Identifying Survival-Changing Sequential Patterns for Employee Attrition Analysis

Youssef Oubelmouh, Frédéric Fargon, Cyril de Runz, Arnaud Soulet, Cyril Veillon

Identifying Survival-Changing Sequential Patterns for Employee Attrition Analysis

Youssef Oubelmouh
Devoteam/University of Tours
Blois, France
youssef.oubelmouh@univ-tours.fr

Frédéric Fargon
Devoteam
Levallois-Perret, France
frederic.fargon@devoteam.com

Cyril de Runz
University of Tours
Blois, France
cyril.derunz@univ-tours.fr

Arnaud Soulet
University of Tours
Blois, France
arnaud.soulet@univ-tours.fr

Cyril Veillon
Devoteam
Levallois-Perret, France
cyril.veillon@devoteam.com

***Abstract*—Employee attrition is a pervasive problem for many organizations, and reducing it has become a key goal in the business world. Although there is a substantial body of literature on predicting customer attrition, the literature on employee attrition is comparatively limited. Moreover, even studies that do address employee attrition often fail to consider the impact of time and duration on attrition rates. In this context, the present paper aims to fill this gap in the literature by combining frequent pattern mining in sequences of events and survival analysis with Kaplan-Meier to examine how event sequences affect employee attrition. We introduce the notion of survival-changing sequential patterns that highlight events that significantly impact the survival estimator. Our findings suggest that certain patterns are associated with a higher rate of employee retention, while the addition of specific events can have a positive or negative impact on employee survival. This research highlights the importance of analyzing event sequences and duration when attempting to reduce employee attrition rates. The practical implications of this research are significant, as it provides a framework for organizations seeking to retain their employees and enhance their overall performance.**

***Index Terms*—Employee attrition, survival analysis, sequential pattern mining, interestingness metrics**

I. Introduction

Employee attrition is a growing problem in technology companies around the world, and particularly in IT companies. According to the DARES[1] [1], France recorded nearly 520,000 resignations per quarter, at the end of 2021 and the beginning of 2022, including 470,000 resignations from permanent contracts, which represents a resignation rate of 2.7%. This rate is much higher in the consulting firms, reaching for example 19% at Accenture, 16% at Atos and even 25% in the technology sector in India. In fact, consultants stay in general between 3 and 8 years in their IT company [2].

Spontaneous resignations imply a decrease in productivity, even more if the employee was specialized in a domain or had seniority. Acquiring new employees involves many costs and lost time, whether it is for searching in the recruitment phase or for training and adaptation to a new environment after hiring. The average cost of disengagement and unavailability in France is 14,580 euros per year and per employee, of which 9,185 euros are controllable, according to the conclusions of the IBET [2] [3]. It is therefore preferable for companies to identify and activate the levers that will allow them to retain their employees who wish to leave.

Many studies based on approaches from the humanities [4] have been conducted for more than 30 years to try to determine what factors drive employees to quit. More recently, especially since IBM released a dummy dataset in 2016 [3], work has focused on using data mining and machine learning approaches for the primary purpose of prediction [5]–[7]. However, this dataset provides limited information about the underlying dynamics. [8] are the only ones who artificially introduced temporality into their data. To do this, they randomly decided for each person who left, whether he or she left in the middle of the year or at the end of the year on the IBM data.

Thus, despite the growing body of research on predicting employee attrition, we observe that little or no consideration is given to the temporal dynamics of employee trajectories, due to the nature of the snapshot datasets. For example, the few time-related variables in the IBM data are not comprehensive and/or precise enough, e.g. Years In Current Role, Years With Current Manager. It should be noted that the reason why IBM's fictional data was used in many works is because the sensitivity of the personal information that can be contained in it does not allow us to propose datasets that are open to the community (GDPR constraints).

To better capture the temporal aspect of attrition, we propose to combine sequential pattern mining techniques in event sequences and survival analysis. Sequential pattern mining aims to discover significant patterns in sequences of events, such

This work has been partially supported by the CIFRE program of the ANRT project 2021/0760

[1]DARES: Direction de l'Animation de la Recherche, des Études et des Statistiques (French Institute)

[2]IBET : Indice du Bien-Être au Travail (Index of Well-Being at Work)

[3]https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

as the sequence of changes that an employee has undergone. Survival analysis is used to model the time to event data, such as the time from hiring to resignation, and to estimate the survival function of a population.

The objective of this paper is to propose an approach to extract employee attrition/retention factors by considering the temporal aspect of the phenomenon through sequential pattern mining and survival analysis. Our approach aims to answer the following questions:

1) What are the successions of events (pattern) leading to a resignation?
2) What are the succession of factors (survival-changing patterns), appearing after a succession of events, that can prevent/delay or accelerate the resignation of an employee?
3) How to link sequential pattern mining and survival analysis?

The contributions of this paper are twofold. First, we propose a new method for extracting survival-changing patterns. Second, we leverage this approach to analyze the phenomenon of attrition using real data from an IT firm, with the goal of identifying patterns that have a significant impact on employee attrition within their internal trajectory in the company.

The structure of the rest of this paper is as follows: Section II focuses on the related work that pertains to the concepts used in this study. Section III provides background information and presents the formal framework. Section IV introduces our proposed method, called survival-changing analysis. This section exposes how to extract survival-changing patterns. Following that, we present the experimental evaluations of our method on our dataset in Section V. Finally, Section VI provides our conclusions and outlines our future work.

II. Related works

This section provides an overview of related work in relation to the concepts utilized in this study. We review relevant prior research to contextualize our approach within the existing literature. By exploring previous contributions, we aim to highlight the relevance and novelty of our method combining survival analysis and sequential pattern mining.

A. Survival analysis

Survival analysis is a general term referring to any analysis of the occurrence of all-or-nothing events over time, such as death, in the presence of censored data [9], [10] (data not fully observed during the study period, e.g., if the event being studied is death, patients who did not die during the study or who were not "lost to follow-up" during the study). This type of analysis is widely used in clinical research. It can be used to describe the survival rate of a group of patients (the time between the start of treatment and the occurrence of death) and also to compare the survival rates of two or more groups of patients to study prognostic factors, i.e., what might explain the occurrence of death (or another event) over time.

Recall that, although survival is the term used for death for historical reasons (a term first used in oncology where patient survival is one of the criteria for treatment efficacy), survival analysis methods are not only applicable to the study of deaths but can also be applied to the study of any "single" event that may occur during a trial: the apparition of a specific type of cancer [11], student dropout [12], click on an ad [13], or attrition which is the event considered in this paper.

In survival analysis, three types of statistical methods are used to calculate survival rates: parametric [14], non-parametric [15]–[17], and semi-parametric [18], [19] methods. Parametric methods require specific assumptions about the distribution, while non-parametric methods do not. Semi-parametric methods combine both approaches. In our study, we have chosen a non-parametric method due to the absence of underlying distribution regarding employee attrition, which allows for a more accurate modeling of survival. The authors in [20] conducted a comprehensive survey on various statistical and non-statistical methods used in survival analysis.

