A Temporal Search Engine for a Massive Multi-Parameter Clinical Information Database

LH Lehman, TH Kyaw, GD Clifford, RG Mark

Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA

Abstract

We describe a novel search engine that is capable of rapid execution of queries concerning changes in the gradients and absolute (and relative) values of multiple irregularly sampled and asynchronous physiological parameters over many time scales. The search engine enables search criteria for multiple physiological parameters using gradient bounds, rates of change, and threshold breeches over various time scales. Multiple signals can be searched and combined in a Boolean manner to form complex queries. Pre-computed ranges and multi-scale gradients are used to significantly reduce the search time for locating temporal events. We have implemented the search engine in MATLAB and tested the algorithm on a massive multi-parameter intensive care unit database (MIMIC II). To illustrate the use of our search approach, a set of numerical search criteria were developed by clinicians to locate evidence for important pathophysiological conditions. .

1. Introduction

The Multi-parameter Intelligent Monitoring for Intensive Care II database (MIMIC II DB) [1] is a massive and growing intensive care unit (ICU) archive collection of over 17,000 patient records. One important challenge in clinical research using MIMIC II is in identifying clinical events of interest and cohorts of patients with similar pathologies. In particular, one main challenge is in the detection of clinical events that may involve complex dynamics of multiple physiological parameters over multiple time scales. Traditional threshold-based searching algorithms are incapable of detecting the complex physiological dynamics.

We describe a search engine that is able to perform multi-parameter, temporal queries on a large-scale time-series medical database, such as MIMIC II. The search engine is designed to serve as a filtering and event detection tool for researchers interested in investigating patient records that meet specific pathophysiological criteria, and in identifying possible onset times of certain clinically significant events. Some example temporal queries that clinicians might like to perform are as follows.

- Find episodes of lactic acidosis, in which pH $\leq$ 7.2, paCO2 $\leq$ 35 mmHg, and lactate $\geq$ 2.5 mmol/L.
- Find evidence of acute kidney injury, where creatinine $\geq$ 1.5 mg/dL and increases by 50% within 48 hours.
- Find episodes of hemodynamic instability in which heart rate (HR) increases by 50% or more in a six hour time window, with a simultaneous drop in arterial blood pressure (ABP) by at least 20% to below 60 mmHg.

Temporal queries on time series data cannot be efficiently implemented in a traditional SQL-based relational database. Temporal query languages, such as TQuel [2] and TSQL2 [3], express time intervals with cumbersome syntax and have limited support for multi-parameter time series data with different temporal resolutions. Saeed *et al.* [4] used a selected set of precomputed wavelet coefficients for efficient temporal searches. In contrast, our approach is to use pre-computed gradient bounds to reduce search time and to employ simple algorithms with little storage overhead for time series searches.

In the rest of the paper, we first characterize the time series data in the MIMIC II DB. Next, we give an overview of the search engine design, and describe a set of example searches. Finally, we demonstrate the utility of the search engine through two simple, illustrative examples of clinical analysis using search engine.

2. Temporal searches on the MIMIC II database

The search engine [5], implemented in MATLAB, currently searches on asynchronously and irregularly sampled MIMIC II lab results, medications, nurse-verified data downloaded from the bedside monitors, and demographic information. We have selected 128 parameters from MIMIC II and imported the data into the search engine environment. The parameter list includes important physiological indicators such as HR, ABP, temperature, cardiac output, laboratory values, IV drug levels, microbiology results, etc. These 128 clinical data items for approximately 17,000 patients together with their demo-

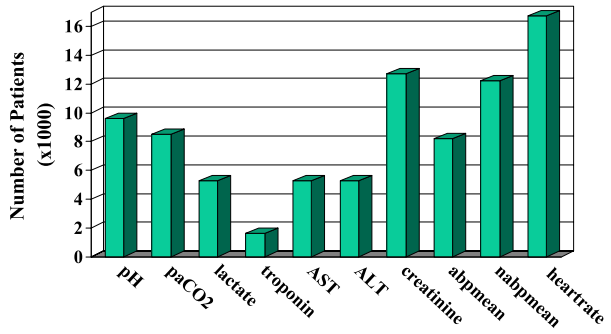

Figure 1. Number of patients with measurements for selected parameters.

