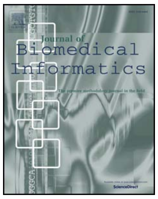
Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Consistent discovery of frequent interval-based temporal patterns in chronic patients' data

Alexander Shknevsky*, Yuval Shahar, Robert Moskovitch

Software and Information Systems Engineering, Ben-Gurion University, Beer Sheva, Israel

ARTICLE INFO

Keywords:
Temporal data mining
Temporal knowledge discovery
Temporal abstraction
Time intervals mining
Frequent pattern mining
Pattern repetition
Pattern consistency
Clustering
Classification
Prediction

ABSTRACT

Increasingly, frequent temporal patterns discovered in longitudinal patient records are proposed as *features* for classification and prediction, and as means to cluster patient clinical trajectories. However, to justify that, we must demonstrate that most frequent temporal patterns are indeed consistently discoverable within the records of different patient subsets within similar patient populations.

We have developed several measures for the consistency of the discovery of temporal patterns. We focus on *time-interval relations patterns* (TIRPs) that can be discovered within different subsets of the same patient population. We expect the discovered TIRPs (1) to be *frequent* in each subset, (2) preserve their "*local*" metrics - the absolute frequency of each pattern, measured by a Proportion Test, and (3) preserve their "*global*" characteristics - their overall distribution, measured by a Kolmogorov-Smirnov test. We also wanted to examine the effect on consistency, over a variety of settings, of varying the minimal frequency threshold for TIRP discovery, and of using a TIRP-filtering criterion that we previously introduced, the *Semantic Adjacency Criterion* (SAC).

We applied our methodology to three medical domains (oncology, infectious hepatitis, and diabetes). We found that, within the minimal frequency ranges we had examined, 70–95% of the discovered TIRPs were consistently discoverable; 40–48% of them maintained their local frequency. TIRP global distribution similarity varied widely, from 0% to 65%. Increasing the threshold usually increased the percentage of TIRPs that were repeatedly discovered across different patient subsets within the same domain, and the probability of a similar TIRP distribution. Using the SAC principle, enhanced, for most minimal support levels, the percentage of repeating TIRPs, their local consistency and their global consistency. The effect of using the SAC was further strengthened as the minimal frequency threshold was raised.

1. Introduction

Analyzing time-oriented data enables researchers to discover new temporal knowledge and gain understanding regarding the temporal behavior and temporal associations of such data, with the further objectives of clustering, classification, and prediction. This functionality is required in many time-oriented domains and tasks, such as finance, information systems security [52,56], intelligence, and, in particular, medicine [43,44]. An example from the medical domain would be supporting a clinical researcher who analyzes the results of a clinical trial of a new drug within a population of chronic patients.

A large number of business intelligence, statistical, and *Data Mining* (DM) or *Temporal Data Mining* (TDM) tools exist. These tools typically focus mainly on the analysis of raw, time-stamped data, such as specific Hemoglobin values on a particular date. However, basing the analysis on meaningful *periods*, or *time intervals*, rather than on time *points*, and on *abstract* (symbolic) concepts that can hold over such time intervals, such as a period of Moderate-Anemia, rather than on *raw* data, has many benefits. The use of symbolic time intervals instead of raw time-stamped data can reduce inherent random noise in the data, avoid problems resulting from sampling the data at different frequencies and at various temporal granularities, and often alleviate the problem of missing data [37].

Thus, to significantly enhance the capabilities of temporal data analysis, a preprocessing step of meaningful summarization and interpretations of the time-stamped raw data (e.g., a series of time-stamped hemoglobin values) into interval-based abstractions (e.g., a period of three months of moderate anemia), known as *temporal abstractions*, can be used. In general, *Temporal Abstraction* is the abstraction and aggregation of a time point series into a succinct, symbolic, time intervals series-based representation, suitable for a human decision-maker or for the purposes of data mining.

Such time intervals are often called *symbolic time intervals*. The temporal-abstraction process can use knowledge-based approaches,

* Corresponding author.
E-mail addresses: sheknabs@bgu.ac.il (A. Shknevsky), yshahar@bgu.ac.il (Y. Shahar), robertmo@bgu.ac.il (R. Moskovitch).

http://dx.doi.org/10.1016/j.jbi.2017.10.002
Received 11 March 2017; Received in revised form 23 August 2017; Accepted 2 October 2017
Available online 04 October 2017

which exploit domain-specific knowledge [48], or data-driven, domain-independent discretization methods [3], [14,30,34,39]. The potential usefulness of temporal abstraction in medical domains, for classification, prediction, visualization, and even natural-language summarization was demonstrated [5,16,27,41,44,57].

The patterns that can be formed by noting the temporal relationships among the symbolic time intervals are at least as interesting as the temporal abstractions that the symbolic intervals represent. These patterns essentially characterize frequent *temporal pathways*, *trajectories*, or *journeys*, within a given population of entities described by longitudinal multivariate data. Thus, in the case of the medical domain, they might be viewed as creating a *temporal clustering* of chronic patients who have a particular condition, according to the course of their disorder. Such a clustering has been demonstrated in the Type II Diabetes domain [36].

Several algorithms have been proposed for the purpose of discovering frequent *time-interval relation patterns* (*TIRPs*) within a set of symbolic time intervals [37]. The potential for using the discovered TIRPs as *classification* or *prediction features* was demonstrated in further studies [4–6,35,38,39].

Not all frequent temporal patterns are equally useful for classification and prediction. In a previous study, we addressed a general problem, namely, that many of the discovered temporal patterns, although *syntactically* valid, are not sufficiently transparent to a domain expert, and in particular, an expert clinician [53]. The lack of transparency refers to the sense in which the discovered patterns do not conform to the expert's basic *semantic* intuitions, such as regarding potential causality relations among the symbolic intervals that form the components of the pattern. For example, some instances of a frequent temporal pattern consisting of the administration of a high dose of a certain medication, followed by a decreasing systolic blood pressure, might in fact include between instances of these two symbolic intervals also an instance denoting a period of *increasing* blood pressure. Such instances of the temporal pattern do conform, syntactically, to the overall constraint of the discovered pattern, but semantically seem very different from an interval of a high-dose medication administration followed by an interval of dropping blood pressure, with no potentially conflicting symbolic time interval between them. Furthermore, many of the discovered patterns might be less useful as features for classification and prediction.

Earlier studies have noticed a potential drawback inherent in redundant patterns [4,13,38]. These studies have mostly considered the issue from a computational perspective, i.e., for the purpose of the discovery of all frequent temporal patterns. However, they have not considered that issue from a *functional* point of view, namely, when considering the possible uses to which the discovered TIRPs can be put (e.g., for classification).

We have therefore introduced a new, formally defined constraint, the *Semantic Adjacency Criterion* SAC), to address explicitly functional considerations, such as the need for transparency of the discovered patterns from the point of view of the domain experts, and the effectiveness of using these patterns for classification and prediction [53]. Using the SAC, we have demonstrated a significant reduction (up to 97%) in the number of the discovered SAC-obeying TIRPs and in the runtime required to discover them (up to 98%), when compared with the original runtime. Nevertheless, *the reduced patterns set, when used as features, resulted in classification and prediction models that were quantitatively at least as good,* with respect to their performance, *as those that used the complete set of discoverable patterns* [53–54].

As we shall see in the current study, using the SAC principle turned out to be one of the approaches that can alleviate the problem of consistent discovery of the same TIRPs across different subsets of the same subjects' population.

Indeed, a thorny issue plagues most of the previous TDM studies. Despite the fact that, as we shall see, many frequent temporal pattern discovery algorithms exist, as well as multiple applications of these algorithms to classification and prediction, most of the studies ignore the issue of *validating the consistent repeatability of the results of the discovery process*, and thus the very issue of *whether any, some, most, or all of the TIRP-based features exist in each subset of the target [patient] population.*

Unlike common examples in classification, in which the features are *static* variables, e.g., age, gender, or weight, and thus can be found in all instances of subjects in the population of interest, the newly discovered knowledge in the case of TDM, is *dynamic*, e.g., sufficiently frequent relationships between periods of red blood cell counts and periods of white blood cell counts. Typically, a set of frequent temporal patterns, discovered within a *training* [patient] population, is used to induce a classifier, and that classifier is applied to a different population of subjects [patients] of the same type; or perhaps even to another subset of the same population as the training population, used as a *testing* set (e.g., when performing a *k*-fold cross validation).

But if some of the patterns found in the *training* set are not sufficiently frequent to be discovered in the *testing* set, or even if most patterns discovered within the training set do appear in the testing set, but in significantly different proportions, we cannot even effectively use them as features to classify the subjects within the testing set, let alone validate the accuracy of the classifier using them.

Such a situation might undermine the main assumption underlying the whole TDM framework, and in general, the *Inductive Learning Hypothesis* of the Machine Learning field [32]; namely, that the features, in this case, the frequent temporal patterns discovered within the *training-set* population, exist (i.e., will also be discovered) within the *testing-set* population (or at least, a significant portion of these features). The Inductive Learning Hypothesis essentially postulates that, given a sufficiently large training set, the same classification model ("Hypothesis") will hold in the new set of subjects. But of course, it cannot even be *assessed*, let alone be *validated*, when the features of the model *do not exist* in the test set.

In the current study, we developed several methods to assess and validate the consistent discovery of a set of frequent temporal patterns (i.e., TIRPs) in populations that involve similar subjects (patients, in this case, since we examined databases from three different medical domains). That is, we are validating that the TIRPs that were discovered in one subset of the given patient population, are indeed similar in a formal sense to the TIRPs discovered in another subset of the same patient population. If so, the discovered TIRPs can indeed be used for various tasks, such as classification or prediction. To be *similar* in our sense, the discovered TIRPs need to preserve their *local* characteristics (in particular, absolute frequency of each TIRP), as well as the *global* characteristics of the whole set of discovered temporal patterns (in particular, their overall distribution).

