Temporal Mining of Integrated Healthcare Data: Methods, Revealings and Implications

Rui Henriques* Sílvia Moura Pina* Cláudia Antunes*
{rmch,silvia.pina,claudia.antunes}@ist.utl.pt

Abstract

The increasing integration and availability of healthcare data triggers new opportunities for an adequate discovery and use of temporal patterns to support medical decisions. However, adequate data mappings are still lacking for the application of temporal mining methods over healthcare databases. Additionally, existing methods commonly do not allow for flexible knowledge representation to guide time partitioning and to support the use of multiple temporal granularities. Finally, existing predictors are not able to rely on these temporal patterns. In this paper, cyclic rules that integrate multiple medical aspects are discovered using an expressive data mapping. Second, temporal constraints are proposed to guide the discovery of patterns under multiple time scales. Third, a classifier relying on cyclic patterns is defined to predict healthcare conditions. The conducted experiments hold evidence for the utility and efficiency of the proposed methods in characterizing and predicting integrated healthcare profiles.

Keywords: temporal pattern mining, cyclic rules, integrated data, pattern-based classification, time constraints

1 Introduction

The mining of temporal patterns over integrated healthcare databases represents an unprecedented opportunity to support a wide-range of medical and administrative decisions. Predictive tasks are becoming increasingly triggered by the growing amount, quality, temporal range and integration of healthcare data through multi-dimensional structures. Medical predictive tasks aim to classify upcoming healthcare needs for a better planning of resources [7] and for the development of care plans before emergencies occur, preventively increasing health while decreasing costs of care. Our hypothesis is that the knowledge-guided discovery of temporal patterns is critical to support medical decisions.

Although there are data mappings for the ready-application of time-sensitive pattern miners, the majority of these mappings are not able to deal simultaneously with the resultant attribute-multiplicity and with the temporal sparsity of healthcare databases. The mining of healthcare dynamics have been centered on single time series as administrative feature's vectors, physiological signals or genomic-proteomic sequences. Although multiple temporal patterns of interest have been defined for these structures (including calendric rules, motifs, episodes, sequential patterns or partially-ordered tones [21, 20, 19]), they have not been applied over integrated healthcare databases. Additional challenges include the need to deal with different time scales simultaneously [2, 7] and to use temporal constraints for a guided discovery of more interesting patterns [3].

The first contribution of this work is on defining an expressive data mapping for the discovery of temporal patterns in integrated healthcare databases. In particular, we focus on cyclic patterns. Second, we use these patterns to compose cyclic rules and a new time-sensitive classifier. Third, temporal constraints are adopted to guide the target mining methods defining time partitions. Finally, the integrated medical patterns and rules are analyzed, the classification metrics collected, and their implications synthesized.

The document is structured as follows. In *section 2*, the problem of mining cyclic rules over integrated healthcare data is motivated and the critical contributions from related research streams covered. *Section 3* formalizes the target task and describes the proposed solution. Finally, results and key implications are synthesized in *section 4*.

2 Background and Motivation

Time is a critical dimension when extracting knowledge from large-scale healthcare databases. In this context, hidden patterns across the monitored healthcare aspects do not necessarily exist or hold throughout the whole time period covered by the database, but only in some time intervals with recurrent and, potentially, periodic nature. Their discovery can be used to improve care pathways, detect inefficiencies, allocate resources, and study drug reactions and treatment effects.

Temporal pattern mining in healthcare.

Multiple types of temporal patterns have been

*Department of Computer Science and Engineering, Instituto Superior Técnico, Technical University of Lisbon

adopted for healthcare domains, including sequential patterns, episodes, calendric patterns, relations among interval-based events, cyclic rules and motifs [9, 21, 25]. Temporal patterns have been used to mine temporal associations, clinical prediction tasks, clustering of health profiles, among other tasks [7]. Challenges of mining temporal patterns in large repositories of patient records are synthesized in [22] and assessed using summarization techniques. Disease anticipation is addressed in [28] using sequential patterns relying on administrative health records tracking drug prescriptions, hospitalizations, and daily hospital activities. The combination of administrative data with omic data is pointed as a decisive step to characterize disease progression [7].

Minimal sets of temporal patterns have been proposed to predict the risk of developing thrombocytopenia [5]. State-based characterization using Markov models was, among others, applied to predict the risk of stroke in sickle cell anemia [27]. Approaches relying on temporal abstractions have been largely used for both physiological signals and coarse-grained medical data [30]. Time-annotated sequences was proposed to extend sequential patterns, where precedences between events is annotated with a duration, to describe the follow-up of liver transplantations [8]. Bellazzi et al. [7] provide an additional wide-set of temporal data mining applications, with an heightened incidence on physiological signals and molecular data.

On the need to mine integrated healthcare data.

In the last decade, new integrated patient-centric data sources emerged. Countries as United Kingdom and Netherlands, already track patients' movements across health providers, payors and suppliers. The changing landscape has been shaped by: *i)* consumer-pushed demand through direct-access to risk and diagnosis information outside of the hospital setting, *ii)* new requirements for drug and treatment development, *iii)* the support of medical decisions for quality compliance, and *iv)* remote home monitoring. Databases are increasingly less fragmented, with appearing both cross-country and cross-player offerings, as provided by Cegedim and IMS[1]. Health records are commonly the means to organize the wide variety of health-related events (as laboratory results, prescriptions, treatments or diagnostics) into a single and compact fact [15].