B. Sequential pattern mining

In the data mining domain, a sequence refers to a series of ordered events or transactions that occur over time. These sequences can be represented in various formats, such as strings, arrays, or graphs, depending on the nature of the data and the context of the analysis.

Sequential pattern mining is a specialized subfield of data mining that is focused on identifying recurrent patterns in sequences. The main objective of sequential pattern mining is to discover patterns that occur frequently in the data, which can then be used for analytical purposes such as prediction, classification [21], and more. The mining of frequent patterns in sequences is a powerful tool for identifying recurring patterns in time-ordered data. Sequential pattern mining can be found across many domains such as financial transaction [22], biotechnology [23], Web Usage [24], and so on. For instance, it can be employed to identify common customer purchase patterns, detect anomalies in medical data, or forecast future web browsing behavior. Sequential pattern mining is a highly active research area, with numerous surveys conducted to explore this domain comprehensively [25], [26]. Few works have specifically addressed the analysis of attrition in companies. Our approach combines sequential pattern mining with survival analysis.

C. Sequential pattern mining for survival analysis

In this subsection, we examine related works combining sequential pattern mining and survival analysis. The authors of [27] propose an approach combining pattern mining and survival analysis for finding meaningful links between events in sequences. Traditional association rule mining methods rely primarily on event frequency, but this may be insufficient to capture links between infrequent but highly associated events. This work, therefore, propose a duration model approach and use a semi-parametric proportional hazards model to estimate the influence of events on each other. This method handles censored data and takes into account statistical significance for rule selection. In our context, this approach allows us to

extract the successions of events that may produce resignation and it will give the risk to produce it, therefore it could be useful to predict the resignation but it will not give how an event sequence will delay the resignation.

[28] have proposed a novel data mining algorithm that combines survival analysis with sequence mining to evaluate the survival associated with sequential medical treatments. Their algorithm efficiently analyzes sequences of renal replacement therapies, providing valuable insights into patient survival patterns. The algorithm incorporates frequent gaps between treatments, evaluates survival time during treatment execution, and applies pruning strategies based on support and median survival. It focuses on prefixed temporal windows and the extracted patterns are subsequences in these temporal windows. This approach does not allow extracting patterns, and thus survival-changing patterns, that may group events that are not directly consecutive which is also our goal.

In [29], survival prediction using health insurance data is enhanced through graph pattern mining. This approach leverages frequent patterns extracted from a representation of patient data. An improved random forest model is employed to handle censored data and predict survival time. Experimental results demonstrate superior performance compared to traditional survival prediction models. This paper is focused on graph pattern mining and its general idea is more on the prediction of the survival time than on the identification of survival-changing events/patterns.

A data-driven approach is proposed in [30] to identify frequent sets of course failures that increase the risk of student dropout. Survival analysis determines the overall probability distribution of dropout, while event analysis incorporates the impact of different course failures. This paper has some weaknesses for our case study as it does not consider the sequential aspect of the events (not based on sequential pattern mining), and the analysis is more global than contextualized.

Using interval sequential pattern mining, [31] analyzed hospitalization records to evaluate disease progression and identify potential factors leading to short-term death after myocardial infarction diagnosis. The analysis revealed five disease pathways, which covered a significant portion of the cohort. These trajectory patterns provide insights for early identification of high-risk individuals and the potential for more aggressive interventions to lower mortality rates.

[32] used an extension of subgroup discovery which is the field of exceptional model mining [33], which applies subgroup discovery algorithms to investigate heterogeneous groups in statistical models that somehow deviate from the norm. The survival response significantly differs from the overall behavior. Therefore, these last two works [31], [32] result in a global asset of the patterns' interestingness than a contextual one that is wanted to identify the survival-changing patterns.

In conclusion, to the best of our knowledge, there is a gap in the literature about the extraction of survival-changing patterns (and not rules) in sequences, and there is no work that exploits this concept for employee attrition analysis. The following section will present the background necessary to introduce our method that fills this gap.

III. Background

A. Basic definitions for sequential patterns

Let $\mathcal{A} = \{A_1, A_2, ..., A_n\}$ be a set of attributes. Let $\mathcal{C} = \{C_1, C_2, ..., C_n\}$ represent the set of changes. A change corresponds to the variation of an attribute. For example, attribute A_1 transitioning from one state to another is a change of attribute (C_1 occurs). We define $\mathcal{E}$ as the set of events, where an event is a pair containing a change and a time. Thus, $\mathcal{E} \subseteq \mathcal{C} \times \mathbb{N}$. For instance, $E_4 = (C_4, t_1)$ is an event, indicating that attribute A_4 changed state at time t_1. A sequence of events $Seq \in \mathcal{S}$ is a set of events ordered in time, such as $Seq_1 = \langle E_3, E_8, ..., E_2 \rangle = \langle (C_3, t_0), (C_8, t_1), ..., (C_2, t_f) \rangle$, where $\forall i$, $E_i \in \mathcal{E}$, and for all t_i, t_j, $i < j \Rightarrow t_i \leq t_j$.

To illustrate our definitions, Table I provides a toy dataset containing 8 sequences described by the changes $\{S, M, P, A, B, C, D, E, F, G, H, J\}$ where we removed the timestamps for sake of simplicity by keeping only the duration between the hiring and the resignation.

TABLE I
Toy Dataset

ID	Sequences	Duration (in days)
1	⟨**S**,**M**,**P**,A⟩	500
2	⟨**S**,B,**C**⟩	1,500
3	⟨**S**,D,E,**M**,**P**⟩	500
4	⟨**S**,**C**⟩	1,500
5	⟨**S**,**M**,F,G,**P**⟩	1,000
6	⟨**S**,**M**,**P**,**J**⟩	2,000
7	⟨**S**,**P**,**M**,**P**,**J**⟩	2,000
8	⟨**S**,H,**C**⟩	2,000

TABLE II
Changes

Items	Meanings
S	Hiring
C	Compensation
M	Mission
P	Pricing profile
J	Job
A,B,D,E,F,G,H	Other changes

A (sequential) pattern X is an ordered sequence of changes (e.g., $\langle S, M, P \rangle$). $\mathcal{P}$ denotes the set of all sequential patterns. A sequential pattern $X = \langle X_1, \ldots, X_n \rangle$ is included in a sequential pattern $Y = \langle Y_1, \ldots, Y_m \rangle$, denoted by $X \sqsubseteq Y$, iff there exist n indexes i_k for $k \in \{1, \ldots, n\}$ such that $X_k = Y_{i_k}$ and $i_k < i_{k+1}$ for $k \in \{1, \ldots, n-1\}$. The support of the pattern is the proportion of sequences that contain the pattern: $supp(X) = \frac{|S \in \mathcal{S} | X \sqsubseteq S|}{|\mathcal{S}|}$. A pattern X is frequent if its support exceeds a user-defined threshold. Let α be this threshold, we define $\mathcal{F}$ as the set of frequent patterns: $\mathcal{F} = \{X \in \mathcal{P} | supp(X) \geq \alpha\}$. It is better to restrict ourselves to

