informatics

Review

Visual Analytics for Electronic Health Records: A Review

Neda Rostamzadeh, Sheikh S. Abdullah and Kamran Sedig *

Insight Lab, Western University, London, ON N6A 3K7, Canada; nrostamz@uwo.ca (N.R.);
sabdul9@uwo.ca (S.S.A.)
* Correspondence: sedig@uwo.ca; Tel.: +1-519-661-2111

Abstract: The increasing use of electronic health record (EHR)-based systems has led to the generation of clinical data at an unprecedented rate, which produces an untapped resource for healthcare experts to improve the quality of care. Despite the growing demand for adopting EHRs, the large amount of clinical data has made some analytical and cognitive processes more challenging. The emergence of a type of computational system called visual analytics has the potential to handle information overload challenges in EHRs by integrating analytics techniques with interactive visualizations. In recent years, several EHR-based visual analytics systems have been developed to fulfill healthcare experts' computational and cognitive demands. In this paper, we conduct a systematic literature review to present the research papers that describe the design of EHR-based visual analytics systems and provide a brief overview of 22 systems that met the selection criteria. We identify and explain the key dimensions of the EHR-based visual analytics design space, including visual analytics tasks, analytics, visualizations, and interactions. We evaluate the systems using the selected dimensions and identify the gaps and areas with little prior work.

Keywords: electronic health records; visual analytics; interaction design; visual analytics tasks; analytics techniques; visualization

check for updates

Citation: Rostamzadeh, N.; Abdullah, S.S.; Sedig, K. Visual Analytics for Electronic Health Records: A Review. *Informatics* **2021**, *8*, 12. https://doi.org/10.3390/informatics8010012

Received: 17 December 2020
Accepted: 20 February 2021
Published: 23 February 2021

1. Introduction

In recent years, medical organizations are increasingly deploying electronic health record (EHR)-based systems that generate, store, and manage their data. Therefore, the amount of data available to clinical researchers and clinicians continues to grow at an unprecedented rate, creating an untapped resource with the capacity to improve the healthcare system [1]. The EHR-based systems are used to detect hidden patterns and trends, monitor patient conditions [2], reduce medical errors [3], detect adverse drug events [4,5], and ultimately improve quality of care [6–8]. However, despite the evidence showing the benefits of EHR-based systems, they rarely improve healthcare experts' ability to make better clinical decisions by having access to more comprehensive information [9,10]. Access to large volumes of clinical data has made some analytical and cognitive processes more difficult for healthcare experts. As the amount of data stored in EHRs continues to grow exponentially, and new EHR-based systems are implemented for those already overrun with too much data, there is a growing demand for computational systems that can handle the huge amount of clinical data.

Visual analytics (VA) systems have shown significant promise in addressing information overload challenges in EHRs by combining analytics techniques with interactive visualizations [11,12]. For a VA system to work well, there must be a strong coupling among all its components [13,14]. Such components include but are not limited to tasks, interactive visual representations, and analytics techniques. Analytics has the potential to facilitate healthcare experts' clinical decision-making process by using techniques from various fields such as statistics, machine learning, and data mining. Completing analytics, interactive visualizations allow healthcare experts to explore the underlying data, alter the representations, and guide the analytics techniques to accomplish their tasks [15–17]. VA systems fuse the strengths of both analytics techniques and interactive visualizations to support the execution of EHR-driven tasks. VA is needed to support the intuitive analysis of EHRs for healthcare experts while masking the data's underlying complexity. Clinical researchers can use VA to perform population-based analysis and gain insights from large volumes of patient data. Moreover, VA can also support physicians in tracking symptom evolution during disease progression and creating and visualizing detection models for disease surveillance [18–21]. The complex and diverse challenges and applications of VA in the analysis and exploration of EHRs have led to the development of several EHR-based VA systems, which aim to fulfill the computational and cognitive needs of healthcare experts. The design and development of such systems require collaboration with healthcare experts to assess their requirements and challenges and to better understand EHR-driven tasks from their perspective. This motivates us to systematically study and gather healthcare experts' needs and expectations and get an overview of the state-of-the-art EHR-based VA systems.

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

The rest of the review is organized as follows. Section 2 presents the strategy for searching relevant literature and explains the selection criteria. Section 3 provides a brief overview of the EHR-based VA systems that met the selection criteria. In Section 4, we identify and explain the key dimensions of the EHR-based VA design space. In Section 5, we discuss how the selected EHR-based VA systems support these dimensions and identify the gaps. Finally, Section 6 concludes the paper.

2. Methods

2.1. Search Strategy

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

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

KEYWORDS: (K1) AND (K2) AND (K3)	
within each group, the keywords are combined by the "OR" operator	
K1 (Visualization)	Visualization or visual
K2 (Analytics)	Analytics or analysis or data mining or machine learning
K3 (EHR [1])	EHR or electronic health record or electronic medical record or EMR [2] or healthcare record or patient record or clinical data

[1] Electronic Health Records; [2] Electronic Medical Records.

2.2. Inclusion and Exclusion Criteria

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

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

2.3. Article Selection and Analysis

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

2.4. Results

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

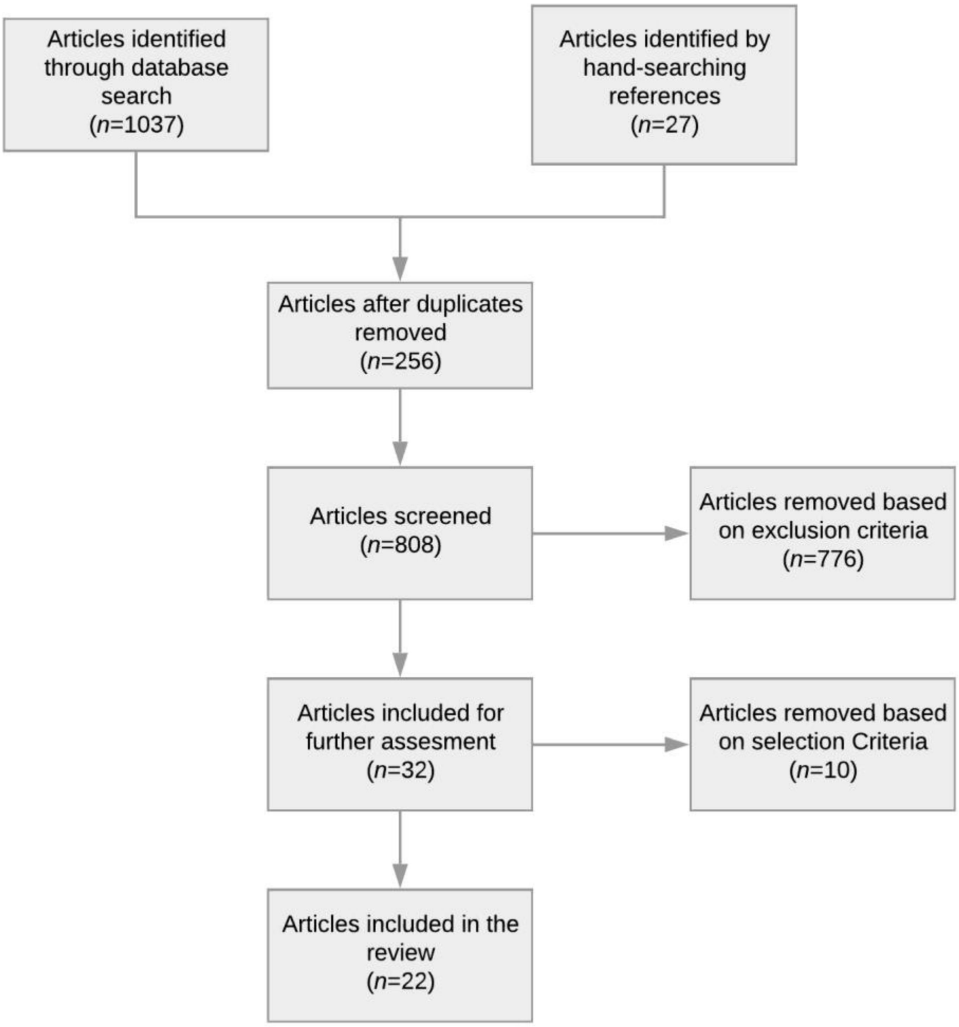

Articles identified through database search (*n*=1037)
Articles identified by hand-searching references (*n*=27)
Articles after duplicates removed (*n*=256)
Articles screened (*n*=808)
Articles removed based on exclusion criteria (*n*=776)
Articles included for further assesment (*n*=32)
Articles removed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

Figure 1. Flow diagram of literature search results.

3. EHR-Based Visual Analytics Systems

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

Overview of Systems

DecisionFlow [22] is a VA system that supports the analysis and exploration of temporal event sequences in high-dimensional datasets. It allows users to test different hypotheses regarding the factors that might affect the patient outcome and compare multiple complex patient event pathways by integrating on-demand statistical analysis techniques with interactive flow-based visualization. DecisionFlow helps users to specify a subsequence of interest with a milestone-based query interface. Then the matching data is aggregated to generate a DecisionFlow graph that contains a linear sequence of nodes (i.e., milestones) connected by directed edges. The system then analyzes the graph to extract multiple statistics (e.g., gender and age distributions and edge summary statistics). The system includes three main linked views-namely, the temporal flow view, edge overview view, and event statists view. The temporal flow view visualizes the DecisionFlow graph using a directed graph of nodes representing milestones where nodes are mapped to grey rectangles and are arranged in temporal order from left to right. The edges that connect these nodes are represented by two marks—namely, the time edges and the link edges, and they are color-coded to encode the average outcome. The edge overview panel summarizes the subsequence of interest that are returned from the query interface by showing multiple aggregate statistics. The event statistic view displays a color-coded bubble chart that represents different edge summary statistics.

RetainVIS [23] is a VA system that assists healthcare experts in the exploration of patient medical records in the context of risk prediction tasks. It provides users with the means to investigate common patterns in a patient's history to identify which medical codes or patient visits (i.e., sequence and timing) contribute to the prediction score. It can also help users to conduct different what-if analyses by testing hypothetical scenarios on patients (e.g., edit/add/remove medical code, alter visit intervals). Furthermore, RetainVIs allows users to provide feedback to the model based on their domain knowledge if the model acts in an undesirable manner. RetainVIS generates prediction scores based on the RetainEX technique, a bidirectional recurrent neural networks (RNN) model that harnesses the temporal information stored in patient records (e.g., time intervals between patient visits). It increases the interpretability and interactivity of models by calculating code-level and visit-level contribution scores.

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

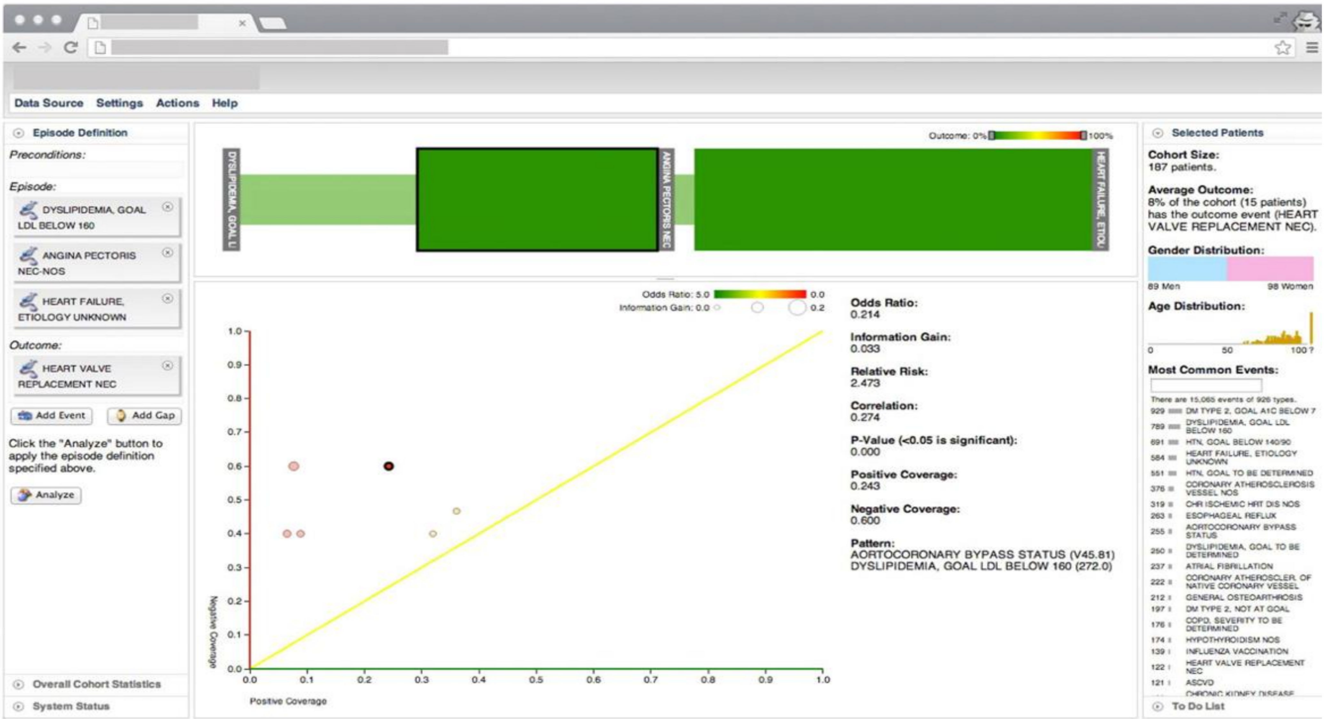
DPvis [24] is a VA system that supports clinical researchers in interactively discovering and exploring disease progression patterns and studying interactions between such patterns and patient's characteristics. It also allows users to test and refine hypotheses for multiple clinically relevant subgroup cohorts in an ad hoc manner. DPVis models disease progression pathways by characterizing a patient's clinical course as a sequence of transitions between multiple states where each state describes a co-occurring pattern of observed symptoms and variables. Then, it uses a class of unsupervised models, namely-continuous-time hidden Markov models (CT-HMMs), to discover these hidden states and state transitions from large-scale longitudinal patient records. These models identify associations between disease progression patterns and various observed variables and predict a patient's future states. DPVis combines the outcome of HMM models with interactive visualizations to assist medical experts in interpreting these models and clinically make sense of the discovered patterns.