1.1. Our contributions

As we shall show, the existence, frequency, and distribution measures of TIRP discovery consistency depend on the frequency threshold (minimal support) used to discover the temporal patterns (increasing the threshold enhances similarity), and on whether the SAC constraint was used when discovering the TIRPs: Discovering only SAC-obeying TIRPs in each of the patient population subsets increases similarity.

Thus, our main contributions in the current study include:

(1) Measures for assessing the similarity among sets of frequent temporal patterns that can be discovered in different subject populations of a similar type.
(2) The rigorous application of these measures to three different medical domains; how the minimal support level and the application of three versions of the SAC constraint affects the existence, local frequency, and global distribution similarity measures.
(3) The demonstration that the existence, frequency, and distribution similarity measures among the discovered TIRP sets within

different subsets of the same patient populations increase:
(a) As the minimal support threshold increases.
(b) When applying the SAC principle to filter out TIRPs.
(c) Even more so, when both methods are used.

1.2. Outline of the paper

In Section 2, we provide the necessary background for the rest of the paper, and in particular, discuss the temporal-abstraction task and methods, the TIRP discovery task and method and in particular the particular methods that we have chosen, and the formal definition of the SAC principle. In Section 3, we describe our computational framework, and the definition of the TIRP consistent discovery validation methods. Section 4 describes the design of the experiments that we had performed within three different medical domains, while Section 5 describes the results. Section 6 discusses the results and presents the main conclusions of this study.

2. Background and related work

The *temporal* aspect, which is crucial in clinical data, is often insufficiently addressed in knowledge-discovery systems; nevertheless, it often requires more complex models and reasoning mechanisms. Time-oriented data can be measured at various temporal granularities and frequencies, and it is often useful to aggregate the raw, time-stamped data into meaningful, abstract concepts that hold over time intervals.

2.1. Temporal abstraction

Temporal Abstraction (TA) is an aggregation or interpretation of (typically) time-stamped data into a succinct (usually with symbolic values) time intervals, which we refer to as *temporal abstractions* or as *symbolic time intervals*. For example, in the medical domain, a 36.6 °C value of the temperature concept might be interpreted as "normal", while the value of 39.7 °C might be interpreted as a "very high" value of the state of that concept, and might even be referred to using the label *Fever*. Such an abstraction might hold not only over a time point, but over a whole time interval, such as "four days of fever". A comprehensive survey of temporal abstraction in intelligent clinical data analysis has been previously conducted [55].

The use of symbolic time intervals to abstract the raw, time-stamped data can reduce inherent random noise in the data (due to the discretization), avoid problems resulting from sampling the data at different frequencies and at various temporal granularities (due to the aggregation of the data into continuous intervals representing significant time periods), and overcome missing data (due to an interpolation process that is, to some extent, a part of most TA methods).

TA can use knowledge-based approaches, such as the *Knowledge-Based Temporal Abstraction* KBTA) method [48], which exploits domain-specific knowledge, to generate *state* abstractions, as well as more complex abstractions, such as *gradients*. The KBTA method has been applied in multiple domains, such as medicine, transportation, and cyber-security, for purposes such as interpretation, monitoring, visual exploration, classification, and prediction [27,31,47,49,50,51]. However, when no suitable domain knowledge exists, various automatic discretization methods exist [3,14], which typically focus on finding cut-off values (using various heuristics) for discretizing continuous data; *Equal Width Discretization* (*EWD*) method, which was been demonstrated to often be sufficient for purposes of classification or prediction [3,38]; the *SAX* method for discretization of time series [30]; and even discretization methods designed classification tasks, such as the *Temporal Discretization for Classification* (*TD4C*) method [39].

2.2. Time intervals mining

Symbolic time intervals mining [33], which discovers frequently repeating patterns of *temporal* relations amongst the symbolic intervals, is a relatively recent TDM subfield, and most of the time interval mining methods use some subset or variation of Allen's temporal relations [2]. Allen defined 13 temporal relations and their transition table, based on seven basic temporal relations and their inverses (e.g., the inverse of *before* is *after*, while *equals* is its own inverse). Fig. 1 illustrates these seven temporal relations, as well as the option of abstracting these relations into only three relations, as was previously examined [37]. However, there are also other approaches, such as the paradigm of applying *Time-Annotated Sequences* TAS) [7].

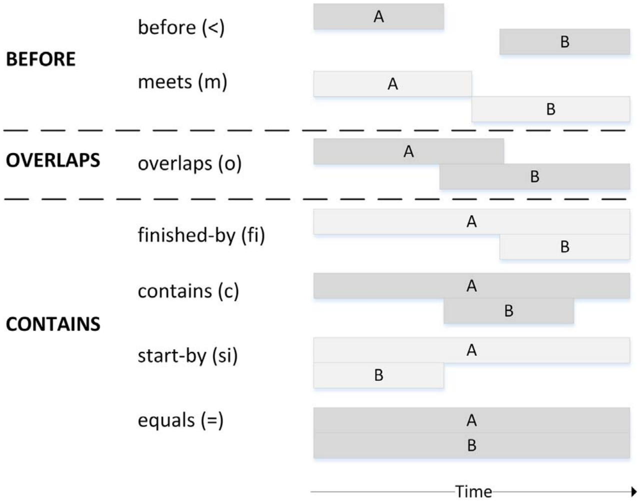


Fig. 1. Three abstract temporal relations, consisting of conjunctions of Allen's seven basic temporal relations.

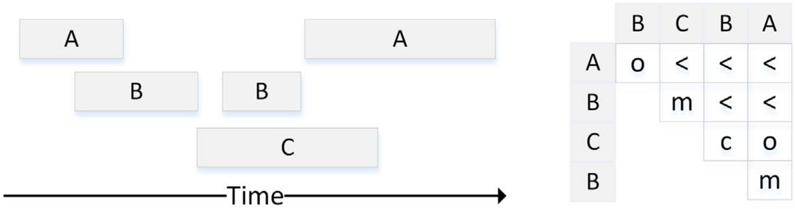
Höppner [22] introduced an advanced method to mine rules in symbolic time intervals sequences using Allen's temporal relations and used a non-ambiguous matrix of all the temporal relations defining the pattern. Winarko and Roddick's [58] approach corresponds to Höppner's representation, but these researchers had used only half of the matrix. Papapetrou et al. [40] presented a time interval mining method that corresponds to Höppner's definition, indexing first all the pairs of symbolic time intervals by their temporal relations, generating an enumeration tree that spans all of the discovered and extended patterns.

2.2.1. Time intervals mining and the KarmaLego algorithm

Following Moskovitch and Shahar, we define a symbolic time interval, $I = \langle s, e, sym \rangle$, as an ordered pair of time points, start-time (*s*) and end-time (*e*), and a symbol (*sym*), which represents one of the domain's symbolic concepts set *S* [37]. A non-ambiguous lexicographic *Time Intervals Related Pattern* (TIRP) $P = \{I, R\}$ is defined as a set I of *k* symbolic time intervals ($I_1..I_k$ ordered lexicographically by start time, then end time, then symbol; and a set R of all of the temporal relations among each of the $(k^2 - k)/2$ pairs of symbolic time intervals in I.

Fig. 2 presents a typical TIRP, represented as a half-matrix of temporal relations. Note that half-matrix representation (as opposed to a full matrix) is possible due to the canonical lexicographic ordering of the symbolic time [37], which leads to a unique half matrix for each TIRP.

Following again Moskovitch and Shahar, we also define the concept of *Vertical Support* [37]: Given a database of $|E|$ distinct entities (e.g., different patients), each represented as a set of symbolic time intervals (in which one or more symbols might repeat over different symbolic time intervals), the *vertical support* of a TIRP *P* is denoted by the cardinality of the set E^P of distinct entities within which *P* holds at least once, divided by the cardinality of $|E|$. When a TIRP has a vertical support above a given minimal predefined threshold *min_ver_sup*, it is referred to as *frequent*. Thus, the time intervals mining task is defined



	B	C	B	A
A	o	<	<	<
B		m	<	<
C			c	o
B				m

Fig. 2. An example of a non-ambiguous lexicographic TIRP represented by 5 lexicographic ordered symbolic time intervals and all of their pair-wise temporal relations. Modified from Moskovitch and Shahar [37].

accordingly: given a set of |E| entities and a *min_ver_sup* threshold, the goal is to find all of the frequent TIRPs.

When multiple instances of a TIRP are discovered within the same entity's longitudinal record, we can calculate also the TIRP's *Horizontal Support* within the record, which is the number of instances discovered per entity. We can also calculate the *Mean Duration*, which is the average duration, from start to end, of the TIRP instances within that record. Both these metrics were used successfully for classification, and showed a better classification performance than the mere existence of the TIRP [38].

The KarmaLego algorithm is a TIRPs discovery algorithm that has an advanced data structure, and exploits the transitivity of temporal relations to significantly decrease the number of TIRP candidates generated at runtime [37]. It is the algorithm that we use in our experiments, mostly due to its availability and efficiency. The KarmaLego algorithm, like similar algorithms, outputs an *enumeration tree* of all of the frequent TIRPs discovered in the given database. The enumeration tree lists explicitly the discovered TIRPs as well as the specific TIRP-instances supporting them, each composed of a set of symbolic time intervals. Each discovered TIRP can be found at a leaf of the tree, while its components (symbolic intervals and temporal relations) appear on the path from the root of the tree to the leaf. Each TIRP (path) represents a cluster of entities (patients, in our experiments) having similar qualitative temporal relations among their multivariate variables.

In our study, two sets of temporal relations were used, Allen's original seven temporal relations and a set of three abstract temporal relations, shown in Fig. 1.

