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

Leonardo Duarte Rodrigues Alexandre

Thesis to obtain the Master Degree in
Information Systems and Computer Engineering

Supervisor(s): Prof. Rui Miguel Carrasqueiro Henriques
Prof. Rafael Sousa Costa

Examination Committee

Chairperson: Prof. Francisco António Chaves Saraiva de Melo
Supervisor: Prof. Rui Miguel Carrasqueiro Henriques
Member of the Committee: Prof. Cátia Raquel Jesus Vaz

January 2021

Acknowledgments

Quero começar por agradecer aos meus orientadores, Rafael Costa e Rui Henriques, e, Doutor Lúcio Santos, pela oportunidade incrivel que me deram de poder trabalhar neste projecto e todo o apoio que me deram ao longo do mesmo.

Quero agradecer a todos os meus amigos em especial aos que conheci no meu percurso académico no ISEL e no IST, que ao lado dos mesmos partilhei muitos momentos intensos de trabalho e felicidade. Quero agradecer ao Cláudio Sardinha, Inês Santos, Pedro Palma e Sérgio Tavares que me apoiaram bastante nos últimos meses. E por fim gostaria de agradecer ao António Hermenegildo, Pedro Palma e Rui Gerardo que sempre tiveram um grande impacto na minha vida.

Para último guardo o maior agradecimento para a minha mãe, Eugénia Rodrigues, pelo apoio e amor incondicional que sempre me deu neste e noutros percursos da minha vida.

This work was supported by by Fundação para a Ciência e a Tecnologia (FCT), through IDMEC, under LAETA, project UIDB/50021/2020 and project IPOscore DSAIPA/DS/0042/2018.

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) enabling a more robust pattern discovery on non-identically distributed variables. 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.

Keywords: surgical risk, biclustering, oncology, post-surgical complications, discriminative pattern mining, data analysis, biostatistics, data mining, software tool

Resumo

Compreender os riscos da realização de procedimentos cirúrgicos é essencial para personalizar os protocolos preparatórios, de intervenção e pós-cirurgico. Este conhecimento é fundamental na área da oncologia, dada a natureza das intervenções, o perfil frágil dos pacientes com comorbilidades e exposição a quimioterapia, e o possível reaparecimento do cancro. Apesar da sua relevância, a descoberta de padrões discriminativos de risco pós-cirúrgico apresenta alguns desafios: 1) o perfil individual fisiológico e demográfico, bem como os cuidados pós-cirúrgicos diferenciados do paciente, 2) a crescente alta dimensionalidade e natureza heterogenea dos dados disponíveis 3) a necessidade de aprender com as populações onde os tumores têm diferenças significativas e os indivíduos realizam procedimentos cirúrgicos únicos (dados estruturalmente esparsos), 4) a necessidade de foco em padrões não triviais de risco cirúrgico, ao mesmo tempo que garantem sua significância estatística e poder discriminativo dos resultados pós-cirúrgicos, e 5) a falta de interpretabilidade e capacidade de ação das abordagens atuais.

Esta tese propõe o uso de *biclustering*, a descoberta de grupos de indivíduos correlacionados em subconjuntos de variáveis, devido às suas propriedades únicas que satisfazem os desafios previamente mencionados, e também um método de discretização, DI2 (Distribution Discretizer), tornando a procura de padrões mais robusta, quando na presença de variáveis não identicamente distribuidas. Os padrões encontrados oferecem uma visão abrangente sobre como o perfil do paciente, a histopatologia do cancro e os procedimentos cirúrgicos envolvidos com a capacidade de determinar: 1) complicações pós-cirúrgicas, 2) sobrevivencia, e 3) necessidades de hospitalares.

Os resultados obtidos confirmam o papel fundamental do *biclustering* em encontrar de forma abrangente padrões interpretáveis, com capacidade de ação e estatisticamente significativos. Os padrões encontrados podem ajudar os profissionais de saúde a estabelecerem protocolos especializados de pré-habilitação e decisões de cuidados hospitalares.

Keywords: risco cirúrgico, biclustering, oncologia, complicações pós-cirurgicas, padrões discriminativos, análise de dados, bioestatística, ferramenta de software

Contents

List of Tables

List of Figures

Nomenclature

American College of Surgeons National Surgical Quality Improvement Program.

American College of Surgeons National Surgical Quality Improvement Program Surgical Risk Calculator.

Artificial Neural Networks.

Analysis of variance.

Assess Respiratory Risk in Surgical Patients in Catalonia.

Society of Anaesthesiologists.

Cheng and Church.

Data Envelopment Analysis.

Data Discretizer.

Frequent Itemset Mining.

High Dependency Unit.

International Statistical Classification of Diseases and Related Health Problems.

Intensive care unit.

Portuguese Oncology Institute.

Inter-Quartiles Range.

k-Nearest Neighbors.

Sum of Squares of the Layer.

Nursing Activities Score.

Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity.

Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity.

Chapter 1

Introduction

Cancer is a disease[1] with genetic and/or epigenetic precedence [26, 61]. It is primarily caused by functional or transcriptional changes that control how our cells function. Normally cells grow and divide to form new cells as the body needs them but when they become older or damaged apoptosis is activated to program their death and other existing cells take their place. Nonetheless, these silent changes together lead to the accumulation of cells that create their microenvironment, gaining their independence, and can invade nearby tissue. The behavior of those cells will depend on the place they started, their cell type and other host conditions, the human health state. These extra cells can divide without stopping and may form a neoplasm. This neoplasm can be benign, but, they can also be malignant which means they can spread into, or invade nearby tissues.

In Portugal, according to data retrieved by the 2018 National Cancer Registry[2], in 2018, the number of new cancer cases was 58 199 and the number of deaths from cancer was 28 960.

Compared with other areas of biological research, the science of molecular oncology is a recent arrival. It began near the year 1975 and, since then, the access to demographic, clinical and molecular data of patients undertaking oncological surgical procedures is growing [58, 82]. This is important because as more data is collected, more analysis can be done. For example, one way to clinically address cancer is to operate the patient and remove the affected cells. This can lead to post-operative complications due to the procedure, physiological response, or external factors (such as infections). If information about the patient is collected, a comparison can be made with previous patients, and the doctors might be able to predict said negative impacts. But this is not an easy process, it requires the data to be collected, consolidated, analyzed and visualized to pinpoint patterns.

One way to determine if a previous subset of conditions of a subset of patients was frequent, and led to a negative impact, is to search for patterns using pattern mining techniques, and, test if they are discriminative. But despite the relevance of discriminative pattern mining approaches, the discovery of patterns discriminating surgical outcomes and other variables of interest 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. Sec-

[1] https://www.cancer.gov/about-cancer/understanding/what-is-cancer, accessed January 2020

[2] https://gco.iarc.fr/today/data/factsheets/populations/620-portugal-fact-sheets.pdf, accessed January 2020

ond, 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.

Due to the nature of interventions, cancer recurrence, and fragile profile of patients (generally debilitated by the tumor effects and common need for chemotherapy) can cause small to life-threatening post-surgical complications [25, 44]. Thus, 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, and other variables and outcomes of interest.

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. Although biclustering has been largely used in the biological domain, its potential to assess surgical and post-surgical care remains untapped. To this end, we provide illustrative patterns of surgical risk, with a particular emphasis on their sensitivity to the unique clinical record of the individuals and ability to discriminate outcomes and variables of interest. We propose a structured view on why, when and how to use biclustering for their effective and efficient discovery. We show how each of the identified challenges can be addressed by extending state-of-the-art principles on pattern-based biclustering. Due to a considerable number of data mining approaches for biomedical data analysis, including state-of-the-art associative models, requiring a form of data discretization and despite multiple discretization approaches already have been proposed, they generally work under a strict set of statistical assumptions which are arguably insufficient to handle the inherent heterogeneity associated with clinical and molecular variables. In addition, an increasing number of symbolic approaches in bioinformatics support the assignment of multiple items for values occurring near discretization boundaries for superior robustness. We propose a fully autonomous, non-parametric and prior-free discretization method, DI2, supporting multi-item assignments. Finally, we guarantee the actionability, usability and statistical significance of the target patterns, thus providing a trustworthy context for healthcare professionals to design pre-surgical, surgical and post-surgical care protocols.

The thesis is structured as follows. Background chapter 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. Related work surveys state-of-the-art pattern discovery and other approaches on analyzing oncological data, it also surveys the use of the traditional risk scores in various cohort studies. Solution chapter describes the approached solution, the data used and its preprocessing, the algorithm used, the post-processing and visualization of the results. Results and Discussion chapter presents the results obtained, their interpretation, and actionability. The Conclusion chapter presents concluding remarks synthesized. Finally, Future work chapter presents the work yet to be done.

Chapter 2

Background

This chapter lays out fundamental concepts of pattern mining, biclustering and pattern-based biclustering, while also providing a brief introduction to post-surgical traditional risk scores for patients undergoing surgery. We will start by giving an introduction to the concept of a tabular dataset followed by the section Traditional Risk Scores presenting the various variables contained within our tabular dataset. Then in section Pattern Mining we will give an introduction to concepts such as transactional dataset and frequent itemsets, and present some of the best known algorithms. Section Biclustering introduces the concept of biclustering, the various types and structures of biclusters, and some algorithmic approaches to biclustering. Finally, Pattern-based biclustering section presents the concept of combining pattern mining and biclustering to search for patterns.

Before we dive into the sections introduced let us establish some main concepts. A dataset is a collection of data. In the case of tabular data, a dataset corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given sample of the dataset in question. A dataset can be analyzed and the analysis may belong to one of two classes: 1) univariate, and 2) multivariate / bivariate.

Univariate data consists of data with only one variable. For example the height of a group of people. In this type of data, conclusions can be drawn by calculating the mean, median, mode, and the dispersion of the data (range, minimum, maximum, quartiles, variance, and standard deviation).

When considering two variables, bivariate, or more, multivariate, the analysis is done to find out the relationship between the two or more variables. If we consider a supervised scenario, one variable is independent and the others are considered dependent. If we consider an unsupervised scenario, all the variables are considered dependent.

The data in this work is in the form of a tabular dataset and its structure can be observed in Table 2.1. Each column represents a variable, each row represents a patient, and a_{ij} represents the value for j variable of i patient.

Table 2.1: Tabular dataset structure

	Variable 1	...	Variable j	...	Variable m
Patient 1	a_{11}	...	a_{1j}	...	a_{1m}
Patient ...	...	...	...	...	...
Patient i	a_{i1}	...	a_{ij}	...	a_{im}
Patient ...	...	...	...	...	...
Patient n	a_{n1}	...	a_{nj}	...	a_{nm}

Table 2.2: Tabular dataset example

	y_1	y_2	y_3	y_4
Patient 1	8	4	2	1
Patient 2	5	4	3	1
Patient 3	8	4	3	1

2.1 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. In this subsection, four clinical risk scores contained within our data will be introduced as they play a fundamental role in this work: 1) P-POSSUM (Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity) [1], 2) ACS NSQIP [2] (American College of Surgeons National Surgical Quality Improvement Program), 3) ARISCAT [3] (Assess Respiratory Risk in Surgical Patients in Catalonia), and 4) Charlson comorbidity index [4]. Later in the Related Work section some works where these scores were applied are presented.

POSSUM score, proposed by Copeland *et al.* [23], and its extension, P-POSSUM score, proposed by Prytherch *et al.* [83], are methods for normalizing patient data so that direct comparisons of patient outcome could be made despite differing patterns of referral and population. These methods were validated by multiple studies [18, 47, 59, 71, 76, 78]. In our work, P-POSSUM consists in 12 *physiological* and 6 operative factors detailed in Table 2.4. These factors are then inserted into two formulas of both morbidity and mortality rates in order to predict said outcomes.

ACS NSQIP surgical risk calculator, presented by Bilimoria *et al.* [11], uses preoperative factors (demographics and comorbidities) to predict 18 outcomes, originally only 8, within 30-days following surgery. The ACS NSQIP subset of universal preoperative factors considered in the target cohort study are described in Table 2.4.

ARISCAT, proposed by Canet *et al.* [14], is a predictive score to identify postoperative pulmonary complications. The authors inputted multiple variables into a regression model and ended up with 7 predictive variables. The score was later validated by Mazo *et al.* [57] in a large European cohort.

Charlson *et al.* [16] developed a comorbidity index based on the 1-year mortality. This index contains 19 binary weighted variables and was validated by multiple studies [69, 70]. The index aims to define a taxonomy of comorbid conditions which singly or in combination might alter the risk of short-term

[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

mortality for patients enrolled in longitudinal studies.

All these simple risk scores are based on doctor-entered data.

Table 2.3: Outcomes for each of the risk scores introduced.

Risk scores	**Outcomes predicted**
P-POSSUM	Mortality, Morbidity
ACS NSQIP	Mortality, Morbidity, Pneumonia, Cardiac, Surgical site infection, Urinary tract infection, Deep venous thrombosis, Renal failure
ARISCAT	Postoperative pulmonary complication
Charlson Comorbidity Index	Risk of death from comorbid disease, Survival 10-years after surgery

2.2 Pattern Mining

A transactional dataset consists in n transactions containing a variable number of variables. A tabular dataset can be mapped to a transactional dataset through discretization and dummyfication. Table 2.5 shows the conversion to a transactional dataset of the tabular dataset example presented in Table 2.2.

Pattern mining discovers patterns within a transactional dataset. These patterns can come in the form of: 1) itemsets, 2) association rules, 3) substructures. They appear with frequency no less than a specified threshold and contain varying numbers of variables.

Definition 1. *Let* $\mathcal{L}$ *be a finite set of items, and* P *be an itemset* $P \subseteq \mathcal{L}$*. A transaction t is a pair* (t_{id}, P) *with* $id \in \mathbb{N}$*. An itemset database* D *over* $\mathcal{L}$ *is a finite set of transactions* $\{t_1, ..., t_n\}$*.*

Definition 2. *A transaction* (t_{id}, P) *contains* P'*, denoted* $P' \subseteq (t_{id}, P)$ *if* $P' \subseteq P$*. The coverage* ϕ_P *of an itemset P occurs:* $\phi_P = \{t \in D | P \subseteq t\}$*. The support of an itemset* P *in* D*, denoted* sup_P*, can either be absolute, being its coverage size* $|\phi_P|$*, or a relative threshold given by* $|\phi_P|/|D|$*.*

Definition 3. *Give an itemset database* D *and a minimum support threshold* θ*, the frequent itemset mining* (FIM) *problem consists of computing the set* $\{P | P \subseteq \mathcal{L}, \theta \leq sup_P\}$*.*

- **Support:** $Support(X) = \dfrac{|t \in T; X \subseteq t|}{|T|}$, where T is a set of transactions of a given database, X is an itemset, t is a set of transactions which contain the itemset X

An accepted pattern is a frequent itemset that satisfies any other placed constraints over D. For example considering Table 2.5 we have $|\mathcal{L}| = |\{1, ..., 8\}| = 8$, $\phi_{\{y_1(8), y_2(4)\}} = \{t_1, t_3\}$ and $sup_{\{y_1(8), y_2(4)\}} = |\{t_1, t_3\}|/3 = 0.(6)$. For $\theta = 2$, the FIM tasks returns $\{\{y_1(8)\}$, $\{y_2(4)\}$, $\{y_3(3)\}$, $\{y_4(1)\}$, $\{y_1(8), y_2(4)\}\}$, $\{y_2(4), y_4(1)\}\}$, $\{y_3(3), y_4(1)\}\}$, $\{y_2(4), y_3(3)\}\}$, $\{y_1(8), y_4(1)\}\}$, $\{y_1(8), y_2(4), y_4(1)\}\}$.

Table 2.4: Input variables and corresponding categories for each traditional risk score

P-POSSUM

Variable	Categories
Age (years)	$\leq$ 60 | 61-70 | $\geq$ 70
Cardiac signs	No failure | Diuretic, digoxin, antianginal or hypertensive therapy | Peripheral oedema; warfarin therapy | Raised jugular venous pressure
Chest radiograph	Borderline cardiomegaly | cadiomegaly
Respiratory history	No dyspnoea | Dyspnoea on exertion | Limiting dyspnoea | Dyspnoea at rest
Blood pressure (systolic) (mmHg)	110-130 | 100-109, 131-170 | $\geq$ 171, 90-99 | $\leq$ 89
Pulse (beats/min)	50-80 | 81-100, 40-49 | 101-120 | $\geq$ 121, $\leq$ 39
Glasgow coma score	15 | 12-14 | 9-11 | $\leq$ 8
Haemoglobin (g/100 ml)	13-16 | 11.5 - 12.9, 16.1 - 17.0 | 10.0 - 11.4, 17.1 - 18.0 | $\leq$ 9.9, $\geq$ 18.1
White cell count ($\times 10^{12}/l$)	4-10 | 10.1-20.0, 3.1 - 4.0 | $\geq$ 20.1, $\leq$ 3.0
Urea (mmol/l)	$\leq$ 7.5 | 7.6-10.0 | 10.1 - 15.0 | $\geq$ 15.1
Sodium (mmol/l)	$\geq$ 136 | 131-135 | 126 - 130 | $\leq$125
Potassium (mmol/l)	3.5 - 5.0 | 3.2 -3.4, 5.1 - 5.3 | 2.9 - 3.1, 5.4 - 5.9 | $\leq$ 2.8, $\geq$ 6.0
Electrocardiogram	Normal | Atrial fibrallation (rate 60 - 90) | Any other abnormal rhythm or $\geq$ 5 ectopocs/min Q waves or ST/T wave changes
Operative severity	Minor | Moderate | Major | Major +
Multiple procedures	1 | 2 | $>$ 2
Total blood loss (ml)	$\leq$ 100 | 101-500 | 501-999 | $\geq$ 1000
Peritoneal soiling	None | Minor (serous fluid) | Local pus | Free bowel content, pus or blood
Presence of malignancy	None | Primary only | Nodal metatases | Distant metastases
Mode of surgery	Elective | Emergency ressuscitation of $\leq$ 2 h possible + Operation $\leq$ 24 h after admission | Emergency (immediate surgery $\leq$ 2 h needed)

ACS NSQIP

Variable	Categories
Age (years)	$<$ 65 | 65-74 | 75-84 | $\geq$ 85
Sex	Male | Female
Functional status	Independent | Partially dependent | Totally dependent
Emergency case	Yes | No
ASA class	1 | 2 | 3 | 4 | 5
Steroid use for chronic condition	Yes | No
Ascites within 30 d preoperatively	Yes | No
System sepsis within 48 h preoperatively	None | SIRS | sepsis | septic shock
Ventilator dependent	Yes | No
Disseminated cancer	Yes | No
Diabetes	No | Oral | Insulin
Hypertension requiring medication	Yes | No
Previous cardiac event	Yes | No
Congestive heart failure in 30 d preoperatively	Yes | No
Dyspnea	Yes | No
Current smoker within 1 y	Yes | No
History of COPD	Yes | No
Dialysis	Yes | No
Acute renal failure	Yes | No
BMI class	Underweight | Normal | Overweight | Obese 1 | Obese 2 | Obese 3

ARISCAT

Variable	Categories
Age in years	$\leq$ 50 | 51-80 | $>$ 80
Preoperative Oxygen %	$\geq$ 96 | 91-95 | $\leq$ 90
Respiratory infection in the last month	Yes | No
Preoperative anemia (<11g/dl)	Yes | No
Surgical incision	Peripheral | Upper abdominal | Intrathoracic
Duration of surgery in hours	$\leq$ 2 | 2-3 | $\geq$ 3
Emergency procedure	Yes | No

CHARLSON

Variable	Categories
Myocardial infart	Yes | No
Congestive heart failure	Yes | No
Peripheral vascular disease	Yes | No
Cerebrovascular disease	Yes | No
Dementia	Yes | No
Chronic pulmonary disease	Yes | No
Connective tissue disease	Yes | No
Ulcer disease	Yes | No
Mild liver disease	Yes | No
Diabetes	Yes | No
Hemiplegia	Yes | No
Moderate or severe renal disease	Yes | No
Diabetes with end organ damage	Yes | No
Any tumor	Yes | No
Leukemia	Yes | No
Lymphoma	Yes | No
Moderate or severe liver disease	Yes | No
Metastatic solid tumor	Yes | No
AIDS	Yes | No

Table 2.5: Transactional dataset structure based on tabular dataset in Table 2.2

Transaction id	Items
1	$\{y_1(8), y_2(4), y_3(2), y_4(1)\}$
2	$\{y_1(5), y_2(4), y_3(3), y_4(1)\}$
3	$\{y_1(8), y_2(4), y_3(3), y_4(1)\}$

Definition 4. *Given an itemset matrix, a support threshold θ, and the coverage function $\phi : 2^{\mathcal{L}} \to 2^{D}$ that maps an itemset P to its set of supporting transactions:*

- *A frequent itemset P is an itemset that satisfies $|\phi(P)| \geq \theta$*
- *A closed frequent itemset is a frequent itemset with no superset with same support $(\forall_{P \subset P'} |P| > |P'|)$*
- *A maximal frequent itemset is a frequent itemset with all supersents being infrequent $(\forall_{P \subset P'} |\phi(P')| < \theta)$*

If we consider the transactional database in Table 2.5 and a given threshold $\theta = 2$ and $P \geq 2$, there are two maximal frequent itemset ($\{y_1(8), y_2(4), y_4(1)\}$, $\{y_2(4), y_3(3), y_4(1)\}$) and there are three closed frequent itemsets ($\{y_1(8), y_2(4), y_4(1)\}$, $\{y_2(4), y_3(3), y_4(1)\}$ and $\{y_2(4), y_4(1)\}$).

Definition 5. *Consider two itemsets $P \in 2^{\mathcal{L}}$ and $P' \in 2^{\mathcal{L}}$, where $P' \subseteq P$, and a predicate M. M is monotonic when $M(P) \Rightarrow M(P')$ and M is anti-monotonic when $\neg M(P') \Rightarrow \neg M(P)$.*

The previously introduced properties are the basis of FIM. Finding patterns is critical to derive relations from data. Association rules are a way to direct the search for itemsets in an informative way as they discriminate the values along specific variables (rule's consequent) based on the occurrence of other items in the transaction (rule's antecedent). To evaluate each rule a set of standard metrics is used, including:

- **Confidence:** $Confidence(X \implies Y) = \frac{Support(X \cup Y)}{Support(X)}$, which tells us how often the rule has been found to be true
- **Lift:** $Lift(X \implies Y) = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)}$, gives us the ratio of the observed support to that expected if X and Y were independent.

If we consider the example in Table 2.5 and the association rule $\{y_2(4); y_3(3)\} \to \{y_1(8)\}$:

- $Confidence(\{y_2(4), y_3(3)\} \to \{y_1(8)\})$ = $\frac{(1/3)}{(2/3)} = \frac{1}{2}$
- $Lift(\{y_2(4), y_3(3)\} \to \{y_1(8)\})$ = $\frac{(1/3)}{(2/3) \times (2/3)} = \frac{3}{4}$

Since the earlier introduction of FIM and association rule mining by Agrawal in [1], various algorithms have been proposed to do frequent itemset mining [84]. Among the best-known algorithms are: 1)

Apriori, it applies a breadth-first search and the downward closure property (any superset of a non-frequent itemset is non-frequent), to prune the search tree, is exhaustive in search for pattern and requires a large memory space due to candidate generation. 2) *FPGrowth*, it applies divide-and-conquer strategy and an FP-tree data structure (condensed representation of the transactional database), making it possible to only use two database ID scans (one to build the FP-Tree, second to extract the frequent itemsets), and requires a large memory space to build said tree. 3) *Eclat*, it applies a depth-first search and a vertical layout (each item is represented by a set of transaction IDs, tidset), the support of an itemset is the size of the tidset representing it, the size of tidsets is one of the main factors affecting the running time and memory usage of eclat.

2.3 Biclustering

Pattern mining approaches have three major problems. First, efficiency decreases when the dimensionality of the data increases. Second, they don't handle well numerical variables. Third, a high volume of outputs is generated.

The term biclustering was first used by Cheng and Church [21] in gene expression data analysis. It refers to a distinct class of clustering algorithms that perform simultaneous sample-variable clustering. Variants such as coclustering, bidimensional clustering, and subspace clustering, among others, are often used in the literature to refer to the same problem formulation.

Then what is the difference between clustering and biclustering? While clustering can be applied to either the samples or the variables of the data matrix separately, biclustering, on the other hand, performs clustering in two dimensions simultaneously. This means that clustering derives a global model while biclustering produces a local model. In this context when clustering algorithms are used, each patient in a given patient cluster is defined using all the variables. Similarly, each variable in a variable cluster is characterized by the activity of all the patients that belong to it. However, each patient in a bicluster is selected using only a subset of the variables and each variable in a bicluster is selected using only a subset of the patients. The goal of biclustering techniques is thus to identify subgroups of patients and subgroups of variables, by performing simultaneous clustering of both samples and variables of the patients matrix, instead of clustering these two dimensions separately.

Definition 6. *Given a matrix,* A=(X, Y)*, with a set of rows* X=$\{x_1, , ..., x_n\}$*, columns* Y=$\{y_1, ..., y_m\}$*, and elements* $a_{ij} \in \mathbb{R}$ *relating row i and column j:*

- We define a *cluster of rows* as a subset of rows that exhibit a similar behavior across that the set of all columns. This means that a row cluster A_{IY}=(I, Y) is a subset of rows defined over the set of all columns Y, where I=$\{i_1, ..., i_k\}$ is a subset of rows ($I \subseteq X$ and $k \leq n$). A cluster of rows (I, Y) can thus be defined as a k by m submatrix of the matrix A.

- We define a *cluster of columns* as a subset of columns that exhibit similar behavior across the set of all rows. A column cluster A_{XJ}=(X, J) is a subset of columns defined over the set of all rows X, where J=$\{j_1, ..., j_s\}$ is a subset of columns ($J \subseteq Y$ and $s \leq m$). A cluster of columns (X, J)

can then be defined as an n by s submatrix of the matrix A.

- A *bicluster* is a subset of rows that exhibit similar behavior across a subset of columns, and vice versa. The bicluster A_{IJ}=(I, J) is thus a subset of rows and a subset of columns where I=$\{i_1, ..., i_k\}$ is a subset of rows $(I \subseteq X$ and $k \leq n)$, and J=$\{j_1, ..., j_s\}$ is a subset of columns $(J \subseteq Y$ and $s \leq m)$. A bicluster (I, J) can be defined as a k by s submatrix of the matrix A.

- The **biclustering task** aims to identify a set of biclusters B_k=(I_k, J_k) such that each bicluster B_k satisfies specific criteria of homogeneity, where $I_k \subset X, J_k \subset Y$ and $k \in \mathbb{N}$.

Approaches to solve the biclustering task rely on a merit function to define the homogeneity criteria (the variance of the bicluster's values is an example of such function). Merit functions guarantee one or both of the following:

- Intra-bicluster's homogeneity.

- Inter-bicluster homogeneity (overall homogeneity of the output set of biclusters).

The merit function defines the type, quality and structure of biclustering solutions. Alternatively, merit functions can be defined to locally maximize greedy iterative searches, to combine row- and column-based clusters, to exploit matrices recursively, or to guide the space exploration in exhaustive searches. In exhaustive searches, which commonly rely on constrained formulation, merit functions are the heuristics that guide the space exploration.

The pursued homogeneity determines the coherence, quality and structure of a biclustering solution [35]. 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. These concepts are discussed in the following sections.

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, presented later on as either α_i or β_j depending on the the type of coherence, is the adjustment for observation x_i, and η_{ij} is the noise factor of a_{ij}. 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 $\eta_{ij} \in [-\delta/2, \delta/2]$.

2.3.1 Bicluster types

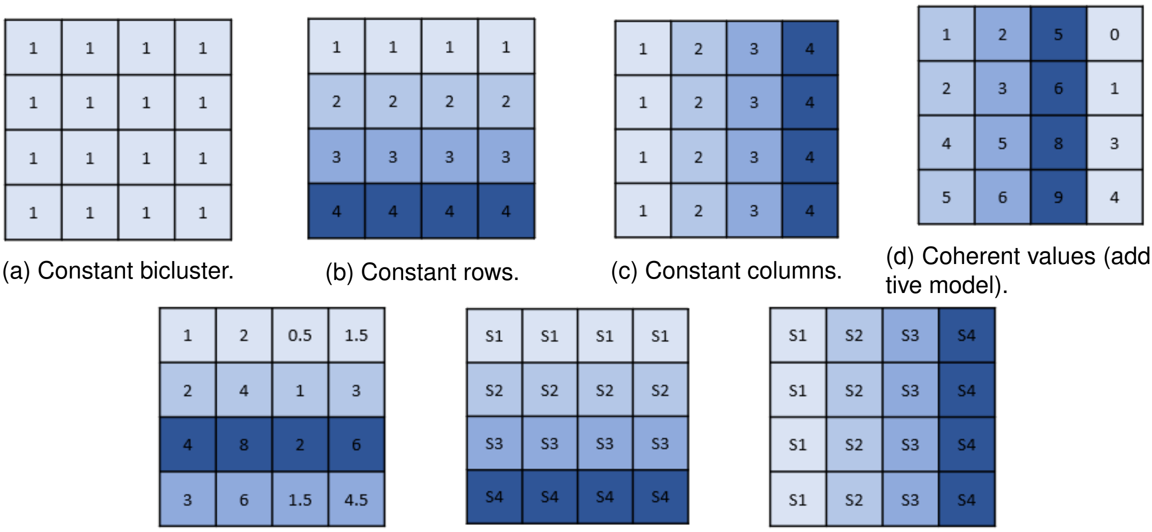
As mentioned before merit functions define the type of biclusters found, Figure 2.1 illustrates different types of biclusters.These different types of biclusters can be categorized into four major classes:

- Constant values. (Figure 2.1a).

- Constant values on rows or columns. (Figure 2.1b and 2.1c).

- Coherent values. (Figure 2.1d and 2.1e).

- Coherent evolutions. (Figure 2.1f, 2.1g and 2.1h).



1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

(a) Constant bicluster.

1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4

(b) Constant rows.

1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4

(c) Constant columns.

1	2	5	0
2	3	6	1
4	5	8	3
5	6	9	4

(d) Coherent values (addictive model).

1	2	0.5	1.5
2	4	1	3
4	8	2	6
3	6	1.5	4.5

(e) Coherent values (multiplicative model).

S1	S1	S1	S1
S2	S2	S2	S2
S3	S3	S3	S3
S4	S4	S4	S4

(f) Coherent evolution on the rows.

S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4

(g) Coherent evolution on the columns.

Figure 2.1: Examples of different types of biclusters.

Each of these four major classes of biclusters has one or multiple expressions that define a *perfect* bicluster. For example a *perfect* constant bicluster is a submatrix (I, J), where all values are equal, for all $i \in I$ and $j \in J$:

$$a_{ij} = \mu \, . \tag{2.1}$$

A *perfect* bicluster with constant rows is a submatrix (I, J), where all the values within the bicluster can be obtained using one of the following expressions:

$$a_{ij} = \mu + \alpha_i \, , \tag{2.2}$$

$$a_{ij} = \mu \times \alpha_i \, , \tag{2.3}$$

where μ is the typical value within the bicluster and α_i is the adjustment for row $i \in I$. This adjustment can be obtained either in an additive (Fig. 2.1d) or multiplicative way (Fig. 2.1e).

Similarly, a perfect bicluster with constant columns is a submatrix (I, J), where all the values within the bicluster can be obtained using one of the following expressions:

$$a_{ij} = \mu + \beta_j \, , \tag{2.4}$$

$$a_{ij} = \mu \times \beta_j \, , \tag{2.5}$$

where μ is the typical value within the bicluster and β_j is the adjustment for column $j \in J$.

If we consider the tabular dataset introduced earlier, Table 2.2, an example of a perfect bicluster with constant columns is shown on Table 2.6 where $\mu = 1$ and $\beta_1 = 7, \beta_2 = 3, \beta_4 = 0$

Table 2.6: An example of a perfect bicluster within the tabular dataset in Table 2.2

	y_1	y_2	y_4
Patient 1	8	4	1
Patient 3	8	4	1

A *perfect* bicluster with coherent values, considering an *additive model*, is a subset of rows and a subset of columns, whose values a_{ij} can be predicted using the following expression:

$$a_{ij} = \mu + \alpha_i + \beta_j \, , \tag{2.6}$$

where μ is the typical value within the bicluster, α_i is the adjustment for row $i \in I$, and β_j is the adjustment for column $j \in J$.

A *perfect* bicluster with coherent values, considering a *multiplicative model*, are given by the following expression:

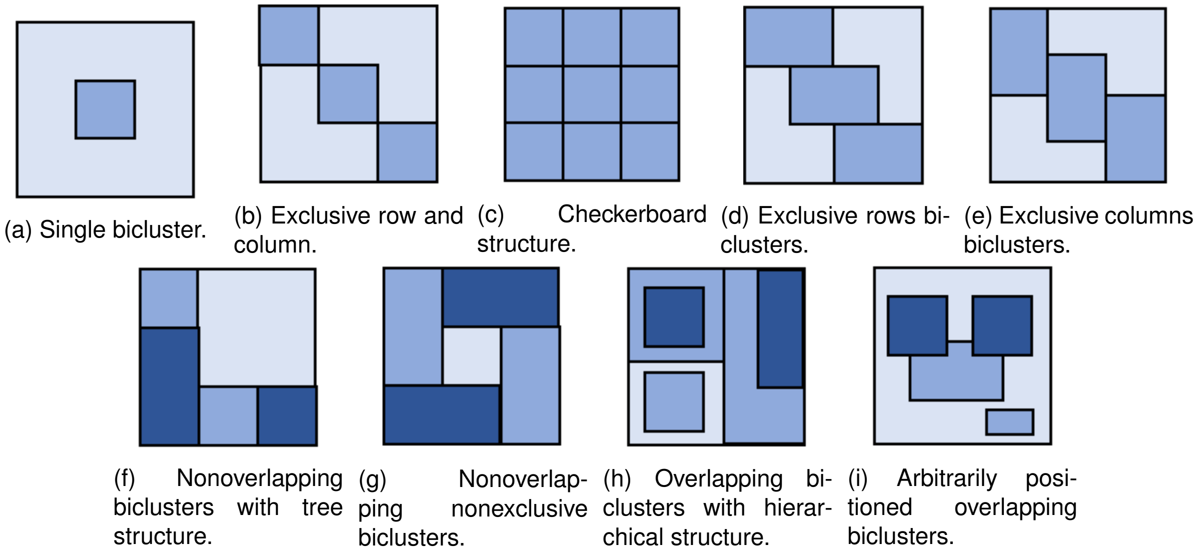
$$a_{ij} = \mu' \times \alpha'_i \times \beta'_j \, . \tag{2.7}$$

Finally, for the fourth category, coherent evolutions, can be present as an order-preserving submatrix where, for example, $a_{j4} \leq a_{j2} \leq a_{j3} \leq a_{j1}$ represent a bicluster with coherent evolutions on its columns. An example using the tabular dataset previously introduced is the whole dataset where $y_1 \geq y_2 \geq y_3 \geq y_4$.

2.3.2 Bicluster structure

The existence of biclusters can be viewed as a single submatrix or as multiple submatrixes inside a matrix. When considering the existence of several biclusters in the data matrix, the following bicluster structures can be obtained:

- exclusive row and column biclusters (Fig. 2.2b).
- nonoverlapping biclusters with checkerboard structure (Fig. 2.2c).
- exclusive-rows biclusters (Fig. 2.2d) or exclusive-columns biclusters (Fig. 2.2e).
- nonoverlapping biclusters with tree structure (Fig. 2.2f).
- nonoverlapping nonexclusive biclusters (Fig. 2.2g).
- overlapping biclusters with hierarchical structure (Fig. 2.2h).
- arbitrarily positioned overlapping biclusters (Fig. 2.2i).



(a) Single bicluster.

(b) Exclusive row and column.

(c) Checkerboard structure.

(d) Exclusive rows biclusters.

(e) Exclusive columns biclusters.

(f) Nonoverlapping biclusters with tree structure.

(g) Nonoverlapping nonexclusive biclusters.

(h) Overlapping biclusters with hierarchical structure.

(i) Arbitrarily positioned overlapping biclusters.

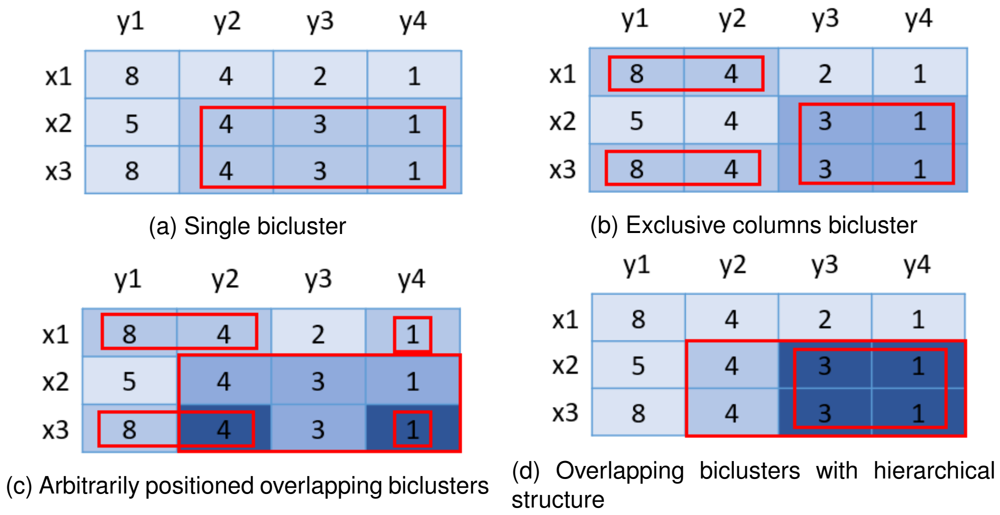
Figure 2.2: Bicluster structures.

An ideal reordering of the matrix would produce an image, similar to Fig.2.2b, with some number K of rectangular blocks on the diagonal. This ideal corresponds to the existence of K mutually exclusive and exhaustive clusters of rows, and a corresponding $K-way$ partitioning of the columns. Every row and every column in the matrix belongs exclusively to one of the K biclusters. Figure 2.2c structure is produced if we consider that rows and columns may belong to more than one bicluster, and assume a checkerboard structure, we allow the existence of K nonoverlapping and nonexclusive biclusters. In this case each row in the data matrix belongs to exactly K biclusters. If we assume that rows (or columns) can only belong to one bicluster, while columns (or rows), can belong to several biclusters, the structure produced is similar to Fig.2.2d and 2.2e. Other bicluster structures include the tree structure depicted in figures 2.2f and 2.2g.

The previously discussed structures assume that the biclusters are exhaustive, that is, every row and every column belongs to at least one bicluster (there are nonexhaustive variations of these structures that make it possible that some rows and columns do not belong to any bicluster), and are restrictive in many ways. Some assume that, for visualization purposes, all the identified biclusters should be observed directly on the data matrix. Others assume that the biclusters are exhaustive. However, it is more likely that, in real data, 1) some rows or columns do not belong to any bicluster at all, and, 2) that the biclusters overlap in some places. If we consider a hierarchical structure, depicted in figure 2.2h, that requires that either the biclusters are disjoint or that one includes the other, it is possible to have the two previously mentioned properties without relaxing the visualization property. A more general bicluster structure allows the existence of K possibly overlapping biclusters without taking into account their direct observation on the data matrix with a common reordering of its rows and columns, figure 2.2i.

Considering the tabular dataset example from earlier, Table 2.2, figure 2.3 displays the found structures. Biclusters have the cells painted with a darker blue and red boxes, if the cell contains a darker blue and is inside two red boxes then it belongs to two biclusters.

When considering overlapping structures of biclusters we can consider an additive or multiplicative overlap. Figure 2.4 illustrates examples of different types of biclusters overlapped with a general addictive model, and, figure 2.5 with a multiplicative model.



	y1	y2	y3	y4
x1	8	4	2	1
x2	5	4	3	1
x3	8	4	3	1

(a) Single bicluster

	y1	y2	y3	y4
x1	8	4	2	1
x2	5	4	3	1
x3	8	4	3	1

(b) Exclusive columns bicluster

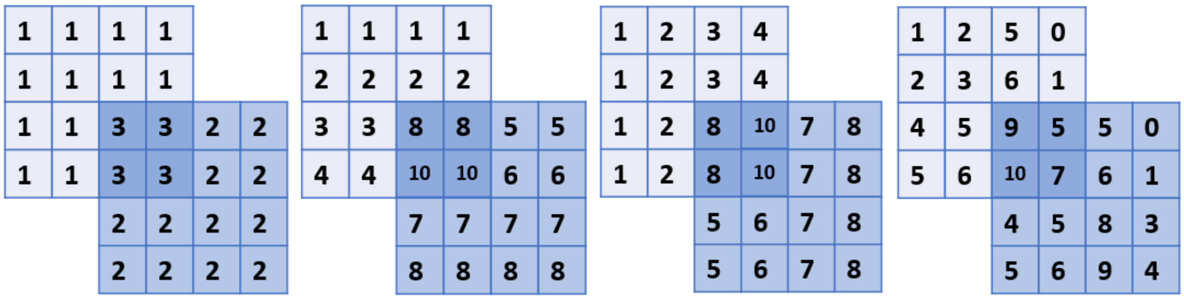
	y1	y2	y3	y4
x1	8	4	2	1
x2	5	4	3	1
x3	8	4	3	1

(c) Arbitrarily positioned overlapping biclusters

	y1	y2	y3	y4
x1	8	4	2	1
x2	5	4	3	1
x3	8	4	3	1

(d) Overlapping biclusters with hierarchical structure

Figure 2.3: Example of bicluster structures found within tabular dataset 2.2.



