Discovery of discriminative patterns in oncological data to understand surgical risk factors

Leonardo Duarte Rodrigues Alexandre
Instituto Superior Técnico, Universidade de Lisboa

***Abstract*—Understanding the individualized risks of undertaking surgical procedures is essential to personalize preparatory, intervention and post-care protocols for minimizing post-surgical complications. This knowledge is key in oncology given the nature of interventions, the fragile profile of patients with comorbidities and drug exposure, and the possible cancer recurrence. Despite its relevance, the discovery of discriminative patterns of post-surgical risk is hampered by major challenges: 1) the unique physiological and demographic individual profile, as well as their differentiated post-surgical care, 2) the increasing high-dimensionality and heterogeneous nature of available biomedical data, combining non-identically distributed risk factors, clinical and molecular variables, 3) the need to learn from populations where tumors have significant histopathological differences and individuals undertake unique surgical procedures (structurally sparse data), 4) the need to focus on non-trivial patterns of surgical risk, while guaranteeing their statistical significance and discriminative power of post-surgical outcomes, and 5) the lack of interpretability and actionability of current approaches.**

This work proposes the use of biclustering, the discovery of groups of individuals correlated on subsets of variables, due to its unique properties of interest able to satisfy the aforementioned challenges, and a discretization method, DI2 (Distribution Discretizer) enablying a more robust pattern discovery on non-identically distributed variables. In this context, this work proposes a structured view on why, when and how to apply biclustering to mine discriminative patterns of post-surgical risk with guarantees of usability, a subject remaining unexplored up to date, and a fully autonomous, non-parametric and prior-free discretization method, DI2, for mixed variables with arbitrarily skewed distributions with support for multi-item assignments. Results show its relevance to improve classic discretization choices. The patterns offer a comprehensive view on how the patient's profile, cancer histopathology and entailed surgical procedures determine: 1) post-surgical complications, 2) survival, and 3) hospitalization needs.

The results confirm the role of biclustering in comprehensively finding interpretable, actionable and statistically significant patterns with a comprehensive view on how the patient's profile, cancer histopathology and entailed surgical procedures determine: 1) post-surgical complications, 2) survival, and 3) hospitalization needs. The patterns can be assisting healthcare professionals to establish specialized pre-habilitation protocols and support healthcare management decisions.

I. Introduction

Despite the relevance of discriminative pattern mining approaches, the discovery of patterns discriminating surgical outcomes is hampered by major challenges. First, individuals undertake personalized surgical procedures and differentiated post-surgical care, as well as show unique demographic, physiological, and tumor histopathological profiles. Second, the high-dimensionality and heterogeneous nature of available biomedical data, combining non-identically distributed risk factors, clinical records and biophysiological variables which contain structural sparsity, where the characterization of the interventions and outcomes are highly specific, yet relevant for the target end. Third, available data is inherently noisy and show arbitrarily-high levels of missing values. Fourth, there is the need to focus on non-trivial patterns of surgical risk able to discriminate post-surgical complications. In addition, the target patterns should strictly be statistically significant, thus minimizing susceptibility of false positive and negative discoveries. Finally, there is the need to guarantee the actionability and interpretability of the target patterns.

The nature of interventions, cancer recurrence, and fragile profile of patients (generally debilitated by the tumor effects and common need for cytotoxic chemotherapy) can cause small to life-threatening post-surgical complications [1], [2]. This work aims at exploring patterns of pre-surgical profiles to help professionals assess the various post-surgical outcomes of patients in need of surgical interventions. This knowledge is then translated into pre-surgical, surgical and post-surgical care protocols. This work proposes a methodology for the discovery of actionable pre-surgical patterns from available clinical data, with particular incidence on patterns able to discriminate the nature and severity of post-surgical complications, amount of required time in the HDU (high dependency unit) after surgery, and death susceptibility within the first year after surgery.

To address the aforementioned limitations of existing approaches, we propose the use of biclustering, the discovery of coherent subspaces, to comprehensively explore discriminative associations from heterogeneous oncological data.

The work is structured as follows. Section II introduces the theoretical concepts on the techniques used in the solution and the results obtained, and it also introduces traditional risk scores on surgical patients contained within the data made available to us. Section III surveys state-of-the-art pattern discovery and other approaches. Section IV describes the approached solution, the data used and its preprocessing, the algorithm used, the post-processing and visualization of the results. Section V presents the results obtained, their interpretation, and actionability. Finally, Section VI presents concluding remarks synthesized.

II. Background

The data in this work is in the form of a tabular dataset where each column represents a variable, each row represents a patient, and a_{ij} represents the value for j variable of i patient.

A. Traditional Risk scores

To facilitate perioperative risk assessment for the selection of patients benefiting from surgery, a variety of traditional scoring systems are used by the physicians: 1) P-POSSUM[1] (Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity), proposed by Prytherch *et al.* [3], 2) ACS NSQIP [2](American College of Surgeons National Surgical Quality Improvement Program), presented by Bilimoria *et al.* [4], 3) ARISCAT [3] (Assess Respiratory Risk in Surgical Patients in Catalonia), proposed by Canet *et al.* [5], and 4) Charlson comorbidity index[4], proposed by Charlson *et al.* [6]. All these simple risk scores are based on doctor-entered data.

B. Biclustering

Given a dataset defined by a set of observations X=$\{x_1, .., x_n\}$, variables Y=$\{y_1, .., y_m\}$, and elements $a_{ij} \in \mathbb{R}$ observed for observation x_i and variable y_j:

- a **bicluster** B=(I,J) is a $n \times m$ subspace, where $I = (i_1, .., i_n) \subseteq X$ is a subset of observations and $J = (j_1, .., j_m) \subseteq Y$ is a subset of variables;
- the **biclustering** task aims at identifying a set of biclusters $\mathcal{B} = (B_1, .., B_s)$ such that each bicluster $B_k = (I_k, J_k)$ satisfies specific criteria of *homogeneity*, *dissimilarity* and *statistical significance*.

Homogeneity criteria are commonly guaranteed through the use of a merit function, such as the variance of the values in a bicluster. Merit functions are typically applied to guide the formation of biclusters in greedy and exhaustive searches.

The pursued homogeneity determines the coherence, quality and structure of a biclustering solution [7]. The *coherence* of a bicluster is determined by the observed form of correlation among its elements (coherence assumption) and by the allowed value deviations from perfect correlation (coherence strength). The *quality* of a bicluster is defined by the type and amount of accommodated noise. The *structure* of a biclustering solution is defined by the number, size, shape and positioning of biclusters. A flexible structure is characterized by an arbitrary number of (possibly overlapping) biclusters.

Given a dataset, the elements within a bicluster $a_{ij} \in (I, J)$ have coherence across variables (**pattern on observations**) if a_{ij}=c_j+γ_i+η_{ij}, where c_j is the expected value of variable y_j, γ_i is the adjustment for observation x_i, and η_{ij} is the noise factor of a_{ij}. A bicluster has **constant coherence** when γ_i=0, and **additive coherence** otherwise, $\gamma_i \neq 0$.

Let r be the amplitude of values of the input data, **coherence strength** is a value $\delta \in [0, r]$ such that $a_{ij} = c_j + \gamma_i + \eta_{ij}$ where $\eta_{ij} \in [-\delta/2, \delta/2]$.

The bicluster **pattern** φ_J is the set of expected values in the absence of adjustments and noise $\{c_j \mid y_j \in J\}$.

Given a real-valued dataset, a bicluster $B = (I, J)$ satisfies the **order-preserving coherence** assumption if the values for each observation in I follow the same ordering π along the subset of variables J.

An example of constant and order-preserving type biclusters can be seen in Figure 1.