2.3. The semantic adjacency criterion

Many discovered temporal patterns, although *syntactically* correct, are not *transparent*, in the sense that they do not conform to basic *semantic* intuitions of domain experts (i.e., to the medical domain experts) such as regarding an association that might imply a potential causality. Note that a symbolic interval is always composed of a *concept*, such as "the trend of the Hemoglobin level" and a *value*, such as "Increasing". Fig. 3 presents the TIRP composed of two symbolic intervals and one temporal relation: "〈Medication-dose-level = High〉 *before* 〈HGB-level = Low〉". The TIRP seems to intuitively imply that administering a medication at a High dose is [temporally] associated with a Low value of the hemoglobin level concept (a *state* abstraction of the HGB-value raw-data concept) following that administration. However, the TIRP might in fact include instances of intervals that represent values that are semantically contradicting to those appearing in the formal TIRP definition. Such contradictions might be instances in which, between the TIRP's first and the second symbols, there is a contradicting concept and value, such as "HGB-level = High" (or "Medication-dose-level = Very_Low"). Either of these contradictory associations might change, and in this case even reverse, the intuitive interpretation of the original temporal association, since it now seems that a High medication-dose level might be actually associated with a High Hemoglobin value (or conversely, that a Low medication dose level is associated with Low hemoglobin level).

We have therefore previously developed a new filtering constraint to be used during the pattern mining process, the *Semantic Adjacency Criterion* SAC) [53], which disallows the existence of potentially contradictory symbolic time intervals between pairs of symbolic time intervals within the discovered pattern.

A *contradictory symbolic interval* that is found between a pair of symbolic intervals, is any symbolic time interval that has the same concept type as one of the members of the pair (e.g., "Medication-dose-level"), but that has a different value (e.g., "Very_Low" instead of "High"). The contradiction is easily found by exploiting the domain's ontology (i.e., the fact that we know not only the overall symbol denoting each symbolic interval, but also its actual concept type and value, which hold, in the logical sense, over the temporal span of the symbolic interval).

We have defined three versions of the SAC criterion:

(1) The *Sequential SAC* (SSAC) disallows contradictory intervals between any two successive symbolic intervals; the succession relation between two symbolic intervals is defined only when the TIRP is represented as a sequence of symbolic intervals.
(2) The *Conservative SAC* (CSAC) disallows a contradiction between *any* two symbolic intervals, regardless if they appear in succession in the TIRP's definition.
(3) The *Liberal SAC* (LSAC) only checks the SAC constraint between symbolic intervals over which hold *different* types of concepts (e.g., between a medication-dose level and a Hemoglobin level, but *not* between two periods of some Hemoglobin level).

All three SAC versions were tested within several medical domains oncology, infectious hepatitis, and type II diabetes), by embedding them within an efficient framework for mining frequent temporal patterns and for exploiting the resulting TIRPs as features for classification and prediction tasks [38].

The results of using the SAC constraint to prune potential TIRP candidates when mining an interval-based data set in the three different medical domains demonstrated a significant reduction, up to 97%, in the number of discovered TIRPs, and a significant reduction in the

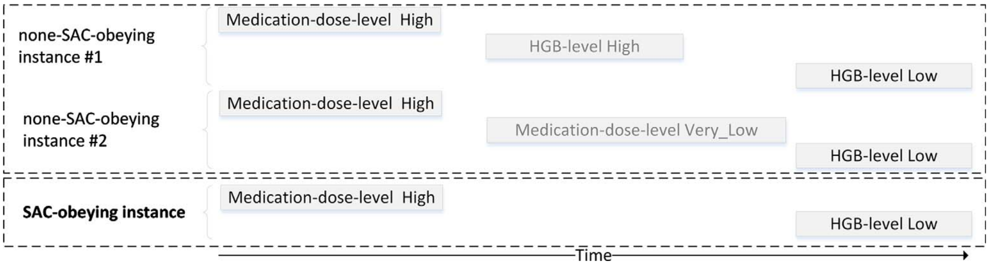

none-SAC-obeying instance #1
Medication-dose-level High
HGB-level High
HGB-level Low
none-SAC-obeying instance #2
Medication-dose-level High
Medication-dose-level Very_Low
HGB-level Low
SAC-obeying instance
Medication-dose-level High
HGB-level Low
Time

Fig. 3. Three syntactically equal instances of the same TIRP, which includes several symbolic intervals, "〈Medication-dose-level = High〉 Before 〈HGB-level = Low〉". Instance #1 describes a situation in which the pair of intervals is not semantically adjacent, since there is a High hemoglobin-level value between them that contradicts the pattern's semantics; Instance #2 similarly, contains different medication dose. The bottom instance describes a semantically adjacent instance, since no contradicting value exists between the relevant pair of intervals.

runtime of the TIRP discovery process, up to 98%, when compared with the original runtime (both measures depending, of course, on the minimal vertical support threshold used). The most conservative criterion version, CSAC, resulted in the smallest number of discovered patterns and in the shortest runtime.

Nevertheless, the significantly reduced set of discovered TIRPs, used as features for four classifier-induction algorithms belonging to four different machine-learning algorithm families, resulted in classification and prediction models that were quantitatively at least as good or better, with respect to their performance, as those that used the complete set of discoverable TIRPs [54].

This insight led to our current motivation for examining the effects of using it as one of the measures that might enhance TIRP discovery consistency among different subsets of the same [patient] population.

2.4. Related work in the medical domain

In Sections 1 and 2.2, we have looked at the TDM task in general; but we will focus on TDM approaches that were applied to the medical field, due to the intrinsic longitudinal nature of clinical data. The main challenge in Electronic Health Records analysis is the sparsity and irregular sampling nature of medical data, since clinical data are commonly recorded only when patients enter the healthcare system, providing a rather sparse and biased view of the patient's clinical history [24,25]. In addition, the data appear in several forms, from numerical values occurring at a particular time e.g. laboratory tests) to events that may span many days, months, or years e.g. conditions, and drug exposures). These challenges complicate the use of existing temporal modelling strategies [1,26,36,48,55].

Although the study of TDM has made much progress over the past several decades, relatively little work has been dedicated to the unique challenges of the medical domain [26]. In the medical domain, thinking in time *durations*, or *intervals*, is intrinsic to the domain. For example, the period of time over which some medication or treatment is prescribed for the patient. The inherent importance of durations to the medical domain is one of the reasons that we focus our methodology and experiments on mining symbolic time intervals [10,38,45,53].

Nevertheless, other interesting approaches exist. For example, there is an increasing interest in *Temporal Functional Dependencies* TFDs), e.g., in the domains of psychiatry and pharmacovigilance [9,10,45]. In particular, *Interval-based TFDs*, in a manner similar to the patterns discovered by time intervals mining, are also intended to represent relations among time intervals, using Allen's temporal relations. However, in the case of TFDs, the events' durations typically refer to the specific *quantitative* constraints over the shortest and longest event durations, rather than the *qualitative* constraints in time intervals mining, in which typically the durations of the time intervals are *not* part of the pattern definition [37,53]. We focused in the current study on the repeatability of the patterns discovered by time intervals mining, when using qualitatively all of Allen's temporal relations, or an abstract subset of these relations. In addition, we also used a domain-specific temporal-abstraction knowledge base (for computing the temporal abstractions, such as “Low systolic blood pressure”).

Other related directions in TDM in the medical domain include *process mining* and *association-rule mining*. Unlike our study, these approaches do not attempt to discover general repeating temporal patterns, but they do bear some resemblance to the approaches on which we focus in the current study, due to their attempt to discover various types of meaningful repeating patterns. For example, mining patterns of meaningful healthcare activities in Type II diabetes patients data using knowledge-based temporal abstractions [11], or processing various clinical measures and guessing the patient's possible state of disease in brain tumor patients data using temporal-association rules [46]. Other approaches involve using combinations of several techniques; an example is the use of *Multiscale Structure Matching* to match time sequences between patients, and *Rough Clustering* to separate the [temporal] sequences that are definite members of a cluster from the sequences that are only possible members of a cluster; in that study, the clusters represented interesting temporal patterns in the data of chronic hepatitis patients [19].

Our contribution is in examining the *consistency* of the frequent temporal pattern discovery task. This consistency directly affects the use of the discovered patterns as features for classification of multivariate time series, and directly affects any approach that attempts to discover frequent temporal patterns of some type, e.g., sequential patterns, to characterize these data in some meaningful fashion. In our study, we have performed the TDM and the assessment tasks in several medical domains, using a general consistency assessment methodology that we present in detail in the next section (See more about the generality of our approach in the Discussion).

The following section will delve deeply into the methods we propose to use to assess the validity of the consistent discovery of the temporal patterns in terms of TIRPs, without assuming anything about their internal structure or how they were discovered. For example, the patterns can be represented by different definitions of temporal patterns, which involve a certain flexibility regarding the semantics, for example, of the *meets*, *before*, or *overlaps* temporal relations.

3. Methods: local and global TIRP set similarity assessment

A variety of TIRPs can be discovered in time oriented data, and in particular, in multivariate longitudinal clinical data, but can we really use them all to characterize the data? Are any of the patterns spurious? Or redundant? [29].

One way to check the consistency of the TIRP discovery process across two different subsets of a similar [patient] population is to examine which TIRPs had managed to cross the minimal support threshold in both patient-record subsets, and to measure the size of the intersection (the portion of the frequent TIRPS found in one set that are also found sufficiently frequently in the second set). Each TIRP set is complemented by zero frequencies for the frequent TIRPs discovered in the other set that were not discovered in the current set. That is a pure *existence* test.

Indeed, we shall use this simple similarity concept, but we shall also define and use two additional similarity measures, *local* and *global* frequency similarity.