1	1	1	1		
1	1	1	1		
1	1	3	3	2	2
1	1	3	3	2	2
		2	2	2	2
		2	2	2	2

(a) Constant biclusters

1	1	1	1		
2	2	2	2		
3	3	8	8	5	5
4	4	10	10	6	6
		7	7	7	7
		8	8	8	8

(b) Constant rows

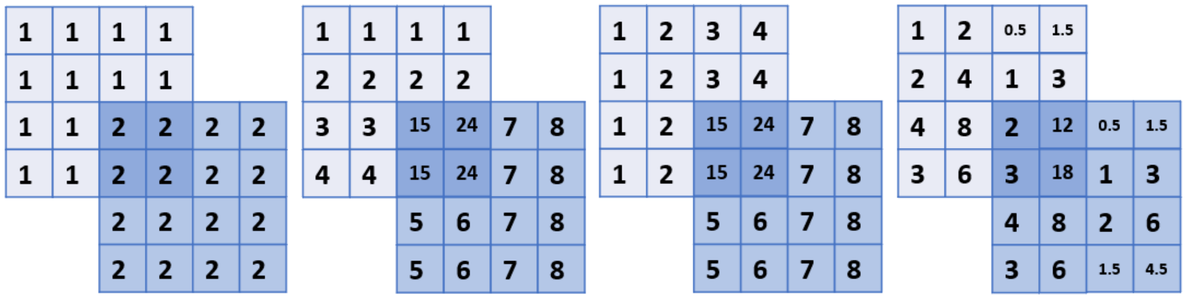
1	2	3	4		
1	2	3	4		
1	2	8	10	7	8
1	2	8	10	7	8
		5	6	7	8
		5	6	7	8

(c) Constant columns

1	2	5	0		
2	3	6	1		
4	5	9	5	5	0
5	6	10	7	6	1
		4	5	8	3
		5	6	9	4

(d) Coherent values

Figure 2.4: Overlapping biclusters with general additive model.



1	1	1	1		
1	1	1	1		
1	1	2	2	2	2
1	1	2	2	2	2
		2	2	2	2
		2	2	2	2

(a) Constant biclusters

1	1	1	1		
2	2	2	2		
3	3	15	24	7	8
4	4	15	24	7	8
		5	6	7	8
		5	6	7	8

(b) Constant rows

1	2	3	4		
1	2	3	4		
1	2	15	24	7	8
1	2	15	24	7	8
		5	6	7	8
		5	6	7	8

(c) Constant columns

1	2	0.5	1.5		
2	4	1	3		
4	8	2	12	0.5	1.5
3	6	3	18	1	3
		4	8	2	6
		3	6	1.5	4.5

(d) Coherent values

Figure 2.5: Overlapping biclusters with general multiplicative model.

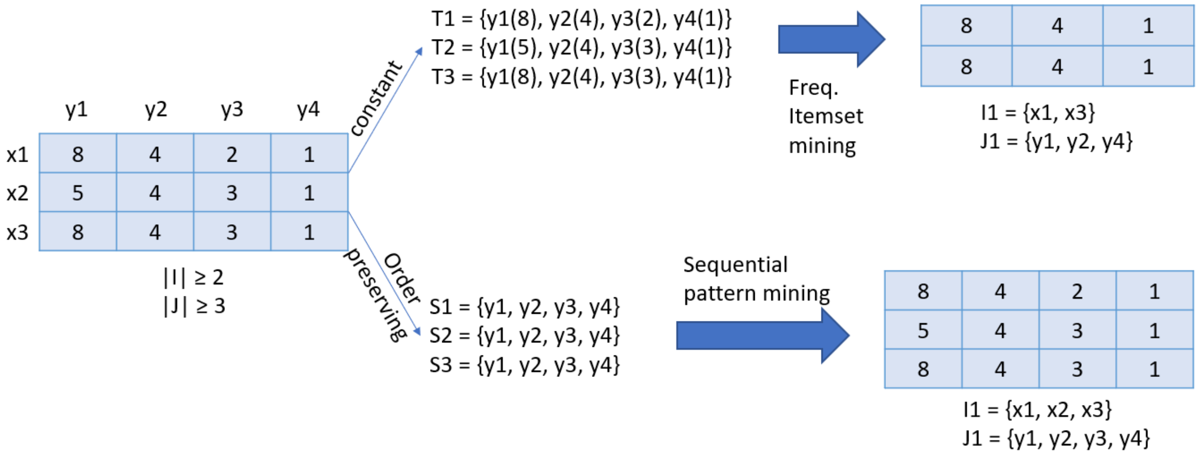
2.3.3 Algorithms

Some approaches attempt to identify one bicluster at a time, others one set of biclusters at a time. Algorithms can perform simultaneous bicluster identification, which means that the biclusters are discovered all at the same time. Given the complexity of the problem, a number of different heuristic approaches have been used:

- Iterative row and column clustering combination.
  - They apply clustering algorithms to the rows and columns of the data matrix separately, and then combine the results using some sort of iterative procedures to combine the two cluster arrangements.
- Divide and conquer.
  - They break the problem into several subproblems that are similar to the original problem but smaller in size, solve the problems recursively, and then combine the solutions to create a solution to the original problem. It has the significant advantage of being potentially very fast, however, it is likely to miss good biclusters that may be split before they can be identified.
- Greedy iterative search.
  - They always make a locally optimal choice in the hope that this choice will lead to a globally good solution. They create biclusters by adding or removing rows/columns from them, using a criterion that maximizes the *local* gain.
- Exhaustive bicluster enumeration.
  - They are based on the idea that the best biclusters can only be identified using an exhaustive enumeration of all possible biclusters existent in the matrix. Due to their high complexity, they can only be executed by assuming restrictions on the size of the biclusters.
- Distribution parameter identification.
  - They assume a given statistical model and try to identify the distribution parameters used to generate the data by iteratively minimizing a certain criterion.

2.4 Pattern-based biclustering

As more biclustering algorithms started to emerge, a new approach to biclustering, pattern-based biclustering, appeared to address some of the commonly observed limitations of peer approaches[34]. A pattern-based biclustering approach allows for an efficient and exhaustive space search. It relies on an itemization step, where the original matrix is transformed into a transactional dataset, followed by the application of frequent itemset mining and sequential pattern mining methods (for real value matrices normalization and discretization procedures are applied). This approach produces a non-fixed number of biclusters within a flexible structure. By using a transactional dataset, pattern-based biclustering deals with missing and noisy values, as they search transactions with varying length, making it possible to remove missings or be associated with zero or multiple values. Despite pattern-based biclustering being inherent oriented to search for biclusters with constant model, it was extended to search for additive, multiplicative, symmetric, order-preserving and plaid models. The search for biclusters can also be extended using discriminative pattern mining or incorporate pattern mining-based constraints to guide the search by pruning the search space and focus on non-trivial biclusters.



	y1	y2	y3	y4
x1	8	4	2	1
x2	5	4	3	1
x3	8	4	3	1

T1 = {y1(8), y2(4), y3(2), y4(1)}
T2 = {y1(5), y2(4), y3(3), y4(1)}
T3 = {y1(8), y2(4), y3(3), y4(1)}

8	4	1
8	4	1

I1 = {x1, x3}
J1 = {y1, y2, y4}

S1 = {y1, y2, y3, y4}
S2 = {y1, y2, y3, y4}
S3 = {y1, y2, y3, y4}

8	4	2	1
5	4	3	1
8	4	3	1

I1 = {x1, x2, x3}
J1 = {y1, y2, y3, y4}

Figure 2.6: 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.

Definition 7. *Given a matrix A and a minimum support threshold θ, a set of biclusters $\cup_k \mathcal{B}_k$, where $\mathcal{B}_k = (I_k, J_k)$, can be derived from the set of frequent itemsets $\cup_k P_k$ by either mapping $(I_k, J_k) = (\phi_{P_k}, P_k)$ to compose biclusters with coherency on rows, or by mapping $(I_k, J_k) = (P_k, \phi_{P_k})$ to compose biclusters with coherency on columns.*

Pattern-based biclustering can be viewed as a sequence of three steps: 1) Mapping, handling of outliers, noise and missings, real-value matrix handling, and itemization. 2) Mining, pattern mining approaches, application schemas for non-constant models, target patterns and algorithms. 3) Closing, merging structures and overlapping, noise tolerance, filtering biclusters.

Two classes of pattern-based biclustering approaches can be considered:

- Directly apply pattern miners over discrete matrices.

- Target numeric matrices by customizing the support metric (itemization step is optional).

Figure 2.6 maps a matrix into two distinct transactional databases (given by index concatenations and orderings) for the subsequent discovery of constant and order-preserving biclusters derived from frequent patterns.

Chapter 3

Related work

This section surveys the usage of the traditional risk scores introduced in the previous section and state-of-the-art advances on pattern discovery as well as applied efforts of bridging these contributions in the oncological domains. Accordingly, this chapter is divided into sections: 1) traditional risk scores, and 2) pattern discovery. In the risk scores section, for each risk score presented in the Background chapter, there is a subsection. In each of these subsections, multiple works where the score was applied for a healthcare system are presented. In each case different types of patients, and the effectiveness of the scores are presented. Pattern discovery contains four subsections: 1) Unsupervised analysis of oncology data, where multiple works using other techniques besides classical pattern mining and biclustering is used in the oncological domain, 2) Classic pattern mining, where extensions/modifications to classical pattern mining approaches are presented, 3) Biclustering, where multiple algorithms, cohorts, and toolboxs are presented, and 4) Pattern visualization, where state of the art visualization of patterns is presented.

3.1 Traditional Risk Scores

3.1.1 P-POSSUM

Emergency laparotomy[1] is considered a high-risk surgical procedure with high morbidity and mortality. Therefore, it is important to have an accurate pre-operative assessment of the associated risks, morbidity and mortality. Cao *et al.* [15] evaluated the predictive power of mortality scores of patients, 65 or older, undergoing emergency laparotomy surgery. Patients who had a conversion from laparoscopic, a minimally invasive surgery, to a laparotomy or those subjected to laparotomy due to traumatic injury were not included in the study. They tested the predictive power using 2 models, a logistic regression and random forests, and 3 scalers, standard scaler, min-max, and, robust scaler. Their results showed that the models used considered P-POSSUM variables to always be between first and fifth most important predictive variables (in all of the 3 different scalers). In the logistic model P-POSSUM morbidity and mortality were always present in the top 5 predictive variables, and, in the random forest model P-POSSUM morbidity

[1] https://www.medicalnewstoday.com/articles/laparotomy

was always present in the top 5 predictive variables. Thahir *et al.* [77] evaluated the accuracy of the mortality given by two scores, P-POSSUM and NELA, in the same emergency procedure. They considered a wider range of patients, ranging from 18 to 99 years old, and don't mention any exclusive type procedures. They compared the mortality predicted versus mortality observed and their results showed that P-POSSUM tended to over-predict mortality. Chen *et al.* [19] present an extension to the P-POSSUM score, the West Australian Categorisation of Operative Severity (WA classification (WA classification (WA classification) created for all neurosurgical procedures. Their study showed that P–POSSUM models are predictive of the overall mortality in a general neurosurgical service, but, POSSUM consistently overestimated the mortality rates in all groups of patients. They showed that P–POSSUM using the WA system is highly predictive of overall mortality. Johns *et al.* [39] tested P-POSSUM score in hip fracture mortalities. They conducted study on all hip fracture mortalities over a 2-year period and only used patients who died after surgery. When evaluated with only patients who died after surgery their results showed that P-POSSUM underpredicted the observed mortality rate. However P-POSSUM scoring system was able to predict morbidity effectively.

In the oncology domain, Bakshi *et al.* [6] investigates if the current practices of pain management after emergency laparotomy in cancer patients are the most appropriate. They analyze factors influencing the choice of pain management techniques such as time of surgery, patient factors including American Society of Anaesthesiologists (ASA) physical status scores and P-POSSUM scores. Pain management was recorded as a numeric value between 1 and 10 (being 10 the most satisfied). P-POSSUM predicted mortality and pain management technique were compared using a one-way ANOVA test. Their results showed that there was no association between P-POSSUM predicted mortality and pain management technique. Karan *et al.* [41] conducted a study to evaluate the discriminative power of the P-POSSUM mortality and morbidity scores. They defined a composite endpoint of 30-day morbidity as complications occurring within 30 days of surgery or discharge, mortality was documented within 30 days of surgery. In their results P-POSSUM predicted morbidity was higher than the observed morbidity. The mortality score could not be assessed as the observed 30 day mortality was nil, at 90 days was 4, and, at 180 days was 7. P-POSSUM in their study was not able to discriminate well the likelihood of perioperative complications. The authors propose that a reason for their study showing conflicting results compared to other studies is the differences in the study populations, and inclusion of both elective and emergency surgeries in other studies.

3.1.2 ACS NSQIP SRC

O'Neill *et al.* [63] conducted a study to evaluate the predictive value of the ACS NSQIP calculator in patients undergoing microvascular breast reconstruction. The study used 515 patients and demonstrated that the ACS NSQIP surgical risk calculator accurately predicted the proportion of patients that developed post-operative complications but it couldn't identify the individual patents who were at risk of complications.

In the oncology domain, Sahara *et al.* [72] conducted a study to validate and examine the accuracy of the ACS NSQIP SRC to predict outcomes among elderly patients undergoing liver resection (due to cancer). Their study concluded that ACS NSQIP SRC failed to estimate accurately the risk of many adverse outcomes, such as complications after resection of the liver, while it overestimated the risk for 30-day readmission and non-home discharge. Cusworth *et al.* [24] aimed to determine the ability of ACS NSQIP SRC to accurately predict risk of complications and length of hospital stay who underwent Pancreaticoduodenectomy (remove cancerous tumors off the head of the pancreas). If the patient developed a pancreatic-specific post-surgical complication then the underestimated the risk, but if the patient developed a non-pancreatic-specific post-surgical complication then the prediction matched with the observed.

3.1.3 ARISCAT

Kupeli *et al.* [46] conducted a study to compare the ASA scale and ARISCAT to predict pulmonary complications after renal transplant. The study used 172 patients primarily male. The authors had a previous study where they considered ARISCAT a useful tool, Kupeli *et al.* [45], in predicting the aforementioned complications on renal transfer patients, in this study they concluded again that ARISCAT was reliable and useful to stratify risk when advising patients before surgery.

3.1.4 Charlson

Birim *et al.* [12] evaluated the impact of the Charlson Comorbidity Index in early stage lung cancer and its predicting capability on long-term survival. The study used 433 patients and concluded that Charlson is a better predictor of survival and validated its ability to stratify comorbidity severity for patients undergoing early stage lung cancer surgery. Voskuijl *et al.* [81] studied if a higher Charlson Comorbidity Index was associated with readmission of the patient, an increased risk of surgical infection or mortality. They concluded that Charlson Comorbidity index was not associated with infection but was associated with a higher risk of death within 30 days of discharge for patients receiving oncologic treatment. Charlson index was also associated with readmission after surgery for spine and trauma surgeries, thus having a significant influence on readmission.

3.2 Pattern discovery

3.2.1 Unsupervised analysis of oncology data

Li *et al.* [49] introduces a new method to discover statistically significant association rules in high-dimensional profiling data, to aggregate the discriminative power of these rules for reliable predictions. In this approach, they use decision trees to discover rules but use committees of trees, instead of a single tree, where every leaf is a collection of rules. They weight the rules according to their coverage in the original training dataset which causes the rules to reflect precisely the nature of the original training data. The final decision to classify a test sample is done by voting, in a weighted manner, the rules in the k trees of the committee that the test sample satisfies. The authors present three facts from real examples of high-dimensional profiling data. First, statistically significant association rules often contain globally low-ranked features. Second, if the construction of a tree is confined to a set of globally top-ranked features, the rules in the resulting tree may be less significant than those rules derived by using the whole feature space. Third, alternative trees can often outperform or compete with the performance of the 'optimal' tree when the same set of test data is applied. The authors report that their method provides a highly competitive accuracy compared to current approaches and highly comprehensible rules that help translate raw data into knowledge.

Bellazzi *et al.* [8] presents a review on the main features of predictive clinical pattern mining such as methods able to deal with temporal data and the efforts performed to translate molecular medicine results into clinically useful pattern mining models. The authors refer to the application of pattern mining in clinical medicine as being related to a predictive (supervised) and descriptive (unsupervised) task. These tasks also employ a feature selection to better the predictive model created. Time series have been also studied with traditional signal processing and feature extraction methods, which are devoted to summarizing the temporal information in attributes suitable for classification algorithms. A predictive model can then be constructed to forecast a response variable. This variable can be categorical or numerical, so that predictive pattern mining may deal with classification and regression problems. Feature selection can occur before the model estimation. This is done using feature ranking based on the attribute predictive capability, for example, information gain.

Pendharkar *et al.* [67] shows how pattern mining using association rules and classification approaches can be a viable tool to predict and diagnose the occurrence of breast cancer. The paper focuses on exploring data envelopment analysis (DEA), for binary classification problems, and artificial neural networks (ANN) using discriminative analysis for mining breast cancer patterns. DEA seeks to determine a subset of k decision-making units that determine the envelopment surface when all k decision-making units consist of m inputs and s outputs. The authors start by filling in the missing values to ensure the data was complete and valid. Then a division was made, the data was split in two, one for learning the patterns and the other for testing the predictive performance. Once the records were divided, extraneous data associated with each record was eliminated and the models were trained. The authors found that comparing the results obtained with other studies on an equitable basis difficult as 1) DEA uses information about one class to determine the discriminant function whereas, other techniques use information about two classes to determine the discriminant function, 2) The performance of DEA is likely to vary if DEA and its variant is used for the 2 classes separately, 3)The datasets are different and attributes considered in this study are different from the attributes considered in other studies. Despite this, the solo evaluation reveals that both DEA and ANN outperform the traditional statistical discriminant analysis, with ANN being superior as DEA assumes the convexity of the acceptable cases and neural networks relax this assumption.

3.2.2 Classic pattern mining

Fang *et al.* [27] presents an approach to mine low-support discriminative patterns from dense and high-dimensional data. They do this by proposing a family of antimonotonic measures named *SupMaxK*. *SupMaxK* conceptually organizes the set of discriminative patterns into nested layers of subsets. As it does, they become progressively more complete in their coverage but require increasingly more computation for their discovery. In particular, a special member of this family is presented by the authors as best suitable for dense and high-dimensional data and can serve a complementary role to the existing approaches by helping to discover low-support discriminative patterns. It is named *SupMaxPair*, because $K = 2$. To evaluate the patterns discovered using *SupMaxPair* they used synthetic datasets and a cancer gene expression dataset, and two criteria, 1) pattern-based biological relevance, 2) Gene collection-based biological relevance. The experimental results they obtained showed that *SupMax1* generally provides very poor approximation of the absolute difference of the relative supports (if the absolute difference is bigger than a given r, then the itemset is discriminative), the approximation is improved substantially when K goes to 2 and 3 but when K is increased further to 3 and 4, the computation time increases exponentially, in spite of the approximation improving much slower when compared to the improvement obtained when K goes from 1 to 2.

Borgelt *et al.* [13] presents an algorithm to find fragments in a set of molecules that help to discriminate between different classes of activity in a drug discovery context. The proposed algorithm maintains parallel embeddings of a fragment into all molecules throughout the growth process and exploits a local order of the atoms and bonds of a fragment to prune the search tree, this results in a fast search and a restricted depth-first search algorithm, similar to the *Eclat* algorithm. They use a pruning technique not based on support nor size but a third type which they call structural pruning. Structural pruning ensures that every itemset is considered in one branch only, even though adding items in different orders can yield the same itemset. They do not define a global order of the atoms they number the atoms in a substructure and record how a substructure was constructed in order to constrain its extensions. The results obtained from applying the algorithm found relevant fragments using data from a well-known HIV-screening compound database.

3.2.3 Biclustering

The *Bimax* algorithm, proposed by Prelic *et al.* [68], finds subgroups in a binary matrix where all entries are one. In a divide-and-conquer fashion, this is done by iterating two steps: 1) Rearrange the samples and variables to concentrate entries with ones in the upper right corner of the matrix, and 2) Divide the matrix into two submatrices. Submatrices with only ones are returned. To obtain satisfying results, the method needs to be restarted several times with different starting points.

The *CC* (Cheng and Church) algorithm, proposed by Cheng and Church [21], it defines a bicluster as a subset of samples and variables with a high similarity score (equation 3.1). In particular, the authors aim at finding large and maximal biclusters with scores below a certain threshold δ. This means that the values in each sample or variable can be generated by shifting the values of other samples or variables by a common offset. Unfortunately, due to noise in data, δ-biclusters may not always be perfect. The concept of residue was thus introduced to quantify the difference between the actual value of an element a_{ij} and its expected value predicted from the corresponding sample mean, variable mean, and bicluster mean. To assess the overall quality of a δ-bicluster, Cheng and Church defined the mean squared residue, H, of a bicluster (*I*, *J*) as the sum of the squared residues, equation 3.2.

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2, \tag{3.1}$$

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} r(a_{ij})^2, \tag{3.2}$$

where a_{iJ} represents the row mean, a_{Ij} represents the column mean, a_{IJ} the bicluster mean.

The algorithm is divided into three major steps:

1. Deleting samples and variables with a score larger than alpha times the matrix score;
2. Deleting samples and variables with largest scores;
3. Adding samples or variables until alpha level is reached;

These steps are repeated until a maximum number of biclusters is reached or no bicluster is found.

Spectral algorithm, proposed by Kluger *et al.* [43], addresses the problem of identifying biclusters with coherent values. It looks for checkerboard structures in the data matrix by integrating biclustering of samples and variables with normalization of the data matrix. The authors assume that after a particular normalization, which was designed to accentuate biclusters if they exist, the contribution of a bicluster is given by a multiplicative model. To identify the structure the algorithm goes through the following steps:

1. Reorder the data matrix and choose a normalization method
2. Compute a singular value singular value decomposition to get eigenvalues and eigenvectors
3. Depending on the chosen normalization methods, construct biclusters beginning from the largest or second largest eigenvalue

The number of biclusters correlates with the number and value of the eigenvalues. The biclusters found have a checkerboard structure and have higher or lower values than the samples and variables around them.

The previous biclustering approaches evaluate separately the contribution of each bicluster without taking into consideration the interactions between biclusters. In particular, they do not explicitly take into account that the value of a given element, a_{ij}, in the data matrix can be seen as a sum of the contributions of the different biclusters to whom the sample *i* and the variable *j* belong. Plaid model, proposed by Lazzeroni *et al.* [48], addressed this limitation by viewing the value of an element in the data matrix as a sum of terms called layers. In the plaid model, the data matrix is described as a linear function of variables (layers) corresponding to its biclusters. The plaid model is defined as follows:

$$a_{ij} = \sum_{k=1}^{K} \theta_{ijk} + \rho_{ik} + \kappa_{jk}, \tag{3.3}$$

where *K* is the number of layers (biclusters) and the value of θ_{ijk} specifies the contribution of each bicluster *k* specified by ρ_{ik} and κ_{jk}. The terms ρ_{ik} and κ_{jk} are binary values that represent, respectively, the membership of sample *i* and variable *j* in bicluster *k*.The algorithm follows these three steps:

1. Update all parameters one after another *S* (user defined) times

2. Calculate the sum of squares of the layer (LSS) using the resulting parameters

3. Compare Result with random permutation and return bicluster if LSS is higher

The steps repeat until no new bicluster is found.

The *Xmotifs* algorithm, proposed by Murali *et al.* [60], searches for samples with constant values over a set of variables. The main aspect of this algorithm is to define a sample state where a sample is called conserved if it has the same state in all variables. Once the data matrix represents the states, the algorithm chooses a random number of variables n times and performs the following steps:

1. Choose a subset from these variables and collect all samples with equal state in this subset

2. Collect all variables where these samples have the same state

3. Return the bicluster if it has the most samples from all found and is also larger than a alpha fraction of the data

The algorithm finds submatrices where all samples have the same value structure over the variables. To collect more than 1 bicluster the calculation can be rerun without the samples and variables already found, or just return small combinations that were found. It is possible to find groups with a large variance in their values in the sample direction.

Veroneze *et al.* [79] 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.

The work done by Harpaz *et al.* [30] used the Bimax algorithm to identify drug groups that share a common set of adverse events. The data used was collected from Adverse Event Reporting System reports which contain a total of 441 009 individual reports, 65 975 unique drugs and 10 886 unique adverse events. These numbers were reduced when the authors mapped to generic drugs and removed the duplicate reports. To use the Bimax algorithm the authors defined as parameters the minimum number of drugs and the minimum number of adverse events, this defines the minimum size for a bicluster found. The biclusters defined consisted of a small set of drugs, each of which is associated with the same small set of adverse events. The results obtained demonstrated that a significant number of biclusters, relating adverse drug events by grouping similar drugs with a common set of adverse events, were identified. Additionally, the authors ran two statistical tests to prove that the biclusters found were extremely unlikely to have occurred by chance. The first statistical test used a standard graph-theoretic approach to compute the expected number and probability of random 3 by 3 complete bipartite graphs. The second statistical test conducted a hypothesis test by creating a set of 100 random 3 by 3 biclusters sampled from the Adverse Event Reporting System drug and adverse events distributions and by comparing the number of known biclusters identified in the random set with the number of known biclusters identified they identified.

The Plaid model biclustering algorithm was highly emphasized by Alavi and co-workers [2], as being one of the most flexible algorithms proposed, and, in their work, they provided an evaluation of the Plaid models in both simulation data and real data. The simulation data generated consists of two matrices with different degrees of overlap and noise with two embedded biclusters. The real dataset contains information related to breast cancer (docetaxel resistance) [42]. They begin by normalizing the information with a median approach and then filling the missing values using KNN. Plaid model is then applied to both data. The authors concluded that in big datasets with little noise, Plaid model could provide useful information.

A method to efficiently select relevant genes using Spectral biclustering was proposed by Liu *et al.* [51]. The key idea of their method is to use the best class partitioning eigenvectors, given by spectral biclustering, and select the top 100-200 genes, from these genes, they select the best 2-gene combinations, which can accurately divide the cancer data. The results achieved for the lymphoma data was a selection of two genes which managed to separate samples perfectly. For liver cancer data, the best two-gene combinations selected separated samples well, with only two samples misclassified.

An exhaustive study, presented by Mandal *et al.* [55], to identify biomarkers using two approaches, frequency-based (Frequency is nothing but several occurrences of a gene or miRNA in all the biclusters) and network-based (this technique incorporates external biological knowledge with biclustering results) use several biclustering techniques. They applied over seventeen different biclustering algorithms to four different single type cancer gene expression datasets such as blood, lung, colon, prostate, one multi-tissue microarray cancer dataset, and one miRNA breast cancer dataset. As far as preprocessing done, to the miRNA expression data for breast cancer they filtered out extremely low expression values across all samples then applied z-score normalization, to the other datasets preprocessing was not done as the data used was already treated. After this they had to specify the biclustering parameters for each biclustering technique, these were selected from previous studies using these algorithms. The fundamental goal of the publication was to do a systematic comparison among a few prominent biclustering algorithms and to evaluate their effectiveness based on their ability to provide relevant information such as biomarkers and subtypes for a given disease. The conclusions they reached were that of the biclustering algorithms one of the best suited for finding subtypes for blood, colon, prostate and breast cancer was Cheng and Church algorithm. For the microarray data, Cheng and Church was also one of the best performing algorithms throughout all the datasets except prostate cancer. For biomarker identification methods Cheng and Church and Xmotifs both proved to be well suited.

To utilize some of the the biclustering algorithms previously mentioned both Kaiser *et al.* [40] and Barkow *et al.* [7] implemented toolboxes. These provide the user with a number of preprocessing, biclustering and cluster validation functions. Table 3.1 lists the functions available in each toolbox.

Table 3.1: List of functions for preprocessing, biclustering, and bicluster analysis of each toolbox.

	biclust	BicAT
Preprocessing	normalization, discretization, independent scaling, bistochastization, log interactions	normalization (log2, mean,centric), discretization
Biclustering Algorithms	Bimax, Cheng and Church, Xmotifs, plaid model, Spectral	Bimax, Cheng and Church, Xmotifs, Iterative Signature [37] [38], Orderpreserving Submatrix [9]
Biclusters Validation	Jaccard index, Variation index [54], Scoring function [22], F Statistic	Analysis of gene pair occurrence to derive gene interconnection graphs

3.2.4 Pattern-based biclustering

Henriques *et al.* [34] 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.* [75], BiModule proposed by Okada *et al.* [62], GenMiner proposed by Martinez *et al.* [56], BicPAM proposed by Henriques *et al.* [36], RAP proposed by Pandey *et al.* [66], RCB Discovery proposed by Atluri [5], and ET-Bicluster proposed by Gupta *et al.* [29]. 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 beneficial 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.

3.2.5 Pattern visualization

Visualizing correctly and with a clean layout the results found through pattern mining is essential to get the correct interpretation of the data. Some standard visualizations were discussed in the Background section and in this subsection, some work that uses them and some new approaches are presented.

The work presented by Liu *et al.* [52] is a system called *AssocExplorer* to support exploratory data analysis via association rule visualization and exploration. They use a scatter plot to provide a global view of the rules. The X-axis represents the coverage of rules and the Y-axis represents confidence. The authors also use colors to help highlight the relevant results to the user. They highlight length-1 rules as they are simple and easy to comprehend. They also allow the user to color rules based on a selected attribute, rules that do not contain the attribute are colored gray, filter out or zoom in on rules that are interesting to them.

Santamaria *et al.* [73] present BicOverlapper, a tool to visualize biclusters from gene-expression matrices using a graph visualization. Nodes represent genes or conditions, and edges join nodes that are grouped by one or more biclusters. Each bicluster is represented as an undirected complete subgraph. The overlap between biclusters is visualized by means of intersecting hulls. The use of glyphs on genes and conditions nodes improves our understanding of instances of overlapping when the representation becomes complex. With these details the tool helps to compare biclustering methods, to unravel trends and to highlight relevant genes and conditions.

The open-source software tool BiVisu, presented by Cheng *et al.* [20] focus on detecting and visualizing biclusters embedded in gene expression matrix. Apart from the preprocessing, biclustering and filtering the tool offers it also presents to the user a way to visualize the biclusters found through Parallel Coordinates visualization. They use the yeast dataset to prove the effectiveness of the tool. Parallel Coordinates together with the mean square residual score and average correlation value, subjective and objective judgment of bicluster homogeneity can be achieved.

In Santamaria *et al.* [74] the authors present an interactive framework that helps to infer differences or similarities between biclustering results. The visualization presented contains multiple representations/display options to help bring forth the results such as Parallel Coordinates, Heatmaps, Bubble charts, TRN graphs, and Overlapper [73]. In this work, the visualizations are connected to each to help the user visualize different properties. The heatmap presented represents a single bicluster and is reordered to represent the similarities between the rows and columns, in hierarchical biclustering the heatmap is accompanied by a dendrogram. The Bubble chart is used to represent the biclusters and their properties where its position and size depending on characteristics such as size and their homogeneity. With the different visualizations working together, the researchers can extract interesting features from the biclustering results, especially the highlighting of overlapping zones that usually represent robust groups of genes and/or conditions.

Chapter 4

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 (Fig.5.3d). 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.

This chapter is structured as follows: 1) we will present a structured view on why, when and how to bicluster oncological data to understand post-surgical outcomes and variables of interest in cancer patients, 2) the dataset characteristics will be presented, 3) the data preprocessing done such as removing errors, dataset divisions, and discretization will be presented, 4) BicPAMS algorithm will be introduced, 5) the output produced and how it is presented, and visualizations implemented to facilitate the exploration of the dataset as well as feature ranking.

Figure 4.1 presents the three main steps, as a pipeline, to produce the results presented in the next chapter.

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-

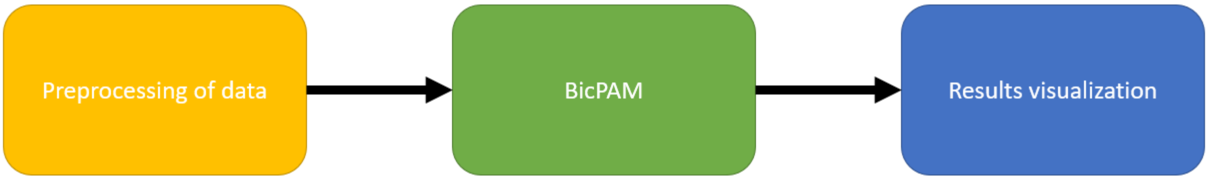


Figure 4.1: Main steps in the solution

boundaries problems[1], inability to comprehensively explore heterogeneous biomedical data), 2) find non-trivial patterns discriminative of post-surgical outcomes with constant and order-preserving coherence, and 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 [36]. 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 [36], 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 [36] 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 [33] 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 [36]. Illustrating, a quality of 80% indicates that an upper limit given 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

[1] The possibility to allow deviations from value expectations (under limits defined by the placed coherence strength) together with multi-item assignments [36] are placed to prevent discretization problems from occurring

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) [31], 2) the possibility to remove uninformative elements in data to guarantee a focus, for instance, on complications [32], and 3) support classification and regression tasks using associative models composed by discriminative patterns [35].

4.1 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.

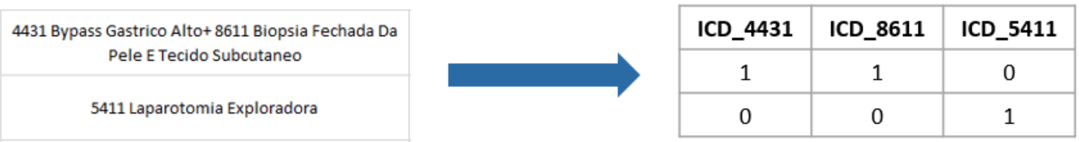
Two informative text variables, named ICD-10 and ACS procedures, exist in the dataset. These indicate the undertaken surgical standard procedures and are discussed in the next section.

4.2 Data preprocessing

This section presents the undertaken data transformations. It describes the data cleaning applied to the data, the data normalization done to some columns, and, how the data was handled for the first results and for the second results.

4.2.1 Data transformation

The dataset contained typing mistakes which were fixed, for example '$5.5.$' which should be '5.5'. A non uniform representation of missing values across the columns existed, for example 'sem dados', 'nan', 'n/a', and were all converted to a global representation '?'. Finally, text columns, which contained important information regarding the surgical interventions each patient was subjected to, were normalized into binary columns. These columns are ’ICD10 interventions’ and ’ACS procedures’. Figure 4.2 shows how the information present in ’ICD10 interventions’ variable was transformed to binary columns.



4431 Bypass Gastrico Alto+ 8611 Biopsia Fechada Da Pele E Tecido Subcutaneo
5411 Laparotomia Exploradora

ICD_4431	ICD_8611	ICD_5411
1	1	0
0	0	1

Figure 4.2: Normalization of column ’ICD10 interventions’

4.2.2 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* is above a certain threshold (later defined and discussed in Results chapter).

The dataset was partitioned 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. Figure 4.3 displays the aforementioned partitioning. A total of sixteen sub-datasets were created, four sub-datasets for each outcome considered.

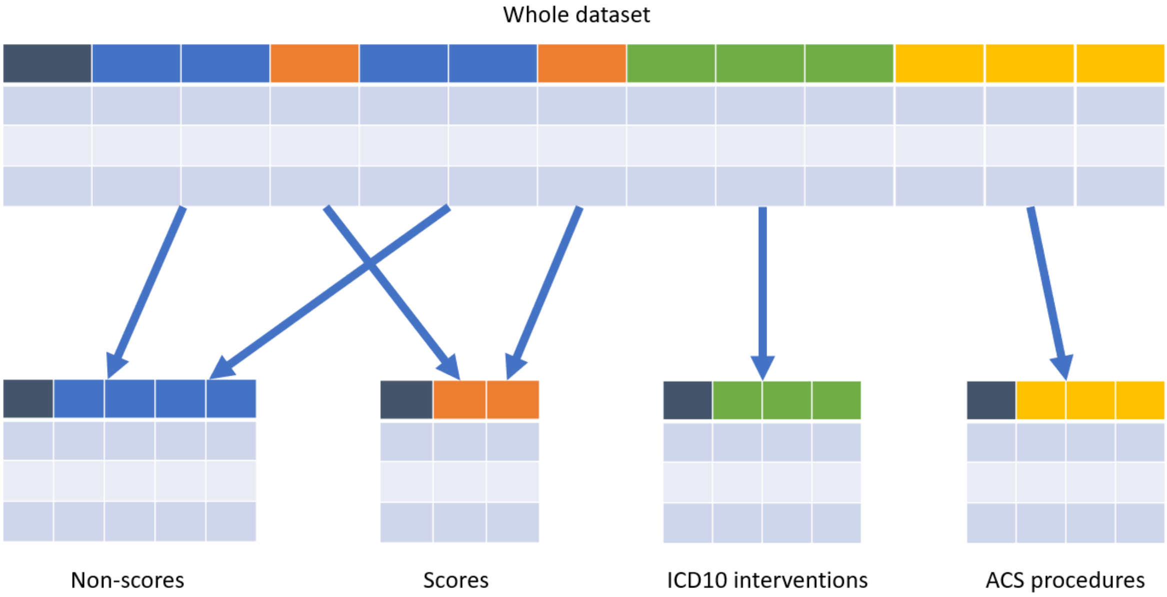


Figure 4.3: Partitioning of the dataset. Black represents a given outcome variable, blue represents non-score variables, orange represents score output variables, green represents ICD10 interventions, yellow represents ACS procedures.

Feature ranking was applied in the non-score and score output datasets to reduce the number of attributes. This preserves only the attributes most correlated with the corresponding outcome. The selected feature ranking approaches are later discussed in the Results section.

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 symbolize 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 before applying BicPAMS algorithm. The proposed DI2 (Data Discretizer) is discussed in next section. This allows for a more careful discretization of variables, as some variables follow skewed distributions, and more robust pattern discovery by BicPAMS algorithm. Finally we also created a version of the dataset where numerical variables are categorized using the range-based discretization. In range-based discretization numerical variables are put into categories of equal width based on range of the variable (from min to max).

4.3 DI2: Data Discretizer Approach

Approaches to discretization of continuous variables have long been discussed alongside their pros and cons. Altman [3] and Bennette *et al.* [10] both discuss the relevance and impact of categorizing continuous variables and reducing the cardinality of categorical variables. Liao *et al.* [50] 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 [36] and associative models such as XGBoost [17].

In this context, we propose DI2 [2], 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 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 [36].

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 [53], and 2) Kolmogorov-Smirnov goodness-of-fit test [28]. 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 correction guarantees the absence of penalizations caused by abrupt yet spurious deviations driven by

[2] https://github.com/JupitersMight/DI2

the selected histogram granularity.

In the aforementioned tests the observed distribution is matched with a theoretical continuous distribution[3] provided by the SciPy open-source library [80]. 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 [4], 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.

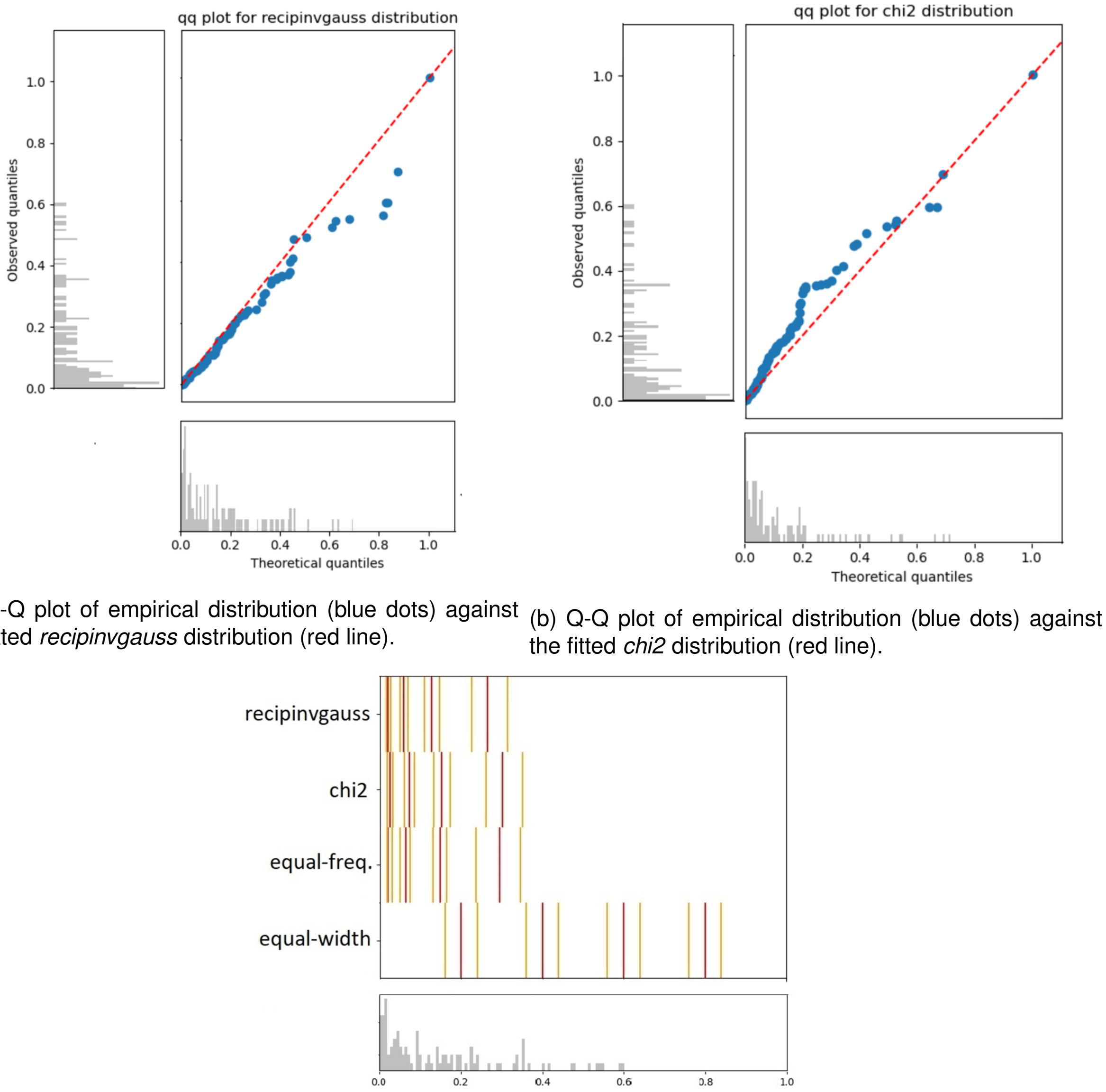
Table 4.1 shows the best fitting distribution for each continuous variable of the dataset without and with Kolmogorov-Smirnov outlier removal. Variables 'I0', 'PA500', 'A/DA', 'DR', and 'P' remained unchanged with a removal of up to 5% of outlier points. Variables 'HFS' and 'Area' produced better results 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 4.1: 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

[3] 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.