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

The VA system for pharmacovigilance in electronic medical records developed by Ledieu et al. [25] integrates a modified version of the Smith–Waterman (SW) sequence alignment algorithm with an interactive web interface to detect inappropriate drug administration and inadequate treatment decisions in patient sequences. The SW algorithm is used to compare a reference sequence (i.e., a sequence specified by the user) and a patient's sequence, where each sequence is considered a string of characters. Each character in the sequence represents a clinical event, such as a laboratory test result or a drug administration. The algorithm calculates a similarity score for each comparison. A high similarity score corresponds to a higher similarity between the reference and the patient sequences. This VA system allows users to create the reference sequence(s) in a query interface. It provides them with a visual dictionary of event types (e.g., the discretized numerical events are encoded by color-coded squares or the direction of arrows represents the trend of change) in a grey rectangular area. To form a pattern, users can drag and drop these icons down to a query line. The system also enables the user to indicate time-constraint events in the query. The adopted SW algorithm returns the search result, which is displayed as a list of patients and their corresponding sequences, sorted based on their similarity score to the reference sequence. Each sequence is aligned to the reference pattern or its closest match. The time interval between the time-constraint event and the aligned events is shown by a vertical line along with the time duration in days on top of it.

Gotz et al. [26] develop a VA system to explore and query clinical event sequences stored in EHRs by combining on-demand analytics with visual queries and interactive visualizations (Figure 2). The visual query module provides an intuitive user interface that enables users to retrieve cohorts of patients that satisfy complex clinical episode specifications. Users can define a clinical episode by specifying milestones, time gaps, preconditions (i.e., a set of constraints that should be satisfied before the starting milestone), and outcome measures in the query interface. Upon submission of the query, the system returns a set of matching patient event sequences. The returned event sequence for each patient includes the specified milestones and several intermediate events that occur between milestones. Each episode is subdivided into a series of intermediate episodes at each milestone.

Frequent pattern mining (FPM) is then performed first on the overall episode as well as on each of the intermediate episodes that are retrieved by the visual query module. The FPM engine includes two main components-namely, the frequent pattern miner and the statistical pattern analyzer. The frequent pattern miner uses the bitmap-based Sequential PAttern Miner (SPAM) [27] algorithm for pattern discovery. SPAM employs a search strategy that combines a depth-first traversal of the search space with an efficient pruning mechanism. It takes a set of event sequences and a user-specified support as inputs and returns a set of frequent patterns as an output. Then the statistical pattern analyzer computes correlations (e.g., Pearson correlation, odds ratio, and information gain) between the mined patterns and the outcome measure. Finally, an interactive visualization allows users to explore the results and discover temporal patterns. The interactive visualization component is composed of three linked views. The cohort overview shows the age and gender distributions for patients that satisfy the query specifications. The milestone timeline represents the sequence of milestones using a series of ordered, vertical grey bars. The bars are connected by color-coded edges, where each edge has two parts-namely, the time edge, and the link edge. The time edge maps the mean duration between the milestones while the link edge connects the bars to show sequentially. The pattern diagram shows the set of patterns mined from the part of the episode that is selected in the milestone timeline in a scatter plot where the x and y axes encode the level of support for a specific pattern for patients with positive and negative outcomes, respectively.



Preconditions:

Episode:

DYSLIPIDEMIA, GOAL LDL BELOW 160

ANGINA PECTORIS NEC-NOS

HEART FAILURE, ETIOLOGY UNKNOWN

Outcome:

HEART VALVE REPLACEMENT NEC

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

Odds Ratio:
0.214

Information Gain:
0.033

Relative Risk:
2.473

Correlation:
0.274

P-Value (<0.05 is significant):
0.000

Positive Coverage:
0.243

Negative Coverage:
0.600

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

Cohort Size:
187 patients.

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

Gender Distribution:

Age Distribution:

Most Common Events:

There are 15,065 events of 926 types.

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

Figure 2. The screenshot of the VA system developed by Gotz et al. [26] including, the visual query panel, the milestone timeline, the cohort overview, and the pattern diagram. Source: Reprinted with permission from ref. [26], Copyright (2014), with permission from Elsevier.

The VA system developed by Simpao et al. [28] facilitates the dynamic and continuous monitoring of medication alerts and care providers' responses through an automated, user-friendly dashboard. It allows pharmacists and care providers to examine and filter the alert data based on patient location and ordering provider type and to identify which specific orders triggered the drug-drug interaction alerts. This VA dashboard is an integral part of a hospital quality improvement initiative to improve medication safety and reduce alert fatigue by deactivating irrelevant alert rules. The system is developed in collaboration with a clinical decision support committee that is asked to perform three interventions to deactivate irrelevant drug-drug interaction alert rules. The impact of these interventions on pharmacists' alerts and override rates is analyzed using an interrupted time-series framework with piecewise regression. Baseline IQRs and median rates are compared to IQRs and median rates following three intervention phases of drug-drug interaction deactivations and are tested for statistical significance using the Wilcoxon rank-sum test. The user interface of this system includes a central display area with graphical and tabular data representations. Medication alert and override rates, different alert types, and various care providers, and patient characteristics are displayed and explored at a specific time point or across a user-defined time interval using multiple filters and limits.

The MOSAIC dashboard system [29] aims to support the prediction and diagnosis of type 2 diabetes mellitus (T2DM) by analyzing clinical and home monitoring data. The system integrates a data querying and mining technique with an interactive user interface to assist caregivers in devising management strategies and therapeutic interventions for T2DM complications. The mining techniques are triggered by the query module that is responsible for retrieving the data from the i2b2 data warehouse, calling the proper data mining technique, and sending the results back to the user interface. The data mining module implements several temporal analytics models such as temporal abstractions, the care flow mining algorithm, drug exposure pattern detection, and risk prediction models for T2DM complications. Temporal abstractions are extracted using the Time Series Abstractor (JTSA) tool that provides a library of techniques that can be employed for time-series processing and abstraction [30]. The care flow mining technique uses the temporal sequence of events to determine the most frequent clinical pathways patients experience during their care process, automatically generating groups of patients with similar care histories [31]. The proportion of days covered is used to summarize the dug purchase patterns using the data gathered from administrative resources. Finally, several risk prediction models are generated to estimate the risk of T2DM complications [29].

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

VisualDecisionLinc [32] is a VA system that helps clinicians to identify subpopulations of patients with similar clinical characteristics and to understand the risks and effectiveness of different treatment options for these subpopulations using psychiatric patients' data with major depressive disorder (MDD). The system aims to improve and simplify the decision-making process by reducing the number of available therapeutic options to those that have proven to be most effective with minimal side-effects. To define the MDD comparative population, VisualDecisionLinc uses a patient data-driven approach where the patient's medical profile is used as 'seed' data (i.e., patients with a primary diagnosis of MDD and their last prescribed medications) to identify a comparable group of patients with similar clinical characteristics. At the computational level, the system creates a bin for each medication and inserts patients into bins of their prescribed medication. At the same time, the system tags patients based on their treatment outcome response, which is reported in the database in the form of a clinical global impression (CGI) score. CGI score is a seven-point scale that offers a brief score of the clinician's assessment of the severity of the patient's illness prior to and after starting treatment. A lower CGI score indicates a better treatment outcome for the patient. After the binning process is done, the system uses additional computational processes to quantify the collective comparative MDD patient response into a '% Patient Improved' score.

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

Care Pathway Explorer [33] is an interactive hierarchical information exploration system that can help physicians analyze patients' longitudinal medical records. The system provides an overview of the frequent patterns that are mined from patient event sequences. The physician then studies these patterns and interactively selects patterns of interest for more details. The system computes the group of patients that match the physician's specified sub-traces. Then the event traces for those patients are extracted using a deeper level of the user-specified hierarchy. The system feeds these traces to the frequent pattern miner engine, which mines frequent patterns and analyzes how these patterns are associated with outcomes using a modified version of the SPAM algorithm [27]. The patterns are then visualized alongside meaningful statistics.

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

RegressionExplorer [34] is an interactive VA system that enables clinical researchers to quickly generate, compare, and evaluate many regression models. It also helps to formulate new hypotheses and steer the development of models by allowing the user to compare candidate models across several subpopulations. Upon loading the dataset and selecting the appropriate responder that captures the condition of interest, the system allows the researcher to analyze the one-to-one relationships between each covariate and the responder by performing a univariate analysis. The results are displayed as colored rectangles next to the variable names in the univariate analysis view. The significance level of an effect is determined using p-value, where a lower p-value results in a higher level of significance and a more saturated color. Red represents a positive effect, while blue represents a negative effect. Next, the system allows the user to perform stepwise multivariate analysis by dragging variables from the list of variables to the variable selection view. After each selection, the system generates a new model displayed as a single row of the multivariate model matrix. Columns in the matrix show the levels of significance for the included covariates following the same convention as for the univariate view. The system also displays histograms, along with some basic descriptive statistics for all the covariate distributions to provide basic checks and interpretation during analysis.

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

The VA system developed by Mica et al. [35] helps guide patient assessment and therapeutic decisions for physicians using severely injured patients' clinical data in a trauma center (Figure 3). The system allows the user to filter cohorts of patients based on multiple parameters, including age, body temperature, injury severity score (ISS), multiple lab results, and abbreviated injury scale (AIS) score. With every change of the filtering criteria, a query is sent to the server to extract a group of patients that satisfy the query specifications using several algorithms such as statistical frequency grouping, time interval simplification, and consecutive event merging. The system enables the user to explore the results using a variation of the Sankey diagram. Each node in the graph encodes a medical state (e.g., treatment or outcome), and each link encodes transitions between consecutive states in the cohort of interest. The height of nodes and links represents the relative number of patients that share the state and transition, respectively. The color encodes the ratio of patients that develop the outcome of interest. Statistically, to justify the distribution of patients based on clinical scores, the system integrates binary logistic retrogression along with receiver operating characteristic (ROC).

Visual Temporal Analysis Laboratory (ViTA-Lab) [36] integrates temporal data mining techniques with query-driven interactive visualizations to support a knowledge-based exploration of time-oriented clinical data and the discovery of interesting patterns within it. ViTA-Lab is composed of three main interfaces. The main visualization interface provides an overview of the longitudinal concepts and the distribution of derived temporal abstractions (TA) for individual and multiple patients at different temporal granularities. It provides the user with a knowledge-based browser and a graphical widget for selecting an individual patient or a group of patients. It uses a scatter diagram over time and a modified version of the bar chart visualization technique to show the distribution of TAs and help the user discover trends in these distributions.

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

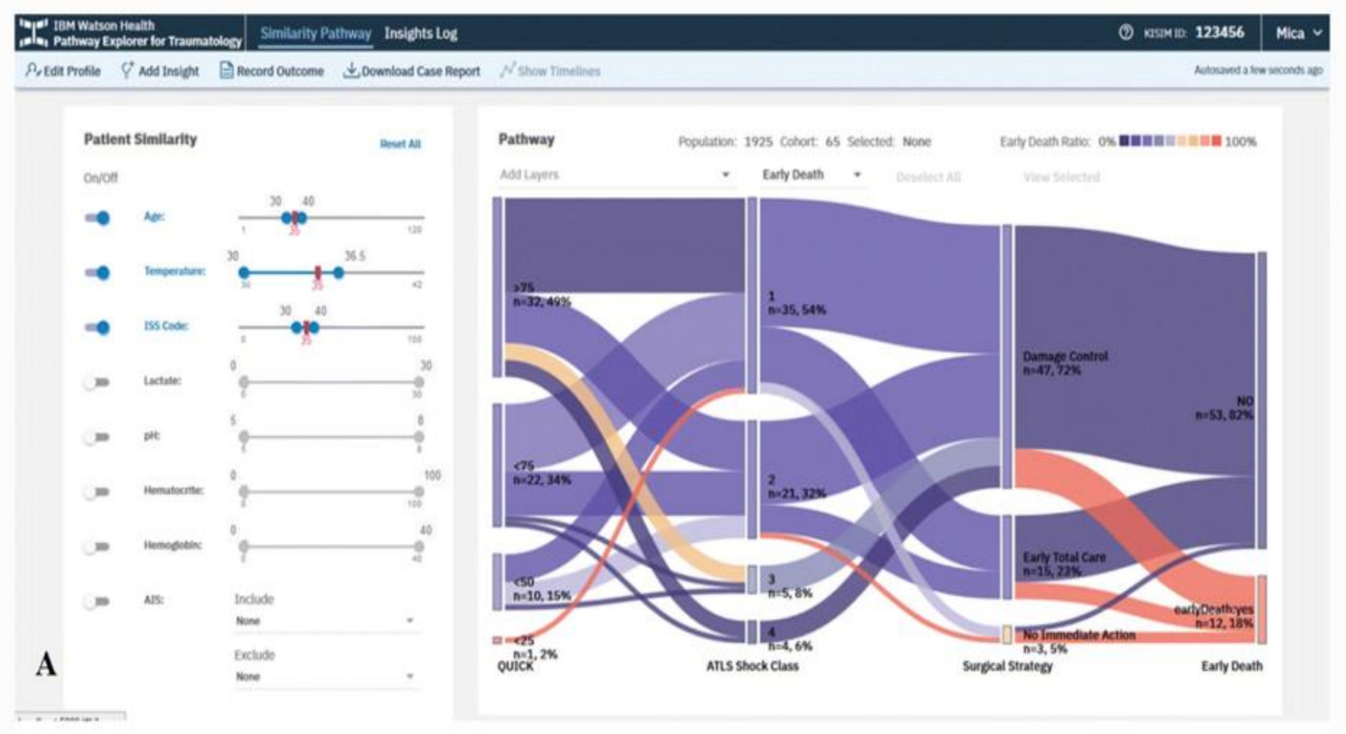
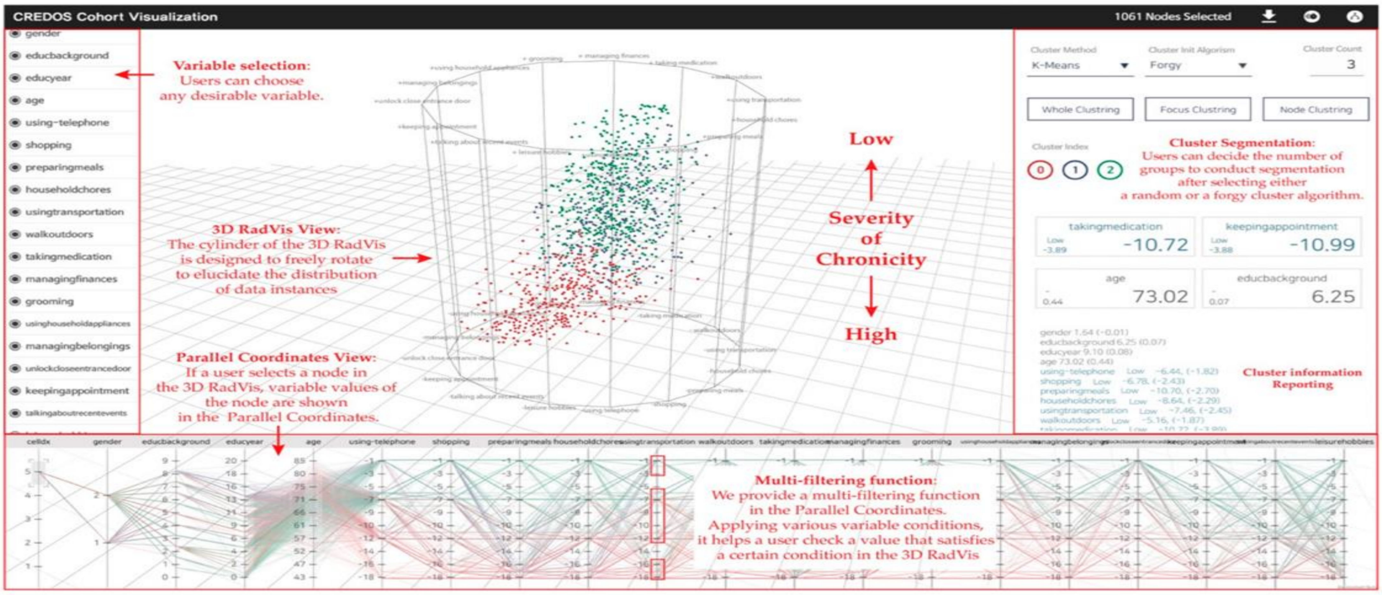


