F.; Jain, A.K.; Zhou, J. PhenoTree: Interactive Visual Analytics for Hierarchical Phenotyping From Large-Scale Electronic Health Records. *IEEE Trans. Multimed.* **2016**, *18*, 2257–2270. [CrossRef]
73. Ha, H.; Lee, J.; Han, H.; Bae, S.; Son, S.; Hong, C.; Shin, H.; Lee, K. Dementia Patient Segmentation Using EMR Data Visualization: A Design Study. *Int. J. Environ. Res. Public. Health* **2019**, *16*, 3438. [CrossRef] [PubMed]

74. Guo, R.; Fujiwara, T.; Li, Y.; Lima, K.M.; Sen, S.; Tran, N.K.; Ma, K.-L. Comparative Visual Analytics for Assessing Medical Records with Sequence Embedding. *Vis. Inform.* **2020**, *4*, 72–85. [CrossRef]
75. Hund, M.; Böhm, D.; Sturm, W.; Sedlmair, M.; Schreck, T.; Ullrich, T.; Keim, D.A.; Majnaric, L.; Holzinger, A. Visual analytics for concept exploration in subspaces of patient groups. *Brain Inform.* **2016**, *3*, 233–247. [CrossRef]
76. Huang, C.-W.; Lu, R.; Iqbal, U.; Lin, S.-H.; Nguyen, P.A.; Yang, H.-C.; Wang, C.-F.; Li, J.; Ma, K.-L.; Li, Y.-C.; et al. A richly interactive exploratory data analysis and visualization tool using electronic medical records. *BMC Med. Inform. Decis. Mak.* **2015**, *15*, 92. [CrossRef] [PubMed]
77. Levy, A.R.; O'Brien, B.J.; Sellors, C.; Grootendorst, P.; Willison, D. Coding accuracy of administrative drug claims in the Ontario Drug Benefit database. *Can. J. Clin. Pharmacol. J. Can. Pharmacol. Clin.* **2003**, *10*, 67–71.
78. Collister, D.; Pannu, N.; Ye, F.; James, M.; Hemmelgarn, B.; Chui, B.; Manns, B.; Klarenbach, S. Health Care Costs Associated with AKI. *Clin. J. Am. Soc. Nephrol. CJASN* **2017**, *12*, 1733–1743. [CrossRef] [PubMed]
79. Liangos, O.; Wald, R.; O'Bell, J.W.; Price, L.; Pereira, B.J.; Jaber, B.L. Epidemiology and outcomes of acute renal failure in hospitalized patients: A national survey. *Clin. J. Am. Soc. Nephrol. CJASN* **2006**, *1*, 43–51. [CrossRef]
80. Thongprayoon, C.; Qureshi, F.; Petnak, T.; Cheungpasitporn, W.; Chewcharat, A.; Cato, L.D.; Boonpheng, B.; Bathini, T.; Hansrivijit, P.; Vallabhajosyula, S.; et al. Impact of Acute Kidney Injury on Outcomes of Hospitalizations for Heat Stroke in the United States. *Dis. Basel Switz.* **2020**, *8*, 28. [CrossRef]
81. Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Garg, A.X.; McArthur, E. Predicting Acute Kidney Injury: A Machine Learning Approach Using Electronic Health Records. *Information* **2020**, *11*, 386. [CrossRef]
82. Han, H.; Wang, W.-Y.; Mao, B.-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *Advances in Intelligent Computing*; Huang, D.-S., Zhang, X.-P., Huang, G.-B., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887.
83. Ferri, C.; Hernández-Orallo, J.; Modroiu, R. An experimental comparison of performance measures for classification. *Pattern Recognit. Lett.* **2009**, *30*, 27–38. [CrossRef]
84. Garcıa, V.; Sánchez, J.S.; Mollineda, R.A. On the suitability of numerical performance measures for class imbalance problems. In Proceedings of the International Conference in Pattern Recognition Applications and Methods, Algarve, Portugal, 6–8 February 2012; pp. 310–313.
85. Parikh, R.; Mathai, A.; Parikh, S.; Chandra Sekhar, G.; Thomas, R. Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.* **2008**, *56*, 45–50. [CrossRef] [PubMed]
86. Rostamzadeh, N. Visual Analytics for Performing Complex Tasks with Electronic Health Records. Ph.D. Thesis, University of Western Ontario, London, ON, Canada, 2021.