Figure 4.4: 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

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.

Tables 4.2 to 4.5 show the best distributions for each numeric variable of *IPOscore* dataset and the corresponding results for each statistical test, with and without Kolmogorov optimization.

Table 4.2: DI2 best distribution and statistical test results for each variablo for 3 and 4 categories (part 1/2).

	3 labels						4 labels					
	Without optimization			With optimization			Without optimization			With optimization		
Variables	distribution	$\tilde{\chi}^2$	KS	distribution	$\tilde{\chi}^2$	KS	distribution	$\tilde{\chi}^2$	KS	distribution	$\tilde{\chi}^2$	KS
P-Possum physiological score (%)	genhalflogistic	1.40	0.07	halfnorm	4.67	0.05	genhalflogistic	3.57	0.07	norm	3.80	0.09
P-Possum surgical severity score (%)	loggamma	1.90	0.19	dweibull	0.46	0.16	exponnorm	118.77	0.18	genhalflogistic	4.76	0.12
P-Possum morbidity (%)	dweibull	1.67	0.06	foldnorm	0.58	0.03	dweibull	11.92	0.06	gengamma	2.82	0.04
P-Possum mortality (%)	mielke	0.03	0.04	betaprime	0.02	0.02	mielke	0.55	0.04	mielke	1.54	0.02
ACS altura	frechet_r	0.68	0.07	frechet_r	0.79	0.05	foldcauchy	3.35	0.09	foldcauchy	1.92	0.09
ACS peso	chi	0.45	0.03	mielke	0.08	0.02	crystalball	0.41	0.45	norm	0.35	0.03
Serious complications (%)	kappa3	0.57	0.07	chi	1.09	0.02	frechet_r	0.11	0.03	beta	0.62	0.02
Average risk of serious complications (%)	betaprime	1.41	0.07	genlogistic	0.18	0.07	cauchy	3.52	0.11	lognorm	2.24	0.05
Any complication (%)	kappa3	0.05	0.06	kappa3	0.82	0.06	beta	0.42	0.04	beta	1.25	0.04
Average risk of any complications (%)	foldnorm	0.03	0.08	mielke	2.03	0.05	chi	1.29	0.06	beta	1.25	0.04
Pneumonia (%)	pearson3	0.06	0.03	genpareto	0.12	0.02	exponnorm	3.03	0.04	betaprime	1.38	0.03
Average risk of pneumonia (%)	cauchy	2.17	0.15	alpha	0.08	0.07	logistic	1.26	0.11	logistic	1.49	0.11
Cardiac complications (%)	pearson3	0.05	0.09	pearson3	0.24	0.07	alpha	9.99	0.06	lognorm	12.37	0.04
Average risk of cardiac complications (%)	maxwell	0.94	0.15	foldnorm	1.23	0.10	rayleigh	7.26	0.15	dgamma	6.28	0.09
Surgical infection (%)	halfcauchy	1.64	0.12	beta	4.76	0.03	mielke	8.19	0.07	chi	8.92	0.04
Average risk of surgical infection (%)	halfcauchy	6.43	0.13	halfcauchy	1.59	0.12	halfcauchy	10.32	0.13	lomax	15.36	0.07

Table 4.3: DI2 best distribution and statistical test results for each variablo for 3 and 4 categories (part 2/2).

	3 labels						4 labels					
	Without optimization			With optimization			Without optimization			With optimization		
Variables	distribution	$\tilde{\chi}^2$	KS	distribution	$\tilde{\chi}^2$	KS	distribution	$\tilde{\chi}^2$	KS	distribution	$\tilde{\chi}^2$	KS
ITU (%)	chi	0.38	0.06	nakagami	0.42	0.04	chi	1.01	0.06	nakagami	0.61	0.04
Average risk of ITU (%)	dgamma	0.81	0.11	foldcauchy	0.22	0.09	dgamma	1.39	0.11	dgamma	0.16	0.11
Venous thromboembolism (%)	hypsecant	0.27	0.12	frechet_r	0.26	0.04	foldcauchy	0.95	0.08	fatiguelife	4.24	0.04
Average risk of venous thromboembolism (%)	halfcauchy	4.59	0.17	maxwell	0.38	0.09	loggamma	2.99	0.10	loggamma	10.98	0.09
Kidney failure (%)	gilbrat	1.14	0.09	gilbrat	0.40	0.07	loglaplace	13.82	0.07	lognorm	6.56	0.07
Average risk of kidney failure (%)	moyal	4.40	0.16	beta	0.02	0.08	loglaplace	15.59	0.17	foldnorm	11.02	0.15
Ileus (%)	dgamma	0.08	0.06	triang	0.02	0.05	gennorm	20.01	0.05	gennorm	0.01	0.05
Average risk of ileus (%)	foldcauchy	10.71	0.28	gennorm	3.98	0.29	foldcauchy	10.71	0.28	gennorm	3.98	0.29
Anastomotic leak (%)	foldcauchy	0.09	0.12	foldcauchy	0.08	0.12	foldnorm	0.33	0.07	beta	0.25	0.08
Average risk of anastomotic leak (%)	loglaplace	1.37	0.38	loglaplace	1.37	0.38	trapz	44.05	0.26	dweibull	26.18	0.49
Readmission (%)	beta	0.76	0.02	chi	0.02	0.01	gengamma	0.54	0.02	nakagami	0.25	0.01
Average risk of readmission (%)	beta	2.19	0.06	frechet_r	0.67	0.04	genlogistic	10.49	0.07	t	2.20	0.06
Reoperation (%)	dgamma	0.08	0.11	mielke	0.07	0.02	mielke	0.38	0.03	pearson3	0.41	0.03
Average risk of reoperation (%)	hypsecant	0.38	0.10	hypsecant	0.33	0.09	laplace	2.61	0.11	dweibull	1.35	0.10
Death (%)	lognorm	4.52	0.08	lognorm	3.45	0.06	levy	3.10	0.06	levy	3.09	0.05
Average risk of death (%)	moyal	0.19	0.13	dgamma	0.76	0.16	loglaplace	1.01	0.08	loglaplace	2.91	0.08
Discharge to nursing or rehad facility (%)	chi	1.55	0.07	gengamma	1.60	0.03	gengamma	1.91	0.03	gengamma	0.84	0.03
Average risk of discharge to nursing or rehad facility (%)	halfcauchy	0.20	0.14	halfcauchy	1.19	0.13	logistic	1.66	0.12	dgamma	3.76	0.11
ACS forecast of hospitalization days (%)	dweibull	1.21	0.08	dgamma	2.24	0.07	foldcauchy	0.77	0.07	foldcauchy	1.68	0.06
ARISCAT total score	foldcauchy	0.04	0.40	chi2	7.83×10^{-15}	0.99	kappa3	6.44	0.38	chi2	1.6×10^{-14}	0.99
Charlson Comorbidity Index	anglit	3.22	0.19	f	8.06×10^{-15}	0.07	kappa3	23.38	0.27	rdist	2.43×10^{-14}	0.77
Survivability (10 years)	gengamma	0.03	0.41	frechet_r	0.24	0.38	frechet_r	3.48	0.41	loglaplace	26.42	0.37

Table 4.4: DI2 best distribution and statistical test results for each variablo for 5 categories (part 1/2).

	5 labels					
	Without optimization			With optimization		
Variables	distribution	$\tilde{\chi}^2$	KS	distribution	$\tilde{\chi}^2$	KS
P-Possum physiological score (%)	genhalflogistic	10.97	0.07	maxwell	4.12	0.10
P-Possum surgical severity score (%)	exponnorm	79.88	0.18	exponnorm	77.53	0.15
P-Possum morbidity (%)	genpareto	7.99	0.06	gengamma	2.77	0.04
P-Possum mortality (%)	mielke	0.94	0.03	lomax	21.17	0.03
ACS altura	t	2.23	0.05	loggamma	4.10	0.05
ACS peso	chi	3.54	0.03	crystalball	1.08	0.03
Serious complications (%)	beta	6.96	0.02	mielke	1.99	0.03
Average risk of serious complications (%)	loggamma	10.80	0.08	gennorm	12.0	0.10
Any complication (%)	pearson3	10.85	0.03	chi	1.52	0.03
Average risk of any complications (%)	gennorm	2.84	0.10	rdist	1.67	0.10
Pneumonia (%)	betaprime	8.82	0.04	expon	3.45	0.03
Average risk of pneumonia (%)	frechet_l	36.61	0.07	logistic	1.49	0.11
Cardiac complications (%)	alpha	5.88	0.06	alpha	7.03	0.04
Average risk of cardiac complications (%)	foldcauchy	8.07	0.11	dgamma	10.18	0.10
Surgical infection (%)	chi2	9.02	0.06	pareto	9.24	0.04
Average risk of surgical infection (%)	exponnorm	24.49	0.10	beta	3.70	0.07

Table 4.5: DI2 best distribution and statistical test results for each variablo for 3 and 4 categories (part 2/2).

	5 labels					
	Without optimization			With optimization		
Variables	distribution	$\tilde{\chi}^2$	KS	distribution	$\tilde{\chi}^2$	KS
ITU (%)	genpareto	6.27	0.04	nakagami	0.94	0.04
Average risk of ITU (%)	dgamma	13.33	0.11	dgamma	2.17	0.11
Venous thromboembolism (%)	foldcauchy	4.74	0.08	frechet_r	3.12	0.04
Average risk of venous thromboembolism (%)	dgamma	23.79	0.10	dweibull	20.28	0.02
Kidney failure (%)	loglaplace	7.37	0.07	halfcauchy	5.55	0.06
Average risk of kidney failure (%)	exponnorm	102.0	0.14	halflogistic	38.46	0.13
Ileus (%)	rice	0.25	0.05	frechet_r	0.07	0.05
Average risk of ileus (%)	gennorm	74.51	0.32	genpareto	7.61	0.29
Anastomotic leak (%)	cauchy	5.80	0.11	foldnorm	5.72	0.07
Average risk of anastomotic leak (%)	trapz	24.71	0.26	dweibull	15.56	0.50
Readmission (%)	beta	3.84	0.02	chi	0.05	0.01
Average risk of readmission (%)	maxwell	20.81	0.05	pearson3	1.99	0.05
Reoperation (%)	betaprime	7.47	0.02	mielke	1.09	0.02
Average risk of reoperation (%)	dgamma	1.10	0.10	fatiguelife	6.13	0.05
Death (%)	invweibull	23.44	0.06	kappa3	10.0	0.06
Average risk of death (%)	moyal	17.50	0.13	dgamma	13.11	0.16
Discharge to nursing or rehad facility (%)	gengamma	8.87	0.04	frechet_r	22.18	0.04
Average risk of discharge to nursing or rehad facility (%)	gengamma	44.44	0.08	mielke	9.83	0.05
ACS forecast of hospitalization days (%)	gumbel_r	5.47	0.05	gumbel_r	22.44	0.04
ARISCAT total score	powerlaw	157.39	0.38	chi2	2.34×10^{-14}	0.99
Charlson Comorbidity Index	kappa3	23.76	0.27	betaprime	4.71	0.72
Survivability (10 years)	beta	15.49	0.41	kappa3	31.46	0.41

4.4 BicPAMS

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 [36] proposed BicPAMS biclustering. It integrates existing principles made available by state-of-the-art pattern-based approaches with two new contributions. First, BicPAMS exhaustively mines non-constant types of biclusters, including additive and multiplicative coherencies in the presence or absence of symmetries. Second, BicPAMS provides strategies to effectively compose different biclustering structures. BicPAMS is an ordered composition of three stages: *mapping*, *mining* (pattern discovery), *closing* (post-processing).

Mapping: In *mapping* BicPAMS handles missing values and tackles noise. The normalization criteria can be applied in the context of a row, a column or the overall matrix, it also makes available a zero-mean value to allow for symmetries. Three discretization methods are available in BicPAMS, the use of fixed ranges, bins, and Gaussian, with key implications on the target solution.

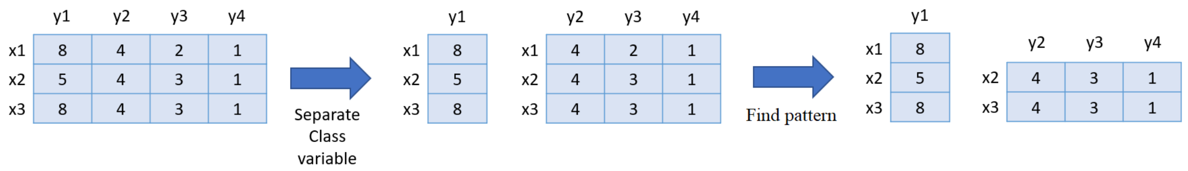
Mining: In *mining* to have adequate use of pattern mining for biclustering it relies on three points, 1) the adopted pattern-based approach to biclustering, 2) the target pattern representation, and 3) the search strategy. BicPAMS uses frequent itemsets as the default pattern-based option to biclustering. If the database is purely categorical FIM-based biclusters are perfect biclusters. If the database contains real-value variables the biclusters can handle noise since two elements with the same item may be numerically distant (due to category boundaries), but sometimes items with a close numerical distance can belong to different items. The default pattern representation used by BicPAMS is a frequent closed pattern, the set of all and maximal frequent patterns are also an option within BicPAMS. By using closed itemsets BicPAMS allows for overlapping biclusters only if a reduction on the number of columns from a specific bicluster results in a higher number of rows. The search strategy by default of BicPAMS is a variant of FP-growth that traces the set of transactions per frequent pattern, but there are other available options to deal with a large number of columns and largescale datasets are Carpenter [64] and Cobbler [65].

Closing: Finally, in the *closing* step BicPAMS has post-processing criteria that can be used to minimize two challenges of the noise dilemma, 1) being too restrictive when it comes to noise tolerance and 2) heightened levels of noise allowance. The criteria is composed of three stages: 1) *extension*, 2) *merging* and 3) *filtering*. To extend the discovery of biclusters BicPAMS has three options, 1) the use of statistical tests, 2) traditional approaches and merit functions and 3) use patterns discovered under more relaxed criteria. The merging operations will control the noise allowance and overall biclustering structure manipulation. Filtering is made possible at two levels: 1) at the bicluster level, 2) at the row-column level. The first type is required to remove duplicates and biclusters that are contained in larger biclusters. The second type can be used to exclude rows or columns from a particular bicluster to intensify its homogeneity. Clinical variables are structurally sparse, especially variables describing surgical interventions (locations and selected procedures) and post-surgical outcomes. As a consequence, an arbitrarily-high fraction of elements is missing, creating a new requirement: ability to discover patterns in the presence of highly sparse data. BicPAMS also handles missing values with three possible strategies: 1) removal, 2) replacement, and 3) handling as a special value. This is particularly relevant to guarantee that missing values do not affect our ability to discover patterns with attributes from surgical interventions and post-surgical outcomes. Patterns found for surgical interventions source in the dataset consist manly of missing values.

4.5 Extending pattern-based biclustering searches

4.5.1 Guarantees of discriminative power

BicPAMS [36] 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. To do this we first separate the class column from the rest of the dataset. Then we search for biclusters and test if they discriminate a selected class value. This pipeline is exemplified in Figure 4.5.



	y1	y2	y3	y4
x1	8	4	2	1
x2	5	4	3	1
x3	8	4	3	1

	y1
x1	8
x2	5
x3	8

	y2	y3	y4
x1	4	2	1
x2	4	3	1
x3	4	3	1

	y1
x1	8
x2	5
x3	8

	y2	y3	y4
x2	4	3	1
x3	4	3	1

Figure 4.5: Separation of class variable from rest of the dataset then discovery of a bicluster with constant assumption

The bicluster discovered in Figure 4.5 depending on the selected class value he can be discriminative of that class. If we select the class value $y_1(8)$ then the bicluster will not be discriminative of this class ($lift < 1$), according to the lift formula introduced in the Background section, but if we select $y_1(5)$ then the pattern found is discriminative of this class ($lift > 1$).

- $lift(bicluster \implies y_1(8)) = \frac{\frac{1}{3}}{\frac{2}{3}\times\frac{2}{3}} = \frac{3}{4} = 0.75$
- $lift(bicluster \implies y_1(5)) = \frac{\frac{1}{3}}{\frac{2}{3}\times\frac{1}{3}} = \frac{3}{2} = 1.5$

Interestingly, as BicPAMS dynamically changes the support threshold when exploring the data space, the presence of discriminative criteria (e.g. lift above 2.0) does not necessarily restrict the number of found patterns. If, amongst candidate patterns, only few are discriminative of a given post-surgical outcome, then BicPAMS will further explore the search space at lower support thresholds.

4.5.2 Biclustering mixed variables

The original version of BicPAMS [36] provides two important principles for handling mixed variable data: 1) 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, and 2) numeric entries per variable belong to the same pattern if they satisfy a given coherence strength (a_{ij}=α_i+η_{ij} or a_{ij}=β_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. Considering numeric variables are scaled in [0,1], a coherence strength of δ=0.2 is translated into a 5-symbol cardinality for ordinal variables with higher cardinalities. In this context, coherence strength is applied over numeric variables ($\eta_{ij} \notin [-\delta/2, \delta/2]$), while value equivalences are pursued for categorical variables ($\eta_{ij} = 0$). This allows pattern-based biclustering to be applicable over mixed variables irrespective of their domain, whether 1) nominal (e.g. demographic variables), 2) ordinal (e.g. risk scales), or 3) numerical (e.g. physiological variables).

4.6 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 (as the example depicted in Figure 4.6), or pattern descriptions. These are generally sufficient to guarantee their usability near healthcare professionals.

Figure 4.6 illustrates a pattern of categorical variables that reveals the presence of a statistically significant group of patients who died within 1 year after surgery and show the following profile: cancer is disseminated, no peritoneal contamination, the tumor is malignant, and the presence of a systemic disease. The pattern is discriminative of the 1-year survivability condition (*lift* = 1.54), and statistically significant (p-value = 4.31$\times$10^{-22} in accordance with [33]), meaning that the probability of this pattern occurring by chance is highly unlikely.

4.6.1 Clustered setting

The patterns found in the clustered setting 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). Figure 4.7 provides illustrative patterns for the first and second type of source.

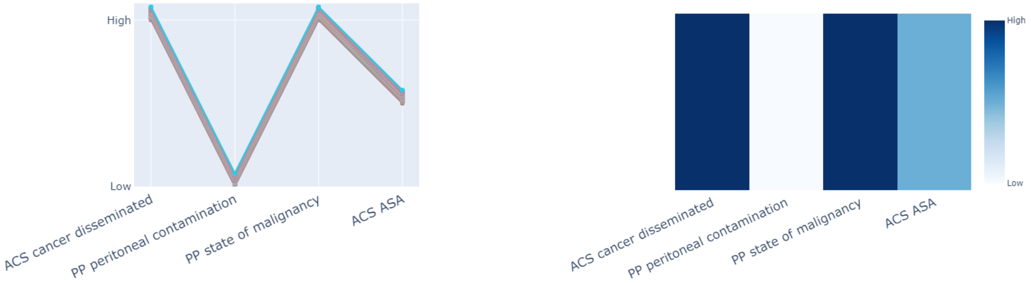
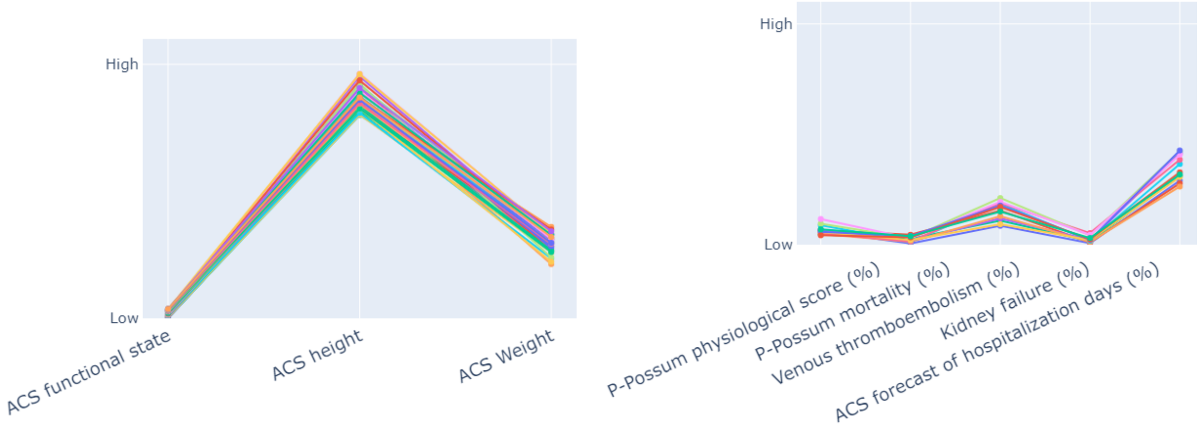


Figure 4.6: Discriminative pattern of death within 1 year after surgery: severe systemic disease (ASA class from ACS scores), high state of malignancy and dissemination (ACS score), and low peritoneal contamination risk. *Lift*=1.54 and *p-value*=4.31$\times$ 10^{-22}. Given the discriminative outcome of this pattern and the patients displaying a severe systemic disease This can potentially be associated with an advanced state of the disease.

An example of a surgical interventions pattern discriminating Clavien-Dindo III.b would be: *variables* = {Lung Lobectomy Not Classifiable Elsewhere, Lung Decortication, Total Pancreatectomy}, lift = 2.96 and *p*-value = 7.2×10^{-12}.

The patterns can be further characterized in accordance with the target variable (including its cardinality and imbalance), figure 5.3. Possible outcomes of interest include:

- complication severity (e.g. Clavien-Dindo);
- presence-absence of surgery-related complications in future or within specific time ranges;
- survivability in a given period (death or alive after a given time period after surgery);
- hospitalization needs: hospitalized period after surgery.



(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 composed of risk scores: low physiological score and less susceptible to death (P-POSSUM mortality), slightly more susceptible to a venous thromboembolism but less to kidney failure, and a medium hospitalization length forecast (ACS score). *Lift* = 2.08 and *p-value* = 1.51×10^{-8}

Figure 4.7: Constant patterns discriminative of Clavien-Dindo III.b (fig. 4.7a) and II (fig. 4.7b) class of complications. The pattern 4.7a might be correlated with malnourished patients. The pattern 4.7b demonstrates that low risk in scores correlates to low severity in complications.

4.6.2 Integrative setting

In the Integrative setting, patterns are characterized in accordance with the target variable:

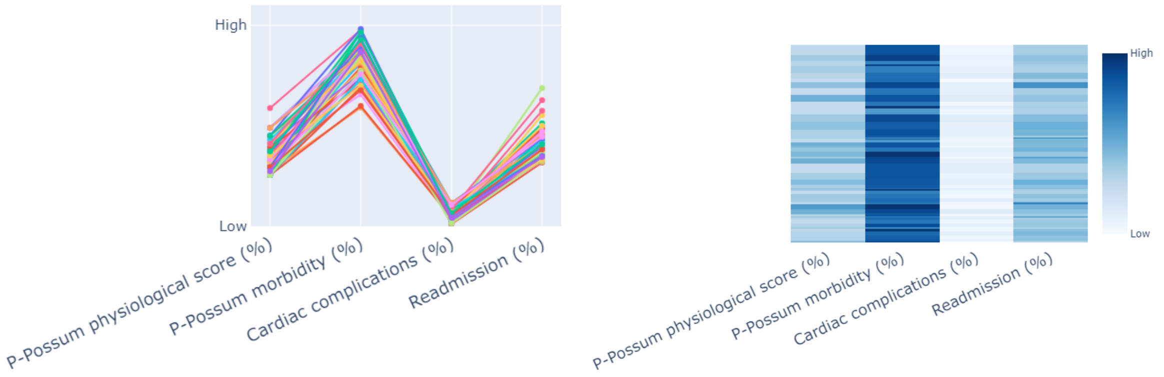
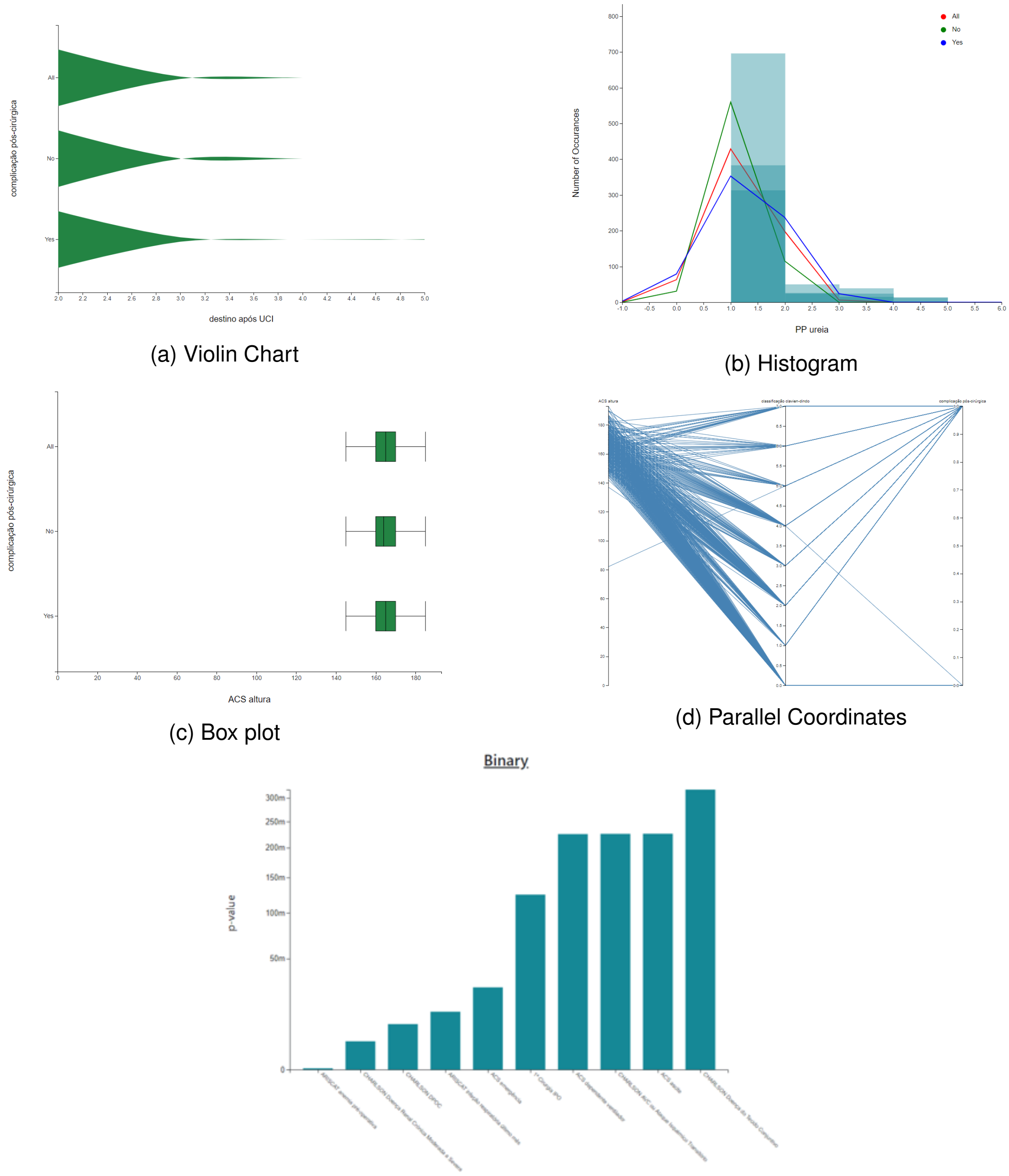


Figure 4.8: Discriminative pattern of death within 1 year after surgery: medium P-POSSUM score and high risk of morbidity (P-POSSUM), low risk of cardiac complications, and medium risk of readmission. *Lift*=2.04 and p-value=4.47$\times$10^{-29}.

- complication severity (e.g. Clavien-Dindo);
- presence-absence of surgery-related complications in future or within specific time ranges;
- survivability in a given period (death or alive after a given time period after surgery);
- hospitalization needs: hospitalized period after surgery in HDU and IPO, if the patient was in intensive care, request type anesthesia.
- provenance of patient
- reason for admission into the HDU or if he had to be readmitted
- destination after HDU/IPO
- average nursery points per day (representative of effort given by nurses to a given patient)

Since the dataset in the Integrative setting is combined there is no separation by source, only by target variables.

Other visualization to help explore the dataset were implemented. Figure 4.9a shows the violin chart visualization where it is possible to see the distribution of a variable in the whole dataset and given an outcome variable. Figure 4.9b shows the histogram visualization where it is possible to see the distribution of the variable aswell as the normal curve. Figure 4.9c displays a box plot visualization where each box plot where the lower and higher whiskers being calculated with $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$ respectively. Figure 4.9d represents a parallel coordinates visualization where in the future pattern will be visualized but as of now you can visualize how the variable selected corresponds to an outcome variable. Figure 4.9e displays a bar chart visualization where the users can visualize the feature ranking tests applied to the data.



(a) Violin Chart

(b) Histogram

(c) Box plot

(d) Parallel Coordinates

(e) Bar chart visualization

Figure 4.9: Data exploration visualizations.

Chapter 5

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 chapter 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.

5.1 Data exploration

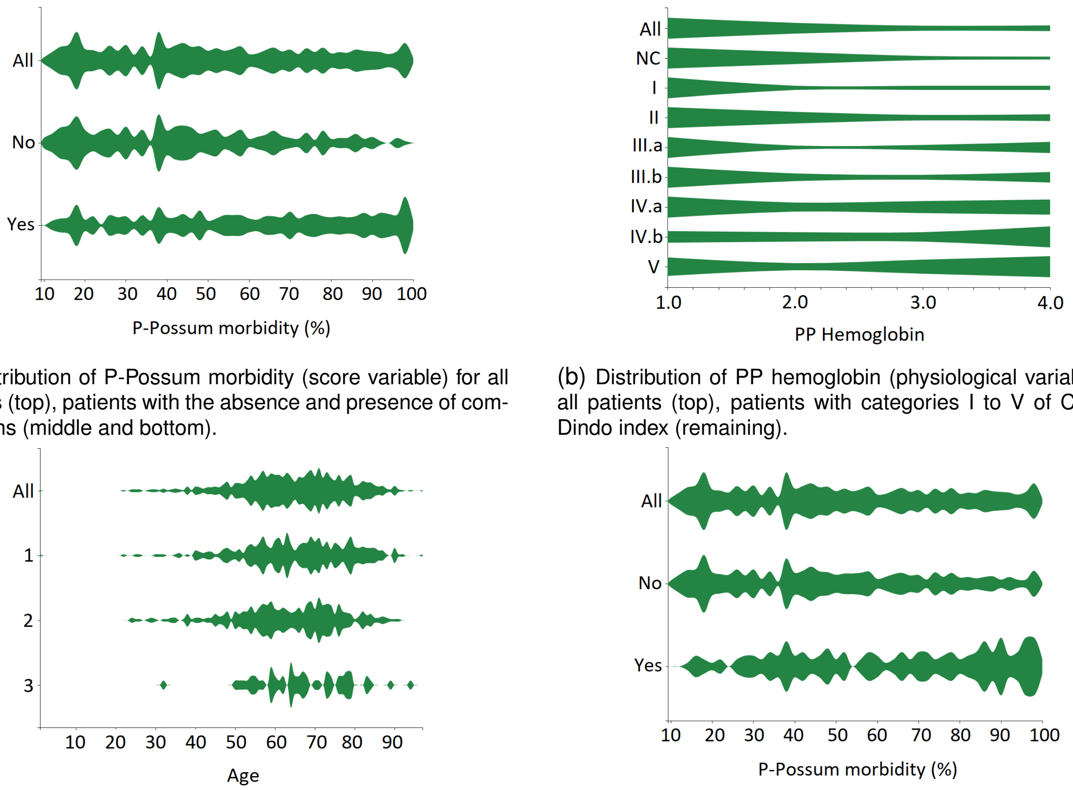
The dataset contains a considerable amount of missings, with 11 variables reaching at least 75% missing values. The data also has 47 variables where a single value occurs for at least 70% observations, two examples of this can be seen in Figures 4.9a and 4.9b. These Figures display the distribution of two attributes ”Destination after HDU” and ”PP ureia”. These Figures serve as an example of data being unbalanced in some variables.

The variables can be clustered in accordance with clinical data, patient characteristics, demographic, surgical proceduredures, and risk scores.

To better understand the impact each variable might have in the output patterns (such as frequent occurance in patterns), feature ranking tests were applied. Table 5.1 shows the top 3 variables that have the best correlation matched with each output variable in the clustered setting. $\tilde{\chi}^2$ 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 (optionally F-Regression test could also be applied to numeric variables). Figure 5.1 provides illustrative class conditioned distributions of some of the input variables in *IPOscore* data, generally showing the difficulty of discriminating post-surgical outcomes.

Table 5.1: Top ranked variables from the dataset in accordance with their ability to discriminate the target four output variables. Only the top three variables per variable type are displayed.

	Post-surgical complication	**Clavien-Dindo post-surgical index**	**Days spent at HDU**	**Death within 1 year**
binary variables $(\tilde{\chi}^2)$	ARISCAT emergent procedure	ARISCAT emergent procedure	ARISCAT emergent procedure	ARISCAT emergent procedure
	ARISCAT pre-surgery anemia	ARISCAT pre-surgery anemia	ARISCAT pre-surgery anemia	ARISCAT pre-surgery anemia
	ARISCAT respiratory infection	ARISCAT respiratory infection	ARISCAT respiratory infection	ARISCAT respiratory infection
nominal variables $(\tilde{\chi}^2)$	ARISCAT surgical incision	ARISCAT surgery duration	ARISCAT surgery duration	ARISCAT surgical incision
	ARISCAT peripheral oxygen saturation	ARISCAT surgical incision	ARISCAT peripheral oxygen saturation	ARISCAT peripheral oxygen saturation
	ARISCAT surgery duration	ARISCAT peripheral oxygen saturation	ARISCAT surgical incision	ARISCAT Age
ordinal variables (Kruskal)	P-Possum surgical severity score	P-Possum surgical severity score	Age	Age
	P-Possum physiological score	P-Possum physiological score	P-Possum surgical severity score	P-Possum physiological score
	ACS height	ACS height	P-Possum physiological score	P-Possum surgical severity score
numeric variables (ANOVA)	ACS forecast of hospitalization days	ACS forecast of hospitalization days	ACS forecast of hospitalization days	ACS forecast of hospitalization days
	Avg. risk of any complications (%)	Avg. risk of reoperation (%)	Risk of reoperation (%)	Discharge to nursing/rehab facility
	Avg. risk of serious complications (%)	Discharge to nursing/rehab facility	Risk of pneumonia (%)	Risk of Death (%)



(a) Distribution of P-Possum morbidity (score variable) for all patients (top), patients with the absence and presence of complications (middle and bottom).

(b) Distribution of PP hemoglobin (physiological variable) for all patients (top), patients with categories I to V of Clavien-Dindo index (remaining).

(c) Age distribution of all patients (top) and patients who stayed in HDU for between 0 and 1 days, 1 and 4 days, and more than 4 days.

(d) Distribution of P-Possum morbidity (risk score) for all patients (top), patients with 1-year survival (middle) and death (bottom).

Figure 5.1: Class-conditional distribution charts for the input variables P-Possum morbidity, PP hemoglobin, NAS-Points considering different outcomes of interest (presence of complication, Clavien-Dindo, hospitalization length, survivability) using violin plots.

5.2 Experimental settings

In the **clustered setting** for discriminative pattern discovery, BicPAMS algorithm is used with default parameters and varying:

- minimum lift of pattern: *lift* $\in \{1.3, 1.7, 2.0\}$
- target classes:
  - *clavien-Dindo* $\in$ {I,II,IIIa, IIIb,IVa,IVb,V}
  - *post-surgical complication* $\in${0,1}
  - *days at HDU* $\in$ {$\leq$1,]1,4], > 4}
  - *1-year death* $\in$ {yes,no}
- coherence strength (δ=$\bar{A}/|\mathcal{L}|$: $|\mathcal{L}| \in \{3, 4, 5, 7\}$)
- decreasing support until $|\mathcal{B}|$ dissimilar biclusters are found: $|\mathcal{B}|$ $\in$ {2,10,50,100,200}
- noise: 0% and up to 30% noisy elements allowed
- coherence assumptions: constant and order-preserving

Three search iterations were considered by masking the biclusters discovered after the Clustered setting to ensure a more comprehensive exploration of the data space and a focus on less-trivial patterns discriminative of surgical outcome.

In the **Integrative setting** for pattern discovery, BicPAMS algorithm is used with default parameteres and varying:

- minimum lift of pattern: *lift* $\in${[1.3,3.0]}
- minimum number of variables in the pattern: *variables* $\in$ {3,8}
- target classes:
  - *clavien-Dindo* $\in$ {I,II,IIIa, IIIb,IVa,IVb,V}
  - *post-surgical complication* $\in \{yes\}$
  - *days at HDU* $\in$ {$\leq$1,]1,2], > 2}
  - *1-year death* $\in$ {yes,no}
  - *request type anesthesia* $\in$ {associated pathology, surgical complexity}
  - *provenance* $\in$ {nursery, intensive care unit, unscheduled service}
  - *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}
  - *days at IPO* $\in$ {$<$ 7, [7, 10], $>$10}

  - *ICU* ∈ {yes}
  - *destination after HDU* ∈ {intensive care unit}
  - *average nursery points per day* ∈ {<60, 60≤}
  - *HDU readmission* ∈ {yes}
  - *destination after IPO* ∈ {death}
  - *moment of death* ∈ {[0,30[, [30-60[, [60, 365]}

- coherence strength (δ=$\bar{A}/|\mathcal{L}|$: $|\mathcal{L}| \in \{3, 4, 5\}$)
- decreasing support until $|\mathcal{B}|$ dissimilar biclusters are found: $|\mathcal{B}|$ ∈ {50, 200, 1000}
- noise: 30% noisy elements allowed
- coherence assumptions: constant
- iterations: between one and three search iterations were considered.

5.3 Clustered setting

Tables 5.2 to 5.5 synthesize the first results produced by biclustering *IPOscore* data with BicPAMS [36]. Confirming the potentialities listed in the previous chapter, BicPAMS was able to efficiently and comprehensively find a large number of homogeneous, dissimilar and statistically significant patterns able to discriminate absence/presence of post-surgical complications, Clavien-Dindo categories, 1-year survivability, and HDU hospitalization-length.