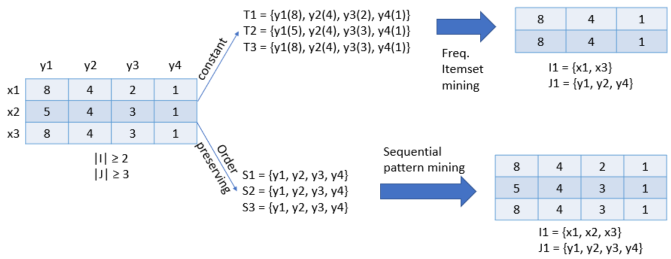


Fig. 1: Pattern-based biclustering: discovery of two illustrative biclusters with constant and order-preserving assumptions based on frequent itemsets and frequent subsequences from transactional data mapped from the input data matrix.

Statistical significance criteria, in addition to homogeneity criteria, guarantees that the probability of a bicluster's occurrence (against a null model) deviates from expectations [8].

In recent years, a clearer understanding of the synergies between biclustering and pattern mining paved the rise of a new class of algorithms, generally referred to as **pattern-based biclustering** algorithms [7]. Pattern-based biclustering algorithms are inherently prepared to efficiently find exhaustive solutions of biclusters and offer the unprecedented possibility to affect their structure, coherency and quality [9]. This behavior explains why this class of biclustering algorithms are receiving an increasing attention in recent years [7].

III. Related work

Veroneze *et al.* [10] presents an enumerative biclustering algorithm that efficiently mines maximal biclusters in mixed-attribute datasets without requiring any preprocessing steps such as discretization or itemization of real-valued attributes. Their proposed solution is an extension of RIn-Close_CVC. They argue that for mixed-attribute datasets only biclusters with constant values on columns are optimal in mixed-attribute datasets and propose a new definition for that type of bicluster, maintaining the monotonicity and anti-monotonicity properties. To select significant biclusters from the enumerative solution the authors propose two filters. One is based on formal concept analysis metrics (support, confidence, lift) to measure the quality of a rule. The second filter is a heuristic that locally maximizes the row-coverage. Their results showed that for five mixed-attribute labeled datasets the biclusters yield a tight set of rules which provide useful and interpretable models.

To utilize some of the the biclustering algorithms both Kaiser *et al.* [11] and Barkow *et al.* [12] implented toolboxes. These provide the user with a number of preprocessing, biclustering and cluster validation functions.

[1] https://www.mdcalc.com/possum-operative-morbidity-mortality-risk, accessed on December 2020

[2] https://riskcalculator.facs.org/RiskCalculator/index.jsp, accessed on December 2020

[3] https://www.mdcalc.com/ariscat-score-postoperative-pulmonary-complications, accessed in December 2020

[4] https://www.mdcalc.com/charlson-comorbidity-index-cci, accessed December 2020

Henriques *et al.* [13] provides a structured view on pattern mining-based approaches to biclustering and applied a qualitative comparison of the state-of-the-art pattern mining-based biclustering approaches supporting their accuracy, efficiency and biological relevance. The pattern mining-based biclustering algorithms analysed were DeBi proposed by Serin *et al.* [14], BiModule proposed by Okada *et al.* [15], GenMiner proposed by Martinez *et al.* [16], BicPAM proposed by Henriques *et al.* [9], RAP proposed by Pandey *et al.* [17], RCB Discovery proposed by Atluri [18], and ET-Bicluster proposed by Gupta *et al.* [19]. Henriques talks about what each of these state-of-the-art algorithms has to offer and the challenges that arise with the use of them. In terms of benificial factors, DeBi offers a complete and statistical rigorous post-processing. BiModule offers multi-level discretization and removal of outliers. GenMiner offers a more robust frame to deal with noisy biclusters. ET-Bicluster offers a parameterizable discovery of biclusters based on noise allowed. BicPAM can search for additive/multiplicative/symmetric/plaid bicluster models and deals with discretization, noise and missing. In terms of difficulties that arise, DeBi has a decrease in efficiency due to post-processing extension procedures, the data is binarized and can miss a large number of potentially significant biclusters due to discovering maximal patterns. BiModule has no merging-extension option to handle noise. RAP is not able to deal with noisy biclusters. RCB Discovery excludes biclusters with meaningful differences across columns when searching for biclusters with constant coherency overall, and has a combinatorial problem that impacts efficiency. ET-Bicluster does not guarantee exhaustive solutions when searching for patterns. BicPAM has efficiency problems for very large matrices when searching for biclusters with non-constant models.

IV. Solution

Our work aims at mining discriminative patterns of post-surgical outcomes from cancer patients and variables of interest. A pattern is a set of co-occurring attributes from surgical, biopathological, physiological and/or demographic variables, discriminative of post-surgical outcomes, and supported by a statistically significant set of individuals. Biclustering, the discovery of subspaces, is in this work suggested to this end. The pattern of a bicluster corresponds to a specific clinical profile, the pattern length corresponds to the number of attributes, and the pattern support corresponds to the individuals sharing the profile. The patterns searched follow either a constant assumption, characterized by a subset of variables on which a statistically significant number of patients have an identical profile, or a non-constant assumption. We seek the non-constant assumption due to the constant assumption suffering from a problem: two individuals need to share the same pattern in order to count as supporting observations for a bicluster. However, variations may be coherently explained by differences on their physiology or comorbidities. In this context, non-constant patterns should be pursued to guarantee a greater robustness to the variability of the profile of individuals, while still guaranteeing the coherence of the target patterns of surgical outcomes. Particularly, the order-preserving relaxation can be placed to find individuals with preserved orders of values observed on risk-measuring variables. Illustrating, if a specific risk score is higher than others for a group of individuals, this ordering can be a pattern irrespectively of the absolute value of the risk scores.

On *WHY*. Biclustering should be considered for mining patterns discriminative of surgical outcomes to: 1) avoid the drawbacks of classic pattern mining methods (including their susceptibility to the item-boundaries problems[5], inability to comprehensively explore heterogeneous biomedical data), 2) find non-trivial patterns discriminative of post-surgical outcomes with constant and order-preserving coherence, 3) pursue patterns with parameterizable properties of interest by customizing the target coherence strength, quality (noise-tolerance), dissimilarity and statistical significance.

On *WHEN*. Similarly, biclustering should be applied when: 1) the target patterns should provide guarantees of discriminative power and/or statistical significance, 2) pursuing non-trivial yet coherent forms of knowledge (including the introduced constant or order-preserving assumptions), 3) discretization drawbacks must be avoided, 4) heterogeneous data sources may be available, and when 5) one seeks to find comprehensive solutions with customizable homogeneity criteria.

On *HOW*: comprehensive exploration of clinical data. Pattern-based biclustering offers principles to find complete pattern solutions by: 1) pursuing multiple homogeneity criteria, including multiple coherence strength thresholds, coherence assumptions and quality thresholds, and 2) exhaustively yet efficiently exploring different regions of the search space, preventing that regions with large patterns jeopardize the search [9]. As a result, non-trivial yet significant correlations within the available clinical data are not neglected.

In addition, pattern-based biclustering does not require the input of support thresholds as it explores the search space at different supports [9], i.e. we do need to place expectations on the minimum number of individuals with a shared profile of surgical risk. Dissimilarity criteria and condensed representations can be also placed [9] to prevent the delivery of redundant patterns.

On *HOW*: statistical significance. A sound statistical testing of the patterns of surgical risk is key to guarantee the absence of spurious relations, and ensure the relevance of the given patterns to support mobility decisions. To this end, the statistical tests proposed in BSig [8] are suggested to minimize false positives (outputted patterns yet not statistically significant) without incurring on false negatives. This is done by approximating a null model of the target clinical data and statistically testing each bicluster against the null model in accordance with its underlying coherence.

On *HOW*: robustness to noise. Pattern-based biclustering can find biclusters with a parameterizable tolerance to noise [9]. Illustrating, a quality of 80% indicates that an upper limit given

[5] The possibility to allow deviations from value expectations (under limits defined by the placed coherence strength) together with multi-item assignments [9] are placed to prevent discretization problems from occurring

by 20% of a_{ij} entries within a bicluster may fail to follow the target clinical profile ($\mu_{ij} \notin [-\delta/2, \delta/2]$). This possibility ensures robustness to the individual-specific variations on a specific variable from a given pattern.

