

Temporal Data Mining

Andrew R. Post, MD, PhD^{a,*},
James H. Harrison, Jr, MD, PhD^b

^a*Division of Clinical Informatics, Department of Public Health Sciences,
University of Virginia, Suite 3181 West Complex, 1335 Hospital Drive, Charlottesville,
VA 22908-0717, USA*

^b*Division of Clinical Informatics, Department of Public Health Sciences and Pathology,
University of Virginia, Suite 3181 West Complex, 1335 Hospital Drive, Charlottesville,
VA 22908, USA*

Clinical data repositories provide a population-based view of phenotypic responses associated with disease presentation and progression and response to therapy. Much of the clinical data in these databases, and particularly laboratory data, are in the form of time-stamped data elements whose temporal relationships may carry significant clinical meaning [1]. For example, the recent redefinition of myocardial infarct includes an increase followed by a decrease in cardiac injury markers according to a characteristic time course, in conjunction with EKG changes and symptoms [2]. The diagnosis of viral hepatitis requires evaluation of temporal relationships in the expression of multiple antibodies and viral antigens [3]. Depending on its temporal relationship with other data, an elevated blood digoxin level could mean overtreatment, but it could also be a signal of declining renal function (progression of disease), an inappropriately timed blood draw or drug dose (a medical process problem), or a drug–drug interaction with quinidine (effectively another medical process problem). The particular arrangement of data elements in time thus conveys clinical meaning that would be lost if the same set of values were randomly ordered. For this reason, most clinical databases can be considered time sequence databases, and data mining methods must take these temporal relationships into account to yield clinically meaningful results.

Conventional data mining methods treat time sequences as unrelated aggregates of individual data elements. Although these techniques can identify associations between individual data values, they cannot identify frequently

* Corresponding author.

E-mail address: arp4m@virginia.edu (A.R. Post).

recurring subsequences with a particular profile of values (eg, trends) or detect sequences with similar profiles that occur over different time scales or have different baseline values. This limits the applicability of data mining methods in clinical laboratories and other settings in which time sequences are important. To make the features of time sequences accessible to data mining, methods are needed that treat time sequences as distinct entities and identify related subsequences within longer time sequences. Temporal data mining seeks to extend conventional data mining methods to incorporate recognition of these temporal features.

Representation of time in clinical information

Temporal relationships are inherent in the accurate expression of clinical histories, therapeutic procedures, and therapeutic outcomes. The importance of time representation in electronic clinical records was explicitly addressed more than 30 years ago in the Time-Oriented Database (TOD) model, which expresses the time of occurrence of clinical observations and events with a timestamp [4]. Standard relational databases support this model, and Structured Query Language (SQL) [5] allows basic comparisons of these timestamps in database queries. The TOD model does not identify time sequences as distinct objects, nor does it support segmentation of time sequences into subsequences, thus limiting its ability to support querying for specific profiles within time series.

An alternative representation of time expresses time sequences as intervals with start and end times, over which a feature or state holds. This representation is believed by many to be more compatible than pure timestamps with typical clinical reasoning about time [1,6]. Intervals may contain a sequence of time-stamped data elements and thus represent an interpretation of those data elements (Fig. 1), or they may represent a period of risk or a disease state that is inferred from a sequence of data elements but is outside of the temporal span of those elements. Extensions to SQL have been proposed that support creation and storage of interval data types that can represent, for example, disease and patient care processes [7]. These extensions also support query for temporal relationships involving the endpoints of intervals or individual timestamps, with relationships specified using a temporal reasoning language based on, for example, Allen's 13 interval relationships (eg, before, after; Fig. 2) [8], or Dechter and colleagues [9] system of numerical relationships (eg, before by at least 2 days). Efforts to incorporate these changes into the SQL standard have thus far been unsuccessful [10].

The proposed SQL extensions do not provide explicit support for computing intervals from data subsequences, but separate development efforts specifically targeting clinical data have produced sophisticated temporal processing systems for relational databases that provide this capability [11–14]. These systems incorporate time sequence processing that allows recognition of intervals representing, for example, states and trends, within

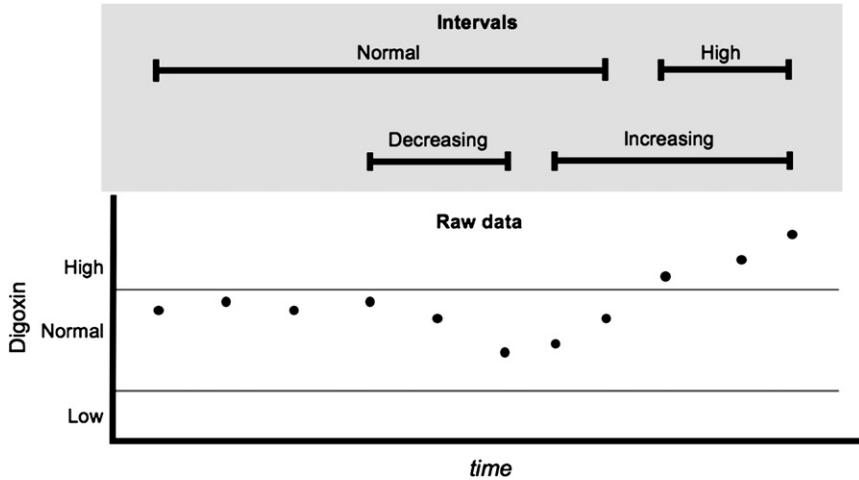


Fig. 1. State intervals of digoxin drug levels. A patient's serum digoxin measurements were used to infer periods of decreasing and increasing values. A period of normal levels can be inferred from the first eight values because they are all within the normal range (between the thin horizontal lines). The last three values are above the upper limit of normal and thus indicate that the patient had toxic levels during that time period. Intervals like these can be computed by temporal abstraction (see Fig. 4).

sequences of clinical data, and can allow recognition of combinations of those intervals across multiple data types [15,16]. The scope of application of these interval identification systems has been limited, however, because the expert knowledge required to specify detailed criteria for clinically relevant states, trends, and other patterns is difficult to marshal [6].