Table 5.2: Properties of the biclustering solutions found in the three partitions of clinical variables for **1-year survivability** using BicPAMS (cf. experimental setting).

configuration		Clinical variables					ICD_10				Scores (%)				
Assumption	quality	$\|L\|$	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$	$\|L\|$	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$
Constant	100%	3	396	65	3.2±0.7	82.9±35.4	7	7	2.7±0.7	21.0±13.2	3	90	86	4.0±1.0	69.4±37.1
Constant	70%	3	367	92	3.8±0.9	77.9±40.9	7	7	2.7±0.7	21.0±13.2	3	79	73	4.3±1.3	76.5±41.6
Constant	70%	4	414	62	3.3±1.1	72.3±36.8	–	–	–	–	5	137	131	3.6±1.1	41.2±21.9
Constant	70%	5	395	68	3.4±1.0	58.9±31.6	–	–	–	–	7	155	142	3.4±0.8	28.6±16.2
Order-preserving	100%	–	272	246	3.7±0.9	63.1±65.7	7	7	2.6±0.7	21.9±12.8	–	93	93	3.6±0.6	28.9±15.8
Order-preserving	70%	–	229	212	3.9±1.1	74.0±72.3	7	7	2.6±0.7	21.9±12.8	–	92	92	3.6±0.6	29.0±15.8

One can check, for instance, in the first row of Table 5.3, 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

Table 5.3: 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	$\|L\|$	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$	$\|L\|$	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$
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 5.4: Properties of the biclustering solutions found in the three partitions of clinical variables for **hospitalization length** at HDU using BicPAMS (cf. experimental setting).

configuration			Clinical variables					ICD_10				Scores (%)				
	Assumption	quality	$\|L\|$	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$	$\|L\|$	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$
days ∈ [0,1]	Constant	100%	3	243	66	3.3±0.6	107.1±55.9	62	48	2.6±0.8	4.1±3.9	3	21	13	7.5±1.3	10.7±1.3
	Constant	70%	3	214	87	3.5±0.8	103.5±72.6	61	47	2.6±0.8	4.2±4.0	3	25	18	7.8±1.8	10.8±2.6
	Constant	70%	4	224	100	3.3±1.0	111.9±87.0	–	–	–	–	5	16	11	6.2±2.3	14.1±4.4
	Constant	70%	5	286	96	3.6±0.9	84.8±62.4	–	–	–	–	7	16	14	4.6±1.7	20.6±6.3
	Order-preserving	100%	–	236	225	4.1±0.9	70.3±85.5	63	49	2.6±0.8	4.1±3.9	–	30	30	4.8±0.8	14.8±5.2
	Order-preserving	70%	–	242	225	4.2±1.1	70.6±85.2	36	32	2.6±0.9	5.1±4.4	–	29	29	4.9±1.0	15.1±5.2
days ∈]1,4]	Constant	100%	3	290	88	4.0±1.0	72.5±43.8	90	73	2.8±0.9	7.0±6.2	3	10	9	7.1±1.7	17.9±6.0
	Constant	70%	3	250	100	4.0±1.2	67.7±48.7	93	76	2.8±0.9	7.2±6.2	3	12	11	6.9±2.0	19.5±6.3
	Constant	70%	4	310	98	4.0±1.3	57.4±45.2	–	–	–	–	5	17	14	4.7±1.4	20.4±3.6
	Constant	70%	5	391	111	3.7±1.2	50.2±37.1	–	–	–	–	7	29	26	4.7±1.1	13.4±2.4
	Order-preserving	100%	–	215	205	4.4±1.0	48.7±63.2	71	60	2.7±0.8	8.2±6.3	–	52	52	4.5±0.8	8.4±3.8
	Order-preserving	70%	–	257	236	4.8±1.1	44.3±65.6	77	66	2.7±0.8	8.1±6.1	–	51	51	4.6±0.9	8.5±3.8
days ∈]4,∞[	Constant	100%	3	364	105	3.4±1.1	55.8±45.1	34	30	2.8±0.9	11.6±10.4	3	121	86	4.0±1.7	63.5±40.4
	Constant	70%	3	357	110	3.5±1.1	55.7±45.4	34	30	2.8±0.8	12.0±10.4	3	126	90	4.3±1.8	66.6±42.4
	Constant	70%	4	391	90	3.2±1.3	52.0±34.1	–	–	–	–	5	244	182	3.3±1.0	33.3±20.4
	Constant	70%	5	411	110	3.1±1.1	47.4±34.0	–	–	–	–	7	235	152	3.0±1.1	29.4±14.3
	Order-preserving	100%	–	182	173	4.0±1.0	81.6±81.6	37	34	2.8±0.8	11.1±9.9	–	40	40	3.2±0.6	61.8±21.5
	Order-preserving	70%	–	230	222	4.1±1.0	69.6±79.8	40	36	2.7±0.8	11.3±9.7	–	40	40	3.2±0.6	61.8±21.5

bicluster on average ($\mu(|I|)$), 3 variables per bicluster on average ($\mu(|J|)$) when considering a constant assumption ($|\mathcal{L}|$=3 and $\delta \in [0, \vec{\bar{A}}/|\mathcal{L}|]$), and a perfect quality (no noise).

These initial 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.

Figure 5.4 provides the details of an illustrative set of four discriminative constant patterns with dif-

Table 5.5: Properties of the biclustering solutions found in the three partitions of clinical variables for different **Clavien-Dindo classes** using BicPAMS (cf. experimental setting).

configurations			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	$\mu(\lvert J\rvert)$ $\pm\sigma(\lvert J\rvert)$	$\mu(\lvert I\rvert)$ $\pm\sigma(\lvert I\rvert)$
type I	Constant	100%	3	153	49	3.2±0.7	85.6±41.1	14	11	2.8±0.9	4.5±3.3	3	65	65	4.4±1.3	86.9±16.7
	Constant	70%	3	134	52	3.3±1.0	90.9±53.3	14	12	2.7±0.9	4.3±3.2	3	52	52	4.9±1.7	99.0±13.8
	Constant	70%	4	145	79	3.3±1.1	89.3±61.8	–	–	–	–	5	34	34	3.9±1.0	87.3±8.6
	Constant	70%	5	194	64	3.5±1.0	86.2±61.2	–	–	–	–	7	55	55	3.5±0.8	68.0±11.1
	Order-preserving	100%	–	163	163	3.7±0.6	57.6±59.7	7	7	2.6±0.7	6.3±3.0	–	135	134	3.3±0.5	64.8±13.9
	Order-preserving	70%	–	159	159	3.9±1.0	53.1±65.8	7	7	2.6±0.7	6.3±3.0	–	106	105	3.4±0.5	67.3±14.8
type II	Constant	100%	3	157	38	3.9±1.0	54.7±28.5	26	15	3.1±0.9	3.0±1.3	3	7	7	7.4±0.9	12.1±2.6
	Constant	70%	3	139	47	4.0±1.1	52.4±32.0	26	15	3.1±0.9	3.0±1.3	3	7	7	7.4±1.0	12.6±2.3
	Constant	70%	4	178	53	4.6±1.1	35.9±27.9	–	–	–	–	5	7	7	5.7±1.0	12.9±1.1
	Constant	70%	5	199	56	4.3±1.2	31.1±23.2	–	–	–	–	7	91	74	5.5±1.9	7.9±1.7
	Order-preserving	100%	–	167	165	4.5±0.8	21.5±27.4	27	16	2.8±0.5	2.9±1.2	–	8	8	5.5±1.0	15.8±2.3
	Order-preserving	70%	–	208	204	4.8±1.0	17.8±23.7	26	13	2.8±0.5	3.0±1.3	–	8	8	5.6±1.0	15.9±2.2
type III.a	Constant	100%	3	170	48	3.8±1.1	49.5±33.2	14	9	3.4±1.2	3.6±1.9	3	122	109	3.7±0.7	59.0±10.7
	Constant	70%	3	142	52	4.0±1.2	52.5±35.3	14	9	3.4±1.2	3.6±1.9	3	104	91	3.9±0.9	62.4±11.7
	Constant	70%	4	197	55	3.8±0.9	35.4±33.1	–	–	–	–	5	88	79	4.2±1.3	46.9±8.4
	Constant	70%	5	204	53	3.9±1.4	41.7±29.9	–	–	–	–	7	60	52	4.0±1.1	41.9±7.6
	Order-preserving	100%	–	130	129	4.0±1.0	43.2±40.3	8	7	2.9±0.8	4.1±1.8	–	110	102	3.7±0.5	61.8±17.3
	Order-preserving	70%	–	145	144	4.1±1.0	45.2±43.8	9	6	2.7±0.7	4.3±1.9	–	92	81	3.7±0.5	58.6±20.9
type III.b	Constant	100%	3	123	40	4.1±1.3	46.3±21.6	16	13	3.2±1.3	3.7±0.6	3	7	7	4.7±2.0	36.3±10.4
	Constant	70%	3	124	41	4.5±1.6	50.2±26.5	16	13	3.2±1.5	3.8±0.8	3	25	14	6.4±2.6	39.9±13.9
	Constant	70%	4	176	59	4.6±1.6	28.5±19.2	–	–	–	–	5	46	36	5.2±2.1	35.4±6.5
	Constant	70%	5	298	46	4.7±1.9	23.2±19.3	–	–	–	–	7	55	49	5.1±1.9	29.7±6.8
	Order-preserving	100%	–	133	131	4.1±0.9	33.3±27.7	12	8	2.5±0.7	4.1±0.3	–	20	16	3.8±0.4	39.9±4.2
	Order-preserving	70%	–	229	217	4.8±1.4	20.6±26.9	14	10	2.4±0.9	4.3±0.5	–	120	101	4.1±0.7	29.3±8.1
type IV.a	Constant	100%	3	181	52	3.3±0.9	44.8±21.9	10	10	2.3±0.5	11.2±7.3	3	69	44	3.3±0.6	41.9±11.9
	Constant	70%	3	177	49	3.5±1.2	44.4±26.5	10	10	2.3±0.5	11.2±7.3	3	74	46	4.0±0.9	45.0±16.6
	Constant	70%	4	279	73	3.9±1.2	24.3±17.1	–	–	–	–	5	38	21	3.6±0.8	46.8±6.0
	Constant	70%	5	269	66	3.8±1.0	24.0±15.8	–	–	–	–	7	58	35	3.5±0.8	34.0±5.0
	Order-preserving	100%	–	145	141	3.8±0.8	41.1±41.1	8	8	2.3±0.4	12.1±7.9	–	77	73	3.4±0.5	79.9±16.8
	Order-preserving	70%	–	142	139	3.9±1.1	40.7±41.9	8	8	2.3±0.4	12.1±7.9	–	53	51	3.6±0.6	81.8±19.0
type IV.b	Constant	100%	3	165	47	3.6±1.2	33.5±15.2	7	7	2.7±1.0	11.4±8.4	3	17	16	3.3±0.4	72.6±4.1
	Constant	70%	3	150	50	4.1±1.5	35.4±17.7	7	7	2.7±1.0	11.4±8.4	3	20	18	3.8±0.7	74.3±6.7
	Constant	70%	4	223	75	4.1±1.5	27.1±19.7	–	–	–	–	5	33	30	3.6±1.0	43.7±5.4
	Constant	70%	5	261	85	3.7±1.2	25.0±17.8	–	–	–	–	7	61	54	3.8±0.9	35.3±4.7
	Order-preserving	100%	–	141	139	3.9±0.9	47.6±47.6	5	5	4.5±0.8	15.2±7.6	–	114	108	3.5±0.5	78.1±16.8
	Order-preserving	70%	–	154	148	4.0±0.9	45.3±46.2	5	5	4.5±0.8	15.2±7.6	–	105	100	3.7±0.5	74.4±19.1
type V	Constant	100%	3	192	32	3.6±1.0	41.0±23.8	9	7	3.0±1.4	4.6±1.6	3	29	6	3.0±0.0	63.8±5.9
	Constant	70%	3	204	31	3.8±1.2	39.9±24.1	9	7	3.0±1.4	4.6±1.6	3	41	15	3.7±1.2	69.9±9.5
	Constant	70%	4	221	55	3.9±1.1	36.5±20.3	–	–	–	–	5	22	10	3.3±0.6	41.5±4.3
	Constant	70%	5	274	71	3.7±1.0	28.4±15.8	–	–	–	–	7	41	15	3.3±0.6	31.0±2.5
	Order-preserving	100%	–	154	147	3.7±1.0	45.4±44.8	4	4	2.5±0.9	5.8±1.3	–	142	115	3.5±0.5	63.2±19.7
	Order-preserving	70%	–	131	129	3.9±1.0	45.9±46.5	4	4	2.5±0.9	5.8±1.3	–	126	91	3.6±0.5	65.7±18.1

ferent target output variables: Clavien-Dindo, post-surgical complications, and hospitalization-length.

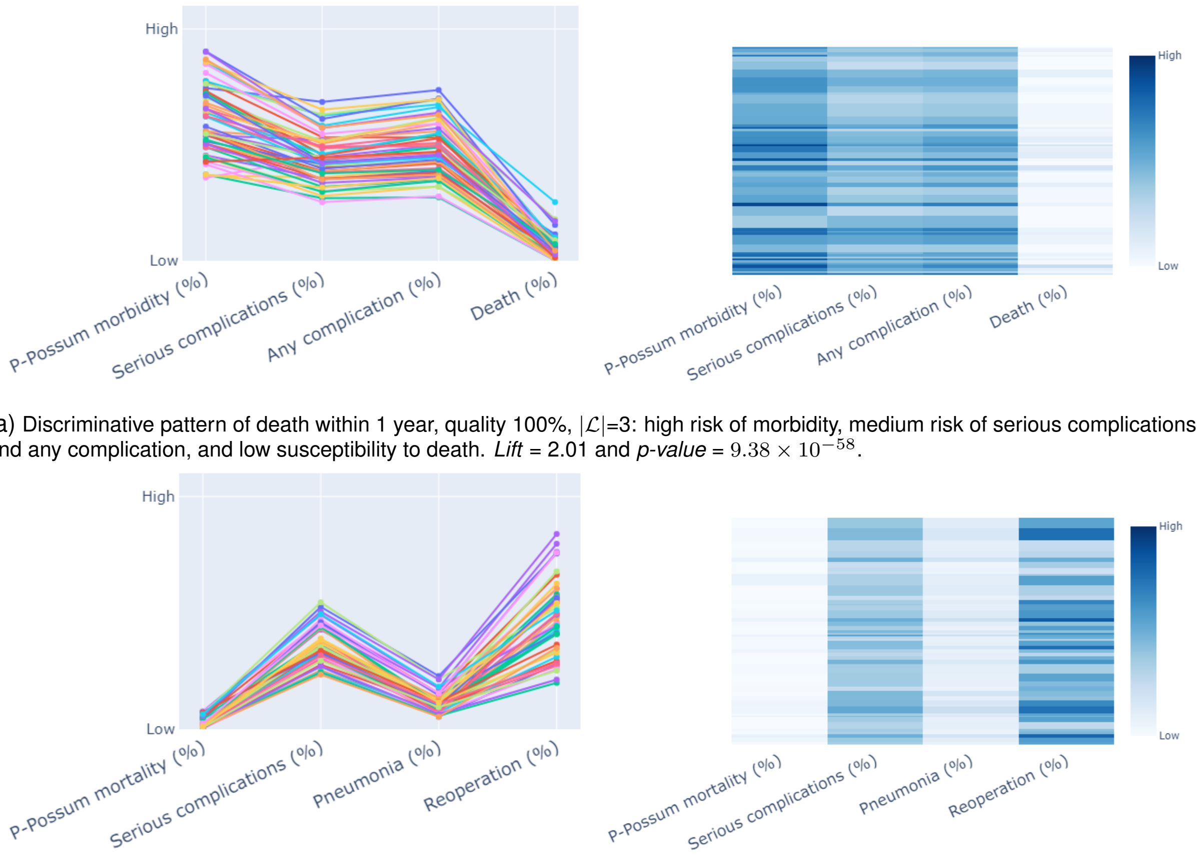
BicPAMS [36] was also applied to find less-trivial yet relevant patterns of surgical risk, patterns with order-preserving coherence assumptions. Figure 5.2 depicts order-preserving patterns for two of the targeted output variables: 1-year survivability and hospitalization-length. These coherence assumptions are useful to accommodate coherent orders and shifts in surgical risk profiles, thus being able to account for coherent differences between individuals, possibly driven by the unique biopathological aspects of the cancer, physiology of individuals and undertaken surgical procedures.

Each bicluster shows a unique pattern of performance. For instance, the constant bicluster from Figure 5.4b 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. Figure 5.4a show a pattern displaying a group of healthy patients with surgery containing only two procedures but with free content of intestine, pus or blood accumulating inside thus causing the death after surgery. This is most likely caused by the surgery going wrong.

These results motivate the relevance of finding both constant and order-preserving biclusters to find coherent factors propelling post-surgical status and hospitalization-length 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.

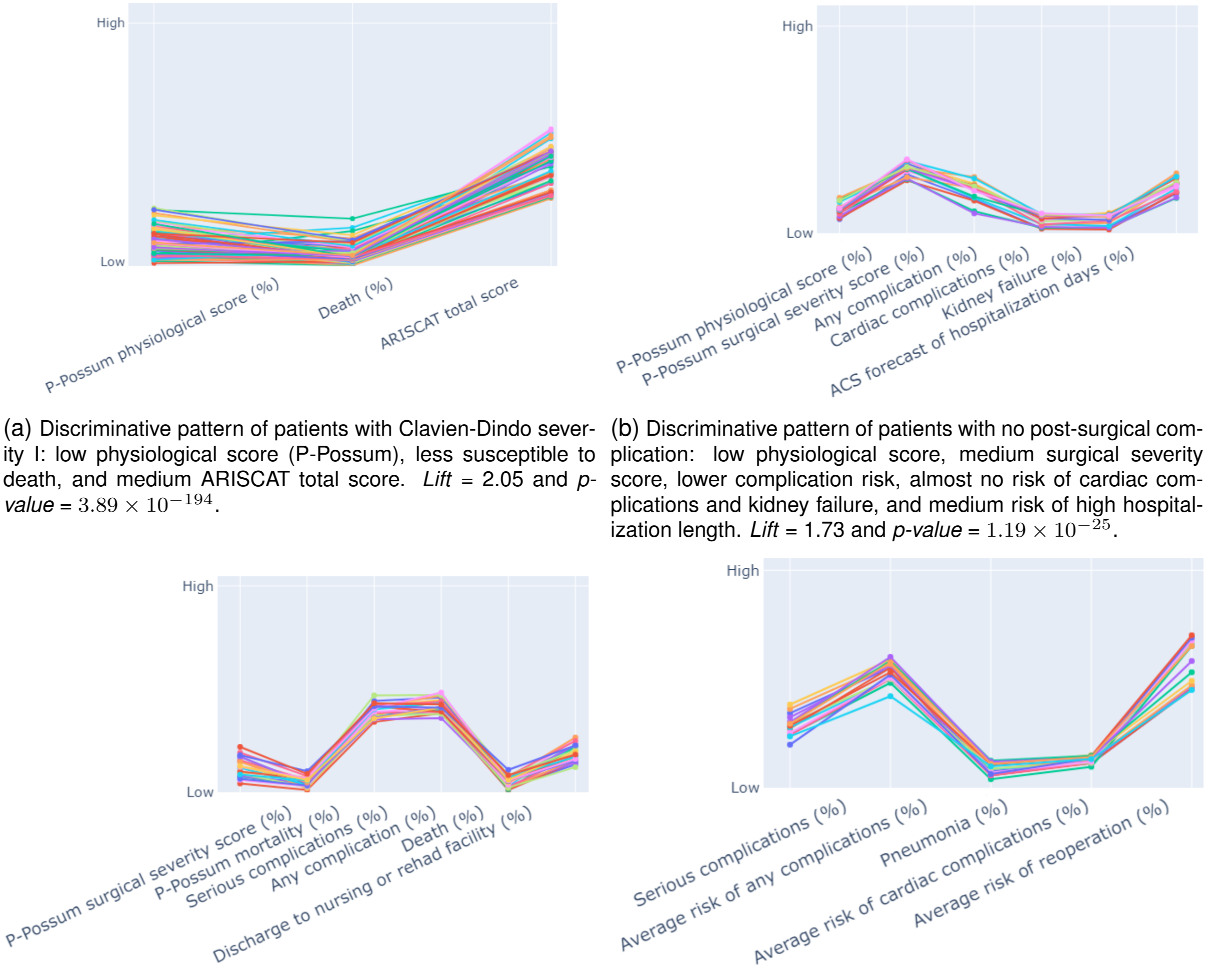
No patterns are presented for the ACS procedures partition. Despite multiple runs of BICPAMS with different criteria applied, no patterns were found. The criteria varied for pattern discovery was: discriminative power, lower number of variables in found biclusters, number of biclusters, bicluster type,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



(a) Discriminative pattern of death within 1 year, quality 100%, $|\mathcal{L}|$=3: high risk of morbidity, medium risk of serious complications and any complication, and low susceptibility to death. *Lift* = 2.01 and *p-value* = 9.38×10^{-58}.

(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}.

Figure 5.2: Two order-preserving patterns of surgical risk found within the IPOscore dataset. Pattern 5.2a portraits patients with high risk of complications but low death risk can be also subjected with post-surgical problems. Pattern 5.2b shows patients who were re-operated need more time in the HDU.



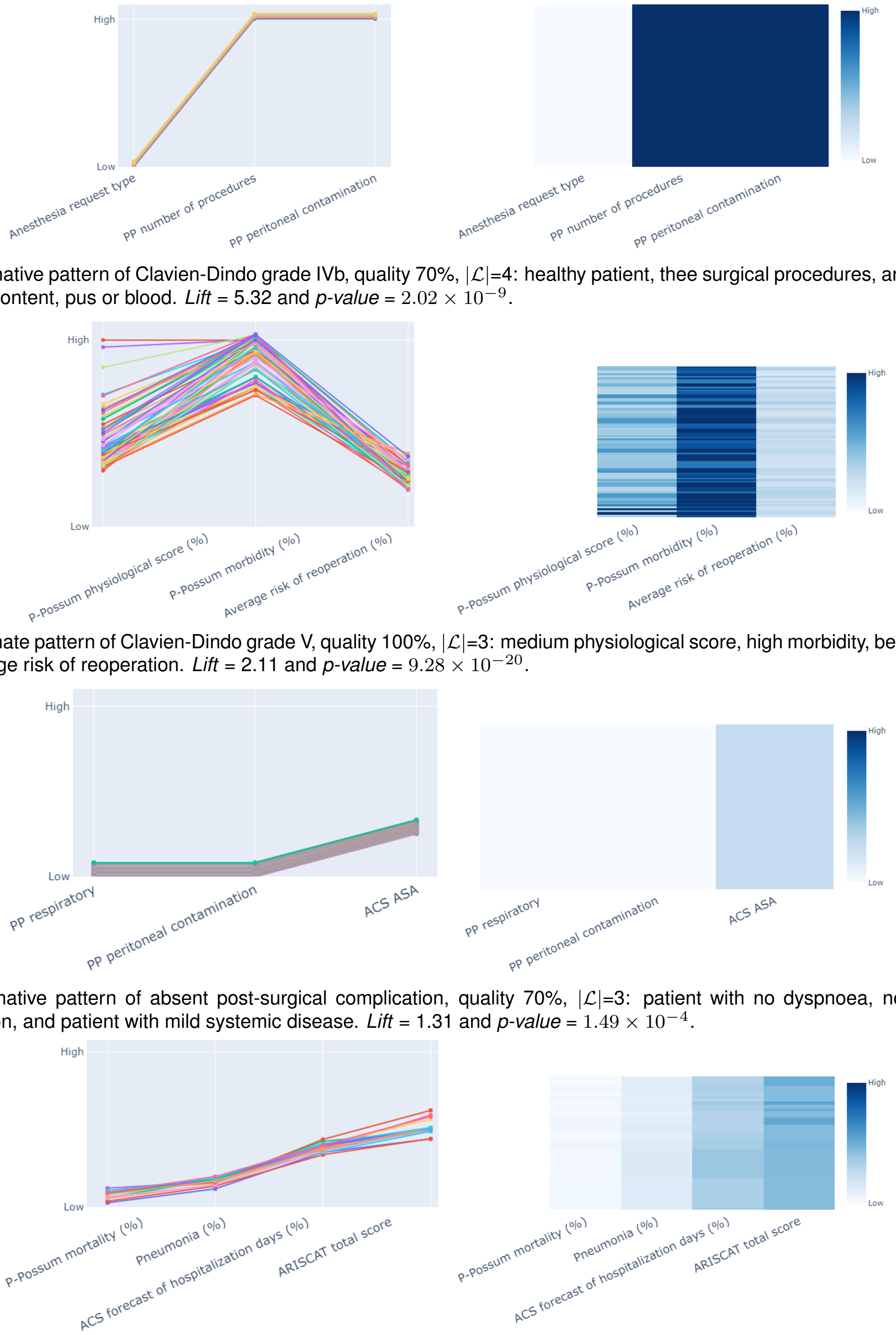
(a) Discriminative pattern of patients with Clavien-Dindo severity I: low physiological score (P-Possum), less susceptible to death, and medium ARISCAT total score. *Lift* = 2.05 and *p-value* = 3.89×10^{-194}.

(b) 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}.

(c) 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}.

(d) 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}.

Figure 5.3: Illustrative discriminative patterns of different post-surgical outcomes: Clavien-Dindo class I (a), no post-surgical complication (b), 1-year death (c) and]1,4] hospitalization-length (d). Patterns 5.3a and 5.3b both demonstrate that low scores correlate with absence of post-surgical complication or low severity in post-surgical complications. Pattern 5.3c might be correlated with patients whose surgeries went wrong turning a healthy patient susceptible to a high mortality risk. Pattern 5.3d shows that patients who have a higher risk of developing post-surgical complications are in observation longer after surgery.



(a) Discriminative pattern of Clavien-Dindo grade IVb, quality 70%, $|\mathcal{L}|$=4: healthy patient, thee surgical procedures, and presence of intestine content, pus or blood. *Lift* = 5.32 and *p-value* = 2.02×10^{-9}.

(b) 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}.

(c) 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}.

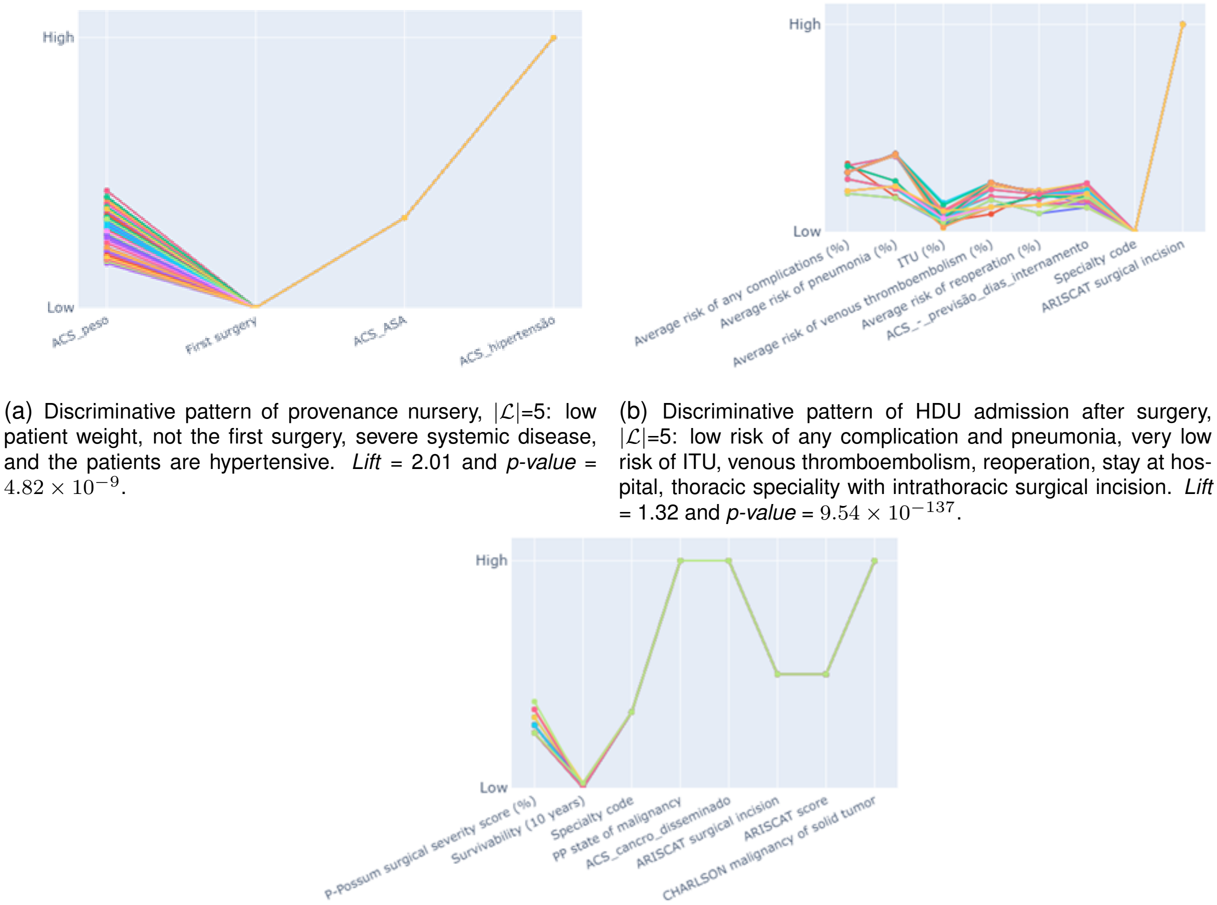
(d) Discriminative pattern of medium hospitalization length, quality 70%, $|\mathcal{L}|$=5: very low risk of mortality, low risk of pneumonia, medium hospitalization length, medium ARISCAT total score. *Lift* = 2.19 and *p-value* = 9.04×10^{-10}.

Figure 5.4: Example of constant patterns of surgical risk found within the IPOscore dataset. Pattern 5.4a correlates with healthy patients whose surgery went wrong in some way. Patterns 5.4b and 5.4c show both ends of the post-surgical complication spectrum: patients with high mortality scores and patients with regular values in clinical variables. Pattern 5.4d shows that patients with a higher risk of developing post-surgical complications need to be observed longer after surgery (in the HDU).

5.4 Integrative setting

Tables 5.6, A.1 and A.2 synthesize the results produced by biclustering *IPOscore* data with range based discretization of numeric variables, with BicPAMS [36]. Tables 5.7 and 5.8 synthesize the results with DI2 discretization of numeric variables. BicPAMS in this setting is also able to efficiently and comprehensively find a large number of homogeneous, dissimilar and statistically significant patterns able to discriminate the various target variables.

A closer inspection of Tables 5.6 to 5.8 and Tables A.1 to A.2 reveals that overall the patterns found with the discretization done by DI2 have more patients per pattern (Clavien-Dindo I,II, III.a, IV.a, IV.b, V, Post-surgery complication, days spent in HDU > 2, days spent in IPO < 7 and > 10, death after IPO, ICU after HDU, death within 30 days and between 30 and 60 days, passed through ICU, readmission into HDU, death within 1 year) and in some outcomes a higher discriminative power (death within 1 year, days spent in HDU < 1).



(a) Discriminative pattern of provenance nursery, $|\mathcal{L}|$=5: low patient weight, not the first surgery, severe systemic disease, and the patients are hypertensive. *Lift* = 2.01 and *p-value* = 4.82×10^{-9}.

(b) Discriminative pattern of HDU admission after surgery, $|\mathcal{L}|$=5: low risk of any complication and pneumonia, very low risk of ITU, venous thromboembolism, reoperation, stay at hospital, thoracic speciality with intrathoracic surgical incision. *Lift* = 1.32 and *p-value* = 9.54×10^{-137}.

(c) Discriminative pattern of anesthesia requested because of surgical complexity, $|\mathcal{L}|$=5: low P-Possum surgical severity, very low expected survivability in 10 years, digestive speciality with disseminated cancer with malignant tumor in distant metastasis, medium ARISCAT score and abdominal surgical incision. *Lift* = 1.84 and *p-value* = 3.39×10^{-126}.

Figure 5.5: Constant patterns, quality 70%, with range based discretization.

Table 5.6: **HDU admission motive**, and **type of requested anesthesia** using BicPAMS with range based discretization.

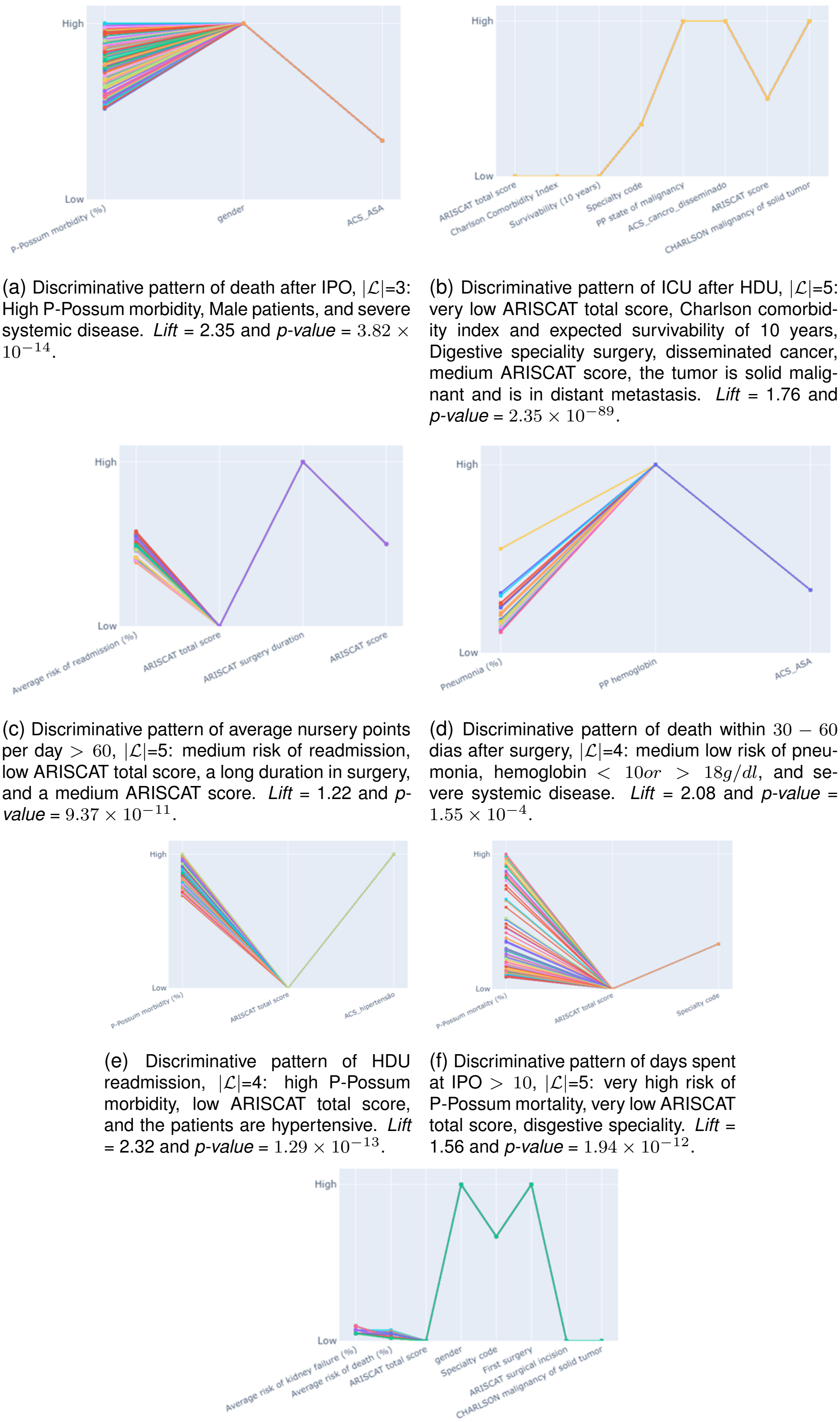
	Assumption	quality	$\lvert L\rvert$	$\lvert C\rvert$	Lift	#bics	p-value <0.001	$\mu(\lvert J\rvert)$ $\pm\sigma(\lvert J\rvert)$	$\mu(\lvert I\rvert)$ $\pm\sigma(\lvert I\rvert)$		Lift	#bics	p-value <0.001	$\mu(\lvert J\rvert)$ $\pm\sigma(\lvert J\rvert)$	$\mu(\lvert I\rvert)$ $\pm\sigma(\lvert I\rvert)$
HDU admission post-surgery	Constant	70%	3	3	1.3	25	24	5.41 ± 2.07	93.29 ± 16.42	HDU admission hemodynamic instability	3	115	104	4.61 ± 1.54	107.13 ± 18.79
	Constant	70%	3	8	1.3	34	34	8.82 ± 1.19	62.32 ± 17.38		1.3	37	37	8.76 ± 0.79	103.51 ± 43.97
	Constant	70%	4	3	3	59	57	4.61 ± 1.57	85.18 ± 16.11		3	356	329	3.92 ± 0.89	89.44 ± 22.98
	Constant	70%	4	8	1.3	12	12	9.33 ± 1.17	84.0 ± 18.16		1.3	128	128	8.63 ± 0.94	63.01 ± 15.30
	Constant	70%	5	3	3	18	18	5.28 ± 1.85	96.94 ± 13.62		3	304	279	4.00 ± 0.94	105.86 ± 17.34
	Constant	70%	5	8	1.3	22	22	9.45 ± 1.83	88.86 ± 19.39		1.3	52	52	8.90 ± 0.97	76.94 ± 26.14
HDU admission Heart	Constant	70%	3	3	2.5	75	75	5.29 ± 1.43	99.08 ± 27.22	HDU admission bleeding	2	74	67	6.37 ± 2.41	109.40 ± 14.71
	Constant	70%	3	8	1.3	43	43	8.67 ± 0.98	91.79 ± 39.47		1.3	7	7	9.0 ± 1.41	157.0 ± 25.65
	Constant	70%	4	3	2.5	55	52	5.58 ± 1.28	97.67 ± 17.12		2	239	218	5.16 ± 1.62	107.33 ± 11.88
	Constant	70%	4	8	1.3	63	63	8.62 ± 0.93	70.40 ± 26.43		1.3	20	20	8.55 ± 0.74	105.47 ± 11.39
	Constant	70%	5	3	2.5	67	65	4.26 ± 0.86	106.97 ± 12.63		2	170	156	4.97 ± 1.72	105.47 ± 11.39
	Constant	70%	5	8	1.3	35	35	8.86 ± 0.96	81.03 ± 22.74		1.3	6	6	8.83 ± 0.90	160.33 ± 40.67
HDU admission respiratory	Constant	70%	3	3	2.5	49	48	4.33 ± 1.19	135.17 ± 12.72	HDU admission post-op re-operation	2.5	41	40	3.65 ± 0.85	143.8 ± 29.82
	Constant	70%	3	8	1.3	47	47	8.57 ± 0.98	80.36 ± 30.57		1.3	42	42	8.67 ± 0.84	82.17 ± 33.80
	Constant	70%	4	3	2.5	102	96	3.83 ± 0.80	125.18 ± 18.92		2.5	83	82	3.38 ± 0.66	122.87 ± 25.78
	Constant	70%	4	8	1.3	180	180	8.81 ± 0.97	52.72 ± 20.11		1.3	47	47	8.60 ± 0.73	67.57 ± 21.54
	Constant	70%	5	3	2.5	152	144	4.56 ± 1.06	106.1 ± 13.95		2.5	72	68	3.47 ± 0.74	132.78 ± 20.98
	Constant	70%	5	8	1.3	13	13	8.62 ± 0.84	106.0 ± 24.73		1.3	50	50	8.74 ± 0.89	80.52 ± 26.17
HDU admission Age	Constant	70%	3	3	3	227	202	4.19 ± 1.03	147.54 ± 51.13	HDU admission Ischemic stroke	3	593	508	4.40 ± 1.30	173.22 ± 57.08
	Constant	70%	3	8	1.3	41	41	8.90 ± 1.20	117.17 ± 66.84		1.3	26	26	8.42 ± 0.74	123.69 ± 74.50
	Constant	70%	4	3	3	286	256	3.43 ± 0.58	158.79 ± 27.93		3	1067	884	3.95 ± 1.08	154.11 ± 46.37
	Constant	70%	4	8	1.3	52	52	8.46 ± 0.60	99.06 ± 39.07		1.3	15	15	8.93 ± 0.85	166.67 ± 39.14
	Constant	70%	5	3	2.5	180	161	3.37 ± 0.66	172.81 ± 19.39		3	964	827	4.20 ± 1.19	150.01 ± 33.54
	Constant	70%	5	8	1.3	39	39	8.59 ± 0.90	103.13 ± 34.10		1.3	13	13	9.15 ± 0.77	164.38 ± 46.19
HDU admission another pathology	Constant	70%	3	3	2.5	82	66	4.54 ± 1.38	111.15 ± 18.93	HDU admission Sepsis / septic shock / BMD	2.5	55	51	4.35 ± 1.28	173.86 ± 40.05
	Constant	70%	3	8	1.3	76	75	8.6 ± 0.73	68.85 ± 23.32		1.3	30	30	8.53 ± 0.72	101.8 ± 63.34
	Constant	70%	4	3	2.5	60	56	4.38 ± 1.16	129.23 ± 18.71		2.5	178	161	3.88 ± 0.98	129.36 ± 20.46
	Constant	70%	4	8	1.3	36	36	8.67 ± 0.85	99.11 ± 23.32		1.3	33	33	8.55 ± 0.78	88.64 ± 26.60
	Constant	70%	5	3	2.5	108	86	4.02 ± 1.05	112.17 ± 19.26		2.5	178	166	3.92 ± 0.91	136.58 ± 15.99
	Constant	70%	5	8	1.3	34	34	8.71 ± 0.86	99.71 ± 24.70		1.3	24	24	8.58 ± 0.70	95.46 ± 18.93
HDU admission co-morbidities	Constant	70%	3	3	2.5	76	72	4.21 ± 1.37	124.43 ± 20.73	anesthesia associated pathology	2.5	74	69	4.03 ± 1.22	106.71 ± 16.72
	Constant	70%	3	8	1.3	55	55	8.51 ± 0.81	65.35 ± 42.81		1.3	35	35	8.51 ± 1.02	100.71 ± 95.97
	Constant	70%	4	3	2.5	92	85	3.56 ± 0.71	109.42 ± 13.61		2.5	78	69	3.75 ± 0.81	112.25 ± 13.45
	Constant	70%	4	8	1.3	37	37	9.41 ± 1.38	77.78 ± 16.02		1.3	28	28	8.64 ± 0.89	112.39 ± 26.53
	Constant	70%	5	3	2.5	89	83	3.88 ± 0.94	106.60 ± 12.87		2.5	71	68	4.06 ± 1.01	121.0 ± 14.02
	Constant	70%	5	8	1.3	48	48	8.83 ± 0.77	69.0 ± 13.16		1.3	31	31	8.94 ± 1.16	97.87 ± 23.19
HDU admission discharge from intensive care	Constant	70%	3	3	2.5	31	31	4.68 ± 1.45	92.65 ± 8.19	anesthesia surgical complexity	1.5	20	19	4.68 ± 1.49	183.95 ± 41.37
	Constant	70%	3	8	1.3	51	51	8.61 ± 0.91	76.14 ± 33.63		1.3	31	31	8.77 ± 1.00	95.97 ± 66.59
	Constant	70%	4	3	2.5	42	38	3.61 ± 0.67	88.71 ± 8.47		1.5	60	58	3.78 ± 1.05	144.26 ± 38.37
	Constant	70%	4	8	1.3	28	28	8.82 ± 0.89	74.0 ± 8.77		1.3	37	37	8.57 ± 0.75	90.24 ± 36.60
	Constant	70%	5	3	2.5	40	39	4.23 ± 1.00	71.05 ± 7.85		1.5	73	69	3.69 ± 1.10	159.94 ± 30.16
	Constant	70%	5	8	1.3	21	21	8.90 ± 0.92	74.90 ± 13.57		1.3	50	50	8.72 ± 0.92	97.86 ± 39.81

Table 5.7: **Clavien-Dindo classes**, **presence of post-surgery complication**, **days spent in HDU** and **days spent in IPO** using BicPAMS with DI2 discretization.