Figure 3. The screenshot of the VA system developed by Mica et al. [35] shows the pathway of the early death outcome of a hypothetical patient with an age of 35 years, an ISS of 35, and a temperature at admission of 35 °C using a Sankey diagram. Source: Reprinted with permission from ref. [35], Copyright (2020).

The pattern explorer supports the exploration of temporal patterns that are discovered by data-driven computational techniques. It works based on a version of the KarmaLego algorithm, which is used for the discovery of frequent temporal patterns [37,38]. Components of the output's temporal pattern (a pair of concept and value) are represented by horizontal lines that are ordered according to each component's start time, maintaining, in a proportional fashion, the mean duration of each component and of the time gaps among components. The color of the same type of component in all patterns stays the same. The pattern explorer allows the user to recognize the meaning of a temporal pattern, that is, which components make up the pattern, and what temporal associations such as overlaps, before, or after hold between them.

RadVis [39] is a VA system that supports psychiatrists in analyzing and exploring multidimensional medical datasets for patients who have dementia (Figure 4). It allows the user to get a better understanding of the characteristics of patient clusters and analyze the variable values of data comprising each cluster at the same time. The system enables the user to select variables of interest from "Variable Selection Menu" and select "Cluster Segmentation Menu" to segment clusters of patients based on their traits. The user can choose the number of clusters for segmentation after selecting either a forgy cluster or a random cluster algorithm. Following either of the clustering algorithms, the cluster's central value is calculated based on the number of clusters. After the Euclidean distance between the central value and each node is calculated, multiple nodes are included to obtain clusters of similar value. This process is repeated until the central value stays constant.



Variable selection: Users can choose any desirable variable.

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

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

Cluster Method: K-Means
Cluster Init Algorism: Forgy
Cluster Count: 3

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

takingmedication Low -3.89 -10.72
keepingappointment Low -3.88 -10.99
age 0.44 73.02
educbackground 0.07 6.25

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

Cluster information Reporting

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

Figure 4. The screenshot of RadVis [39] combing 3D RadVis and parallel coordinates. Source: image used under CC-BY 4.0 License.

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

The predictive VA system developed by Sun et al. [40] aims to predict the risk and timing of deterioration in hypertension control using EHRs. The system is composed of three main modules. The feature engineering module converts clinical data into a feature matrix and a target label vector that can be used to build the predictive model. The target label is derived based on the physician's assessment of blood pressure control status as in-control (i.e., positive) versus out-of-control (i.e., negative). The positive and negative transition points (from an episode of positive (negative) assessment points into negative (positive) points) are considered as target labels for the prediction model. Next, to turn event sequences into feature variables, the system specifies an observation window for each feature concept (e.g., diagnosis concept). It then aggregates all the events of the same feature concept within the observation window into a single value. The system then applies a two-level feature selection process. In the first level, within the same concept, features are chosen based on the information gain. Then a greedy forward selection algorithm is used to choose which concepts to keep. In the next step, the system starts iteratively combining features from different concepts until the combination fails to improve the performance of the prediction. Finally, various techniques, such as naive Bayes, logistic regression, and random forests, are used to generate transition point models. The system allows the user to explore the prediction results and other events through interactive visualization. An individual patient's timeline is represented by a line, and each hypertension control assessment event is represented by a circle. Red and blue circles represent in-control and out-of-control blood pressure episodes, respectively.

The VA system developed by Guo et al. [41] helps clinicians to explore medical records from both multivariate and temporal perspectives and identify and analyze similar records (Figure 5). The system integrates an unsupervised learning-based technique with interactive linked views to support physicians in several tasks such as finding similar records based on a focal patient record, comparing patients' medical feature values at a specific time point, or identifying (dis)similar time stamps among similar records. The system provides two overviews of all patients: One is for patients' similarities according to the combination of tests taken during the collected time period, and the other view shows patient's similarities according to the test values. To create the first overview, the system applies the Jaccard index [42] to compute the similarity. Then it extracts clusters of similar patients by combing a dimensionality reduction (DR) technique (i.e., t-SNE) and a density-based clustering method (i.e., HDBSCAN). For the second overview, the system first calculates the similarity of each pair of the test records and then similar to the other overview; it applies t-SNE to visualize the similarity relationships. To visualize the clustering information, each point (i.e., each patient's record) is colored based on the assigned cluster-ID. The system allows the user to select a patient of interest from these overviews. It then automatically searches for the top-3 similar patients based on the pre-computed similarities. The system uses autoencoder-based event embedding [43] and sequence to sequence learning (seq2seq) [44] technique to handle various event types and convert records with different lengths to vectors of the same length. Then, it computes the similarity of each pair of patients using a certain distance metric, such as the Euclidean distance. The system provides multiple line charts to show changes of dissimilarities of test records over time between the patient of interest and top-3 similar patients and to visualize a statistical overview of the focal and top-3 similar patients.

SubVIS [45] is a VA system to support medical experts in interpreting high-dimensional clinical data and exploring subspace clusters from different perspectives (Figure 6). It enables the user to analyze each subspace independent of its association to a certain clustering technique. It allows the use of every subspace clustering technique available at OpenSubspace Framework [46]. SubVis allows a three-level exploration of data and clusters through its interface. The first level provides the user with a general overview of all the detected subspace clusters, their properties, and the distribution of dimensions within each subspace cluster using interactive bar charts. A matrix-based heatmap is also available to give more details on the association between the pair-wise distance. The second exploration level allows the user to choose a subset of relevant clusters in the multidimensional scaling (MDS) [47] plot to get an aggregated overview of the cluster members in an aggregation table. The distance between various clusters in the MDS plot shows their pair-wise similarity. SubVis contains various similarity measures, such as Overlapping, Jaccard Index, and Dice Coefficient. The system enables the user to inspect the distribution of the cluster members in every dimension for each cluster. In the last exploration level, a table-lens-like view [48] supports the exploration of the actual data records and provides interactive coloring and sorting of the record and its dimension.

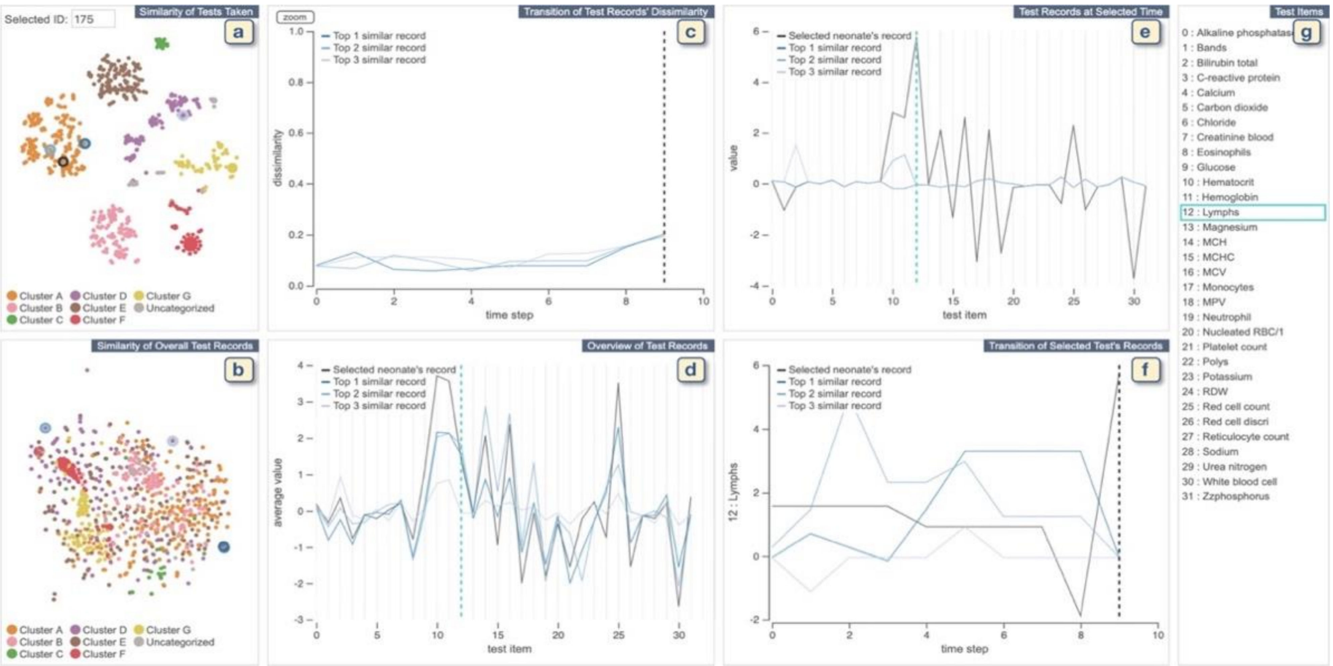


Figure 5. The screenshot of the VA system developed by Guo et al. [41]. (**a**,**b**) display each neonate's similarities of tests taken and the records of test values, respectively. (**c**) shows changes of dissimilarities of test records over time between the neonate chosen from (**a**,**b**) and top-3 similar neonates. (**d**) displays a statistical overview of the chosen neonate and top-3 similar neonates. (**e**) provides all the test results at the selected time in (**c**) or (**f**). (**f**) displays the temporal changes of values of the chosen test in (**d**) or (**e**). (**g**) lists all medical test names. Source: Reprinted with permission from ref. [41], Copyright (2020), with permission from Elsevier.

The VA system developed by Huang et al. [49] supports the interactive exploration of patient trajectories to assist physicians and clinical researchers in identifying chronic diseases and determining how a group of patients with chronic diseases might go on to develop other comorbidities over time (Figure 7). The system first aligns patient trajectories based on the time they are diagnosed with a specific chronic disease. Then once the user specifies the time windows, the patient trajectories are divided based on their timestamps, and patients within the same time window are aggregated into one. The system then clusters the patient records at each time window based on a similarity measure and creates a set of cohorts. The system supports frequency-based cohort clustering and hierarchical cohort clustering techniques. A cohort of patient trajectory network is built based on the clustering result where each node represents a cohort at a time window, and each edge shows the relationship between two cohorts at consecutive time windows where their members overlap. The system allows the user to filter edges using the variance-based association filtering technique by adjusting the entropy threshold. When the threshold is zero, only associations between fully overlapped cohorts are shown; in the case when the threshold is high, all associations are visualized. A Sankey-like timeline then visualizes the output results. The nodes are color-coded based on the unique comorbidities, and the color of the edges is determined by the two nodes it connects (i.e., a gradient for smooth transitions). Each cohort has a label that shows its dominant features. In addition, the cardinality of both nodes and edges are represented by their height.