Both within and outside of medicine, there has been a growing recognition of the need for specialized techniques for mining temporal data [17] to identify anomalies, detect change, and predict future events [18]. These techniques are relatively well developed in the business and engineering communities [19], and have been used in a wide range of applications,

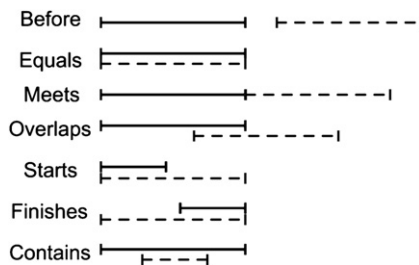


Fig. 2. Temporal relationships between intervals defined in Allen's temporal logic [8]. There are seven basic relationships and their inverses (not shown). In the case of Equals, the basic and inverse relationships are identical.

from the discovery of patterns in stock market data [20] to the identification of anomalies in space shuttle telemetry [21]. The use of temporal data mining with clinical time sequences has emerged more recently [22] because increasing volumes of patient data are being stored electronically [23]. Some kinds of clinical data (electrocardiogram [21] and intensive care unit [24] data, in particular) have been found to have properties resembling these business and engineering data sources and are thus amenable to similar types of analyses.

Temporal data mining techniques in the medical domain [18] are generally designed either for exploration or for prediction. Exploratory methods involve processing a database to identify groups of time series with similar combinations of frequent intervals (see Fig. 1) and temporal relationships (see Fig. 2). Clinical domain knowledge may be applied to these groups or clusters to determine if they represent useful or previously unknown relationships between data types. Predictive techniques may target a diagnosis, therapeutic response, or other clinical or patient care process, and search for combinations of intervals that frequently occur with some temporal relationship to the target. These methods vary in the temporal features of time sequences that they incorporate into pattern discovery and the ease with which they can be configured for use with a particular data set. All of them also depend on effective techniques for measuring the similarity of time series.

Time series similarity measures

The distance metrics used by conventional data mining techniques to determine similarity between data elements can be applied to detecting similarity in some clinical time series. The most commonly used measure for this purpose is the Euclidean distance (Table 1) [25], which computes the square root of the sum-squared differences between sequential pairs of points in two series. Euclidean distance requires each time series to have the same number of values, or new values must be interpolated into the time series to equalize their lengths. Although interpolation preserves the general profile of a time sequence, it distorts the speed with which the profile changes value (ie, the local slope of the graph) and assumes that additional values are accurately predictable from existing data. Euclidean distance is not sensitive to data order, hence randomly shuffling the elements of two time sequences would not change their Euclidean distance, and the difference between time series with significant variation tends to be underestimated. Euclidean distance also disregards the durations of the time spans between data elements, is sensitive to outliers, and poorly estimates the difference between short time series. Results are thus usually unsatisfactory for clinically relevant comparisons between medical time series.

Several distance measures improve on Euclidean distance by correlating successive values of a pair of time series, thus taking into account the

Table 1
Representative methods for evaluating time series similarity

Name	Strategy	Application ^a
Euclidean distance	Evaluates the total distance between data elements	Time series with equal length where data order and spacing are not important
Pearson correlation distance	Evaluates the similarity between profiles of data sequences	Time series with equal length where data order is important but spacing is not
Dynamic time warping	Evaluates similarity after stretching and/or compressing portions of the time axis of one series to achieve best alignment	Time series with possibly different lengths where the spacing of data and duration of temporal features are not important
Fourier transform	Decomposes time series into sets of sine waves of varying frequency and amplitude, which are compared	Time series where the lengths are equal, data are regularly spaced, and temporal features are periodic
Discrete wavelet transform	Decomposes time series into wavelet coefficients determined at multiple time scales, which are compared	Time series where the lengths are equal and data are regularly spaced; temporal features may be irregular

^a Length refers to the number of elements in the series rather than its total duration.

relative order of the data elements. The Pearson correlation distance (see Table 1) [26] recognizes similarities in the shapes of two sequences, as long as those sequences are the same length, and can also capture inverse relationships (an increase in the values of one time sequence with a concurrent decrease in those of another). For data sets in which time sequences all have the same number of data elements, the time spans between data elements are unimportant, and comparisons of entire sequences are of interest, this distance metric can produce satisfactory results. Because routine medical observations are not made with the intent of cross-population comparison, however, medical time series are often composed of data elements of varying numbers that are spaced irregularly and differently across episodes and patients. Traditional metrics, such as Pearson, may classify incorrectly profiles that are similar but are sampled at different times. If the durations of the time spans between data elements are important, the number of data elements in each sequence is variable, or temporal relationships between subsequences (eg, trends and periodicity) are important, more sophisticated methods are needed to incorporate these temporal relationships into similarity measures.

Dynamic time warping

Dynamic time warping (see Table 1) is a robust distance calculation method developed for speech recognition [27]. It computes the distance

between two time series after stretching and compressing subsequences of one of the time series along the time axis until an alignment between the two series is found that minimizes their distance. Time warping works well for computing the similarity of time sequences with different numbers of data elements and irregular data spacing. It does not explicitly take the time spans between data elements into account, however, and thus it is best used when duration constraints are unimportant. In the clinical domain, it has been successfully applied to detection of anomalies in the serum creatinine values of kidney transplant patients [28], demonstrating its ability to find similarities in laboratory test result trends in patients who have different baseline values.