On *HOW*: other opportunities. Additional benefits of pattern-based biclustering that can be carried towards the analysis of surgical risk data include: 1) incorporation of domain knowledge to guide the task in the presence of background information (e.g. focus on a specific type of cancer of surgical procedure) [20], 2) the possibility to remove uninformative elements in data to guarantee a focus, for instance, on complications [21], 3) support classification and regression tasks using associative models composed by discriminative patterns [7].

A. Dataset description

A retrospective cohort of cancer patients undertaken surgery at the Portuguese Institute of Oncology, Porto, Portugal (IPO-Porto) were monitored (2016 to 2018) for this study. The gathered data, termed *IPOscore* dataset, contains information pertaining to the demographic and physiological patient characteristics, cancer location and histopathological determinants, risk scores, surgical procedures, and post-surgical outcomes. The risk scores within the dataset are P-POSSUM, ACS NSQIP, ARISCAT, and Charlson comorbidity Index. The IPO-Porto Ethics Committee approved (CES IPO:91/019) the analysis of the anonymized *IPOscore* data.

The dataset contains 847 patients (samples/observations) with 138 variables (33 binary, 45 nominal, 8 ordinal, 35 numerical, 13 free-text, 4 date). Of these variables in the clustered setting 4 are considered as outcomes of interest: 1) presence-absence of post-surgical complication, 2) Clavien-Dindo index of post-surgical severity, 3) days spent in HDU, and 4) death within 1 year. In the integrative setting 14 variables were considered as target variables, the previous mentioned and 10 new ones: 1) request type anesthesia, 2) provenance, 3) HDU motive of admission, 4) number of days at IPO, 5) admitted into intensive care, 6) average nursery points per day, 7) destination after HDU, 8) readmitted into HDU, 9) destination after IPO, 10) moment of death after surgery. The patients included in this study were selected because they had co-morbidities or because the surgery to be performed was complex, which advocated that the immediate postoperative be monitored in the HDU.

Two informative text variables, named ICD-10 and ACS procedures, exist in the dataset. These indicate the undertaken surgical procedures and are discussed in the next section.

B. Data transformation

The dataset contained typing mistakes which were fixed, a non uniform representation of missing values across the columns existed and were all converted to a global representation. Finally, texts columns, which contained important information regarding the surgical interventions each patient was subjected to, were normalized into binary columns.

Clustered versus integrative setting:

For the Clustered setting we considered that the patterns should be able to discriminate four outcomes: 1) post-surgical complication, 2) clavien-dindo post-surgical index, 3) days spent at HDU, and 4) death within 1 year. The pattern discriminates one of these outcomes if the measure *lift*[6] is above a certain threshold.

The dataset was partioned into four sub-datasets: 1) ICD-10, 2) ACS_procs, both of these two sub-datasets contain only the surgical interventions, 3) Scores, this sub-dataset contains only the output variables of each score within the dataset, 4) Non-score variables, this sub-dataset contains the physiological, demographic and operative variables. A total of sixteen sub-datasets were created, four sub-datasets for each outcome considered. Feature ranking was applied in the non-score and score output datasets to reduce the number of attributes.

In the integrative setting, nine outcomes are considered: 1) post-surgical complication, 2) clavien-dindo post-surgical index, 3) days spent at HDU, 4) death within 1 year, 5) days spent at IPO, 6) destination after HDU, 7) average points NAS per day, 8) HDU readmission, 8) destination after IPO, and 9) moment of death after surgery. We also consider patterns for: 1) request type anesthesia, 2) provenance, 3) HDU motive of admission, and 4) passed by intensive care. In this setting no attributes are removed based on feature ranking tests and the dataset is not partitioned. Values from binary/categorical variable that simbolize the absence of a disease/condition are replaced with missing values, this substitution is also applied to values that occur more than 70% within a variable. We implemented and applied a new form of discretization of numerical variables, DI2, and a range-based discretization, where numerical variables are put into categories of equal width based on range of the variable (from min to max), before applying BicPAMs algorithm.

C. Data Discretizer approach

Approaches to discretization of continuous variables have long been discussed alongside their pros and cons. Altman [22] and Bennette *et al.* [23] both discuss the relevance and impact of categorizing continuous variables and reducing the cardinality of categorical variables. Liao *et al.* [24] compares various categorization techniques in the context of classification tasks in medical domains, without using domain knowledge of field experts. The relevance of discretization meets both descriptive and predictive ends, encompassing state-of-the-art approaches such as pattern-based biclustering [9] and associative models such as XGBoost [25].

In this context, we propose DI2 (Distribution Discretizer), an approach that makes use of non-parametric tests to find the best fitting distribution for a given variable and discretize it accordingly. DI2 offers three major contributions: 1) corrections to the empirical distribution before statistical fitting to guarantee a more robust approximation of candidate

[6] Given an association $A \implies B$ where A is a pattern and B an outcome of interest, lift measures $P(B|A)/P(B)$, a ratio of the target pattern-conditional support to the average support [**?**].

distributions, 2) efficient statistical fitting of over 50 state-of-the-art theoretical distributions, and, 3) assignment of multiple items according to the proximity of values to the boundaries of discretization, a possibility supported by numerous symbolic approaches [9].

DI2 provides three data normalization techniques, which are selected for preprocessing a given variable based on its empirical distribution. The supported techniques are: 1) min-max, 2) z-score, and 3) mean. Before discretizing the data, two non-parametric tests are applied. 1) $\tilde{\chi}^2$ test [26], and 2) Kolmogorov-Smirnov goodness-of-fit test [27]. The Kolmogorov-Smirnov goodness-of-fit test can optionally be used to remove up to 5% outlier points from the observed distribution according to the matched theoretical continuous distribution. The modified observed distribution from the iteration of the Kolmogorov-Smirnov test with the best KS-statistic is used for the subsequent fitting stage. This corrections guarantees the absence of penalizations caused by abrupt yet spurious deviations driven by the selected histogram granularity.

In the aforementioned tests the observed distribution is matched with a theoretical continuous distribution[7] provided by the SciPy open-source library [28]. The binning of the distributions for the $\tilde{\chi}^2$ test is based on the number of categories the user inputs and are built using equal-frequency binning. The user can either choose the $\tilde{\chi}^2$ or the Kolmogorov-Smirnov goodness-of-fit as the *primary* fitting test.

After selecting the theoretical continuous distribution that best fits the continuous variable, DI2 proceeds with the discretization. Given a desirable number of categories (bins), multiple cut-off points are generated using the inverse cumulative distribution function of the theoretical continuous distribution. The cut-off points guarantee an approximately uniform distribution of observation per category, although empirical-theoretical distribution differences can underlie imbalances.

DI2 supports multi-item assignments by identifying border values for each category. To this end, the user can optionally also define a percentage (between 0 and 50% with 20% default) to affect the width of the borders. These borders take an intermediate value which symbolize that it belongs to both upper and lower category. Width extremes, 0% (50%) correspond to none (one) additional category assigned to every observation.

To illustrate some of the DI2 properties, we consider as an example the *breast-tissue* dataset available at the UCI machine learning repository [29], containing electrical impedance measurements in samples of freshly excised tissue from the breast. It contains 106 instances and 9 continuous variables (I0, PA500, HFS, DA, AREA, A/DA, MAX IP, DR, P).

The gathered results show the decisions placed by DI2 in the absence and presence of Kolmogorov-Smirnov optimization. For this analysis, we considered a min-max normalization for all variables, a desirable number of 5 categories per variable, and $\tilde{\chi}^2$ as the primary statistical test.