3.1. Validation of locally consistent TIRP discovery

We refer to the TIRP's *local consistent discovery* as the possibility that *this specific* TIRP will appear in two subsets of the same population with similar characteristics. In most time intervals mining algorithms, the vertical support is used to filter the most frequent temporal patterns [22,37,40]. Thus, to assess the local similarity of a TIRP we should test its vertical support, which is the proportion of patients has this TIRP in their records within the mined group. To assess the *local similarity* of the vertical support of same TIRP in two subsets of the same patient population, we shall use the well-known *Proportion Test.*

Given a database D that is divided into several subsets D = {D_1, D_2..D_n}, the discovery of a particular TIRP in subset D_1 (containing N_1 objects) with a vertical support (proportion) p_1 is compared to the discovery of the same TIRP in subset D_2 (containing N_2 objects) with a proportion p_2 (treating the proportion as 0, if not found in one of the subsets). The proportion test is a test for the significance of *the difference between two proportions* [18], which means rejecting the hypothesis that the two proportions arise from the same population. Given the weighted mean of two samples' proportions defined by

$$P_e = \frac{N_1p_1 + N_2p_2}{N_1 + N_2},$$

one can calculate the significance using a z ratio defined by

$$z=\frac{|p_1-p_2|}{\sqrt{p_e(1-p_e)\left(\frac{N_1+N_2}{N_1N_2}\right)}},$$

and reject the hypothesis (based on significance level, α, generally $\alpha = 0.05$), with a one-tailed reject if $1-\phi(z) < \alpha/2$ or with a two-tailed reject if $1-\phi(z) < \alpha$, where $\phi(z)$ is the probability that the statistic is less than z.

3.2. Validation of globally consistent TIRP discovery

Since we are dealing with a *set* of TIRPs, or a "Bag-of-TIRPs" [38], a statistical analysis of the patterns is even more important [29]. Even if the same set of TIRPs is repeatedly discovered within different subsets of the same population, the *overall* set of TIRPs *distribution* might be quite different.

First, note that *if the same TIRPs are not discoverable within the two sub-populations, we cannot use them* for classification or prediction (or, alternatively, their power for classification and prediction does not exist). We would like the training and tests sets to have, as much as possible, similar features; else, the underlying assumption enabling us to *induce* a classifier, using standard machine learning methods, might be false.

Multiple tests exist for comparing two distributions; however, for the purpose of this study, we chose the *Two-sample Kolmogorov-Smirnov* (*K-S*) test [21]. The K-S test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two given samples.

Other tests for comparing two distributions exist; e.g., the Kullback–Leibler divergence [28] or even tests intended to be applied to a structured data set, such as, in our case, the KarmaLego algorithm's output of an enumerated tree of TIRPs. For example, the *tree kernel* [15] and *tree edit distance* [8]. However, using the Kullback–Leibler test requires a normalization of the distribution, while using the tree methods requires using a predefined weight for each possible deviation (insertion, deletion, and more complex distance measures, such as between different temporal relations). Thus, we decided to stick to the K-S test, whose semantics are well understood, and which requires making only very few assumptions.

The K-S test is a nonparametric test of the similarity of two distributions that can be used to compare a sample with a reference probability distribution, or to compare two samples (i.e., the Two-sample Kolmogorov–Smirnov test). The Two-sample Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the two samples are drawn from the same distribution.

Given two independent random samples $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ (which in our case are the vertical support values of the TIRPs) with distribution functions F and G, respectively, we shall test whether the two distributions arose from the same population (accepting the null hypothesis) or not. To test the hypothesis we first need to calculate the empirical distribution function of the two samples defined by,

$$F_m(t) = \frac{number\ of\ sample\ X's \leqslant t}{\mathrm{m}}$$

and

$$G_n(t) = \frac{number\ of\ sample\ Y's \leqslant t}{n}.$$

Given the two distributions, we can calculate the statistic defined by

$$\mathrm{J}=\frac{\mathrm{mn}}{\mathrm{d}}\max_{i=1,\ldots,N}\{|F_m(Z_i)-G_n(\mathrm{Z_i})|\},$$

when d is the greatest common divisor of m and n, and $Z_1,\ldots,Z_N$ are ordered values for the combined sample $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$.

By specifying the type I error probability as α, and retrieving the critical value J_α [42], we can reject the hypothesis if $J \geq J_\alpha$, otherwise accept it.

Note that the K-S test makes the assumption that the underlying distribution of the features that are discovered in both subsets includes the same features. However, this assumption is not necessarily true in the case of features that are the TIRPs whose vertical support within each subset was above a certain minimal vertical support threshold. Thus, in our assessment of the similarity between two TIRP distributions, we shall always validate first that a sufficient *intersection* exists between the sets of features appearing in the two subsets to be compared. Thus, we shall apply the K-S test only after testing that more than 50% of the overall number of [frequent] TIRPs passing the minimal vertical support threshold, in both sets, are indeed found in both subsets of the population.

4. Assessing the consistency of TIRP discovery in the medical domain: the experimental design

Given our informal hypotheses regarding the *existence*, *local*, and *global* consistency of the TIRP discovery methods, our three specific research questions (for all of the different settings conditions that we shall detail below) were:

1. **Existence**: Are the discovered TIRPs indeed found (i.e., discovered sufficiently frequently) in different subsets of the same patient population? Does using a SAC version during TIRP discovery matter?
2. **Local consistency**: Does each individual frequent TIRP maintain its vertical support level within different subsets of the same patient population? Does using a SAC version during TIRP discovery matter?
3. **Global consistency**: Does the overall set of frequent TIRPs maintain its distribution within different subsets of the same patient population? Does using a SAC version during TIRP discovery matter?

For uniformity reasons and to avoid biases, we used the *Holdout Random Subsampling* technique, by randomly dividing the population into mutually exclusive halves, and used the KarmaLego algorithm to discover TIRPs in both halves and compute the three consistency measures across all possible configuration settings. With respect to the different conditions (configurations) we checked: The evaluation was performed across different discretization/abstraction methods (knowledge-based [KB], EWD, SAX, and TD4C-KL), each with three bins, two different temporal relation sets (either the three abstract temporal relations, or the full seven temporal relations, both shown in Fig. 1), varying the minimal vertical support levels, and, when generating candidate TIRPs, using CSAC, SSAC, LSAC, or no SAC version.

4.1. The data sets

Three clinical datasets were used: (1) an oncology dataset from the Rush Medical Center, Chicago, USA, including patients who had undergone either allogeneic or autologous bone-marrow transplantation; (2) a hepatitis data set describing patients who had either Hepatitis B or C, from a KDD conference challenge [20], which is publicly available [12], and (3) a diabetes dataset from our local academic medical center [17], including patients who had been followed (albeit sporadically) for at least five years.

Table 1 describes the characteristics of the three data sets used throughout all of the evaluations: the number of total records, the number of entities (i.e., patients), the number of concepts (e.g., Hemoglobin State in a particular context, as explained in Section 2.2) and the average number of records per entity. The full description of the data sets and the knowledge base used within each domain, in the case of the knowledge-based temporal-abstraction method, appear in the appendix.

Table 1
Descriptive statistics of the three datasets.

Dataset	Records	Entities	Concepts	Mean Records Per Entity
Oncology	76,468	207	12	369
Hepatitis	368,216	499	10	738
Diabetes	165,199	5178	4	32

4.2. *The experimental process and the consistency evaluation measures*

Our overall computational architecture and the evaluation framework defined on top of it represents a process that is similar to the KarmaLegoSification framework [38] with the addition of the TIRP validation afterwards.

The input time-stamped raw data were abstracted into a set of symbolic time intervals, using the knowledge-based temporal-abstraction method, when the knowledge was available, and several automated discretization methods (see Section 2.1). The result was a set of temporal abstractions, i.e., a set of symbolic time intervals. TIRPs were discovered using varying minimal vertical support levels, with or without using any of the SAC versions. In either case, we examined the effect of using either the abstract three temporal relations, or the full seven temporal relations (see Fig. 1). The absolute frequencies and the distributions of the discovered frequent TIRPs in the different subsets of each medical domain's patient population were then compared using the several consistency measurement methods that were described in Section 3.

To answer the three research questions, we first measured the percentage of repeating frequent TIRPs. We also measured the percentage of TIRPs whose frequencies passed the proportion test (with a significance level of $\alpha = 0.05$) and, when more than 50% of the overall number of frequent TIRPs were found in both sets, the K-S statistic value.

Note that for each domain we used a different value of minimal vertical support (the same value across all of the research questions), which was the smallest value that enabled us in that domain to store all of the domain's data and TIRPs in the memory of the computer that we had used for all of the experiments (We used an AMD Opteron™ Processor 6128 2.00 GHz Machine, with a 32.00 GB RAM). This value was a function of the overall number of TIRPs that could be discovered in each domain.

The KarmaLego algorithm was implemented based on the original publication [37]. With respect to the automated discretization methods, we used the EWD method, the SAX algorithm [30], and the TD4C based on the Kullback–Leibler divergence [39].

The results were averaged over all configuration settings, such as number of temporal relations, so as to focus on the core research questions.

5. Results

We shall now present the results of applying the three consistency measures defined by our research questions. Note that since this study is not about interpretation, classification, prediction, or clustering using TIRPs, but rather about the specific, focused question, "Are the *same* TIRPS discovered in *similar patient groups*?" we shall not extend its scope by delving into any of the specific patterns discovered, or into the interpretations provided for them by the clinicians. Furthermore, note that using the TIRPs as features for automated classification or prediction, as they were indeed effectively used in previous studies [38], does not require the clinicians to provide *any* interpretation for the discovered TIRPs.