Transform-based methods

Transform-based techniques apply mathematic functions to time series that output an unordered list of values encapsulating the temporal characteristics of the data. The output of these functions can generally be input into a reverse transform that reproduces the original time sequence. Parameters for these functions control the degree to which the output faithfully reproduces every aspect of the original time sequence, and can be used to smooth noisy time sequences and reduce the impact of outliers on the output. The time spans between data elements are incorporated into these functions, unlike the distance metrics above. Once these lists of shape characteristics are assembled for each time series in the database, they can be compared for similarity using conventional distance measures such as those described earlier.

The Fourier transform (see Table 1) [29] characterizes the shape of a time sequence as a collection of sinusoidal waves with different amplitudes and phases as a function of frequency, the sum or integral of which reproduces the original time sequence. It assumes that the entire length of a time sequence can be represented by the same set of wave functions, an assumption that may not hold for time sequences reflecting multiple clinical processes occurring sequentially. Fourier transform techniques also assume regularly spaced data elements and require the number of data elements in each time series to be a power of two. These requirements often necessitate padding the data, which can be achieved through interpolation. The Fourier transform is best suited to characterizing data sequences that are periodic, such as diurnal variation in serum glucose levels [30].

A similar approach called the discrete wavelet transform (see Table 1) [31] localizes the wave functions describing a time sequence to periods of time, thus relaxing the Fourier transform's assumption of a single description of the entire time series, and supports repeating the transform at multiple time scales. A member of a class of wave functions called a wavelet, chosen empirically, is stretched and repeated across a time sequence at one or more time scales, transforming the sequence into a set of wavelet

coefficients. The coefficients represent progressively finer levels of detail associated with their corresponding time scales. The wavelet transform can thus represent the general shape of a time sequence and its fine structure, effectively allowing for zooming in and out of the sequence's temporal features. It, like the Fourier transform, assumes regularly spaced data in which the number of data elements is a power of two, and interpolation can be used to satisfy these requirements. Because the functional descriptions produced by the wavelet transform are localized in time, this technique is applicable to detection of nonperiodic patterns in clinical data sequences. Clinical application areas include early detection of hemodynamic deterioration as measured by multiple physiologic variables in ICU patients [25] and early detection of infectious disease outbreaks by monitoring for spikes in emergency department chief complaints [32].

Although the transform-based methods feature significantly improved temporal pattern recognition over plain distance measures, including the ability to identify periodic patterns and similarities between complex profiles, they suffer from the need to prespecify parameters (eg, the wavelet function for the wavelet transform) that may be unintuitive for clinical experts, and, like basic similarity measures and dynamic time warping, they compare the similarity of entire time sequences. In practice, however, interesting clinical features are most often expressed as characteristic subsequences within longer sequences of data. In clinical settings comparison of entire time sequences is often undesirable and methods are needed to split time sequences into shorter, clinically meaningful subsequences.

Subsequencing methods

Sliding window methods

Time series may be directly divided into subsequences by scanning a “sliding window” that views a fixed number of data elements across each sequence, creating the set of all possible subsequences of that length (Table 2). The subsequences overlap, and thus every data element is analyzed in its local temporal context. The sliding window method is illustrated in Fig. 3. The position of a subsequence within a sequence is lost in this technique, thus it is primarily useful for identifying common motifs or combinations of motifs across a collection of time sequence data for which the specific location of a motif within each sequence does not matter. Also, because the size of the sliding window is defined as a number of data elements rather than a duration, the sliding window approach is not appropriate when the duration of a motif is a critical feature, particularly if the time sequences are not regularly spaced or have missing values.

As compared with transform-based methods, the window size parameter is relatively intuitive and the approach allows for clustering time series based on specific temporal features that can be contained by the window, even if

Table 2
Selected methods for subsequencing time series

Name	Strategy	Application
Sliding window	Breaks sequences into all possible subsequences of a specified length ^a	Time series with discrete temporal features of uniform duration, where location in the larger sequence is unimportant
Segmentation	Breaks sequences into adjacent subsequences based on trend patterns and inflection points	Time series with high frequency, regularly-spaced data that accurately represent temporal features
Temporal abstraction	Specifically identifies subsequences of interest based on defined temporal and mathematic relationships between data elements	Time series with features of interest that can be predefined; features may be complex and/or multivariate

^a Length refers to the number of elements in the series rather than its total duration.

the time series differ significantly in other respects. The window size can markedly affect what temporal features are found, however, and empirical testing may be required to pick the best window size. Also complicating the use of this technique is that time sequences from the same source tend