Table I shows the best fitting distribution for each continuous variable of the dataset without and with Kolmogorov-Smirnov outlier removal. Variables 'I0', 'PA500', 'A/DA', 'DR', and 'P' remained unchanged with a removal of up to 5% of outlier points. Variables 'HFS' and 'Area' produced better results in the $\tilde{\chi}^2$ test with the removal of outliers solidifying the distribution choice. Finally, the fitting choice changed for variables 'DA' and 'Max IP' under the $\tilde{\chi}^2$ test, revealing a more solid choice from the analysis of the residuals.

TABLE I: Best fitting distributions for each continuous variable, without and with Kolmogorov-Smirnov correction. Both $\tilde{\chi}^2$ (primary) and KS statistics are shown.

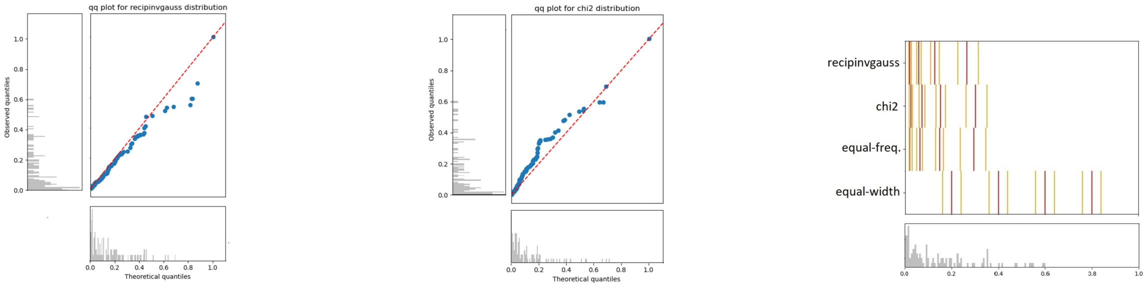
Variables	Without opt.	$\tilde{\chi}^2$	Ks	With opt.	$\tilde{\chi}^2$	Ks
I0	**alpha**	**8.8**	0.12	**alpha**	**8.8**	0.11
PA500	**exponnorm**	**2.98**	0.07	**exponnorm**	**2.98**	0.07
HFS	**foldcauchy**	**2.25**	0.07	**foldcauchy**	**1.57**	0.07
DA	**recipinvgauss**	**1.6**	0.06	**chi2**	**1.01**	0.06
Area	**frechet_r**	**0.5**	0.07	**frechet_r**	**0.25**	0.05
A/DA	**mielke**	**1.17**	0.06	**mielke**	**1.17**	0.05
Max IP	**johnsonsu**	**4.72**	0.05	**alpha**	**1.09**	0.07
DR	**johnsonsb**	**1.2**	0.05	**johnsonsb**	**1.2**	0.05
P	**genextreme**	**5.13**	0.09	**genextreme**	**5.13**	0.09

Considering variable 'DA', Figures 4.4a and 4.4b show its Q-Q (quantile-quantile) plot, offering a view on the adequacy of the statistical fitting. In this context, we depict histograms for the observed data with 100 bins (blue dots) and the best theoretical distribution picked without and with Kolmogorov-Smirnov correction (red line). A moderate improvement from Figure 4.4a to 4.4b can be detected, with the observed quantiles (blue dots) being closer to the theoretical continuous quantiles (red line). After the fitting stage, cut-off points are calculated to produce the final categories. Figure 4.4c compares different discretization options: equal-frequency and the two best fitting theoretical continuous distributions (without and with Kolmogorov-Smirnov optimization). Cut-off points are marked as red lines, and the border cut-off points in yellow. This analysis shows how critical discretization can be, determining the inclusion or exclusion of high density bins. The ability of DI2 to assign multiple items using borders can be explored by symbolic approaches to mitigate vulnerabilities inherent to the discretization process.

D. BicPAM

As surveyed, pattern-based biclustering approaches provide the unprecedented possibility to comprehensively find patterns in real-valued data with parameterizable homogeneity and guarantees of statistical significance. To be able to differentiate different clinical profiles of interest, the coherence strength and coherence assumption of biclustering solutions can be customized in accordance with the desirable patient profile. Henriques and Madeira [9] proposed BicPAM biclustering. It integrates existing principles made available by state-of-the-art pattern-based approaches with two new contributions. First, BicPAM exhaustively mines non-constant types of biclusters, including additive and multiplicative coherencies in the presence or absence of symmetries. Second, BicPAM provides strategies to effectively compose different biclustering structures. BicPAM is an ordered composition of three

[7] https://docs.scipy.org/doc/scipy/reference/stats.html



(a) Q-Q plot of empirical distribution (blue dots) against the fitted *recipinvgauss* distribution (red line).

(b) Q-Q plot of empirical distribution (blue dots) against the fitted *chi2* distribution (red line).

(c) Empirical distribution (gray bins) and corresponding cut-off points using equal-width, equal-frequency and D2I statistical fitting with and without Kolmogorov-Smirnov correction. Red and yellow lines correspond to category and border boundaries.

Fig. 2: Figure 4.4a displays how the observed distribution matched with the theoretical distribution without the Kolmogorov-optimization. Figure 4.4b displays how the observed distribution matched with the theoretical distribution with the Kolmogorov-optimization. Figure 4.4c shows where the category boundaries are depending on the technique

stages: 1) *mapping*, where BicPAM handles missing values and tackles noise as well as apply normalization and discretization methods, 2) *mining* (pattern discovery), where BicPAM discovers patterns using a pattern-based approach to biclustering, the selected pattern representation, and the selected search strategy, and 3) *closing* (post-processing), where BicPAM merges and filters biclusters.

E. Extending pattern-based biclustering searches

Guarantees of discriminative power. BicPAMS [9] is not originally prepared to assess and guarantee the discriminative power of the returning patterns. In this context, in the presence of an output variable, the search was extended to compute interestingness measures, such as lift, for each pattern under formation, and remove patterns with interestingness criteria below a parameterizable threshold.

Biclustering mixed variables. The original version of BicPAMS [9] provides two important principles for handling mixed variable data: i) categorical variables are seen as symbolic, irrespective of whether variables are nominal or ordinal, and occurring symbols per variable need to match to form a pattern (a_{ij}=c_j); and ii) numeric entries per variable belong to the same pattern if they satisfy a given coherence strength (a_{ij}=c_j+η_{ij} with $|\eta_{ij}| \leq \delta/2$). The behavior of BicPAMS was further revised to guarantee a balanced cardinality among ordinal variables, aligned with the chosen coherence strength.

F. Output: discriminative patterns of post-surgical risk

In the context of our work, a discriminative pattern of post-surgical outcomes is an association of pre-surgical variables – comprising biopathological, physiological, demographic factors – that satisfies the two following conditions:

- the pattern is supported by a statistically significant number of individuals in accordance with the characteristics of the population under study;
- the pattern is discriminative of post-surgical outcomes, such as presence/absence of post-surgical complications, ranking of post-surgical complication, survivability aspects or hospitalization needs.

The patterns will be presented in simple visual representations, either as heatmaps or parallel coordinate charts, or pattern descriptions. These are generally sufficient to guarantee their usability near healthcare professionals.

The patterns found can be characterized according to their source, including: 1) demographic and clinical variables, 2) clinical risk scores, 3) and surgical interventions (e.g. ICD_10 tabled procedures). Or be characterized in accordance with the target variable: 1) complication severity (e.g. Clavien-Dindo), 2) presence-absence of surgery-related complications in future or within specific time ranges, 3) survivability in a given period (death or alive after a given time period after surgery), 4) hospitalization needs: hospitalized period after surgery in HDU and IPO, if the patient was in intensive care, request type anesthesia, 5) provenance of patient, 6) reason for admission into the HDU or if he had to be readmitted, 7) destination after HDU/IPO, and 8) average nursery points per day (representative of effort given by nurses to a given patient).

V. Results

Considering the population monitored at IPO-Porto as a study case, the proposed approach was applied to comprehensively discover patterns able to discriminate post-surgical outcomes and additional variables of interest. This section is organized as follows. First, an initial data exploration is presented. Secondly, the experimental setting on how we varied the search for patterns is presented. Then the results for each experimental setting are presented and discussed. Finally, the statistical significance and pattern actionability are discussed.