[1]Other relevant sources include databases derived from claims (Ingenix, D2Hawkeye, CMS), e-health records (McKesson, GE, PracticeFusion), imported health records (GoogleHealth, HealthVault), content aggregators (Walters Kluer, Reed Elsevier, Thomson), patient communities (Alere, Pharos, SilverLink, WebMD, HealthBoards), consumer reports (Anthem, vimo, hospitalcompare), online worksite healthcare (iTrax, webConsult), and physician portals (Medstory, Sermo, Doctors.net.uk).

However, there is still a lack of methods to mine relevant temporal patterns over these databases. Simplistic data mappings require the denormalization of these structures into tabular datasets, which often results in the loss of temporal distances. Alternatively, the discovery of temporal patterns has been mainly proposed over single temporal structures. To deal with multiple temporal structures derived from integrated healthcare databases, some simplifications have been proposed. One direction has been to apply temporal miners to each attribute independently as in [31]. The drawback of this solution is the loss of critical integrated views that does not show up when each attribute is analyzed separately. A second direction is to perform feature selection over sets of events with the goal of loosing both the dimensionality and temporality, as in [6]. Different options, such as the use of feature vectors [18], clustering techniques [14] and generative models [4] can be considered under this goal. A final direction is to use rule generation approaches from sets of events [20].

Shortcomings of temporal pattern mining methods.

The existing temporal pattern mining methods suffer from two critical drawbacks. First, they are not able to effectively deal with different time granularities simultaneously. Second, the few methods able to deal with multiple attributes (transactional data) are limited by non-robust and inflexible representations of time [9]. The work on mining temporal patterns on transactional sequences has been centered on the analysis of precedences, thus neglecting temporal distances with the exception of few gap-based methods [2].

Temporal constraints to guide mining tasks.

Sequential constraints have been captured through regular languages, context-free grammars and acyclic graphs [12, 19]. Taxonomic and temporal constraints were proposed between items in sequences [29], and included during the pruning of large pattern-based structures [26]. Calendric constraints have been also proposed [24]. To avoid the blocking of novel patterns, a hierarchy of constraint relaxations for itemset sequences ranging from conservative to distance-based approximations is introduced in [2]. D2PM framework [3] uses an ontology to impose flexible constraints captured through taxonomical, relational and compositional constraints deep into the mining process. A language for the specification of periodic temporal patterns with non-strict time boundaries is proposed in [10].

Temporal constraints are needed to support the incorporation of time regions and scales of interest. Background knowledge is, thus, of critical value to mine integrated healthcare data as it guides the definition of time windows; provides methods to bridge multiple time granularities and to remove spurious correla-

tions; and allows for incrementally improved results by refining the way domain-knowledge is represented. Understandably, mining methods cannot rely on combinatorial options to compose time windows when dealing with large-scale databases. Currently, time partitioning strategies include clustering, pseudo-items, fuzzy-characterization, split-based sequential-trees, or symbolic-interleaving [20, 19]. However, there is the need for a more flexible constraining of temporal regions and scales of interest to avoid a bias towards regions driven by spurious patterns.

On temporal pattern-based classification.

Few works have proposed the use of temporal patterns to assist classification and prediction tasks. A rule-based classifier using time-interval patterns [25] and alternative predictors based on static patterns [7] are examples. When targeting transactional sequences, as we do in this work, sequential patterns are generally the target temporal patterns. These patterns have been combined with naive Bayesian classifiers [16] and decision trees [13], or simply ranked with scores that identify the patterns more able to discriminate a specific class. However, such methods are not able to capture temporal distances. To the best of our knowledge there are not yet attempts on the combination of transactional and time-sensitive patterns to perform classification. Nevertheless, existing directions provide critical principles for their definition.

3 Methods

Having motivated the need for knowledge-guided temporal mining methods over integrated healthcare data, in this section we incrementally propose methods to answer this problem.

3.1 The Proposed Data Mapping

DEFINITION 3.1. *Let Σ be an alphabet of symbols, and θ be a timestamp. A* **time sequence** *$w \in \mathbb{T}^{n,p}$ is an ordered multi-set of* events *(σ_i, θ_i):*

$$\{(\sigma_i, \theta_i) | \sigma_i = \{\sigma_{i1}, .., \sigma_{ip}\}, \sigma_{ij} \in \Sigma_j, \theta_{i+1} > \theta_i; i = 1, .., n\},$$

where $n \in \mathbb{N}$ is its length, $p \in \mathbb{N}$ its multivariate order*, and $\mathbb{T}^{n,p}$ is the set of all time sequences.*

A time series is a time sequence where the occurring events are temporally equally distant, not allowing for co-occurrences and sparsity. Exemplifying, $y = \{(0, \tau), (3\text{-}5, 2\tau), (>5, 3\tau), (2, 4\tau)\} \in \mathbb{T}^{4,1}$ is an univariate time series, while $y = \{([2\ 21], 2\tau), ([3\ 19], 5\tau), ([3\ 19], 5\tau)\} \in \mathbb{T}^{6,2}$ is a multivariate time sequence of $p = 2$ order. Common time sequences with $p > 1$ order include lab tests or bedside measurements.