TABLE III
Set of changes

Items	Meanings
S	Hiring
C	Compensation
M	Mission
P	Pricing profile
J2	Job
J3	Title
J4	Entity

closed patterns that constitute a lossless representation. Frequent patterns are not numerous but we have still used closed patterns as there is no real benefit in retaining two patterns that have exactly the same support with one included in the other. More precisely, the closure of a pattern X, denoted by $h(X)$, is the largest sequential pattern including X with the same support: $h(X) = \max_{\sqsubseteq}\{Y \sqsupseteq X : supp(Y) = supp(X)\}$. A pattern is closed iff $X = h(X)$. Considering the minimum support threshold $\alpha = 2/8$, the changes in $\{S, C, M, P, J\}$ are the only ones that contribute to the presence of frequent patterns and these changes lead to 4 closed frequent patterns: $\langle S\rangle$ (with $8/8$ as support), $\langle S, M, P\rangle$ $(5/8)$, $\langle S, C\rangle$ $(3/8)$ and $\langle S, M, P, J\rangle$ $(2/8)$. Note that other frequent patterns (e.g., $\langle S, P\rangle$) are not closed.

B. Description of the Kaplan-Meier survival curve

The survival curve is the most commonly used representation to describe the dynamics of the occurrence of death over time. It shows the probability of survival as a function of time. Most survival curves are constructed using the Kaplan-Meier estimator [15] which is a non-parametric method. The Kaplan-Meier estimator, also known as the product-limit estimator, is an estimator for estimating the survival function based on lifetime data. In this method, the observed participation time is divided into time intervals and survival is estimated over each time interval. This estimator takes into account right-censored data. The observed participation time is divided into time intervals, starting at the time t_i when one death occurs and ending just before the next death, and survival is estimated over each time interval, which will give the curve a *stair step* appearance. It will have m steps with m being the number of times that death occurs, not to be confused with the number of deaths (several deaths can happen at the same time). Δ denotes the duration of the employee who stayed the longest in all the population.

The Kaplan-Meier estimator at time t for the individuals that have the pattern X in their sequences, denoted by $\widehat{Sur_X}(t)$, is defined as follows:

$$\widehat{Sur_X}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i^X}{n_i^X}\right)$$

where:

- t_i is the time of the i^{th} event.
- n_i^X is the number of individuals with the pattern X at risk just before the i^{th} event.
- d_i^X is the number of individuals with the pattern X who experience the i^{th} event.

The formula calculates the probability that an individual survives until a given time t, given that they have survived until the previous time t_i and that there have been d_i^X events that occurred at that time. The Kaplan-Meier estimator is often used in medical research to measure the fraction of patients alive for a certain time after treatment, but it is also used in economics and ecology. In this paper, we will use it to plot the survival curve of employees.

Let's visualize the survival curve for the entire population (i.e., $X = \emptyset$) in our toy dataset, assuming that attrition has occurred for each employee. Initially, at time 0, no resignation has taken place, resulting in a survival probability of one on the Kaplan-Meier survival curve until the next resignation event. At time 500, two employees have resigned, resulting in a probability of $\left(1 - \frac{2}{8}\right) = 0.75$ between 500 and the next resignation. Subsequently, one employee quits at time 1000, causing the probability to become $0.75 \times \left(1 - \frac{1}{6}\right) = 0.625$. At time 1500, two additional events occur, reducing the probability to $0.625 \times \left(1 - \frac{2}{5}\right) = 0.375$. Finally, the last employees resign at time 2000, resulting in a survival probability of $0.375 \times \left(1 - \frac{3}{3}\right) = 0$.

Figure 1 represents the visualization of this survival curve.

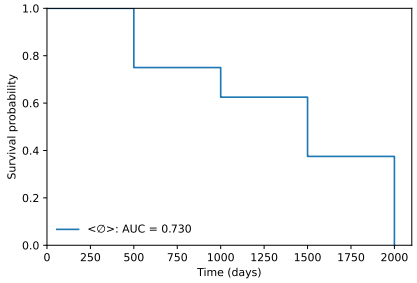

Fig. 1. Survival Curve for the entire population of the toy dataset

IV. Survival-Changing Analysis

A. Overview of the method

In this study, we employ a comprehensive methodology to analyze sequences and the factors that contribute to survival improvements from one sequence to another. Our approach extends traditional frequent pattern mining by benefiting from the Kaplan-Meier estimator. More precisely, we start by identifying a context pattern, which serves as the starting point for the survival analysis. From this context pattern, we aim at finding the simplest suffixes that change the survival rate above a predefined threshold. In this case, the identified suffixes are referred to as *survival-changing patterns* (see Section IV-B). The mining process consists of two steps: extraction of closed frequent sequential patterns and enumeration of pairs formed by a survival-changing pattern and its context (see Section IV-C). This approach allows us to pinpoint the most influential and actionable patterns that have a direct effect on the likelihood of survival within the given context. By emphasizing simplicity and frequency, we can identify meaningful survival-changing patterns that provide valuable insights for making informed decisions and implementing targeted interventions.

B. Interestingness measures for survival analysis

In this section, we introduce the two original interestingness measures of our proposal. The first one evaluates the interest of

a pattern by measuring its area under the survival curve (see Definition 1). The second measure considers the difference between the survival curve of the assessed pattern and that of a contextual pattern (see Definition 2).

Even if there is a loss of information to reduce a curve to a single point, the machine learning community most often uses the AUC statistic for model comparison. The following definition formalizes the notion of *Survival AUC*:

Definition 1 (Survival AUC): The normalized area under the curve of the Kaplan-Meier survival function for a pattern X is defined as below:

$$AUC_{Sur}(X) = \frac{\int \widehat{Sur}_X(t)\,dt}{\Delta} = \frac{\sum_{i=0}^{m-1} \widehat{Sur}_X(t_i) \cdot (t_{i+1} - t_i)}{\Delta}$$

Intuitively, the survival AUC calculates the area under the Kaplan-Meier curve by normalizing it between 0 and 1 using the maximum duration Δ. Overall, the higher the survival AUC for a pattern, the longer the covered population survives. For instance, the survival AUC for the pattern $\langle S \rangle$ corresponds to the AUC of the Kaplan-Meier Curve for the entire population: $AUC_{Sur}(\langle S \rangle) = \frac{1375}{2000} = 0.688$. Similarly, the survival AUC for $\langle S, M, P \rangle$ is slightly worse with 0.60, while the survival AUC for $\langle S, C \rangle$ is slightly better with 0.83. Considering our toy dataset, Figure 2 shows the Kaplan-Meier curves of four closed frequent patterns with their survival AUC. Finally, since a pattern X and its closure $h(X)$ share the same population, they have the same survival AUC: $AUC_{Sur}(X) = AUC_{Sur}(h(X))$. It means that restricting to only closed patterns is sufficient to keep the information complete for all patterns.