A. Data exploration

The dataset contains a considerable amount of missings, with 11 variables reaching at least 75% missing values. The data also has 47 variables where a single value occurs for at least 70% observations. To better understand the impact each variable might have in the output patterns (such as frequent occurance in patterns), feature ranking tests were applied. $\tilde{\chi}^2$ test was applied for binary and nominal input variables. Kruskal-Wallis test was applied in the presence of ordinal and ANOVA one-way test for numeric input variables. Figure 2 provides illustrative class conditioned distributions of some of the input variables in *IPOscore* data, generally showing the difficulty of discriminating post-surgical outcomes.

B. Experimental settings

BicPAMS algorithm is used with default parameters and varying: 1) minimum lift of pattern: *lift* $\in$\{[1.3,3.0]\}, 2) minimum number of variables in the pattern: *variables* $\in$ \{3,8\}, 3) target classes: i) *clavien-Dindo* $\in$ \{I,II,IIIa, IIIb,IVa,IVb,V\}, ii)*post-surgical complication* $\in \{yes, no\}$, iii) *days at HDU* $\in$ \{$\leq$1,]1,2], $>$ 2\}, and $\in$ \{$\leq$1,]1,4], $>$ 4\}, iv) *1-year death* $\in$ \{yes,no\}, v) *request type anesthesia* $\in$ \{associated pathology, surgical complexity\}, vi) *provenance* $\in$ \{nursery, intensive care unit, unscheduled service\}, vii) *HDU reason for admission* $\in$ \{post-surgery, heart, respiratory, age, another pathology, co-morbidities, discharge from intensive care, hemodynamic instability, bleeding, post-op reoperation, ischemic stroke, sepsis/septic shock/BMD\}, viii) *days at IPO* $\in$ \{$<$ 7, [7, 10], $>$10\}, ix)*ICU* $\in$ \{yes\}, x) *destination after HDU* $\in$ \{intensive care unit\}, xi) *average nursery points per day* $\in$ \{<60, 60$\leq$\}, xii) *HDU readmission* $\in$ \{yes\}, xiii) *destination after IPO* $\in$ \{death\}, xiv)*moment of death* $\in$ \{[0,30[, [30-60[, [60, 365]\}, 4) coherence strength (δ=$\bar{A}/|\mathcal{L}|$: $|\mathcal{L}| \in \{3,4,5\}$), 5) decreasing support until $|\mathcal{B}|$ dissimilar biclusters are found: $|\mathcal{B}| \in$ \{2,10,50,100,200,1000\}, 6) noise: 0% and up to 30% noisy elements allowed, 7) coherence assumptions: constant and order-preserving, and 8) iterations: between one and three search iterations were considered.

C. Clustered and integrative setting results

Tables III synthesizes the results for presence/absence of post-surgical complication produced by biclustering *IPOscore* data with BicPAMS [9] in the clustered setting. Table IV synthesizes the results for Clavien-Dindo classes, presence of post-surgery complication, days spent in HDU, and days spent in IPO, in the integrative setting using DI2 discretization. These results confirm the potentialities listed before, BicPAMS was able to efficiently and comprehensively find a large number of homogeneous, dissimilar and statistically significant patterns able to discriminate post-surgical outcomes.

One can check, for instance, in the first row of Table III, that among a total of 153 discovered discriminative biclusters for the major clinical data variables, we found that 49 of them are statistically significant (p-value lower that 0.1%). Given these 49 biclusters, there are approximately 86 patients per bicluster on average ($\mu(|I|)$), 3 variables per bicluster on average ($\mu(|J|)$) when considering a constant assumption ($|\mathcal{L}|$=3 and $\delta \in [0, \vec{A}/|\mathcal{L}|]$), and a perfect quality (no noise).

These results further show the impact of: tolerating noise; placing different coherence assumptions (such as the order-preserving assumption); and parameterizing coherence strength ($\delta \propto \frac{1}{|\mathcal{L}|}$) on the biclustering solution.

BicPAMS [9] was also applied to find less-trivial yet relevant patterns of surgical risk, patterns with order-preserving coherence assumptions. Figure 3c depicts order-preserving patterns for Clavien-Dindo I.

Each bicluster shows a unique pattern of performance. For instance, the constant bicluster from Figure 5a reveals a group of 61 patients who coherently encountered high physiological score and morbidity risk (P-Possum), and medium average risk of reoperation (corresponding to the pattern \{2,2,1\} using 3 bins where 0 denotes low risk score and 2 a high risk score) for Clavien-Dindo type V, showing us that patients who follow this pattern end up dying in surgery.

These results motivate the relevance of finding both constant and order-preserving biclusters to find coherent factors propelling post-surgical status and hospitalization needs for a statistically significant group of individuals. One can check that a bicluster considers both identical physiological values or risk scores values (where lines converge) and more loosely similar values (where lines diverge). The profile of the patient in a specific bicluster can be further analyzed to further understand its influence on the resulting performance.

A closer analysis of the found discriminative patterns shows their robustness to the item-boundaries problem: slightly deviating limits to the expected limit are not excluded from the bicluster. This allows the discovery of patterns without the drawbacks of the traditional discrete views.

No patterns are presented for the ACS procedures partition. Despite multiple runs of BICPAMS with different criteria applied, no patterns were found. The criteria varied for pattern discovery was: discriminative power, lower number of variables in found biclusters, number of biclusters, bicluster type, noise tolerated.

D. Statistical Significance

As previously mentioned, Tables III and IV show the ability of the target biclustering searches to find statistically significant relations within *IPOscore* data. A bicluster is statistically significant if the number of individuals sharing the given pattern is unexpected [8]. Figure 6 provides two scatter plots of the statistical significance (vertical axis) and area $|I|$x$|J|$ (horizontal axis) of constant type biclusters for each target variable considered in the clustered setting, 6a) Post-surgical complication, 6b) Clavien-Dindo, 6. This analysis suggests the presence of a soft correlation between size and statistical significance. A few biclusters with loose statistical significance (left upper dots) can be discarded to not incorrectly bias clinical decisions.

E. Pattern actionability

The found patterns, help healthcare professionals taking decisions to better handle patients who follow the same patterns. For example, Figure 3a suggests malnutrition that

TABLE II: Properties of the biclustering solutions found in the three partitions of clinical variables for presence/absence of **post-surgical complications** classes using BicPAMS (cf. experimental setting).