DEFINITION 3.2. *Let an item σ be an element from an alphabet Σ. An* **itemset** *I is an orderedset of items. A* **transaction** *or* event *is a tuple $e = (I, t)$, where $e.I$ is an itemset and $e.t$ a timestamp.*

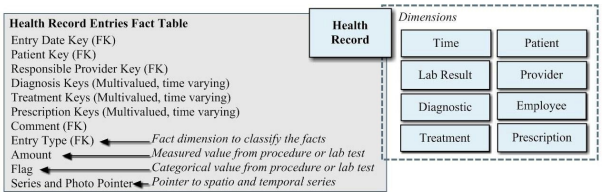


Figure 1: Health record-centered multi-dimensional structure

DEFINITION 3.3. *A transactional sequence or* **itemset sequence**, *$s \in \mathbb{I}$, is an ordered set of itemsets $< e_1.I, .., e_n.I >$, whose timestamps respect: $\forall_{i \in \mathbb{N}} 1 \leq i < n \Rightarrow e_i.t < e_{i+1}.t$, where $n = |s|$ is the sequence length.*

The proposed data mapping method combines multiple time sequences derived from multi-dimensional structures into a single temporal structure, an itemset sequence. This can be viewed as an integrative solution at the input level.

For this mapping, we assume that the target databases rely on health records to organize the wide variety of episodes into a set of compact facts. In Fig.1, an illustrative health record is presented. Beyond the relatively structured diagnosis and treatment dimensions (with contents mandated by the insurance industry and governments), other key dimensions shared by healthcare providers can include the calendar date, patient identity, payer, provider, prescription and location.

Under the presence of similar databases, six critical steps compose the proposed data mapping. First, the split dimension, commonly the patient dimension, is used to group instances from fact occurrences.

Second, health records commonly define what the fact represents and the type of its fields to deal with their large number of entries. Therefore, a denormalization is needed to compose each time sequence.

Third, amounts are discretized into intervals (seen as ordinal symbols), free-text is ignored, and complex data is converted into categorical sets of symbols of fixed p length (to be mapped as time sequences of p-order).

Fourth, the sets of co-occurring measures from each fact are mapped into a multivariate time sequence $\mathbb{T}^{n,p}$ using the time dimension.

Fifth, conflicts between the domains of the considered time sequences are removed (by potentially replacing symbols), and their dimensionality ($|\Sigma|$) balanced by aggregating ordinally related symbols.

Finally, the multiple time sequences are mapped into an itemset sequence by performing the union of their revised domains. An illustration is given in Fig.2.

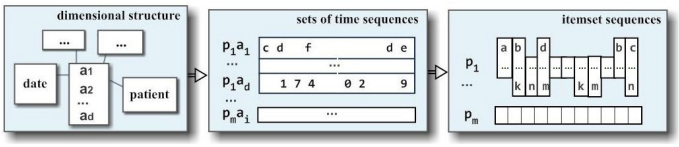


Figure 2: Mapping integrated databases as an itemset sequence

Under an itemset sequential-based formulation, existing temporal mining principles to deal with such structures can be adopted to support the definition of the target solutions.

3.2 Temporal Pattern Mining

To guide the discovery of temporal patterns, we propose methods built upon previous work done in the D2PM framework. The D2PM framework [3] has the goal of supporting pattern mining tasks with the use of domain knowledge represented through an ontology and a set of predefined constraints. An ontology has the expressive power to allow for different time representations, being an explicit specification of a domain. This framework also encompasses a full range of mining methods that receive the ontology and its constraints as the input, and return a set of patterns that satisfy the specified constraints as output. We chose this framework since it seamlessly integrates pattern mining methods with the use of domain knowledge, and due to its extensibility potential, thus allowing the introduction of flexible types of constraints and algorithms.

A constraint has been defined as a predicate $c : 2^{\Sigma} \rightarrow \{true, false\}$. An itemset I satisfies c if $c(I)$ is true. In the context of D2PM, this notion is revised.

Definition 3.4. *A* constraint *is a tuple* $C=(\theta, \mu, \psi, \varphi)$ *where* θ *is the minimum support threshold,* μ *is a mapping function,* ψ *is a predicate called the equivalence function that defines which items contribute for the same support, and* φ *defines the acceptance function. A* **temporal constraint** *is a constraint where* μ *maps a timestamp to a user-specified time granularity, and the acceptance function* φ *works as a filter that eliminates all the transactions that do not fit into the time interval and granularity in consideration.*

The proposed temporal constraints are similar to the ones defined in [2], in order to include temporal criteria from a temporal ontology. The time ontology used in this work is the Reusable Time Ontology [32], based on the notion of time line and centered on Time Point and Time Interval classes. Concepts as Convex Time Interval, which consists of a connected interval on the time line, and Non Convex Time Interval, corresponding to non-connected time intervals as periodically occurring

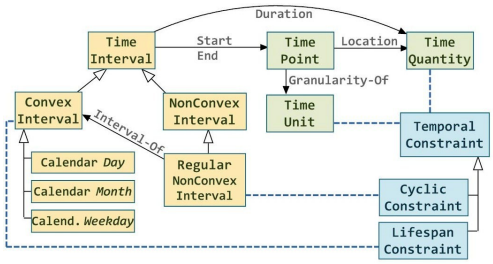


Figure 3: Temporal constraints in the reusable time ontology

events, are allowed. A Time Quantity is represented by a real number and a Time Unit defines the level of granularity (year, month, day, and so on). A simplified representation of this ontology is illustrated in Fig.3.

Introducing a temporal constraint instantiates multiple temporal concepts in the adopted ontology. For example, if one wants to consider only "transactions taking place on the 4th June of 2012", the constraint immediately instantiates "4th", "Monday" and "June". This facilitates the access to different time granularities simultaneously. In this work, we select one specific type of temporal constraints, cyclic constraints, which define a periodicity for an itemset support through the use of a regular non-convex time interval.