	Assumption	quality	$\|L\|$	$\|C\|$	Lift	#bics	p-value	$\mu(\|J\|)$	$\mu(\|I\|) \pm\sigma(\|I\|)$		Lift	#bics	p-value	$\mu(\|J\|)$	$\mu(\|I\|) \pm\sigma(\|I\|)$
							<0.001	$\pm\sigma(\|J\|)$					<0.001	$\pm\sigma(\|J\|)$	
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

Table 5.8: **Destination after IPO**, **destination after HDU**, **moment of death after surgery** and **readmission into HDU**, **death after surgery within 1 year**, **admitted into ICU**, **average nursery care points** 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	$\mu(\lvert J\rvert)$ $\pm\sigma(\lvert J\rvert)$	$\mu(\lvert I\rvert) \pm\sigma(\lvert I\rvert)$
Destination after IPO : Death	Constant	70%	3	3	2	203	198	3.50 ± 0.88	183.52 ± 36.15	readmission into HDU	2	41	41	3.70 ± 0.91	179.78 ± 20.12
	Constant	70%	3	8	1.3	15	15	9.93 ± 1.91	167.4 ± 25.96		1.3	6	6	9.16 ± 1.67	159.5 ± 30.40
	Constant	70%	4	3	2	201	194	3.44 ± 0.76	136.84 ± 30.07		2	68	61	3.68 ± 0.91	140.55 ± 19.64
	Constant	70%	4	8	1.3	18	18	8.55 ± 0.90	88.05 ± 22.77		1.3	27	27	9.40 ± 1.59	106.14 ± 24.84
	Constant	70%	5	3	2	189	174	3.34 ± 0.64	116.29 ± 22.89		2	70	59	3.27 ± 0.54	117.16 ± 19.29
	Constant	70%	5	8	1.3	18	18	8.77 ± 0.92	71.72 ± 13.66		1.3	7	7	11.28 ± 0.87	113.14 ± 15.23
Destination after HDU : Death	Constant	70%	3	3	2	54	52	4.11 ± 1.22	169.81 ± 17.27	Death after surgery within 1 year	2	45	43	5.27 ± 1.71	158.53 ± 18.81
	Constant	70%	3	8	1.3	6	6	9.33 ± 1.37	166.0 ± 32.07		1.3	5	5	10.0 ± 1.26	190.8 ± 21.97
	Constant	70%	4	3	2	65	59	3.52 ± 0.79	138.37 ± 18.28		2	56	52	4.0 ± 1.16	135.42 ± 15.60
	Constant	70%	4	8	1.3	9	9	8.89 ± 1.28	101.11 ± 16.36		1.3	7	7	8.85 ± 1.12	109.85 ± 23.55
	Constant	70%	5	3	2	108	95	3.41 ± 0.69	106.76 ± 19.14		2	63	58	3.44 ± 0.69	112.62 ± 15.99
	Constant	70%	5	8	1.3	29	29	8.93 ± 1.01	71.62 ± 14.52		1.3	16	16	8.87 ± 1.05	68.25 ± 12.70
Death after surgery $0-30$ (days)	Constant	70%	3	3	2	39	38	5.31 ± 2.16	49.55 ± 5.80	Admitted into ICU	2	43	43	4.34 ± 1.39	183.81 ± 25.13
	Constant	70%	3	8	1.3	10	10	9.6 ± 2.2	55.0 ± 10.68		1.3	10	10	9.4 ± 1.74	162.0 ± 33.70
	Constant	70%	4	3	2	45	41	4.63 ± 1.72	39.73 ± 3.90		2	41	40	3.6 ± 1.01	151.17 ± 18.57
	Constant	70%	4	8	1.3	16	16	9.5 ± 1.73	37.38 ± 10.06		1.3	5	5	8.6 ± 0.8	108.6 ± 25.71
	Constant	70%	5	3	2	44	38	4.05 ± 1.32	37.16 ± 3.80		2	74	73	3.41 ± 0.73	120.17 ± 19.59
	Constant	70%	5	8	1.3	16	16	8.94 ± 1.09	30.75 ± 6.83		1.3	25	25	8.72 ± 0.87	67.16 ± 9.95
Death after surgery $30-60$ (days)	Constant	70%	3	3	2	19	17	10.11 ± 3.73	33.05 ± 5.76	Average nursery care points per day : <60	1.7	71	71	5.72 ± 1.85	133.83 ± 15.56
	Constant	70%	3	8	1.3	13	13	9.77 ± 1.80	48.23 ± 12.16		1.3	19	19	9.63 ± 1.46	134.10 ± 41.20
	Constant	70%	4	3	2	104	87	4.85 ± 1.64	22.86 ± 4.19		1.7	50	47	5.19 ± 1.69	120.25 ± 15.58
	Constant	70%	4	8	1.3	34	34	9.24 ± 1.33	26.20 ± 9.43		1.3	18	18	9.22 ± 1.61	90.56 ± 30.80
	Constant	70%	5	3	2	61	56	5.30 ± 1.43	25.88 ± 4.21		1.7	13	13	4.54 ± 1.45	121.0 ± 17.19
	Constant	70%	5	8	1.3	25	25	8.96 ± 1.31	20.72 ± 7.00		1.3	8	8	9.75 ± 1.47	106.88 ± 26.01
Death after surgery $60-365$ (days)	Constant	70%	3	3	1.3	6	6	3.5 ± 0.5	36.0 ± 2.52	Average nursery care points per day : ≥ 60	1.2	5	5	3.6 ± 0.8	148.6 ± 6.71
	Constant	70%	3	8	1.1	2	2	10.0 ± 1.0	37.0 ± 1.0		1.2	1	1	8.0 ± 0.0	116.0 ± 0.0
	Constant	70%	4	3	1.3	49	29	3.72 ± 0.83	26.86 ± 3.5		1.2	117	103	3.73 ± 0.95	95.15 ± 16.05
	Constant	70%	4	8	1.3	1	1	9.0 ± 0.0	21.0 ± 0.0		1.2	2	2	8.5 ± 0.5	83.5 ± 8.5
	Constant	70%	5	3	1.3	44	38	4.05 ± 1.31	37.18 ± 3.80		1.2	96	83	3.57 ± 0.68	82.99 ± 13.34
	Constant	70%	5	8	1.3	6	6	8.83 ± 1.07	16.0 ± 3.21		1.2	2	2	8.5 ± 0.5	59.5 ± 3.5



(a) Discriminative pattern of death after IPO, $|\mathcal{L}|$=3: High P-Possum morbidity, Male patients, and severe systemic disease. *Lift* = 2.35 and *p-value* = 3.82×10^{-14}.

(b) Discriminative pattern of ICU after HDU, $|\mathcal{L}|$=5: very low ARISCAT total score, Charlson comorbidity index and expected survivability of 10 years, Digestive speciality surgery, disseminated cancer, medium ARISCAT score, the tumor is solid malignant and is in distant metastasis. *Lift* = 1.76 and *p-value* = 2.35×10^{-89}.

(c) Discriminative pattern of average nursery points per day > 60, $|\mathcal{L}|$=5: medium risk of readmission, low ARISCAT total score, a long duration in surgery, and a medium ARISCAT score. *Lift* = 1.22 and *p-value* = 9.37×10^{-11}.

(d) Discriminative pattern of death within $30 - 60$ dias after surgery, $|\mathcal{L}|$=4: medium low risk of pneumonia, hemoglobin $< 10or > 18g/dl$, and severe systemic disease. *Lift* = 2.08 and *p-value* = 1.55×10^{-4}.

(e) Discriminative pattern of HDU readmission, $|\mathcal{L}|$=4: high P-Possum morbidity, low ARISCAT total score, and the patients are hypertensive. *Lift* = 2.32 and *p-value* = 1.29×10^{-13}.

(f) Discriminative pattern of days spent at IPO > 10, $|\mathcal{L}|$=5: very high risk of P-Possum mortality, very low ARISCAT total score, disgestive speciality. *Lift* = 1.56 and *p-value* = 1.94×10^{-12}.

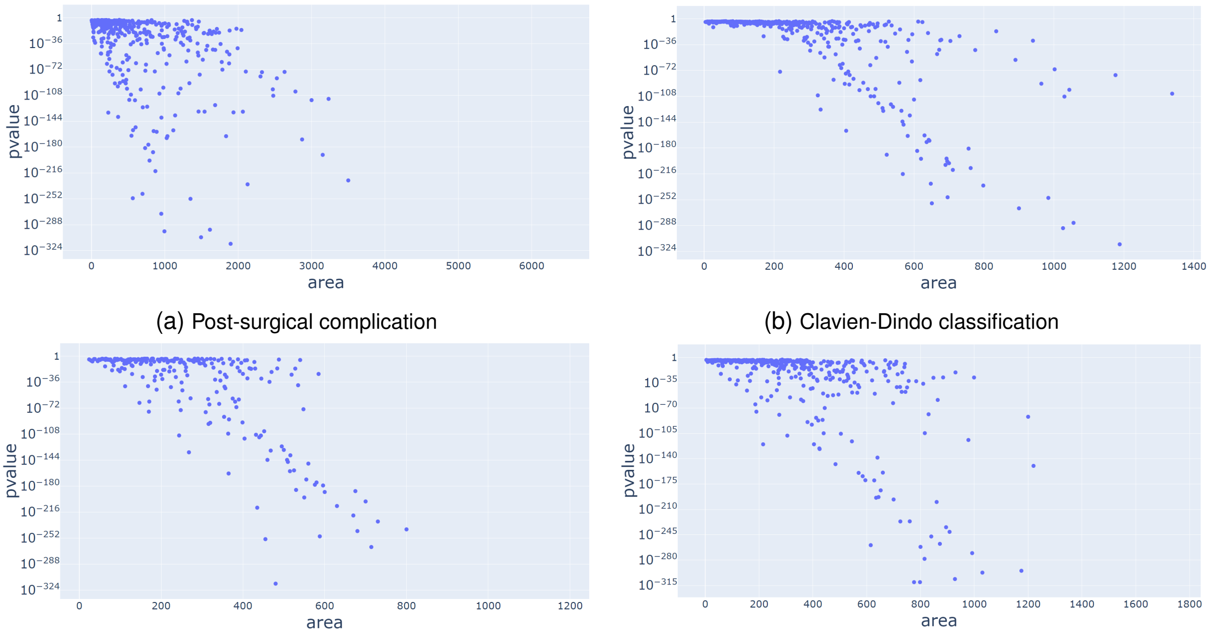
(g) Discriminative pattern of patients who passed through ICU, $|\mathcal{L}|$=5: very low risk of kidney failure, death, ARISCAT total score, male patients in head and neck speciality with a solido tumor subjected to surgery for the first time, the surgical incision is peripheral. *Lift* = 1.34 and *p-value* = 3.06×10^{-98}.

Figure 5.6: Constant patterns, quality 70%, with DI2 discretization.

5.5 Statistical Significance

As previously mentioned, Tables 5.2 to 5.8 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 [33]. To test the statistical significance of a given bicluster B, the regularities of the input matrix needs to be adequately modeled to assess the probability of occurrence of bicluster B. The probability of occurrence can be computed by testing B against approximated distributions computing Binomial tails to estimate the probability of constant bicluster B from the joint probability (for order-preserving patterns we use permutation probability) of a specific pattern to occur for a minimum number of rows.

Figure 5.7 provides four 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, 5.7a) Post-surgical complication, 5.7b) Clavien-Dindo, 5.7c) 1-year Survivability, 5.7d) HDU hospitalization-length. 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.



(a) Post-surgical complication

(b) Clavien-Dindo classification

(c) 1-year Survivability

(d) HDU hospitalization length

Figure 5.7: Statistical significance versus size of constant patterns.

5.6 Pattern actionability

The found patterns, help healthcare professionals taking decisions to better handle patients who follow the same patterns. For example, Figure 4.7a suggests malnutrition that can be tackled with specialized programs before surgery. Figure 4.8 suggests that patients with previous addressable comorbidities can be subjected to pre-habilitation to reduce the risk of death. Patterns such as in figures 5.3d, 5.4d and

5.2b 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 4.7b, 5.3a, 5.3b, 5.4b and 5.4c 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.

Chapter 6

Conclusions

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-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 mixed variables with arbitrarily skewed distributions.

The proposed solution is able to deal with the heterogeneous, structurally sparse, and high-dimensional nature of the available clinical data as the underlying pattern-based biclustering searches hold unique properties of interest: efficient yet exhaustive searches; knowledge from mixed data; ability to discover patterns with parameterizable coherence; tolerance to noise and missing data; ability to incorporate domain knowledge; absence of pattern positioning or overlapping restrictions; and sound statistical testing.

Results confirm the key role of biclustering in finding relevant discriminative patterns sensitive to highly variable physiology and biopathological traits of patients, 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.

6.1 Scientific Contributions

Throughout the development of this work three institutional presentations with IPO-Porto were made, and the following scientific contributions:

- IEEE Journal of Biomedical and Health Informatics (Revision): ”Mining pre-surgical patterns able to discriminate post-surgical outcomes in the oncological domain”. Authors: Leonardo Alexandre, Rafael S. Costa, Lúcio Lara Santos, Rui Henriques
- Bioinformatics (Submitted): ”DI2: prior-free and multi-item discretization of biomedical data and its applications”. Authors: Leonardo Alexandre, Rafael S. Costa, Rui Henriques. Software available at: https://github.com/JupitersMight/DI2

6.2 Future work

For future work we propose:

- break up the dataset into gender base and cancer specialities.
- create a full report of the new patterns found for the healthcare professionals.
- create a score based on matching patterns given a new patient for an outcome.
- implement the pattern into the already implemented view, parallel coordinates, for web visualization by healthcare professionals.
- integrate both the pattern discovery module and pattern visualization module in the online platform.

Bibliography

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[2] H. Alavi Majd, S. Shahsavari, A. R. Baghestani, S. M. Tabatabaei, N. Khadem Bashi, M. Rezaei Tavirani, and M. Hamidpour. Evaluation of plaid models in biclustering of gene expression data. *Scientifica*, 2016, 2016.

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

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

[5] 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)*, page 47, 2000.

[6] S. G. Bakshi, A. Gawri, A. R. Panigrahi, et al. Audit of pain management following emergency laparotomies in cancer patients: A prospective observational study from an indian tertiary care hospital. *Indian Journal of Anaesthesia*, 64(6):470, 2020.

[7] S. Barkow, S. Bleuler, A. Prelić, P. Zimmermann, and E. Zitzler. Bicat: a biclustering analysis toolbox. *Bioinformatics*, 22(10):1282–1283, 2006.

[8] R. Bellazzi, F. Ferrazzi, and L. Sacchi. Predictive data mining in clinical medicine: a focus on selected methods and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):416–430, 2011.

[9] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of computational biology*, 10(3-4):373–384, 2003.

[10] C. Bennette and A. Vickers. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC medical research methodology*, 12(1):21, 2012.

[11] 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*, 217(5): 833–842, 2013.

[12] Ö. Birim, A. P. Kappetein, and A. J. Bogers. Charlson comorbidity index as a predictor of long-term outcome after surgery for nonsmall cell lung cancer. *European journal of cardio-thoracic surgery*, 28(5):759–762, 2005.

[13] C. Borgelt and M. R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 51–58. IEEE, 2002.

[14] 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*, 113(6):1338–1350, 2010.

[15] Y. Cao, G. A. Bass, R. Ahl, A. Pourlotfi, H. Geijer, S. Montgomery, and S. Mohseni. The statistical importance of p-possum scores for predicting mortality after emergency laparotomy in geriatric patients. *BMC medical informatics and decision making*, 20:1–11, 2020.

[16] 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*, 40(5):373–383, 1987.

[17] 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*, pages 785–794, 2016.

[18] T. Chen, H. Wang, H. Wang, Y. Song, X. Li, and J. Wang. Possum and p-possum as predictors of postoperative morbidity and mortality in patients undergoing hepato-biliary-pancreatic surgery: a meta-analysis. *Annals of surgical oncology*, 20(8):2501–2510, 2013.

[19] W. Chen, J. Fong, C. Lind, and N. Knuckey. P–possum scoring system for mortality prediction in general neurosurgery. *Journal of clinical neuroscience*, 17(5):567–570, 2010.

[20] K.-O. Cheng, N.-F. Law, W.-C. Siu, and T. Lau. Bivisu: software tool for bicluster detection and visualization. *Bioinformatics*, 23(17):2342–2344, 2007.

[21] Y. Cheng and G. M. Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.

[22] B. K. H. Chia and R. K. M. Karuturi. Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. *Algorithms for molecular biology*, 5(1):23, 2010.

[23] G. Copeland, D. Jones, and M. Walters. Possum: a scoring system for surgical audit. *British Journal of Surgery*, 78(3):355–360, 1991.

[24] B. M. Cusworth, B. A. Krasnick, T. M. Nywening, C. A. Woolsey, R. C. Fields, M. M. Doyle, J. Liu, and W. G. Hawkins. Whipple-specific complications result in prolonged length of stay not accounted for in acs-nsqip surgical risk calculator. *HPB*, 19(2):147–153, 2017.

[25] 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*, 30(14):1615–1619, 2012.

[26] M. Esteller. Cancer epigenomics: Dna methylomes and histone-modification maps. *Nature reviews genetics*, 8(4):286, 2007.

[27] G. Fang, G. Pandey, W. Wang, M. Gupta, M. Steinbach, and V. Kumar. Mining low-support discriminative patterns from dense and high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):279–294, 2010.

[28] T. Gonzalez, S. Sahni, and W. R. Franta. An efficient algorithm for the kolmogorov-smirnov and lilliefors tests. *ACM TOMS*, 3(1):60–64, 1977.

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

[30] R. Harpaz, H. Perez, H. S. Chase, R. Rabadan, G. Hripcsak, and C. Friedman. Biclustering of adverse drug events in the fda's spontaneous reporting system. *Clinical Pharmacology & Therapeutics*, 89(2):243–250, 2011.

[31] R. Henriques and S. C. Madeira. Bic2pam: constraint-guided biclustering for biological data analysis with domain knowledge. *Algorithms for Molecular Biology*, 11(1):23, 2016.

[32] R. Henriques and S. C. Madeira. Bicnet: Flexible module discovery in large-scale biological networks using biclustering. *Algorithms for Molecular Biology*, 11(1):1–30, 2016. ISSN 1748-7188.

[33] R. Henriques and S. C. Madeira. Bsig: evaluating the statistical significance of biclustering solutions. *Data Mining and Knowledge Discovery*, 32(1):124–161, 2018.

[34] R. Henriques, C. Antunes, and S. C. Madeira. A structured view on pattern mining-based biclustering. *Pattern Recognition*, 48(12):3941–3958, 2015.

[35] R. Henriques, C. Antunes, and S. C. Madeira. A structured view on pattern mining-based biclustering. *Pattern Recognition*, 4(12):3941—-3958, 2015.

[36] R. Henriques, F. L. Ferreira, and S. C. Madeira. Bicpams: software for biological data analysis with pattern-based biclustering. *BMC bioinformatics*, 18(1):82, 2017.

[37] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nature genetics*, 31(4):370, 2002.

[38] J. Ihmels, S. Bergmann, and N. Barkai. Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993–2003, 2004.

[39] W. L. Johns, B. Strong, S. Kates, and N. K. Patel. Possum and p-possum scoring in hip fracture mortalities. *Geriatric Orthopaedic Surgery & Rehabilitation*, 11:2151459320931674, 2020.

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

[41] N. Karan, S. Siddiqui, K. S. Sharma, G. H. Pantvaidya, J. V. Divatia, and A. P. Kulkarni. Evaluation and validation of physiological and operative severity score for the enumeration of mortality and morbidity and portsmouth-possum scores in predicting morbidity and mortality in patients undergoing head and neck cancer surgeries. *Head & Neck*, 42(10):2968–2974, 2020.

[42] K. Kato, R. Yamashita, R. Matoba, M. Monden, S. Noguchi, T. Takagi, and K. Nakai. Cancer gene expression database (cged): a database for gene expression profiling with accompanying clinical information of human cancer tissues. *Nucleic acids research*, 33(suppl_1):D533–D536, 2005.

[43] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research*, 13(4):703–716, 2003.

[44] 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*, 23(2):848–856, 2018.

[45] E. Küpeli, E. Dedekargınoğlu, G. Ulubay, and M. Haberal. Association between preoperative pulmonary risk scores and postoperative complications in renal transplant recipients. *Experimental and Clinical Transplantation: Official Journal of the Middle East Society for Organ Transplantation*, 14(Suppl 3):82–86, 2016.

[46] E. Kupeli, B. Er Dedekarginoglu, G. Ulubay, F. Oner Eyuboglu, and M. Haberal. American society of anesthesiologists classification versus ariscat risk index: Predicting pulmonary complications following renal transplant. *Exp Clin Transplant*, 15(Suppl 1):208–13, 2017.

[47] C.-M. Lam, S.-T. Fan, A.-C. Yuen, W.-L. Law, and K. Poon. Validation of possum scoring systems for audit of major hepatectomy. *British journal of surgery*, 91(4):450–454, 2004.

[48] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica sinica*, pages 61–86, 2002.

[49] J. Li, H. Liu, S.-K. Ng, and L. Wong. Discovery of significant rules for classifying cancer diagnosis data. *Bioinformatics*, 19(suppl_2):ii93–ii102, 2003.

[50] S.-C. Liao and I.-N. Lee. Appropriate medical data categorization for data mining classification techniques. *Medical informatics and the Internet in medicine*, 27(1):59–67, 2002.

[51] B. Liu, C. Wan, and L. Wang. An efficient semi-unsupervised gene selection method via spectral biclustering. *IEEE Transactions on nanobioscience*, 5(2):110–114, 2006.

[52] G. Liu, A. Suchitra, H. Zhang, M. Feng, S.-K. Ng, and L. Wong. Assocexplorer: an association rule visualization system for exploratory data analysis. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1536–1539. ACM, 2012.

[53] R. Lowry. Concepts and applications of inferential statistics. 2014.

[54] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1):24–45, 2004.

[55] K. Mandal, R. Sarmah, and D. K. Bhattacharyya. Biomarker identification for cancer disease using biclustering approach: An empirical study. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(2):490–509, 2018.

[56] 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)*, pages 15–22. IEEE, 2007.

[57] V. Mazo, S. Sabaté, J. Canet, L. Gallart, M. G. de Abreu, J. Belda, O. Langeron, A. Hoeft, and P. Pelosi. Prospective external validation of a predictive score for postoperative pulmonary complications. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 121(2): 219–231, 2014.

[58] B. J. Miriovsky, L. N. Shulman, and A. P. Abernethy. Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care. *Journal of Clinical Oncology*, 30(34):4243–4248, 2012.

[59] R. Mohil, D. Bhatnagar, L. Bahadur, D. Dev, M. Magan, et al. Possum and p-possum for risk-adjusted audit of patients undergoing emergency laparotomy. *British journal of surgery*, 91(4): 500–503, 2004.

[60] T. Murali and S. Kasif. Extracting conserved gene expression motifs from gene expression data. In *Biocomputing 2003*, pages 77–88. World Scientific, 2002.

[61] A. Nebbioso, F. P. Tambaro, C. Dell'Aversana, and L. Altucci. Cancer epigenetics: moving forward. *PLoS genetics*, 14(6):e1007362, 2018.

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

[63] A. C. O'Neill, S. Bagher, M. Barandun, S. O. Hofer, and T. Zhong. Can the american college of surgeons nsqip surgical risk calculator identify patients at risk of complications following microsurgical breast reconstruction? *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 69(10):1356–1362, 2016.

[64] F. Pan, G. Cong, A. K. Tung, J. Yang, and M. J. Zaki. Carpenter: Finding closed patterns in long biological datasets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–642. ACM, 2003.

[65] F. Pan, A. K. Tung, G. Cong, and X. Xu. Cobbler: combining column and row enumeration for closed pattern discovery. In *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.*, pages 21–30. IEEE, 2004.

[66] 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*, pages 677–686, 2009.

[67] P. C. Pendharkar, J. Rodger, G. Yaverbaum, N. Herman, and M. Benner. Association, statistical, mathematical and neural approaches for mining breast cancer patterns. *Expert Systems with Applications*, 17(3):223–232, 1999.

[68] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

[69] H. Quan, B. Li, C. M. Couris, K. Fushimi, P. Graham, P. Hider, J.-M. Januel, and V. Sundararajan. Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *American journal of epidemiology*, 173(6):676–682, 2011.

[70] D. Radovanovic, B. Seifert, P. Urban, F. R. Eberli, H. Rickli, O. Bertel, M. A. Puhan, P. Erne, A. P. Investigators, et al. Validity of charlson comorbidity index in patients hospitalised with acute coronary syndrome. insights from the nationwide amis plus registry 2002–2012. *Heart*, 100(4):288–294, 2014.

[71] T. Ramkumar, V. Ng, L. Fowler, and R. Farouk. A comparison of possum, p-possum and colorectal possum for the prediction of postoperative mortality in patients undergoing colorectal resection. *Diseases of the colon & rectum*, 49(3):330–335, 2006.

[72] K. Sahara, A. Z. Paredes, K. Merath, D. I. Tsilimigras, F. Bagante, F. Ratti, H. P. Marques, O. Soubrane, E. W. Beal, V. Lam, et al. Evaluation of the acs nsqip surgical risk calculator in elderly patients undergoing hepatectomy for hepatocellular carcinoma. *Journal of Gastrointestinal Surgery*, 24(3):551–559, 2020.

[73] R. Santamaría, R. Therón, and L. Quintales. Bicoverlapper: a tool for bicluster visualization. *Bioinformatics*, 24(9):1212–1213, 2008.

[74] R. Santamaría, R. Therón, and L. Quintales. A visual analytics approach for understanding biclustering results from microarray data. *BMC bioinformatics*, 9(1):247, 2008.

[75] A. Serin and M. Vingron. Debi: Discovering differentially expressed biclusters using a frequent itemset approach. *Algorithms for Molecular Biology*, 6(1):1–12, 2011.

[76] M. Tez, Ö. Yoldaş, E. Gocmen, B. Külah, and M. Koc. Evaluation of p-possum and cr-possum scores in patients with colorectal cancer undergoing resection. *World journal of surgery*, 30(12):2266–2269, 2006.

[77] A. Thahir, R. Pinto-Lopes, S. Madenlidou, L. Daby, and C. Halahakoon. Mortality risk scoring in emergency general surgery: Are we using the best tool? *Journal of Perioperative Practice*, page 1750458920920133, 2020.

[78] M. L. van Zeeland, I. P. O. Genovesi, J.-W. R. Mulder, P. R. Strating, A. S. Glas, and A. F. Engel. Possum predicts hospital mortality and long-term survival in patients with hip fractures. *Journal of Trauma and Acute Care Surgery*, 70(4):E67–E72, 2011.

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

[80] 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*, 17(3):261–272, 2020.

[81] T. Voskuijl, M. Hageman, and D. Ring. Higher charlson comorbidity index scores are associated with readmission after orthopaedic surgery. *Clinical Orthopaedics and Related Research®*, 472(5): 1638–1644, 2014.

[82] R. A. Weinberg. *The Biology of Cancer: Second International Student Edition*. WW Norton & Company, 2013.

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

[84] M. J. Zaki, W. Meira Jr, and W. Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

Appendix A

Appendix

A.1 Clustered setting data description:

Missing data:

Score ACS

”Ileus (%)”: 90% missing values

”Average risk of ileus (%)”: 90% missing values

”fuga anastomótica (%)”: 90% missing values

”Average risk of anastomotic leak (%)”: 90% missing values

Score Charson

”Charlson Comorbidity Index”: 97% missing values

Attributes:

”Provenance” contains 85% of value 1,0

”UCI motive of admission” contains 74% of value 1,0

”Type of surgery” contains 87% of value 1,0

”Destination after UCI” contains 96% of value 2,0

”Pre-surgery chemotherapy” contains 73% of value 0,0

”readmission into HDU” contains 91% of value 0,0

”Destination after IPO” contains 96% of value 1,0

”Death within 1 year” contains 83% of value 0,0

Score PP 10/18

”PP respiratory” contains 71% of value 1,0

”PP leukocytes” contains 84% of value 1,0

”PP urea” contains 87% of value 1,0

”PP sodium” contains 93% of value 1,0

”PP potassium” contains 95% of value 1,0

”PP glasgow scale” contains 100% of value 1,0

”PP surgery type” contains 100% of value 3,0

”PP number of procedures” contains 90% of value 1,0

”PP peritoneal contamination” contains 89% of value 1,0

”PP CEPOP surgery classification” contains 89% of value 1,0

Score ACS 13/21

”ACS functional state” contains 81% of value 1,0

”ACS emergency surgery” contains 89% of value 1,0

”ACS steroids” contains 80% of value 1,0

”ACS ascites” contains 94% of value 1,0

”ACS systemic sepsis” contains 94 of value 1,0

”ACS ventilator dependent” contains 96% of value 1,0

”ACS diabetes” contains 75% of value 1,0

”ACS ICC” contains 79% of value 1,0

”ACS dyspnoea” contains 77% of value 1,0

”ACS smoker” contains 77% of value 1,0

”ACS DPOC” contains 81% of value 1,0

”ACS dialysis” contains 98% of value 1,0

”ACS acute renal failure” contains 89% of value 1,0

ARISCAT 2/7

”ARISCAT Age” contains 77% of value 2,0

”ARISCAT SpO2 ” contains 93% of value 1,0

CHARLSON 14/15

”CHARLSON diabetes” contains 73% of value 0,0

”CHARLSON liver disease” contains 95% of value 0,0

”CHARLSON malignancy of solid tumor” contains 79% of value 2,0

”CHARLSON AIDS” contains 100% of value 0,0

”CHARLSON chronic kidney disease (Moderate to severe)” contains 86% of value 0,0

”CHARLSON cardiac insufficiency” contains 88% of value 0,0

”CHARLSON myocardial infarction” contains 98% of value 0,0

”CHARLSON DPOC” contains 87% of value 0,0

”CHARLSON Doença Vascular periférica” contains 97% of value 0,0

”CHARLSON stroke or Transient Ischemic Attack” contains 93% of value 0,0

”CHARLSON dementia” contains 99% of value 0,0

”CHARLSON Doença do Tecido Conjuntivo” contains 100% of value 0,0

”CHARLSON connective Tissue Disease” contains 98% of value 0,0

Numerical attributes

Age: Average 64.72 / Standard deviation 13.13 / Confidence(95) [63.84 , 65.61] / patients 847

Days spent at HDU: Average 1.95 / Standard deviation 1.83 / Confidence(95) [1.83 , 2.081] / patients 847

Days spent at IPO: Average 19.42 / Standard deviation 23.41 / Confidence(95) [17.83 , 21] / patients 840

Total points of NAS: Average 130.12 / Standard deviation 120.27 / Confidence(95) [121.39 , 138.85] / patients 733

Points NAS per day: Average 64.42 / Standard deviation 14.29 / Confidence(95) [63.39 , 65.46] / patients 733

P-Possum physiological score (%): Average 22.05 / Standard deviation 7.6 / Confidence(95) [21.53 , 22.57] / patients 823

P-Possum surgical severity score (%): Average 14.06 / Standard deviation 4.51 / Confidence(95) [13.75 , 14.37] / patients 823

P-Possum morbidity (%): Average 52.18 / Standard deviation 25.74 / Confidence(95) [50.42 , 53.94] / patients 823

P-Possum mortality (%): Average 9.34 / Standard deviation 17.42 / Confidence(95) [8.15 , 10.53] / patients 823

ACS height: Average 164.76 / Standard deviation 8.97 / Confidence(95) [164.15 , 165.38] / patients 823

ACS weight: Average 69.87 / Standard deviation 15.15 / Confidence(95) [68.84 , 70.91] / patients 823

Serious complications (%): Average 20.68 / Standard deviation 12.73 / Confidence(95) [19.81 , 21.55] / patients 820

Average risk of serious complications (%): Average 16.24 / Standard deviation 8.48 / Confidence(95) [15.66 , 16.82] / patients 820

Any complication (%): Average 23.83 / Standard deviation 14.50 / Confidence(95) [22.83 , 24.82] / patients 820

Average risk of any complications (%): Average 18.91 / Standard deviation 9.78 / Confidence(95) [18.23 , 19.58] / patients 820

Pneumonia (%): Average 4.66 / Standard deviation 5.14 / Confidence(95) [4.31 , 5.01430707961717] / patients 817

Average risk of pneumonia (%): Average 2.87 / Standard deviation 2.34 / Confidence(95) [2.71 , 3.028] / patients 817

Cardiac complications (%): Average 1.77 / Standard deviation 2.58 / Confidence(95) [1.59 , 1.95] / patients 820

Average risk of cardiac complications (%): Average 0.8 / Standard deviation 0.66 / Confidence(95) [0.75 , 0.84] / patients 820

Surgical infection (%): Average 7.32 / Standard deviation 6.22 / Confidence(95) [6.89 , 7.75] / patients 818

Average risk of surgical infection (%): Average 6.53 / Standard deviation 5.25 / Confidence(95) [6.17 , 6.89] / patients 818

ITU (%): Average 2.38 / Standard deviation 2.01 / Confidence(95) [2.24 , 2.52] / patients 820

Average risk of ITU (%): Average 1.68 / Standard deviation 1.31 / Confidence(95) [1.58 , 1.76] / patients 820

Venous thromboembolism (%): Average 2.25 / Standard deviation 1.79 / Confidence(95) [2.13 , 2.38] / patients 819

Average risk of venous thromboembolism (%): Average 1.52 / Standard deviation 1.03 / Confidence(95) [1.45 , 1.59] / patients 819

Kidney failure (%): Average 1.39 / Standard deviation 2.33 / Confidence(95) [1.22 , 1.56] / patients 725

Average risk of kidney failure (%): Average 0.76 / Standard deviation 0.706 / Confidence(95) [0.71 , 0.81] / patients 724

Ileus (%): Average 24.33 / Standard deviation 10.19 / Confidence(95) [22.11 , 26.54] / patients 85

Average risk of ileus (%): Average 19.14 / Standard deviation 6.48 / Confidence(95) [17.73 , 20.54] / patients 85

Anastomotic leak (%): Average 4.39 / Standard deviation 1.99 / Confidence(95) [3.96 , 4.82] / patients 85

Average risk of anastomotic leak (%): Average 3.53 / Standard deviation 0.99 / Confidence(95) [3.32 , 3.75] / patients 85

Readmission (%): Average 12.08 / Standard deviation 6.59 / Confidence(95) [11.63 , 12.53] / patients 817

Average risk of readmission (%): Average 8.81 / Standard deviation 3.92 / Confidence(95) [8.54 , 9.08] / patients 816

Reoperation (%): Average 6.52 / Standard deviation 4.76 / Confidence(95) [6.19 , 6.85] / patients 820

Average risk of reoperation (%): Average 5.63 / Standard deviation 3.95 / Confidence(95) [5.36 , 5.90] / patients 819

Death (%): Average 5.94 / Standard deviation 13.77 / Confidence(95) [5 , 6.88] / patients 820

Average risk of death (%): Average 1.13 / Standard deviation 1.45 / Confidence(95) [1.03 , 1.23] / patients 819

Discharge to nursing or rehad facility (%): Average 16.78 / Standard deviation 19.72 / Confidence(95) [15.43 , 18.14] / patients 818

Average risk of discharge to nursing or rehad facility (%): Average 7 / Standard deviation 6.37 / Confidence(95) [6.564 , 7.44] / patients 817

ACS forecast of hospitalization days (%): Average 8.097 / Standard deviation 5.46 / Confidence(95) [7.72 , 8.47] / patients 818

ARISCAT total score: Average 28.74 / Standard deviation 14.61 / Confidence(95) [27.74 , 29.74] / patients 823

Charlson Comorbidity Index: Average 5.96 / Standard deviation 1.88 / Confidence(95) [5.22 , 6.71] / patients 28

Survivability (10 years): Average 0.24 / Standard deviation 0.31 / Confidence(95) [0.21 , 0.27] / patients 379

There are 374 patients (complications yes)
There are 449 patients (complications no)

94 patients (complications yes, quimio antes yes)
125 patients (complications no, quimio antes yes)
280 patients (complications yes, quimio antes no)
324 patients (complications no, quimio antes no)

197 patients (complications yes, first surgery yes)
272 patients (complications no, first surgery yes)
177 patients (complications yes, first surgery no)
177 patients (complications no, first surgery no)

75 patients (complications yes, cirurgia urgente yes)
15 patients (complications no, cirurgia urgente yes)
299 patients (complications yes, cirurgia urgente no)
434 patients (complications yes, cirurgia urgente no)

29 patients (complications yes) (torax speciality)
86 patients (complications no) (torax speciality)
160 patients (complications no) (digestive speciality)
181 patients (complications yes) (digestive speciality)
87 patients (complications yes) (head speciality)
103 patients (complications no) (head speciality)
77 patients (complications yes) (another speciality)
100 patients (complications no) (another speciality)

84 patients (complications yes, Death within 1 year yes)
51 patients (complications no, Death within 1 year yes)
288 patients (complications yes, Death within 1 year no)
398 patients (complications no, Death within 1 year no)

44 patients (complications yes) (readmission into HDU)
13 patients (complications no) (readmission into HDU)
330 patients (complications yes) (no readmission into HDU)
436 patients (complications no) (no readmission into HDU)

Location: colon : 97 patients, stomach : 66 patients, liver : 48 patients, lungs : 102 patients, larynx : 43 patients, rectum : 49 patients, oral cavity : 70 patients speciality: digestive : 349 patients, torax : 115 patients, general : 127 patients, head and neck : 67 patients

A.2 Clustered setting bicluster extra analysis

Each cell represents the % of a variable within all the biclusters found and the value it takes. For example, in Figure A.1 first line after 3 labels we can see that ”P-POSSUM mortality” appears in 100% of the biclusters found and in those biclusters 100% of the time appears with category 1.

