informatics

MDPI

Article

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

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

[1] Insight Lab, Western University, London, ON N6A3K7, Canada; sabdul9@uwo.ca (S.S.A.); nrostamz@uwo.ca (N.R.)
[2] Department of Medicine, Epidemiology and Biostatistics, Western University, London, ON N6A3K7, 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 (ext. 86612)

Received: 11 May 2020; Accepted: 24 May 2020; Published: 27 May 2020

Abstract: Recent advancement in EHR-based (Electronic Health Record) systems has resulted in producing data at an unprecedented rate. The complex, growing, and high-dimensional data available in EHRs creates great opportunities for machine learning techniques such as clustering. Cluster analysis often requires dimension reduction to achieve efficient processing time and mitigate the curse of dimensionality. Given a wide range of techniques for dimension reduction and cluster analysis, it is not straightforward to identify which combination of techniques from both families leads to the desired result. The ability to derive useful and precise insights from EHRs requires a deeper understanding of the data, intermediary results, configuration parameters, and analysis processes. Although these tasks are often tackled separately in existing studies, we present a visual analytics (VA) system, called Visual Analytics for Cluster Analysis and Dimension Reduction of High Dimensional Electronic Health Records (VALENCIA), to address the challenges of high-dimensional EHRs in a single system. VALENCIA brings 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 tasks in such a way that would be difficult without the aid of a VA system. Through a real case study, we have demonstrated how VALENCIA can be used to analyze the healthcare administrative dataset stored at ICES. This research also highlights what needs to be considered in the future when developing VA systems that are designed to derive deep and novel insights into EHRs.

Keywords: visual analytics; dimension reduction; cluster analysis; electronic health records; high-dimensional data; interactive visualization; human-data interaction

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 [1,2]. EHR data includes, but is not limited to, medical and demographic records, healthcare administrative records, and results of laboratory tests [3]. The complex, diverse, and growing information available in EHRs creates promising opportunities for the healthcare providers to drastically improve the healthcare system [2,4]. It is often challenging for healthcare providers to keep pace with the large volumes of heterogeneous data stored in EHRs [5]. Automated data analysis techniques based on data mining and machine learning hold great promise to fulfill the computational requirements of EHRs [6,7]. There are currently a variety of efforts underway

to organize, analyze, and interpret EHRs using unsupervised machine learning techniques such as clustering [6,8–12].

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 [13,14]. 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 [11], implausible clinical observations [12], and risk factors for a disease [15]. 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, a very common occurrence in EHRs [16]. In such situations, the output of CA is not reproducible and meaningful since variances among data elements become sparse and large [17,18]. One solution is to employ dimensionality reduction (DR) techniques that can potentially reduce the number of features to a manageable size before using CA [19].

DR refers to the transformation of the original high-dimensional dataset into a new dataset with reduced dimensionality without loss of much information [20]. DR techniques are developed based on the idea that most high-dimensional datasets contain overlapping information [21]. 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 [22]. 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) [23]. 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 [24]. Similarly, one needs to determine the optimal values for the configuration parameters when using a DR technique [25]. 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 [22]. 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 [26]. 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 [27,28]. They allow exploration of the visualized data to answer user-initiated queries [29]. In recent years, several EHR-based visualization systems have been developed to support healthcare providers in performing various user-driven activities [30]. 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 space on display devices [31–34]. 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 [35–37]. 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 [38].

The purpose of this paper 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 [24,39–43], 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.

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.

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 [27,44]. 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 [27,29,38,45–47]. 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 [48,49].

Human cognition is limited when confronted with data-intensive tasks, especially when the data is high-dimensional and complex [38,50]. 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 [38]. The pre-processing stage is responsible for preparing raw data from different sources, which includes procedures such as cleaning, integration, and reduction [51]. Next, in the

transformation stage, the pre-processed data is transformed into forms suitable for analysis [52]. The transformation stage includes procedures such as smoothing, aggregation, feature generation, discretization, and normalization [53]. 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 [54,55]. 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 [38]. 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 [56,57].

2.2. Dimension Reduction (DR)

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). This has long been one of the leading research topics in statistics, data mining, and machine learning [58]. 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 [59–62]. 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 [63]. 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 [19,64,65]. 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.

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. For instance, 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 [24,39].

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 [66]. 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.

2.4. Healthcare Stakeholders

For the purposes of this paper, 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 [67]. 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 [68], or clinical researchers develop frameworks to discover temporal knowledge from healthcare administrative data [42]. 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 [69,70]). 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.

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.

3.1. DR-Based Visual Analytics Systems

GGobi19 [61] is a visual analytics system that uses a DR technique called grand tour [71] 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—principal component analysis) to represent high-dimensional data is iPCA [62]. 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 [72] 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 [73] 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.

3.2. CA-Based Visual Analytics Systems

The Hierarchical Clustering Explorer (HCE) [74] 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 [75] allows users to arrange and compare multiple clusters simultaneously using heatmaps and parallel coordinates. It shows raw data along with the clustering results. ClusterSculptor [76] 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 [77] 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 [78] resolves this issue by rearranging clusters and their items to minimize visual clutter between features. XCluSim [79] 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.

3.3. DR and CA-Based Visual Analytics Systems

IN-SPIRE [40] 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 [41]. 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 [39] 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 [24]; 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".

3.4. EHR-Based Visual Analytics Systems

MatrixFlow [80] 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 [81]) with interactive visualizations to represent networks of clinical events as a temporal flow of matrices. Another visual analytics system is VisualDecisionLinc [82] 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. [35] 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) [42] 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 [83] 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 [84] allows interactive exploration of patient cohorts and interpretation of hierarchical phenotypes by integrating sparse PCA with an interactive visual interface. VISA_M3R3 [85] 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.

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 Section 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, Section 4.5, and Section 4.6. Finally, Section 4.7 outlines the implementation details of VALENCIA.

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 [86]. 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 [87]. 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 requirement 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 at every stage of the design and development process. This process was essential to build trust between the proposed system and its end-users.

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.

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.

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.

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.

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.

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.