Definition 3.5. *Given an itemset sequence database* D*, the* **coverage** *of an itemset* I *with respect to a temporal region of acceptance* φ*,* $\Phi_{\varphi}(I)$*, is the set of events* e *in* D *where* $e.t \subseteq \varphi$ *in which the itemset* I *occurs* $(I \subseteq e.I)$*. The* **support** *of an itemset* I *in* D *respecting* φ*, denoted* $sup_{\varphi}(I)$ *is its coverage size* $|\Phi_{\varphi}(I)|$*.*

When considering simple partitions as the convex interval $\varphi=[t_i, t_f]$, the inclusion of events on its coverage, $\Phi_{[t_i,t_f]}(I)$, should satisfy $I \subseteq e.I \wedge t_i \leq e.t \leq t_f$.

Definition 3.6. *Given an itemset sequence database* D*, a* **temporal association rule** R *with respect to a temporal partition* φ *is defined as* $A \Rightarrow_{\varphi} B$*, where* A *and* B *are itemsets that occur in the given time partition, and* $A \cap B = \emptyset$*. The* support *of a rule is given by* $sup_{\varphi}(A \Rightarrow B) = sup_{\varphi}(A \cup B)$*, the* confidence *of a rule is given by* $conf_{\varphi}(A \Rightarrow B) = \frac{sup_{\varphi}(A \cup B)}{sup_{\varphi}(A)}$*.*

Definition 3.7. *Given a dataset* $D = \{x_1, ...x_m\}$*, where each instance is an itemset sequence,* $x_i \in \mathbb{I}$*, and a set of temporal constraints* C_i *with a minimum support threshold* θ_i *and temporal regions of acceptance* φ_i*, the* target temporal pattern mining task *aims to discover the set of patterns that occur frequently in* D *for the constrained temporal regions:*

$\{I \mid \forall_i sup_{\varphi_i}(I) > \theta_i\}$

Using these formulations and the proposed data mappings, the research problem can now be decomposed into two subproblems. First, to extend existing methods that work with itemset sequences to deal with time-stamped transactional data, as they commonly only account for precedences. Second, to adapt existing methods to allow for the inclusion of the previously defined temporal constraints to guide the mining process.

We propose an adaptation of the Interleaved method to deal with time and to incorporate time constraints deep into the mining process. In particular, we focus on cyclic constraints.

Since we want the resulting method to be extensible so further constraints can be included at a later time, we propose a separation between the algorithm behavior and constraint instantiation process. We refer the resulting method as (TD)^{2}Interleaved, standing for Transactional Temporal Domain-Driven Interleaved.

DEFINITION 3.8. *A* time partition φ_i *is the set of transactions that take place in* $[i \times t, (i+1)t]$ *region, where* t *is the considered temporal granularity defined according to a constraint-based* μ *mapping function.*

DEFINITION 3.9. *Let a cycle be a tuple (p-period, k-offset) that defines time intervals that start at the* k *time point and that repeats* p *after* p *time points. Given a minimum support threshold* θ*, a* **cyclic pattern** *is defined as* $I(p,k)$*, where* I *is an itemset, and* (p,k) *is a cycle defining regular non-convex intervals where for every temporal partition* φ_i*,* $\forall_{i \in \mathbb{N}} : \varphi_i = [ip+k, ip+k+1]$*, the pattern has* $sup_{\varphi_i}(I) > \theta$*.*

An illustrative cycle is $I(p = 3, o = 1)$, meaning that the $\sigma \in I$ items co-occur periodically from 3 to 3 time units starting on the 2^{nd} partition of the dataset.

DEFINITION 3.10. *Given a minimum support* θ_1 *and confidence* θ_2 *thresholds, a* **cyclic rule** *is defined as* $A \Rightarrow B(p,k)$*, where* A *and* B *are non-overlapping itemsets, and* (p,k) *defines periodic non-convex intervals* φ_i*, where* $sup_{\varphi_i}(A \Rightarrow B) > \theta_1$ *and* $conf_{\varphi_i}(A \Rightarrow B) > \theta_2$*.*

For example, for a given granularity of 2 hours, the cyclic rule highGlucoseLevels⇒insuline(24h,7am) is discovered if the rule displays levels of support and confidence above minimum thresholds repeatedly for the 7-9am period every 24 hours.

The proposed (TD)^{2}Interleaved is described in three stages. The first stage, the preprocessing, receives the integrated database, performs the data mapping steps described previously, and converts strings into contiguous symbolic values for an efficient mining.

The second stage is responsible for the discovery of frequent rules in two steps. First, the pre-processed dataset is partitioned into multiple time partitions according to the given temporal constraints[2]. To consider different levels of granularity, we recur to the mapping function μ (see Def.3.4). Finally, for an efficient grouping of events, since μ operation is only defined in the Reusable Time Ontology for time points, we generalize the notion of TimePoint to be optionally seen as a TimeInterval with the finest duration with regards to a specified granularity.

Second, the Interleaved algorithm is applied over these partitions. In Ozden et al.'s work [23], the problem of finding cyclic association rules is solved over an already partitioned dataset using convex intervals $\{\varphi_0, .., \varphi_{n-1}\}$. The pseudocode for the proposed adaptation of the interleaved variant able to deal with constraints is describe in Alg.1.