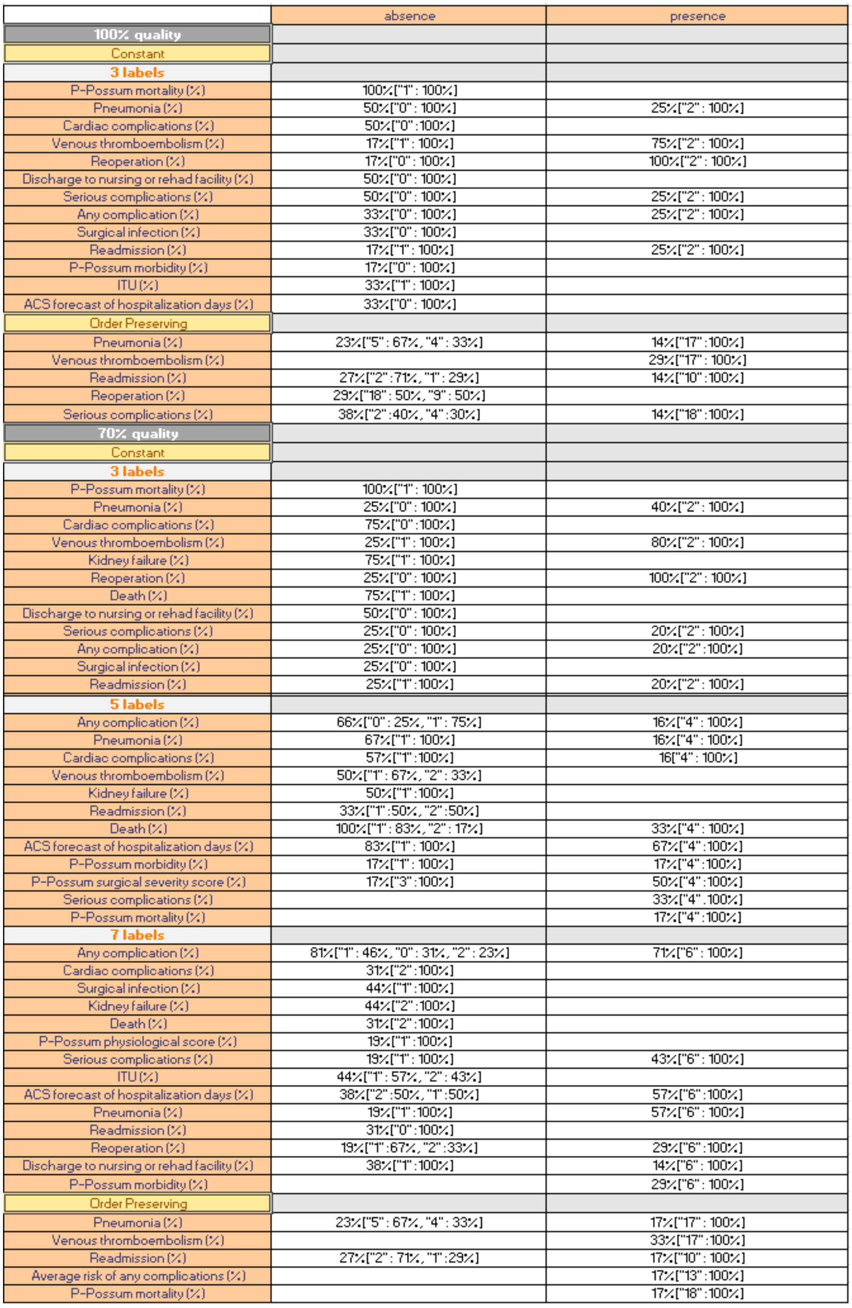
A.2.1 Post-surgical complication

A.2.2 Death within 1 year

A.2.3 Days spent at HDU

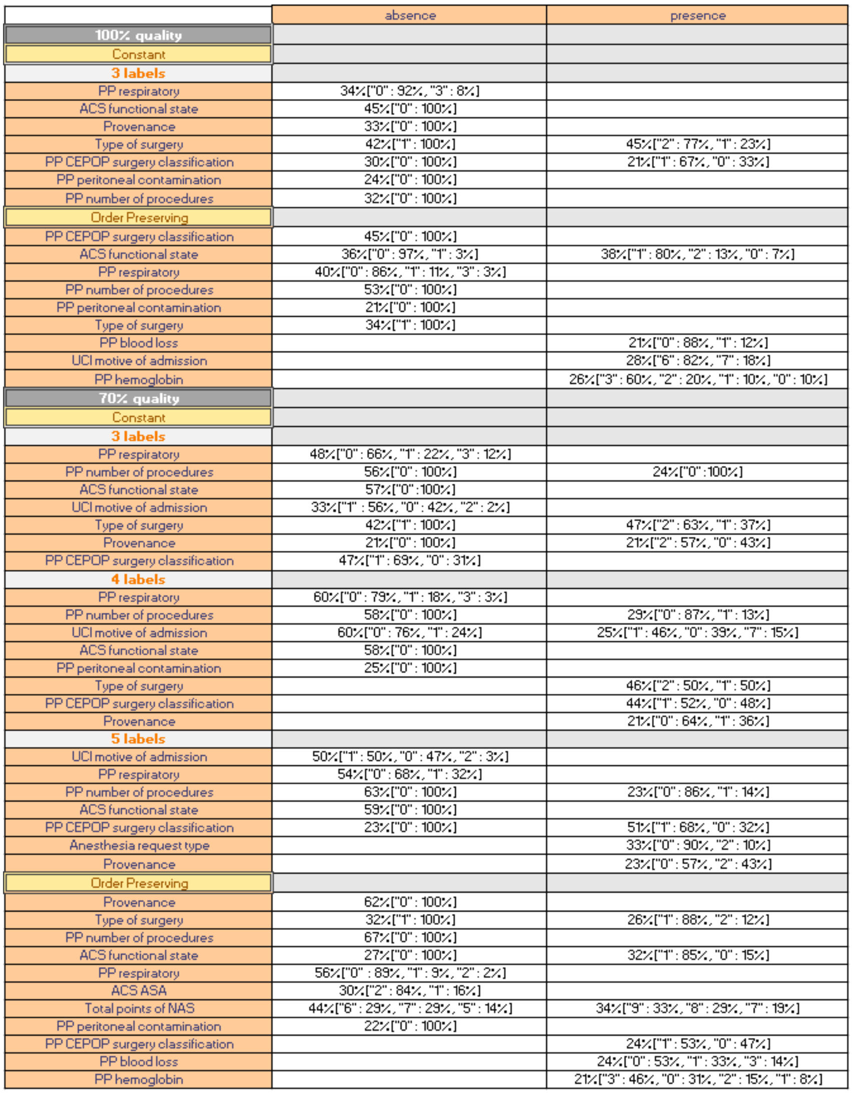
A.2.4 Clavien-Dindo

A.3 Range based extra tables



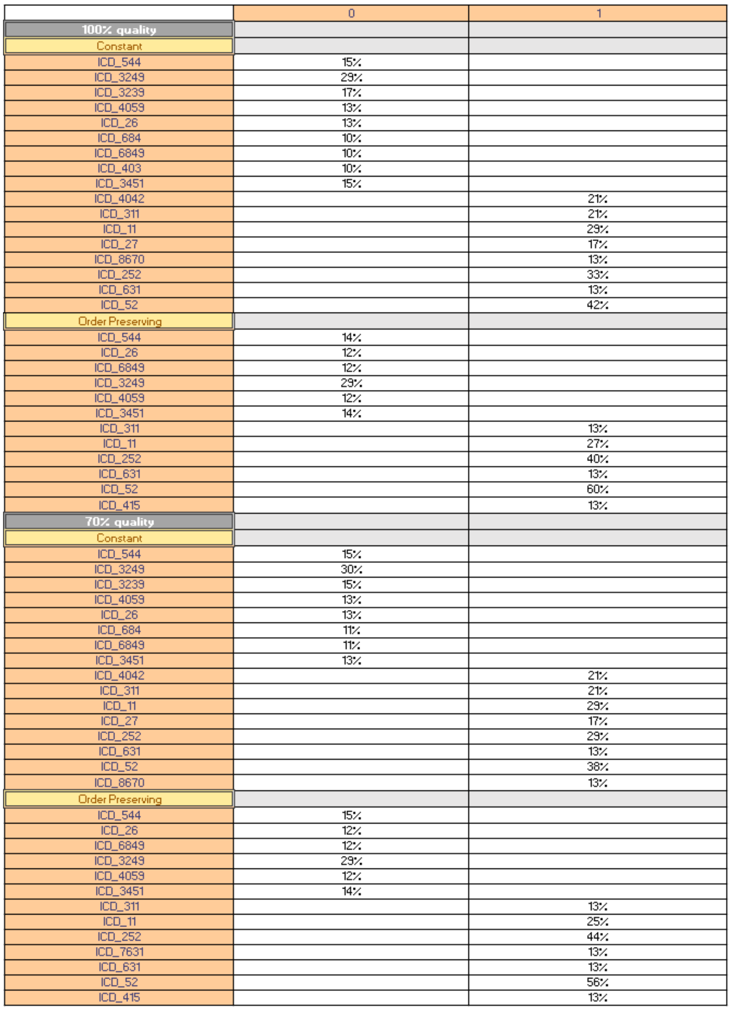
	absence	presence
100% quality		
Constant		
3 labels		
P-Possum mortality (%)	100%["1" : 100%]	
Pneumonia (%)	50%["0" : 100%]	25%["2" : 100%]
Cardiac complications (%)	50%["0" :100%]	
Venous thromboembolism (%)	17%["1" : 100%]	75%["2" : 100%]
Reoperation (%)	17%["0" : 100%]	100%["2" : 100%]
Discharge to nursing or rehad facility (%)	50%["0" : 100%]	
Serious complications (%)	50%["0" : 100%]	25%["2" : 100%]
Any complication (%)	33%["0" : 100%]	25%["2" : 100%]
Surgical infection (%)	33%["0" : 100%]	
Readmission (%)	17%["1" : 100%]	25%["2" : 100%]
P-Possum morbidity (%)	17%["0" : 100%]	
ITU (%)	33%["1" : 100%]	
ACS forecast of hospitalization days (%)	33%["0" : 100%]	
Order Preserving		
Pneumonia (%)	23%["5" : 67%, "4" : 33%]	14%["17" :100%]
Venous thromboembolism (%)		29%["17" : 100%]
Readmission (%)	27%["2" :71%, "1" : 29%]	14%["10" :100%]
Reoperation (%)	29%["18" : 50%, "9" : 50%]	
Serious complications (%)	38%["2" :40%, "4" :30%]	14%["18" :100%]
70% quality		
Constant		
3 labels		
P-Possum mortality (%)	100%["1" : 100%]	
Pneumonia (%)	25%["0" : 100%]	40%["2" : 100%]
Cardiac complications (%)	75%["0" :100%]	
Venous thromboembolism (%)	25%["1" : 100%]	80%["2" : 100%]
Kidney failure (%)	75%["1" : 100%]	
Reoperation (%)	25%["0" : 100%]	100%["2" : 100%]
Death (%)	75%["1" : 100%]	
Discharge to nursing or rehad facility (%)	50%["0" : 100%]	
Serious complications (%)	25%["0" : 100%]	20%["2" : 100%]
Any complication (%)	25%["0" : 100%]	20%["2" :100%]
Surgical infection (%)	25%["0" : 100%]	
Readmission (%)	25%["1" :100%]	20%["2" : 100%]
5 labels		
Any complication (%)	66%["0" : 25%, "1" : 75%]	16%["4" : 100%]
Pneumonia (%)	67%["1" : 100%]	16%["4" : 100%]
Cardiac complications (%)	57%["1" :100%]	16["4" : 100%]
Venous thromboembolism (%)	50%["1" : 67%, "2" : 33%]	
Kidney failure (%)	50%["1" :100%]	
Readmission (%)	33%["1" :50%, "2" :50%]	
Death (%)	100%["1" : 83%, "2" : 17%]	33%["4" : 100%]
ACS forecast of hospitalization days (%)	83%["1" : 100%]	67%["4" :100%]
P-Possum morbidity (%)	17%["1" : 100%]	17%["4" :100%]
P-Possum surgical severity score (%)	17%["3" :100%]	50%["4" :100%]
Serious complications (%)		33%["4" .100%]
P-Possum mortality (%)		17%["4" :100%]
7 labels		
Any complication (%)	81%["1" : 46%, "0" : 31%, "2" : 23%]	71%["6" : 100%]
Cardiac complications (%)	31%["2" :100%]	
Surgical infection (%)	44%["1" :100%]	
Kidney failure (%)	44%["2" :100%]	
Death (%)	31%["2" :100%]	
P-Possum physiological score (%)	19%["1" :100%]	
Serious complications (%)	19%["1" : 100%]	43%["6" : 100%]
ITU (%)	44%["1" : 57%, "2" : 43%]	
ACS forecast of hospitalization days (%)	38%["2" :50%, "1" :50%]	57%["6" :100%]
Pneumonia (%)	19%["1" :100%]	57%["6" : 100%]
Readmission (%)	31%["0" :100%]	
Reoperation (%)	19%["1" :67%, "2" :33%]	29%["6" :100%]
Discharge to nursing or rehad facility (%)	38%["1" :100%]	14%["6" : 100%]
P-Possum morbidity (%)		29%["6" : 100%]
Order Preserving		
Pneumonia (%)	23%["5" : 67%, "4" : 33%]	17%["17" : 100%]
Venous thromboembolism (%)		33%["17" :100%]
Readmission (%)	27%["2" : 71%, "1" :29%]	17%["10" : 100%]
Average risk of any complications (%)		17%["13" :100%]
P-Possum mortality (%)		17%["18" :100%]

Figure A.1: Occurrence of variables and their values in the biclusters of post-surgical complications, scores dataset.



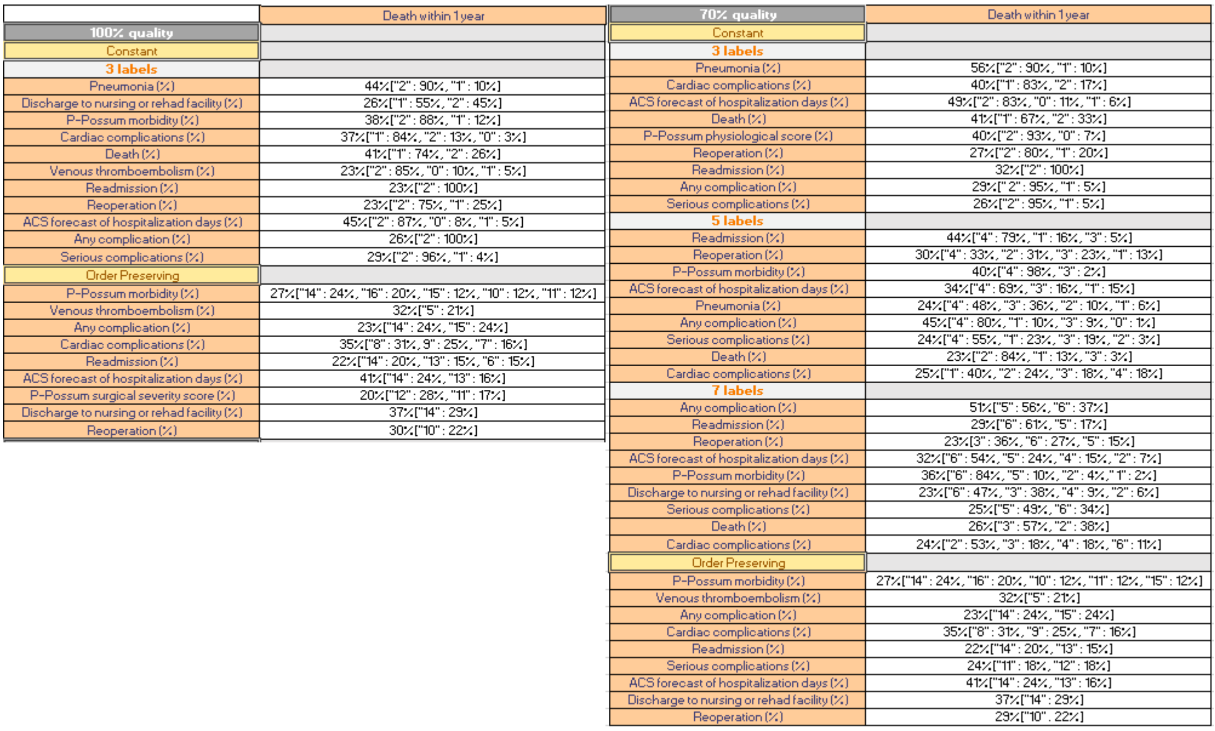
	absence	presence
100% quality		
Constant		
3 labels		
PP respiratory	34%["0" : 92%, "3" : 8%]	
ACS functional state	45%["0" : 100%]	
Provenance	33%["0" : 100%]	
Type of surgery	42%["1" : 100%]	45%["2" : 77%, "1" : 23%]
PP CEPOP surgery classification	30%["0" : 100%]	21%["1" : 67%, "0" : 33%]
PP peritoneal contamination	24%["0" : 100%]	
PP number of procedures	32%["0" : 100%]	
Order Preserving		
PP CEPOP surgery classification	45%["0" : 100%]	
ACS functional state	36%["0" : 97%, "1" : 3%]	38%["1" : 80%, "2" : 13%, "0" : 7%]
PP respiratory	40%["0" : 86%, "1" : 11%, "3" : 3%]	
PP number of procedures	53%["0" : 100%]	
PP peritoneal contamination	21%["0" : 100%]	
Type of surgery	34%["1" : 100%]	
PP blood loss		21%["0" : 88%, "1" : 12%]
UCI motive of admission		28%["6" : 82%, "7" : 18%]
PP hemoglobin		26%["3" : 60%, "2" : 20%, "1" : 10%, "0" : 10%]
70% quality		
Constant		
3 labels		
PP respiratory	48%["0" : 66%, "1" : 22%, "3" : 12%]	
PP number of procedures	56%["0" : 100%]	24%["0" :100%]
ACS functional state	57%["0" :100%]	
UCI motive of admission	33%["1" : 56%, "0" : 42%, "2" : 2%]	
Type of surgery	42%["1" : 100%]	47%["2" : 63%, "1" : 37%]
Provenance	21%["0" : 100%]	21%["2" : 57%, "0" : 43%]
PP CEPOP surgery classification	47%["1" : 69%, "0" : 31%]	
4 labels		
PP respiratory	60%["0" : 79%, "1" : 18%, "3" : 3%]	
PP number of procedures	58%["0" : 100%]	29%["0" : 87%, "1" : 13%]
UCI motive of admission	60%["0" : 76%, "1" : 24%]	25%["1" : 46%, "0" : 39%, "7" : 15%]
ACS functional state	58%["0" : 100%]	
PP peritoneal contamination	25%["0" : 100%]	
Type of surgery		46%["2" : 50%, "1" : 50%]
PP CEPOP surgery classification		44%["1" : 52%, "0" : 48%]
Provenance		21%["0" : 64%, "1" : 36%]
5 labels		
UCI motive of admission	50%["1" : 50%, "0" : 47%, "2" : 3%]	
PP respiratory	54%["0" : 68%, "1" : 32%]	
PP number of procedures	63%["0" : 100%]	23%["0" : 86%, "1" : 14%]
ACS functional state	59%["0" : 100%]	
PP CEPOP surgery classification	23%["0" : 100%]	51%["1" : 68%, "0" : 32%]
Anesthesia request type		33%["0" : 90%, "2" : 10%]
Provenance		23%["0" : 57%, "2" : 43%]
Order Preserving		
Provenance	62%["0" : 100%]	
Type of surgery	32%["1" : 100%]	26%["1" : 88%, "2" : 12%]
PP number of procedures	67%["0" : 100%]	
ACS functional state	27%["0" : 100%]	32%["1" : 85%, "0" : 15%]
PP respiratory	56%["0" : 89%, "1" : 9%, "2" : 2%]	
ACS ASA	30%["2" : 84%, "1" : 16%]	
Total points of NAS	44%["6" : 29%, "7" : 29%, "5" : 14%]	34%["9" : 33%, "8" : 29%, "7" : 19%]
PP peritoneal contamination	22%["0" : 100%]	
PP CEPOP surgery classification		24%["1" : 53%, "0" : 47%]
PP blood loss		24%["0" : 53%, "1" : 33%, "3" : 14%]
PP hemoglobin		21%["3" : 46%, "0" : 31%, "2" : 15%, "1" : 8%]

Figure A.2: Occurrence of variables and their values in the biclusters of post-surgical complications, non-scores dataset.



	0	1
100% quality		
Constant		
ICD_544	15%	
ICD_3249	29%	
ICD_3239	17%	
ICD_4059	13%	
ICD_26	13%	
ICD_684	10%	
ICD_6849	10%	
ICD_403	10%	
ICD_3451	15%	
ICD_4042		21%
ICD_311		21%
ICD_11		29%
ICD_27		17%
ICD_8670		13%
ICD_252		33%
ICD_631		13%
ICD_52		42%
Order Preserving		
ICD_544	14%	
ICD_26	12%	
ICD_6849	12%	
ICD_3249	29%	
ICD_4059	12%	
ICD_3451	14%	
ICD_311		13%
ICD_11		27%
ICD_252		40%
ICD_631		13%
ICD_52		60%
ICD_415		13%
70% quality		
Constant		
ICD_544	15%	
ICD_3249	30%	
ICD_3239	15%	
ICD_4059	13%	
ICD_26	13%	
ICD_684	11%	
ICD_6849	11%	
ICD_3451	13%	
ICD_4042		21%
ICD_311		21%
ICD_11		29%
ICD_27		17%
ICD_252		29%
ICD_631		13%
ICD_52		38%
ICD_8670		13%
Order Preserving		
ICD_544	15%	
ICD_26	12%	
ICD_6849	12%	
ICD_3249	29%	
ICD_4059	12%	
ICD_3451	14%	
ICD_311		13%
ICD_11		25%
ICD_252		44%
ICD_7631		13%
ICD_631		13%
ICD_52		56%
ICD_415		13%

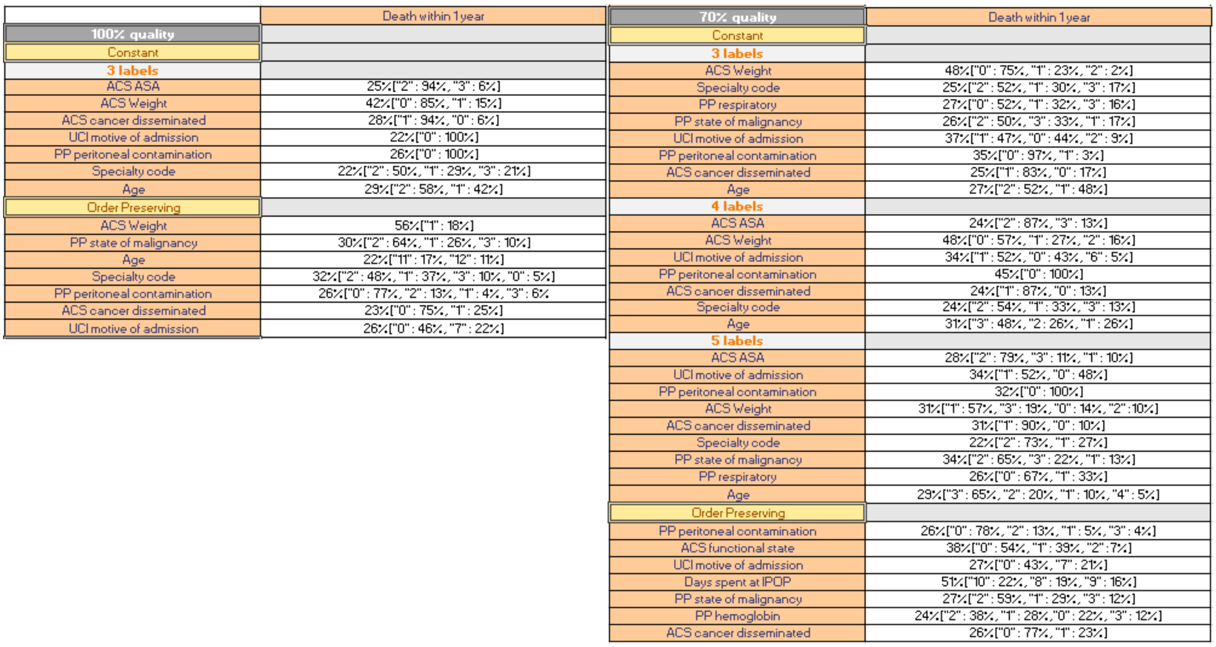
Figure A.3: Occurrence of variables and their values in the biclusters of post-surgical complications, ICD-10 dataset.



	Death within 1 year
100% quality	
Constant	
3 labels	
Pneumonia (%)	44%["2" : 90%, "1" : 10%]
Discharge to nursing or rehad facility (%)	26%["1" : 55%, "2" : 45%]
P-Possum morbidity (%)	38%["2" : 88%, "1" : 12%]
Cardiac complications (%)	37%["1" : 84%, "2" : 13%, "0" : 3%]
Death (%)	41%["1" : 74%, "2" : 26%]
Venous thromboembolism (%)	23%["2" : 85%, "0" : 10%, "1" : 5%]
Readmission (%)	23%["2" : 100%]
Reoperation (%)	23%["2" : 75%, "1" : 25%]
ACS forecast of hospitalization days (%)	45%["2" : 87%, "0" : 8%, "1" : 5%]
Any complication (%)	26%["2" : 100%]
Serious complications (%)	29%["2" : 96%, "1" : 4%]
Order Preserving	
P-Possum morbidity (%)	27%["14" : 24%, "16" : 20%, "15" : 12%, "10" : 12%, "11" : 12%]
Venous thromboembolism (%)	32%["5" : 21%]
Any complication (%)	23%["14" : 24%, "15" : 24%]
Cardiac complications (%)	35%["8" : 31%, 9" : 25%, "7" : 16%]
Readmission (%)	22%["14" : 20%, "13" : 15%, "6" : 15%]
ACS forecast of hospitalization days (%)	41%["14" : 24%, "13" : 16%]
P-Possum surgical severity score (%)	20%["12" : 28%, "11" : 17%]
Discharge to nursing or rehad facility (%)	37%["14" : 29%]
Reoperation (%)	30%["10" : 22%]

70% quality	Death within 1 year
Constant	
3 labels	
Pneumonia (%)	56%["2" : 90%, "1" : 10%]
Cardiac complications (%)	40%["1" : 83%, "2" : 17%]
ACS forecast of hospitalization days (%)	49%["2" : 83%, "0" : 11%, "1" : 6%]
Death (%)	41%["1" : 67%, "2" : 33%]
P-Possum physiological score (%)	40%["2" : 93%, "0" : 7%]
Reoperation (%)	27%["2" : 80%, "1" : 20%]
Readmission (%)	32%["2" : 100%]
Any complication (%)	29%[" 2" : 95%, "1" : 5%]
Serious complications (%)	26%["2" : 95%, "1" : 5%]
5 labels	
Readmission (%)	44%["4" : 79%, "1" : 16%, "3" : 5%]
Reoperation (%)	30%["4" : 33%, "2" : 31%, "3" : 23%, "1" : 13%]
P-Possum morbidity (%)	40%["4" : 98%, "3" : 2%]
ACS forecast of hospitalization days (%)	34%["4" : 69%, "3" : 16%, "1" : 15%]
Pneumonia (%)	24%["4" : 48%, "3" : 36%, "2" : 10%, "1" : 6%]
Any complication (%)	45%["4" : 80%, "1" : 10%, "3" : 9%, "0" : 1%]
Serious complications (%)	24%["4" : 55%, "1" : 23%, "3" : 19%, "2" : 3%]
Death (%)	23%["2" : 84%, "1" : 13%, "3" : 3%]
Cardiac complications (%)	25%["1" : 40%, "2" : 24%, "3" : 18%, "4" : 18%]
7 labels	
Any complication (%)	51%["5" : 56%, "6" : 37%]
Readmission (%)	29%["6" : 61%, "5" : 17%]
Reoperation (%)	23%[3" : 36%, "6" : 27%, "5" : 15%]
ACS forecast of hospitalization days (%)	32%["6" : 54%, "5" : 24%, "4" : 15%, "2" : 7%]
P-Possum morbidity (%)	36%["6" : 84%, "5" : 10%, "2" : 4%, " 1" : 2%]
Discharge to nursing or rehad facility (%)	23%["6" : 47%, "3" : 38%, "4" : 9%, "2" : 6%]
Serious complications (%)	25%["5" : 49%, "6" : 34%]
Death (%)	26%["3" : 57%, "2" : 38%]
Cardiac complications (%)	24%["2" : 53%, "3" : 18%, "4" : 18%, "6" : 11%]
Order Preserving	
P-Possum morbidity (%)	27%["14" : 24%, "16" : 20%, "10" : 12%, "11" : 12%, "15" : 12%]
Venous thromboembolism (%)	32%["5" : 21%]
Any complication (%)	23%["14" : 24%, "15" : 24%]
Cardiac complications (%)	35%["8" : 31%, "9" : 25%, "7" : 16%]
Readmission (%)	22%["14" : 20%, "13" : 15%]
Serious complications (%)	24%["11" : 18%, "12" : 18%]
ACS forecast of hospitalization days (%)	41%["14" : 24%, "13" : 16%]
Discharge to nursing or rehad facility (%)	37%["14" : 29%]
Reoperation (%)	29%["10" . 22%]

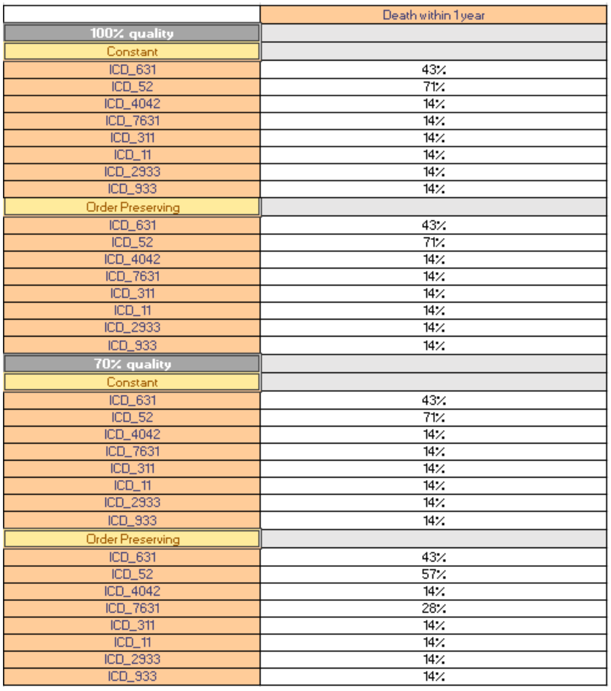
Figure A.4: Occurrence of variables and their values in the biclusters of death within 1 year, scores dataset.



	Death within 1 year
100% quality	
Constant	
3 labels	
ACS ASA	25%["2" : 94%, "3" : 6%]
ACS Weight	42%["0" : 85%, "1" : 15%]
ACS cancer disseminated	28%["1" : 94%, "0" : 6%]
UCI motive of admission	22%["0" : 100%]
PP peritoneal contamination	26%["0" : 100%]
Specialty code	22%["2" : 50%, "1" : 29%, "3" : 21%]
Age	29%["2" : 58%, "1" : 42%]
Order Preserving	
ACS Weight	56%["1" : 18%]
PP state of malignancy	30%["2" : 64%, "1" : 26%, "3" : 10%]
Age	22%["11" : 17%, "12" : 11%]
Specialty code	32%["2" : 48%, "1" : 37%, "3" : 10%, "0" : 5%]
PP peritoneal contamination	26%["0" : 77%, "2" : 13%, "1" : 4%, "3" : 6%
ACS cancer disseminated	23%["0" : 75%, "1" : 25%]
UCI motive of admission	26%["0" : 46%, "7" : 22%]

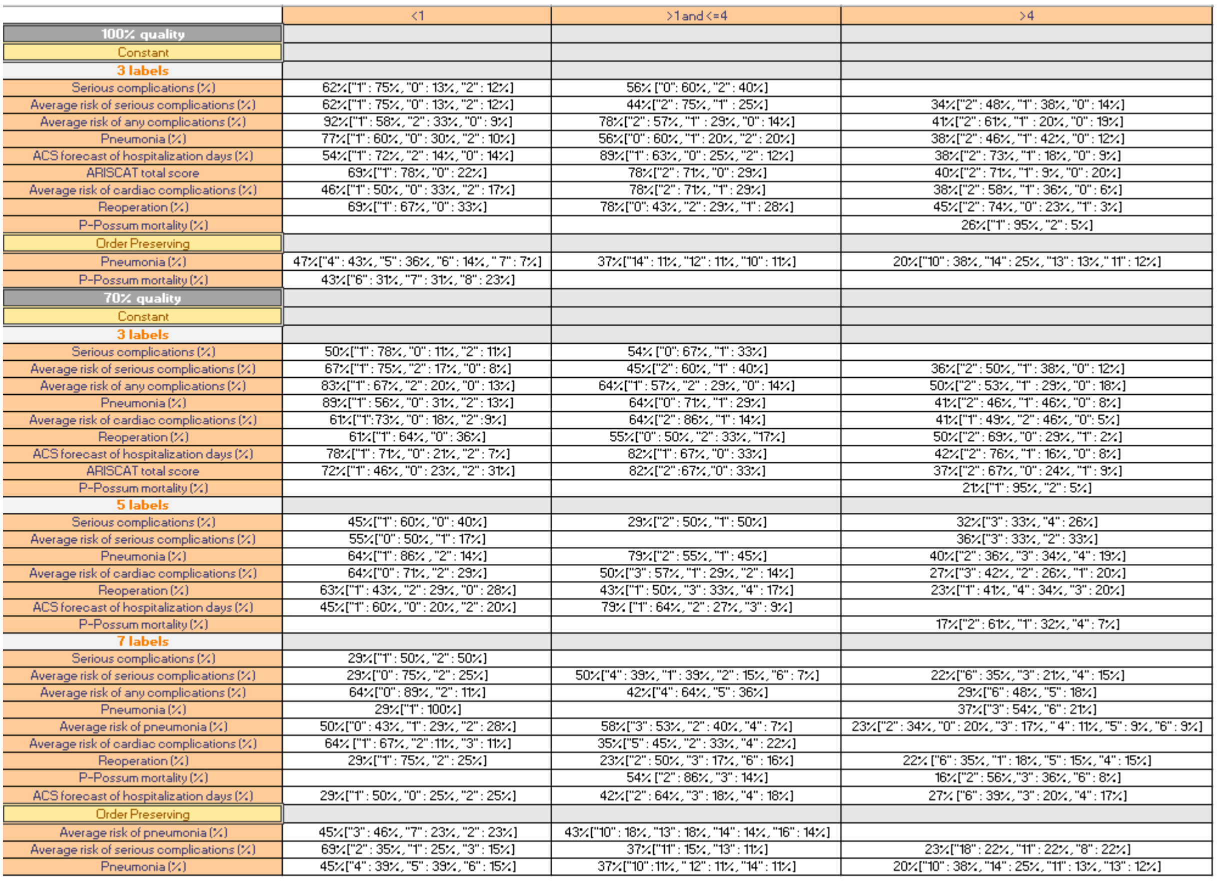
70% quality	Death within 1 year
Constant	
3 labels	
ACS Weight	48%["0" : 75%, "1" : 23%, "2" : 2%]
Specialty code	25%["2" : 52%, "1" : 30%, "3" : 17%]
PP respiratory	27%["0" : 52%, "1" : 32%, "3" : 16%]
PP state of malignancy	26%["2" : 50%, "3" : 33%, "1" : 17%]
UCI motive of admission	37%["1" : 47%, "0" : 44%, "2" : 9%]
PP peritoneal contamination	35%["0" : 97%, "1" : 3%]
ACS cancer disseminated	25%["1" : 83%, "0" : 17%]
Age	27%["2" : 52%, "1" : 48%]
4 labels	
ACS ASA	24%["2" : 87%, "3" : 13%]
ACS Weight	48%["0" : 57%, "1" : 27%, "2" : 16%]
UCI motive of admission	34%["1" : 52%, "0" : 43%, "6" : 5%]
PP peritoneal contamination	45%["0" : 100%]
ACS cancer disseminated	24%["1" : 87%, "0" : 13%]
Specialty code	24%["2" : 54%, "1" : 33%, "3" : 13%]
Age	31%["3" : 48%, "2 : 26%, "1" : 26%]
5 labels	
ACS ASA	28%["2" : 79%, "3" : 11%, "1" : 10%]
UCI motive of admission	34%["1" : 52%, "0" : 48%]
PP peritoneal contamination	32%["0" : 100%]
ACS Weight	31%["1" : 57%, "3" : 19%, "0" : 14%, "2" :10%]
ACS cancer disseminated	31%["1" : 90%, "0" : 10%]
Specialty code	22%["2" : 73%, "1" : 27%]
PP state of malignancy	34%["2" : 65%, "3" : 22%, "1" : 13%]
PP respiratory	26%["0" : 67%, "1" : 33%]
Age	29%["3" : 65%, "2" : 20%, "1" : 10%, "4" : 5%]
Order Preserving	
PP peritoneal contamination	26%["0" : 78%, "2" : 13%, "1" : 5%, "3" : 4%]
ACS functional state	38%["0" : 54%, "1" : 39%, "2" :7%]
UCI motive of admission	27%["0" : 43%, "7" : 21%]
Days spent at IPOP	51%["10" : 22%, "8" : 19%, "9" : 16%]
PP state of malignancy	27%["2" : 59%, "1" : 29%, "3" : 12%]
PP hemoglobin	24%["2" : 38%, "1" : 28%,"0" : 22%, "3" : 12%]
ACS cancer disseminated	26%["0" : 77%, "1" : 23%]

Figure A.5: Occurrence of variables and their values in the biclusters of death within 1 year, non-scores dataset.



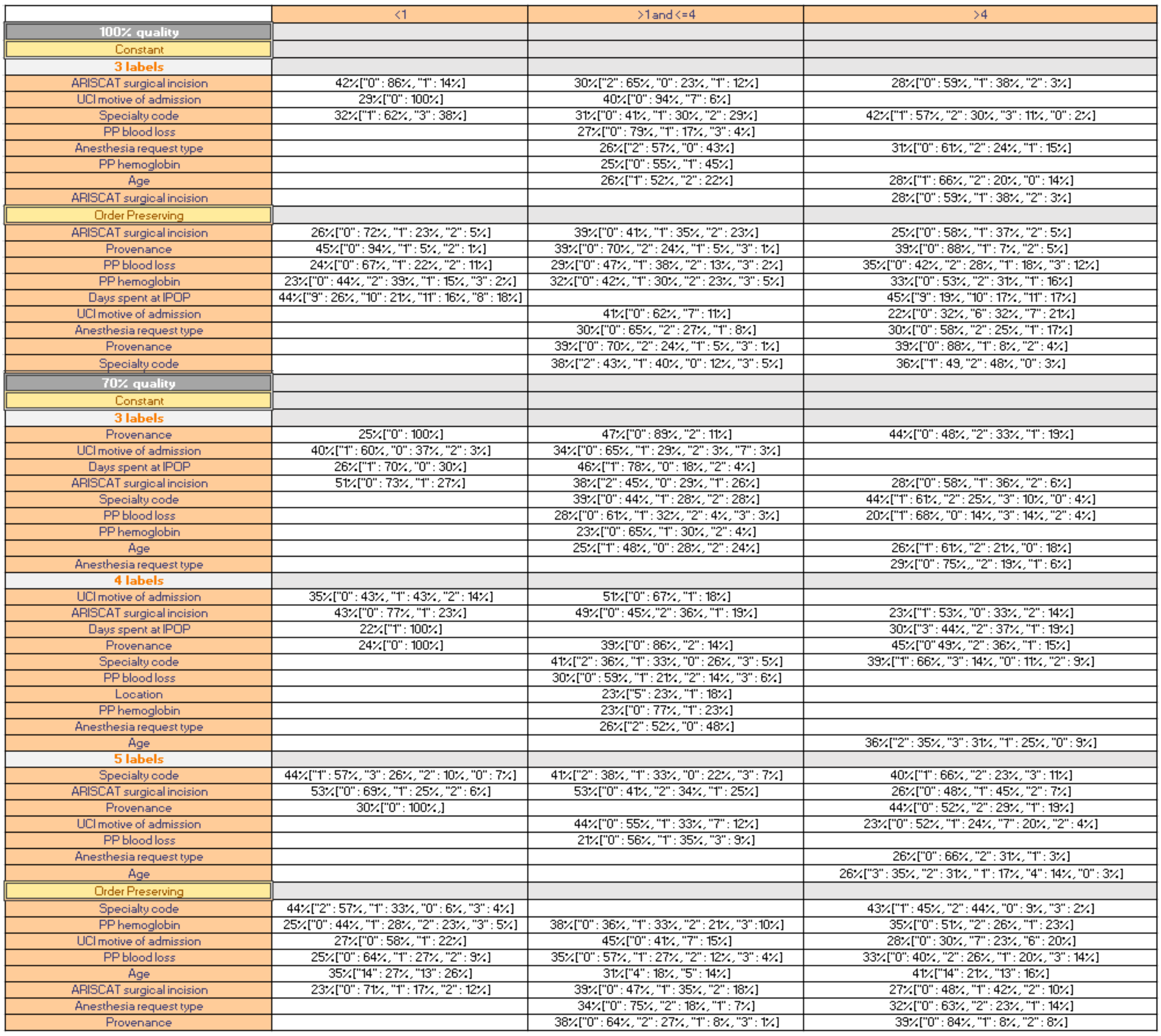
	Death within 1 year
100% quality	
Constant	
ICD_631	43%
ICD_52	71%
ICD_4042	14%
ICD_7631	14%
ICD_311	14%
ICD_11	14%
ICD_2933	14%
ICD_933	14%
Order Preserving	
ICD_631	43%
ICD_52	71%
ICD_4042	14%
ICD_7631	14%
ICD_311	14%
ICD_11	14%
ICD_2933	14%
ICD_933	14%
70% quality	
Constant	
ICD_631	43%
ICD_52	71%
ICD_4042	14%
ICD_7631	14%
ICD_311	14%
ICD_11	14%
ICD_2933	14%
ICD_933	14%
Order Preserving	
ICD_631	43%
ICD_52	57%
ICD_4042	14%
ICD_7631	28%
ICD_311	14%
ICD_11	14%
ICD_2933	14%
ICD_933	14%

Figure A.6: Occurrence of variables and their values in the biclusters of death within 1 year, ICD-10 dataset.