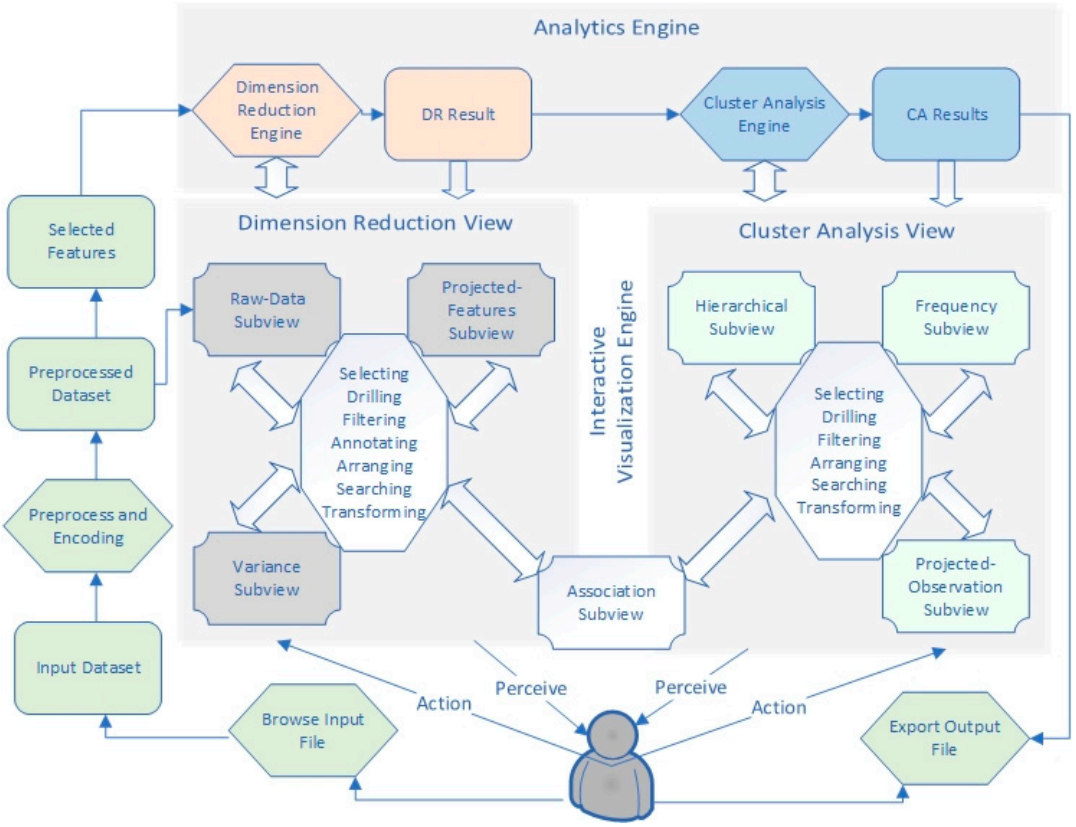

Figure 1. Basic workflow of Visual Analytics for Cluster Analysis and Dimension Reduction of High Dimensional Electronic Health Records (VALENCIA). The backgrounds of the components are color-coded to show the similarity between processes.

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.

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.

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., [19,88–92]). 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.

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 [93], classical multidimensional scaling (CMDS) [94], principal component analysis (PCA) [95,96], multiple correspondence analysis (MCA) [97], and multiple factor analysis (MFA) [98] 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) [99] and nonmetric multidimensional scaling (NMDS) [100,101] 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 [102,103]. 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 [104]. 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 [99]. 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 using PCA 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.

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 [19]. 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 [105,106], 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 [107] does not require users to specify the number of clusters initially. It comes in two forms: agglomerative and divisive [108]. 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 [109] 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 [110] 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 [111]. 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.

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.

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).

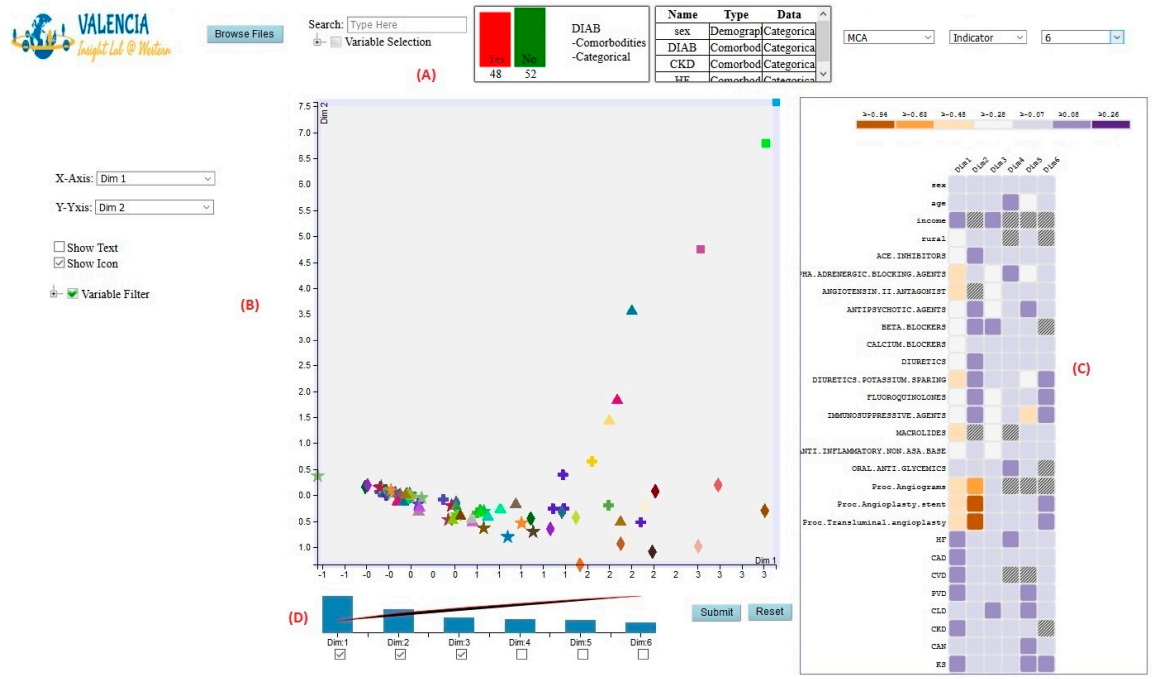

Figure 2. The dimensionality reduction (DR) view containing (**A**) raw-data subview, (**B**) projected-features subview, (**C**) association subview, and (**D**) variance subview.

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.

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.

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.

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.

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.