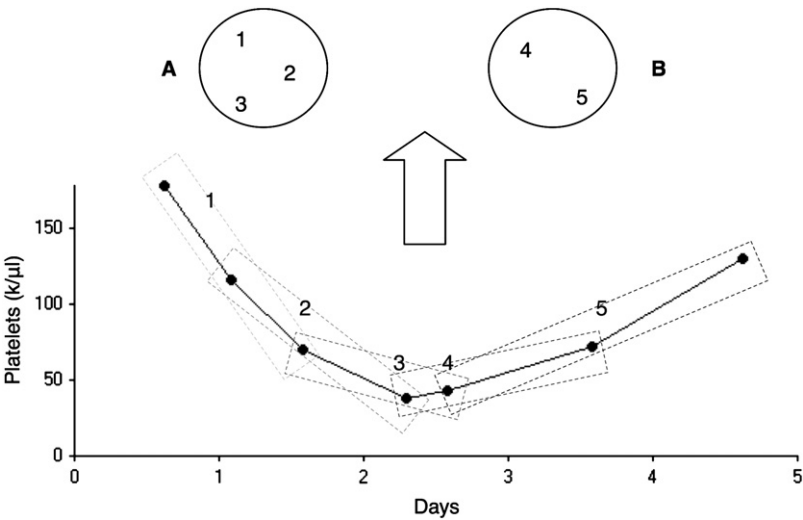


Fig. 3. Illustration of a sliding window of length 3 scanning a time sequence of platelet counts. Subsequences are indicated by dashed rectangles and are labeled by number in the order in which they are processed. The first three scanned subsequences have negative slopes, and the last two subsequences have positive slopes. A data mining algorithm might cluster the first three and last two subsequences into two clusters representing decreasing (*A*) and increasing (*B*) values.

to have spans with many similar, overlapping subsequences, which may “dilute” more distant and interesting subsequences that share similar features. Attempts to cluster time series data with sliding window methods thus may result in “meaningless” clusters if the specified temporal features of interest are too general [33].

Segmentation

The problems caused by overlapping subsequences and a fixed sliding window size can be solved by partitioning or segmenting a time series using statistical techniques into nonoverlapping subsequences with particular trend characteristics (see Table 2). Segmentation algorithms (for a brief review, see Hoppner Ref. [34]) find inflection points in the data at which the data change direction (eg, from increasing to decreasing based on the deviation from a regression line calculated across recent values). Segments do not overlap, and thus segmentation algorithms are less vulnerable to detection of trivially similar motifs than sliding window algorithms. Segmentation algorithms may allow a frequency threshold to be set below which segments are discarded, or they may scan for infrequent subsequences [21] that might reflect anomalies indicating adverse events or patient care processes that deviate from the standard of care.

Basic segmentation algorithms assume high frequency, regularly spaced data, and have been applied successfully to summarizing physiologic variables in ICU data [24] and clustering RR intervals in ECG tracings [22]. They are sensitive to missing data, however, and prone to error for data sampled at low frequency or irregular intervals. A variant of segmentation has been proposed that performs iterative smoothing (eg, by the wavelet transform described earlier) of a time sequence, finds segments in each smoothed profile, and discards segments that are found in less than a prespecified number or percentage of smoothing cycles [35,36]. This approach allows generally similar profiles to be recognized as such, even if data are spaced differently across them, or if there are outliers. An alternative technique weights segments according to the number of data values they span and discards segments below a prespecified threshold length, under the assumption that the slopes of very short segments are more sensitive to outliers and missing data and less likely to reflect meaningful patterns [22].

Temporal abstraction

The segmentation methods described earlier use statistical methods for identifying temporal features within time series subsequences and are typically limited to trend detection. In the 1990s, investigators in artificial intelligence proposed software layers built on top of clinical databases that use expert knowledge to infer clinical states or processes implicit in clinical time series data and represent them as abstractions that specify an

interval of time over which a state or process exists (see Fig. 1) [37]. These temporal abstractions may be inferred from raw time-stamped data elements based on prespecified mathematic relationships (eg, states, trends) in the data, yielding “low-level” abstractions, or from prespecified combinations of previously inferred abstractions based on temporal relationships between their intervals, yielding “high-level” abstractions. Examples of abstractions are shown in Fig. 4. Knowledge elicitation techniques have been developed to facilitate the process of encoding the expert knowledge defining abstractions in a computable form [38]. Temporal abstraction has been developed primarily in the medical domain, and has been applied successfully to pattern detection in laboratory test results [16,30,39] and children’s