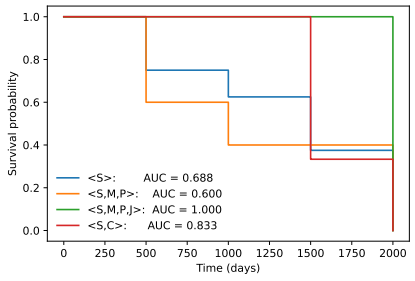

Fig. 2. Survival curves of the closed frequent patterns in the toy dataset

Survival AUC is a relevant interestingness measure to identify whether a pattern increases or decreases survival. In the case of changes that are known to be negative, it is important to be able to remedy them by taking action. In our toy dataset, the changes M and P worsen the situation compared to S with a reduction of $0.60 - 0.687 = -0.087$ of its survival AUC. It would be interesting to know what subsequent changes might remedy this decline. In order to identify such actionable patterns, we propose to measure the difference between the survival AUC of pattern X and that corresponding to a reference pattern, named *contextual pattern* (or context in short):

Definition 2 (Survival gain): Given a contextual pattern C, the survival gain of pattern X corresponds to the difference between the survival AUC of $C \cdot X$ and that of C:

$$Gain_C(X) = AUC_{Sur}(C \cdot X) - AUC_{Sur}(C)$$

It is easy to see that survival gain is a measure whose value is between -1 and 1. A positive (resp. negative) survival gain means that X improves (resp. deteriorates) the survival rate. For instance, as mentioned before, we find that $Gain_{\langle S \rangle}(\langle M, P \rangle) = -0.087$ underling the negative impact of $\langle M, P \rangle$ in the context $\langle S \rangle$. Conversely, pattern $\langle C \rangle$ improves the survival rate as $Gain_{\langle S \rangle}(\langle C \rangle) = AUC_{Sur}(\langle S, C \rangle) - AUC_{Sur}(\langle S \rangle) = 0.833 - 0.687 = 0.146$.

The further the survival gain of a pattern is from zero, the more relevant that pattern is. For this reason, the following definition introduces the notion of *survival-changing pattern* that have a sufficiently strong impact on the survival rate:

Definition 3 (Survival-changing pattern): A pattern X is survival-changing for the context C iff its survival gain in absolute is greater than a user-specified threshold σ: $|Gain_C(X)| > \sigma$.

At first sight, Definition 3 could be reminiscent of sequential association rules [34] where the context C is the body and the survival-changing pattern X is the head. But, the semantics of the survival gain is quite different from traditional measures such as confidence whose objective is to identify a correlation between C and X.

Let us illustrate this definition with our example. The patterns $\langle M, P \rangle$ and $\langle C \rangle$ respectively act as *negative* and *positive* survival-changing patterns within the same context $\langle S \rangle$ for the threshold $\sigma = 0.05$. Interestingly, it is possible that a negative survival-changing pattern is in turn the context of a positive survival-changing pattern. For instance, as we have $Gain_{\langle S,M,P \rangle}(\langle J \rangle) = 1 - 0.687 = 0.313$, the pattern $\langle J \rangle$ is a positive survival-changing pattern for the context $\langle S, M, P \rangle$. In the case of employee attrition, this means that if events $\langle M, P \rangle$ cannot be prevented, then it is relevant to consider $\langle J \rangle$ in order to retain the employee. Such patterns obviously provide very valuable information and the next section details how to extract them.

C. Mining survival-changing patterns

Rather than trying to extract all survival-changing patterns from all possible contexts, we restrict the context either to the empty sequence (i.e., corresponding to the whole population), or to a pattern being itself survival-changing. Indeed, the contexts resulting from survival-changing patterns are those where the end-user can be led to wonder about the next actions to take. For this purpose, we introduce the notion of *relevant* survival-changing pattern:

Definition 4 (Relevant survival-changing pattern): Given a minimum support threshold α and a minimum gain threshold σ, a pattern X is said to be a relevant survival-changing pattern for the context C iff:

- $C \cdot X$ is a closed frequent pattern w.r.t α,
- C is either $\langle\emptyset\rangle$ or equal to $C' \cdot X'$ where X' is a relevant survival-changing pattern for C' and
- X is a survival-changing pattern for C w.r.t σ.

Basically, Definition 4 is a recursive definition where the context of the survival-changing pattern is based itself on a relevant survival-changing pattern. Typically, $\langle M, P\rangle$ is a relevant survival-changing pattern for the context $\langle S\rangle$ and $\langle J\rangle$ is also a relevant survival-changing pattern for the context $\langle S, M, P\rangle$.

From a minimum support threshold α and a minimum gain threshold σ, Algorithm 1 extracts all relevant survival-changing patterns. This algorithm computes the set of closed frequent patterns (line 1). Then, line 2 initializes the initial collection with an empty sequence for both the context and the survival-changing pattern. The main loop (lines 4-6) extracts all the survival-changing patterns, stopping as soon as $\mathcal{L}_i$ is empty. At each iteration of the loop, line 5 extracts the survival-changing patterns before incrementing i. This extraction first builds all the pairs $(C \cdot Y, X)$ by ensuring that Y was previously a survival-changing pattern in $\mathcal{L}_i$ for the context C, $C \cdot Y \cdot X$ is a closed frequent pattern and the gain of pattern X is sufficiently high (with respect to the context $C \cdot Y$). Note that $\mathcal{L}_{i+1}$ only preserve the minimum patterns (e.g., $(C, \langle M, P\rangle)$ will be eliminated by $(C, \langle M\rangle)$). Finally, line 7 returns all the layers of the collection.