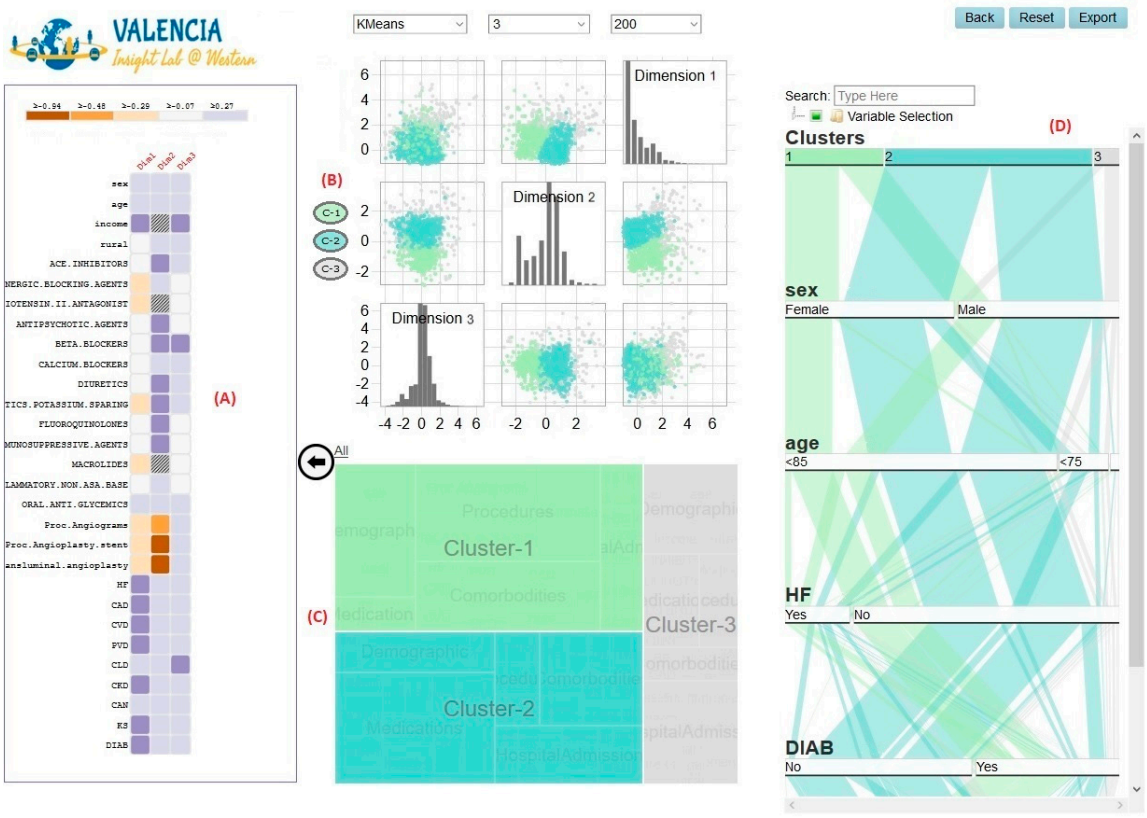

Figure 3. The dimensionality reduction (CA) view containing (**A**) association subview, (**B**) projected-observations subview (**C**) hierarchical subview, and (**D**) frequency subview.

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 Table 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.

Table 1. Showing how the size of the rectangle is computed in the hierarchical subview.

Require: Raw dataset with cluster labels	(1)
compute the number of features in each group in number_of_groupfeatures []	(2)
compute max_groupfeatures = maximum value in number_of_groupfeatures []	(3)
compute frequency of each feature in the dataset	(4)
divide the dataset based on each cluster	(5)
for each cluster C in the dataset	(6)
for each feature F in the dataset	(7)
compute relative frequencies of feature F in cluster C	(8)
feature_weight = (relative frequencies/frequency [F]) × 100	(9)
adjusted_feature_weight [C,F] = (max_groupfeatures/number_of_groupfeatures [F]) × feature_weight	(10)
return adjusted_feature_weight [][]	(11)

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.

Frequency Subview

The frequency subview includes a parallel set, collapsible tree structure, search bar, and checkbox. Parallel Sets [112] 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.

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.

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. 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.

5.1. 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 [113]. We ascertained hospital admission, procedure, baseline comorbidity, and emergency department visit data from the National Ambulatory Care Reporting System (i.e., ED—emergency department 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.

5.2. 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 five-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.

5.3. Cohort Description

There were a total of 47 unique features and about one 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%).

5.4. 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 paper, 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 4A). As shown in Figure 4B, 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 4C.

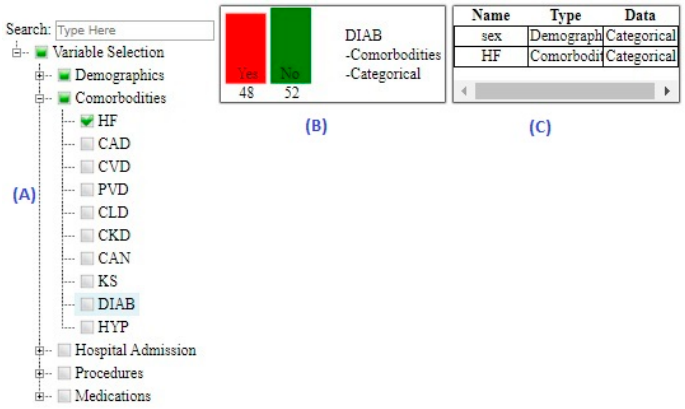

Figure 4. The raw-data subview containing (**A**) collapsible tree structure, (**B**) bar chart, and (**C**) data table.

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.

Then, VALENCIA updates the projected-features, association, and variance subviews when the data items are generated. As shown in Figure 5A, 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 5B, 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 2, 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 5B). 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 5B. 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 5C).