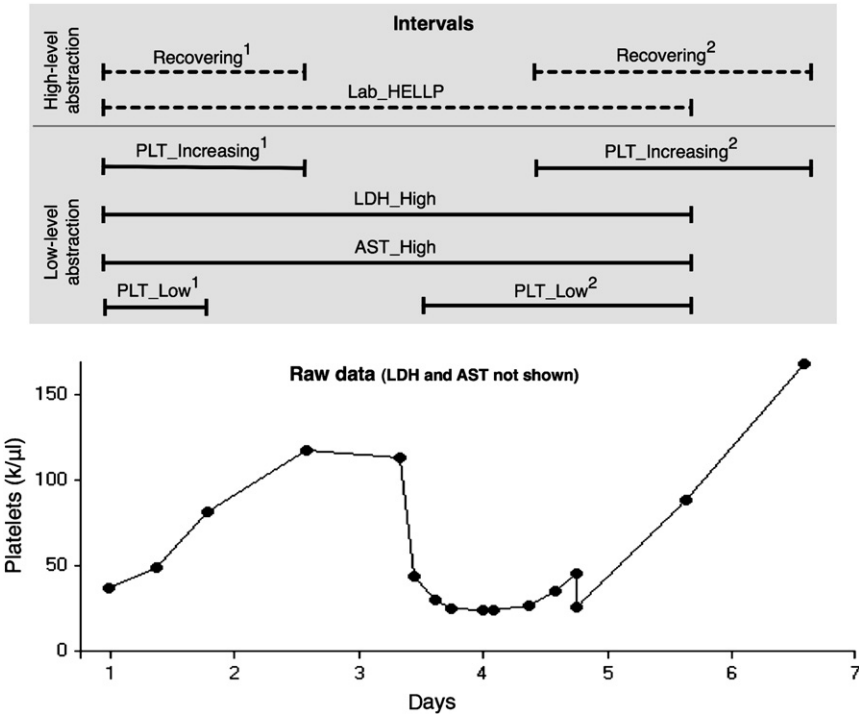


Fig. 4. Example time series of platelet (PLT) counts in HELLP (Hemolysis, Elevated Liver enzymes, Low Platelets) syndrome, and intervals identified by temporal abstraction. HELLP is a dangerous complication of pregnancy that appears during the latter part of the third trimester or after childbirth [50]. HELLP syndrome has been defined as pre-eclampsia with PLT less than 100,000/ μ l, lactate dehydrogenase (LDH) greater than 600 U/L, and aspartate aminotransferase (AST) greater than 70 U/L, and increasing PLT indicates recovery [51,52]. Subsequences of PLT, LDH, and AST values that satisfy these thresholds are labeled as PLT_Low, LDH_High, and AST_High intervals, respectively. Overlapping subsequences of low PLT, high LDH, and high AST are labeled as Lab_HELLP intervals, subsequences of increasing PLT count are labeled as PLT_Increasing intervals, and subsequences of increasing PLT count that begin at or after the start of a Lab_HELLP interval are labeled as Recovering intervals.

growth [40,41] for decision support in inpatient and outpatient settings. A recent extension to temporal abstraction adapts the technique for querying retrospective clinical data repositories for specified temporal patterns and relationships, returning populations of patients who display those patterns [42].

Low-level abstractions can resemble the trend-related subsequences returned by segmentation algorithms, and such abstraction-based segmentation (see Table 2) has been applied to analysis of physiologic data collected during hemodialysis [36]. In addition, low-level abstractions may specify and identify subsequences that satisfy independently defined and data-type-specific constraints in data values, such as whether laboratory test results are within the test's reference range (see Fig. 1). High-level abstractions can represent relationships between multiple subsequences in time series within and across data types that are useful for identifying complex states, such as disease severity and progression or response to therapy, and are particularly useful for this purpose when those states are not represented in standard coding schemes (eg, ICD-9-CM), are poorly expressed by codes, or may be inaccurately or incompletely coded in clinical databases [42]. Existing temporal abstraction methods require prespecification of high-level abstractions, which limits their applicability to clinical domains in which sufficient knowledge exists to define interesting interval relationships. Data mining techniques are available that can learn frequent interval associations that are analogous to high-level abstractions, and feed those associations back into the data mining process (see later discussion).

The low- and high-level intervals of temporal abstraction have the potential to be used in data mining, although we have been unable to find any examples of such an application. They have several advantages over intervals produced by automated segmentation. Using expert knowledge, “detection” algorithms can be designed that recognize specific patterns of interest clinically and identify outliers or unusual states that are clinically important, while ignoring uninteresting “noise”. Time sequences can be processed in multiple, mathematically distinct ways rather than using a general “one size fits all” approach for feature identification and smoothing that may not be optimal for all data types. Patterns that are clinically important could overlap within a time series, which is physiologically possible but difficult to handle with automatic segmentation. Finally, the ability to identify patterns that are highly likely to be clinically relevant as a precursor to data mining would reduce the processing requirements for mining and the volume of uninteresting patterns that must be pruned after mining. Temporal abstraction requires expert knowledge to identify the intervals to be incorporated into an interval database for mining, however, and because these intervals are not mined directly from the raw data, unanticipated low-level temporal relationships (ie, motifs) may be missed. Temporal abstraction is most appropriate as an adjunct to data mining when interesting low- and

high-level intervals can be specified and the goal is discovering novel associations and predictive relationships between those intervals.

Mining interval databases for temporal patterns

Once interesting subsequences of time series are identified by segmentation, temporal abstraction, or other methods, temporal mining strategies can assemble them into multivariate patterns that may have clinical meaning. Methods have been proposed that identify combinations of intervals that frequently co-occur in a data set, and some of these methods allow for mining combinations of intervals that frequently co-occur with some temporal relationship (see Fig. 2). The former methods are generally called association learning methods, and the latter are called temporal rule learning methods.

Association learning methods use a variant of the Apriori algorithm [43], which is an iterative process that first identifies all single instances of a particular type of interval, finds all frequent combinations, or “itemsets,” of the first interval type with a second interval type, and then finds all frequent combinations of the first two interval types with a third interval type, and so on (Fig. 5). These aggregations can either be used for discovery of frequently co-occurring intervals, or they can be thought of as predictive rules in which the presence of the interval types in an itemset predicts the presence of the interval type added to the itemset in each iteration of the algorithm.

When used for discovery of frequent combinations of interval types, Apriori has a parameter called the minimum support threshold that defines the minimum number or percentage of time sequences in which an itemset must occur for the itemset to be passed on to the next iteration of the algorithm (see Fig. 5). For mining interval types that may have multiple occurrences within a time sequence, a useful alternative definition of support is the minimum total time duration of all instances of an interval type in a sequence. For rule learning, there is an additional parameter called the minimum confidence threshold, which defines the minimum conditional probability that a predicted interval type will occur within a sequence, given an already-discovered itemset. This parameter is useful for filtering out rules that have low predictive value. Apriori-based association rule learning has been applied to assess the quality of a hemodialysis service: frequent combinations of temporal abstractions in physiologic parameters were mined and used to improve the understanding of the contribution of physiologic factors to the values of a set of quality indicators [36].