	configuration		Clinical variables					ICD_10				Scores (%)				
	Assumption	quality	$\lvert L\rvert$	#bics	p-value <0.001	$\mu(\lvert J\rvert)$ $\pm\sigma(\lvert J\rvert)$	$\mu(\lvert I\rvert)$ $\pm\sigma(\lvert I\rvert)$	#bics	p-value <0.001	$\mu(\lvert J\rvert)$ $\pm\sigma(\lvert J\rvert)$	$\mu(\lvert I\rvert)$ $\pm\sigma(\lvert I\rvert)$	$\lvert L\rvert$	#bics	p-value <0.001	$\mu(\lvert J\rvert)$ $\pm\sigma(\lvert J\rvert)$	$\mu(\lvert I\rvert)$ $\pm\sigma(\lvert I\rvert)$
absence	Constant	100%	3	341	113	5.1±1.7	144.9±111.1	68	48	2.7±0.7	5.0±5.0	3	8	6	8.2±1.1	34.0±7.8
	Constant	70%	3	278	135	5.4±1.9	121.7±118.3	67	47	2.7±0.7	5.0±5.1	3	5	4	7.8±1.5	37.3±7.9
	Constant	70%	4	136	55	4.9±1.5	154.3±154.9	–	–	–	–	5	6	6	7.7±2.4	33.2±4.3
	Constant	70%	5	358	120	5.3±1.9	150.7±140.4	–	–	–	–	7	16	16	5.8±1.9	27.8±10.0
	Order-preserving	100%	–	98	89	5.7±1.2	68.2±81.8	63	42	2.7±0.7	5.5±5.2	–	26	26	4.6±0.7	29.2±4.9
	Order-preserving	70%	–	81	63	5.8±1.8	86.0±87.6	63	42	2.7±0.7	5.5±5.2	–	26	26	4.7±0.8	29.2±5.0
presence	Constant	100%	3	94	29	2.9±0.9	62.7±24.9	30	24	3.1±0.9	11.3±10.9	3	4	4	3.8±0.8	58.5±1.5
	Constant	70%	3	113	34	3.4±1.4	64.1±27.4	30	24	3.0±0.9	11.5±10.8	3	5	5	3.8±1.2	60.2±2.2
	Constant	70%	4	170	52	3.9±1.7	74.7±32.0	–	–	–	–	5	6	6	3.7±1.7	47.7±12.1
	Constant	70%	5	186	61	3.6±1.6	57.9±34.0	–	–	–	–	7	7	7	3.4±1.0	45.6±4.0
	Order-preserving	100%	–	42	39	3.0±0.8	125.4±51.1	15	15	2.7±0.8	15.2±12.3	–	7	7	3.9±0.3	29.7±7.5
	Order-preserving	70%	–	73	62	3.4±1.1	90.1±61.2	16	16	2.7±0.8	15.0±11.9	–	6	6	4.0±0.6	30.2±6.8

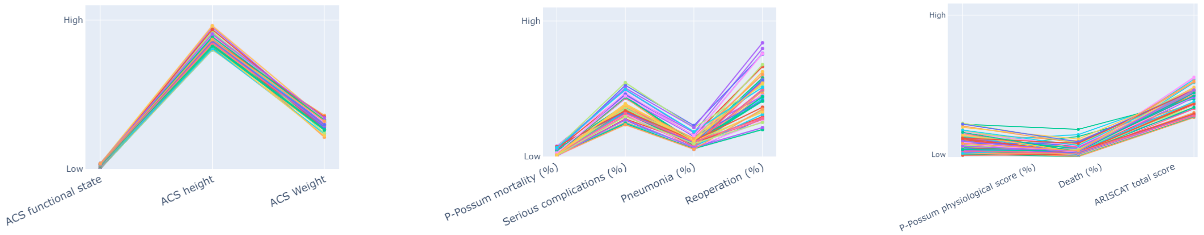
TABLE III: **Clavien-Dindo classes**, **presence of post-surgery complication**, **days spent in HDU** and **days spent in IPO** using BicPAMS with DI2 discretization.

	Assumption	quality	$\lvert L\rvert$	$\lvert C\rvert$	Lift	#bics	p-value <0.001	$\mu(\lvert J\rvert)$ $\pm\sigma(\lvert J\rvert)$	$\mu(\lvert I\rvert)$ $\pm\sigma(\lvert I\rvert)$		Lift	#bics	p-value <0.001	$\mu(\lvert J\rvert)$ $\pm\sigma(\lvert J\rvert)$	$\mu(\lvert I\rvert)$ $\pm\sigma(\lvert I\rvert)$
Clavien-Dindo type I	Constant	70%	3	3	2	101	100	3.66 ± 0.87	149.6 ± 18.43	Presence of Post-surgery comp.	1.5	138	138	4.09 ± 1.19	169.41 ± 24.46
	Constant	70%	3	8	1.3	39	39	9.41 ± 1.61	133.67 ± 41.78		1.3	11	11	9.18 ± 1.33	152.81 ± 38.24
	Constant	70%	4	3	2	124	103	3.69 ± 0.83	131.38 ± 25.38		1.5	112	112	3.76 ± 1.02	150.99 ± 20.38
	Constant	70%	4	8	1.3	20	20	9.15 ± 1.49	107.5 ± 18.61		1.3	8	8	9.12 ± 0.93	111.88 ± 21.83
	Constant	70%	5	3	2	155	127	3.56 ± 0.68	110.12 ± 18.61		1.5	160	157	3.58 ± 0.84	107.95 ± 17.24
	Constant	70%	5	8	1.3	35	35	9.26 ± 1.75	75.68 ± 17.66		1.3	6	6	9.0 ± 0.82	76.33 ± 16.23
Clavien-Dindo type II	Constant	70%	3	3	1.5	116	105	3.78 ± 1.05	154.79 ± 23.54	HDU days < 1	1.7	50	49	5.65 ± 1.64	150.47 ± 16.59
	Constant	70%	3	8	1.3	16	16	8.62 ± 1.11	130.75 ± 22.48		1.3	22	22	9.54 ± 1.50	150.64 ± 45.67
	Constant	70%	4	3	1.5	60	53	3.68 ± 0.80	144.55 ± 24.09		1.7	36	36	5.53 ± 1.58	139.83 ± 20.78
	Constant	70%	4	8	1.3	10	10	8.7 ± 0.78	90.5 ± 13.46		1.3	13	13	9.31 ± 1.94	130.69 ± 28.47
	Constant	70%	5	3	1.5	115	96	3.39 ± 0.67	113.14 ± 26.39		1.7	77	52	3.73 ± 0.76	159.03 ± 25.61
	Constant	70%	5	8	1.3	15	15	8.47 ± 0.62	68.2 ± 11.08		1.3	10	10	9.5 ± 1.86	113.8 ± 25.35
Clavien-Dindo type III.a	Constant	70%	3	3	2	95	92	3.42 ± 0.74	165.53 ± 22.34	HDU days $1 - 2$	1.3	22	22	3.95 ± 1.11	123.32 ± 10.60
	Constant	70%	3	8	1.3	11	11	8.45 ± 0.65	130.45 ± 20.89		1.3	5	5	10.0 ± 1.67	146.0 ± 13.1
	Constant	70%	4	3	2	120	108	3.22 ± 0.46	137.93 ± 30.14		1.3	46	40	4.53 ± 1.18	93.63 ± 7.04
	Constant	70%	4	8	1.3	9	9	8.78 ± 0.63	81.22 ± 6.27		1.3	16	16	9.0 ± 1.12	93.06 ± 19.05
	Constant	70%	5	3	2	101	76	3.18 ± 0.39	127.61 ± 32.95		1.3	107	91	3.79 ± 0.82	81.58 ± 12.70
	Constant	70%	5	8	1.3	8	8	8.64 ± 0.70	65.38 ± 11.97		1.3	45	45	8.62 ± 0.87	63.49 ± 13.29
Clavien-Dindo type III.b	Constant	70%	3	3	2	37	32	4.44 ± 1.06	141.28 ± 16.67	HDU days > 2	1.5	27	26	4.57 ± 1.50	152.73 ± 24.74
	Constant	70%	3	8	1.3	32	32	8.97 ± 1.21	121.38 ± 34.44		1.2	5	5	9.2 ± 1.47	175.6 ± 20.44
	Constant	70%	4	3	2	66	56	3.83 ± 0.72	119.73 ± 13.92		1.5	113	105	3.2 ± 0.51	141.07 ± 25.38
	Constant	70%	4	8	1.3	13	13	9.54 ± 1.64	117.77 ± 25.93		1.2	5	5	8.4 ± 0.8	109.6 ± 25.34
	Constant	70%	5	3	2	125	94	3.47 ± 0.66	88.52 ± 19.33		1.5	111	104	3.42 ± 0.64	90.15 ± 17.94
	Constant	70%	5	8	1.3	2	2	11.0 ± 0.0	147.0 ± 0.0		1.2	6	6	8.66 ± 0.74	74.83 ± 15.02
Clavien-Dindo type IV.a	Constant	70%	3	3	2	80	79	3.24 ± 0.53	210.09 ± 31.65	IPO days < 7	2	288	284	4.27 ± 1.36	192.96 ± 41.15
	Constant	70%	3	8	1.3	14	14	9.07 ± 1.39	155.43 ± 25.87		1.3	18	18	9.83 ± 1.67	181.28 ± 37.47
	Constant	70%	4	3	2	79	69	3.36 ± 0.59	173.61 ± 24.37		2	73	57	3.50 ± 0.77	173.96 ± 30.36
	Constant	70%	4	8	1.3	13	13	9.23 ± 1.12	110.15 ± 21.83		1.3	24	24	9.16 ± 1.49	114.5 ± 32.69
	Constant	70%	5	3	2	55	44	3.36 ± 0.68	153.79 ± 17.80		2	62	52	3.65 ± 0.96	163.17 ± 16.80
	Constant	70%	5	8	1.3	22	22	9.14 ± 1.49	75.18 ± 15.25		1.3	10	10	9.5 ± 1.8	113.8 ± 25.36
Clavien-Dindo type IV.b	Constant	70%	3	3	2	57	57	3.65 ± 0.85	195.73 ± 31.48	IPO days $7 - 10$	1.7	48	46	4.83 ± 1.46	165.0 ± 25.83
	Constant	70%	3	8	1.3	7	7	9.41 ± 1.12	166.0 ± 32.98		1.3	3	3	9.0 ± 0.82	161.0 ± 34.32
	Constant	70%	4	3	2	76	69	3.45 ± 0.77	159.72 ± 21.09		1.7	72	72	3.80 ± 0.93	122.68 ± 16.56
	Constant	70%	4	8	1.3	21	21	8.85 ± 1.03	95.19 ± 21.45		1.3	2	2	8.5 ± 0.5	103.0 ± 18.0
	Constant	70%	5	3	2	118	85	3.11 ± 0.34	120.4 ± 25.13		1.7	71	66	3.66 ± 0.78	109.51 ± 19.77
	Constant	70%	5	8	1.3	34	34	9.14 ± 1.19	67.18 ± 9.94		1.3	5	5	8.4 ± 0.48	75.6 ± 8.8
Clavien-Dindo type V	Constant	70%	3	3	2	84	83	3.36 ± 0.72	194.18 ± 31.79	IPO days > 10	1.5	9	9	5.0 ± 1.49	195.22 ± 18.91
	Constant	70%	3	8	1.3	11	11	9.45 ± 1.62	160.90 ± 23.58		1.3	17	17	9.23 ± 1.51	163.47 ± 30.44
	Constant	70%	4	3	2	58	52	3.29 ± 0.66	152.02 ± 28.24		1.5	33	33	3.69 ± 0.99	155.84 ± 23.30
	Constant	70%	4	8	1.3	8	8	9.13 ± 1.17	99.5 ± 13.51		1.3	6	6	8.83 ± 0.89	108.0 ± 28.85
	Constant	70%	5	3	2	151	135	3.24 ± 0.57	102.36 ± 24.38		1.5	35	35	3.74 ± 0.93	127.51 ± 19.03
	Constant	70%	5	8	1.3	16	16	8.94 ± 1.30	63.94 ± 10.84		1.3	10	10	9.1 ± 1.13	78.7 ± 12.81