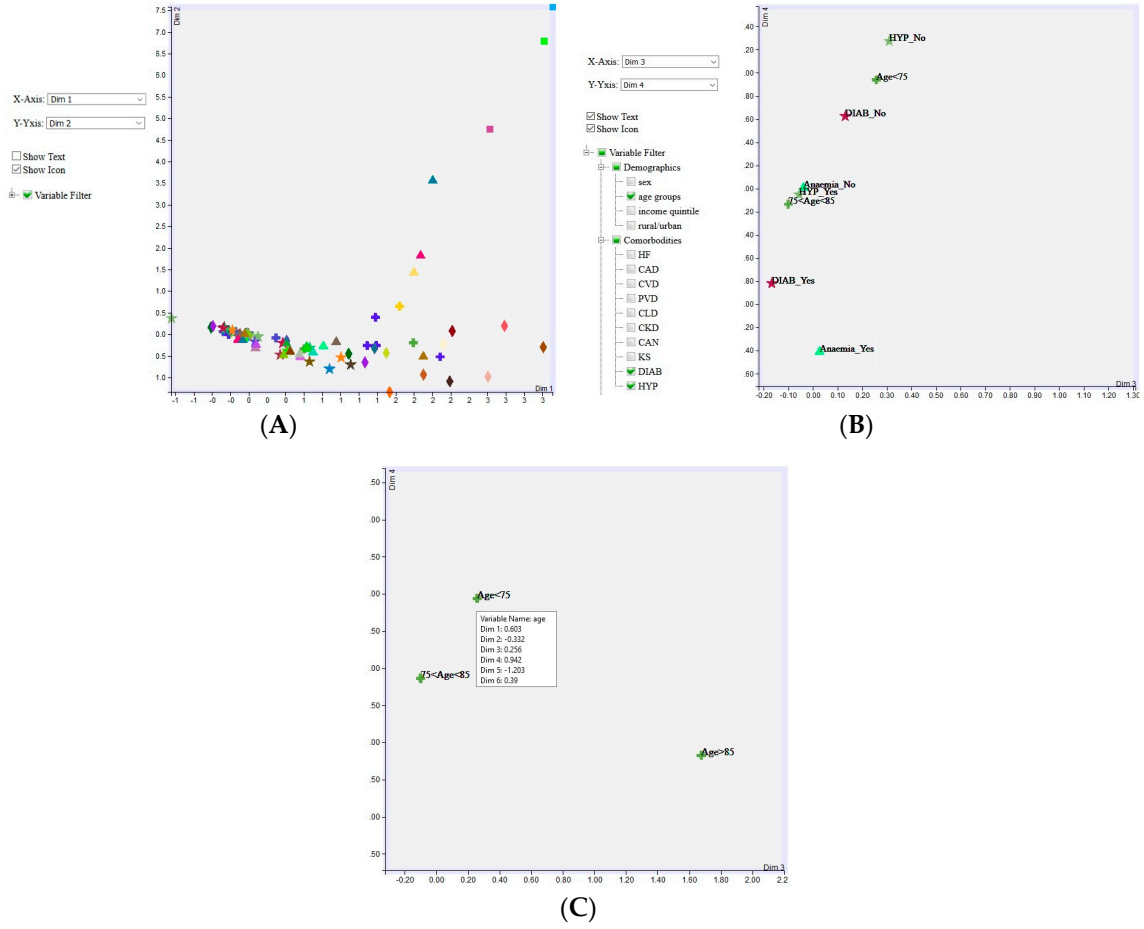

Figure 5. Showing an overview of the projected-features subview, which includes (**A**) all glyphs with respect to dimensions one and two, (**B**) some selected glyphs and labels with regard to dimensions three and four, and (**C**) all the glyphs and labels representing age upon drilling.

Table 2. 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 5B 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 6A, 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 6B). Then the user can select these features in the projected-features subview to investigate these correlations more closely.

(**A**) (**B**)

Figure 6. 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.

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.

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 7A,

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 7A). 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 and two). This helps the user to investigate how the observations in the selected region are distributed in other scatter plots (Figure 7B). As shown in Figure 7C, when the user clicks on button "C-3", the system removes all the observations belonging to cluster three. 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.

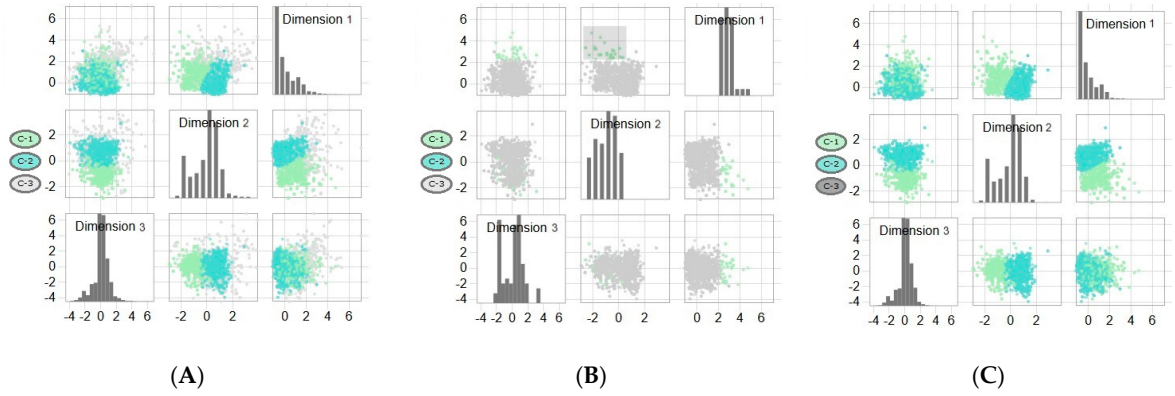

Figure 7. Showing the overview of the projected-observation subview, which displays (**A**) all the observations color-coded based on clusters, (**B**) the brushing interaction, and (**C**) observations in cluster-1 and cluster-2 because cluster-3 is filtered out.

The hierarchical subview allows the user to detect which clusters cover the maximum amount of variation of the data (Figure 8A). 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 8B–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 8E).

The user can consult the frequency subview to get the frequency distribution of clusters and features. As shown in Figure 9A, 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 9B displays the frequencies of the comorbidities when the user selects the "cluster-3" -> "comorbidities" rectangles (Figure 8E) 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.