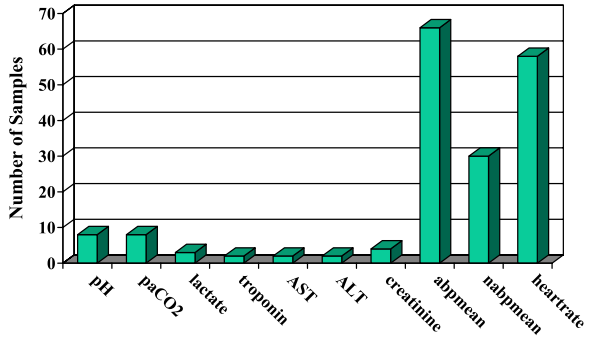

Figure 2. Median number of samples per patient for selected parameters.

graphic information (age, gender and patient ID number) were retrieved and stored in MATLAB compatible structures that enable fast and efficient access to a patient's data (see [5] for details). Figures 1, 2, and 3 show sampling statistics for a selected set of physiological measurements in MIMIC II. In these figures, $abpmean$ and $nabpmean$ represent invasive and non-invasive mean arterial blood pressure levels respectively. AST and ALT represent the concentration of aspartate and alanine aminotransferase respectively, both of which are indicators of the liver function. paCO2 is the partial pressure of arterial carbon dioxide.

3. Search engine system overview

One important design goal for the search engine is to provide a set of temporal query types that can be used to characterize the temporal patterns of the physiological events of interest. Threshold breeches, such as those used in traditional alarm settings, are useful in flagging occurrences of certain clinical events, such as hypotension. However, many pathophysiological conditions can be better characterized by the rate of change of physiological signals over time. Further, the time scales over which these parameter values change can sometimes provide useful in-

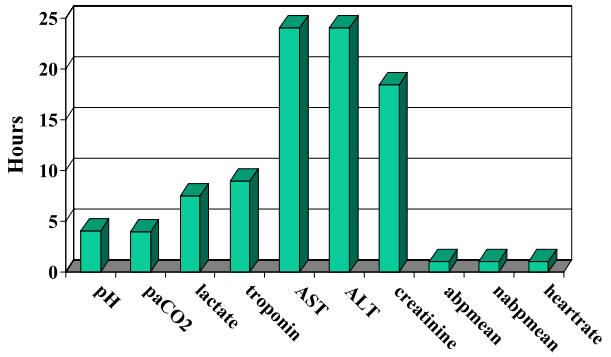

Figure 3. Median sampling interval for selected parameters.

sight into the underlying physiological events.

Our search engine [5] allows the users to construct temporal queries using both *thresholds* and *gradient bounds* over a specified range of time scales. More specifically, the search engine supports three types of basic time series search criteria.

- Threshold breeches, e.g., $7.36 \leq$ pH ≤ 7.44. In the threshold search criteria, either a lower, or an upper bound, or both are specified for a parameter.
- Gradient search *by value* over a specified time range, e.g., creatinine increases ≥ 3 mg/dL within 48 hours. In gradient search by value, the lower and/or upper bounds for the nominal change in item of interest, and the smallest and/or largest time intervals for which the change in values should be calculated must be specified.
- Gradient search *by percent* over a specified time range, e.g., heart rate increases by at least 50% within 6 hours. Gradient search by percent requires the same search parameters as gradient search by value, except that the lower and upper bounds on value changes are interpreted as the bounds of *percentage* change in each parameter.

It is possible to construct queries that search for multiple physiological parameters using a combination of threshold and gradient bounds over specified time scales. Multiple signals can be searched and combined in a Boolean manner to form complex queries (see [5] for details).

We have implemented two options in combining searches from multiple parameters. In an unsynchronized search, multiple parameters are searched separately; event intervals are defined by the intersecting periods in which criteria for all parameters are met. In a synchronized search, criteria for multiple parameters must be satisfied in the same time window. The output of the search engine from both types of searches is a list of patient records that satisfy the criteria and the time intervals in which the criteria are satisfied.

Demographic information (such as age and gender) can be included in the search criteria to reduce the number of patient records that need to be searched during runtime.

Additionally, the search engine uses pre-computed ranges and multi-scale gradients to reduce the search time for locating temporal events. More specifically, for each parameter, we pre-compute the gradients over all possible pairs of samples of a patient, and store the min/max nominal change, percent change, and sampling interval in a MATLAB structure. At runtime, these values are used to pre-filter patient records that do not satisfy a given trend query.

