Similarity Search and Correlation-Based Exploratory Analysis in EHRs: A Case Study with COVID-19 Databases

Mirela T. Cazzolato[1], Lucas S. Rodrigues[1], Marcela X. Ribeiro[2], Marco A. Gutierrez[3], Caetano Traina Jr.[1], Agma J. M. Traina[1]

[1] Institute of Mathematics and Computer Sciences
University of São Paulo (ICMC-USP), São Carlos, Brazil

[2]Computer Sciences Department
Federal University of São Carlos (DC-UFSCar), São Carlos, Brazil

[3]Heart Institute (InCor) – Clinical Hospital of Faculty of Medicine
University of São Paulo (HC-FMUSP), São Paulo, Brazil

mirela@usp.br, agma@icmc.usp.br

Abstract. *With the COVID-19 pandemic, many hospitals have collected Electronic Health Records (EHRs) from patients and shared them publicly. EHRs include heterogeneous attribute types, such as image exams, numerical, textual, and categorical information. Simply posing similarity queries over EHRs can underestimate the semantics and potential information of particular attributes and thus would be best supported by exploratory data analysis methods. Thus, we propose the Sketch method for comparing EHRs by similarity to provide a tool for a correlation-based exploratory analysis over different attributes. Sketch computes the overall data correlation considering the distance space of every attribute. Further, it employs both ANOVA and association rules with lift correlations to study the relationship between variables, allowing a deep data analysis. As a case study, we employed two open databases of COVID-19 cases, showing that specialists can benefit from the inference modules of Sketch to analyze EHRs. Sketch found strong correlations among tuples and attributes, with statistically significant results. The exploratory analysis has shown to complement the similarity search task, identifying and evaluating patterns discovered from heterogeneous attributes.*

1. Introduction

According to the World Health Organization[1], the Coronavirus disease (COVID-19) is an infectious disease that makes most infected people experience mild to moderate respiratory illness. According to the Brazilian Ministry of Health, COVID-19 has infected more than 18 million people in Brazil, with more than half a million confirmed deaths in the country[2]. With great effort, diverse health institutions have collected, organized, and shared public information from COVID-19 patients, aimed at supporting studies in the understanding and analysis of such cases [FAPESP 2020, Cohen et al. 2020]. Electronic Health Records (EHRs) store patients' personal information of heterogeneous types [Yadav et al. 2018, Jensen et al. 2012], which are acquired and organized by health

[1]https://www.who.int/health-topics/coronavirus, accessed on July 2, 2021.
[2]https://covid.saude.gov.br/, accessed on July 2, 2021.

institutions. Examples of such information range from simple data, *e.g.*, dates, textual observations and diagnosis, and exam results, such as blood counts, to complex data, such as, electrocardiograms, X-Ray images, and Computed Tomography (CT). EHRs may also include data acquired from different hospitals and clinics, requiring the applications to consider the data interoperability issue [Gansel et al. 2019, Jensen et al. 2012]. The growing amount of available data is also a relevant issue. Query and analysis tasks should provide timely results, aiding specialists in their everyday routine. Accordingly, Database Management Systems (DBMS) can organize the available data and perform queries timely, supporting analytical and exploratory tools.

In this paper, we explore similarity search and correlation-based exploratory analysis over tuples with heterogeneous attribute types acquired from diverse sources. That is, we show how to evaluate, visually present, and take advantage of correlations among different types of attributes to learn meaningful information from heterogeneous data.

Similarity queries have been a relevant topic approached by the Database community for decades now, since the introduction of the multimedia data [Farias et al. 2019, Samet 2006]. Attribute types can be scalars, such as dates, numbers, and small strings, or complex, such as images and time series. When posing queries over tuples with such attributes, the query engine must employ a specific representation and comparison measure for every data dimension. Distance functions compare attributes according to their types and application requirements [Deza and Deza 2009, Samet 2006]. For instance, we can compute the difference in unities from numbers and dates or compare images according to their similarity of color distribution.

Beyond comparing tuples of data, specialists need to access meaningful patterns in the database to comprehend the available information and make decisions. In the exploratory data analysis, correlation measures can uncover, expose and show relationships between attributes of unlabeled data [Hoshen and Wolf 2018]. Analyzing the correlation between variables and their underlying interactions is essential for multivariable datasets [Zhang et al. 2016] and has been the subject of study for decades now [DSouza and Velan S. 2020, Yang et al. 2019, Kaieski et al. 2016]. This work aims to take advantage of correlation measures and visual tools to support the exploratory analysis and data understanding.

Working with similarity-based comparisons of EHRs (or tuples with heterogeneous types of attributes) and their exploratory analysis encompasses different problems. Similarity functions employed to compare pairs of tuples attribute-wise can somehow underestimate the semantic meaning of the information, for instance, when considering categorical attributes. Correlation analysis can complement the analysis by employing different coefficients and metrics according to the data type. While exploring tuples, the approach must identify the proper methods to employ, given the data characteristics. Moreover, the correlation-based analysis should be grounded by metrics such as confidence intervals and frequencies [Han et al. 2011]. Finally, the discovered correlations, patterns, and findings are not always intelligible and trivial for domain specialists to understand. Visual tools can improve knowledge readability and understanding, considering distinct data characteristics.