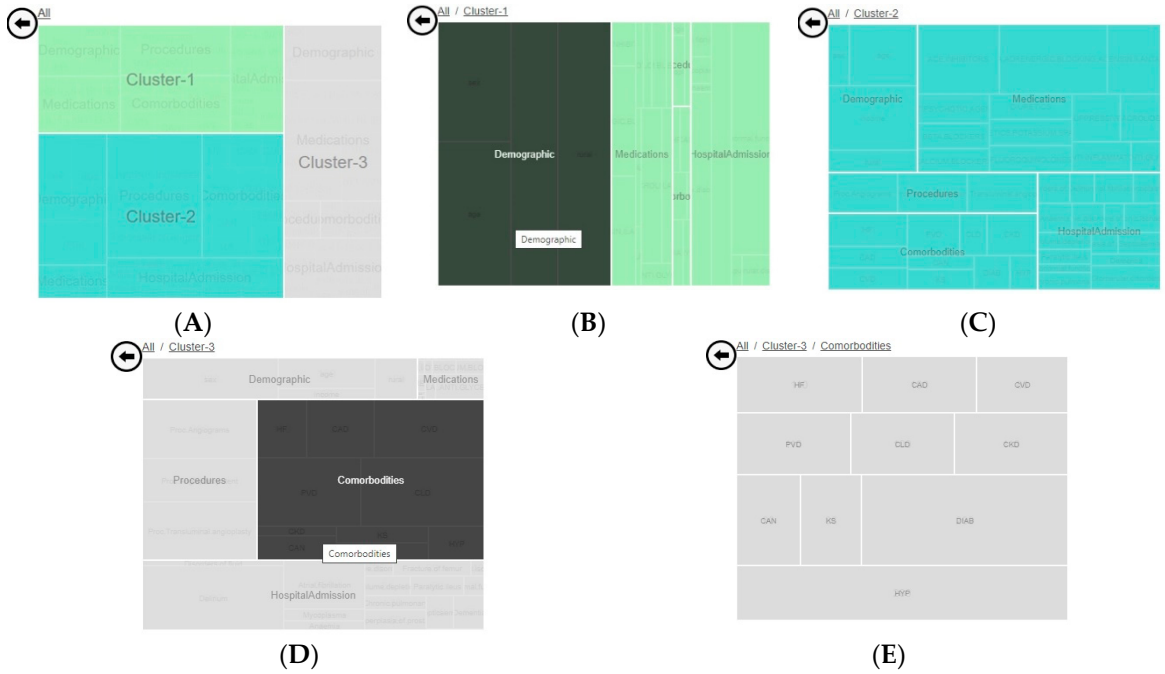

Figure 8. Showing the overview of the hierarchical subview, which displays (**A**) all the clusters, (**B**) feature groups within cluster-1, (**C**) feature groups within cluster-2, (**D**) feature groups within cluster-3, and (**E**) features within the comorbidity group in cluster-3.

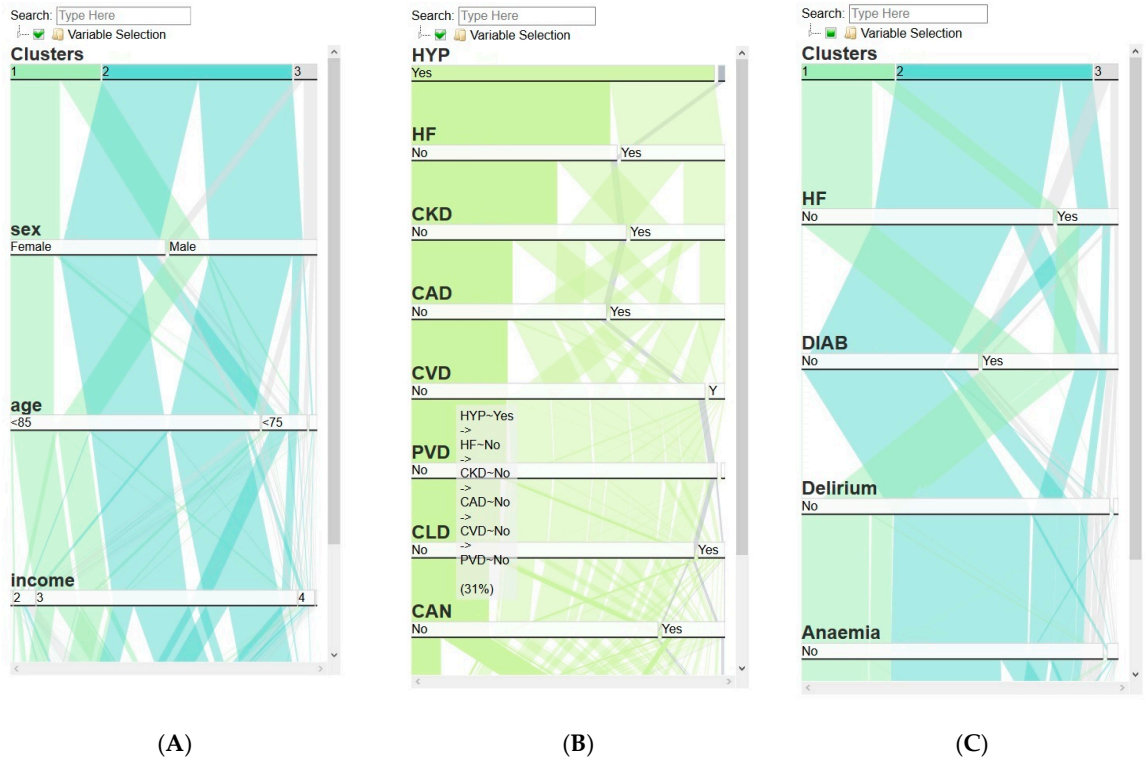

Figure 9. Showing the overview of the frequency subview, which shows the distribution of (**A**) different clusters and demographics, (**B**) all the comorbidities within a particular cluster, and (**C**) clusters and some user-selected features.

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 9C 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).

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.

6. Conclusions

In this paper, 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 paper 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.

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

Funding: This research received no external funding.

Acknowledgments: We would like to thank all ICES and Western staff who helped us throughout the process.