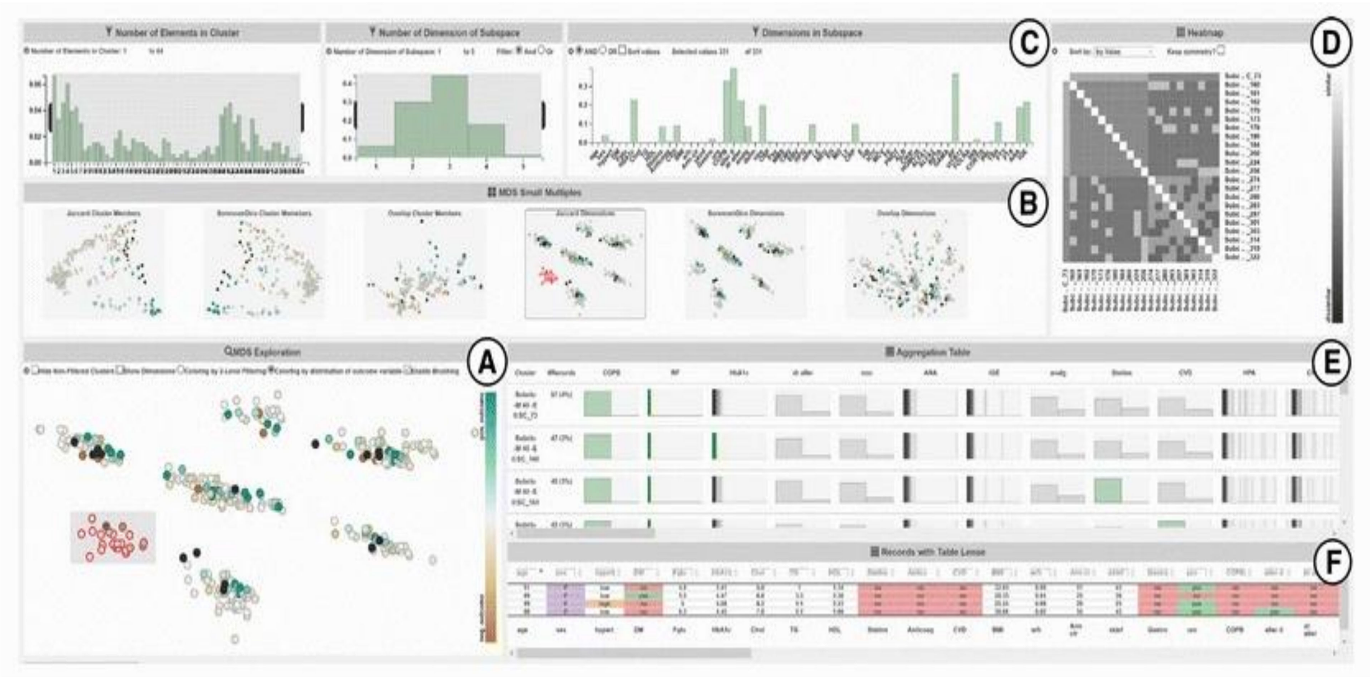
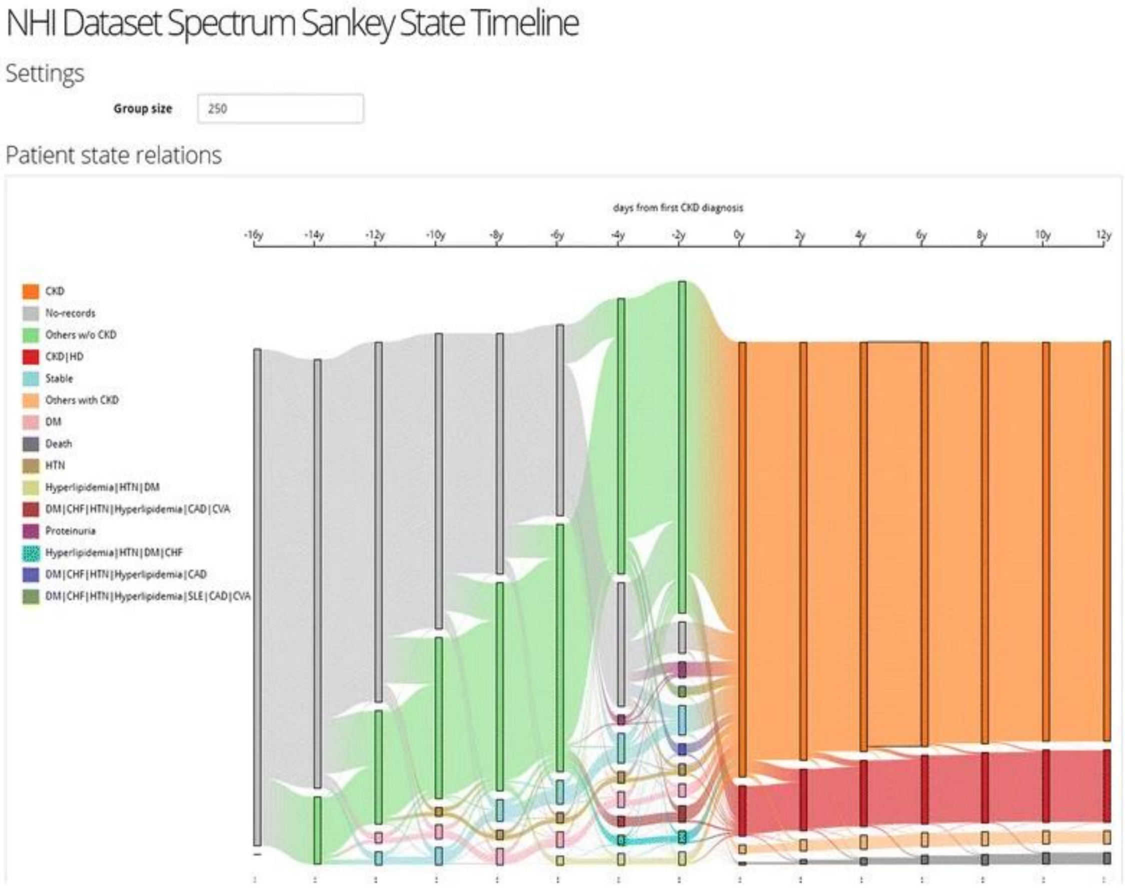


Figure 6. A screenshot of SubVIS [45] including (**A**) MDS projection plot, (**B**) MDS small multiples, (**C**) barcharts showing the distribution properties of the subspaces, (**D**) heatmap, (**E**) aggregation table, and (**F**) table lens. Source: image used under CC-BY License.



NHI Dataset Spectrum Sankey State Timeline

Settings

Group size 250

Patient state relations

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

Figure 7. The screenshot of the VA system developed by Huang et al. [49] shows the result of frequency-based cohort clustering using a Sankey diagram. Source: image used under CC-BY 4.0 License.

CarePre [50] is a clinical decision assistance system that supports the exploration and interpretation of deep learning prediction models that are developed to predict future diagnosis events for a focal patient based on their medical background. It assists physicians in making more informed decisions by letting them analyze contributing factors in prediction results and explore the outcomes of possible treatments through interactive visualizations. CarePre allows the physician to input potential diagnoses (based on the patient's symptoms and tests) for a focal patient into the system. The system then automatically estimates the risk of future diseases for the patient based on their medical history using a state-of-the-art deep learning technique and allows the physician to explore the results and the details of the historical medical records in the prediction view. The prediction view shows the patient's event sequence leading up to the time point of prediction, which is represented by rectangular nodes arranged horizontally in order of their occurrence. The predicted likelihood of each diagnosis is also displayed as a series of rectangular nodes where the color saturation for each node shows the prevalence of the predicted diagnosis across the records for a population of similar patients.

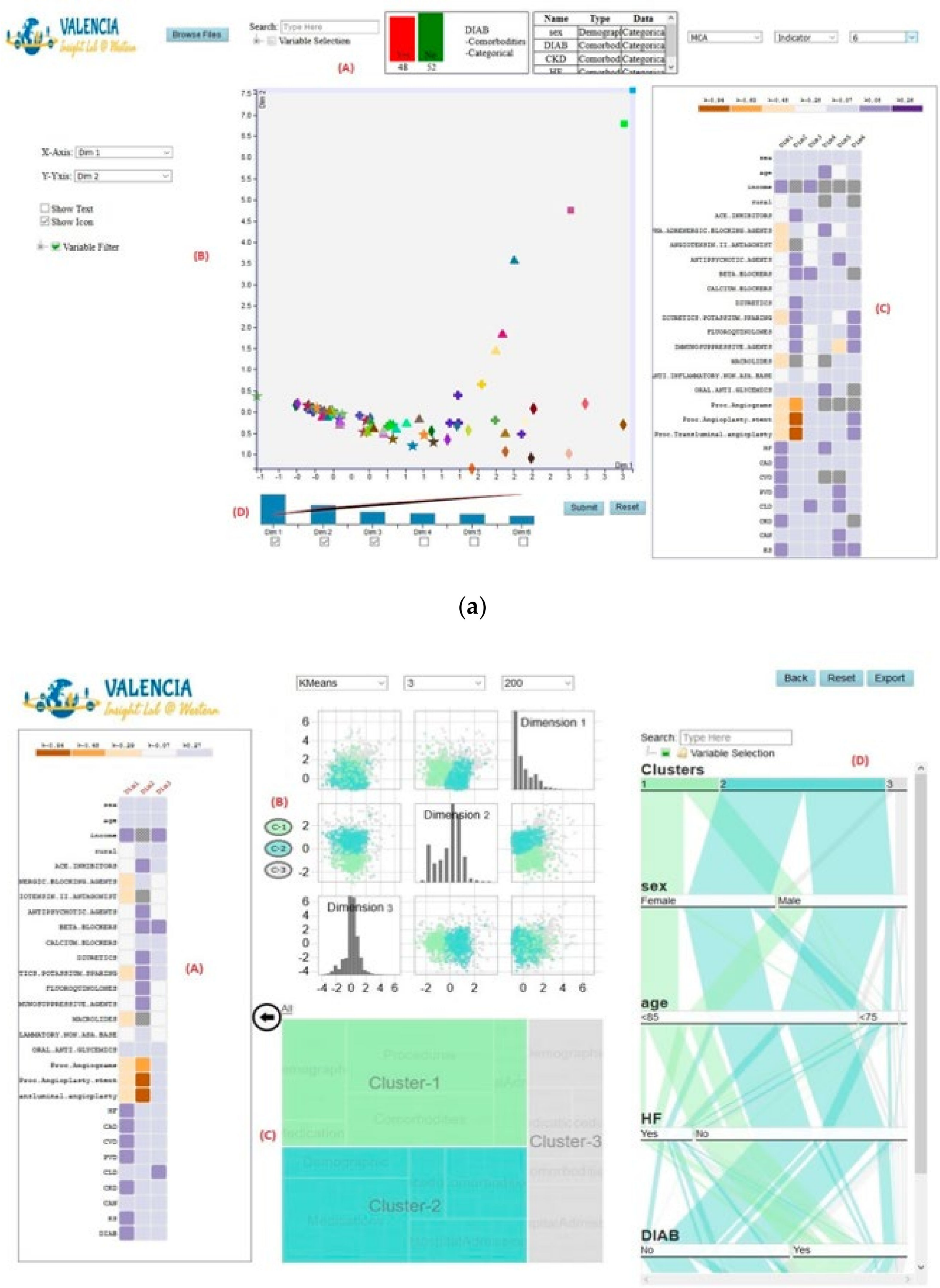
In the next step, the physician can specify a query to retrieve a group of similar patients to help interpret the prediction results. CarePre measures similarity between sequences by computing the similarity between each pair of events using the Euclidean distance of the corresponding event vectors.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

Peekquence [51] is a VA system that aims to make the frequent sequence mining results more interpretable by allowing the user to explore the patterns by ranking them based on their variability or correlation to the outcome. It can also integrate patterns with a patient timeline to help the user understand where the patterns occur in the actual data. Peekquence uses the SPAM [27] frequent sequence mining algorithm to detect the most frequent sequences. The system uses four linked views to visualize the result of SPAM on the patient's medical records. All the views use an event glyph to visualize the event sequences. The event glyph represents each unique event type appearing in the mined patterns by a circle and is color-coded based on a categorical ontology. These event glyphs are labeled with an abbreviation of the name of the event type. The sequence network view displays the frequency of co-occurring events within patterns that are mined using SPAM. The event types are represented by the nodes, and the two co-occurring nodes within patterns are connected by an edge. The pattern list view displays all the mined patterns, aligned vertically. Each row represents a pattern that is visualized as a sequence of circular event glyphs. Furthermore, the association of the patterns with the outcome is represented by the stacked bar chart next to the sequence. The event co-occurrence histogram view shows the frequency of co-occurring events with a selected pattern from the pattern list view. Each event type is represented by a bar partitioned into three blocks to show events occurring before, within, and after the chosen pattern. Lastly, the timeline view displays the patient's event sequences aligned according to the selected pattern.

PHENOTREE [52] is a hierarchical and interactive phenotyping VA system that allows physicians to participate in the phenotyping process of large-scale patient records. It enables the user to explore patient cohorts, and to create, interpret, and evaluate phenotypes by generating and navigating a phenotype hierarchy. The system uses the sparse principal component analysis (SPCA) to identify key clinical features that describe the population given a cohort or sub-cohorts of patients. These key clinical features are used to build deeper phenotypes at finer granularities by expanding the phenotype hierarchy. Patients that are associated with each key feature are grouped into individual sub-cohorts. The system then iteratively applies the SPCS to each sub-cohort of patients created in the previous step. PHENOTREE assists physicians in identifying groups of phenotypes and their corresponding patient sub-cohorts at different granularities through this process. The system utilizes the radial Reingold-Tilford tree to visualize the results. Each node in the tree represents a structured phenotype and a sub-cohort characterized by this phenotype.

VALENCIA [53] is a VA system that aims to address the challenges of high-dimensional EHRs by integrating several dimensionality reduction (DR) and cluster analysis (CA) techniques with real-time analytics and interactive visualizations (Figure 8). VALENCIA's analytics engine has two components—namely, DR and CA engines. The DR engine incorporates several DR techniques to transform EHRs from the high-dimensional space to one with lower dimensions. The CA engine then uses several clustering techniques to classify the data points in this low-dimensional space into meaningful groups with similar characteristics. VALENCIA allows the user to choose the most appropriate combination of DR and CA techniques and explore the results through two main views—namely, DR and CA views. The DR view has four subviews, including raw-data, projected-features, association, and variance subviews. These subviews allow the user to choose their features of interest, select the DR technique, adjust the configuration parameters, investigate how features are associated with transformed dimensions, and choose dimensions to be included in the CA engine. The CA view has three subviews—namely, hierarchical subview, frequency subview, and projected-observation subview. These subviews allow the user to examine the hierarchical structure of the CA results, choose the CA technique and configuration parameters, and observe the distribution of features in each subset of the data.