We approach an exploratory analysis supplemented by interestingness measures

based on correlation and visualization tools. Unlike existing works, we aim to take advantage of the different types of attributes to enrich the analysis and support of similarity queries. We propose the *Sketch* approach (***S**imilarity and **E**xploratory **T**asks with **C**orrelation-based **H**euristics*) for the exploratory analysis and similarity search over EHRs. The main contributions of *Sketch* are two-fold. First, (i) the method allows posing similarity-based queries of tuples considering heterogeneous attributes. For tuples, we propose a new heuristic (*Sketch-Corr*) to compute the correlation between variables, considering the distance space of every attribute. We use scatter plots to show the multi-dimensional distance space of tuples, and heatmaps to show the global correlations found in the data. An important characteristic of EHRs is the use of categorical attributes to describe different values that have specific semantics to the medical domain. Generic distance functions may fail to compare and analyze such data adequately. Thus, in the second contribution of this work (ii) we focus on improving the semantics of exploratory analysis obtained by categorical attributes. *Sketch* discovers association rules (AR) from different categories and analyzes the corresponding lift correlation scores. Sankey diagrams visually show the discovered rules, with transitions between correlated items. Finally, we employ the Analysis of Variance (*ANOVA*) correlation for the combination of categorical and numerical values. Box-plots visually show the relationship of categories regarding the numerical variable. The experimental analysis over two open COVID-19 datasets shows that *Sketch* can find significant patterns for all analysis tools employed. We also provide *Sketch-GUI*, which implements all functionalities and visual tools of *Sketch*. It is open-source and available for download in a public repository aimed at supporting future researches.

This paper is organized as follows. Section 2 describes the relevant background and related work. Section 3 presents the proposed approach. Section 4 shows the experimental analysis and discussion. Finally, Section 5 gives the conclusions.

2. Background and Related Work

Similarity Search. Data retrieval in Database Management Systems (DBMSs) compares pairs of objects based on operators of identity ($=$ and $\neq$) and order ($<, \leq, >$ and $\geq$). Similarity-based comparisons can work with both scalar (such as numbers, dates, and small strings) and complex (such as images, time series, and text) data. When posing similarity searches over a database, the specialist must define the best feature extraction method (for complex attributes) to represent the feature vectors of objects and the distance functions to compare pairs of objects. Distance Functions (δ) measure the dissimilarity between both feature vectors. Given two objects s_1, s_2 and a distance function δ, the (dis)similarity among s_1 and s_2 is measured as $\delta(s_1, s_2)$, resulting in a real value in $\mathbb{R}^+$. Several distance functions are suitable for similarity comparisons, such as those from the Minkowski family for numerical data and Levenshtein (L_{Edit}) for textual data [Samet 2006].

The two basic similarity queries are Similarity Range and *k*-Nearest Neighbors (*k*NN). Let $\mathcal{D}$ be a dataset of complex objects $S \in \mathbb{S}$, where $\mathbb{S}$ is the complex domain, δ be a distance function, and s_q and s_i be elements in domain $\mathbb{S}$, where s_q is the query center. A Similarity Range Query retrieves every element $s_i \in S$ where the distance to s_q is less or equal than a similarity radius ξ, *i.e.* $\delta(s_q, s_i) \leqslant \xi$. The *k*-NN Query retrieves the k objects $s_i \in S$ that are most similar to s_q, measured by a given distance function δ.

Correlation Heuristics. Correlation coefficients provide the existing association between variables in a dataset [DSouza and Velan S. 2020]. Examples of well-known correlations employed in the literature are Pearson, Spearman, and *ANOVA*.

ANOVA (**AN**alysis **O**f **VA**riance) analyzes two or more populations described by a numeric variable and at least one categorical variable [Han et al. 2011]. *ANOVA* test can show significant differences between numerical values and two or more categorical groups. *ANOVA* returns two values, *F-test* and *p-value*. F-test is a correlation score that informs how much the actual means of the groups deviate from the primary assumption that the means of all groups are the same. The higher the F-test score, the larger the difference between the means. As a complement, the *p-values* inform the statistical significance of the given score.