Temporal rule learning extends Apriori by storing temporal relationships between interval types as additional attributes of each itemset [35,44], using a temporal reasoning language (see later discussion) to encode relationships. Many itemsets are created for a given group of interval types, one for each temporal configuration of those interval types that is found in the dataset based on the selected support and confidence parameters. Temporal Apriori

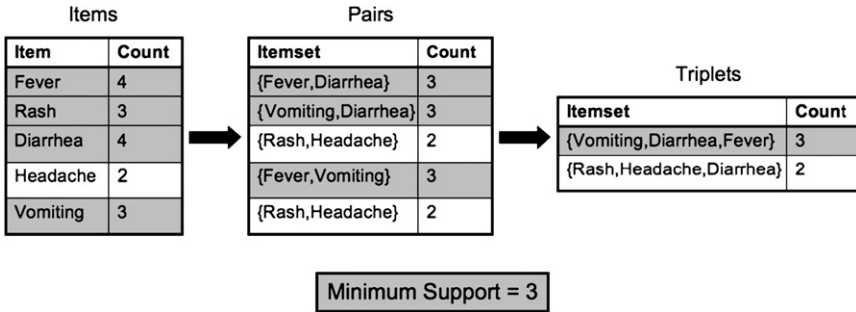


Fig. 5. Illustration of the Apriori algorithm processing a hypothetical database of clinical findings for frequent associations between findings, with a minimum support threshold of 3 occurrences in the database. In the left table (*Items*), findings and number of occurrences are listed. In the center table (*Pairs*), pairs of findings that co-occur in the same patient record are shown. In the right table (*Triplets*), combinations of three findings that co-occur in the same patient record are shown. Findings and combinations of findings that satisfy the minimum support threshold are highlighted in gray.

thus aggregates frequent interval types into frequent higher-level pattern specifications similar to high-level temporal abstractions, effectively discovering high-level abstractions with potential clinical relevance. Temporal rules are similar to the association rules described earlier, but store a temporal relationship between each interval type and the itemset to which it is added, and thus can predict not only whether an interval type will occur within a sequence, but when in that sequence it will occur. This approach has been applied in the hemodialysis domain similarly to the association discovery method described earlier [36].

Although Allen's temporal language (see Fig. 2) has been used previously for expressing temporal relationships in temporal rule learning [35,36], recent theoretic work indicates that distinguishing temporal configurations of interval types based on Allen's relationships may result in arbitrary distinctions between itemsets that are actually very similar. For example, in an appropriate context a "Starts" pattern may differ little from an "Overlaps" pattern with a small overlap at the start (see Fig. 2). An alternative language has been proposed with a smaller set of temporal relationships than Allen and greater flexibility between the boundaries of each relationship [45]. In an initial evaluation of a temporal Apriori implementation using this language, similar patterns were found as compared with an implementation of temporal Apriori using Allen's language, but with fewer itemsets that more closely matched the expectations of a domain expert.

Current status and future directions

Clinical data repositories contain records from millions of patients, and these patients' data may reflect unexpected responses to therapy, previously

unknown relationships between disease states, and deviations from standard clinical practice that may be of interest to clinicians and researchers. The patient data stored in these repositories are high dimensional, and many of the data, such as care encounters, clinical orders, laboratory test results, diagnostic and procedure codes, and records of therapy, include time sequences of variable length that are sampled irregularly and at different frequencies. These databases are of great value for better understanding biomedical and operational aspects of health care, but the data they contain are difficult to process using traditional time series analysis and data mining methods. Temporal data mining methods are under development and have been used successfully for analyzing limited subsets of clinical data repositories that are characterized by few data types and high-frequency or regularly spaced timestamps [22,24,25,28,32,36]. These methods have yet to be applied more generally, and implementations thus far have been site specific. One temporal data mining system has been reported to be in production use [36], but otherwise these systems have been implemented only in research environments.

Because clinical repositories contain a broad range of data types with different characteristics, the wider application of temporal data mining in medicine will probably require software systems that aggregate a number of features currently found in several experimental systems. In particular, systems will likely need to support multiple methods of low-level time sequence processing that identify intervals of interest. These methods will allow appropriate low-level processing approaches to be applied to different data types, with the ability to group the intervals detected across a range of data types into meaningful higher-level relationships. Such systems might use, for example, statistical time series segmentation for high-frequency data sets or those with highly characteristic inflection points, and knowledge-based detection of predefined motifs for low-frequency, irregular time sequences with specific duration constraints. Found intervals and interval groups will be stored in an interval database that may be mined for associative or predictive relationships using a variant of the Apriori algorithm. Recently, the temporal abstraction method has been incorporated into a system that supports pluggable low-level processing algorithms and aggregation of intervals found by those algorithms [42]. This type of system could support multiple data processing strategies to create a database of intervals for mining.

Adapting interval and relationship mining algorithms for use with large clinical databases requires improving their abilities to find unexpected patterns and optimizing their performance. Although existing algorithms are designed to find the most frequent patterns in a data set, unexpected patterns are likely to be frequent only within relatively small patient subsets (eg, a small group of patients all with a relationship between drug administration and laboratory test result intervals reflecting an unexpected response to a drug). Existing algorithms could be extended with methods from the

information retrieval [46] literature that weight infrequent, interesting words in documents higher than words that are frequent, where “documents” are patients, and “words” are temporal patterns found in the patients’ data [25]. Alternatively, found patterns could be filtered against a database of trivial and already-known temporal features that is populated with a combination of expert knowledge, literature values, and mined high-frequency patterns found off-line by existing algorithms, analogous to the co-mining approach [47] described in the article by Siadaty and Harrison, elsewhere in this issue.