4. Example MIMIC II searches

To illustrate the use of our search approach, a set of numerical search criteria were developed by clinicians using our temporal query syntax to locate evidence for important pathophysiological conditions such as acute myocardial infarction (AMI), lactic acidosis (LA), acute kidney injury (AKI), hemodynamic instability (HI), multi-organ failure (M-Org), and paroxysmal tachyarrhythmia (Tachy).

Type	Criteria
AMI	troponin ≥ 0.1
LA	pH ≤ 7.2 & paCO2 ≤ 35 & lactate ≥ 2.5
AKI	Cr ≥ 1.5 & Δ Cr $\uparrow \geq 100\%$ in 48 hours
HI	(Δ HR $\uparrow \geq 50\%$) & (Δ ABP $\downarrow \geq 20\%$ to ≤ 60 mmHg) in a 5 to 6 hr window
Tachy	Δ HR $\uparrow \geq 40$ in 5 minutes
M-Org	(Cr ≥ 1.5 & Δ Cr $\uparrow \geq 100\%$ in 48 hrs) & (AST ≥ 40 & Δ AST $\uparrow \geq 50\%$ in 48 hrs) & (ALT ≥ 40 & Δ ALT $\uparrow \geq 50\%$ in 48 hrs)

Table 1. Search Types: HR, heart rate; Cr, serum creatinine; ABP, mean arterial blood pressure. The criteria for HI specifies a synchronized search in which the heart rate increase and blood pressure drop occur in the same 5-6 hour time window. Units of measurement: paCO2 (mmHg), lactate (mmol/L), Cr (mg/dL), HR (bpm), ABP (mmHg), AST/ALT (units/mL).

Table 1 lists example criteria used for each type of pathophysiological event of interest. To remove artifacts and invalid entries, each parameter is pre-filtered with a set of min/max values so that only sample values in a feasible physiological range are used for searching.

4.1. Search results

Table 2 lists the search results performed on the MIMIC II DB with over 17,000 patients. Note that the pre-filtering strategy implemented in search engine can significantly reduce the number of patient records that need to be searched during runtime. For example, the pre-filtering procedure ruled out up to 90% of the available patient records in searching for acute kidney injury and multi-organ failure, and significantly reduced the search time.

Type	N	M	Hits	Ratio
AMI	1,610	1,092	1,092	67.83%
LA	5,009	676	203	4.05%
AKI	10,423	1,096	294	2.82%
HI	12,694	5,111	477	3.76%
Tachy	15,679	5,840	527	3.36%
M-Org	2,439	266	25	1.03%

Table 2. MIMIC II search engine search results. N is the number of patients with at least 1 sample (for threshold search) or 2 samples (for gradient search) of all the required search criteria. M is the number of potential patients remaining after the pre-filtering step has been performed. 'Hits' represent the number of patients that the search engine reported as having at least one episode of the specified event type. 'Ratio' = $Hits/N$, and is the number of potential patients with sufficient data that satisfied the criteria. The HI result shows the union of patients from 5 searches of varying window sizes between 5 and 6 hours with 15-minute increment in window size.

Except for Tachy and HI, the searches in Table 2 took between 5 and 15 seconds on a P4 3 GHz processor to complete. Queries on Tachy and HI (for each synchronized window size search) took approximately 270 and 420 seconds respectively to complete. In general, queries that involve gradient searches on items with high sampling rate such as HR and ABP require longer time to complete.

It should be noted that the search hits from these example searches do not necessarily represent the actual number of patients with the specified pathology in the MIMIC II DB. Gradient searches are limited to patients with at least two samples of the items searched on, which might rule out some patients with the pathophysiological conditions of interest. As an example, there are approximately 5,200 patients with AST measurements in the MIMIC II DB. However, only half of those patients have two or more samples of AST measurements.

Moreover, current implementation of the search engine requires at least some overlapping interval in which the criteria for all parameters must be satisfied in order for a search hit to occur. In the search for multi-organ failure, for example, the liver failure (indicated by the AST and ALT criteria) and the kidney failure (indicated by the creatinine criteria) must occur in some overlapping time intervals for the search engine to return a search hit for the patient.