Association Rules (AR) look for items that co-occur in a database. Let I be the itemset with all possible items in a transactional database D. AR are implications in the form of $A \Rightarrow B$, where $A \subset I$ and $B \subset I$ are non-empty itemsets, and $A \cap B = \emptyset$. The *support* of a rule is given by the proportion of D that contains $A \cup B$, and the *confidence* refers to the proportion of transactions in D containing A that also contain B:

$$\textit{sup}(A \Rightarrow B) = P(A \cup B)$$

$$\textit{conf}(A \Rightarrow B) = P(B|A) = \textit{sup}(A \cup B)/\textit{sup}(A)$$

We supplement the support-confidence evaluation of AR with the lift correlation measure. The occurrence of A is said to be independent of B if $P(A \cup B) = P(A)P(B)$. Otherwise, both itemsets are dependent and correlated as events. In practice, the *lift* correlation is measured from the support and confidence values as $\textit{lift} = \textit{conf}(A \Rightarrow B)/\textit{sup}(B)$. It assesses the degree to which the occurrence of one item "*lifts*" the occurrence of the other [Han et al. 2011]. If $\textit{lift} = 1$, A and B are independent and there is no correlation between them; $\textit{lift} < 1$ indicates a negative correlation between the itemsets, where the presence of one implies the absence of the other; and $\textit{lift} > 1$ indicates a positive correlation between A and B, where the occurrence of one implies the occurrence of the other. As AR work with transactions, the EHRs must be converted before the pattern discovery step. Therefore, every combination of $\{\textit{categorical\_attribute}, \textit{value}\}$ is converted to an attribute, which can be present or absent in a given tuple.

Exploratory Analysis with Correlation. Several studies in the literature propose exploratory data analysis tools to support data understanding and pattern discovery. In [Xiao et al. 2016], the authors present an exploratory analysis using the Spearman coefficient to find the most significant attributes over a total of 207,880 available. They aimed at detecting the operational condition of pumps and at predicting mechanical faults. Using scatter plots and histograms over the mapped correlation matrix, the study selected only highly correlated attributes. However, the work does not provide further details on the impact of the feature selection in the analysis results.

In [Kaieski et al. 2016], the authors propose the Vis-Health visual tool to analyze Dengue incidence in Brazil, crossing public health information with climatic factors in the corresponding regions. They applied a correlation analysis, and selected the most relevant attributes such as rainfall, temperature, and the number of Dengue cases. Experimental results show the same pattern in 6 out of 7 analyzed state capitals, relying on maps and

pie charts to understand the found results and patterns. The data correlation was computed using the Principal Component Analysis technique over the original data, which changes the semantics of attributes in the exploratory analysis. In [Huang et al. 2019], the authors present a recommendation model based on data entropy and decision trees optimized by correlation analysis between Pearson and Kendall. The results over medical datasets show patterns found by employing visualization tools based on Chord and Sankey diagrams. Specifically, the authors present the relationship of pharmaceutical compositions and recommend frequent rules when the model maps strong relationships. However, the data integration step of their pipeline performs a dimensionality reduction in the data, which can discard useful information or introduce bias in the model.

The authors of [Yang et al. 2019] studied how the human perception of visual features relates to correlation patterns using scatter plots. Although the study does not provide an exploratory data analysis, they found indications that differences in visual features are more predictive of human judgments than the correlation itself. Their findings indicate that people pay attention to only a few visual features discriminating between different scatter-plot representations. More recently, in [DSouza and Velan S. 2020] the authors presented an exploratory analysis with visual tools to find patterns over a dataset with COVID-19 cases from Italy. The work analyzes the relationship between variables, crossing with statistical data from each region of the country. The visualization tools assisted the authors in identifying possible trends and insights based on more relevant attributes. However, the work does not discuss correlation measures and the statistical significance of the identified relationships.

3. *Sketch*: The Proposed Method

In this section, we introduce *Sketch* (***S**imilarity and **E**xploratory **T**asks with **C**orrelation-based **H**euristics*). *Sketch* creates a multifunctional environment, providing a holistic approach that is able to evaluate the correlation between discrete and continuous values, representing categorical, numerical, and complex attributes. Additionally, the method provides correlations and association rules to explore individual samples and multiresolution clusters of samples, including similarity-based comparisons.

Figure 1 illustrates *Sketch* and its six main modules. In (a), the specialist collects patients' information and provides it to the query engine. By "*sketch*" we also mean that the user can choose to have only partial information for data exploring, selecting the relevant attributes, and analyzing the correlation scores of a subset of attributes according to their types. In (b), the query engine receives the query arguments (*i.e.* attributes) and verifies the available tools for exploratory analysis regarding the attribute types. *Sketch* compares tuples (c) with the given distance functions for every attribute, allowing the analyst to weigh the important variables according to their correlation, computed in module (d). For this, *Sketch* implements the *Sketch-Corr* correlation heuristic, which computes the correlation between every pair of attributes. *Sketch* uses the provided information (e) to find correlations between data records, and retrieves the most similar tuples. The analyst can continue to explore the data using AR with lift and *ANOVA* correlations from the similarity results. Finally, (f) the user can evaluate the results using the available visualization tools and analysis results. When visualizing different attributes, *Sketch* considers the variables being used and the adequate tools to employ. We explain the heuristics employed in the Similarity Search and Correlation modules in the next subsections.