Existing temporal relationship detection algorithms have sufficient performance for mining small, targeted subsets of a clinical data repository that are reduced in the number of patients or the number of variables, but are too computationally expensive to apply to an entire repository. Temporal relationship detection requires scanning every pair of previously found intervals within a patient [16,35]. Improving performance may require pre-processing a repository into subsets, and applying temporal data mining algorithms separately to each subset. One approach is to divide a data set along the values of application-specific variables (eg, gender, age ranges) [48], but appropriate variables have to be chosen carefully for a data mining task so that interesting patterns involving those variables are not lost. Alternatively, theoretic work in nontemporal data mining has identified a class of “expensive” data mining algorithms for which the data can be preprocessed by a “cheap” algorithm that outputs large but manageable clusters for expensive processing, with a guarantee that similar data will not be separated prematurely into different clusters by the cheap algorithm [49]. Research is needed to determine whether temporal data mining algorithms can be developed that also have this property.

Summary

Data mining of clinical time series has evolved over the past decade from simple algorithms that only consider the order of the data and not their timestamps, to sophisticated algorithms that aim to discover prediction rules. Temporal data mining offers the potential for detecting previously unknown combinations of clinical observations and events that reflect novel patient phenotypes and useful information about care delivery processes, but clinically relevant patterns of interest may occur in only a small number of patients. Large-scale processing of clinical data repositories may be needed to detect these patterns. Temporal data mining algorithms have thus far been applied to low-dimensional, homogeneous data sets. Although these experiments have yielded useful information, the major benefits of data mining will come from its application to large-scale, high-dimensional, heterogeneous data in general clinical repositories. We believe that the potential for combining interval identification methods (temporal abstraction and statistical time sequence segmentation) with interval mining methods, and recent research successfully scaling data mining algorithms to large

datasets, suggest an optimistic view that general-purpose temporal data mining software applicable to clinical data repositories is practical and likely will become an important tool for clinical research and health care quality assurance.

References

- [1] Shahar Y. Dimensions of time in illness: an objective view. *Ann Intern Med* 2000;132(1): 45–53.
- [2] Alpert JS, Thygesen K, Antman E, et al. Myocardial infarction redefined—a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. *J Am Coll Cardiol* 2000;36(3): 959–69.
- [3] Imperial JC. Natural history of chronic hepatitis B and C. *J Gastroenterol Hepatol* 1999; 14(Suppl):S1–5.
- [4] Wiederhold G, Fries JF. Structured organization of clinical data bases. Proceedings of the American Federation of Information Processing Societies National Computer Conference (AFIPS) 1975;44:479–85.
- [5] Elmasri R, Navathe SB. Fundamentals of database systems. 3rd edition. New York: Addison-Wesley; 2000.
- [6] Keravnou ET, Shahar Y. Temporal reasoning in medicine. In: Fisher M, Gabbay D, Vila L, editors. Handbook of temporal reasoning in artificial intelligence. New York: Elsevier; 2005. p. 587–653.
- [7] Snodgrass R, Bohlen MH, Jensen CS, et al. In: Etzion O, Jajodia S, Sripada S, editors. Temporal databases: research and practice. Transitioning temporal support in TSQL2 to SQL3, vol. 1399. Berlin: Springer; 1998. p. 150–94.
- [8] Allen JF. Maintaining knowledge about temporal intervals. *Commun ACM* 1983;26(11): 832–43.
- [9] Dechter R, Meiri I, Pearl J. Temporal constraint networks. *Artif Intell* 1991;49:61–95.
- [10] Adlassnig KP, Combi C, Das AK, et al. Temporal representation and reasoning in medicine: research directions and challenges. *Artif Intell Med* 2006;38(2):101–13.
- [11] Dorda W, Gall W, Duftschmid G. Clinical data retrieval: 25 years of temporal query management at the University of Vienna Medical School. *Methods Inf Med* 2002;41(2): 89–97.
- [12] Nigrin DJ, Kohane IS. Temporal expressiveness in querying a time-stamp-based clinical database. *J Am Med Inform Assoc* 2000;7(2):152–63.
- [13] O'Connor MJ, Tu SW, Musen MA. The Chronus II temporal database mediator. *Proc AMIA Symp* 2002;567–71.
- [14] Spokoyny A, Shahar Y. A knowledge-based time-oriented active database approach for intelligent abstraction, querying and continuous monitoring of clinical data. *Medinfo* 2004; 11(Pt 1):84–8.
- [15] O'Connor MJ, Grosso WE, Tu SW, et al. RASTA: a distributed temporal abstraction system to facilitate knowledge-driven monitoring of clinical databases. *Medinfo* 2001;10(Pt 1): 508–12.
- [16] Shahar Y. A framework for knowledge-based temporal abstraction. *Artif Intell* 1997;90: 79–133.
- [17] Roddick JF, Spiliopoulou M. A survey of temporal knowledge discovery paradigms and methods. *Knowledge and Data Engineering, IEEE Transactions on* 2002;14(4):750–67.
- [18] Roddick JF, Fule P, Warwick JG. Exploratory medical knowledge discovery: experiences and issues. *SIGKDD Explor. Newsl* 2003;5(1):94–9.

