ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

The physiology analysis system: An integrated approach for warehousing, management and analysis of time-series physiology data

*Thomas M. McKenna, Gagandeep Bawa, Kamal Kumar, Jaques Reifman**

Bioinformatics Cell, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, MD 21702, United States

ARTICLE INFO

Article history:
Received 27 June 2006
Accepted 3 January 2007

Keywords:
Physiology
Medical informatics
Data display
Time-series data analysis
Data base management system

ABSTRACT

The physiology analysis system (PAS) was developed as a resource to support the efficient warehousing, management, and analysis of physiology data, particularly, continuous time-series data that may be extensive, of variable quality, and distributed across many files. The PAS incorporates time-series data collected by many types of data-acquisition devices, and it is designed to free users from data management burdens. This Web-based system allows both discrete (attribute) and time-series (ordered) data to be manipulated, visualized, and analyzed via a client's Web browser. All processes occur on a server, so that the client does not have to download data or any application programs, and the PAS is independent of the client's computer operating system. The PAS contains a library of functions, written in different computer languages that the client can add to and use to perform specific data operations. Functions from the library are sequentially inserted into a function chain-based logical structure to construct sophisticated data operators from simple function building blocks, affording *ad hoc* query and analysis of time-series data. These features support advanced mining of physiology data.

1. Introduction

Advances in physiology sensors and data acquisition technology increasingly support collection of time-series data from patients in locations other than the clinic or laboratory. These data provide new opportunities to analyze the physiological state of individuals at times when such information is most valuable, both in real-time and for *post hoc* data mining. However, as the capability to collect time-series data advances, data may be collected under suboptimal conditions, such as during the monitoring of subjects engaged in various states of physical activity, or during transport of patients from a site of injury to hospital. These conditions can degrade the quality of the collected data. For example, cursory examination of physiology waveform data, such as electrocardiograms (ECG) collected during transport of patients shows that the waveforms are subject to patient movement artifacts, and yield records that exhibit intermittent periods of good- and poor-quality recordings [1]. Before time-series data can be comprehensively mined, it is necessary to efficiently warehouse, manage, and analyze the data files, which can, generally, be characterized as extensive, numerous, and of variable quality. A description of such files, based on data collected from trauma patients, will be given below.

The physiology analysis system (PAS) was designed as a research platform to facilitate the extraction of knowl-

* *Corresponding author at*: U.S. Army Medical Research and Materiel Command, MRMC/TATRC, Building 363, Miller Drive, Fort Detrick, MD 21702-5012, United States. Tel.: +1 301 619 7915; fax: +1 301 619 1983.
E-mail address: jaques.reifman@us.army.mil (J. Reifman).

0169-2607/$ – see front matter
doi:10.1016/j.cmpb.2007.01.003

edge from physiology time-series data. Its essential features reduce the data management task faced by investigators, while providing data analysis capabilities in an environment that promotes *ad hoc* queries of the databases. The system includes: (1) a platform to integrate storage, query and analysis of attribute (i.e., discrete) and time-series data, (2) a structure in which all system operations are executed via the Internet, and (3) an interface and logical structure that is as flexible and user-friendly as possible. The PAS requires only that clients have Internet access and a Web-browser; all data storage, access, analysis, and graphics functions reside on a server. Besides query and analysis, the client can visualize all attribute and time-series data, output results in reports, export the results or raw data as files, and share data and analytic routines amongst other users.

Ultimately, it is expected that the system will be used to develop knowledge that can be incorporated into algorithms that will, for instance, continuously monitor physiology variables during transport of a patient and provide a continuous read-out of the physiologic stress the patient is experiencing. A similar application is to use the mining-derived knowledge to synthesize algorithms that will allow diagnosis of injury type and severity, and prognosis of outcome based on real-time physiology data inputs from vital sign monitors, as commonly used during transport of injured civilians.

2. Data specification

The PAS is independent of specific physiology data acquisition devices; it can incorporate data collected by any device, as long as the time-series data can be expressed as ASCII files. The PAS hosts multiple physiology databases in a common interface, and it currently incorporates three databases: (1) data collected during transport of patients from the site at which they were injured to a hospital [2], (2) data collected from subjects while they were engaged in varying intensities of physical activity [3], and (3) data collected from subjects during transport on an instrumented platform similar to a litter [4]. The time-series data were collected by a variety of data acquisition devices, including the Propaq Encore, Model 206-EL (Welch Allyn, Skaneateles Falls, NY), Lifeshirt (VivoMetrics, Ventura, CA), Schiller Cardiovit AT-6 (Schiller Inc., Baar, Switzerland), SensorMedics Model 2900 (SensorMedics, Yorba Linda, CA), and a respiratory flow track board (Novametrics Medical Systems, Wallingford, CT).

A data mining example, drawn from the trauma patient database, will be used throughout this paper to illustrate features of the PAS. This database consists of attribute and time-series vital signs data collected from approximately 900 patients during helicopter transport from the site of injury to a Level-1 trauma center at the University of Texas Health Science Center at Houston, TX. The attribute data include items such as patient demographic information, injury description, and treatments. There are 100 variables of this type for each patient, and these data have already been subjected to a mining exercise [2]. In addition, twelve time-series variables were collected by Propaq 206 vital sign monitors [5] during transport; these include blood pressures, ECG, pulse oximetry, respiration, end-tidal CO_2, and additional time-series data files that are derived from the original waveforms, such as heart and respiratory rates, and arterial blood oxygen saturation.

The types of data described above have data management and analysis challenges. These include: (1) variable data frequencies, which result from data acquisition and output limits specific to vital signs monitors and the algorithms used to calculate derived values. In the trauma database, time-series frequencies range from 182 Hz for the ECG waveform to 1 Hz for derived values, such as heart rates, to sub Hz for blood pressures (i.e., single-point blood pressures are measured by employing a cuff, at intervals between 2 and 10 min), (2) the sheer volume of data—the record for an average patient is approximately a half million data points (approximately 2MB), (3) both attribute and time-series data that must be analyzed at the same time, and (4) a large number of data files to manage (e.g., more than 14,500 time-series files, which consist of the original waveform data and their associated derived values, currently stored in the database).

Existing database management systems are not optimal for the management and analysis of time-series data, since they were not built with this objective [6,7]. Their storage architecture is not structured to effectively handle time-series data and simple time-series operations are poorly supported. Querying the data requires writing statements in some form of query language [7]; this is acceptable in business applications where the same queries are performed repetitively, but becomes burdensome when the database is subject to *ad hoc* queries, as is typical of medical data mining research applications.

3. Computational methods and theory

3.1. *Function chain-based data query and analysis*

The PAS works via function chains in which data are passed through distinct functions that are sequentially executed until the analytical objective is attained. For example, a simple function chain consisting of two functions could be constructed by first applying a function to data in the trauma database to select only individuals that suffer blunt injury, followed by a second function to compute heart rate from their ECG waveforms. There can be many functions to calculate heart rate, each using a different method to do so, and the user determines which one to insert into the chain.

A function chain accomplishes two objectives. The first is to select a subset of subjects from the database, i.e., a query, based on client-supplied constraints that can be applied to attribute and/or time-series data. In the example, a constraint was used to select only individuals with blunt injuries in the trauma database.

A second function chain objective is the mathematical analysis of time-series data that is associated with each subject, to yield new time-series or scalar data. In the example, the calculation of heart rate from an ECG waveform yields new heart rate time-series data that did not previously exist.

Function chain architecture, because of its inherent modularity, allows substantial flexibility in performing *ad hoc*