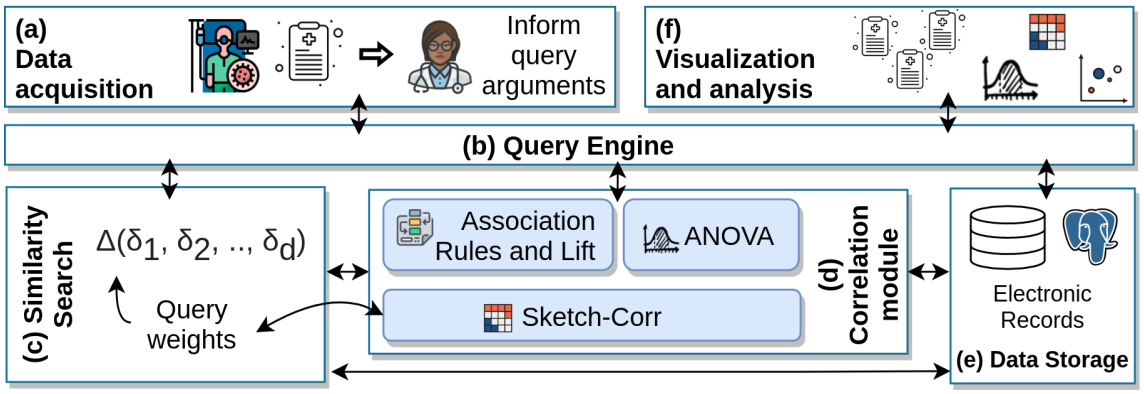


Figure 1. The *Sketch* method.

3.1. *Sketch-Corr* for Tuples

Let D be a dataset with EHRs of d attributes and n tuples. *Sketch* employs distance functions according to every attribute type. Equation 1 gives the *Sketch-Dist*, the *global* distance function Δ to compare a pair of tuples $< t_i, t_j >$, such that $0 \leq i, j \leq n$:

$$\Delta_{i,j} = \frac{1}{w} \sqrt[p]{\sum_{a=1}^{d} |\delta(t_i^a, t_j^a)|^p \times w_a}, \tag{1}$$

where w is the sum of attribute weights, such that $w = \sum_{a=1}^{d} w_a$. The global function Δ works as a Minkowski distance of order p. For every attribute a, such that $1 \leq a \leq d$, the *local* function δ gives the distance between the pair of tuples, given every specific attribute type. Alternatively, the global function can weigh every attribute according to its corresponding relevance in the comparison, represented by w_a.

While correlation coefficients, such as Spearman and Kendall, rely on global references, we aim at determining the correlation locally, that is, in every region of the data space. Also, although such coefficients can work with more than one attribute, the attributes need to be concatenated while preserving the lexicographical order. To overcome these limitations, we propose *Sketch-Corr*, which builds over those monotonic correlation coefficients to handle correlations between the variation of distances among attributes.

Algorithm 1 details *Sketch-Corr*. It receives as input S (a sample of m tuples from D), with d attributes, the set of distance functions F, that will compare every type of attribute among $Tp = \{numeric, textual, categorical, date, complex\}$, and the correlation coefficient Φ, such as Pearson and Spearman. In Line 1 *Sketch-Corr* initializes the variables. For every attribute a and every tuple t_i in S (Lines 3 and 4), *Sketch-Corr* computes the array of distances $D_a[i]$ of the tuple t_i to all other tuples in the dataset (Line 5). As a result, D_a has the mean distance variation of the tuples concerning attribute a. For every pair of attributes $< a_r, a_s >$ in S (Line 6), such that $1 \leq r, s \leq d$, the algorithm computes the correlation between the arrays of distance variations D_r and D_s (Line 7). *Sketch-Corr* returns the correlation matrix M of dimensions $d \times d$ (Line 8). Every row q ($1 \leq q \leq d$) of M has the correlation scores of attribute a_q to all other attributes. The values correspond to the weights w of Δ (see Eq. 1), with the reference attribute a_q.

Sketch employs two visualization tools to evaluate tuples: a heatmap and a scatter plot. The heatmap represents the correlation matrix M with a color pattern, where more

Algorithm 1: *Sketch-Corr* to compute the correlation between attributes

Data: S: a sample dataset with d attributes and m tuples
F: a set of distance functions
Φ: the correlation coefficient
Result: M: Matrix of correlations of dimensions $d \times d$

begin
  Initialization
  foreach *attribute a in* S **do**
    for i *from 1 to* m **do**
      /* Mean distance of t_i to all the other tuples */
      $D_a[i] \leftarrow \frac{1}{m} \sum_{j=1}^{m} \delta_a(t_i, t_j)$ // where $\delta_a \in F$
  foreach *pair of attributes* $< a_r, a_s >$ *in* S **do** // where $1 \leq r, s \leq d$
    /* Compute the correlation between the attributes */
    $M[r][s] \leftarrow \Phi\,(D_r, D_s)$
  return M

saturated colors represent strong correlations (positive or negative), and colors close to white represent weak correlations. The scatter plot shows the distribution of tuple distances. *Sketch* employs the Manifold Multidimensional Scaling (MDS) method to display the data distribution of objects (in this case, tuples) in a two-dimensional space, using the distances between them. We provide the original and correlation-weighted spaces when posing queries, showing the different results for both options.