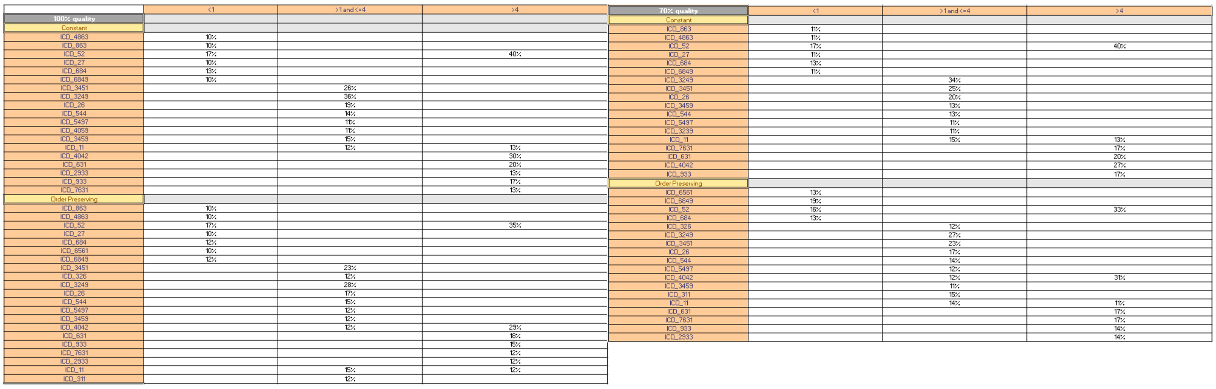
	<1	>1 and <=4	>4
100% quality			
Constant			
3 labels			
Serious complications (%)	62%["1" : 75%, "0" : 13%, "2" : 12%]	56% ["0": 60%, "2" : 40%]	
Average risk of serious complications (%)	62%["1" : 75%, "0" : 13%, "2" : 12%]	44%["2" : 75%, "1" : 25%]	34%["2" : 48%, "1" : 38%, "0" : 14%]
Average risk of any complications (%)	92%["1" : 58%, "2" : 33%, "0" : 9%]	78%["2" : 57%, "1" : 29%, "0" : 14%]	41%["2" : 61%, "1" : 20%, "0" : 19%]
Pneumonia (%)	77%["1" : 60%, "0" : 30%, "2" : 10%]	56%["0" : 60%, "1" : 20%, "2" : 20%]	38%["2" : 46%, "1" : 42%, "0" : 12%]
ACS forecast of hospitalization days (%)	54%["1" : 72%, "2" : 14%, "0" : 14%]	89%["1" : 63%, "0" : 25%, "2" : 12%]	38%["2" : 73%, "1" : 18%, "0" : 9%]
ARISCAT total score	69%["1" : 78%, "0" : 22%]	78%["2" : 71%, "0" : 29%]	40%["2" : 71%, "1" : 9%, "0" : 20%]
Average risk of cardiac complications (%)	46%["1" : 50%, "0" : 33%, "2" : 17%]	78%["2" : 71%, "1" : 29%]	38%["2" : 58%, "1" : 36%, "0" : 6%]
Reoperation (%)	69%["1" : 67%, "0" : 33%]	78%["0": 43%, "2" : 29%, "1" : 28%]	45%["2" : 74%, "0" : 23%, "1" : 3%]
P-Possum mortality (%)			26%["1" : 95%, "2" : 5%]
Order Preserving			
Pneumonia (%)	47%["4" : 43%, "5" : 36%, "6" : 14%, " 7" : 7%]	37%["14" : 11%, "12" : 11%, "10" : 11%]	20%["10" : 38%, "14" : 25%, "13" : 13%," 11" : 12%]
P-Possum mortality (%)	43%["6" : 31%, "7" : 31%, "8" : 23%]		
70% quality			
Constant			
3 labels			
Serious complications (%)	50%["1" : 78%, "0" : 11%, "2" : 11%]	54% ["0": 67%, "1" : 33%]	
Average risk of serious complications (%)	67%["1" : 75%, "2" : 17%, "0" : 8%]	45%["2" : 60%, "1" : 40%]	36%["2" : 50%, "1" : 38%, "0" : 12%]
Average risk of any complications (%)	83%["1" : 67%, "2" : 20%, "0" : 13%]	64%["1" : 57%, "2" : 29%, "0" : 14%]	50%["2" : 53%, "1" : 29%, "0" : 18%]
Pneumonia (%)	89%["1" : 56%, "0" : 31%, "2" : 13%]	64%["0" : 71%, "1" : 29%]	41%["2" : 46%, "1" : 46%, "0" : 8%]
Average risk of cardiac complications (%)	61%["1":73%, "0" : 18%, "2":9%]	64%["2" : 86%, "1" : 14%]	41%["1" : 49%, "2" : 46%, "0": 5%]
Reoperation (%)	61%["1" : 64%, "0" : 36%]	55%["0" : 50%, "2" : 33%, "17%]	50%["2" : 69%, "0" : 29%, "1" : 2%]
ACS forecast of hospitalization days (%)	78%["1" : 71%, "0" : 21%, "2" : 7%]	82%["1" : 67%, "0" : 33%]	42%["2" : 76%, "1" : 16%, "0" : 8%]
ARISCAT total score	72%["1" : 46%, "0" : 23%, "2" : 31%]	82%["2" :67%,"0" : 33%]	37%["2" : 67%, "0" : 24%, "1" : 9%]
P-Possum mortality (%)			21%["1" : 95%, "2" : 5%]
5 labels			
Serious complications (%)	45%["1" : 60%, "0" : 40%]	29%["2" : 50%, "1" : 50%]	32%["3" : 33%, "4" : 26%]
Average risk of serious complications (%)	55%["0" : 50%, "1" : 17%]		36%["3" : 33%, "2" : 33%]
Pneumonia (%)	64%["1" : 86% , "2" : 14%]	79%["2" : 55%, "1" : 45%]	40%["2" : 36%, "3" : 34%, "4" : 19%]
Average risk of cardiac complications (%)	64%["0" : 71%, "2" : 29%]	50%["3" : 57%, "1" : 29%, "2" : 14%]	27%["3" : 42%, "2" : 26%, "1" : 20%]
Reoperation (%)	63%["1" : 43%, "2" : 29%, "0" : 28%]	43%["1" : 50%, "3" : 33%, "4" : 17%]	23%["1" : 41%, "4" : 34%, "3" : 20%]
ACS forecast of hospitalization days (%)	45%["1" : 60%, "0" : 20%, "2" : 20%]	79% ["1" : 64%, "2" : 27%, "3" : 9%]	
P-Possum mortality (%)			17%["2" : 61%, "1" : 32%, "4" : 7%]
7 labels			
Serious complications (%)	29%["1" : 50%, "2" : 50%]		
Average risk of serious complications (%)	29%["0" : 75%, "2" : 25%]	50%["4" : 39%, "1" : 39%, "2" : 15%, "6" : 7%]	22%["6" : 35%, "3" : 21%, "4" : 15%]
Average risk of any complications (%)	64%["0" : 89%, "2" : 11%]	42%["4" : 64%, "5" : 36%]	29%["6" : 48%, "5" : 18%]
Pneumonia (%)	29%["1" : 100%]		37%["3" : 54%, "6" : 21%]
Average risk of pneumonia (%)	50%["0" : 43%, "1" : 29%, "2" : 28%]	58%["3" : 53%, "2" : 40%, "4" : 7%]	23%["2" : 34%, "0" : 20%, "3" : 17%, " 4" : 11%, "5" : 9%, "6" : 9%]
Average risk of cardiac complications (%)	64% ["1" : 67%, "2" :11%, "3" : 11%]	35%["5" : 45%, "2" : 33%, "4" : 22%]	
Reoperation (%)	29%["1" : 75%, "2" : 25%]	23%["2" : 50%, "3" : 17%, "6" : 16%]	22% ["6" : 35%, "1" : 18%, "5" : 15%, "4" : 15%]
P-Possum mortality (%)		54% ["2" : 86%, "3" : 14%]	16%["2" : 56%,"3" : 36%, "6" : 8%]
ACS forecast of hospitalization days (%)	29%["1" : 50%, "0" : 25%, "2" : 25%]	42%["2" : 64%, "3" : 18%, "4" : 18%]	27% ["6" : 39%, "3" : 20%, "4" : 17%]
Order Preserving			
Average risk of pneumonia (%)	45%["3" : 46%, "7" : 23%, "2" : 23%]	43%["10" : 18%, "13" : 18%, "14" : 14%, "16" : 14%]	
Average risk of serious complications (%)	69%["2" : 35%, "1" : 25%, "3" : 15%]	37%["11" : 15%, "13" : 11%]	23%["18" : 22%, "11" : 22%, "8" : 22%]
Pneumonia (%)	45%["4" : 39%, "5" : 39%, "6" : 15%]	37%["10" :11%, " 12" : 11%, "14" : 11%]	20%["10" : 38%, "14" : 25%, "11" : 13%, "13" : 12%]

Figure A.7: Occurrence of variables and their values in the biclusters of days spent at HDU, scores dataset.



	<1	>1 and <=4	>4
100% quality			
Constant			
3 labels			
ARISCAT surgical incision	42%["0" : 86%, "1" : 14%]	30%["2" : 65%, "0" : 23%, "1" : 12%]	28%["0" : 59%, "1" : 38%, "2" : 3%]
UCI motive of admission	29%["0" : 100%]	40%["0" : 94%, "7" : 6%]	
Specialty code	32%["1" : 62%, "3" : 38%]	31%["0" : 41%, "1" : 30%, "2" : 29%]	42%["1" : 57%, "2" : 30%, "3" : 11%, "0" : 2%]
PP blood loss		27%["0" : 79%, "1" : 17%, "3" : 4%]	
Anesthesia request type		26%["2" : 57%, "0" : 43%]	31%["0" : 61%, "2" : 24%, "1" : 15%]
PP hemoglobin		25%["0" : 55%, "1" : 45%]	
Age		26%["1" : 52%, "2" : 22%]	28%["1" : 66%, "2" : 20%, "0" : 14%]
ARISCAT surgical incision			28%["0" : 59%, "1" : 38%, "2" : 3%]
Order Preserving			
ARISCAT surgical incision	26%["0" : 72%, "1" : 23%, "2" : 5%]	39%["0" : 41%, "1" : 35%, "2" : 23%]	25%["0" : 58%, "1" : 37%, "2" : 5%]
Provenance	45%["0" : 94%, "1" : 5%, "2" : 1%]	39%["0" : 70%, "2" : 24%, "1" : 5%, "3" : 1%]	39%["0" : 88%, "1" : 7%, "2" : 5%]
PP blood loss	24%["0" : 67%, "1" : 22%, "2" : 11%]	29%["0" : 47%, "1" : 38%, "2" : 13%, "3" : 2%]	35%["0" : 42%, "2" : 28%, "1" : 18%, "3" : 12%]
PP hemoglobin	23%["0" : 44%, "2" : 39%, "1" : 15%, "3" : 2%]	32%["0" : 42%, "1" : 30%, "2" : 23%, "3" : 5%]	33%["0" : 53%, "2" : 31%, "1" : 16%]
Days spent at IPOP	44%["9" : 26%, "10" : 21%, "11" : 16%, "8" : 18%]		45%["9" : 19%, "10" : 17%, "11" : 17%]
UCI motive of admission		41%["0" : 62%, "7" : 11%]	22%["0" : 32%, "6" : 32%, "7" : 21%]
Anesthesia request type		30%["0" : 65%, "2" : 27%, "1" : 8%]	30%["0" : 58%, "2" : 25%, "1" : 17%]
Provenance		39%["0" : 70%, "2" : 24%, "1" : 5%, "3" : 1%]	39%["0" : 88%, "1" : 8%, "2" : 4%]
Specialty code		38%["2" : 43%, "1" : 40%, "0" : 12%, "3" : 5%]	36%["1" : 49, "2" : 48%, "0" : 3%]
70% quality			
Constant			
3 labels			
Provenance	25%["0" : 100%]	47%["0" : 89%, "2" : 11%]	44%["0" : 48%, "2" : 33%, "1" : 19%]
UCI motive of admission	40%["1" : 60%, "0" : 37%, "2" : 3%]	34%["0" : 65%, "1" : 29%, "2" : 3%, "7" : 3%]	
Days spent at IPOP	26%["1" : 70%, "0" : 30%]	46%["1" : 78%, "0" : 18%, "2" : 4%]	
ARISCAT surgical incision	51%["0" : 73%, "1" : 27%]	38%["2" : 45%, "0" : 29%, "1" : 26%]	28%["0" : 58%, "1" : 36%, "2" : 6%]
Specialty code		39%["0" : 44%, "1" : 28%, "2" : 28%]	44%["1" : 61%, "2" : 25%, "3" : 10%, "0" : 4%]
PP blood loss		28%["0" : 61%, "1" : 32%, "2" : 4%, "3" : 3%]	20%["1" : 68%, "0" : 14%, "3" : 14%, "2" : 4%]
PP hemoglobin		23%["0" : 65%, "1" : 30%, "2" : 4%]	
Age		25%["1" : 48%, "0" : 28%, "2" : 24%]	26%["1" : 61%, "2" : 21%, "0" : 18%]
Anesthesia request type			29%["0" : 75%,, "2" : 19%, "1" : 6%]
4 labels			
UCI motive of admission	35%["0" : 43%, "1" : 43%, "2" : 14%]	51%["0" : 67%, "1" : 18%]	
ARISCAT surgical incision	43%["0" : 77%, "1" : 23%]	49%["0" : 45%,"2" : 36%, "1" : 19%]	23%["1" : 53%, "0" : 33%, "2" : 14%]
Days spent at IPOP	22%["1" : 100%]		30%["3" : 44%, "2" : 37%, "1" : 19%]
Provenance	24%["0" : 100%]	39%["0" : 86%, "2" : 14%]	45%["0" 49%, "2" : 36%, "1" : 15%]
Specialty code		41%["2" : 36%, "1" : 33%, "0" : 26%, "3" : 5%]	39%["1" : 66%, "3" : 14%, "0" : 11%, "2" : 9%]
PP blood loss		30%["0" : 59%, "1" : 21%, "2" : 14%, "3" : 6%]	
Location		23%["5" : 23%, "1" : 18%]	
PP hemoglobin		23%["0" : 77%, "1" : 23%]	
Anesthesia request type		26%["2" : 52%, "0" : 48%]	
Age			36%["2" : 35%, "3" : 31%, "1" : 25%, "0" : 9%]
5 labels			
Specialty code	44%["1" : 57%, "3" : 26%, "2" : 10%, "0" : 7%]	41%["2" : 38%, "1" : 33%, "0" : 22%, "3" : 7%]	40%["1" : 66%, "2" : 23%, "3" : 11%]
ARISCAT surgical incision	53%["0" : 69%, "1" : 25%, "2" : 6%]	53%["0" : 41%, "2" : 34%, "1" : 25%]	26%["0" : 48%, "1" : 45%, "2" : 7%]
Provenance	30%["0" : 100%,]		44%["0" : 52%, "2" : 29%, "1" : 19%]
UCI motive of admission		44%["0" : 55%, "1" : 33%, "7" : 12%]	23%["0" : 52%, "1" : 24%, "7" : 20%, "2" : 4%]
PP blood loss		21%["0" : 56%, "1" : 35%, "3" : 9%]	
Anesthesia request type			26%["0" : 66%, "2" : 31%, "1" : 3%]
Age			26%["3" : 35%, "2" : 31%, " 1" : 17%, "4" : 14%, "0" : 3%]
Order Preserving			
Specialty code	44%["2" : 57%, "1" : 33%, "0" : 6%, "3" : 4%]		43%["1" : 45%, "2" : 44%, "0" : 9%, "3" : 2%]
PP hemoglobin	25%["0" : 44%, "1" : 28%, "2" : 23%, "3" : 5%]	38%["0" : 36%, "1" : 33%, "2" : 21%, "3" :10%]	35%["0" : 51%, "2" : 26%, "1" : 23%]
UCI motive of admission	27%["0" : 58%, "1" : 22%]	45%["0" : 41%, "7" : 15%]	28%["0" : 30%, "7" : 23%, "6" : 20%]
PP blood loss	25%["0" : 64%, "1" : 27%, "2" : 9%]	35%["0" : 57%, "1" : 27%, "2" : 12%, "3" : 4%]	33%["0" : 40%, "2" : 26%, "1" : 20%, "3" : 14%]
Age	35%["14" : 27%, "13" : 26%]	31%["4" : 18%, "5" : 14%]	41%["14" : 21%, "13" : 16%]
ARISCAT surgical incision	23%["0" : 71%, "1" : 17%, "2" : 12%]	39%["0" : 47%, "1" : 35%, "2" : 18%]	27%["0" : 48%, "1" : 42%, "2" : 10%]
Anesthesia request type		34%["0" : 75%, "2" : 18%, "1" : 7%]	32%["0" : 63%, "2" : 23%, "1" : 14%]
Provenance		38%["0" : 64%, "2" : 27%, "1" : 8%, "3" : 1%]	39%["0" : 84%, "1" : 8%, "2" : 8%]

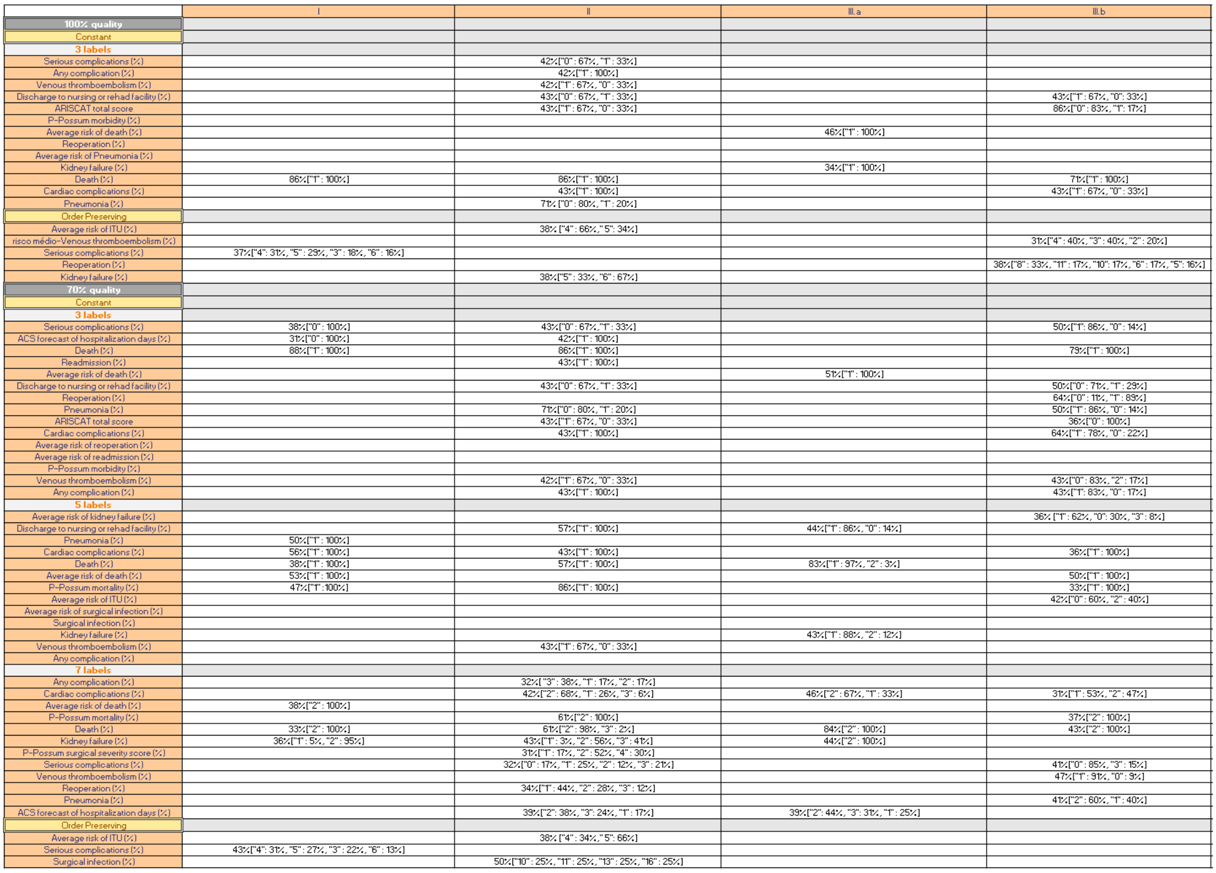
Figure A.8: Occurrence of variables and their values in the biclusters of days spent at HDU, non-scores dataset.



	<1	>1 and <=4	>4
100% quality			
Constant			
ICD_4863	10%		
ICD_863	10%		
ICD_52	17%		40%
ICD_27	10%		
ICD_684	13%		
ICD_6849	10%		
ICD_3451		26%	
ICD_3249		36%	
ICD_26		19%	
ICD_544		14%	
ICD_5497		11%	
ICD_4059		11%	
ICD_3459		15%	
ICD_11		12%	13%
ICD_4042			30%
ICD_631			20%
ICD_2933			13%
ICD_933			17%
ICD_7631			13%
Order Preserving			
ICD_863	10%		
ICD_4863	10%		
ICD_52	17%		35%
ICD_27	10%		
ICD_684	12%		
ICD_6561	10%		
ICD_6849	12%		
ICD_3451		23%	
ICD_326		12%	
ICD_3249		28%	
ICD_26		17%	
ICD_544		15%	
ICD_5497		12%	
ICD_3459		12%	
ICD_4042		12%	29%
ICD_631			18%
ICD_933			15%
ICD_7631			12%
ICD_2933			12%
ICD_11		15%	12%
ICD_311		12%	

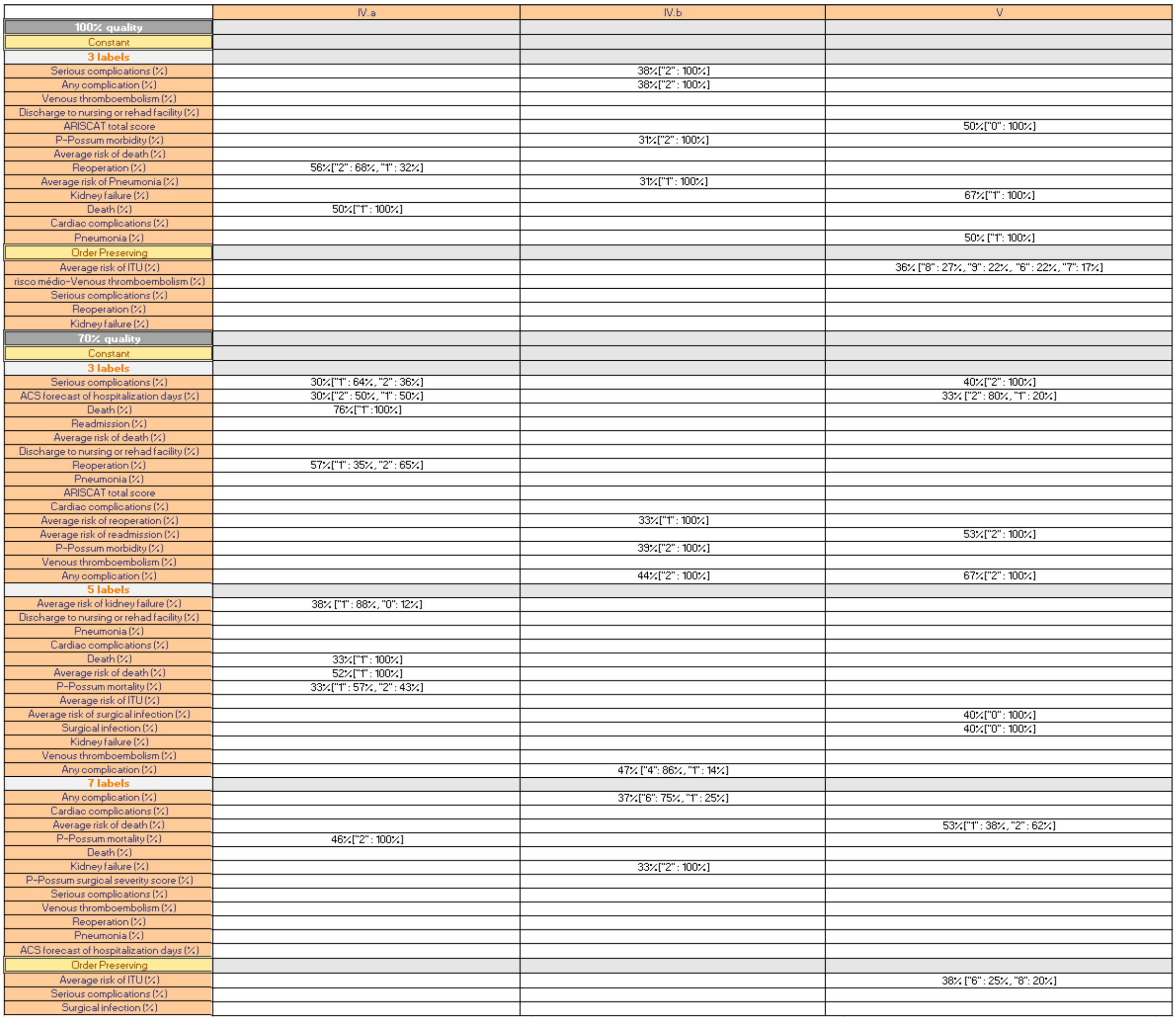
70% quality	<1	>1 and <=4	>4
Constant			
ICD_863	11%		
ICD_4863	11%		
ICD_52	17%		40%
ICD_27	11%		
ICD_684	13%		
ICD_6849	11%		
ICD_3249		34%	
ICD_3451		25%	
ICD_26		20%	
ICD_3459		13%	
ICD_544		13%	
ICD_5497		11%	
ICD_3239		11%	
ICD_11		15%	13%
ICD_7631			17%
ICD_631			20%
ICD_4042			27%
ICD_933			17%
Order Preserving			
ICD_6561	13%		
ICD_6849	19%		
ICD_52	16%		33%
ICD_684	13%		
ICD_326		12%	
ICD_3249		27%	
ICD_3451		23%	
ICD_26		17%	
ICD_544		14%	
ICD_5497		12%	
ICD_4042		12%	31%
ICD_3459		11%	
ICD_311		15%	
ICD_11		14%	11%
ICD_631			17%
ICD_7631			17%
ICD_933			14%
ICD_2933			14%

Figure A.9: Occurrence of variables and their values in the biclusters of days spent at HDU, ICD-10 dataset.



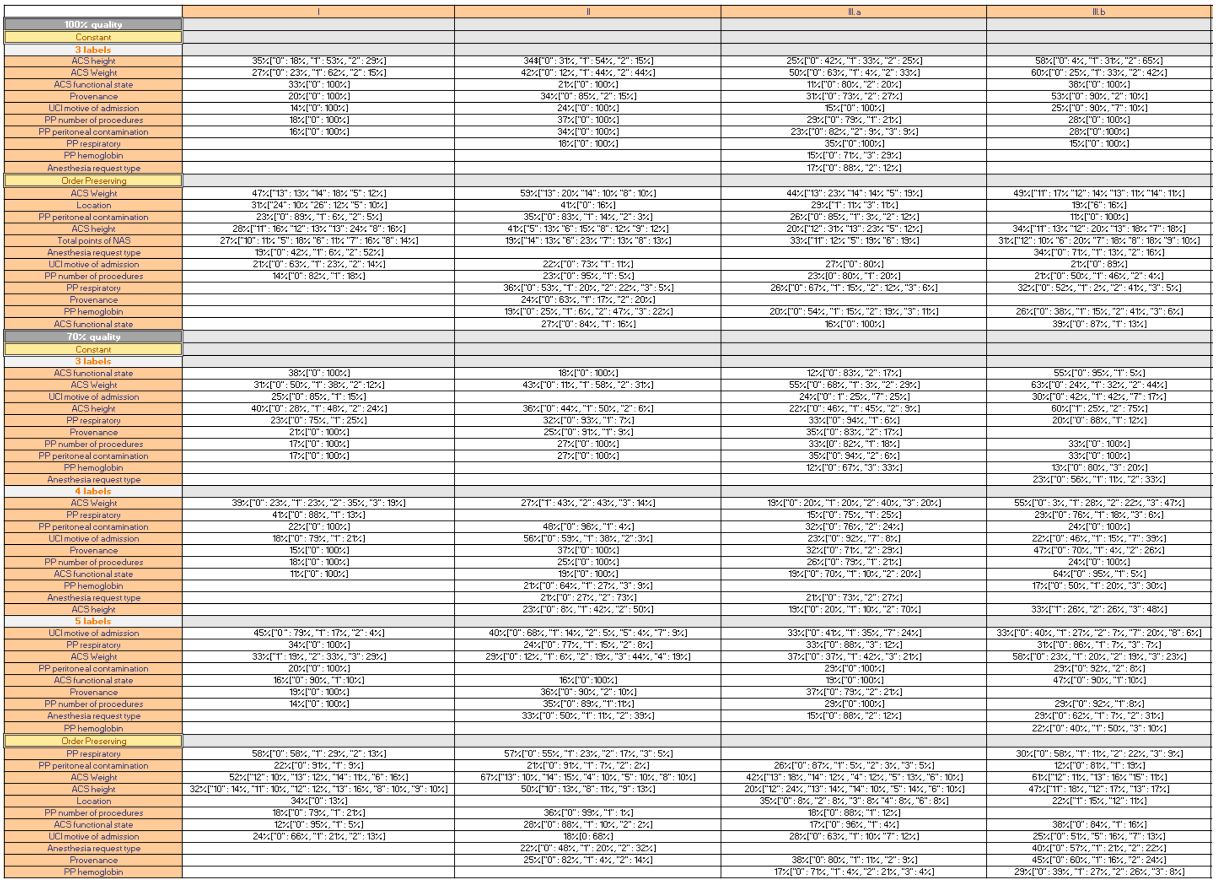
	I	II	III.a	III.b
100% quality				
Constant				
3 labels				
Serious complications (%)		42%["0" : 67%, "1" : 33%]		
Any complication (%)		42%["1" : 100%]		
Venous thromboembolism (%)		42%["1" : 67%, "0" : 33%]		
Discharge to nursing or rehad facility (%)		43%["0" : 67%, "1" : 33%]		43%["1" : 67%, "0": 33%]
ARISCAT total score		43%["1" : 67%, "0" : 33%]		86%["0" : 83%, "1": 17%]
P-Possum morbidity (%)				
Average risk of death (%)			46%["1" : 100%]	
Reoperation (%)				
Average risk of Pneumonia (%)				
Kidney failure (%)			34%["1" : 100%]	
Death (%)	86%["1" : 100%]	86%["1" : 100%]		71%["1" : 100%]
Cardiac complications (%)		43%["1" : 100%]		43%["1" : 67%, "0" : 33%]
Pneumonia (%)		71% ["0" : 80%, "1" : 20%]		
Order Preserving				
Average risk of ITU (%)		38% ["4" : 66%, "5": 34%]		
risco médio-Venous thromboembolism (%)				31%["4" : 40%, "3" : 40%, "2" : 20%]
Serious complications (%)	37%["4": 31%, "5" : 29%, "3" : 18%, "6" : 16%]			
Reoperation (%)				38%["8" : 33%, "11" : 17%, "10": 17%, "6" : 17%, "5": 16%]
Kidney failure (%)		38%["5" : 33%, "6" : 67%]		
70% quality				
Constant				
3 labels				
Serious complications (%)	38%["0" : 100%]	43%["0" : 67%, "1" : 33%]		50%["1": 86%, "0" : 14%]
ACS forecast of hospitalization days (%)	31%["0" : 100%]	42%["1" : 100%]		
Death (%)	88%["1" : 100%]	86%["1" : 100%]		79%["1" : 100%]
Readmission (%)		43%["1" : 100%]		
Average risk of death (%)			51%["1" : 100%]	
Discharge to nursing or rehad facility (%)		43%["0" : 67%, "1" : 33%]		50%["0" : 71%, "1" : 29%]
Reoperation (%)				64%["0" : 11%, "1" : 89%]
Pneumonia (%)		71%["0" : 80%, "1" : 20%]		50%["1" : 86%, "0" : 14%]
ARISCAT total score		43%["1" : 67%, "0" : 33%]		36%["0" : 100%]
Cardiac complications (%)		43%["1" : 100%]		64%["1" : 78%, "0" : 22%]
Average risk of reoperation (%)				
Average risk of readmission (%)				
P-Possum morbidity (%)				
Venous thromboembolism (%)		42%["1" : 67%, "0" : 33%]		43%["0" : 83%, "2" : 17%]
Any complication (%)		43%["1" : 100%]		43%["1": 83%, "0" : 17%]
5 labels				
Average risk of kidney failure (%)				36% ["1" : 62%, "0": 30%, "3" : 8%]
Discharge to nursing or rehad facility (%)		57%["1" : 100%]	44%["1" : 86%, "0" : 14%]	
Pneumonia (%)	50%["1" : 100%]			
Cardiac complications (%)	56%["1" : 100%]	43%["1" : 100%]		36%["1" : 100%]
Death (%)	38%["1" : 100%]	57%["1" : 100%]	83%["1" : 97%, "2" : 3%]	
Average risk of death (%)	53%["1" : 100%]			50%["1" : 100%]
P-Possum mortality (%)	47%["1" : 100%]	86%["1" : 100%]		33%["1" : 100%]
Average risk of ITU (%)				42%["0" : 60%, "2" : 40%]
Average risk of surgical infection (%)				
Surgical infection (%)				
Kidney failure (%)			43%["1" : 88%, "2" : 12%]	
Venous thromboembolism (%)		43%["1" : 67%, "0" : 33%]		
Any complication (%)				
7 labels				
Any complication (%)		32%["3" : 38%, "1" : 17%, "2" : 17%]		
Cardiac complications (%)		42%["2" : 68%, "1" : 26%, "3" : 6%]	46%["2" : 67%, "1" : 33%]	31%["1" : 53%, "2" : 47%]
Average risk of death (%)	38%["2" : 100%]			
P-Possum mortality (%)		61%["2" : 100%]		37%["2" : 100%]
Death (%)	33%["2" : 100%]	61%["2" : 98%, "3" : 2%]	84%["2" : 100%]	43%["2" : 100%]
Kidney failure (%)	36%["1" : 5%, "2" : 95%]	43%["1" : 3%, "2" : 56%, "3" : 41%]	44%["2" : 100%]	
P-Possum surgical severity score (%)		31%["1" : 17%, "2" : 52%, "4" : 30%]		
Serious complications (%)		32%["0" : 17%, "1" : 25%, "2" : 12%, "3" : 21%]		41%["0" : 85%, "3" : 15%]
Venous thromboembolism (%)				47%["1" : 91%, "0" : 9%]
Reoperation (%)		34%["1" : 44%, "2" : 28%, "3" : 12%]		
Pneumonia (%)				41%["2" : 60%, "1" : 40%]
ACS forecast of hospitalization days (%)		39%["2": 38%, "3": 24%, "1" : 17%]	39%["2": 44%, "3": 31%, "1" : 25%]	
Order Preserving				
Average risk of ITU (%)		38% ["4" : 34%, "5": 66%]		
Serious complications (%)	43%["4": 31%, "5" : 27%, "3" : 22%, "6" : 13%]			
Surgical infection (%)		50%["10" : 25%, "11" : 25%, "13" : 25%, "16" : 25%]		

Figure A.10: Occurrence of variables and their values in the biclusters of Clavien-Dindo score, scores dataset.



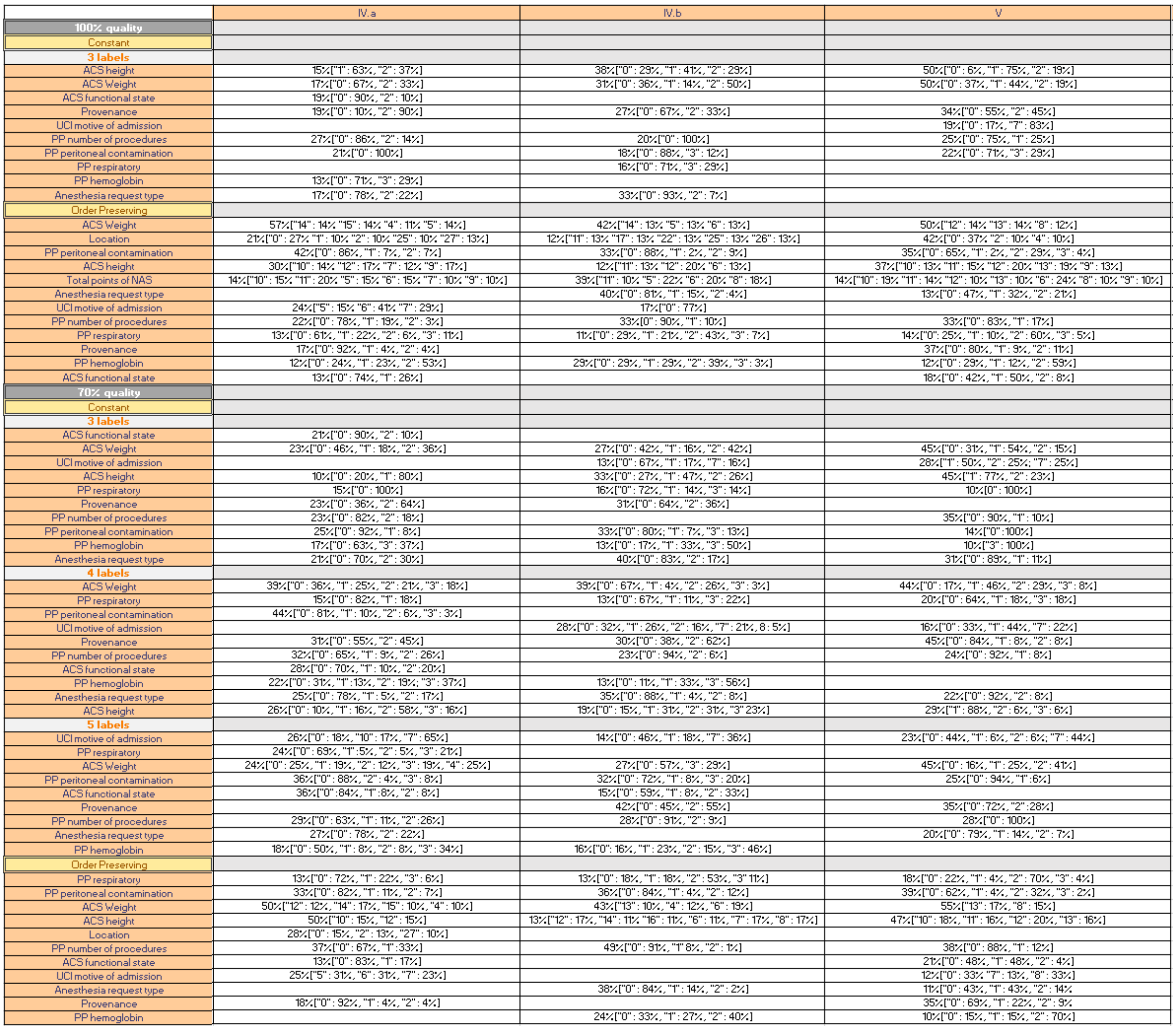
	IV.a	IV.b	V
100% quality			
Constant			
3 labels			
Serious complications (%)		38%["2" : 100%]	
Any complication (%)		38%["2" : 100%]	
Venous thromboembolism (%)			
Discharge to nursing or rehad facility (%)			
ARISCAT total score			50%["0" : 100%]
P-Possum morbidity (%)		31%["2" : 100%]	
Average risk of death (%)			
Reoperation (%)	56%["2" : 68%, "1" : 32%]		
Average risk of Pneumonia (%)		31%["1" : 100%]	
Kidney failure (%)			67%["1" : 100%]
Death (%)	50%["1" : 100%]		
Cardiac complications (%)			
Pneumonia (%)			50% ["1": 100%]
Order Preserving			
Average risk of ITU (%)			36% ["8" : 27%, "9" : 22%, "6" : 22%, "7": 17%]
risco médio-Venous thromboembolism (%)			
Serious complications (%)			
Reoperation (%)			
Kidney failure (%)			
70% quality			
Constant			
3 labels			
Serious complications (%)	30%["1" : 64%, "2" : 36%]		40%["2" : 100%]
ACS forecast of hospitalization days (%)	30%["2" : 50%, "1" : 50%]		33% ["2" : 80%, "1" : 20%]
Death (%)	76%["1" :100%]		
Readmission (%)			
Average risk of death (%)			
Discharge to nursing or rehad facility (%)			
Reoperation (%)	57%["1" : 35%, "2" : 65%]		
Pneumonia (%)			
ARISCAT total score			
Cardiac complications (%)			
Average risk of reoperation (%)		33%["1" : 100%]	
Average risk of readmission (%)			53%["2" : 100%]
P-Possum morbidity (%)		39%["2" : 100%]	
Venous thromboembolism (%)			
Any complication (%)		44%["2" : 100%]	67%["2" : 100%]
5 labels			
Average risk of kidney failure (%)	38% ["1" : 88%, "0": 12%]		
Discharge to nursing or rehad facility (%)			
Pneumonia (%)			
Cardiac complications (%)			
Death (%)	33%["1" : 100%]		
Average risk of death (%)	52%["1" : 100%]		
P-Possum mortality (%)	33%["1" : 57%, "2" : 43%]		
Average risk of ITU (%)			
Average risk of surgical infection (%)			40%["0" : 100%]
Surgical infection (%)			40%["0" : 100%]
Kidney failure (%)			
Venous thromboembolism (%)			
Any complication (%)		47% ["4": 86%, "1" : 14%]	
7 labels			
Any complication (%)		37%["6": 75%, "1" : 25%]	
Cardiac complications (%)			
Average risk of death (%)			53%["1" : 38%, "2" : 62%]
P-Possum mortality (%)	46%["2" : 100%]		
Death (%)			
Kidney failure (%)		33%["2" : 100%]	
P-Possum surgical severity score (%)			
Serious complications (%)			
Venous thromboembolism (%)			
Reoperation (%)			
Pneumonia (%)			
ACS forecast of hospitalization days (%)			
Order Preserving			
Average risk of ITU (%)			38% ["6" : 25%, "8": 20%]
Serious complications (%)			
Surgical infection (%)			

Figure A.11: Occurrence of variables and their values in the biclusters of Clavien-Dindo score, scores dataset.