Input: dataset of itemset sequences with partitions <s_1..s_n>
Input: Lmin>0, Lmax>0 *//min and max p-period of the cycles*
Input: θ_1 */*min support*/* θ_2 */*min confidence*/*
Output: All cyclic association rules $R(p,k)$:
$\{R \mid \forall_{L_{min} \leq j \leq L_{max}} \forall_{i=jp+k} sup_{s_i}(R) > \theta_1, conf_{s_i}(R) > \theta_2\}$
ForEach partition s_i
Run Apriori to find all rules R with $sup(R) > \theta_1 \wedge conf(R) > \theta_2$
ForEach rule R: compute bitmap *//1 in* i^{th} *pos if* R *exists in* s_i
$n = 1$: cycles are assumed to exist for each single itemset
ForEach $n > 1$ *//cycle detection*
Generate cycles for n-itemsets from $(n\text{-}1)^{th}$ cycles *//pruning*
ForEach time unit t_i
Collect n-itemsets from candidates in s_i *//skipping*
Remove cycles containing n-itemset I if $sup_{s_i}(I) < \theta_1$

Algorithm 1: Interleaved algorithm

For each partition φ_i, an algorithm as Apriori [1] is applied to find all the frequent temporal association rules $X \Rightarrow_{\varphi_i} Y$. For each association rule, a bit sequence of size n is created (for example 0010101), where a 0 in position i means that the rule does not stand in φ_i and a 1 states that the rule stands in φ_i. Finally, the algorithm performs cycle detection relying on three properties: cycle skipping, cycle pruning and cycle elimination.

Finally, the third stage takes as input the cyclic patterns obtained in the previous stage, and translates them according to the temporal ontology in a meaningful way for the user.

3.3 Discriminative Temporal Patterns

The use of temporal patterns and rules is critical to classify health states and to predict upcoming medical

[2]Each timestamp corresponds to an instance of TimePoint. To compute partitions we need to consider the TimeInterval class, which includes sets of two or more TimePoints. Additionally, when the time intervals do not correspond to connected intervals, they are instanced as members of the NonConvexTimeInterval class, as the previously introduced RegularNonConvexTimeInterval constraints (e.g. "every Friday in December").

conditions as the need for a surgery. For this we propose a method to mine discriminative sets of patterns for a specific condition, which receives labeled instances as input, mines temporal patterns by applying the previously introduced methods, and delivers a discriminative model as output.

First, the complete set of cyclic patterns is generated for each medical condition and their confidence evaluated to compose a new type of rules of the form $I \Rightarrow c_i$, where I is a cyclic pattern and c_i is the class. Second, and similarly to CMAR [17], the cyclic-based rules are inserted in a tree structure if: *i)* the χ^2 test over the rule is above an user specified α significance level, and if *ii)* the tree does not contain a rule with higher priority. A rule $R_1 : I_1 \Rightarrow c$ is said to have priority over $R_2 : I_2 \Rightarrow c$ if $I_1 \supseteq I_2$or if:

$$conf(R_1) > conf(R_2) \vee (conf(R_1){=}conf(R_2) \wedge sup(R_1) > sup(R_2)) \vee (conf(R_1){=}conf(R_2) \wedge sup(R_1){=}sup(R_2) \wedge |I_1|{<}|I_2|)$$

Finally, the tree is pruned using the rules' priority[3].

In addition to this discriminative method, we propose a classifier, C^{2}P (Classification from Cyclic Patterns), that relies on a discriminative set of patterns for each condition/class to compose the learning model – (pattern,weight,class). The great challenge for this classifier is the testing step. Cyclic patterns emerge from a population of patients, and, therefore, one should not expect to find discriminative cyclic patterns within a given patient. The need for periodic care interventions is not a common individual profile but an emerging collective profile. For this reason, we consider a weighting criteria, cycles matching score, per patient's pattern proportional to the number of matching occurrences against discriminative cyclic patterns.

Given a specific testing patient, for all the discriminative cyclic patterns we use the learner to determine the instantiated patterns and their cycles matching score. The strength of each group of conditions is calculated using a Weighted Chi Squared (WCS)[4] measure as in [17], which is finally combined with the cycles matching score.

The strongest condition is outputted as the estimated class if we want a deterministic classifier, otherwise the computed strength for each class can be seen as its probabilistic value.

[3]Alternative scoring methods include probabilist induction [31] and optimization metrics based on confusion matrices [11]

[4]$WCS = \Sigma_I(\chi^2(I) \times \chi^2(I))/MCS(I)$, where MCS is the Maximum Chi Squared: $MCS(I \Rightarrow c) = (min(sup(I), sup(c)) - sup(I)sup(c)/N)^2 \times N \times e$, where N is the number of testing instances and $e = 1/sup(A)sup(c) + 1/sup(A)N - sup(c) + 1/N - sup(A)sup(c) + 1/(N - sup(A))(N - sup(c))$

4 Results

To assess the performance of the proposed solution, we first describe the properties of the target healthcare database. Second, we present key observations from the collected cyclic patterns, and perform an experimental evaluation of the proposed method using results from pattern-based classification against medical conditions. Finally, their implications are synthesized.

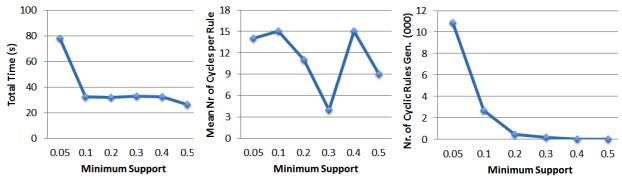
4.1 Target Dataset

To evaluate the proposed solutions, a database that integrates healthcare data across hospitals, clinics, pharmacies and laboratories was adopted. This database monitors several aspects per patient across multiple specialties. According to the proposed data mapping, such scheme was denormalized using the patient dimension, multiple time sequences with discretized domains were defined using the time dimension, and the target time-sensitive itemset sequence per patient composed. For simplicity, in this section we only present the integrated observations for the available records of the following time sequences: health condition, provider, illness severity index, specialty and required procedures.