5. MIMIC II search results clinical analysis

One use of the search engine is in identifying cohorts of patients with similar pathologies. In this section, we demonstrate the utility of the search engine through two

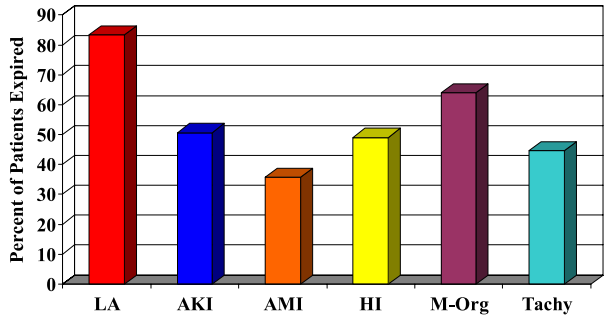

Figure 4. Mortality of each group of patients identified using the MIMIC II search engine.

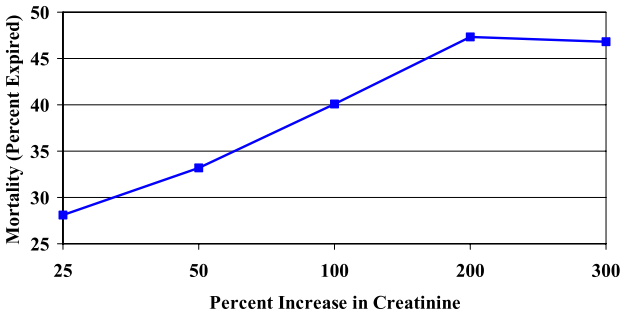

Figure 5. MIMIC II search results clinical analysis on AKI: mortality as a function of percentage increase in creatinine in 48 hours.

simple, illustrative examples of clinical analysis that study mortality among groups of patients identified through different search criteria.

In the first example, we demonstrate the use of the search engine in studying the mortality in each group of patients with similar pathologies. Figure 4 illustrates the mortality of patient groups identified using the search criteria defined in Table 1. The overall mortality rate in the MIMIC II ICU is approximately 16% (2,817 expired patients). The lactic acidosis group has the highest mortality rate – of the 203 patients with lactic acidosis, 169 of these patients expired.

In the second example, we illustrate the use of the search engine in studying the relationship between changes in patients' serum creatinine level and the mortality rate. We use the search engine to identify patients with different levels of acute kidney injury defined in terms of percentage increase in creatinine greater than or equal to 25%, 50%, 100%, 200%, and 300% in 48 hours. Figure 5 presents mortality rate as a function of percentage increase in creatinine. We observe that an increase in creatinine is associated with significant increase in mortality (the mortality of patients with 50% increase in creatinine in 48 hours is twice the general ICU mortality in MIMIC II).

6. Conclusions and future work

We have described the design, implementation, and use of a search engine for identification of clinically significant events and episodes of physiological interest for a large-scale, multi-parameter time series medical database. We have demonstrated the use of the search engine in identifying cohorts of patients in the MIMIC II DB with similar pathologies, and the mortality in each group of patients.

For future work, we plan to expand the types of temporal patterns that can be detected by the search engine, including increasing flexibility for defining searches for multiple asynchronous events from multiple parameters. We also intend to explore methods of implementing searching algorithms that use techniques such as regressions for event detection, and a more robust mechanism for dealing with noise and artifacts.

Acknowledgements

This work was supported in part by the U.S. National Institute of Biomedical Imaging and Bioengineering (NIBIB) and the National Institutes of Health (NIH) under Grant Number R01 EB001659, and Philips Medical Systems. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIBIB, the NIH, or Philips Medical Systems.

References

[1] Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Computers in Cardiology 2002; 29:641–644.

[2] Snodgrass R. The temporal query language TQuel. ACM Transactions on Database Systems June 1987;12(2):247–298.

[3] Snodgrass R. TSQL2 Temporal Query Language. Kluwer Academic Publishers, 1995.

[4] Saeed M, Mark RG. Efficient hemodynamic event detection utilizing relational databases and wavelet analysis. Computers in Cardiology 2001;28:153–156.

[5] Kyaw HT. Formatting and Searching a Massive, Multi-Parameter Clinical Information Database. Master's thesis, MIT, Cambridge, MA, September 2005. URL `http://hdl.handle.net/1721.1/36788`.

Address for correspondence:

Li-wei H. Lehman
Harvard-MIT Health Sciences and Technology
E25-505, Cambridge, MA 02139 USA
lilehman@mit.edu