Algorithm 1 Survival-changing mining algorithm

Require: A minimum support threshold α and a minimum gain threshold σ

Ensure: The collection $\mathcal{L}$ containing all the pairs (C, X) where X is a survival-changing pattern and C is its context

1: $\mathcal{F} \leftarrow \{X \in \mathcal{P} : supp(X) \geq \alpha \wedge X = h(X)\}$
2: $\mathcal{L}_0 \leftarrow \{(\langle\emptyset\rangle, h(\langle\emptyset\rangle))\}$
3: $i \leftarrow 0$
4: **while** $\mathcal{L}_i \neq \emptyset$ **do**
5: $\quad \mathcal{L}_{i+1} \leftarrow \min_{\subseteq}\{(C \cdot Y, X) \in \mathcal{F} \times \mathcal{P} : (C, Y) \in \mathcal{L}_i \wedge C \cdot Y \cdot X \in \mathcal{F} \wedge |Gain_{C \cdot Y}(X)| > \sigma\}$
6: $\quad i \leftarrow i + 1$
7: **return** $\mathcal{L} \leftarrow \bigcup_i \mathcal{L}_i$

The sequential pattern mining in line 1 has exponential complexity, whereas the complexity of the while loop in lines 4 to 6 is polynomial. By applying Algorithm 1 with a threshold of 0.05, we obtain the following collection of layers :

$\mathcal{L}$=$\{(\langle\emptyset\rangle, \langle S\rangle), (\langle S\rangle, \langle M, P\rangle), (\langle S\rangle, \langle C\rangle), (\langle S, M, P\rangle, \langle J\rangle)\}$

V. Experimentation on our data

This experimental section evaluates the quantity and the type of the relevant survival-changing sequential patterns. After presenting our real-world dataset, we provide answers to the following key questions:

- What is the influence of the frequency threshold, α, on the results of the algorithms?
- What is the influence of the gain threshold, σ, on our findings?
- How to combine both thresholds to obtain the best results in our context of attrition?

Evaluating the effectiveness of our approach is out of the scope of this paper where we focus instead on a real-world dataset (where all experiments were performed within a few minutes).

A. Description of the data

In this work, we have restricted ourselves to employees who are or have been on permanent contracts. In order to study their attrition, we are interested in 4 tables obtained through 3 different data sources from an IT company including two human resources management software. The constitution of our data warehouse has a non-negligible cost because of the different *Extract-Transform-Load* processes to manage while respecting the General Data Protection Regulation (GDPR) which is particularly sensitive to this type of personal data.

The 4 tables used in this paper are *Job History*, *Mission History*, *Pricing Profile History*, and *Compensation History*. Job History contains changes related to the position such as the date of hire, and/or the name of new positions (e.g. Business Analyst, Controller, Product Owner) or the new Business Title. Mission History contains the different missions carried out for the different clients of the company with the start and end dates. Pricing Profile History has the different changes linked to the types of profiles sold to the client (Junior Consultant, Senior Consultant, Expert Director, etc.). Finally, Compensation History contains the salary history. Each change for an employee leads to create a new record in the corresponding table.

Due to some bias in our data (not having resignation date before 2020 in the new software and some other issues) and in order to have the most complete data possible, we decided for the rest of the paper that we would restrict ourselves only to employees who were hired after January 1, 2020. Because of these restrictions, we go from a total of approximately 16,000 employees to only 7,527. Note that our data ends on December 19, 2022.

Table IV succinctly describes these different tables by indicating the number of records, the number of employees still in the company and the number of employees who have resigned.

TABLE IV
DESCRIPTION OF DATASETS (ACTIVE EMPLOYEE DENOTED BY $Active = 1$ / RESIGNED EMPLOYEE DENOTED BY $Active = 0$).

Name	#records	#Employees $Active = 1$	#Employees $Active = 0$
Job History	19,159	4,213	2,281
Mission History	1,483	787	323
Pricing Profil History	7,003	2,216	1,125
Compensation History	8,615	4,763	2,105

Once our data was cleaned and prepared according to the steps presented previously, we proceeded to merge our data

tables. This merge allowed us to consolidate all relevant information into a single sequence of events for each employee. Each row in our merged table represents a change associated with a specific date and employee. By performing this merge, we obtained a total of 7,527 sequences, each with varying duration and length based on the events that have occurred for each employee. The minimum length is 2, the maximum length is 26 and the mean length is 5.25. This sequence representation provides us with a comprehensive view of individual paths within the company and will allow us to analyze patterns and trends in these sequences.

There are a total of 7 changes considered in our analysis: S denotes the hiring event, C represents a compensation change, M signifies a mission change, P indicates a pricing profile change, $J2$ represents a job change, $J3$ denotes a title change, and $J4$ signifies a transfer within the company.

Among all the employees in our population, the maximum duration of employment within the company is 1185 days. Figure 3 is the Kaplan-Meier curve representing the survival curve of all the considered employees.

B. Study of the threshold frequent impact

We initially decided to explore the impact of different support thresholds on the frequency of patterns. To investigate this, we applied the CloSpan algorithms [35] (but any algorithm that extracts sequential closed patterns can be used) where we varied the support threshold in increments of 1% starting from 1%. We then analyzed the results and visualized them using an informative histogram in Figure 4. The histogram illustrates the number of closed frequent patterns discovered for each support threshold. We observed a significant increase in the number of patterns identified as the support threshold is increasing. This finding highlights the sensitivity of pattern discovery to changes in the support frequency threshold, indicating that even small variations can lead to a substantial growth in the number of identified closed frequent patterns. Figure 5 plots the survival curve of the top 5 closed frequent patterns with their corresponding pattern and AUC.

Afterward, we applied our Survival-changing mining algorithm (Algorithm 1) to our dataset, considering various values of α and σ ranging from 0.01 to 0.1 in increments of 0.01. For each combination of α and σ, we obtained a collection of survival-changing patterns. Each collection consisted of multiple layers, where each layer $\mathcal{L}_i$ contained patterns where the prefix is contained in the previous layer $\mathcal{L}_{i-1}$ (except for $\mathcal{L}_0$) and the remainder is a survival-changing pattern. Table V displays the counts of distinct survival-changing patterns discovered by our algorithm, with respect to different values of α and σ.

Table V offers valuable insights into the relationship between the frequency support threshold (α) and the gain threshold (σ) in determining the number of distinct survival-changing patterns. By analyzing the table, we can observe that: as the frequency support threshold decreases and the gain threshold increases, the number of unique survival-changing patterns significantly grows. This finding suggests that imposing more gain in term of survival leads to a broader range of survival-changing patterns being identified.

TABLE V
NUMBER OF DISTINCT SURVIVAL-CHANGING PATTERN VARYING WITH α AND σ

σ / α	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
0.01	39	29	25	22	18	17	16	15	14	13
0.02	48	31	27	24	21	20	19	19	17	16
0.03	65	46	40	37	35	34	30	28	24	23
0.04	80	60	52	45	43	38	34	31	26	23
0.05	105	81	66	57	51	46	40	36	30	26
0.06	151	116	95	81	74	58	50	45	41	35
0.07	186	139	114	97	86	67	58	52	47	41
0.08	232	180	148	123	106	84	73	62	53	43
0.09	290	216	182	141	117	95	82	67	55	40
0.1	401	277	216	158	130	107	80	61	50	35

As expected, when the frequency support threshold is set to a lower value, it allows for the discovery of a larger number of closed frequent patterns. This increased pool of patterns provides more opportunities to uncover diverse survival-changing patterns. Each closed frequent pattern serves as a potential candidate for a survival-changing pattern.