- [19] Antunes CM, Oliveira AL. Temporal data mining: an overview. Paper presented at the Proceedings of the Knowledge Discovery and Data Mining Workshop on Temporal Data Mining (KDD 01); Aug 26–29, 2001. San Francisco.
- [20] Fu T-c, Chung F-l, Luk R, et al. Preventing meaningless stock time series pattern discovery by changing perceptually important point detection. *Fuzzy Systems and Knowledge Discovery* 2005;1171–4.
- [21] Keogh E, Lin J, Fu A. HOT SAX: finding the most unusual time series subsequence: algorithms and applications. Paper presented at the 5th IEEE International Conference on Data Mining. New Orleans (LA), November 27–30, 2005.
- [22] Keogh E, Pazzani M. An enhanced representation of time series which allows fast and accurate classification, clustering, and relevance feedback. AAAI Press; Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. 1998. p. 239–41.
- [23] Haux R. Health information systems—past, present, future. *Int J Med Inform* 2006;75(3–4): 268–81.
- [24] Li J, Leong TY. Using linear regression functions to abstract high-frequency data in medicine. *Proc AMIA Symp* 2000;492–6.
- [25] Saeed M, Mark R. A novel method for the efficient retrieval of similar multiparameter physiologic time series using wavelet-based symbolic representations. *AMIA Annu Symp Proc* 2006;679–83.
- [26] Altıparmak F, Ferhatosmanoglu H, Erdal S, et al. Information mining over heterogeneous and high-dimensional time-series data in clinical trials databases. *IEEE Trans Inf Technol Biomed* 2006;10(2):254–63.
- [27] Ratanamahatana CA, Keogh E. Everything you know about dynamic time warping is wrong. Paper presented at the Third Workshop on Mining Temporal and Sequential Data (KDD-2004). Seattle (WA), August 22–25, 2004.
- [28] Fritsche L, Schlaefer A, Budde K, et al. Recognition of critical situations from time series of laboratory results by case-based reasoning. *J Am Med Inform Assoc* 2002; 9(5):520–8.
- [29] Chatfield C. Analysis of time series. 4th edition. New York: Chapman and Hall; 1989.
- [30] Bellazzi R, Larizza C, Riva A. Temporal abstractions for interpreting diabetic patients monitoring data. *Intelligent Data Analysis* 1998;2(1–4):97–122.
- [31] Graps A. An introduction to wavelets. *IEEE Comput Sci Eng* 1995;2(2):50–61.
- [32] Zhang J, Tsui FC, Wagner MM, et al. Detection of outbreaks from time series data using wavelet transform. *Proc AMIA Symp* 2003;748–52.
- [33] Keogh E, Lin J, Truppel W. Clustering of time series subsequences is meaningless: implications for previous and future research. Paper presented at The Third IEEE International Conference on Data Mining (ICDM '03). Melbourne (FL), November 19–22, 2003.
- [34] Hoppner F. Time series abstraction methods—a survey. Paper presented at the Dortmund, Germany: GI Jahrestagung; September 30–Oct 3; 2002.
- [35] Hoppner F. Learning dependencies in multivariate time series. Paper presented at the ECAI'02 Workshop on Knowledge Discovery in (Spatio-) Temporal Data. Lyon: France; July 22–23, 2002.
- [36] Bellazzi R, Larizza C, Magni P, et al. Temporal data mining for the quality assessment of hemodialysis services. *Artif Intell Med* 2005;34(1):25–39.
- [37] Stacey M, McGregor C. Temporal abstraction in intelligent clinical data analysis: a survey. *Artif Intell Med* 2007;39(1):1–24.
- [38] Shahar Y, Chen H, Stites DP, et al. Semi-automated entry of clinical temporal-abstraction knowledge. *J Am Med Inform Assoc* 1999;6(6):494–511.
- [39] Larizza C, Moglia A, Stefanelli M. M-HTP: a system for monitoring heart transplant patients. *Artif Intell Med* 1992;4:111–26.

- [40] Kuilboer MM, Shahar Y, Wilson DM, et al. Knowledge reuse: temporal-abstraction mechanisms for the assessment of children's growth. *Proc Annu Symp Comput Appl Med Care* 1993;449–53.
- [41] Kohane IS, Haimowitz IJ. Hypothesis-driven data abstraction with trend templates. *Proc Annu Symp Comput Appl Med Care* 1993;444–8.
- [42] Post AR, Harrison JH Jr. PROTEMPA: a method for specifying and identifying temporal sequences in retrospective data for patient selection. *J Am Med Inform Assoc* 2007;14(5): 674–83.
- [43] Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. Paper presented at the 20th Int Conf Very Large Databases. Santiago de Chile, Chile: VLDB; September 12–15. 1994. Santiago de Chile, Chile.
- [44] Morchen F, Ultsch A. Discovering temporal knowledge in multivariate time series. Paper presented at the Gesellschaft für Klassifikation (GfKI). Dortmund, Germany; March 9–11, 2004.
- [45] Morchen F. A better tool than Allen's relations for expressing temporal knowledge in interval data. Paper presented at the Theory and Practice of Temporal Data Mining (TPTDM 2006). Philadelphia; August 20–23, 2006.
- [46] Korfhage RR. Information storage and retrieval. New York: Wiley; 1997.
- [47] Siadat MS, Knaus WA. Locating previously unknown patterns in data-mining results: a dual data- and knowledge-mining method. *BMC Med Inform Decis Mak* 2006;6:13.
- [48] Tsoi AC, Zhang S, Hagenbuchner M. Pattern discovery on Australian medical claims data—a systematic approach. *IEEE Trans Know Data Eng* 2005;17(10):1420–35.
- [49] McCallum A, Nigam K, Ungar LH. Efficient clustering of high-dimensional data sets with application to reference matching. Paper presented at the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Boston, MA: August 20–23, 2000.
- [50] Sibai BM. The HELLP syndrome (hemolysis, elevated liver enzymes, and low platelets): much ado about nothing? *Am J Obstet Gynecol* 1990;162(2):311–6.
- [51] Sibai BM, Barton JR. Dexamethasone to improve maternal outcome in women with hemolysis, elevated liver enzymes, and low platelets syndrome. *Am J Obstet Gynecol* 2005;193(5): 1587–90.
- [52] Fonseca JE, Mendez F, Catano C, et al. Dexamethasone treatment does not improve the outcome of women with HELLP syndrome: a double-blind, placebo-controlled, randomized clinical trial. *Am J Obstet Gynecol* 2005;193(5):1591–8.