Sketch-Corr gives the *overall* correlation among the attributes. Notice that some distance functions can underestimate the semantics of specific attribute values, such as categorical ones, and the relationships of those with other attributes. The overall data analysis informs the analyst of potential patterns that should be further investigated. Existing correlation heuristics can also take advantage of specific attribute types (and values) to provide meaningful data semantics, as we show next.

3.2. Association Rules (AR) and Lift Correlation for Categorical Attributes

Sketch takes advantage of AR to find categories of attributes that frequently occur together in the database, given the minimum support and confidence values. In this context, the lift metric gives the correlation between the antecedent and consequent items (in our case, category values) of discovered AR. Importantly, *Sketch-GUI* provides SQL-query support for users to select the relevant attributes and filter tuples, given a criterion. For a large number of tuples, *Sketch* can pre-process the dataset and store the discovered AR to use in further analysis. In this step, the minimum support and confidence values give the strength of the rules, which should be provided by the data analyst, and the corresponding lift value, which gives the correlation of every found rule.

The analyst informs the attribute and corresponding value to be used as a reference. *Sketch* searches for all patterns that include the given arguments and shows the corresponding rules with the support, confidence, and lift correlation. *Sketch* employs Sankey diagrams to visualize the AR returned by the algorithm. In the diagram, every item corresponds to a vertical bar, linkages represent items that co-occur in the discovered AR, and the line thickness of the item linkages corresponds to the rule confidence.

3.3. *ANOVA* for the Combination of Numerical and Categorical Attributes

Finally, *Sketch* employs *ANOVA* to analyze pairs of attributes, showing existing differences between categorical and numerical variables. In our context, *ANOVA* allows the specialists to test and check which variables are important to take into account when analyzing a specific numerical measurement. We employ Box-plots to visually show the differences between different groups of the same variable.

3.4. *Sketch-GUI*: A Visual Tool for Information Retrieval

We provide *Sketch-GUI*, an application that implements correlation heuristics and visualization tools of *Sketch*. *Sketch-GUI* prototype allows users to query the tuples by similarity and perform an exploratory analysis over the available data using *Sketch-Corr*, AR-Lift, and *ANOVA*. The application is available in a Git repository (ref. to Section 4), and currently supports numerical, categorical, text, and complex attributes.

4. Experiments

COVID-19 Public Datasets. We evaluate *Sketch* with two public datasets, namely, *Ds-IEEE* (the IEEE Covid-19 [Cohen et al. 2020]) and *Ds-FAPESP* (FAPESP COVID-19 DataSharingBR [FAPESP 2020]). *Ds-IEEE* combines data acquired from diverse sources, also containing a complex attribute (chest X-Ray or CT):

- **Covid_Cases**, 950 tuples. `NUMERIC`: offset_at, age, temperature, po2_saturation, leukocyte_count, neutrophil_count, lymphocyte_count; `TEXTUAL`: location, clinical_notes, other_notes; `CATEGORICAL`: sex, finding, rt_pcr_positive, survival, intubated, intubation_present, went_icu, in_icu, needed_supplemental_o2, extubated, view, modality; `COMPLEX`: chest_xray_ct.

Ds-FAPESP has data of five hospitals from São Paulo State: Israelita Albert Einstein, Fleury Group, Clinical Hospital of the School of Medicine of USP, Sírio-Libanês, and Beneficência Portuguesa of São Paulo. *Ds-FAPESP* has a standard data schema, with three tables and attributes of different types. We consider only numerical, textual, and categorical attributes:

- **Patient**, 602,552 tuples. `NUMERICAL`: aa_birth; `CATEGORICAL`: de_sex, cd_city, cd_state, cd_country. **Exam**, with 32,981,024 tuples. `TEXTUAL`: de_result, de_exam; `CATEGORICAL`: de_orign, de_analyte, cd_unity, de_reference_value. **Outcome**, with 260,681 tuples. `CATEGORICAL`: de_attendance_type, de_clinic, de_outcome.

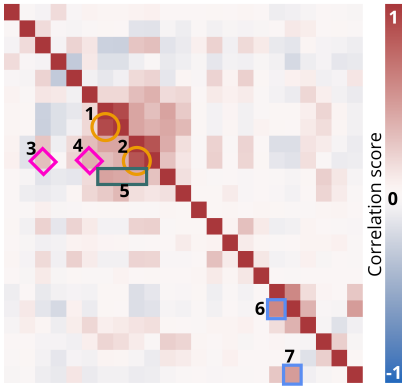
Both datasets were collected on May 13, 2021. Scripts for data pre-processing and insertion in the PostgreSQL database are available in a Git repository[3], together with the image features of *Ds-IEEE* (*chest_xray_ct*), the discovered AR patterns and pre-processed information. *Sketch-Corr* uses the Levenshtein (LEdit) distance function for categorical and textual attributes, and Euclidean for the remaining ones. Due to space limitations, we focus our quantitative evaluation on the correlation-based exploratory analysis.