Besides, as the gain threshold is raised, it indicates a higher requirement for improving survival. This leads to the identification of more complex survival-changing patterns. These patterns reflect intricate combinations of events that have a significant impact on survival outcomes. With a higher gain threshold, the algorithm focuses on capturing patterns that contribute to substantial improvements in survival, resulting in a greater diversity of survival-changing patterns.

Therefore, the combination of a lower frequency support threshold and a higher gain threshold offers a favorable setting for discovering diverse and impactful survival-changing patterns. The larger pool of closed frequent patterns facilitates the exploration of different patterns, while the higher gain threshold emphasizes the identification of patterns with significant survival implications. This creates an environment where the algorithm is more likely to identify a variety of unique survival-changing patterns.

In the subsequent subsection of the paper, we decided to compare the results obtained by varying the parameter σ while keeping α at 0.05. We chose this value for α because it corresponds to a support threshold of 376, which is a suitable number for representing employees. Below this threshold, the patterns may not be sufficiently representative.

C. Study of the gain threshold impact

Table VI provides the depth of each layer of the collection returned by our algorithm for different minimum gain thresholds σ when $\alpha = 0.05$.

We can observe that regardless of the variables, the first layer $\mathcal{L}_0$ always contains a single element, here, the closure of the empty set $h(\langle\emptyset\rangle) = \langle S\rangle$ representing the "hire" event that is common to everyone. The number of layers varies depending on the value of σ. The lower the σ value, the

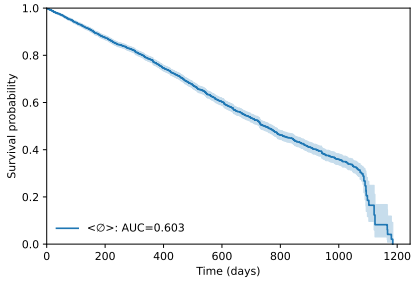

Fig. 3. Survival Curve of all the population

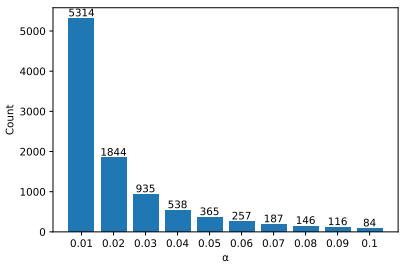

Fig. 4. Number of frequent pattern against α

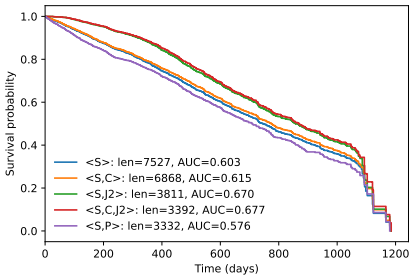

Fig. 5. Survival curve of the top 5 frequent pattern

TABLE VI
DEPTH OF EACH LAYER OF THE CONTEXT/SURVIVAL-CHANGING PATTERN COLLECTION VARYING WITH σ

σ / layer	$\mathcal{L}_0$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_6$	$\mathcal{L}_7$
0.01	1	6	38	99	100	51	4	0
0.02	1	11	66	97	69	12	0	0
0.03	1	25	111	87	7	0	0	0
0.04	1	32	122	50	1	0	0	0
0.05	1	39	127	25	0	0	0	0
0.06	1	57	107	9	0	0	0	0
0.07	1	74	89	0	0	0	0	0
0.08	1	98	51	0	0	0	0	0
0.09	1	112	23	0	0	0	0	0
0.1	1	126	7	0	0	0	0	0

easier it is to increase the AUC to reach the fixed gain threshold, resulting in smaller survival-changing patterns in general. This allows for multiple successive survival-changing patterns within a single frequent pattern, resulting in a larger number of layers. For instance, with σ equal to 0.01, we have 7 layers, whereas with σ equal to 0.1, we only have 3 layers.

Let us now examine an example graphically with $\sigma = 0.01$ and $\sigma = 0.05$. We will take an initial pattern from the last layer of the collection and trace back through the layers to reconstruct the sequence of survival-changing patterns in the correct order. Then, we will display the corresponding AUC curve for each pattern state.

On Figure 6, where the gain threshold is set to 0.01, we observe 7 curves that accurately represent the number of layers in the collection. Since we selected a pattern from $\mathcal{L}_6$, we have 6 successive survival-changing pattern within the chosen pattern. Starting from the sequence $\langle S \rangle$, we eventually reach the sequence $\langle S, C, J2, J3, P, J2, J3 \rangle$ composed of 7 events. Consequently, each survival-changing pattern in this case consists of a single event.

In contrast, on Figure 7 with a gain threshold of 0.05, we have 4 layers, indicating 3 successive survival-changing patterns. Starting from $\langle S \rangle$, we progress towards $\langle S, C, J2, J2, J3 \rangle$ which is also a pattern of length 7. Here, we observe slightly longer survival-changing patterns in general.

This can be attributed to the fact that with a higher gain threshold, a more efficient survival-changing pattern is required. As a result, the survival-changing patterns become more complex.

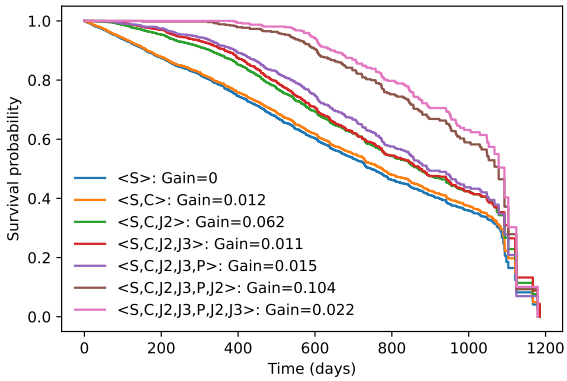

Fig. 6. Survival curves for successive modified patterns with $\sigma = 0.01$

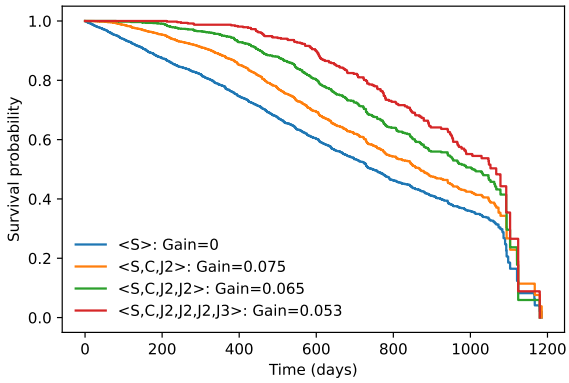

Fig. 7. Survival curves for successive modified patterns with $\sigma = 0.05$

In the first example, where the gain threshold is set at 0.01, the range of gains observed is from 0.012 to 0.104. Consequently, the corresponding survival curves are more or less close from each other, as they are influenced by the magnitude of the gain. Conversely, in the second example with a higher gain threshold of 0.05, the range of gains narrows down to 0.053 to 0.075. As a result, we observe distinct and well-separated survival curves, highlighting the impact of the gain threshold on the shape and characteristics of the curves.