5.1. *The oncology dataset*

Fig. 4 presents the results for validating the consistent discovery in the oncological dataset. For minimal vertical support thresholds in the range that we could examine, i.e., from 0.3 to 0.6, the overall percentage of repeating TIRPs (without using any SAC version) ranged from 74% to 78%. Thus, most of the TIRPs did repeat across different patient

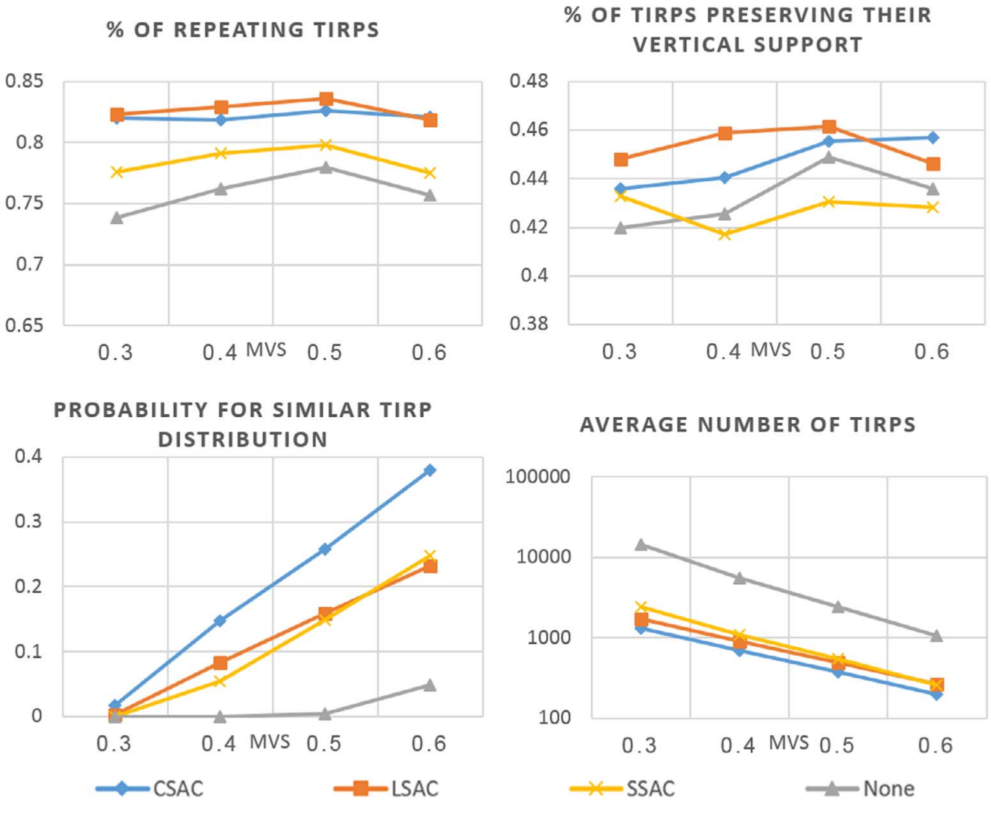


Fig. 4. The results for validating the consistent discovery of frequent TIRPs in the **oncological** dataset. The X-axis in all graphs presents the minimal vertical support. The top left graph presents the percentage of repeating (re-discoverable) frequent TIRPs. The top right graph presents the percentage of frequent TIRPs preserving their "local" vertical support level. The bottom left graph presents the K-S tests on the "global" distributions of the TIRPs. The bottom right graph presents the average number of TIRPs in a logarithmic scale. The SAC version legend for all four graphs is presented in the bottom of the diagram.

sub-groups, although definitely not all of them.

Furthermore, we can see an increasing trend in the percentage of repeating TIRPs, as the minimal support threshold increases, although the percentage of repeating TIRPs somewhat decreased at a minimal vertical support of 0.6.

When considering the use of the various SAC versions to prune the TIRPs during the discovery process, we can clearly see that using the LSAC and CSAC versions resulted in the highest percentages of repeating TIRPs (i.e., the *existence* measure), with 82–84% of frequent TIRPs found in one patient subset being found also in the other subset; and resulted also in the highest percentages of TIRPs (around 46%) that preserved the same level of vertical support without a significant change (i.e., *local* consistency). These two versions of the SAC also induced the smallest number of TIRPs [53–54], and as expected, they also discovered more consistent (repeatable) sets of TIRPs. However, roughly 42–46% of the TIRPs did preserve their vertical support across all patient sub-populations, even when not using any version of the SAC.

We then checked for *global* consistency. As can been seen in the bottom graph in Fig. 4, as the minimal vertical support threshold increased, the probability, across all experimental configuration settings, that the TIRPs distribution has indeed remained the same, i.e., was not significantly different (using the K-S statistic), has *increased*, at least when using any of the SAC version. Note that the overall TIRP repetition rate used in the calculation appears in the top left hand part of Fig. 4.Using the CSAC version was consistently preferable at every minimal support threshold level. *Not* using SAC was always worse than using any SAC version: For almost every minimal vertical support threshold level, when the SAC constraint was *not* used to discover TIRPs, the distributions of the discovered TIRPs were significantly different.

5.2. The hepatitis dataset

Fig. 5 presents the results for validating the consistent discovery in the hepatitis dataset. For minimal vertical support thresholds in the range that we could examine, i.e., from 0.6 to 0.9, the overall percentage of repeating TIRPs (*without using any SAC version*) ranged from 70% to 84%. Thus, most of the TIRPs did repeat across different patient sub-groups, although definitely not all of them.

We can see a trend of increasing *existence* consistency when using all SAC versions, as the minimal support threshold increases; in this case, we did not find that trend when not using SAC. The percentage of TIRPs that preserved their "*local*" vertical support levels was usually higher when using any version of the SAC version (and in particular, LSAC or CSAC), versus when not using it. However, roughly 40% of the TIRPs did preserve their vertical support, even when not using any version of the SAC.

Regarding the "*global*" distribution of the TIRPs, as the minimal vertical support increased, the probability that the distribution of the TIRPs has remained the same (i.e., there was no significant difference, using the K-S statistic) increased as well, whether we used the SAC or not. However, when not using any SAC version and the minimal vertical support was at the lowest (i.e., 0.6), the distributions were significantly different.

When considering the use of the SAC pruning criterion to maintain global consistency, using any SAC version was consistently preferable at almost every minimal support threshold level. In particular, it seemed that using either the CSAC or the LSAC versions resulting in discovering more consistently repeating frequent TIRPs, except at the highest (90%) level of minimal vertical support. Note, however, that at such a high level of support, the number of discovered TIRPs was very low, and the percentage was thus highly sensitive to very small absolute-number differences.

5.3. The diabetes dataset

Fig. 6 presents the results for validating the consistent discovery of TIRPs within the diabetes dataset. For minimal vertical support thresholds in the range that we could examine, i.e., from 0.025 to 0.2, the overall percentage of repeating TIRPs (without using any SAC version) ranged from 89% to 96%.

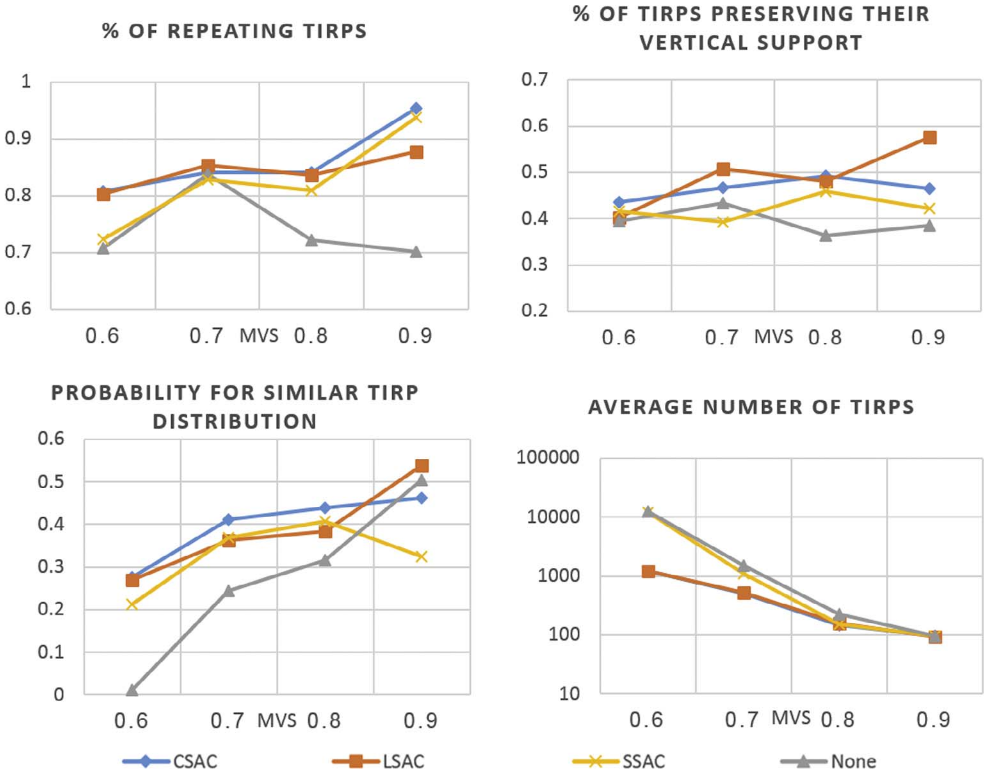


Fig. 5. The results for validating the consistent discovery of frequent TIRPs in the **hepatitis** dataset. The X-axis in all graphs presents the minimal vertical support. The top left graph presents the percentage of repeating TIRPs. The top right graph presents the percentage of TIRPs preserved their vertical support. The bottom left graph presents the K-S tests. The bottom right graph presents the average number of TIRPs in a logarithmic scale. The legend is presented on the bottom.