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

References

1. Caban, J.J.; Gotz, D. Visual analytics in healthcare-opportunities and research challenges. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 260–262. [CrossRef] [PubMed]
2. Murdoch, T.B.; Detsky, A.S. The inevitable application of big data to health care. *Jama J. Am. Med. Assoc.* **2013**, *309*, 1351–1352. [CrossRef] [PubMed]
3. Cowie, M.R.; Blomster, J.I.; Curtis, L.H.; Duclaux, S.; Ford, I.; Fritz, F.; Goldman, S.; Janmohamed, S.; Kreuzer, J.; Leenay, M.; et al. Electronic health records to facilitate clinical research. *Clin. Res. Cardiol.* **2017**, *106*, 1–9. [CrossRef] [PubMed]
4. Kamal, N. Big Data and Visual Analytics in Health and Medicine: From Pipe Dream to Reality. *J. Health Med. Inform.* **2014**, 5. [CrossRef]
5. Rind, A.; Wagner, M.; Aigner, W. Towards a Structural Framework for Explicit Domain Knowledge in Visual Analytics. In Proceedings of the 2019 IEEE Workshop on Visual Analytics in Healthcare (VAHC), Vancouver, BC, Canada, 20–20 October 2019; pp. 33–40. [CrossRef]
6. 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, Miami, FL, USA, 28–30 January 2012; ACM Press: Miami, FL, USA, 2012; p. 389.
7. Wetzel, R.C. The virtual pediatric intensive care unit: Practice in the new millennium. *Pediatric Clin.* **2001**, *48*, 795–814.
8. Haraty, R.A.; Dimishkieh, M.; Masud, M. An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data. *Int. J. Distrib. Sens. Netw.* **2015**, *11*, 615740. [CrossRef]
9. 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, Portugal, 19–21 January 2018; SCITEPRESS—Science and Technology Publications: Funchal, Portugal, 2018; pp. 266–273.
10. Liao, M.; Li, Y.; Kianifard, F.; Obi, E.; Arcona, S. Cluster analysis and its application to healthcare claims data: A study of end-stage renal disease patients who initiated hemodialysis. *BMC Nephrol.* **2016**, *17*, 25. [CrossRef]
11. Foguet-Boreu, Q.; Violán, C.; Rodriguez-Blanco, T.; Roso-Llorach, A.; Pons-Vigués, M.; Pujol-Ribera, E.; Cossio Gil, Y.; Valderas, J.M. Multimorbidity Patterns in Elderly Primary Health Care Patients in a South Mediterranean European Region: A Cluster Analysis. *PLoS ONE* **2015**, *10*, e0141155. [CrossRef]
12. Estiri, H.; Klann, J.G.; Murphy, S.N. A clustering approach for detecting implausible observation values in electronic health records data. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 142. [CrossRef]
13. Dilts, D.; Khamalah, J.; Plotkin, A. Using cluster analysis for medical resource decision making. *Med. Decis. Mak.* **1995**, *15*, 333–346. [CrossRef]
14. McLachlan, G.J. Cluster analysis and related techniques in medical research. *Stat. Methods Med. Res.* **1992**, *1*, 27–48. [CrossRef] [PubMed]

15. Doust, D.; Walsh, Z. Data Mining Clustering: A Healthcare Application. In Proceedings of the Mediterranean Conference on Information Systems (MCIS), Limassol, Cyprus, 3–5 September 2011.
16. Ruan, T.; Lei, L.; Zhou, Y.; Zhai, J.; Zhang, L.; He, P.; Gao, J. Representation learning for clinical time series prediction tasks in electronic health records. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 259. [CrossRef] [PubMed]
17. Adachi, S. Rigid geometry solves "curse of dimensionality" effects in clustering methods: An application to omics data. *PLoS ONE* **2017**, 12. [CrossRef]
18. Ronan, T.; Qi, Z.; Naegle, K.M. Avoiding common pitfalls when clustering biological data. *Sci. Signal* **2016**, *9*, re6. [CrossRef] [PubMed]
19. Mitsuhiro, M.; Yadohisa, H. Reduced k-means clustering with MCA in a low-dimensional space. *Comput. Stat.* **2015**, *30*, 463–475. [CrossRef]
20. Siwek, K.; Osowski, S.; Markiewicz, T.; Korytkowski, J. Analysis of medical data using dimensionality reduction techniques. *Przegląd Elektrotechniczny* **2013**, *89*, 279–281.
21. Wilke, C.O. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*, 1st ed.; O'Reilly Media: Sebastopol, CA, USA, 2019; ISBN 978-1-4920-3108-6.
22. Wenskovitch, J.; Crandell, I.; Ramakrishnan, N.; House, L.; Leman, S.; North, C. Towards a Systematic Combination of Dimension Reduction and Clustering in Visual Analytics. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 131–141. [CrossRef]
23. Sembiring, R.W.; Zain, J.M.; Embong, A. Dimension Reduction of Health Data Clustering. *arXiv* **2011**, arXiv:1110.3569.
24. Demiralp, Ç. Clustrophile: A tool for visual clustering analysis. *arXiv* **2017**, arXiv:1710.02173.
25. Halpern, Y.; Horng, S.; Nathanson, L.A.; Shapiro, N.I.; Sontag, D. A comparison of dimensionality reduction techniques for unstructured clinical text. In Proceedings of the Icml 2012 Workshop on Clinical Data Analysis, Edinburgh, UK, 30 June–1 July 2012; Volume 6.
26. Yoo, I.; Alafaireet, P.; Marinov, M.; Pena-Hernandez, K.; Gopidi, R.; Chang, J.F.; Hua, L. Data mining in healthcare and biomedicine: A survey of the literature. *J. Med. Syst.* **2012**, *36*, 2431–2448. [CrossRef]
27. Keim, D.A.; Mansmann, F.; Thomas, J. Visual analytics: How much visualization and how much analytics? *Sigkdd Explor. Newsl.* **2010**, *11*, 5–8. [CrossRef]
28. Cook, K.A.; Thomas, J.J. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*; Pacific Northwest National Lab (PNNL): Richland, WA, USA, 2005.
29. 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]
30. Rind, A.; Aigner, W.; Miksch, S.; Wiltner, S.; Pohl, M.; Turic, T.; Drexler, F. Visual exploration of time-oriented patient data for chronic diseases: Design study and evaluation. In *Symposium of the Austrian HCI and Usability Engineering Group*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 301–320.
31. Aimone, A.M.; Perumal, N.; Cole, D.C. 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. *Int. J. Health Geogr.* **2013**, 12. [CrossRef] [PubMed]
32. Faisal, S.; Blandford, A.; Potts, H.W. Making sense of personal health information: Challenges for information visualization. *Health Inform. J.* **2013**, *19*, 198–217. [CrossRef] [PubMed]
33. Kosara, R.; Miksch, S. Visualization methods for data analysis and planning in medical applications. *Int. J. Med. Inform.* **2002**, *68*, 141–153. [CrossRef]
34. Lavado, R.; Hayrapetyan, S.; Kharazyan, S. *Expansion of the Benifits Package: The Experience of Armenia*; World Bank: Washington, DC, USA, 2018; pp. 1–36.
35. 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. 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.* **2015**, *22*, 361–369. [CrossRef]
36. Saffer, J.D.; Burnett, V.L.; Chen, G.; van der Spek, P. Visual analytics in the pharmaceutical industry. *IEEE Comput. Graph. Appl.* **2004**, *24*, 10–15. [CrossRef]
37. Parsons, P.; Sedig, K.; Mercer, R.E.; 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; Association for Computing Machinery: Chicago, IL, USA, 2015; pp. 1–8.