D. Interpretation of results in the context of attrition

The findings of our study hold particular relevance in the context of employee attrition within an organization. When aiming to improve the survival prospects of an individual employee based on their context pattern, it becomes crucial to identify a survival-changing pattern that is not only highly effective in terms of gain but also manageable in terms of complexity. Implementing a multitude of events within a short period of time for a single employee can be challenging, if not impossible such as applying the same event multiple times.

We proceeded to plot boxplots representing the 10 most prevalent survival-changing patterns in each collection of the two previously chosen values of σ. The number at the top of each boxplot represents the number of occurences of the survival-changing pattern in the whole collection of layers (i.e., $|\{C \in \mathcal{P} | (C, X) \in \mathcal{L}\}|$).

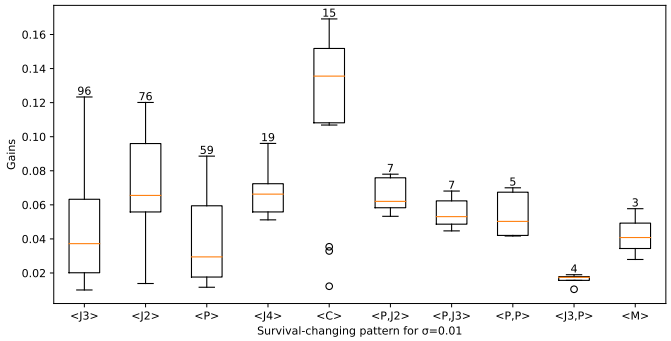

Fig. 8. Boxplots of the top 10 survival-changing pattern when $\sigma = 0.01$

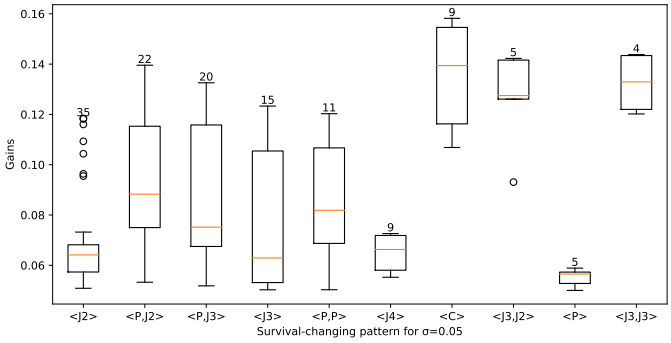

Fig. 9. Boxplots of the top 10 survival-changing pattern when $\sigma = 0.05$

According to Table V, we observe 18 distinct survival-changing patterns for $\sigma = 0.01$ and 51 for $\sigma = 0.05$. However, when examining the numbers displayed at the top of the boxplots of Figures 8 and 9, we can see that beyond the first 5 survival-changing patterns, the remaining patterns have fewer than 10 occurrences in the whole collection of layers. This suggests that, in general, the first 5 patterns are survival-changing patterns for many contexts compared to the others. We can observe that in both figures, the survival-changing pattern with the highest average gain is the pattern $\langle C \rangle$ (change of compensation). We noticed that only the pattern $\langle P \rangle$ (change of pricing profile) occasionally appears as a negative survival-changing pattern when $\sigma = 0.01$ (but not visible in the Figure 8). In this case, it occurs three times, whereas in the remaining $59-3 = 56$ instances, it is a positive survival-changing pattern.

Figure 9 shows that in average the survival-changing patterns in top-10 are slightly more complex, suggesting that achieving a significant gain in survival AUC for a given context may require multiple events rather than a single event. For instance, consider the survival-changing pattern $\langle P, J2 \rangle$ (change of pricing profile and job change), which is the second most occurring survival-changing pattern in the collection. This pattern alone enables at least 22 different context patterns to improve the survival AUC by more than 0.05.

The identified survival-changing patterns provide a strategic approach to addressing employee turnover. By focusing on the most simple patterns with the highest average gain, such as the pattern $\langle C \rangle$, organizations can prioritize interventions or strategies that directly target these events or event sequences. This enables them to make targeted efforts towards improving the survival rate of individual employees.

Moreover, the emphasis on simplicity in our methodology is of great significance. The patterns that exhibit the most significant gains are not excessively complex (thanks to the constraint of minimum line 5 of Algorithm 1), making them more feasible to implement within a short time frame. This consideration aligns with the practical constraints faced by organizations when attempting to improve employee survival. It acknowledges the need for interventions that are efficient, achievable, and realistic within the context of day-to-day operations.

By leveraging the insights gained from our survival-changing patterns, organizations can design tailored interventions and strategies to mitigate attrition risks. These may include adjustments in job responsibilities or work environment, or personalized career development plans. Such interventions, when based on data-driven survival-changing patterns, have the potential to enhance employee satisfaction, engagement, and overall retention within the organization.

It is important to note that the interpretation and application of these results should be done with careful consideration of the specific organizational context and employee characteristics. The identified survival-changing patterns provide valuable guidance, but their implementation should be contextualized and aligned with the organization's culture, resources, and strategic goals. Additionally, monitoring the effectiveness of implemented interventions and continuously refining the approach based on empirical feedback is crucial for maximizing the impact on attrition reduction.

In summary, our study contributes to the understanding of attrition dynamics within organizations by providing a methodology to identify survival-changing patterns based on employee context patterns. These patterns offer insights into event sequences that significantly influence an employee's likelihood of survival. By prioritizing patterns with high gain and manageable complexity, organizations can develop tar-

geted interventions to improve employee retention and reduce attrition rates. However, the application of these findings should be context-specific, considering the unique characteristics and constraints of the organization and its workforce.

VI. Conclusion

In this study, we presented a novel approach for sequence analysis by identifying survival-changing patterns. These sequential patterns highlight events that significantly impact the Kaplan-Meier survival estimator with respect to a given context. By adjusting these thresholds appropriately, we were able to identify a wide range of unique patterns, from simple to complex patterns with significant impact on survival. In the context of employee attrition, our methodology offers practical insights. Using our results, organizations can target effective and manageable survival change motives to improve employee retention prospects. Of course, our methodology could also be applied to other domains where survival analysis makes sense. It would be interesting to generalize our results by applying them to different domains and using larger and more varied datasets.