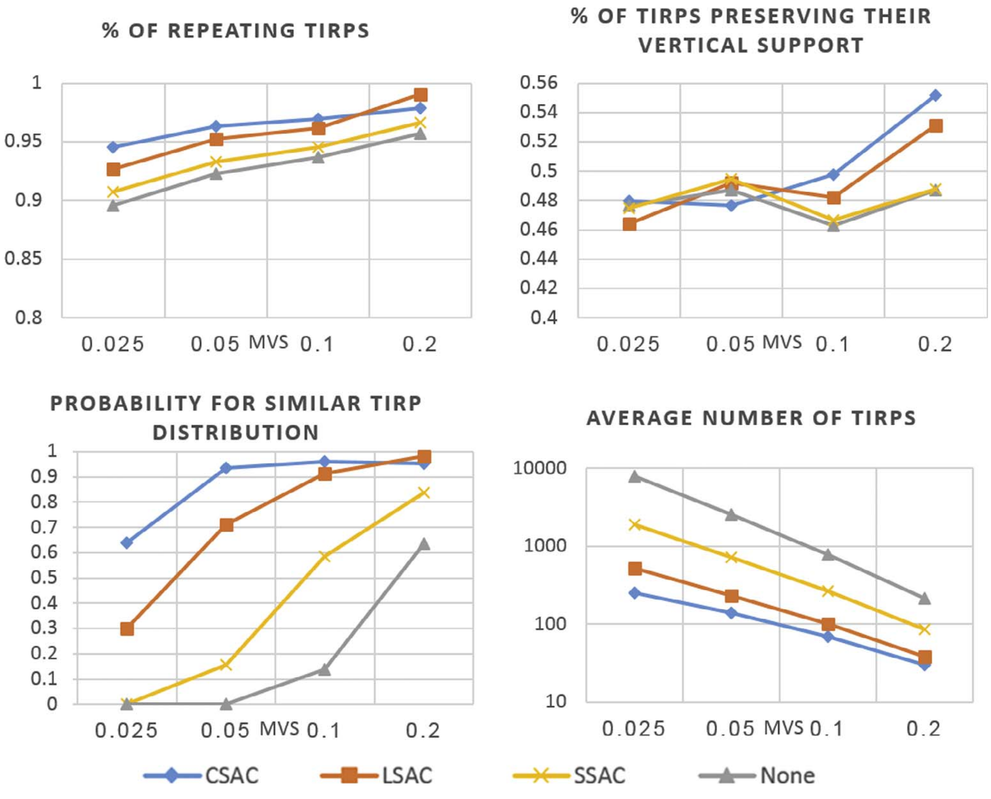


Fig. 6. The results for validating the consistent discovery of frequent TIRPs in the **diabetes** dataset. The X-axis in all graphs presents the minimal vertical support. The top left graph presents the percentage of repeating TIRPs. The top right graph presents the percentage of TIRPs preserved their vertical support. The bottom left graph presents the K-S tests. The bottom right graph presents the average number of TIRPs in a logarithmic scale. The legend is presented in the bottom.

We can see a clear trend in the *existence* consistency; as the minimal vertical support is increased, the percentage of repeating TIRPs increases as well, whether when using or when not using the SAC constraint, and the actual percentage of repeating TIRPs is always higher when using any version of the SAC constraint to discover TIRPs, especially when using CSAC or LSAC.

However, as for the percentage of TIRPs that maintained their vertical supports without a significant change, the monotonic relationship between that percentage and the increasing minimal vertical support threshold held only when using LSAC and CSAC. Using the SSAC version resulted in almost the same percentage of TIRPs who have a similar vertical support as when not using any SAC version.

Considering our *global* consistency measures, as the minimal vertical support increased, the probability that the distribution of the TIRPs has remained the same (i.e., there was no significant difference, using the K-S statistic) increased as well, whether we used the SAC or not, at least for minimal vertical-support thresholds higher than 0.05.

However, to maintain global distributional consistency of discovered TIRPs, using *any* SAC version was consistently preferable to not using it, at practically every minimal support threshold level. This was especially true when using the LSAC and CSAC versions. Indeed, when using CSAC, the distributions of the discovered TIRPs were highly similar even when using extremely low minimal vertical supports, as can be seen from the bottom graph in Fig. 6. When using the CSAC version to constrain the discovered TIRPs, even when the minimal vertical support threshold was as low as 0.05, the probability for a similar (i.e., not significantly different) TIRP distribution was higher than 90%. Only when not using SAC, or when using the SSAC version with a relatively low minimal vertical support level (5% or less), the distributions were significantly different.

6. Summary and discussion

A temporal pattern, and in particular a TIRP, typically characterizes a subset of the patient population. In fact, it can be viewed as a *temporal cluster* of patients who behave in a similar fashion [37]. Such clinical courses are often referred to as “patient journeys”, or “temporal pathways”, and can serve to characterize (e.g., clinically and economically) the population associated with that disorder. Furthermore, as explained in the Introduction, TIRPs can also be used as features for classification and prediction [6,23,27,38,39,41].

As explained in Sections 1 and 2.4, most of the studies to which our approach is relevant are similar to general TDM, in the sense that they try to discover various types of repeating temporal patterns, whether by examining functional dependencies, temporal associations, or longitudinal processes, or by using other methods. Common to all of these methods is the implicit assumption that the discovered repeating patterns *do* repeat also in other similar data sets; and in particular, can be exploited as classification or prediction features in additional instances of longitudinal records, and especially, in patient records. In the current study, we happened to use, for frequent temporal-pattern discovery, the KarmaLego algorithm. However, note that our three consistent-discovery measures can be applied to other types of patterns (not only TIRPs), including most of the patterns discovered by the methods mentioned in Section 2.4; all our proposed measures are purely logical or statistical, and assume *nothing* about the patterns' inner structure, or how they were discovered. For example, one can evaluate our methodology on an [approximate] temporal functional dependencies set [9,10,45], or on other types of patterns, such as sequential patterns, to examine whether the same patterns repeat within similar populations or sub-populations.

Thus, our current study has focused on an issue that is at the core of the whole TDM and related frameworks: To what extent are frequent interval-based temporal patterns discoverable consistently and repeatedly across different subsets of the same [patient] population?

To answer that question, we introduced three measures for *TIRP consistent discovery validation*, which tested the existence, local, and global repeatability of TIRPs across different subsets of a population. Using our consistent discovery validation methods, we tested whether similar sets of frequent TIRPs are repeatedly discovered, and in similar absolute proportions, across different subsets of three different patient populations, and measured quantitatively the similarity between the sets of frequent TIRPs in each case.

Perhaps the most encouraging findings of the study are that, within the minimal frequency ranges that we had examined, 70–95% of the discovered TIRPs were consistently discoverable; 40–48% of them maintained their local frequency. TIRP global distribution similarity varied widely, from 0% to 65%, depending on the minimal frequency discovery threshold.

The overall results, across the different datasets and configurations, showed a clear trend of an increasing *existence* consistency, regardless of using the SAC constraint to filter out TIRPs, when the minimal vertical support threshold was increased. Using the SAC principle, though, further enhanced the repeatability of the discovery.

The same relationship to the minimal frequency threshold held true also for the probability of finding a similar (i.e., insignificantly different) *overall distribution* of the TIRPs, regardless of using or not using the SAC principle. The enhancement in global consistency was mostly apparent, though, when using the CSAC version, which in most cases resulted in the highest probability of maintaining the same TIRP distribution across all patient subsets and experimental configurations.

However, we did *not* see a clear-cut, consistent trend of an increasing percentage of TIRPs preserving a similar vertical support in different subsets when the minimal vertical support was increased, although this "local" consistency was usually enhanced when using the LSAC and CSAC versions.

Thus, it seems that while the similarity in the *distribution* of discovered frequent TIRPs might well be a reliable measure, and possibly even used as a part of a characterization of a particular patient population, especially when the appropriate SAC versions are used, the *absolute frequency* of specific TIRPs might change in different patient subpopulations – in all domains, only half of the TIRPs maintained their absolute frequency.

Thus, when working in a new medical domain, or within a new context within a familiar domain, and intending to discover frequent temporal patterns for the purpose of characterizing the patients' clinical trajectories, or for use as classification and prediction features, it might be wise to start with high minimal frequency thresholds. The absolute value of such a threshold, however, might vary for each domain. Therefore, using the SAC versions might well enable researchers to produce repeatable results for considerably lower thresholds, and with much less computational effort and time.

It is also interesting to consider, *why* using the SAC versions reduces TIRP variability. One possibility is that the very removal of TIRPs that contain potentially contradicting symbolic intervals, as defined in Section 2.3, often removes, in fact, TIRPs that do *not* characterize any meaningful temporal cluster of patients, or patient behaviors. The result might well be a much smaller set of "core" [temporal] patterns that are more likely to characterize the patients' behavior over time. For example, the pattern denoting that a certain medication, given at a high dose, reduces blood pressure, "high-dose medication before Low blood pressure", would now stand out and would not be diluted by cases in which, after the *decrease* in blood pressure, there was eventually an *increase*, since the resulting spurious pattern "high-dose medication before High blood pressure" will be filtered out when using the SAC principle.

This explanation raises another issue – can there be other core [temporal] patterns hidden within longer patterns, which can serve as the common ground among several TIRPs? Indeed, there might be, and we consider these patterns as *Abstract TIRPs* that can have several extensions, or instantiations. Usually, the *a priori* principle used by most TDM algorithms, including KarmaLego, will detect these shorter core patterns and place them at a higher node of the output TIRP tree (in which the longest and least frequent TIRPs are placed in the leaves). But there might be cases in which an abstract TIRP such as "A overlaps B, and B overlaps C" might not be discovered, since any instantiation of the relation between A and C (necessary for a complete description of the extended TIRP) reduces the vertical support for the extended, more specific TIRP below the threshold necessary for its discovery as frequent in the patient population. (Note that three temporal relations [*before, meets, overlaps*] are possible, in theory, between A and C). Thus, in this case, using the "core" abstract [partial] pattern might indeed be preferable, although that pattern is technically not completely defined.