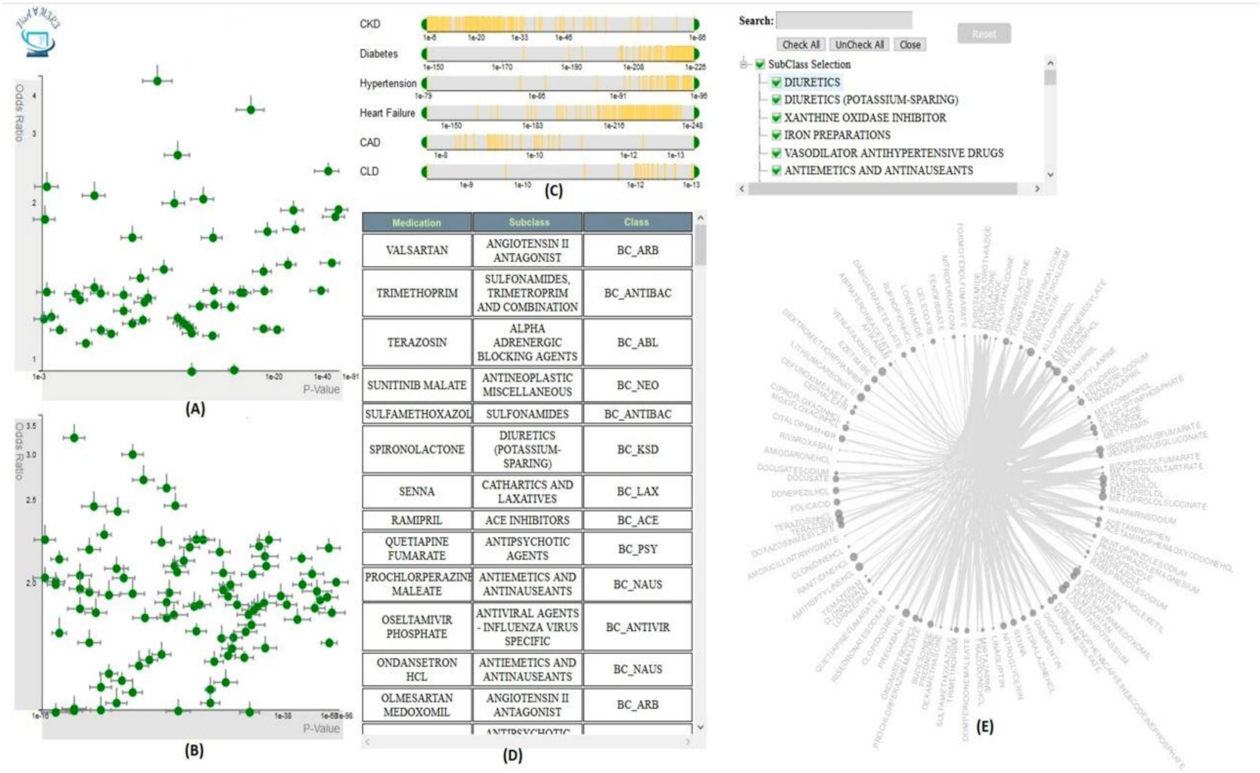
VISA_M3R3 [54] is a VA system that allows clinical researchers to identify medications or medication combinations that are associated with a higher risk of acute kidney injury (AKI) (Figure 9). The system incorporates regression, frequent itemset mining, and interactive visualization to help the user explore the relationship between medications and AKI. The analytics module of Visa_M3R3 is composed of two components. The first component is the single-medication analyzer that focuses on finding associations between individual medications and AKI using multivariate regression. The multiple-medications analyzer aims to identify associations between medication combinations and AKI using frequent itemset mining and regression. All models are validated through Bonferroni correction and represented in multiple interactive views. The regression models generated from single-medication and multiple-medications analyzers are represented in two scatter plots in the single-medication and multiple-medication views. The output of the frequent itemset mining is shown using a chord diagram in the frequent-itemset view. The user can filter and control the information presented in other views using sliders in the covariates view. Finally, the medication-hierarchy view displays additional information regarding data elements using a data table.



(**a**)

(**b**)

Figure 8. The screenshot of VALENCIA [53] showing (**a**) the DR view and (**b**) the CA view. Source: image by authors.



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

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

Figure 9. The screenshot of VISA_M3R3 [54] showing (**A**) the single-medication view, (**B**) the multiple-medication view, (**C**) the covariates view, (**D**) the data table, and (**E**) the frequent-itemset view. Source: Image by authors.

4. Design Space

In this section, we introduce the primary dimensions of the design space of EHR-based VA systems and highlight the key elements in each dimension that are frequently used for designing and developing these systems. For a VA system to work well, there must be harmonious functioning among all of its components. Such components include VA tasks, analytics and data models, and visual representations. One way in which we can investigate the strength of the coupling among components of VA systems is through the lens of interaction. Therefore, the four key dimensions that are used to evaluate the existing systems include VA tasks, analytics, visualizations, and interactions (for comparison see [13,55]).

4.1. VA Tasks

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

Understanding the Progression of Diseases: VA techniques can be used to model and visualize a patient's medical condition over time, which is known as the patient's medical trajectory. Research studies have shown that different patient trajectories can have different associated risks for the same outcome [56,57]. For instance, a patient may die due to cardiovascular complications, kidney complications, or peripheral complications. Although the outcome is the same, disease progression paths that lead to the outcome are typically different. Thus, studying such various trajectories can result in the development of more tailored treatment plans, the discovery of biomarkers, and the development of different risk estimation indices. Comorbidity analysis, which is the process of analyzing and exploring associations among diseases, is another key factor in improving the quality of care, especially for older patients who suffer from multiple diseases. Therefore, understanding the incidence, prevalence, and coincidence of diseases is the foundation for making important policy decisions. Thus, there have been many EHR-based VA systems developed to accomplish this task. For instance, Huang et al. [49] have developed a system that supports the exploration of patient trajectories to help clinical researchers detect chronic diseases and determine how a set of patients with multiple chronic diseases might go on to develop other comorbidities over time. DPvis [24] supports the interactive exploration of disease progression patterns and the discovery of interactions between such patterns and patient's characteristics. Similarly, DecisionFlow [22] and Peekquence [51] allow comparison of multiple complex patient event pathways by combining statistical analysis processes with interactive flow-based visualizations.

Discovering and Exploring Cohorts of Interest: The identification of a cohort (group) of patients who meet predefined criteria from a large patient population has various use cases, including survival analysis, clinical trial recruitment, and other retrospective studies [58,59]. Cohort identification forms a platform for future clinical research studies in areas such as predicting complications, pharmacovigilance, and detecting adverse events. Traditionally, this process is carried out through chart reviews by primary care staff and research staff in individual practices to query the clinical systems for patients matching a specific set of criteria. However, manual cohort identification can be extremely challenging and time-consuming, depending on the complexity of the criteria. This is because the patient data satisfying these criteria is buried within large volumes of data stored in EHRs. Thus, there is a need for electronic phenotyping algorithms to replace the manual chart reviews for cohort identification.

A phenotype can be defined as a specification of an observable state of an organism. It can be applied to patient characteristics that are inferred from EHRs, such as clinical conditions, blood type, or physical traits. Phenotype algorithms that characterize or identify phenotypes can be used for the direct identification of cohorts based on clinical or medical characteristics, risk factors, and complications, thereby allowing clinical researchers to improve patient outcomes. These algorithms can be generated using various forms of machine learning techniques. However, an integrated approach that combines these analysis techniques with visualization is more likely to facilitate the process of creating and comparing different patient cohorts, determining risk factors associated with a particular disease, and discovering hidden structures in the patient data. As a result, several VA systems have been developed recently to address this issue. For instance, Mane et al. [32] developed VisualDecisionLinc to help clinical researchers identify subpopulations of patients with similar clinical characteristics to help them evaluate the risks and effectiveness of different treatment options. Similarly, PHENOTREE [52] is another VA system that allows clinical researchers to explore patient cohorts and create and evaluate phenotypes by generating a phenotype hierarchy.

Learning and Understanding Prediction Models: The focus on creating prediction models is increasing in many areas of clinical research. These models aim to assist physicians in personalized decision making with regards to diagnosis, prognosis, and treatment. Examples of successful risk prediction models are the Apache system that estimates the risk of hospital mortality, the Framingham heart score that predicts cardiovascular mortality, and the Nottingham Prognostic Index that allocates patients with breast cancer to different risk groups [60–65]. Despite the strong performance of these models, it is often challenging for physicians to understand how the prediction models arrive at an estimated risk. The black-box nature of most of these models can impede their wide adoption in clinical practice since there is little tolerance for errors in medical decision making.

Thus, providing interpretability and transparency in prediction models is critical in validating the resulting predictions. To address these needs, VA systems provide clinical researchers with accurate, fast, and trustworthy interpretation of prediction models by integrating effective visual representations with machine learning techniques [66–68]. For instance, RetainVIS [23] combines interpretable and interactive RNN-based models and interactive visualization to allow exploration of patient records in the context of prediction tasks.

Discovering adverse events: Adverse events can be defined as the harmful effects of medical care on a patient's medical condition. They are caused by medical management rather than the patient's underlying condition [69]. For instance, an infection developed during the treatment of a different condition is considered an adverse event. Adverse events are responsible for 2.9–16.6% of all acute hospitalizations and studies have shown that 30–58% of all these events are preventable [70–74]. Adverse events can also often be linked to drugs. Adverse drug events cause 3.5 million physician visits, 125,000 hospitalizations, and 98,000 drug-related deaths each year [75]. Even though drugs are tested for any potential adverse events and are cleared for marketing to the medical community, unsuspected adverse events are occasionally detected. This is due to the fact that clinical trials are usually limited to short time periods and include only a small test cohort. In addition, the frequency of these adverse events may be so low that they are hard to detect in clinical trials. Another issue in detecting adverse drug events is confounding by indication. For instance, insulin is prescribed for diabetes. Myocardial infarction is a common comorbid disease for patients who have diabetes and thus, detecting the adverse event myocardial infarction for insulin is a false positive ("a confounding effect"). There are several approaches to detect significant adverse drug events using automatic analysis techniques; however, most of these approaches overlook low-frequency events. Furthermore, the domain knowledge regarding the confounding effect should be included in these automatic analysis techniques. VA systems can address these issues by involving domain experts in the analysis process. For instance, Ledieu et al. [25] developed a VA system for pharmacovigilance in electronic medical records to detect inappropriate drug administration and inadequate treatment decisions in patient sequences. VISA_M3R3 [54] is another EHR-based VA system that helps healthcare providers to identify medications that are associated with a higher risk of acute kidney injury.

4.2. Analytics

There have been several analytics methods used for visual analysis of EHR data. These methods include (1) classification, (2) clustering, (3) Pattern discovery, (4) regression, (5) inference, and (6) dimensionality reduction.

Classification: Classification is used to classify data points into predefined categorical class labels. "Class" is the feature in a dataset in which users are most interested. In statistics, it can be defined as the dependent variable. In order to classy data points, a classification technique generates a model, including classification rules. Classification is a two-step process, including training and testing. In the training step, a classification model is built by analyzing training data that contains class labels. The accuracy of the classification model depends on the degree to which classifying rules are correct. In the testing step, the classifier's (i.e., classification model) ability to classify unknown data points for prediction is examined. Some of the most common classification techniques that are used in the analysis of EHRs are support vector machine [76,77], decision tree [78], naïve Bayes [79], and neural network [80]. For instance, RetainVIS [23] uses the RetainEX technique (i.e., a bidirectional recurrent neural networks (RNN) model) to generate prediction scores based on the temporal information stored in EHRs and help the user identify which medical codes or patient visits contribute to the prediction score.

Clustering: Clustering is an unsupervised learning technique that occurs by analyzing only independent variables. In other words, unlike classification techniques, clustering techniques do not use "class". Thus, clustering is best used for exploratory studies, especially if those studies include large volumes of data, but very little is known about the data. Clustering groups the data points into a certain number of clusters so that points within a cluster have high similarity and points from different clusters have a low similarity. The similarities between the data points are measured using their feature values. Some of the most commonly used clustering techniques in exploring EHRs are k-means [81,82], hierarchical clustering [83], model-based clustering [84], and density-based clustering [85]. For instance, the VA system developed by Guo et al. [41] uses HDBSCAN which is a density-based clustering technique to cluster similar patients according to the combination of tests taken during a specific time period.

Pattern discovery: Pattern discovery aims to identify statistically significant associations and frequently occurring patterns in the data. In the analysis of EHR data, pattern discovery can be further classified into frequent pattern mining and association rule mining techniques [86]. The purpose of frequent pattern mining is to identify the inherent regularities in the EHR data. In other words, these techniques can be used to find common subsequences in the clinical event sequence dataset. Frequent pattern mining can be further extended to other problems such as sequential pattern mining and time-series mining that are very common when dealing with clinical event sequences. Association rules can be considered as a second-stage output of frequent pattern mining. Association rule mining is often used to discover relationships among data items. Association rule mining techniques can be employed to identify underlying relationships among health conditions, symptoms, and diseases in the healthcare field. For instance, the VA system developed by Gotz et al. [26] helps the user to explore the clinical event sequences using a Frequent Pattern Mining (FPM) engine. The FPM engine has two main components, including the Frequent Pattern Miner and the Statistical Pattern Analyzer. The frequent pattern miner uses the SPAM [27] algorithm for pattern discovery. Then the Statistical Pattern Analyzer computes correlations between the mined patterns and the outcome measure.