can be tackled with specialized programs before surgery, other previous addressable comorbidities can also be subjected to pre-habilitation. Patterns such as 3bb are helpful logistic-wise as they identify groups of patients susceptible to longer monitoring periods after surgery, showing the possibility to reserve beds in the HDU. Finally, patterns in Figures 5b and 5c help professionals identifying the possible nature of post-surgical complications (Clavien-Dindo) and, accordingly, revise surgical procedures and modes of pre- and post-operative care.

VI. Conclusion

This work proposes a comprehensive set of principles on how to mine discriminative patterns of post-surgical outcomes from heterogeneous oncological data with guarantees of usability. State-of-the-art contributions on pattern-based biclustering are extended towards this end, offering the unprecedented possibility to comprehensively discover non-

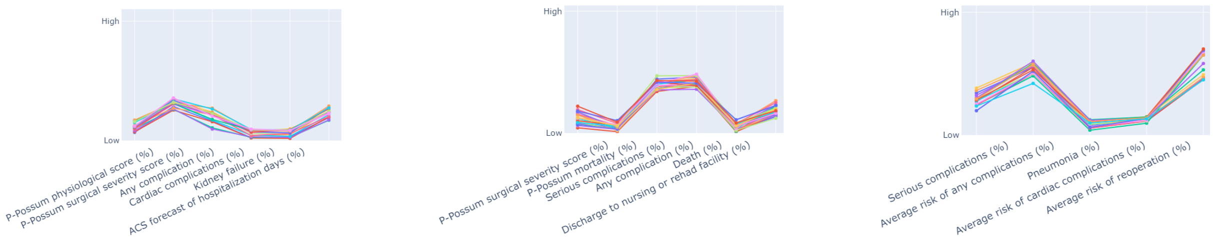


(a) Discriminative pattern composed of demographic and physiological variables: patients in a good and independent functional state, above average height, and average weight. *Lift* = 1.71 and *p-value* = 3.58×10^{-5}

(b) Discriminative pattern of high hospitalization length, quality 70%, $|\mathcal{L}|$=3: low risk of mortality, medium risk of serious complications, low risk of pneumonia, medium risk of reoperation. *Lift* = 2.18 and *p-value* = 3.21×10^{-172}.

(c) Discriminative pattern of patients with Clavien-Dindo severity I: low physiological score (P-Possum), less susceptible to death, and medium ARISCAT total score. *Lift* = 2.05 and *p-value* = 3.89×10^{-194}.

Fig. 3: Constant pattern discriminative of Clavien-Dindo III.b (fig. 3a) and Clavien-Dindo I (fig. 3c), and order-preserving pattern discriminative of high hospitalization length (fig. 3b).

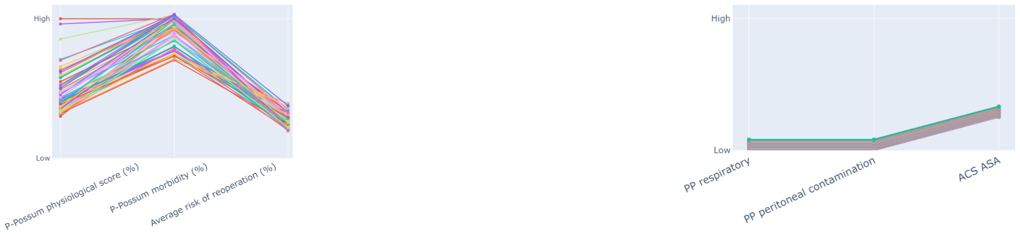


(a) Discriminative pattern of patients with no post-surgical complication: low physiological score, medium surgical severity score, lower complication risk, almost no risk of cardiac complications and kidney failure, and medium risk of high hospitalization length. *Lift* = 1.73 and *p-value* = 1.19×10^{-25}.

(b) Discriminative pattern of patients who died within 1 year of surgery: low surgical severity score, low risk of mortality (P-Possum), medium susceptibility to serious complications, low death probability, and slightly higher probability of rehab needs. *Lift* = 2.01 and *p-value* = 7.07×10^{-36}.

(c) Discriminative pattern of patients who stayed between 1 and 4 days in the HDU: medium risk of serious complications, average risk for any complication, low probability of pneumonia, average risk of cardiac complications, and medium average risk of reoperation. *Lift* = 2.05 and *p-value* = 4.63×10^{-65}.

Fig. 4: Illustrative discriminative patterns of different post-surgical outcomes: no post-surgical complication (a), 1-year death (b) and]1,4] hospitalization-length (c). Patterns 4a.



(a) Discriminate pattern of Clavien-Dindo grade V, quality 100%, $|\mathcal{L}|$=3: medium physiological score, high morbidity, below average risk of average risk of reoperation. *Lift* = 2.11 and *p-value* = 9.28×10^{-20}.