Quantitative Evaluation. Figure 2 (a) shows the correlation among attributes obtained by *Sketch-Corr* from *Ds-IEEE* (executed in 25*s* in a sample of 200 tuples). Some very

[3]Git repository of Sketch: *https://github.com/mtcazzolato/sketch*.

high correlations indicate obvious relationships, such as (1) *intubation_present* with *intubated* (*score*= 0.9) and (2) *in_icu* with *went_icu* (*score*= 0.85). We can also observe medium correlations with interesting patterns, such as (3) *in_icu* with *age* (*score*= 0.3), (4) *in_icu* with *survival* (*score*= 0.4), (5) *needed_supplemental_O2* with *intubated* (*score*= 0.43), *intubation_present* (*score*= 0.4) and *went_icu* (*score*= 0.4), (6) *modality* with *view* (*score*= 0.58), and (7) with *chest_xray_ct* (*score*= 0.5). For cases (6) and (7), the attributes are related to the type and position of the imaging exam. All correlations show *p-value* < 0.01. Figure 2 (b) shows the AR discovered from categorical values with lift correlation different from 1. The AR discovery takes only few seconds. The links between items refer to the confidence values of rules. The discovered patterns show that the "RT-PCR" exam with a positive result is correlated to the "Pneumonia/Viral/COVID-19" finding, and the majority of cases are from male patients. Another interesting pattern is that most rules indicate an implication of "Pneumonia/Viral/COVID-19" with "X-Rays" images (instead of "CT scans"). Transforming the input data into transactions may take several minutes, depending on the selected sample. Thus, we recommend saving the generated file for future analysis using *Sketch-GUI*.

(a) *Sketch-Corr* Correlations



(b) AR with lift correlations

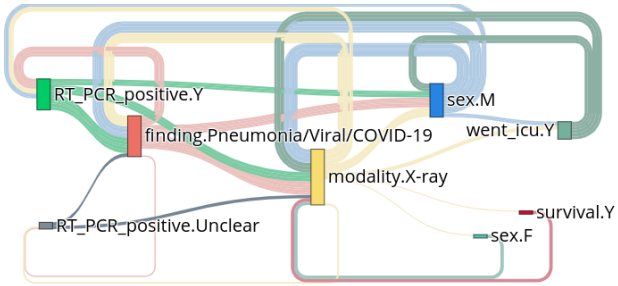


Figure 2. Exploring *Ds-IEEE*: (a) Heatmap of *Sketch-Corr* correlations and (b) the Sankey diagram with AR discovered from categorical values.

ANOVA checks the distribution of categorical values regarding a specific numerical variable. Figure 3 shows correlation scores of *ANOVA* (executed in 15*s*). Categories presenting different means regarding the numerical variable (Cases 1 and 2) will show low correlation scores, identifying groups with different behavior. The low-correlated patterns show interesting findings, where (1) the patients' temperature tends to be high when they need supplemental O2, and (2) the "leukocyte count" tends to be higher for patients in the ICU. Case 3, on the other hand, has a high correlation score with statistically significant *p-value* (< 0.01), which indicates that the difference in age of patients in and out of the ICU has a similar mean. We observed that *Sketch-Corr* did not present high correlation scores between the attributes of all three cases, showing the importance of employing different tools to provide a more diverse exploration analysis of the data.

Figure 4 (a) shows the correlation heatmap obtained by applying *Sketch-Corr* over *Ds-FAPESP* (executed in 38*s* in a sample of 500 tuples). The map shows (1) medium correlations of attribute *de_origin* with *de_exam* (*score*= 0.5) and *de_analyte* (*score*= 0.5). A medium correlation was also observed in (2) between attributes *de_outcome* and

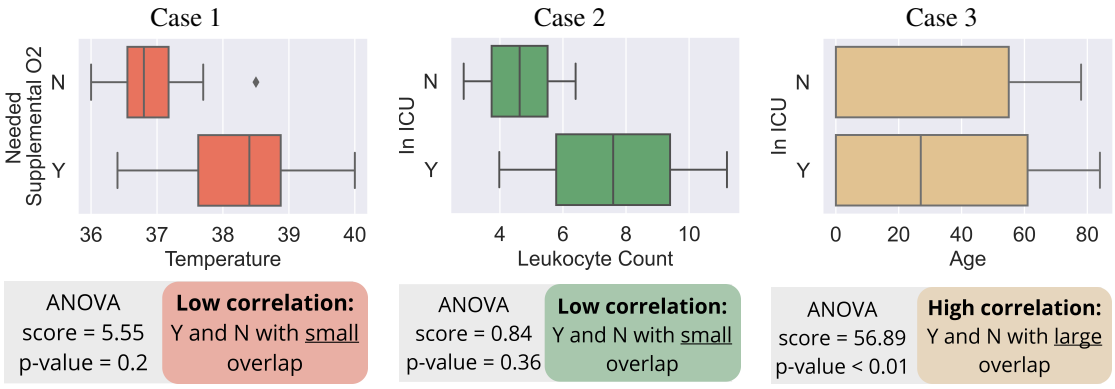


Figure 3. *ANOVA* correlations obtained from from *Ds-IEEE*.