Regression: Regression techniques are often used to identify associations between features, such as the extent to which feature A affects feature B. Logistic regression is a special type of regression that is commonly used in the analysis of clinical data [87]. It draws a separating line among classes using the training data; then, it applies the line to classify the test data's unknown data points. Logistic regression is often used to analyze the relationship between a dependent feature (e.g., patient outcomes) and one or more independent features (e.g., patient comorbidities, symptoms, and laboratory test results). For instance, RegressionExplorer [34] allows the user to formulate a new hypothesis and steer the development of models by creating, comparing, and evaluating multiple regression models.

Inference: Inference refers to the process of reaching conclusions based on the evidence found in the existing data. However, conclusions drawn from inference are only justifiable under some specific conditions and can be false when applied to unobserved data. One of the inference techniques used in the analysis of the clinical event sequences is graphical models. Graphical models show the conditional dependence between clinical events using an event correlation graph, such as the Markov chain [88] and Bayesian Networks [89]. For instance, DPvis [24] uses continuous-time hidden Markov models to learn how various diseases go through different states, discover biomarkers (i.e., observed variables) that can characterize the disease progression, and to use these biomarkers to identify diseases earlier in patients.

Dimensionality Reduction: Dimensionality reduction is the process of transforming a high-dimensional dataset into a dataset with reduced dimensionality without losing too much information [90]. Dimensionality reduction techniques help the user to get a better understanding of the underlying structure of the data by removing multicollinearity and creating a dataset with a smaller volume. In clinical settings, dimensionality reduction is often required, as EHRs are often high dimensional. Thus, by reducing the dimensions, one can mitigate this issue and possibly decrease the computational time for analysis and visualization of the EHR data. For instance, PHENOTREE [52] uses sparse principal component analysis (SPCA) to identify primary clinical features that describe the patient population to assist the user in building and navigating a phenotype hierarchy and exploring patient cohorts.

4.3. Visualizations

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

Relation-based: Relation-based visualizations show connections and relationships between two or more attributes. They are inherent to the clustering and association tasks within VA. A variety of visualization techniques can be used to display relations, such as scatter plots, parallel coordinates plots, bubble charts, bar charts, and heatmaps. For instance, Gotz et al. [26] use a scatter plot to show the distribution of the most frequent patterns with respect to their level of support for patients with positive and negative outcomes.

Time-based: Time-based visualizations show data or the sequence of clinical events over a time period. The main function of these visualizations is to assist the analysis and reasoning process of healthcare experts when investigating patients' clinical history. The primary time-based visualization technique is Timeline. Timeline displays a series of clinical events in a temporal order where each event is generally represented by an icon and is encoded by size, shape, or color to distinguish events with different characteristics. For instance, Peekquence [51] displays each patient's entire event sequence in a timeline to assist users in discovering patterns in the patient's event sequences.

Hierarchy-based: Hierarchy-based visualizations show how data items are ordered and ranked in a system. Several visualization techniques can be used to display the hierarchical structure of the data, such as tree diagrams, treemaps, and icicle plots. For instance, DecisionFlow [22] aggregates clinical event sequences with a similar occurrence of milestone events into a tree of sequences, where each node encodes an event positioned according to its prefix in the sequence. VALENCIA [53] uses a treemap to display the distribution of patient characteristics in different clusters.

Flow-based: Flow-based visualizations show flows and their quantities with respect to one another. Sankey diagrams and parallel sets are two of the main flow-based visualization techniques that are used in EHR-based VA systems to provide an overview of transitions between different types of clinical events. For instance, Care Pathway Explorer [33] uses a Sankey diagram to show how clinical events in the most frequent patterns are connected to each other, where each event is represented by a node and nodes belonging to the same pattern are connected by edges.

4.4. Interactions

Interaction is an integral part of VA and plays a vital role in the success of EHR-based VA systems. We adapted the epistemic actions introduced as part of the framework proposed by Sedig et al. [91] to classify and evaluate interactions used in EHR-based VA systems. Epistemic actions are actions that are taken to alter the visualizations in a manner that supports the user's analytical and cognitive needs (mental processes). The subset of the actions identified in the framework commonly used in EHR-based VA systems [91] include: (1) arranging, (2) comparing, (3) drilling, (4) filtering, (5) searching, (6) selecting, (7) transforming, (8) translating, (9) animating/freezing, (10) collapsing/expanding, (11) inserting/removing, and (12) linking/unlinking.

Arranging: Arranging refers to acting upon visualizations to change their ordering, either temporally or spatially. Some variants of this epistemic action that are commonly used in EHR-based VA systems are sorting, ordering, organizing, and ranking. For instance, VALENCIA [53] supports arranging by allowing the user to sort the heatmap that represents the result of dimensionality reduction techniques based on either a dimension or a feature by clicking on the corresponding column or row.

Comparing: Comparing refers to acting upon visualizations to determine their degree of similarity or dissimilarity. The degree of similarity is often defined as the distance between or proximity of value or meaning in EHRs-based VA systems. For instance, RegressionExplorer [34] allows the user to investigate a regression model's behavior on a specific subpopulation and compare it with its behavior on a different subpopulation. It supports this action by letting the user drag and drop a feature from the variable selection view to the population view, which results in the partition of the patient population.

Drilling: Drilling is acting upon visualizations to bring out deep information that is currently not displayed. Its main functionality is to make perceptually inaccessible information available for further investigation. Drilling is a fundamental action in EHR-based VA systems as it helps the user process and examines desired information more deeply when dealing with a large volume of data stored in EHRs. For instance, Visa_M3R3 [54] supports this action by allowing the user to hover their mouse over a glyph representing a regression model in the scatter plot to get additional information about the corresponding model.

Filtering: Filtering refers to acting upon visualizations to show a subset of their elements based on specific criteria. It allows the user to adjust the level of details, which is an essential feature of the process of abstraction in the exploration of complex high-dimensional EHRs. Thus, filtering is integral to many EHR-based VA tasks. For instance, the VA system developed by Mica et al. [35] allows the user to filter a cohort of patients according to various parameters (e.g., body temperature, age, and lab results).

Searching: Searching refers to acting upon visualizations to locate or seek out the existence of position of certain relationships, items, or structures. Some variants of this action are querying and seeking. Searching is commonly used in EHR-based VA systems. For instance, Visa_M3R3 [54] supports this action by allowing the user to enter the name of the medication of interest in a search bar, allowing that medication to get highlighted in other views.

Selecting: Selecting refers to acting upon visualizations to focus on or choose them either individually or as a group. This action is necessary for performing other actions in VA systems. By selecting an information item and making it visually distinctive, the user can keep track of it within a large volume of information, even when it is going through some changes. Most of the EHR-based VA systems support selecting. For instance, DecisionFlow [22] allows the user to perform edge selection by clicking on time edges in the temporal flow view. The system then outlines the corresponding rectangular mark representing the edge and updates the overview and the edge statistics view to display information regarding the selected edge.

Transforming: Transforming refers to acting upon visualizations to modify their geometric form. This epistemic action can change the look, size, or orientation of visualizations by scaling, rotating, magnifying, and/or distorting them. Magnifying visualizations is the most common variant of this action in EHR-based VA systems. For instance, Visa_M3R3 [54] applies the cartesian fisheye distortion technique on both axes of the scatter plot representing regression models to help the user distinguish between models when the glyphs are densely clustered.

Translating: Translating is acting upon visualizations to convert them into alternative conceptually- or informationally-equivalent forms. This action has a high degree of utility for all EHR-based VA tasks as each alternative visualization form reveals different aspects of the data. For instance, SubVIS [45] supports translating by allowing the user to choose a more advanced representation of glyphs to show the subspace's underlying dimensions where each dimension is encoded by a small line around the border of the dot.

Animating/Freezing: Animating/Freezing refers to acting upon dynamic visualizations to create movement in constituent parts or oppositely to stop. Animating can be used to observe temporal trends and show complex relationships among clinical events in EHRs. For instance, the VA system developed by Gotz et al. [26] uses a three-staged animation process to transition between different views in the Pattern Diagram view to highlight temporal patterns by allowing comparison between mining results at different parts of a clinical episode.

Collapsing/Expanding: Collapsing/Expanding is acting upon visualizations to make them compact or, oppositely, make them diffuse. These actions can facilitate investigating the associations among data items when dealing with complex high-dimensional EHRs. Collapsing enables the user to condense a set of data items into one, thus reducing complexity and facilitating the understanding of overall associations and patterns within EHRs while expanding allows the user to explore them in more detail. For instance, DPvis [24] allows the user to convert Pathway over Observation diagram (i.e., a stacked Sankey diagram displaying pathways for subjects) into a bipartite Sankey diagram, which includes two stacked bars with paths between them. The converted view simplifies the transition by only displaying the changeover from a start to an end, thereby allowing the user to follow the state pathways between two consecutive patient visits.

Inserting/Removing: Inserting/Removing refers to acting upon visualizations to add new visualizations into them, or oppositely to take out unwanted or unnecessary parts. Such actions can facilitate the exploration of EHRs, allowing the user to create hypothetical scenarios by interjecting or getting rid of clinical events in patient trajectories and observing the effect. For instance, CarePre [50] enables the user to edit the focal patients' event sequence within the prediction view by inserting new events and removing existing ones. This allows the user to determine key factors that affect the prediction results and explore how changes in the patient's record (i.e., a novel treatment or the absence of a comorbidity) impact those results.

Linking/Unlinking: linking/Unlinking refers to acting upon visualizations to establish an association between them, or oppositely to disconnect their associations. In general, EHR-based VA systems with multiple coordinated views are assumed to support these actions. For instance, in VisualDecisionLinc [32], all the views are linked together to create a coordinated display, where filtering updates on one of the views prompts relevant updates to data items in other views. This aids the user in their decision-making process and evaluation of multiple treatment options by helping them to better understand the relationship between different data elements.

5. Discussion and Limitations

There are many challenges that designers might face when developing EHR-based VA systems. In addition to common data-related problems such as data integration, ease of use, and interpretability, EHRs introduce several domain-specific challenges. One of the main challenges is that EHRs often contain different data types such as medications, procedures, diagnosis codes, unstructured clinical codes, radiology results, and laboratory tests. Each of these types of data is composed of a large number of features. For instance, there are around 68,000 unique diagnostic codes in the ICD-10 (International Classification of Diseases—Tenth Revision) coding system. Across all data types, the number of features can increase to the hundreds of thousands. The other issue is the high proportion of missing data in EHRs caused by either documentation issues (e.g., human errors) or data collection issues. Adding to this challenge is that the absence of clinical events is often not recorded in EHRs. This makes it very difficult to differentiate between clinical events that have not occurred and missing events. Data sparsity in EHRs is one of the other challenges that is unavoidable because most patients take only specific medical examinations and treatments. In addition, there is a challenge of evaluating EHR-based VA systems as designers have to refine and validate their prototypes by testing them in realistic environments with overloaded physicians. Finally, another critical issue is that many clinical features are temporal. VA systems should therefore be able to uncover novel temporal patterns involving discrete and interval clinical events. Therefore, the design of EHR-based VA systems requires a deep understanding of users' needs and expectations, VA tasks, individual components of VA systems, and how to best integrate them.

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

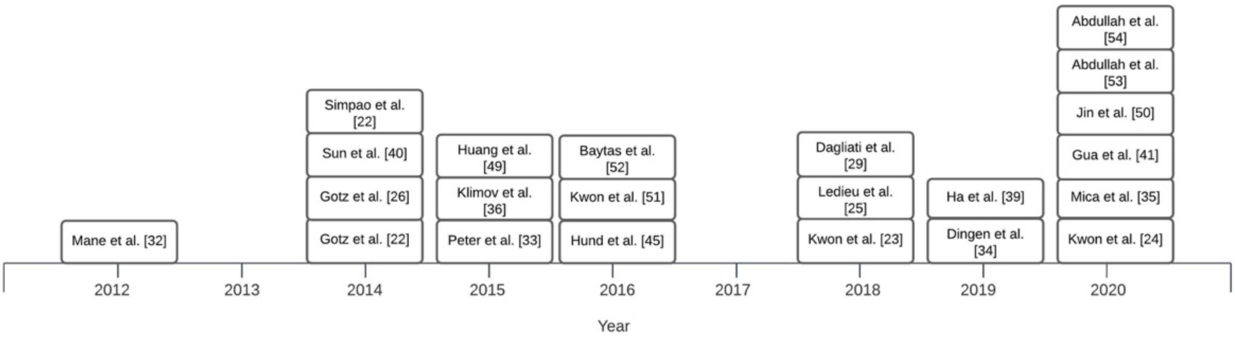
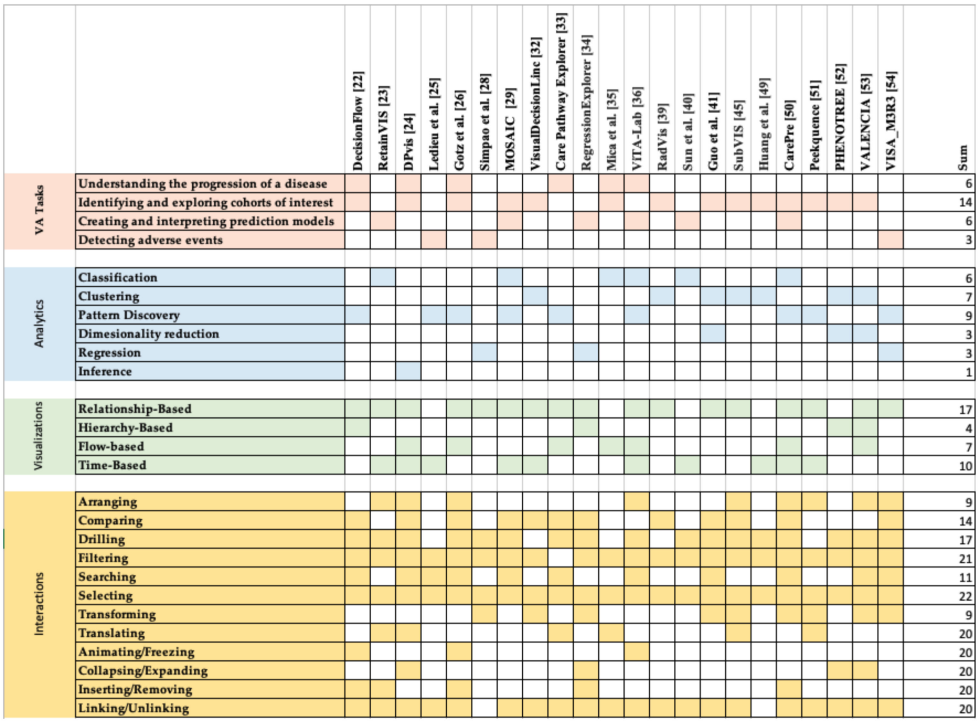


Figure 10. The figure shows the original order of the creation of the VA systems in a timeline [22–26,29,32–36,39–41,45,49–54].