Note that in another study, we measured the classification performance by the mean *Area Under the Curve* AUC), using classifiers from four different classifier-induction families Random Forest, Naïve Bayes, SVM, and Logistic Regression), for three classification and prediction tasks defined, respectively, within each of the three medical domains explored in the current study. When discovering only SAC-obeying TIRPs, classification and prediction performance did not vary, compared to using the full, unfiltered set of discoverable TIRPs, even though the number of TIRPs discovered, and the time needed to discover them were both reduced by one to two orders of magnitude [54].

Furthermore, the SAC-obeying TIRPs are more semantically transparent to an expert clinician who wishes to understand, *which* frequent patient journeys were [repeatedly] discovered in her patient population. For example, the pattern denoting that an ACE inhibitor medication, given at a high dose, reduces blood pressure, "high-dose ACE-Inhibitor medication before Low blood pressure", is clear to a clinician, conforming to her clinical and pharmacological intuitions. But if sometimes, after the decrease in blood pressure, there was eventually an increase in that blood pressure, the additional temporal pattern "high-dose ACE Inhibitor medication before High blood pressure" (skipping the intermediate decrease in blood pressure) would be quite mystifying to a clinician, as characterizing her patients, although, syntactically, in the technical sense, it does exist.

In summary, for multiple clustering, classification, and prediction tasks, we want, on one hand, a small, compact set of TIRPs that is quickly discoverable and that we know will repeat itself with a high probability in other subsets of the patient population. On the other hand, we do not want to lose important information, and thus reduce the clustering, classification, or prediction performance. The current study suggests that frequent TIRPs usually do repeat within similar patient populations, and that using higher minimal support thresholds and some domain-specific semantic information might enhance that repetition, at least in the three medical domains that we have looked at, although examination of additional medical domains is necessary to validate our conclusions.

Conflict of Interest

Authors declare that there is no conflict of interest.

Acknowledgments

The authors wish to thank all their clinical collaborators for assisting in developing the clinical knowledge bases. A. Shknevsky and Y. Shahar were partially supported by the MobiGuide project, funded by the European Commission's 7 th Framework Program, grant #287811, and were partially funded by the Israeli Ministry of Defense.

Appendix A. Data sets and knowledge-base definitions

We describe here the data sets used in our experiments, and the definitions we used in the case of the knowledge-based temporal abstraction method.

Table 2
The knowledge base for the oncology dataset.

Platelet		HGB		WBC	
High	≥ 400	High	≥ 16	Very_High	≥ 20
Normal	100–400	Normal	11–16	High	12–20
Moderately_Low	50–100	Moderately_Low	9–11	Normal	2.5–12
Low	20–50	Low	7–9	Moderately_Low	0.5–2.5
Very_Low	< 20	Very_Low	< 7	Low	0.1–0.5
				Very_Low	< 0.1
Glucose		Total Bilirubin		Alkaline Phosphatase	
Very_High	≥ 250	Very_High	≥ 10	Very_High	≥ 225
High	151–250	High	3–10	High	110–225
Normal	75–151	Normal	1.5–3	Normal	35–110
Low	< 75	Low	< 1.5	Low	< 35
HCT		Monocytes		Lymphs	
High	≥ 46.9	High	≥ 10	High	≥ 52
Normal	34.9–46.9	Normal	3–10	Normal	18–52
Low	< 34.9	Low	< 3	Low	< 18
EOS		Bands		Basos	
Very_High	≥ 12	High	> =6	High	> =3
High	6–12	Normal	< 6	Normal	< 3
Normal	< 6				

Table 3
The knowledge base for the hepatitis dataset.

GOT		GPT		LDH	
High	≥ 40	High	≥ 40	High	≥ 450
Normal	7–40	Normal	7–40	Normal	216–450
Low	< 7	Low	< 7	Low	< 216
TP		ALP		ALB	
High	≥ 8.2	High	≥ 206	High	≥ 5.1
Normal	6.5–8.2	Normal	72–206	Normal	3.9–5.1
Low	< 6.5	Low	< 72	Low	< 3.9
UA		T-BIL		I-BIL	
High	≥ 8	High	≥ 1.2	High	≥ 0.9
Normal	2.5–8	Normal	0.2–1.2	Normal	0.2–0.9
Low	< 2.5	Low	< 0.2	Low	< 0.2
D-BIL					
High	≥ 3				
Normal	< 3				

A.1. The oncology dataset

The data used for the evaluation was of anonymous bone-marrow transplantation patients who were followed for 2–4 years at the Rush Medical Center, Chicago, USA, during the early 1990 s. The data were already abstracted into symbolic intervals as part of the KBTA studies [48,51]. The oncology knowledge data base used to perform the abstraction was specific to the bone-marrow transplantation domain, and included in total more than 350 concepts: more than 200 raw concepts (e.g. Glucose) and internal events (e.g., bone marrow transplantations).

Only a small portion of the data and knowledge were used for the purposes of this study (see the relevant knowledge in Table 2). Note that in case of an overlap in value ranges, the maximum value was used.

We used the records of 207 patients who had a bone marrow transplantation and data for the following 12 laboratory tests: Platelet count (PLETALET), Neutrophilic band forms (BANDS), Basophil granulocyte count (BASOS), Eosinophil granulocyte count (EOS), Hematocrit (HCT), Hemoglobin (HGB), Lymphocytes (LYMPHS), Monocytes (MONOS), Alkaline Phosphatase (ALK_PHOS), Total Bilirubin (T_BILI), White Blood Cell count (WBC), and Glucose levels (GLUCOSE). We used 7 days as the maximum gap between intervals when extracting TIRPs.

A.2. The hepatitis dataset

The hepatitis dataset contains the results of laboratory examinations on hepatitis B and C patients who were admitted to Chiba University Hospital in Japan. Hepatitis A, B, and C are viral infections that affect the liver of the patient. Hepatitis B and C chronically inflame the hepatocyte,

Table 4
The knowledge base for the diabetes dataset.

Albuminuria ACR					
Female	Macro	> 300	Male	Macro	> 300
	Micro	30–300		Micro	30–300
	Normo-High	15–30		Normo-High	13–30
	Normo-Low	0–15		Normo-Low	0–13
Albuminuria U24h					
Female	Macro	> 300	Male	Macro	> 300
	Micro	30–300		Micro	30–300
	Normo-High	15–30		Normo-High	13–30
	Normo-Low	0–15		Normo-Low	0–13
CREATININE					
Female	Very_High	> 4	Male	Very_High	> 4
	High	2–4		High	2–4
	Moderately_High	1–2		Moderately_High	1.2–2
	Normal	< 1		Normal	< 1.2
HBA1C					
Very_High	> 10.5				
High	9–10.5				
Moderately_High	7–9				
Normal	< 7				

whereas hepatitis A acutely inflames it. Hepatitis B and C are especially important because they have a potential risk for developing liver cirrhosis or hepatocarcinoma. The dataset contains long time-series data of laboratory examinations. The subjects are 771 patients with hepatitis B and C who were examined between 1982 and 2001. The relevant knowledge was extracted from a public KDD conference challenge [20] and was used for our evaluation (see Table 3; In case of an overlap the maximum value was taken. The data set is publicly available [12].

We used only 499 patients who had a biopsy result of hepatitis B (204 patients) or C (295 patients) and the ten most frequent tests (occurring in most of the patients), including: Glutamic-Oxaloacetic Transminase (GOT), Glutamic-Pyruvic Transminase (GPT), Lactate DeHydrogenase (LDH), Total Protein (TP), ALkaline Phosphatase (ALP), Albumin (ALB), Uric Acid (UA), Total BILirubin (T-BIL), Direct BILirubin (D-BIL), and Indirect BILirubin (I-BIL). We used 4 months as the maximum gap between intervals when extracting TIRPs.

A.3. The diabetes dataset

The diabetes data set was provided as part of a joint study with our local academic medical center [17]. The subjects included 26,000 anonymous patients (and about 12 million raw data records) who had diabetes and various laboratory tests between 2004 and 2008. The data include static information (e.g., gender) and temporal records (e.g., High-density lipoprotein, Low-density lipoprotein, Triglycerides, Glucose, Hemoglobin A1c, Creatinine, Total cholesterol, and Albuminuria). The main interest in this data was on the investigation of factors associated with changes in renal function (mostly focusing on the level of albuminuria, or secretion of protein in the urine), exploring its predictive risk factors.

We used 5178 patients who had Albumin-to-creatinine ratio or Albumin-24 h from urine in the last, fifth, year of the data set, and who also had these tests and also Glycosylated hemoglobin HbA1c), and Creatinine CREATININE) in the first four years of the data set. The task was to predict Albuminuria-normo 3231 patients) versus micro or above normal albuminuria 1947 patients) in the fifth year based on TIRPs discovered in the first four years. The relevant knowledge was supplied by our collaborator [17] and other clinicians who worked on other projects as well (see Table 4). We used 4 years as the maximum gap between intervals when extracting TIRPs

References

[1] K.P. Adlassnig, C. Combi, A.K. Das, E.T. Keravnou, G. Pozzi, Temporal representation and reasoning in medicine: research directions and challenges, Artif. Intell. Med. 38 (2) (2006) 101–113.

[2] J.F. Allen, Maintaining knowledge about temporal intervals, Commun. ACM 26 (11) (1983) 832–843.

[3] R. Azulay, R. Moskovitch, D. Stopel, M. Verduijn, E. De Jonge, Y. Shahar, Temporal discretization of medical time series – a comparative study, in: Proceedings of the 11th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology IDAMAP, 2007.

[4] I. Batal, D. Fradkin, J. Harrison, F. Moerchen, M. Hauskrecht, Mining recent temporal patterns for event detection in multivariate time series data, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '12, 2012, pp. 280–288.