de_attendance_type (*score*= 0.5). A very high correlation is shown in (3) between *de_analyte* and *de_exam* (*score*= 0.99), which we found to be trivial because both variables are semantically complementary. This is also the case of (4), correlating attribute *de_reference_value* with *de_exam* (*score*= 0.78) and *de_analyte* (*score*= 0.77). All correlations have *p-value*< 0.01. For further analyzing the available correlations, we filtered the *Outcome* table of *Ds-FAPESP* to analyze only tuples presenting an analyte related to COVID-19, selecting attributes *ic_sex*, *de_analyte*, *de_attendance_type*, and *de_outcomes* to provide better AR analysis. Figure 4 (b) shows the discovered AR patterns, with lift correlation different from 1. The diagram items show two analyte values found as frequent. We notice reasonable transitions between *de_outcome* and *de_attendance_type* (pattern 2 of *Sketch-Corr* heatmap), which mostly relate to cases of administrative discharge.

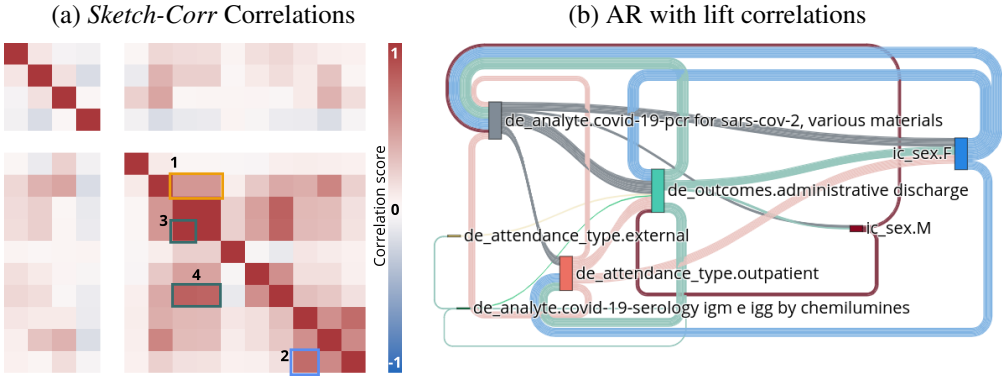


Figure 4. Exploring *Ds-FAPESP*: (a) Heatmap of *Sketch-Corr* correlations and (b) the Sankey diagram with AR discovered from categorical values.

To further analyze pattern (1) of the heatmap, we included the *de_result* attribute to the *ANOVA* analysis, aimed at evaluating the behavior of categorical values regarding the results of the exams. Figure 5 depicts the analyses of two interesting cases we found (executed in 2*s*). In the first example, we analyze the "Urea (plasma)" results among patients tested with the "coronavirus covid-19" analyte, observing a different dispersion for patients tested positive for COVID-19. In the second example, we verified the "Vitamin B12 (Blood)" analyte results regarding different origin places. *ANOVA* correlation scores were high and statistically significant for many combinations. This result is reasonable since the categories show major overlapping regarding the numerical variable, but with

different means. However, the chart visually shows a large value deviation for patients from the ICU and inpatient unities, showing a high concentration of both analytes. Again, the combination of correlation measures and visual tools improved the data analysis, allowing us to identify interesting patterns.

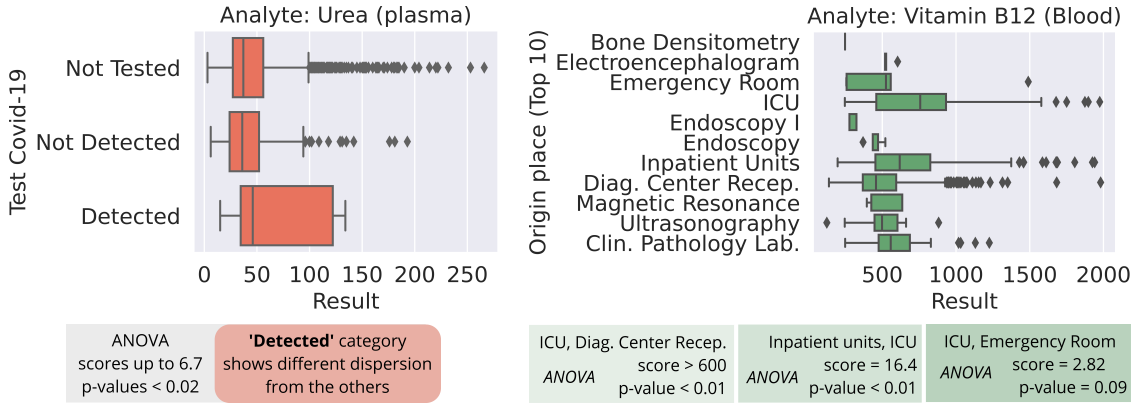


Figure 5. Analyte results and *ANOVA* correlation from *Ds-FAPESP*.