		DecisionFlow [22]	RetainVIS [23]	DPvis [24]	Ledieu et al. [25]	Gotz et al. [26]	Simpao et al. [28]	MOSAIC [29]	VisualDecisionLinc [32]	Care Pathway Explorer [33]	RegressionExplorer [34]	Mica et al. [35]	VITA-Lab [36]	RadVis [39]	Sun et al. [40]	Guo et al. [41]	SubVIS [45]	Huang et al. [49]	CarePre [50]	Peekquence [51]	PHENOTREE [52]	VALENCIA [53]	VISA_M3R3 [54]	Sum
VA Tasks	Understanding the progression of a disease	■		■		■				■		■	■											6
	Identifying and exploring cohorts of interest	■		■		■		■	■			■		■		■	■	■	■	■	■	■		14
	Creating and interpreting prediction models		■					■			■		■		■				■					6
	Detecting adverse events				■		■																■	3
Analytics	Classification		■					■				■	■		■				■					6
	Clustering								■					■		■	■	■			■	■		7
	Pattern Discovery	■			■	■		■		■			■						■	■			■	9
	Dimesionality reduction															■					■	■		3
	Regression						■				■												■	3
	Inference			■																				1
Visualizations	Relationship-Based	■	■	■		■	■	■	■	■	■		■	■		■	■		■	■		■	■	17
	Hierarchy-Based	■									■										■	■		4
	Flow-based			■		■				■		■	■						■			■		7
	Time-Based		■	■	■			■	■				■		■			■	■	■				10
Interactions	Arranging		■	■		■							■				■		■	■		■	■	9
	Comparing	■		■		■		■	■	■	■			■		■	■		■				■	14
	Drilling	■	■	■		■	■	■		■	■		■		■	■	■	■	■		■	■	■	17
	Filtering	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	21
	Searching	■		■	■	■		■		■			■			■			■			■	■	11
	Selecting	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	22
	Transforming						■		■		■					■	■		■	■		■	■	9
	Translating		■	■						■		■					■			■				20
	Animating/Freezing	■				■							■											20
	Collapsing/Expanding			■							■										■	■		20
	Inserting/Removing	■	■			■					■								■					20
	Linking/Unlinking	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	20

Figure 11. The reviewed EHR-based visual analytics system. Each system is labeled by the relevant element in the design space. The rows are grouped and colored by dimensions of the design space: VA tasks, analytics, visualization, and interaction [22–26,28,29,32–36,39–41,45,49–54].

The most common VA task that is supported by the systems is discovering and exploring patient cohorts. This task's popularity is mostly because of its numerous use cases, including survival analysis, clinical trial recruitment, and other kinds of retrospective studies [58,59]. Moreover, identifying patients who satisfy pre-defined criteria can form the platform for future studies in areas such as predicting patient outcomes, pharmacovigilance, and understanding patient trajectories. Thus, as seen in Figure 11, discovering and exploring patient cohorts is supported by most systems that support other VA tasks. Other VA tasks supported by several systems are understanding the progression of a disease and learning and exploring prediction models. Understanding the incidence, prevalence, and coincidence of diseases is the foundation on which important policy decisions are made, and thus researchers have spent considerable efforts on this task. Similarly, learning and exploring prediction models is the most natural and immediately impactful task. Conversely, discovering adverse events is not as popular as the other VA tasks. This could be due to the fact that this task requires extensive collaborations.

Pattern discovery is one of the most widely used analytics methods in EHR-based VA systems. It is often used to uncover common subsequences in the clinical event sequences and quantify the similarity between event sequences. This analytics method is commonly used in most of the systems that support understanding the progression of diseases. The other popular analytics method is clustering. For instance, the VA system developed by Gotz et al. [26] and Care Pathway Explorer [33] use the SPAM [27] frequent sequence mining algorithm to detect the most frequent patterns and examines how these patterns are associated with patient outcomes. Clustering often aims to organize patients into several groups, where patients within each group have similar characteristics. Most of the systems that support cohort discovery tasks use different clustering techniques. For instance, the VA system developed by Guo et al. [41], VALENCIA [53], and RadVis [39] apply several clustering techniques to assists clinical researchers in the identification of cohorts of patients with similar clinical characteristics. One of the other commonly used analytics methods is classification. It is often used in VA systems that support learning and exploring prediction models. For instance, RetainVIS [23] and CarePre [50] support this VA task by creating deep learning prediction models and allowing the user to explore and interpret the results. Dimensionality reduction, regression, and inference are the other analytics methods that are used in EHR-based VA techniques. Dimensionality reduction techniques are usually used as a pre-processing step, followed by clustering techniques.

Most of the systems use multiple visualizations to allow the user to explore the data and the analytics results from different perspectives. The most common visualization used in the systems is relation-based visualizations, including scatter plots, bar charts, heatmaps, and parallel coordinates plots. Scatter plots are most suitable in representing clustering techniques, while bar charts are usually used to show the distribution of clinical features and their contributions to the prediction models. The other common visualization used in the EHR-based VA systems is time-based. The reviewed systems frequently adopt time-based visualizations such as timelines to display temporal distribution of clinical events in different time granularities and reveal temporal information among clinical event sequences. CarePre [50] and RetainVIS [23] represent patient's clinical events in a temporal order in a time-based visualization that allows the user to conduct what-if analyses by modifying these events and get the newly generated predicted risks for the patient. Thus, the systems can display the result of classification and pattern discovery techniques by adopting time-based visualizations. Flow-based visualizations such as Sankey diagrams and parallel sets have also been adopted by many EHR-based VA systems. They are mostly used to provide an overview of the progression pathways of clinical events within a cohort and thus, help the user to understand which clinical features, pathways, or other structures are more associated with the outcome of interest. Finally, a small number of EHR-based VA systems adopt hierarchy-based visualizations such as icicle plots and treemap to reveal the hierarchical organization of features or event sequences within EHRs. For instance, while DecisionFlow [22] uses a hierarchy-based tree to display the aggregated progression patterns of interest, VALENCIA [53] allows the user to explore the hierarchical structure of clustering results using a tree map.

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

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

This review has a few key limitations. First, we do not investigate the usability, and the user base of these systems as our review relies only on the descriptions of the VA systems found in publications. Second, we could not examine the accuracy and completeness of the data sources the reviewed systems are using. Finally, we excluded the EHR-based VA systems with static visualizations as interaction is one of the main dimensions of our proposed design space.

6. Conclusions

In this paper, we conduct a systematic literature review to gather research papers that describe the design and development of EHR-based VA systems and provide a comprehensive overview of these systems. We then propose a design space, including four primary dimensions used to characterize and evaluate the existing EHR-based VA systems. These key dimensions include VA tasks, analytics, visualizations, and interactions. This review shows the major application of analytics, visualizations, and interactions in supporting the execution of EHR-driven VA tasks. We connect and unify the existing work using the dimensions identified in the proposed design space. Furthermore, we identify the challenges that a designer might confront when developing EHR-based VA systems. Finally, we discuss areas of little prior work and identify promising future research directions.

Author Contributions: Conceptualization, N.R., S.S.A., and K.S.; methodology, N.R., S.S.A., and K.S.; investigation, N.R. and S.S.A.; writing—original draft preparation, N.R.; writing—review and editing, N.R., S.S.A., and K.S.; supervision, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: We would like to thank all authors and publishers who shared images of their tools with us.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. 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]
2. Doupi, P. Using EHR Data for Monitoring and Promoting Patient Safety: Reviewing the Evidence on Trigger Tools. *Stud. Health Technol. Inf.* **2012**, *180*, 786–790.
3. Agrawal, A. Medication Errors: Prevention Using Information Technology Systems. *Br. J. Clin. Pharmacol.* **2009**, *67*, 681–686. [CrossRef]
4. Dey, S.; Luo, H.; Fokoue, A.; Hu, J.; Zhang, P. Predicting Adverse Drug Reactions through Interpretable Deep Learning Framework. *BMC Bioinform.* **2018**, *19*, 476. [CrossRef] [PubMed]
5. Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Lizotte, D.J.; Garg, A.X.; McArthur, E. Machine Learning for Identifying Medication-Associated Acute Kidney Injury. *Informatics* **2020**, *7*, 18. [CrossRef]
6. Tang, P.C.; McDonald, C.J. Electronic health record systems. In *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*; Shortliffe, E.H., Cimino, J.J., Eds.; Health Informatics; Springer: New York, NY, USA, 2006; pp. 447–475. ISBN 978-0-387-36278-6.
7. Christensen, T.; Grimsmo, A. Instant Availability of Patient Records, but Diminished Availability of Patient Information: A Multi-Method Study of GP's Use of Electronic Patient Records. *BMC Med. Inform. Decis. Mak.* **2008**, *8*, 12. [CrossRef]
8. Rostamzadeh, N.; Abdullah, S.S.; Sedig, K. Data-Driven Activities Involving Electronic Health Records: An Activity and Task Analysis Framework for Interactive Visualization Tools. *Multimodal Technol. Interact.* **2020**, *4*, 7. [CrossRef]
9. Heisey-Grove, D.; Danehy, L.N.; Consolazio, M.; Lynch, K.; Mostashari, F. A National Study of Challenges to Electronic Health Record Adoption and Meaningful Use. *Med. Care* **2014**, *52*, 144–148. [CrossRef] [PubMed]
10. Lau, F.; Price, M.; Boyd, J.; Partridge, C.; Bell, H.; Raworth, R. Impact of Electronic Medical Record on Physician Practice in Office Settings: A Systematic Review. *BMC Med. Inform. Decis. Mak.* **2012**, *12*, 10. [CrossRef]
11. Ola, O.; Sedig, K. The Challenge of Big Data in Public Health: An Opportunity for Visual Analytics. *Online J. Public Health Inf.* **2014**, *5*, 223. [CrossRef]
12. Keim, D.A.; Mansmann, F.; Thomas, J. Visual Analytics: How Much Visualization and How Much Analytics? *ACM SIGKDD Explor. Newsl.* **2010**, *11*, 5. [CrossRef]
13. Sedig, K.; Parsons, P.; Babanski, A. Towards a Characterization of Interactivity in Visual Analytics. *J. Multimed. Process. Technol.* **2012**, *3*, 12–28.
14. Ribarsky, W.; Fisher, B.; Pottenger, W.M. Science of Analytical Reasoning. *Inf. Vis.* **2009**. [CrossRef]
15. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477. [CrossRef] [PubMed]
16. Cortez, P.; Embrechts, M.J. Using Sensitivity Analysis and Visualization Techniques to Open Black Box Data Mining Models. *Inf. Sci.* **2013**, *225*, 1–17. [CrossRef]
17. Keim, D.A.; Munzner, T.; Rossi, F.; Verleysen, M. Bridging Information Visualization with Machine Learning (Dagstuhl Seminar 15101). *Dagstuhl Rep.* **2015**, *5*, 1–27. [CrossRef]
18. Rajwan, Y.G.; Barclay, P.W.; Lee, T.; Sun, I.-F.; Passaretti, C.; Lehmann, H. Visualizing Central Line –Associated Blood Stream Infection (CLABSI) Outcome Data for Decision Making by Health Care Consumers and Practitioners—An Evaluation Study. *Online J. Public Health Inf.* **2013**, *5*, 218. [CrossRef]
19. Goldsmith, M.-R.; Transue, T.R.; Chang, D.T.; Tornero-Velez, R.; Breen, M.S.; Dary, C.C. PAVA: Physiological and Anatomical Visual Analytics for Mapping of Tissue-Specific Concentration and Time-Course Data. *J. Pharm. Pharm.* **2010**, *37*, 277–287. [CrossRef]
20. 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]
21. Lo, Y.-S.; Lee, W.-S.; Liu, C.-T. Utilization of Electronic Medical Records to Build a Detection Model for Surveillance of Healthcare-Associated Urinary Tract Infections. *J. Med. Syst.* **2013**, *37*, 9923. [CrossRef]
22. Gotz, D.; Stavropoulos, H. Decisionflow: Visual Analytics for High-Dimensional Temporal Event Sequence Data. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1783–1792. [CrossRef]
23. Kwon, B.C.; Choi, M.-J.; Kim, J.T.; Choi, E.; Kim, Y.B.; Kwon, S.; Sun, J.; Choo, J. Retainvis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 299–309. [CrossRef] [PubMed]
24. Kwon, B.C.; Anand, V.; Severson, K.A.; Ghosh, S.; Sun, Z.; Frohnert, B.I.; Lundgren, M.; Ng, K. DPVis: Visual Analytics with Hidden Markov Models for Disease Progression Pathways. *IEEE Trans. Vis. Comput. Graph.* **2020**. [CrossRef] [PubMed]
25. Ledieu, T.; Bouzille, G.; Plaisant, C.; Thiessard, F.; Polard, E.; Cuggia, M. Mining Clinical Big Data for Drug Safety: Detecting Inadequate Treatment with a DNA Sequence Alignment Algorithm. *AMIA Annu. Symp. Proc.* **2018**, *2018*, 1368–1376. [PubMed]
26. Gotz, D.; Wang, F.; Perer, A. A Methodology for Interactive Mining and Visual Analysis of Clinical Event Patterns Using Electronic Health Record Data. *J. Biomed. Inform.* **2014**, *48*, 148–159. [CrossRef]