38. 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]
39. Choo, J.; Lee, H.; Liu, Z.; Stasko, J.; Park, H. 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: Washington, DC, USA, 2013; Volume 8654, p. 865402.
40. Wise, J.A. The ecological approach to text visualization. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 1224–1233. [CrossRef]
41. Stasko, J.; Görg, C.; Liu, Z. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. *Inf. Vis.* **2008**, *7*, 118–132. [CrossRef]
42. Klimov, D.; Shknevsky, A.; Shahar, Y. Exploration of patterns predicting renal damage in patients with diabetes type II using a visual temporal analysis laboratory. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 275–289. [CrossRef] [PubMed]
43. Ninkov, A.; Sedig, K. VINCENT: A visual analytics system for investigating the online vaccine debate. *Online J. Public Health Inform.* **2019**, *11*, e5. [CrossRef] [PubMed]
44. Thomas, J.J.; Cook, K.A. A visual analytics agenda. *IEEE Comput. Graph. Appl.* **2006**, *26*, 10–13. [CrossRef] [PubMed]
45. Cui, W. Visual Analytics: A Comprehensive Overview. *IEEE Access* **2019**, *7*, 81555–81573. [CrossRef]
46. 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. [CrossRef]
47. Parsons, P.; Sedig, K. Distribution of information processing while performing complex cognitive activities with visualization tools. In *Handbook of Human Centric Visualization*; Springer: New York, NY, USA, 2014; pp. 693–715, ISBN 978-1-46-147485-2.
48. Sears, A.; Jacko, J.A. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2007; ISBN 978-1-41-061586-2.
49. Sedig, K.; Parsons, P. Design of visualizations for human-information interaction: A pattern-based framework. *Synth. Lect. Vis.* **2016**, *4*, 1–185. [CrossRef]
50. Green, T.M.; Maciejewski, R. A role for reasoning in visual analytics. In Proceedings of the Annual Hawaii International Conference on System Sciences, Wailea, HI, USA, 7–10 January 2013, pp. 1495–1504.
51. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; The Morgan Kaufmann Series in Data Management Systems; Elsevier: Amsterdam, The Netherlands, 2011.
52. Kusiak, A. Feature transformation methods in data mining. *IEEE Trans. Electron. Packag. Manuf.* **2001**, *24*, 214–221. [CrossRef]
53. Han, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Elsevier: Amsterdam, The Netherlands, 2011.
54. Agrawal, R.; Swami, A.; Imielinski, T. Database Mining: A Performance Perspective. *IEEE Trans. Knowl. Data Eng.* **1993**, *5*, 914–925. [CrossRef]
55. Sahu, H.; Shrma, S.; Gondhalakar, S. A Brief Overview on Data Mining Survey. *Int. J. Comput. Technol. Electron. Eng. (IJCTEE)* **2008**, *1*, 114–121.
56. Keim, D.A.; Mansmann, F.; Schneidewind, J.; Thomas, J.; Ziegler, H. Visual Analytics: Scope and Challenges. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*; Simoff, S.J., Böhlen, M.H., Mazeika, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 76–90, ISBN 978-3-540-71080-6.
57. Kehrer, J.; Hauser, H. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 495–513. [CrossRef]
58. Sorzano, C.O.S.; Vargas, J.; Montano, A.P. A survey of dimensionality reduction techniques. *arXiv* **2014**, arXiv:1403.2877.
59. Geng, X.; Zhan, D.-C.; Zhou, Z.-H. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Trans. Syst. ManCybern. Part B (Cybern.)* **2005**, *35*, 1098–1107. [CrossRef] [PubMed]
60. Fujiwara, T.; Chou, J.-K.; Shilpika; Xu, P.; Ren, L.; Ma, K.-L. An Incremental Dimensionality Reduction Method for Visualizing Streaming Multidimensional Data. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 418–428. [CrossRef] [PubMed]
61. Cook, D.; Swayne, D.F.; Buja, A. *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*; Springer Science & Business Media: New York, NY, USA, 2007.
62. Hege, H.C.; Hotz, I.; Muntzner, T. iPCA: An Interactive System for PCA-Based Visual Analytics. Available online: https://viscenter.uncc.edu/sites/viscenter.uncc.edu/files/CVC-UNCC-09-05_0.pdf (accessed on 11 May 2020).