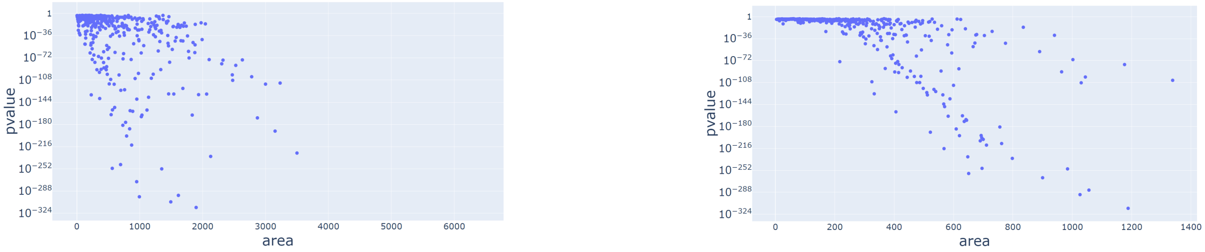
(b) Discriminative pattern of absent post-surgical complication, quality 70%, $|\mathcal{L}|$=3: patient with no dyspnoea, no peritoneal contamination, and patient with mild systemic disease. *Lift* = 1.31 and *p-value* = 1.49×10^{-4}.

Fig. 5: Example of constant patterns of surgical risk found within the IPOscore dataset. Pattern 5a correlates with healthy patients whose surgery went wrong in some way. Patterns 5b and 5c show both ends of the post-surgical complication spectrum: patients with high mortality scores and patients with regular values in clinical variables. Pattern 5d shows that patients with a higher risk of developing post-surgical complications need to be observed longer after surgery (in the HDU).

trivial, yet actionable and statistically significant associations between cancer morphology, individual's profile, undertaken surgery and post-operatory outcomes. It also proposes a fully autonomous, non-parametric and prior-free discretization method, DI2, for numerical variables with arbitrarily skewed distributions.

Results confirm the unique role of biclustering in finding relevant discriminative patterns sensitive to highly variable physiology and biopathological traits of individuals, as well as the singularity of undertaken surgeries and post-surgical care. In particular, the search for non-constant patterns (order-preserving coherence assumptions) show a delineate ability to tolerate individual differences, while still guaranteeing the coherence and interpretability of the target patterns.

Results further show evidence of the ability to comprehensively unveil actionable and statistically significant patterns of post-surgical outcomes, thus providing a trustworthy context for healthcare professionals to support the design of surgical interventions, pre-surgical and post-surgical care.



(a) Post-surgical complication

(b) Clavien-Dindo classification

Fig. 6: Statistical significance versus size of constant patterns.

References

[1] M. Derogar, N. Orsini, O. Sadr-Azodi, and P. Lagergren, "Influence of major postoperative complications on health-related quality of life among long-term survivors of esophageal cancer surgery," *Journal of Clinical Oncology*, vol. 30, no. 14, pp. 1615–1619, 2012.

[2] A. Kumar and H. Anjomshoa, "A two-stage model to predict surgical patients' lengths of stay from an electronic patient database," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 848–856, 2018.

[3] M. Whiteley, D. Prytherch, B. Higgins, P. Weaver, W. Prout, and S. Powell, "Possum and portsmouth possum for predicting mortality," *British Journal of Surgery*, vol. 85, pp. 1217–1220, 1998.

[4] K. Y. Bilimoria, Y. Liu, J. L. Paruch, L. Zhou, T. E. Kmiecik, C. Y. Ko, and M. E. Cohen, "Development and evaluation of the universal acs nsqip surgical risk calculator: a decision aid and informed consent tool for patients and surgeons," *Journal of the American College of Surgeons*, vol. 217, no. 5, pp. 833–842, 2013.

[5] J. Canet, L. Gallart, C. Gomar, G. Paluzie, J. Vallès, J. Castillo, S. Sabate, V. Mazo, Z. Briones, J. Sanchis *et al.*, "Prediction of postoperative pulmonary complications in a population-based surgical cohort," *The Journal of the American Society of Anesthesiologists*, vol. 113, no. 6, pp. 1338–1350, 2010.

[6] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation," *Journal of Clinical Epidemiology*, vol. 40, no. 5, pp. 373–383, 1987.

[7] R. Henriques, C. Antunes, and S. C. Madeira, "A structured view on pattern mining-based biclustering," *Pattern Recognition*, vol. 4, no. 12, pp. 3941—-3958, 2015.

[8] R. Henriques and S. C. Madeira, "Bsig: evaluating the statistical significance of biclustering solutions," *Data Mining and Knowledge Discovery*, vol. 32, no. 1, pp. 124–161, 2018.

[9] R. Henriques, F. L. Ferreira, and S. C. Madeira, "Bicpams: software for biological data analysis with pattern-based biclustering," *BMC bioinformatics*, vol. 18, no. 1, p. 82, 2017.

[10] R. Veroneze and F. J. Von Zuben, "Efficient mining of maximal biclusters in mixed-attribute datasets," *arXiv preprint arXiv:1710.03289*, 2017.

[11] S. Kaiser and F. Leisch, "A toolbox for bicluster analysis in r," 2008.

[12] S. Barkow, S. Bleuler, A. Prelić, P. Zimmermann, and E. Zitzler, "Bicat: a biclustering analysis toolbox," *Bioinformatics*, vol. 22, no. 10, pp. 1282–1283, 2006.

[13] R. Henriques, C. Antunes, and S. C. Madeira, "A structured view on pattern mining-based biclustering," *Pattern Recognition*, vol. 48, no. 12, pp. 3941–3958, 2015.

[14] A. Serin and M. Vingron, "Debi: Discovering differentially expressed biclusters using a frequent itemset approach," *Algorithms for Molecular Biology*, vol. 6, no. 1, pp. 1–12, 2011.

[15] Y. Okada, W. Fujibuchi, and P. Horton, "A biclustering method for gene expression module discovery using a closed itemset enumeration algorithm," *IPSJ Digital Courier*, vol. 3, pp. 183–192, 2007.

[16] R. Martinez, C. Pasquier, and N. Pasquier, "Genminer: mining informative association rules from genomic data," in *2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007)*. IEEE, 2007, pp. 15–22.

[17] G. Pandey, G. Atluri, M. Steinbach, C. L. Myers, and V. Kumar, "An association analysis approach to biclustering," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 677–686.

[18] G. Atluri, J. Bellay, G. Pandey, C. Myers, and V. Kumar, "Discovering coherent value bicliques in genetic interaction data," in *Proceedings of 9th International Workshop on Data Mining in Bioinformatics (BIOKDD'10)*, 2000, p. 47.

[19] R. Gupta, N. Rao, and V. Kumar, "Discovery of error-tolerant biclusters from noisy gene expression data," 2011.

[20] R. Henriques and S. C. Madeira, "Bic2pam: constraint-guided biclustering for biological data analysis with domain knowledge," *Algorithms for Molecular Biology*, vol. 11, no. 1, p. 23, 2016.

[21] ——, "Bicnet: Flexible module discovery in large-scale biological networks using biclustering," *Algorithms for Molecular Biology*, vol. 11, no. 1, pp. 1–30, 2016.

[22] D. G. Altman, "Categorizing continuous variables," *Wiley StatsRef: Statistics Reference Online*, 2014.

[23] C. Bennette and A. Vickers, "Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents," *BMC medical research methodology*, vol. 12, no. 1, p. 21, 2012.

[24] S.-C. Liao and I.-N. Lee, "Appropriate medical data categorization for data mining classification techniques," *Medical informatics and the Internet in medicine*, vol. 27, no. 1, pp. 59–67, 2002.

[25] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[26] R. Lowry, "Concepts and applications of inferential statistics," 2014.

[27] T. Gonzalez, S. Sahni, and W. R. Franta, "An efficient algorithm for the kolmogorov-smirnov and lilliefors tests," *ACM TOMS*, vol. 3, no. 1, pp. 60–64, 1977.

[28] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[29] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.