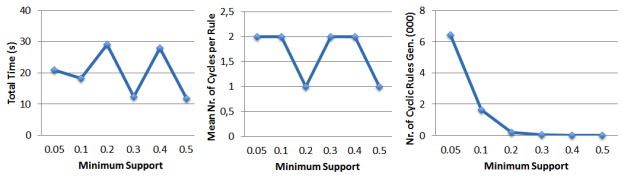
4.2 Temporal Patterns and Rules

The proposed (TD)^{2}Interleaved method was run with constraint-guided varying levels of granularity (month, trimester and semester), support and confidence. Each granularity determines the number of time partitions to be considered in the data. From the finer (month) to the coarser (semester) granularity, there is a corresponding increase in the mean number of transactions assigned to each partition. For month, quarter and semester scale, we have 36, 13 and 6 partitions with a mean number of transactions of 2161, 5985, and 12967, respectively.

Fig.4 shows the total running time, the number of cycles per pattern and the total number patterns derived from the application of the (TD)^{2}Interleaved algorithm using alternative support levels and two illustrative granularities – month and quarter – for a fix confidence level (0.75). Three observations can be retrieved. First, when comparing the running time across the several granularities, we observe that the performance degrades for finer granularities. This means that the complexity of this task is more affected by the cycles computation (inter-transactional) than by the discovery of rules within a given time partition (intra-transactional). Second, the number of cycles per rule was found to be higher for finer granularities and, understandably, to decrease with an increasing support. For finer partitions the number of options for the p cycle period is higher, which, when combined with alternative offsets, gives rise to this observation. Finally, the number of cyclic

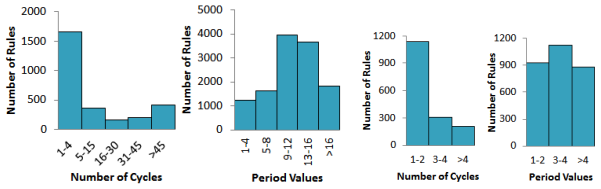


(a) Month granularity



(b) Quarter granularity

Figure 4: Running time, number of cycles per rule, and number of rules for (TD)^{2}Interleaved algorithm over an integrated healthcare database with ∼80.000 transactions



(a) Month granularity (b) Quarter granularity

Figure 5: Grouping of rules per number of cycles and per period values for a minimum support of 0.1 and confidence of 0.8

rules strongly increases for lower support levels, and become hardly available for supports thresholds above 10%. This is a consequence of the nature of integrated healthcare databases, where a cyclic medical condition of interest is commonly related with a specific specialty and, therefore, has a significantly low support.

Fig.5 characterizes the discovered rules based on their number of cycles (excluding different offsets) and periodicity (including different offsets). Two key observations are retrieved. First, for a month granularity, there is a large portion of rules with a high number of cycles. These are rules that sustain levels of support for a high number of partitions. Lifespan rules can be adopted to remove rules with such behavior. Second, periodicity values do not only coincide with trivial cycles (as yearly or semester periodicities), but capture other ranges with potential healthcare significance.

Table 1 shows a small set of illustrative frequent patterns and rules retrieved using the target database[5]. The first cyclic pattern, P1, informs that a combined visit of two specialties A=1 and A=2 by patients

ID	Cycles	Granularity
P1	{A=1,A=2,D=3}(p=12,o=0)(p=12,o=11)	Month
P2	{A=5,B=8,B=2,C=1}(p=2,o=0)(p=4,o=1)	Quarter
R1	{A=2,D=2}⇒{A=7,B=3}(p=6,o=3)(p=6,o=4)	Month
R2	{B=5}⇒{C=7}(p=4,o=1)	Quarter

Table 1: Illustrative sanitized set of cyclic patterns and rules obtained from the target database

Rule	Conf.	Score
{B=3,C=1,D=3}(p=12,o=11)⇒Surgery	92%	85
{B=2,D=2}(p=2,o=0)(p=4,o=1)⇒Surgery	90%	84
{B=4,D=0}(p=2,o=0)(p=2,o=1)⇒NoSurgery	72%	82
{B=5,B=1}(p=4,o=3)⇒Surgery	89%	81

Table 2: Illustrative discriminative rules to predict a condition

with medium-to-high severity indexes[6] (D=3) appears cyclically in the 1^{st} (period=12, offset=0) and 11^{th} month (period=12, offset=10) of the year.

The R2 cyclic rule informs that the need for a specific procedure C=7 is implied for patients with respiratory defficiencies (B=5) with heightened incidence on the second quarter of every year. By disclosing the support of a pattern across the time scale we can test its overall strength. This rule has the following quarters support (three years) – 37, *120*, 30, 32, 35, *132*, 28, 29, 39, *131*, 41 and 28 – meaning that there is clearly stronger than alternative cycles for the same rule.

The use of these temporal patterns and rules motivate the importance of retrieving temporal rules using the proposed data mapping. Note that cyclic rules are not illustrated to convey how healthcare value systems can benefit from their discovery, but, rather, as an illustrative case of a temporal rule that sustains integrated healthcare profiles.

4.3 Pattern-based Classification

Multiple temporal medical conditions as the need for a specific treatment or monitoring can be subjected to characterization or prediction. An illustrative discriminative set of cycle-based rules that support the characterization of the need for a surgery on the last quarter is depicted in Table 2. This table presents how different cyclic patterns induct with varying levels of strength the class under assessment.