63. Cunningham, P. Dimension Reduction. In *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*; Cord, M., Cunningham, P., Eds.; Cognitive Technologies; Springer: Berlin/Heidelberg, Germany, 2008; pp. 91–112, ISBN 978-3-54-075171-7.
64. Yan, J.; Zhang, B.; Liu, N.; Yan, S.; Cheng, Q.; Fan, W.; Yang, Q.; Xi, W.; Chen, Z. Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 320–333. [CrossRef]
65. Obaid, H.S.; Dheyab, S.A.; Sabry, S.S. The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning. In Proceedings of the 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON), Jaipur, India, 13–15 March 2019; pp. 279–283.
66. Kameshwaran, K.; Malarvizhi, K. Survey on Clustering Techniques in Data Mining. *Int. J. Comput. Sci. Inf. Technol.* **2014**, *5*, 2272–2276.
67. Davis, E. What is a health care contract? *Health Values* **2019**, *4*, 82–86, 89.
68. Soyiri, I.N.; Reidpath, D.D. An overview of health forecasting. *Environ. Health Prev Med.* **2013**, *18*, 1–9. [CrossRef]
69. SAS Enterprise BI Server. Available online: https://www.sas.com/en_ca/software/enterprise-bi-server.html (accessed on 19 February 2020).
70. Weka 3—Data Mining with Open Source Machine Learning Software in Java. Available online: https://www.cs.waikato.ac.nz/ml/weka/courses.html (accessed on 12 March 2020).
71. Asimov, D. The grand tour: A tool for viewing multidimensional data. *SIAM J. Sci. Stat. Comput.* **1985**, *6*, 128–143. [CrossRef]
72. Cavallo, M.; Demiralp, Ç. A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal QC, Canada, 21–26 April 2018; Association for Computing Machinery: Montreal, QC, Canada, 2018; pp. 1–13.
73. Ali, M.; Jones, M.W.; Xie, X.; Williams, M. TimeCluster: Dimension reduction applied to temporal data for visual analytics. *Vis Comput* **2019**, *35*, 1013–1026. [CrossRef]
74. Seo, J.; Shneiderman, B. Interactively Exploring Hierarchical Clustering Results. In *The Craft of Information Visualization*; Elsevier: Amsterdam, The Netherlands, 2003; pp. 334–340, ISBN 978-1-55-860915-0.
75. Lex, A.; Streit, M.; Partl, C.; Kashofer, K.; Schmalstieg, D. Comparative Analysis of Multidimensional, Quantitative Data. *IEEE Trans. Vis. Comput. Graph.* **2010**, *16*, 1027–1035. [CrossRef]
76. Nam, E.J.; Han, Y.; Mueller, K.; Zelenyuk, A.; Imre, D. ClusterSculptor: A Visual Analytics Tool for High-Dimensional Data. In Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology, Sacramento, CA, USA, 30 October–1 November 2007; pp. 75–82.
77. Ding, H.; Wang, C.; Huang, K.; Machiraju, R. iGPSe: A visual analytic system for integrative genomic based cancer patient stratification. *BMC Bioinform.* **2014**, *15*, 203. [CrossRef] [PubMed]
78. Zhou, J.; Konecni, S.; Grinstein, G. Visually comparing multiple partitions of data with applications to clustering. In *Visualization and Data Analysis 2009*; International Society for Optics and Photonics: Orlando, FL, USA, 2009; Volume 7243, p. 72430J.
79. L'Yi, S.; Ko, B.; Shin, D.; Cho, Y.-J.; Lee, J.; Kim, B.; Seo, J. XCluSim: A visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. *BMC Bioinform.* **2015**, *16*, S5.
80. Perer, A.; Sun, J. MatrixFlow: Temporal network visual analytics to track symptom evolution during disease progression. *AMIA Annu. Symp. Proc.* **2012**, *2012*, 716–725. [PubMed]
81. Heer, J.; Perer, A. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. *Inf. Vis.* **2014**, *13*, 111–133. [CrossRef]
82. 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]
83. Perer, A.; Wang, F.; Hu, J. Mining and exploring care pathways from electronic medical records with visual analytics. *J. Biomed. Inform.* **2015**, *56*, 369–378. [CrossRef]
84. 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]

85. 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]
86. Varga, M.; Varga, C. Visual Analytics: Data, Analytical and Reasoning Provenance. In Building Trust in Information. Springer: Cham, Switzerland, 2016; pp. 141–150.
87. Leighton, J.P. (Ed.) Defining and Describing Reason. In *The Nature of Reasoning*; Cambridge University Press: Cambridge, UK, 2004; pp. 3–11, ISBN 0-521-81090-6.
88. Arabie, P. Cluster analysis in marketing research. *Adv. Methods Mark. Res.* **1994**, 160–189.
89. De Soete, G.; Carroll, J.D. K-means clustering in a low-dimensional Euclidean space. In *New Approaches in Classification and Data Analysis*; Springer: Berlin/Heidelberg, Germany, 1994; pp. 212–219.
90. Vichi, M.; Kiers, H.A. Factorial k-means analysis for two-way data. *Comput. Stat. Data Anal.* **2001**, *37*, 49–64. [CrossRef]
91. Timmerman, M.E.; Ceulemans, E.; Kiers, H.A.; Vichi, M. Factorial and reduced K-means reconsidered. *Comput. Stat. Data Anal.* **2010**, *54*, 1858–1871. [CrossRef]
92. Rocci, R.; Gattone, S.A.; Vichi, M. A new dimension reduction method: Factor discriminant k-means. *J. Classif.* **2011**, *28*, 210–226. [CrossRef]
93. Hirschfeld, H.O. A Connection between Correlation and Contingency. *Math. Proc. Camb. Philos. Soc.* **1935**, *31*, 520–524. [CrossRef]
94. Torgerson, W.S. *Theory and Methods of Scaling*; Wiley: Oxford, UK, 1958.
95. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417–441. [CrossRef]
96. Pearson, K., LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
97. Greenacre, M.; Blasius, J. *Multiple Correspondence Analysis and Related Methods*; CRC press: Boca Raton, FL, USA, 2006.
98. Escofier, B.; Pagès, J. Multiple factor analysis (AFMULT package). *Comput. Stat. Data Anal.* **1994**, *18*, 121–140. [CrossRef]
99. Van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
100. Shepard, R.N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* **1962**, *27*, 219–246. [CrossRef]
101. Kruskal, J.B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **1964**, *29*, 115–129. [CrossRef]
102. Leeuw, J.D. Multivariate Analysis with Optimal Scaling. In Proceedings of the International Conference on Advances in Multivariate Statistical Analysis, Calcutta, India; Indian Statistical Institute: Calcutta, Indian, 1988; pp. 127–160.
103. Gifi, A. *Nonlinear Multivariate Analysis*; Wiley: Hoboken, NJ, USA, 1990.
104. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
105. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108. [CrossRef]
106. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
107. Nielsen, F. Hierarchical Clustering. In *Introduction to HPC with MPI for Data Science*; Nielsen, F., Ed.; Undergraduate Topics in Computer Science; Springer International Publishing: Cham, Switzerland, 2016; pp. 195–211, ISBN 978-3-31-921903-5.
108. Rokach, L.; Maimon, O. Clustering Methods. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer: Boston, MA, USA, 2005; pp. 321–352, ISBN 978-0-38-725465-4.
109. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd* **1996**, *96*, 226–231.
110. Fraley, C.; Raftery, A.E. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *J. Am. Stat. Assoc.* **2002**, *97*, 611–631. [CrossRef]
111. Feldman, R.; Sanger, J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*; Cambridge University Press: Cambridge, UK, 2007; ISBN 978-0-52-183657-9.

112. Kosara, R. Turning a table into a tree: Growing parallel sets into a purposeful project. In *Beautiful Visualization: Looking at Data through the Eyes of Experts*; Steele, J., Iliinsky, N., Eds.; O'Reilly: Sebastopol, CA, USA, 2010; pp. 193–204.
113. 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.* **2003**, *10*, 67–71.