[5] I. Batal, L. Sacchi, R. Bellazzi, Multivariate time series classification with temporal abstractions, in: Proceedings of the Twenty-Second International FLAIRS Conference, 2009, pp. 344–349.

[6] I. Batal, H. Valizadegan, G.F. Cooper, M. Hauskrecht, A temporal pattern mining approach for classifying electronic health record data, in: ACM Transaction on Intelligent Systems and Technology (ACM TIST), (Special Issue on Health Informatics), 2012b.

[7] M. Berlingerio, F. Bonchi, F. Giannotti, F. Turini, Mining clinical data with a temporal dimension: a case study, in: Proceedings - 2007 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2007, 2007, pp. 429–436.

[8] P. Bille, A survey on tree edit distance and related problems, Theoret. Comput. Sci. 337 (1–3) (2005) 217–239.

[9] C. Combi, M. Mantovani, A. Sabaini, P. Sala, F. Amaddeo, U. Moretti, G. Pozzi, Mining approximate temporal functional dependencies with pure temporal grouping in clinical databases, Comput. Biol. Med. 62 (2015) 306–324.

[10] C. Combi, P. Sala, Mining approximate interval-based temporal dependencies, Acta Informatica 53 (6–8) (2016) 547–585.

[11] A. Dagliati, L. Sacchi, C. Cerra, P. Leporati, P. De Cata, L. Chiovato, J.H. Holmes, R. Bellazzi, Temporal data mining and process mining techniques to identify cardiovascular risk-associated clinical pathways in Type 2 diabetes patients, in: IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2014, pp. 240–243.

[12] ["ECML/PKDD 2002 Discovery Challenge," 2002], 2002. Retrieved from < http://lisp.vse.cz/challenge/ecmlpkdd2002/ > .

[13] D. Fradkin, F. Mörchen, Mining sequential patterns for classification, Knowl. Inf. Syst. 45 (3) (2015) 731–749.

[14] S. Garcıa, J. Luengo, J.A. Saez, V. Lopez, F. Herrera, A Survey of discretization techniques : taxonomy and empirical analysis in supervised learning, IEEE Trans. Knowl. Data Eng. 25 (4) (2013) 734–750.
[15] T. Gartner, A Survey of kernels for structured data, ACM SIGKDD Explor. Newsl. 5 (1) (2003) 49–58.
[16] A. Goldstein, Y. Shahar, An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data, J. Biomed. Inform. 61 (2016) 159–175.
[17] M. Gordon, Development and Implementation of Computational Methodologies for a Systems Level Analysis of Bio-Medical Data, 2012.
[18] J. Guilford, B. Fruchter, Fundamental Statistics in Psychology and Education, McGrew-Hill Book Company, 1978.
[19] S. Hirano, S. Tsumoto, Mining similar temporal patterns in long time-series data and its application to medicine, in: 2002 IEEE International Conference on Data Mining, 2002, pp. 219–226.
[20] T.B. Ho, T.D. Nguyen, Mining hepatitis data with temporal abstraction, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 369–377.
[21] M. Hollander, D.A. Wolfe, E. Chicken, Nonparametric Statistical Methods. Nonparametric Statistical Methods vol. 7, John Wiley & Sons, 2013.
[22] F. Höppner, Learning temporal rules from state sequences, in: IJCAI Workshop on Learning from Temporal and Spatial Data, 25, 2001.
[23] F. Höppner, S. Peter, M.R. Berthold, Enriching multivariate temporal patterns with context information to support classification, Comput. Intell. Intell. Data Anal. 195–206 (2013).
[24] G. Hripcsak, Physics of the medical record: handling time in health record studies, in: Artificial Intelligence in Medicine (AIME), Pavia, Italy, 2015.
[25] G. Hripcsak, D.J. Albers, A. Perotte, Exploiting time in electronic health record correlations, J. Am. Med. Informat. Assoc. 18 (1) (2011) 109–115.
[26] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (6) (2012) 395–405.
[27] D. Klimov, A. Shknevsky, Y. Shahar, Exploration of patterns predicting renal damage in diabetes type II patients using a visual temporal analysis laboratory, J. Am. Med. Inform. Assoc. 22 (2) (2015) 275–289.
[28] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[29] S. Laxman, P.S. Sastry, A survey of temporal data mining, Sadhana 31 (2) (2006) 173–198.
[30] J. Lin, E. Keogh, S. Lonardi, B. Chiu, A symbolic representation of time series, with implications for streaming algorithms, in: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery - DMKD '03, ACM Press, New York, USA, 2003, pp. 2–11.
[31] S.B. Martins, Y. Shahar, D. Goren-Bar, M. Galperin, H. Kaizer, L.V. Basso, D. McNaughton, M.K. Goldstein, Evaluation of an architecture for intelligent query and exploration of time-oriented clinical data, Artif. Intell. Med. 43 (1) (2008) 17–34.
[32] T. Mitchel, Machine Learning, McGraw Hill, 1997.
[33] F. Mörchen, D. Fradkin, Robust mining of time intervals with semi-interval partial order patterns, Sdm 315–326 (2010).
[34] F. Mörchen, A. Ultsch, Optimizing time series discretization for knowledge discovery, in: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM Press, New York, USA, 2005, pp. 660–665.
[35] R. Moskovitch, N. Peek, Y. Shahar, Classification of ICU patients via temporal abstraction and temporal patterns mining, in: IDAMAP, Verona, Italy, 2009.
[36] R. Moskovitch, Y. Shahar, Medical temporal-knowledge discovery via temporal abstraction, in: AMIA 2009 Symposium Proceedings, 2009, pp. 452–456.
[37] R. Moskovitch, Y. Shahar, Fast time intervals mining using the transitivity of temporal relations, Knowl. Inf. Syst. 42 (1) (2015) 21–48.
[38] R. Moskovitch, Y. Shahar, Classification of multivariate time series via temporal abstraction and time intervals mining, Knowl. Inf. Syst. 45 (1) (2015) 35–74.
[39] R. Moskovitch, Y. Shahar, Classification-driven temporal discretization of multivariate time series, Data Min. Knowl. Disc. 29 (4) (2015) 871–913.
[40] P. Papapetrou, G. Kollios, S. Sclaroff, discovering frequent arrangements of temporal intervals, in: Fifth IEEE International Conference on Data Mining (ICDM'05), 2005.
[41] D. Patel, W. Hsu, M.L. Lee, Mining relationships among interval-based events for classification, in: Proceedings of the ACM SIGMOD international conference on Management of data, 2008, pp. 393–404.
[42] E. Pearson, H. Hartley, Biometrika Tables for Statisticians vol. 2, Cambridge University Press, 1972.
[43] L. Sacchi, A. Dagliati, R. Bellazzi, Analyzing complex patients' temporal histories: new frontiers in temporal data mining, Data Min. Clin. Med. 89–105 (2015).
[44] L. Sacchi, C. Larizza, C. Combi, R. Bellazzi, Data mining with temporal abstractions: learning rules from time series, Data Min. Knowl. Disc. 15 (2) (2007) 217–247.
[45] P. Sala, C. Combi, M. Cuccato, A. Galvani, A. Sabaini, A framework for mining evolution rules and its application to the clinical domain, in: Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015, 2015, pp. 293–302.
[46] D. Sengupta, P.K. Naik, SN algorithm: analysis of temporal clinical data for mining periodic patterns and impending augury, J. Clin. Bioinformatics 3 (1) (2013) 24.
[47] A. Shabtai, Y. Fledel, Y. Elovici, Y. Shahar, Using the KBTA method for inferring computer and network security alerts from time-stamped, raw system metrics, J. Comput. Virol. 6 (2010) 239–259.
[48] Y. Shahar, A framework for knowledge-based temporal abstraction, Artif. Intell. 90 (1–2) (1997) 79–133.
[49] Y. Shahar, D. Goren-bar, D. Boaz, G. Tahan, Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions, Artif. Intell. Med. 38 (2) (2006) 115–135.
[50] Y. Shahar, M.A. Musen, RÉSUMÉ: a temporal-abstraction system for patient monitoring, Comput. Biomed. Res. 26 (3) (1993) 255–273.
[51] Y. Shahar, M.A. Musen, Knowledge-based temporal abstraction in clinical domains, Artif. Intell. Med. 8 (3) (1996) 287–298.
[52] T. Shimshon, R. Moskovitch, L. Rokach, Y. Elovici, Clustering di-graphs for continuously verifying users according to their typing patterns, in: 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel, 2010, pp. 445–449.
[53] A. Shknevsky, R. Moskovitch, Y. Shahar, Semantic considerations in time-intervals mining, in: Proceedings of ACM SIGKDD workshop on Connected Health at Big Data Era (BigCHat2014), New York, USA, 2014.
[54] A. Shknevsky, Y. Shahar, R. Moskovitch, The Semantic Adjacency Criterion in Time Intervals Mining. SISE-TechReport-2017-24410, 2017. Retrieved from < http://www.ise.bgu.ac.il/engineering/ShowMore.aspx?id = 24410 > .
[55] M. Stacey, C. McGregor, Temporal abstraction in intelligent clinical data analysis: a survey, Artif. Intell. Med. 39 (1) (2007) 1–24.
[56] D. Stopel, Z. Boger, R. Moskovitch, Y. Shahar, Y. Elovici, Improving worm detection with artificial neural networks through feature selection and temporal analysis techniques, Int. J. Comput. Sci. Eng. 15 (2006) 202–208.
[57] M. Verduijn, L. Sacchi, N. Peek, R. Bellazzi, E. de Jonge, B.A.J.M. de Mol, Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data, Artif. Intell. Med. 41 (1) (2007) 1–12.
[58] E. Winarko, J.F. Roddick, Discovering richer temporal association rules from interval-based data, Data Warehousing and Knowledge Discovery, vol. 3589, 2005, pp. 315–325.