27. Ayres, J.; Flannick, J.; Gehrke, J.; Yiu, T. Sequential Pattern Mining Using a Bitmap Representation. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23 July 2002; pp. 429–435.
28. 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.* **2014**, *22*, 361–369. [CrossRef]
29. Dagliati, A.; Sacchi, L.; Tibollo, V.; Cogni, G.; Teliti, M.; Martinez-Millana, A.; Traver, V.; Segagni, D.; Posada, J.; Ottaviano, M.; et al. A Dashboard-Based System for Supporting Diabetes Care. *J. Am. Med. Inf. Assoc.* **2018**, *25*, 538–547. [CrossRef]
30. Sacchi, L.; Capozzi, D.; Bellazzi, R.; Larizza, C. JTSA: An Open Source Framework for Time Series Abstractions. *Comput. Methods Programs Biomed.* **2015**, *121*, 175–188. [CrossRef] [PubMed]
31. Dagliati, A.; Sacchi, L.; Zambelli, A.; Tibollo, V.; Pavesi, L.; Holmes, J.H.; Bellazzi, R. Temporal Electronic Phenotyping by Mining Careflows of Breast Cancer Patients. *J. Biomed. Inf.* **2017**, *66*, 136–147. [CrossRef]
32. 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] [PubMed]
33. 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] [PubMed]
34. Dingen, D.; van't Veer, M.; Houthuizen, P.; Mestrom, E.H.J.; Korsten, E.H.H.M.; Bouwman, A.R.A.; van Wijk, J. RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 246–255. [CrossRef] [PubMed]
35. Mica, L.; Niggli, C.; Bak, P.; Yaeli, A.; McClain, M.; Lawrie, C.M.; Pape, H.-C. Development of a Visual Analytics Tool for Polytrauma Patients: Proof of Concept for a New Assessment Tool Using a Multiple Layer Sankey Diagram in a Single-Center Database. *World J. Surg.* **2020**, *44*, 764–772. [CrossRef]
36. 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]
37. Moskovitch, R.; Shahar, Y. Classification of Multivariate Time Series via Temporal Abstraction and Time Intervals Mining. *Knowl. Inf. Syst.* **2015**, *45*, 35–74. [CrossRef]
38. Moskovitch, R.; Shahar, Y. Fast Time Intervals Mining Using the Transitivity of Temporal Relations. *Knowl. Inf. Syst.* **2015**, *42*, 21–48. [CrossRef]
39. Ha, H.; Lee, J.; Han, H.; Bae, S.; Son, S.; Hong, C.; Shin, H.; Lee, K. Dementia Patient Segmentation Using EMR Data Visualization: A Design Study. *Int. J. Environ. Res. Public Health* **2019**, *16*, 3438. [CrossRef]
40. Sun, J.; McNaughton, C.D.; Zhang, P.; Perer, A.; Gkoulalas-Divanis, A.; Denny, J.C.; Kirby, J.; Lasko, T.; Saip, A.; Malin, B.A. Predicting Changes in Hypertension Control Using Electronic Health Records from a Chronic Disease Management Program. *J. Am. Med. Inf. Assoc.* **2014**, *21*, 337–344. [CrossRef] [PubMed]
41. Guo, R.; Fujiwara, T.; Li, Y.; Lima, K.M.; Sen, S.; Tran, N.K.; Ma, K.-L. Comparative Visual Analytics for Assessing Medical Records with Sequence Embedding. *Vis. Inform.* **2020**, *4*, 72–85. [CrossRef]
42. Gower, J.C.; Warrens, M.J. Similarity, Dissimilarity, and Distance, Measures Of. *Wiley StatsRef Stat. Ref. Online* **2014**, 1–11. [CrossRef]
43. Kramer, M.A. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AICHE J.* **1991**, *37*, 233–243. [CrossRef]
44. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3104–3112.
45. Hund, M.; Böhm, D.; Sturm, W.; Sedlmair, M.; Schreck, T.; Ullrich, T.; Keim, D.A.; Majnaric, L.; Holzinger, A. Visual Analytics for Concept Exploration in Subspaces of Patient Groups. *Brain Inf.* **2016**, *3*, 233–247. [CrossRef] [PubMed]
46. Müller, E.; Günnemann, S.; Assent, I.; Seidl, T. Evaluating Clustering in Subspace Projections of High Dimensional Data. *Proc. VLDB Endow.* **2009**, *2*, 1270–1281. [CrossRef]
47. Cox, M.A.A.; Cox, T.F. Multidimensional Scaling. In *Handbook of Data Visualization*; Chen, C., Härdle, W., Unwin, A., Eds.; Springer Handbooks Comp. Statistics; Springer: Berlin/Heidelberg, Germany, 2008; pp. 315–347. ISBN 978-3-540-33037-0.
48. Rao, R.; Card, S.K. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+ Context Visualization for Tabular Information. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 24–28 April 1994; pp. 318–322.
49. Huang, C.-W.; Lu, R.; Iqbal, U.; Lin, S.-H.; Nguyen, P.A.A.; Yang, H.-C.; Wang, C.-F.; Li, J.; Ma, K.-L.; Li, Y.-C.J.; et al. A Richly Interactive Exploratory Data Analysis and Visualization Tool Using Electronic Medical Records. *BMC Med. Inform. Decis. Mak.* **2015**, *15*, 92. [CrossRef]
50. Jin, Z.; Cui, S.; Guo, S.; Gotz, D.; Sun, J.; Cao, N. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Trans. Comput. Healthc.* **2020**, *1*, 1–20. [CrossRef]
51. Kwon, B.C.; Verma, J.; Perer, A. Peekquence: Visual Analytics for Event Sequence Data. In Proceedings of the ACM SIGKDD 2016 Workshop on Interactive Data Exploration and Analytics, San Francisco, CA, USA, 14 August 2016; Volume 1.

52. 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]
53. Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Garg, A.X.; McArthur, E. Visual Analytics for Dimension Reduction and Cluster Analysis of High Dimensional Electronic Health Records. *Informatics* **2020**, *7*, 17. [CrossRef]
54. 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]
55. Sedig, K.; Parsons, P. Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework. *Synth. Lect. Vis.* **2016**, *4*, 1–185. [CrossRef]
56. Yadav, P.; Pruinelli, L.; Hangsleben, A.; Dey, S.; Hauwiller, K.; Westra, B.L.; Delaney, C.W.; Kumar, V.; Steinbach, M.S.; Simon, G.J. Modelling Trajectories for Diabetes Complications. In Proceedings of the 4th Workshop on Data Mining for Medicine and Healthcare. 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, 30 April–2 May 2015.
57. Oh, W.; Kim, E.; Castro, M.R.; Caraballo, P.J.; Kumar, V.; Steinbach, M.S.; Simon, G.J. Type 2 Diabetes Mellitus Trajectories and Associated Risks. *Big Data* **2016**, *4*, 25–30. [CrossRef]
58. Mathias, J.S.; Gossett, D.; Baker, D.W. Use of Electronic Health Record Data to Evaluate Overuse of Cervical Cancer Screening. *J. Am. Med. Inf. Assoc.* **2012**, *19*, e96–e101. [CrossRef]
59. Strom, B.L.; Schinnar, R.; Jones, J.; Bilker, W.B.; Weiner, M.G.; Hennessy, S.; Leonard, C.E.; Cronholm, P.F.; Pifer, E. Detecting Pregnancy Use of Non-Hormonal Category X Medications in Electronic Medical Records. *J. Am. Med. Inf. Assoc.* **2011**, *18*, i81–i86. [CrossRef] [PubMed]
60. Galea, M.H.; Blamey, R.W.; Elston, C.E.; Ellis, I.O. The Nottingham Prognostic Index in Primary Breast Cancer. *Breast Cancer Res Treat.* **1992**, *22*, 207–219. [CrossRef]
61. Knaus, W.A.; Wagner, D.P.; Draper, E.A.; Zimmerman, J.E.; Bergner, M.; Bastos, P.G.; Sirio, C.A.; Murphy, D.J.; Lotring, T.; Damiano, A.; et al. The APACHE III Prognostic System: Risk Prediction of Hospital Mortality for Critically III Hospitalized Adults. *Chest* **1991**, *100*, 1619–1636. [CrossRef] [PubMed]
62. Timmerman, D.; Testa, A.C.; Bourne, T.; Ferrazzi, E.; Ameye, L.; Konstantinovic, M.L.; Van Calster, B.; Collins, W.P.; Vergote, I.; Van Huffel, S.; et al. Logistic Regression Model to Distinguish Between the Benign and Malignant Adnexal Mass Before Surgery: A Multicenter Study by the International Ovarian Tumor Analysis Group. *JCO* **2005**, *23*, 8794–8801. [CrossRef] [PubMed]
63. Nashef, S.A.; Roques, F.; Michel, P.; Gauducheau, E.; Lemeshow, S.; Salamon, R.; EuroSCORE Study Group. European System for Cardiac Operative Risk Evaluation (EuroSCORE). *Eur. J. Cardiothorac. Surg.* **1999**, *16*, 9–13. [CrossRef]
64. Chalmers, J.; Pullan, M.; Fabri, B.; McShane, J.; Shaw, M.; Mediratta, N.; Poullis, M. Validation of EuroSCORE II in a Modern Cohort of Patients Undergoing Cardiac Surgery. *Eur. J Cardiothorac. Surg.* **2013**, *43*, 688–694. [CrossRef]
65. Gaziano, T.A.; Bitton, A.; Anand, S.; Abrahams-Gessel, S.; Murphy, A. Growing Epidemic of Coronary Heart Disease in Low- and Middle-Income Countries. *Curr. Probl. Cardiol.* **2010**, *35*, 72–115. [CrossRef]
66. Munzner, T. *Visualization Analysis and Design*; CRC Press: Boca Raton, FL, USA, 2014; ISBN 978-1-4987-5971-7.
67. Treisman, A. Preattentive Processing in Vision. *Comput. Vis. Graph. Image Process.* **1985**, *31*, 156–177. [CrossRef]
68. Ware, C. *Information Visualization: Perception for Design*; Morgan Kaufmann: Burlington, MA, USA, 2019; ISBN 978-0-12-812876-3.
69. Institute of Medicine (US) Committee on Quality of Health Care in America; Kohn, L.T.; Corrigan, J.M.; Donaldso, M.S. *To Err Is Human: Building a Safer Health System*; National Academies Press: Washington, DC, USA, 2000; ISBN 978-0-309-26174-6.
70. Brennan, T.A.; Leape, L.L.; Laird, N.M.; Hebert, L.; Localio, A.R.; Lawthers, A.G.; Newhouse, J.P.; Weiler, P.C.; Hiatt, H.H. Incidence of Adverse Events and Negligence in Hospitalized Patients. *N. Engl. J. Med.* **1991**, *324*, 370–376. [CrossRef]
71. Leape, L.L.; Brennan, T.A.; Laird, N.; Lawthers, A.G.; Localio, A.R.; Barnes, B.A.; Hebert, L.; Newhouse, J.P.; Weiler, P.C.; Hiatt, H. The Nature of Adverse Events in Hospitalized Patients. *N. Engl. J. Med.* **1991**, *324*, 377–384. [CrossRef]
72. Thomas, E.J.; Studdert, D.M.; Burstin, H.R.; Orav, E.J.; Zeena, T.; Williams, E.J.; Howard, K.M.; Weiler, P.C.; Brennan, T.A. Incidence and Types of Adverse Events and Negligent Care in Utah and Colorado. *Med. Care* **2000**, *38*, 261–271. [CrossRef]
73. Wilson, R.M.; Runciman, W.B.; Gibberd, R.W.; Harrison, B.T.; Newby, L.; Hamilton, J.D. The Quality in Australian Health Care Study. *Med. J. Aust.* **1995**, *163*, 458–471. [CrossRef]
74. Thomas, E.J.; Studdert, D.M.; Newhouse, J.P.; Zbar, B.I.W.; Howard, K.M.; Williams, E.J.; Brennan, T.A. Costs of Medical Injuries in Utah and Colorado. *Inquiry* **1999**, *36*, 255–264. [PubMed]
75. Torio, C.M.; Elixhauser, A.; Andrews, R.M. Trends in Potentially Preventable Hospital Admissions among Adults and Children, 2005–2010: Statistical Brief #151. In *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs*; Agency for Healthcare Research and Quality (US): Rockville, MD, USA, 2006.
76. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000; ISBN 978-0-521-78019-3.
77. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
78. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
79. Lewis, D.D. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In *Proceedings of the European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 4–15.
80. Daniel, G.G. Artificial Neural Network. In *Encyclopedia of Sciences and Religions*; Runehov, A.L.C., Oviedo, L., Eds.; Springer: Dordrecht, The Netherlands, 2013; p. 143. ISBN 978-1-4020-8265-8.

81. 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]
82. Jain, A.K. Data Clustering: 50 Years beyond K-Means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
83. 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 978-3-319-21903-5.
84. Fraley, C.; Raftery, A.E. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *J. Am. Stat. Assoc.* **2002**, *97*, 611–631. [CrossRef]
85. 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.
86. Agrawal, R.; Imielinski, T.; Swami, A. Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, 26–28 May 1993; pp. 207–216.
87. Ismail, B.; Anil, M. Regression Methods for Analyzing the Risk Factors for a Life Style Disease among the Young Population of India. *Indian Heart J.* **2014**, *66*, 587–592. [CrossRef]
88. Stopar, L.; Skraba, P.; Grobelnik, M.; Mladenic, D. StreamStory: Exploring Multivariate Time Series on Multiple Scales. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 1788–1802. [CrossRef] [PubMed]
89. Bhattacharjya, D.; Shanmugam, K.; Gao, T.; Mattei, N.; Varshney, K.; Subramanian, D. Event-Driven Continuous Time Bayesian Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 3 April 2020; Volume 34, pp. 3259–3266.
90. Siwek, K.; Osowski, S.; Markiewicz, T.; Korytkowski, J. Analysis of Medical Data Using Dimensionality Reduction Techniques. *Przegląd Elektrotechniczny* **2013**, *89*, 279–281.
91. 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]