For future work, it would be promising to expand our analysis by considering not only the time between hiring and quitting, but also the time between individual events within each sequence. This could be accomplished by forming itemsets that group events that are very close to each other. We will look more finely at sequences of specific events and analyze how the temporal proximity (or not) between these events affects survival outcomes. Furthermore, an interesting extension of our work would be to incorporate more specific elements about employees, such as gender, age, or other relevant characteristics, while complying with GDPR. With this additional information, we could examine how these individual factors interact with sequence patterns and influence employee survival. This would provide deeper insights into differences in survival based on these variables, and identify potential mismatches or biases in the context of attrition in a company. By combining fine-grained event timing analysis with specific individual data, we could better understand the complex mechanisms underlying employee career paths, paving the way for more targeted interventions and more equitable human resource management policies.

References

[1] V. B. Adrien Lagouge, Ismaël Ramajo, "La france vit-elle une "grande démission"," DARES : Direction de l'animation de la recherche, des études et des statistiques, Tech. Rep., Oct 2022. [Online]. Available: https://dares.travail-emploi.gouv.fr/publication/la-france-vit-elle-une-grande-demission

[2] G. Morrison, "How long do people stay at their firms?" Consulting Point, Tech. Rep., Mar 2021. [Online]. Available: https://www.consultingpoint.com/market-information/2021/3/29/regional-attrition-and-tenure

[3] G. APICIL, "Ibet – désengagement des salariés : un coût de 14580€/an/salarié," Groupe APICIL, Tech. Rep., June 2019. [Online]. Available: https://www.groupe-apicil.com/newsroom/presse/desengagement-des-salaries/

[4] W. H. Mobley, "Intermediate linkages in the relationship between job satisfaction and employee turnover." *Journal of applied psychology*, vol. 62, no. 2, p. 237, 1977.

[5] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: A reliable approach," in *Proc. of SAI intelligent systems conference*. Springer, 2018, pp. 737–758.

[6] İ. O. Yiğit and H. Shourabizadeh, "An approach for predicting employee churn by using data mining," in *2017 International IDAP*. IEEE, 2017, pp. 1–4.

[7] M. L. Kane-Sellers, "Predictive models of employee voluntary turnover in a north american professional sales force using data-mining analysis," Ph.D. dissertation, Texas University, 2007.

[8] N. Brockett, C. Clarke, M. Berlingerio, and S. Dutta, "A system for analysis and remediation of attrition," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 2016–2019.

[9] J. P. Klein and M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*. Springer, 2003, vol. 1230.

[10] E. T. Lee and J. Wang, *Statistical methods for survival data analysis*. John Wiley & Sons, 2003, vol. 476.

[11] D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas *et al.*, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature medicine*, vol. 8, no. 8, pp. 816–824, 2002.

[12] P. A. Murtaugh, L. D. Burns, and J. Schuster, "Predicting the retention of university students," *Research in higher education*, vol. 40, no. 3, pp. 355–371, 1999.

[13] N. Barbieri, F. Silvestri, and M. Lalmas, "Improving post-click user engagement on native ads via survival analysis," in *Proc. of the 25th International Conference on World Wide Web*, 2016, pp. 761–770.

[14] J. Buckley and I. James, "Linear regression with censored data," *Biometrika*, vol. 66, no. 3, pp. 429–436, 1979.

[15] E.L.Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[16] S. J. Cutler and F. Ederer, "Maximum utilization of the life table method in analyzing survival," *Journal of chronic diseases*, vol. 8, no. 6, pp. 699–712, 1958.

[17] W. Nelson, "Hazard plotting for incomplete failure data," *Journal of Quality Technology*, vol. 1, no. 1, pp. 27–52, 1969.

[18] J. L. Powell, "Estimation of semiparametric models," *Handbook of econometrics*, vol. 4, pp. 2443–2521, 1994.

[19] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[20] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.

[21] M. S. Nawaz, P. Fournier-Viger, M. Z. Nawaz, G. Chen, and Y. Wu, "Malspm: Metamorphic malware behavior analysis and classification using sequential pattern mining," *Computers & Security*, vol. 118, p. 102741, 2022.

[22] A. Verma, A. Taneja, and A. Arora, "Fraud detection and frequent pattern matching in insurance claims using data mining techniques," in *2017 tenth international conference on contemporary computing (IC3)*. IEEE, 2017, pp. 1–7.

[23] C.-M. Hsu, C.-Y. Chen, B.-J. Liu, C.-C. Huang, M.-H. Laio, C.-C. Lin, and T.-L. Wu, "Identification of hot regions in protein-protein interactions by sequential pattern mining," *BMC bioinformatics*, vol. 8, no. 5, pp. 1–15, 2007.

[24] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *Acm Sigkdd Explorations Newsletter*, vol. 1, no. 2, pp. 12–23, 2000.

[25] C. H. Mooney and J. F. Roddick, "Sequential pattern mining–approaches and algorithms," *ACM Computing Surveys (CSUR)*, vol. 45, no. 2, pp. 1–39, 2013.

[26] T. Truong-Chi and P. Fournier-Viger, *A Survey of High Utility Sequential Pattern Mining*. Cham: Springer International Publishing, 2019, pp. 97–129.

[27] G. Ritschard, A. Gabadinho, N. S. Muller, and M. Studer, "Mining event histories: A social science perspective," *International Journal of Data Mining, Modelling and Management*, vol. 1, no. 1, pp. 68–90, 2008.

[28] A. Silva, W. Meira Jr, O. Queiroz, and M. Cherchiglia, "Sequential medical treatment mining for survival analysis," in *SBBD*, 2009, pp. 166–180.

[29] Y. Ren, K. Zhang, and Y. Shi, "Survival prediction from longitudinal health insurance data using graph pattern mining," in *2019 IEEE*

International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019, pp. 1104–1108.

[30] R. Csalódi and J. Abonyi, "Integrated survival analysis and frequent pattern mining for course failure-based prediction of student dropout," *Mathematics*, vol. 9, no. 5, p. 463, 2021.

[31] Y.-S. Wu, D. Taniar, K. Adhinugraha, C.-H. Wang, and T.-W. Pai, "Progression to myocardial infarction short-term death based on interval sequential pattern mining," 2022.

[32] J. B. MATTOS, "A supervised descriptive local pattern mining approach to the discovery of subgroups with exceptional survival behaviour," Master's thesis, Universidade Federal de Pernambuco, 2021.

[33] D. Leman, A. Feelders, and A. Knobbe, "Exceptional model mining," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008*. Springer, 2008, pp. 1–16.

[34] P. Fournier-Viger, R. Nkambou, and V. S.-M. Tseng, "Rulegrowth: mining sequential rules common to several sequences by pattern-growth," in *Proc. of the ACM symposium on applied computing*, 2011, pp. 956–961.

[35] X. Yan, J. Han, and R. Afshar, "Clospan: Mining: Closed sequential patterns in large datasets," in *Proc. of the 2003 SIAM international conference on data mining*. SIAM, 2003, pp. 166–177.