Additionally, Fig.6 illustrates key classification metrics when predicting the need for a specific procedure on the last quarter. Note that there are not yet classifiers able to use the target databases with whom we can establish comparisons. Comparison against tabular classifiers was not included as a simple denormalization of events per patient would result in a database

[5]Results were sanitized to protect data confidentiality issues

[6]Indexes were discretized and balanced (section 3.1)

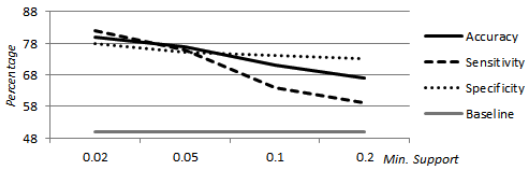


Figure 6: Accuracy, sensitivity and specificity of the target classifier to predict the need for a surgery on the last quarter

with more than 1000 attributes[7] , which turns the learning task impracticable and leads to accuracy rates near 50% (similarly to a random classifier). Additionally, sequence classifiers are only able to receive a single time series as input, which is a structurally different task.

To compute the metrics for this prediction task, the data from the last quarter was removed and the classes per patient were computed using the medical condition. We adopted a 10-fold cross-validation. Three observations can be made. First, performance improves when classification relies on cyclic patterns mined under lower thresholds. This results from the fact that generally the patterns with lower support generate classification rules with higher confidence, since spurious patterns (as co-informative specialties and procedures) tend to no longer compose the majority of patterns.

Second, the observed sensitivity and specificity levels are interestingly similar. This is because the positive condition (having surgery) is related with the set of rules with higher confidence (since it relies on a wider set of very specific patterns). Despite this behavior favors true positives, such unbalanced discriminative set of rules turns also the classifier more susceptible to Type II errors (false negatives).

Finally, accuracy rates seem to have space for improvement. That is, we hypothesize that the accuracy levels not only result from the natural unpredictability of health profiles or from the task complexity. A core reason is the fact that cyclic patterns emerge for a population, and, therefore, can hardly be fully find within a given patient. Although, within the testing stage, we use a weighting criteria to consider patterns that do not necessarily repeat for a patient under evaluation[8], the isolated use of cyclic patterns still provides an incomplete temporal picture. For this, the complementary use of temporal rules (beyond cyclic) for prediction tasks is a promising direction.

4.4 Implications

Methodological Implications

Common integrated healthcare profiles of interest are only shared by a small portion of the overall target population. For this reason, low support and high confidence thresholds should be adopted to find temporal rules of interest. Additionally, constraints relying on equivalence functions (Def.3.4) should be complementary defined to remove spurious patterns as multiple temporal patterns (as periodicities) that derive from a single pattern (as a lifespan-holding pattern).

Two implications result from the efficiency observations. First, when mining temporal rules from large-scale healthcare databases, it is critical to bound their lifespan to avoid a high number of partitions. This can be done recurring to sliding time windows. Second, an exploitation of multiple granularities should start from coarser to finer scales, with the patterns obtained for coarser time scales potentially adopted to compose ψ and φ functions (Def.3.4) that guide the search space over finer granularities. This can be done by extending monotonicity principles across time scales.

Finally, the high number of temporal patterns and of their internal configurations (as multiple periodicities and offsets for cyclic patterns) claim for an extension of existing condensed representations to include the time dimension. This results in reduced discriminative sets of patterns more prone to medical validation.

Healthcare Implications

As individuals move across locations, specialties and care providers, a coherent historical picture becomes available. At a global level, emerging temporal patterns, as periodic patterns, are critical for planning resources and programs, or to discriminate integrated profiles related with a medical condition (e.g. to assess the impact of procedures and prescriptions in upcoming health states). Temporal patterns can also be use to support personalized medical decisions. Here, temporal patterns as periodicities that only gain significance for a population should be complemented with additional temporal patterns to maximize predictive performance.

In the conducted experiment, integrated cyclic rules were discovered using multiple coarse-grained temporal granularities. However, their discovery using finer time scales are also allowed by the proposed method, through the simple inclusion of new temporal constraints. This is increasingly required for hourly and daily analysis of integrated healthcare data collected from tele-wearables as remote glucose monitors, pacemakers, spiro-meter and smart-shirts within mobile-health programs.

5 Discussion

This work motivates the task of mining temporal patterns in healthcare databases using temporal constraints, and synthesizes the limitations and contributions of related research streams to answer the problem.

[7]Mean of 10 occurrences for 5 time sequences over 36 months

[8]Approximates the task to traditional rule-based classification

Cyclic rules are selected to illustrate how temporal patterns can integrate multiple properties when mined over time-sensitive itemset sequences.

The key contributions of this work are: *i)* the definition of an expressive data mapping for the discovery of temporal patterns in integrated healthcare databases, *ii)* the proposal of a method for the inclusion of expressive temporal constraints to support the pattern discovery under multiple time scales, *iii)* the extension of cyclic-rules mining methods to deal with timestamped transactional data, and *iv)* the definition of a new classifier that uses cyclic patterns to characterize, discriminate and predict healthcare conditions.

The conducted experiments hold evidence for the utility and efficiency of mining cyclic rules using non-convex time constraints. Additionally, the observed performance for the selected prediction task opens a door for the inclusion of new temporal patterns to anticipate medical conditions from integrated healthcare profiles. This extension is simple as the proposed data mapping, time-partitioning guidance, and pattern-ranking scheme are independent from the underlying type of patterns.

Acknowledgment

This work was partially supported by *Fundação para a Ciência e Tecnologia* under the research project D2PM (PTDC/EIA-EIA/110074/2009) and the PhD grant SFRH/BD/75924/2011.