Lessons Learned. Both datasets presented in the experiments contain many tuples with missing data. We observed relevant differences in the *Sketch-Corr* results when comparing the correlation obtained with incomplete tuples, tuples with data imputation, and working only with complete tuples. Furthermore, the available data, especially from *Ds-FAPESP*, lacks value standardization, which is expected since the information comes from five hospitals. For instance, attributes *de_analyte* and *de_unity* are co-dependent, and have compatible information but with distinct nomenclature. Attribute *de_result* has a mixture of numbers and texts, many of them with the same meaning, resulting in an extensive range of unique values that hurts algorithms based on frequency. Besides pre-processing the available data, transforming the current modeling into a Common Data Model widely used in the medical domain could improve data readability, semantic interoperability and, consequently, carry more profound and systematic pattern recognition processes.

5. Conclusion

We presented *Sketch* to allow similarity search of tuples with heterogeneous attributes, providing correlation-based exploratory data analysis tools. The method includes *Sketch-Corr*, an algorithm that computes the correlation among attributes based on individual distance spaces. *Sketch* evaluates the overall correlation among heterogeneous attributes with *Sketch-Corr*, categorical ones with AR and lift correlation, and also the relationship between categorical and numerical attributes with *ANOVA* correlation. The experimental analysis performed over two COVID-19 databases shows the application of correlation and visual tools of *Sketch* in analyzing heterogeneous data. Also, we provide *Sketch-GUI* as a visual tool that implements *Sketch* and is openly available for research purposes.

The correlation-based analysis proposed in this work aimed at objectively improving the interestingness of pattern recognition in EHRs. The next step of this work consists of assessing the subjective quality of our analysis with domain specialists, assisted by *Sketch-GUI*. Also, a relevant future study is the longitudinal evaluation of patient-wise patterns in EHRs over time, considering different exam results of the same patient. The available data contains potential information such as admission dates and exams, contributing to the extension of the correlation analyses for time series analysis.

Acknowledgments. This research was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, by the São Paulo Research Foundation – FAPESP (grants no 2020/11258-2, 2020/07200-9, 2020/10902-5, 2016/17078-0), and the National Council for Scientific and Technological Development (CNPq).

References

Cohen, J. P. et al. (2020). Covid-19 image data collection: Prospective predictions are the future. *arXiv 2006.11988*.

Deza, M. M. and Deza, E. (2009). Encyclopedia of distances. In *Encyclopedia of distances*, pages 1–583. Springer. DOI: 10.1007/978-3-642-00234-2.

DSouza, J. and Velan S., S. (2020). Using exploratory data analysis for generating inferences on the correlation of covid-19 cases. In *ICCCNT Conference*, pages 1–6. IEEE. DOI:10.1109/ICCCNT49239.2020.9225621.

FAPESP (2020). FAPESP COVID-19 Data Sharing/BR. `https://repositoriodatasharingfapesp.uspdigital.usp.br`.

Farias, J. d., Barioni, M. C., and Rezende, H. (2019). Explorando o uso de árvores b+ na indexação de dados por similaridade. In *SBBD Conference*, pages 163–168, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2019.8817.

Gansel, X., Mary, M., and van Belkum, A. (2019). Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review. *EJCMID Journal*, 38(6):1023–1034. DOI:10.1007/s10096-019-03501-6.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann.

Hoshen, Y. and Wolf, L. (2018). Unsupervised correlation analysis. In *CVPR Conference*, pages 3319–3328. DOI: 10.1109/CVPR.2018.00350.

Huang, H., Zhang, R., and Lu, X. (2019). A recommendation model for medical data visualization based on information entropy and decision tree optimized by two correlation coefficients. In *ACM ICICM Conference*, page 52–56. DOI: 10.1145/3357419.3357436.

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

Kaieski, N., de Oliveira, L. P. L., and Villamil, M. B. (2016). Vis-health: Exploratory analysis and visualization of dengue cases in brazil. In *HICSS Conference*, pages 3063–3072. IEEE. DOI: 10.1109/HICSS.2016.385.

Samet, H. (2006). *Foundations of multidimensional and metric data structures*. M. K. series in data management systems. Academic Press.

Xiao, C. et al. (2016). Using spearman's correlation coefficients for exploratory data analysis on big dataset. *CCPE Journal*, 28(14):3866–3878. DOI: 10.1002/cpe.3745.

Yadav, P., Steinbach, M., Kumar, V., and Simon, G. (2018). Mining electronic health records (EHRs): A survey. *ACM Computing Surveys*, 50(6). DOI: 10.1145/3127881.

Yang, F. et al. (2019). Correlation judgment and visualization features: A comparative study. *IEEE TVCG Journal*, 25(3):1474–1488. DOI:10.1109/TVCG.2018.2810918.

Zhang, H., Hou, Y., Qu, D., and Liu, Q. (2016). Correlation visualization of time-varying patterns for multi-variable data. *IEEE Access*, 4:4669–4677. DOI: 10.1109/ACCESS.2016.2601339.