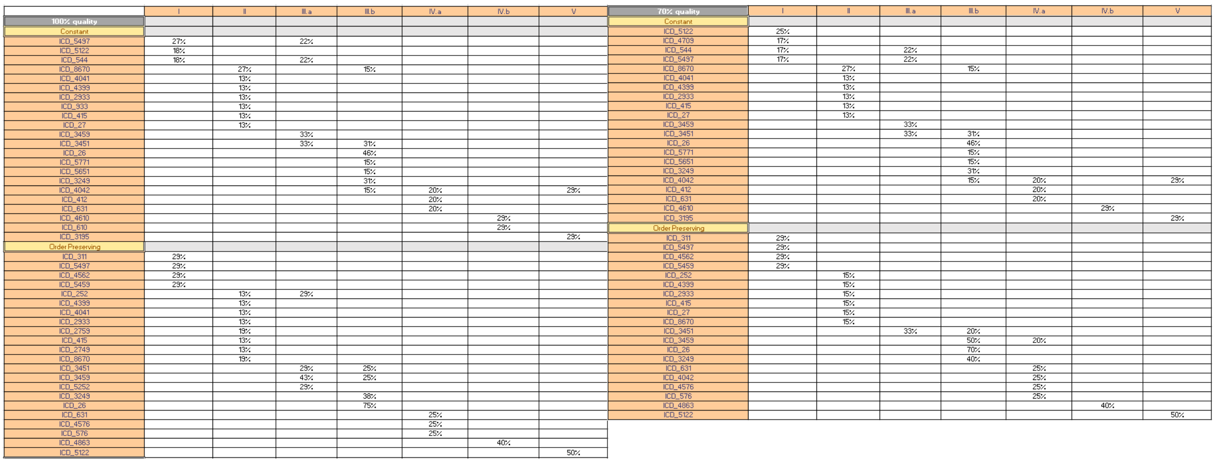
	I	II	III.a	III.b
100% quality				
Constant				
3 labels				
ACS height	35%["0" : 18%, "1" : 53%, "2" : 29%]	34$["0" : 31%, "1" : 54%, "2" : 15%]	25%["0" : 42%, "1" : 33%, "2" : 25%]	58%["0": 4%, "1" : 31%, "2" : 65%]
ACS Weight	27%["0" : 23%, "1" : 62%, "2" : 15%]	42%["0" : 12%, "1" : 44%, "2" : 44%]	50%["0" : 63%, "1" : 4%, "2" : 33%]	60%["0" : 25%, "1" : 33%, "2" : 42%]
ACS functional state	33%["0" : 100%]	21%["0" : 100%]	11%["0" : 80%, "2" : 20%]	38%["0" : 100%]
Provenance	20%["0" : 100%]	34%["0" : 85%, "2" : 15%]	31%["0" : 73%, "2" : 27%]	53%["0" : 90%, "2" : 10%]
UCI motive of admission	14%["0": 100%]	24%["0" : 100%]	15%["0" : 100%]	25%["0" : 90%, "7" : 10%]
PP number of procedures	18%["0" : 100%]	37%["0" : 100%]	29%["0" : 79%, "1" : 21%]	28%["0" : 100%]
PP peritoneal contamination	16%["0" : 100%]	34%["0" : 100%]	23%["0" : 82%, "2" : 9%, "3" : 9%]	28%["0" :100%]
PP respiratory		18%["0" : 100%]	35%["0" :100%]	15%["0" : 100%]
PP hemoglobin			15%["0" : 71%, "3" : 29%]	
Anesthesia request type			17%["0" : 88%, "2" : 12%]	
Order Preserving				
ACS Weight	47%["13" : 13% "14" : 18% "5" : 12%]	59%["13" : 20% "14" : 10% "8" : 10%]	44%["13" : 23% "14" : 14% "5" : 19%]	49%["11" : 17% "12" : 14% "13" : 11% "14" : 11%]
Location	31%["24" : 10% "26" : 12% "5" : 10%]	41%["0" : 16%]	29%["1" : 11% "3" : 11%]	19%["6" : 16%]
PP peritoneal contamination	23%["0" : 89%, "1" : 6%, "2" : 5%]	35%["0" : 83%, "1" : 14%, "2" : 3%]	26%["0" : 85%, "1" : 3%, "2" : 12%]	11%["0" : 100%]
ACS height	28%["11" : 16% "12" : 13% "13" : 24% "8" : 16%]	41%["5" : 13% "6" : 15% "8" : 12% "9" : 12%]	20%["12" : 31% "13" : 23% "5" : 12%]	34%["11" : 13% "12" : 20% "13" : 18% "7" : 18%]
Total points of NAS	27%["10" : 11% "5" : 18% "6" : 11% "7" : 16% "8" : 14%]	19%["14" : 13% "6" : 23% "7" : 13% "8" : 13%]	33%["11" : 12% "5" : 19% "6" : 19%]	31%["12" : 10% "6" : 20% "7" : 18% "8" : 18% "9" : 10%]
Anesthesia request type	19%["0" : 42%, "1" : 6%, "2" : 52%]			34%["0" : 71%, "1" : 13%, "2" : 16%]
UCI motive of admission	21%["0" : 63%, "1" : 23%, "2" : 14%]	22%["0" : 73% "1" : 11%]	27%["0" : 80%]	21%["0" : 89%]
PP number of procedures	14%["0" : 82%, "1" : 18%]	23%["0" : 95%, "1" : 5%]	23%[0" : 80%, "1" : 20%]	21%["0" : 50%, "1" : 46%, "2" : 4%]
PP respiratory		36%["0" : 53%, "1" : 20%, "2" : 22%, "3" : 5%]	26%["0" : 67%, "1" : 15%, "2" : 12%, "3" : 6%]	32%["0" : 52%, "1" : 2%, "2" : 41%, "3" : 5%]
Provenance		24%["0" : 63%, "1" : 17%, "2" : 20%]		
PP hemoglobin		19%["0" : 25%, "1" : 6%, "2" : 47%, "3" : 22%]	20%["0" : 54%, "1" : 15%, "2" : 19%, "3" : 11%]	26%["0" : 38%, "1" : 15%, "2" : 41%, "3" : 6%]
ACS functional state		27%["0" : 84%, "1" : 16%]	16%["0" : 100%]	39%["0" : 87%, "1" : 13%]
70% quality				
Constant				
3 labels				
ACS functional state	38%["0" : 100%]	18%["0" : 100%]	12%["0" : 83%, "2" : 17%]	55%["0" : 95%, "1" : 5%]
ACS Weight	31%["0" : 50%, "1" : 38%, "2" :12%]	43%["0" : 11%, "1" : 58%, "2" : 31%]	55%["0" : 68%, "1" : 3%, "2" : 29%]	63%["0" : 24%, "1" : 32%, "2" : 44%]
UCI motive of admission	25%["0" : 85%, "1" : 15%]		24%["0" : 1" : 25%, "7" : 25%]	30%["0" : 42%, "1" : 42%, "7" : 17%]
ACS height	40%["0" : 28%, "1" : 48%, "2" : 24%]	36%["0" : 44%, "1" : 50%, "2" : 6%]	22%["0" : 46%, "1" : 45%, "2" : 9%]	60%["1" : 25%, "2" : 75%]
PP respiratory	23%["0" : 75%, "1" : 25%]	32%["0" : 93%, "1" : 7%]	33%["0" : 94%, "1" : 6%]	20%["0" : 88%, "1" : 12%]
Provenance	21%["0" : 100%]	25%["0" : 91%, "1" : 9%]	35%["0" : 83%, "2" : 17%]	
PP number of procedures	17%["0" : 100%]	27%["0" : 100%]	33%[0" : 82%, "1" : 18%]	33%["0" : 100%]
PP peritoneal contamination	17%["0" : 100%]	27%["0" : 100%]	35%["0": 94%, "2" : 6%]	33%["0" : 100%]
PP hemoglobin			12%["0" : 67%, "3" : 33%]	13%["0" : 80%, "3" : 20%]
Anesthesia request type				23%["0" : 56%, "1" : 11%, "2" : 33%]
4 labels				
ACS Weight	39%["0" : 23%, "1" : 23%, "2" : 35%, "3" : 19%]	27%["1" : 43%, "2" : 43%, "3" : 14%]	19%["0" : 20%, "1" : 20%, "2" : 40%, "3" : 20%]	55%["0" : 3%, "1" : 28%, "2" : 22%, "3" : 47%]
PP respiratory	41%["0" : 88%, "1" : 13%]		15%["0" : 75%, "1" : 25%]	29%["0" : 76%, "1" : 18%, "3" : 6%]
PP peritoneal contamination	22%["0" : 100%]	48%["0" : 96%, "1" : 4%]	32%["0" : 76%, "2" : 24%]	24%["0" : 100%]
UCI motive of admission	18%["0" : 79%, "1" : 21%]	56%["0" : 59%, "1" : 38%, "2" :3%]	23%["0" : 92%, "7" : 8%]	22%["0" : 46%, "1" : 15%, "7" : 39%]
Provenance	15%["0" : 100%]	37%["0" : 100%]	32%["0" : 71%, "2" : 29%]	47%["0" : 70%, "1" : 4%, "2" : 26%]
PP number of procedures	18%["0" : 100%]	25%["0" : 100%]	26%["0" : 79%, "1" : 21%]	24%["0" : 100%]
ACS functional state	11%["0" : 100%]	19%["0" : 100%]	19%["0" : 70%, "1" : 10%, "2" : 20%]	64%["0" : 95%, "1" : 5%]
PP hemoglobin		21%["0" : 64%, "1" : 27%, "3" : 9%]		17%["0" : 50%, "1" : 20%, "3" : 30%]
Anesthesia request type		21%["0" : 27%, "2" : 73%]	21%["0" : 73%, "2" : 27%]	
ACS height		23%["0" : 8%, "1" : 42%, "2" : 50%]	19%["0" : 20%, "1" : 10%, "2" : 70%]	33%["1" : 26%, "2" : 26%, "3" : 48%]
5 labels				
UCI motive of admission	45%["0 " : 79%, "1" : 17%, "2" : 4%]	40%["0" : 68%, "1" : 14%, "2" : 5%, "5" : 4%, "7" : 9%]	33%["0" : 41%, "1" : 35%, "7" : 24%]	33%["0" : 40%, "1" : 27%, "2" : 7%, "7" : 20%, "8" : 6%]
PP respiratory	34%["0" : 100%]	24%["0" : 77%, "1" : 15%, "2" : 8%]	33%["0" : 88%, "3" : 12%]	31%["0" : 86%, "1" : 7%, "3" : 7%]
ACS Weight	33%["1" : 19%, "2" : 33%, "3" : 29%]	29%["0" : 12%, "1" : 6%, "2" : 19%, "3" : 44%, "4" : 19%]	37%["0" : 37%, "1" : 42%, "3" : 21%]	58%["0" : 23%, "1" : 20%, "2" : 19%, "3" : 23%]
PP peritoneal contamination	20%["0" : 100%]		29%["0" :100%]	29%["0": 92%, "2" : 8%]
ACS functional state	16%["0" : 90%, "1" :10%]	16%["0" :100%]	19%["0" :100%]	47%["0" : 90%, "1" :10%]
Provenance	19%["0" : 100%]	36%["0" : 90%, "2" : 10%]	37%["0" : 79%, "2" : 21%]	
PP number of procedures	14%["0" : 100%]	35%["0" : 89%, "1" :11%]	29%["0" :100%]	29%["0" : 92%, "1" :8%]
Anesthesia request type		33%["0" : 50%, "1" : 11%, "2" : 39%]	15%["0" : 88%, "2" : 12%]	29%["0" : 62%, "1" : 7%, "2" : 31%]
PP hemoglobin				22%["0" : 40%, "1" : 50%, "3" : 10%]
Order Preserving				
PP respiratory	58%["0" : 58%, "1" : 29%, "2" : 13%]	57%["0" : 55%, "1" : 23%, "2" : 17%, "3" : 5%]		30%["0" : 58%, "1" : 11%, "2" : 22%, "3" : 9%]
PP peritoneal contamination	22%["0" : 91%, "1" : 9%]	21%["0" : 91%, "1" : 7%, "2" : 2%]	26%["0" : 87%, "1" : 5%, "2" : 3%, "3" : 5%]	12%["0" : 81%, "1" : 19%]
ACS Weight	52%["12" : 10%, "13" : 12%, "14" : 11%, "6" : 16%]	67%["13" : 10%, "14" : 15%, "4" : 10%, "5" : 10%, "8" : 10%]	42%["13" : 18%, "14" : 12% , "4" : 12%, "5" : 13%, "6" : 10%]	61%["12" : 11%, "13" : 16% "15" : 11%]
ACS height	32%["10" : 14%, "11" : 10%, "12" : 12%, "13" : 16%, "8" : 10%, "9" : 10%]	50%["10" : 13%, "8" : 11%, "9" : 13%]	20%["12" : 24%, "13" : 14%, "14" : 10%, "5" : 14%, "6" : 10%]	47%["11" : 18%, "12" : 17%, "13" : 17%]
Location	34%["0" : 13%]		35%["0" : 8%, "2" : 8%, "3" : 8% "4" : 8%, "6" : 8%]	22%["1" : 15%, "12" : 11%]
PP number of procedures	18%["0" : 79%, "1" : 21%]	36%["0" : 99%, "1" : 1%]	18%["0" : 88%; "1" : 12%]	
ACS functional state	12%["0" : 95%, "1" : 5%]	28%["0" : 88%, "1" : 10%, "2" : 2%]	17%["0" : 96%, "1" : 4%]	38%["0" : 84%, "1" : 16%]
UCI motive of admission	24%["0" : 66%, "1" : 21%, "2" : 13%]	18%[0 : 68%]	28%["0" : 63%, "1" : 10% "7" : 12%]	25%["0" : 51%, "5" : 16%, "7" : 13%]
Anesthesia request type		22%["0" : 48%, "1" : 20%, "2" : 32%]		40%["0" : 57%, "1" : 21%, "2" : 22%]
Provenance		25%["0" : 82%, "1" : 4%, "2" : 14%]	38%["0": 80%, "1" : 11%, "2" : 9%]	45%["0" : 60%, "1" : 16%, "2" : 24%]
PP hemoglobin			17%["0" : 71%, "1" : 4%, "2" : 21%, "3" : 4%]	29%["0" : 39%, "1" : 27%, "2" : 26%, "3" : 8%]

Figure A.12: Occurrence of variables and their values in the biclusters of Clavien-Dindo score, variables dataset.



	IV.a	IV.b	V
100% quality			
Constant			
3 labels			
ACS height	15%["1" : 63%, "2" : 37%]	38%["0" : 29%, "1" : 41%, "2" : 29%]	50%["0" : 6%, "1" : 75%, "2" : 19%]
ACS Weight	17%["0" : 67%, "2" : 33%]	31%["0" : 36%, "1" : 14%, "2" : 50%]	50%["0" : 37%, "1" : 44%, "2" : 19%]
ACS functional state	19%["0" : 90%, "2" : 10%]		
Provenance	19%["0" : 10%, "2" : 90%]	27%["0" : 67%, "2" : 33%]	34%["0" : 55%, "2" : 45%]
UCI motive of admission			19%["0" : 17%, "7" : 83%]
PP number of procedures	27%["0" : 86%, "2" : 14%]	20%["0" : 100%]	25%["0" : 75%, "1" : 25%]
PP peritoneal contamination	21%["0" : 100%]	18%["0" : 88%, "3" : 12%]	22%["0" : 71%, "3" : 29%]
PP respiratory		16%["0" : 71%, "3" : 29%]	
PP hemoglobin	13%["0" : 71%, "3" : 29%]		
Anesthesia request type	17%["0" : 78%, "2" :22%]	33%["0" : 93%, "2" : 7%]	
Order Preserving			
ACS Weight	57%["14" : 14% "15" : 14% "4" : 11% "5" : 14%]	42%["14" : 13% "5" : 13% "6" : 13%]	50%["12" : 14% "13" : 14% "8" : 12%]
Location	21%["0" : 27% "1" : 10% "2" : 10% "25" : 10% "27" : 13%]	12%["11" : 13% "17" : 13% "22" : 13% "25" : 13% "26" : 13%]	42%["0" : 37% "2" : 10% "4" : 10%]
PP peritoneal contamination	42%["0" : 86%, "1" : 7%, "2" : 7%]	33%["0" : 88%, "1" : 2%, "2" : 9%]	35%["0" : 65%, "1" : 2%, "2" : 29%, "3" : 4%]
ACS height	30%["10" : 14% "12" : 17% "7" : 12% "9" : 17%]	12%["11" : 13% "12" : 20% "6" : 13%]	37%["10" : 13% "11" : 15% "12" : 20% "13" : 19% "9" : 13%]
Total points of NAS	14%["10" : 15% "11" : 20% "5" : 15% "6" : 15% "7" : 10% "9" : 10%]	39%["11" : 10% "5" : 22% "6" : 20% "8" : 18%]	14%["10" : 19% "11" : 14% "12" : 10% "13" : 10% "6" : 24% "8" : 10% "9" : 10%]
Anesthesia request type		40%["0" : 81%, "1" : 15%, "2" :4%]	13%["0" : 47%, "1" : 32%, "2" : 21%]
UCI motive of admission	24%["5" : 15% "6" : 41% "7" : 29%]	17%["0" : 77%]	
PP number of procedures	22%["0" : 78%, "1" : 19%, "2" : 3%]	33%[0" : 90%, "1" : 10%]	33%["0" : 83%, "1" : 17%]
PP respiratory	13%["0" : 61%, "1" : 22%, "2" : 6%, "3" : 11%]	11%["0" : 29%, "1" : 21%, "2" : 43%, "3" : 7%]	14%["0": 25%, "1" : 10%, "2" : 60%, "3" : 5%]
Provenance	17%["0": 92%, "1" : 4%, "2" : 4%]		37%["0" : 80%, "1" : 9%, "2" : 11%]
PP hemoglobin	12%["0" : 24%, "1" : 23%, "2" : 53%]	29%["0" : 29%, "1" : 29%, "2" : 39%, "3" : 3%]	12%["0" : 29%, "1" : 12%, "2" : 59%]
ACS functional state	13%["0" : 74%, "1" : 26%]		18%["0" : 42%, "1" : 50%, "2" : 8%]
70% quality			
Constant			
3 labels			
ACS functional state	21%["0" : 90%, "2" : 10%]		
ACS Weight	23%["0" : 46%, "1" : 18%, "2" : 36%]	27%["0" : 42%, "1" : 16%, "2" : 42%]	45%["0" : 31%, "1" : 54%, "2" : 15%]
UCI motive of admission		13%["0" : 67%, "1" : 17%, "7" : 16%]	28%["1" : 50%, "2" : 25%, "7" : 25%]
ACS height	10%["0" : 20%, "1" : 80%]	33%["0" : 27%, "1" : 47%, "2" : 26%]	45%["1" : 77%, "2" : 23%]
PP respiratory	15%["0" : 100%]	16%["0" : 72%, "1" : 14%, "3" : 14%]	10%[0" : 100%]
Provenance	23%["0" : 36%, "2" : 64%]	31%["0" : 64%, "2" : 36%]	
PP number of procedures	23%["0" : 82%, "2" : 18%]		35%["0" : 90%, "1" : 10%]
PP peritoneal contamination	25%["0" : 92%, "1" : 8%]	33%["0" : 80%; "1" : 7%, "3" : 13%]	14%["0" :100%]
PP hemoglobin	17%["0" : 63%, "3" : 37%]	13%["0" : 17%, "1" : 33%, "3" : 50%]	10%["3" : 100%]
Anesthesia request type	21%["0" : 70%, "2" : 30%]	40%["0" : 83%, "2" : 17%]	31%["0" : 89%, "1" : 11%]
4 labels			
ACS Weight	39%["0" : 36%, "1" : 25%, "2" : 21%, "3" : 18%]	39%["0" : 67%, "1" : 4%, "2" : 26%, "3" : 3%]	44%["0" : 17%, "1" : 46%, "2" : 29%, "3" : 8%]
PP respiratory	15%["0" : 82%, "1" : 18%]	13%["0" : 67%, "1" : 11%, "3" : 22%]	20%["0" : 64%, "1" : 18%, "3" : 18%]
PP peritoneal contamination	44%["0" : 81%, "1" : 10%, "2" : 6%, "3" : 3%]		
UCI motive of admission		28%["0" : 32%, "1" : 26%, "2" : 16%, "7" : 21%, 8 : 5%]	16%["0" : 33%, "1" : 44%, "7" : 22%]
Provenance	31%["0" : 55%, "2" : 45%]	30%["0" : 38%, "2" : 62%]	45%["0" : 84%, "1" : 8%, "2" : 8%]
PP number of procedures	32%["0" : 65%, "1" : 9%, "2" : 26%]	23%["0" : 94%, "2" : 6%]	24%["0" : 92%, "1" : 8%]
ACS functional state	28%["0" : 70%, "1" : 10%, "2" :20%]		
PP hemoglobin	22%["0" : 31%, "1" :13%, "2" : 19%; "3" : 37%]	13%["0" : 11%, "1" : 33%, "3" : 56%]	
Anesthesia request type	25%["0" : 78%, "1" : 5%, "2" : 17%]	35%["0" : 88%, "1" : 4%, "2" : 8%]	22%["0" : 92%, "2" : 8%]
ACS height	26%["0" : 10%, "1" : 16%, "2" : 58%, "3" : 16%]	19%["0" : 15%, "1" : 31%, "2" : 31%, "3" 23%]	29%["1" : 88%, "2" : 6%, "3" : 6%]
5 labels			
UCI motive of admission	26%["0" : 18%, "10" : 17%, "7" : 65%]	14%["0" : 46%, "1" : 18%, "7" : 36%]	23%["0" : 44%, "1" : 6%, "2" : 6%; "7" : 44%]
PP respiratory	24%["0" : 69%, "1" :5%, "2" : 5%, "3" : 21%]		
ACS Weight	24%["0" : 25%, "1" : 19%, "2" : 12%, "3" : 19%, "4" : 25%]	27%["0" : 57%, "3" : 29%]	45%["0" : 16%, "1" : 25%, "2" : 41%]
PP peritoneal contamination	36%["0" : 88%, "2" : 4%, "3" : 8%]	32%["0" : 72%, "1" : 8%, "3" : 20%]	25%["0" : 94%, "1" :6%]
ACS functional state	36%["0" :84%, "1" :8%, "2" : 8%]	15%["0" : 59%, "1" : 8%, "2" : 33%]	
Provenance		42%["0" : 45%, "2" : 55%]	35%["0" :72%, "2" :28%]
PP number of procedures	29%["0" : 63%, "1" : 11%, "2" :26%]	28%["0" : 91%, "2" : 9%]	28%["0" : 100%]
Anesthesia request type	27%["0" : 78%, "2" : 22%]		20%["0" : 79%, "1" : 14%, "2" : 7%]
PP hemoglobin	18%["0" : 50%, "1" : 8%, "2" : 8%, "3" : 34%]	16%["0": 16%, "1" : 23%, "2" : 15%, "3" : 46%]	
Order Preserving			
PP respiratory	13%["0" : 72%, "1" : 22%, "3" : 6%]	13%["0" : 18%, "1" : 18%, "2" : 53%, "3" 11%]	18%["0" : 22%, "1" : 4%, "2" : 70%, "3" : 4%]
PP peritoneal contamination	33%["0" : 82%, "1" : 11%, "2" : 7%]	36%["0" : 84%, "1" : 4%, "2" : 12%]	39%["0" : 62%, "1" : 4%, "2" : 32%, "3" : 2%]
ACS Weight	50%["12" : 12%, "14" : 17%, "15" : 10%, "4" : 10%]	43%["13" : 10%, "4" : 12%, "6" : 19%]	55%["13" : 17%, "8" : 15%]
ACS height	50%["10" : 15%, "12" : 15%]	13%["12" : 17%, "14" : 11% "16" : 11%, "6" : 11%, "7" : 17%, "8" : 17%]	47%["10" : 18%, "11" : 16%, "12" : 20%, "13" : 16%]
Location	28%["0" : 15%, "2" : 13%, "27" : 10%]		
PP number of procedures	37%["0" : 67%, "1" :33%]	49%["0" : 91%, "1" 8%, "2" : 1%]	38%["0" : 88%, "1" : 12%]
ACS functional state	13%["0" : 83%, "1" : 17%]		21%["0" : 48%, "1" : 48%, "2" : 4%]
UCI motive of admission	25%["5" : 31%, "6" : 31%, "7" : 23%]		12%["0" : 33% "7" : 13%, "8" : 33%]
Anesthesia request type		38%["0" : 84%, "1" : 14%, "2" : 2%]	11%["0" : 43%, "1" : 43%, "2" : 14%
Provenance	18%["0" : 92%, "1" : 4%, "2" : 4%]		35%["0" : 69%, "1" : 22%, "2" : 9%
PP hemoglobin		24%["0" : 33%, "1" : 27%, "2" : 40%]	10%["0" : 15%, "1" : 15%, "2" : 70%]

Figure A.13: Occurrence of variables and their values in the biclusters of Clavien-Dindo score, variables dataset.



	I	II	III.a	III.b	IV.a	IV.b	V
100% quality							
Constant							
ICD_5497	27%		22%				
ICD_5122	18%						
ICD_544	18%		22%				
ICD_8670		27%		15%			
ICD_4041		13%					
ICD_4399		13%					
ICD_2933		13%					
ICD_933		13%					
ICD_415		13%					
ICD_27		13%					
ICD_3459			33%				
ICD_3451			33%	31%			
ICD_26				46%			
ICD_5771				15%			
ICD_5651				15%			
ICD_3249				31%			
ICD_4042				15%	20%		29%
ICD_412					20%		
ICD_631					20%		
ICD_4610						29%	
ICD_610						29%	
ICD_3195							29%
Order Preserving							
ICD_311	29%						
ICD_5497	29%						
ICD_4562	29%						
ICD_5459	29%						
ICD_252		13%	29%				
ICD_4399		13%					
ICD_4041		13%					
ICD_2933		13%					
ICD_2759		19%					
ICD_415		13%					
ICD_2749		13%					
ICD_8670		19%					
ICD_3451			29%	25%			
ICD_3459			43%	25%			
ICD_5252			29%				
ICD_3249				38%			
ICD_26				75%			
ICD_631					25%		
ICD_4576					25%		
ICD_576					25%		
ICD_4863						40%	
ICD_5122							50%

70% quality	I	II	III.a	III.b	IV.a	IV.b	V
Constant							
ICD_5122	25%						
ICD_4709	17%						
ICD_544	17%		22%				
ICD_5497	17%		22%				
ICD_8670		27%		15%			
ICD_4041		13%					
ICD_4399		13%					
ICD_2933		13%					
ICD_415		13%					
ICD_27		13%					
ICD_3459			33%				
ICD_3451			33%	31%			
ICD_26				46%			
ICD_5771				15%			
ICD_5651				15%			
ICD_3249				31%			
ICD_4042				15%	20%		29%
ICD_412					20%		
ICD_631					20%		
ICD_4610						29%	
ICD_3195							29%
Order Preserving							
ICD_311	29%						
ICD_5497	29%						
ICD_4562	29%						
ICD_5459	29%						
ICD_252		15%					
ICD_4399		15%					
ICD_2933		15%					
ICD_415		15%					
ICD_27		15%					
ICD_8670		15%					
ICD_3451			33%	20%			
ICD_3459				50%	20%		
ICD_26				70%			
ICD_3249				40%			
ICD_631					25%		
ICD_4042					25%		
ICD_4576					25%		
ICD_576					25%		
ICD_4863						40%	
ICD_5122							50%

Figure A.14: Occurrence of variables and their values in the biclusters of Clavien-Dindo score, ICD-10 dataset.

Table A.1: **Destination after IPO**, **destination after HDU**, **moment of death after sugery categories**, **admitted into ICU**, **readmission into HDU**, **death within 1 year after surgery**, **provenance of patient** using BicPAMS with range based 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	$\mu(\lvert J\rvert)$ $\pm\sigma(\lvert J\rvert)$	$\mu(\lvert I\rvert)$ $\pm\sigma(\lvert I\rvert)$
Destination after IPO : Death	Constant	70%	3	3	2.5	76	71	3.97 ± 1.00	121.01 ± 20.38	readmission into HDU	2.5	74	70	3.76 ± 0.78	114.53 ± 17.09
	Constant	70%	3	8	1.3	47	47	8.59 ± 0.79	90.29 ± 41.28		1.3	50	50	8.76 ± 1.21	77.72 ± 30.69
	Constant	70%	4	3	2.5	190	168	3.54 ± 0.81	92.11 ± 19.21		2.5	84	72	3.28 ± 0.51	92.42 ± 11.81
	Constant	70%	4	8	1.3	40	40	8.53 ± 0.77	73.98 ± 27.52		1.3	39	39	8.62 ± 1.03	70.54 ± 19.20
	Constant	70%	5	3	3	109	91	3.60 ± 0.86	100.54 ± 21.51		2.5	94	75	3.48 ± 0.67	77.45 ± 10.36
	Constant	70%	5	8	1.3	49	49	8.77 ± 1.11	83.08 ± 16.26		1.3	60	60	8.65 ± 0.87	64.57 ± 18.32
Destination after HDU : ICU	Constant	70%	3	3	2.5	100	99	3.63 ± 0.78	105.41 ± 18.94	Death after surgery within 1 year	1.5	104	101	3.50 ± 0.93	141.54 ± 22.0
	Constant	70%	3	8	1.3	52	52	8.71 ± 0.81	111.89 ± 38.07		1.3	41	41	8.76 ± 1.00	87.27 ± 42.68
	Constant	70%	4	3	2.5	106	98	3.58 ± 0.83	97.69 ± 14.89		1.5	124	98	3.34 ± 0.79	125.06 ± 29.76
	Constant	70%	4	8	1.3	68	68	8.44 ± 0.69	71.05 ± 23.98		1.3	41	41	8.49 ± 0.63	79.66 ± 21.42
	Constant	70%	5	3	2.5	153	130	3.59 ± 0.77	75.01 ± 14.82		1.5	117	81	3.25 ± 0.56	116.41 ± 20.15
	Constant	70%	5	8	1.3	33	33	8.75 ± 0.85	97.58 ± 18.65		1.3	31	31	8.61 ± 0.79	86.26 ± 19.74
Death after surgery $0-30$ (days)	Constant	70%	3	3	2.5	51	41	4.0 ± 1.05	22.27 ± 2.53	Provenance nursery	2	33	27	5.85 ± 1.51	111.93 ± 14.53
	Constant	70%	3	8	1.3	40	40	8.68 ± 0.82	20.05 ± 8.74		1.3	38	38	8.74 ± 0.96	77.5 ± 44.51
	Constant	70%	4	3	2.5	57	39	4.02 ± 0.86	23.72 ± 3.69		2	46	31	4.32 ± 1	91.19 ± 10.70
	Constant	70%	4	8	1.3	41	41	8.88 ± 1.02	18.39 ± 4.43		1.3	98	98	8.64 ± 0.87	68.15 ± 22.08
	Constant	70%	5	3	2.5	43	38	3.58 ± 0.75	22.39 ± 3.07		2	92	57	4.44 ± 1.08	86.6 ± 12.10
	Constant	70%	5	8	1.3	31	31	9.29 ± 1.11	22.74 ± 3.80		1.3	33	33	8.73 ± 0.86	85.91 ± 17.29
Death after surgery $30-60$ (days)	Constant	70%	3	3	2.5	27	24	5.88 ± 1.45	18.54 ± 2.25	Provenance ICU	2	81	78	3.67 ± 0.87	105.79 ± 22.05
	Constant	70%	3	8	1.3	35	35	8.71 ± 0.85	23.28 ± 7.34		1.3	44	44	8.55 ± 0.78	81.95 ± 39.65
	Constant	70%	4	3	2.5	31	31	5.45 ± 1.72	17.94 ± 1.72		2	147	138	3.36 ± 0.63	86.75 ± 21.18
	Constant	70%	4	8	1.3	31	31	8.52 ± 0.76	18.77 ± 4.90		1.3	69	69	8.59 ± 0.84	61.65 ± 21.79
	Constant	70%	5	3	2.5	25	19	4.37 ± 0.81	17.11 ± 0.79		2	132	112	3.46 ± 0.72	80.66 ± 16.83
	Constant	70%	5	8	1.3	11	11	8.82 ± 1.19	25.27 ± 5.45		1.3	38	38	8.76 ± 0.93	74.68 ± 11.95
Death after surgery $60-365$ (days)	Constant	70%	3	3	1.3	94	74	4.77 ± 1.21	24.80 ± 3.39	Provenance unscheduled service	2	74	74	3.78 ± 0.99	221.73 ± 45.01
	Constant	70%	3	8	1.3	29	29	9.03 ± 1.3	16.93 ± 5.59		1.3	40	40	8.63 ± 0.83	104.95 ± 57.26
	Constant	70%	4	3	1.3	40	30	4.4 ± 1.01	26.50 ± 3.35		2.5	480	440	3.67 ± 0.78	138.04 ± 27.23
	Constant	70%	4	8	1.3	21	21	9.05 ± 1.13	15.95 ± 2.44		1.3	137	137	8.78 ± 0.88	91.34 ± 21.35
	Constant	70%	5	3	1.3	40	35	4.2 ± 0.91	27.0 ± 3.07		2.5	470	411	3.74 ± 0.89	125.86 ± 26.93
	Constant	70%	5	8	1.3	48	48	8.85 ± 1.10	13.40 ± 3.17		1.3	34	34	8.59 ± 0.73	96.65 ± 29.57
Admitted into ICU	Constant	70%	3	3	2	104	98	3.60 ± 0.75	111.29 ± 19.49		–	–	–	–	–
	Constant	70%	3	8	1.3	57	57	8.56 ± 0.82	75.58 ± 35.07		–	–	–	–	–
	Constant	70%	4	3	2	129	113	3.22 ± 0.47	85.15 ± 12.05		–	–	–	–	–
	Constant	70%	4	8	1.3	60	60	8.55 ± 0.78	71.27 ± 21.63		–	–	–	–	–
	Constant	70%	5	3	2	36	28	3.14 ± 0.35	85.75 ± 7.35		–	–	–	–	–
	Constant	70%	5	8	1.3	67	67	8.69 ± 0.85	62.82 ± 11.84		–	–	–	–	–

Table A.2: **Clavien-Dindo classes**, **presence of post-surgery complication**, **days spent in HDU** and **days spent in IPO** using BicPAMS with range based discretization.

	Assumption	quality	$\|L\|$	$\|C\|$	Lift	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|) \pm\sigma(\|I\|)$		Lift	#bics	p-value <0.001	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$\mu(\|I\|) \pm\sigma(\|I\|)$
Clavien-Dindo type I	Constant	70%	3	3	2.5	76	67	5.74 ± 1.73	96.56 ± 17.60	Presence of Post-surgery comp.	1.5	54	54	4.09 ± 1.15	117.11 ± 21.21
	Constant	70%	3	8	1.3	32	32	8.43 ± 0.66	124.78 ± 70.52		1.3	42	42	8.52 ± 0.66	77.38 ± 37.51
	Constant	70%	4	3	2.5	54	51	4.6 ± 1.17	122.01 ± 19.60		1.5	143	139	3.35 ± 0.67	77.85 ± 12.92
	Constant	70%	4	8	1.3	83	83	8.44 ± 0.56	107.36 ± 42.12		1.3	69	69	8.49 ± 0.81	62.94 ± 21.37
	Constant	70%	5	3	2.5	127	106	5.19 ± 1.49	92.18 ± 19.31		1.5	112	107	3.45 ± 0.73	78.11 ± 11.47
	Constant	70%	5	8	1.3	117	117	8.67 ± 0.81	99.84 ± 36.72		1.3	22	22	8.82 ± 0.94	79.82 ± 15.91
Clavien-Dindo type II	Constant	70%	3	3	1.7	70	66	4.12 ± 1.08	108.89 ± 20.01	HDU days < 1	1.5	67	55	5.83 ± 1.73	119.0 ± 25.62
	Constant	70%	3	8	1.3	46	46	8.63 ± 0.81	86.78 ± 33.61		1.3	31	31	8.61 ± 0.97	102.12 ± 58.91
	Constant	70%	4	3	1.7	82	80	3.85 ± 0.71	109.5 ± 12.09		1.5	113	105	5.84 ± 1.76	106.93 ± 30.57
	Constant	70%	4	8	1.3	120	120	8.63 ± 0.94	60.82 ± 20.31		1.3	53	53	8.64 ± 0.91	107.26 ± 40.77
	Constant	70%	5	3	1.7	103	97	4.14 ± 1.04	99.73 ± 9.89		1.5	166	143	4.92 ± 1.32	113.91 ± 37.28
	Constant	70%	5	8	1.3	48	48	8.75 ± 0.92	74.02 ± 19.73		1.3	18	18	8.83 ± 0.89	150.5 ± 55.82
Clavien-Dindo type III.a	Constant	70%	3	3	2.5	141	140	4.48 ± 1.32	120.68 ± 29.08	HDU days $1-2$	1.3	75	56	5.03 ± 1.68	92.08 ± 16.75
	Constant	70%	3	8	1.3	55	55	8.83 ± 0.90	86.90 ± 44.27		1.3	11	11	9.45 ± 1.59	101.91 ± 15.19
	Constant	70%	4	3	3	74	74	4.35 ± 1.16	105.35 ± 23.55		1.3	68	54	4.62 ± 1.43	98.62 ± 16.94
	Constant	70%	4	8	1.3	43	43	8.65 ± 0.83	80.41 ± 30.94		1.3	48	48	8.71 ± 0.93	71.58 ± 21.09
	Constant	70%	5	3	2.5	170	165	3.73 ± 0.93	115.12 ± 21.45		1.3	99	74	4.22 ± 1.11	88.0 ± 9.34
	Constant	70%	5	8	1.3	64	64	8.78 ± 0.85	80.47 ± 30.42		1.3	80	80	9.03 ± 1.21	59.81 ± 10.51
Clavien-Dindo type III.b	Constant	70%	3	3	2.5	93	82	5.44 ± 1.63	79.68 ± 10.23	HDU days > 2	2	11	11	5.36 ± 2.01	104.81 ± 15.93
	Constant	70%	3	8	1.3	57	57	8.38 ± 0.61	78.33 ± 31.96		1.3	48	48	8.81 ± 1.13	86.33 ± 39.70
	Constant	70%	4	3	2.5	26	25	4.36 ± 1.26	95.32 ± 9.78		2	37	37	4.21 ± 1.37	78.16 ± 8.29
	Constant	70%	4	8	1.3	44	44	8.84 ± 0.97	96.93 ± 21.43		1.3	32	32	8.65 ± 1.04	82.5 ± 30.33
	Constant	70%	5	3	2.5	48	40	4.33 ± 1.10	77.18 ± 5.98		1.5	123	104	3.31 ± 0.59	102.78 ± 22.47
	Constant	70%	5	8	1.3	11	11	8.73 ± 0.75	119.27 ± 32.09		1.3	37	37	8.78 ± 0.90	84.91 ± 27.21
Clavien-Dindo type IV.a	Constant	70%	3	3	3	138	134	3.91 ± 1.25	127.57 ± 20.50	IPO days < 7	2.5	46	40	5.55 ± 1.90	124.65 ± 41.55
	Constant	70%	3	8	1.3	32	32	8.88 ± 0.99	102.94 ± 40.93		1.3	32	32	8.37 ± 0.59	112.68 ± 81.85
	Constant	70%	4	3	3	149	126	3.88 ± 0.83	98.60 ± 9.74		2.5	102	89	5.03 ± 1.56	106.02 ± 31.02
	Constant	70%	4	8	1.3	55	55	8.43 ± 0.68	71.94 ± 19.22		1.3	39	39	8.56 ± 0.74	115.41 ± 44.99
	Constant	70%	5	3	3	153	126	3.57 ± 0.77	88.13 ± 11.9		2.5	113	107	5.28 ± 1.70	121.42 ± 29.25
	Constant	70%	5	8	1.3	15	15	8.6 ± 1.02	92.27 ± 24.00		1.3	78	78	8.74 ± 0.85	123.08 ± 43.81
Clavien-Dindo type IV.b	Constant	70%	3	3	3	143	133	4.18 ± 0.97	107.54 ± 17.95	IPO days $7-10$	1.7	18	16	5.31 ± 1.44	126.56 ± 37.47
	Constant	70%	3	8	1.3	39	39	8.71 ± 0.96	98.71 ± 29.78		1.3	37	37	8.83 ± 0.94	118.54 ± 66.30
	Constant	70%	4	3	3	75	70	3.74 ± 0.87	96.92 ± 9.90		1.7	49	48	4.72 ± 1.03	127.58 ± 23.26
	Constant	70%	4	8	1.3	73	73	8.46 ± 0.68	58.86 ± 16.27		1.3	38	38	8.55 ± 0.87	90.44 ± 32.88
	Constant	70%	5	3	3	81	75	3.71 ± 0.98	91.32 ± 9.21		2	37	37	5.59 ± 1.51	109.81 ± 11.88
	Constant	70%	5	8	1.3	49	49	8.63 ± 0.72	71.61 ± 20.03		1.3	42	42	8.78 ± 0.96	90.66 ± 31.21
Clavien-Dindo type V	Constant	70%	3	3	3	66	59	4.35 ± 1.23	96.18 ± 19.23	IPO days > 10	1.5	71	69	4.04 ± 1.21	102.63 ± 20.90
	Constant	70%	3	8	1.3	64	64	8.46 ± 0.68	77.29 ± 35.04		1.3	28	28	8.78 ± 0.72	91.96 ± 47.29
	Constant	70%	4	3	2.5	117	95	3.58 ± 0.79	109.31 ± 32.28		1.5	77	77	3.61 ± 0.72	106.77 ± 19.84
	Constant	70%	4	8	1.3	63	63	8.44 ± 0.68	69.39 ± 26.35		1.3	29	29	8.55 ± 0.72	76.13 ± 16.01
	Constant	70%	5	3	3	137	123	3.59 ± 0.87	78.82 ± 15.56		1.5	81	78	3.71 ± 0.87	85.5 ± 12.0
	Constant	70%	5	8	1.3	48	48	8.48 ± 0.65	81.48 ± 22.86		1.3	16	16	8.62 ± 0.92	82.18 ± 14.69