References

[1] R. Agrawal, T. Imieliński, and A. Swami, *Mining association rules between sets of items in large databases*, SIGMOD Rec., 22 (1993), pp. 207–216.

[2] C. Antunes, *Pattern Mining over Nominal Event Sequences using Constraint Relaxations*, PhD thesis, IST, 2005.

[3] C. Antunes, *D2pm: Domain driven pattern mining, project report*, Tech. Report 1530, IST, Lisboa, 2011.

[4] E. Arjas, H. Mannila, M. Salmenkivi, R. Suramo, and H. Toivonen, *Bass: Bayesian analyzer of event sequences*, in COMPSTAT, Barcelona, Spain, 1996, pp. 199–204.

[5] I. Batal, Valizadegan, Cooper, and M. Hauskrecht, *A pattern mining approach for classifying multivariate temporal data*, in IEEE BIBM, 2011, pp. 358 –365.

[6] R. A. Baxter, G. J. Williams, and H. He, *Feature selection for temporal health records*, in PAKDD, London, UK, 2001, Springer-Verlag, pp. 198–209.

[7] R. Bellazzi, F. Ferrazzi, and L. Sacchi, *Predictive data mining in clinical medicine: a focus on selected methods and applications*, Data Min. Know. Disc., 1 (2011), pp. 416–430.

[8] M. Berlingerio, F. Bonchi, F. Giannotti, and F. Turini, *Mining clinical data with a temporal dimension: A case study*, in IEEE BIBM, IEEE CS, 2007, pp. 429–436.

[9] G. Bruno and P. Garza, *Temporal pattern mining for medical applications*, in Data Min.: Found. and Int. Paradigms, vol. 25 of ISRL, Springer Heidelberg, 2012, pp. 9–18.

[10] S. Chakravarty and Y. Shahar, *A constraint-based specification of periodic patterns in time-oriented data*, in TIME, IEEE CS, 1999, pp. 29–40.

[11] T. P. Exarchos, M. G. Tsipouras, C. Papaloukas, and D. I. Fotiadis, *A two-stage methodology for sequence classification based on sequential pattern mining and optimization*, Data Knowl. Eng., 66 (2008), pp. 467–487.

[12] M. N. Garofalakis, R. Rastogi, and K. Shim, *Spirit: Sequential pattern mining with regular expression constraints*, in VLDB, San Francisco, CA, USA, 1999, Morgan Kaufmann Publishers Inc., pp. 223–234.

[13] P. Geurts, *Pattern extraction for time series classification*, in PKDD, London, UK, 2001, Springer-Verlag, pp. 115–127.

[14] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[15] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, John Wiley & Sons, Inc., USA, 2nd ed., 2002.

[16] N. Lesh, M. J. Zaki, and M. Ogihara, *Mining features for sequence classification*, in KDD, ACM, 1999, pp. 342–346.

[17] W. Li, J. Han, and J. Pei, *Cmar: Accurate and efficient classification based on multiple class-association rules*, in ICDM, IEEE CS, 2001, pp. 369–376.

[18] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[19] H. Mannila, H. Toivonen, and A. Inkeri Verkamo, *Discovery of frequent episodes in event sequences*, Data Min. Knowl. Discov., 1 (1997), pp. 259–289.

[20] F. Mörchen, *Time series knowledge mining*, Wissenschaft in Dissertationen, Görich & Weiershäuser, 2006.

[21] A. Mueen, E. J. Keogh, Q. Zhu, S. Cash, and M. B. Westover, *Exact discovery of time series motifs*, in SDM, SIAM, 2009, pp. 473–484.

[22] G. Norén, J. Hopstadius, Bate, Star, and I. Edwards, *Temporal pattern discovery in longitudinal electronic patient records*, Data Min. Knowl. Discov., 20 (2010), pp. 361–387.

[23] B. Özden, S. Ramaswamy, and A. Silberschatz, *Cyclic association rules*, in ICDE, Washington, DC, USA, 1998, IEEE CS, pp. 412–421.

[24] B. Padmanabhan and A. Tuzhilin, *Pattern discovery in temporal databases: A temporal logic approach*, in KDD, ACM, 1996, pp. 351–354.

[25] D. Patel, W. Hsu, and M. L. Lee, *Mining relationships among interval-based events for classification*, in SIGMOD, New York, NY, USA, 2008, ACM, pp. 393–404.

[26] J. Pei, J. Han, and W. Wang, *Mining sequential patterns with constraints in large databases*, in CIKM, New York, NY, USA, 2002, ACM, pp. 18–25.

[27] P. Sebastiani, M. Ramoni, V. Nolan, C. Baldwin, and M. Steinberg, *Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia*, Nature Genetics, 37 (2005), pp. 435–440.

[28] A. Silberschatz and A. Tuzhilin, *What makes patterns interesting in knowledge discovery systems*, IEEE Trans. on Knowl. and Data Eng., 8 (1996), pp. 970–974.

[29] R. Srikant and R. Agrawal, *Mining sequential patterns: Generalizations and performance improvements*, in EDBT, London, UK, 1996, Springer-Verlag, pp. 3–17.

[30] M. Stacey and C. McGregor, *Temporal abstraction in intelligent clinical data analysis: A survey*, Artif. Intell. Med., 39 (2007), pp. 1–24.

[31] V. Tseng and C.-H. Lee, *Effective temporal data classification by integrating sequential pattern mining and probabilistic induction*, Expert Sys.App., 36 (2009), pp. 9524–9532.

[32] Q. Zhou and Q. Z. R. Fikes, *A reusable time ontology*, in AAAI IW on Ontologies for the Semantic Web, 2002.