Information Mining Over Heterogeneous and High-Dimensional Time-Series Data in Clinical Trials Databases

Fatih Altiparmak, Hakan Ferhatosmanoglu, Selnur Erdal, and Donald C. Trost

***Abstract*—An effective analysis of clinical trials data involves analyzing different types of data such as heterogeneous and high dimensional time series data. The current time series analysis methods generally assume that the series at hand have sufficient length to apply statistical techniques to them. Other ideal case assumptions are that data are collected in equal length intervals, and while comparing time series, the lengths are usually expected to be equal to each other. However, these assumptions are not valid for many real data sets, especially for the clinical trials data sets. An addition, the data sources are different from each other, the data are heterogeneous, and the sensitivity of the experiments varies by the source. Approaches for mining time series data need to be revisited, keeping the wide range of requirements in mind. In this paper, we propose a novel approach for information mining that involves two major steps: applying a data mining algorithm over homogeneous subsets of data, and identifying common or distinct patterns over the information gathered in the first step. Our approach is implemented specifically for heterogeneous and high dimensional time series clinical trials data. Using this framework, we propose a new way of utilizing frequent itemset mining, as well as clustering and declustering techniques with novel distance metrics for measuring similarity between time series data. By clustering the data, we find groups of analytes (substances in blood) that are most strongly correlated. Most of these relationships already known are verified by the clinical panels, and, in addition, we identify novel groups that need further biomedical analysis. A slight modification to our algorithm results an effective declustering of high dimensional time series data, which is then used for "feature selection." Using industry-sponsored clinical trials data sets, we are able to identify a small set of analytes that effectively models the state of normal health.**

***Index Terms*—Clinical trials, information mining, time series.**

I. Introduction

THE most expensive parts of drug development are the Phase III clinical trials, which are performed to prove efficacy via statistical significance tests. Traditionally, data collection and analytical rigor have been applied to statistical analysis of efficacy, while the safety measurements are counted and presented only as descriptive statistics, leaving the conclusions about safety of a new drug to clinical judgment. These safety signals tend to be detected when millions of patients have been exposed after the drug goes on the market. An effective mining of drug data collected under the current conditions is crucial to find subtle signals that heretofore went undernoticed. However, mining such clinical trial data is a challenging task, given its high dimensionality, the amount of missing data, noise in data, heterogeneity, differences in its data sources, and also differences in the number of data points and attributes. Within the mining process of clinical trials, medical professionals note themselves that "There are no diseases, but there are patients," which means diseases may occur in many varieties due to both patient biology and environmental effects on the patient. Even the psychological state of the patient could effect the measurements from one patient to another.

Recently, the U.S. Food and Drug Administration (FDA) reported that "examples of tools that are urgently needed include better predictors of human immune responses to foreign antigens" [1]. For this reason, one of the actions taken by FDA was to mine available databases to identify molecular substructures with potentially negative toxicologic properties early in development process. The purpose was to enhance the safety of transplanted human tissues, and find new techniques for accessing drug liver toxicity. The FDA also highlighted that computational approaches, such as computer modeling, and the gained information from such analysis, when combined with predictive toxicology, may reduce the overall cost of new drug discovered by 50%. Many investigational new molecular entities are tested in laboratories and ultimately evaluated by industry and government-funded clinical trials every year. Retrieving as much information as possible from such trials is very important, since the cost of investigating a drug is on the order of a billion dollars. Even though studies are designed with detailed protocols, where particular patient populations, disease types, and drug regimens are specified in detail, the actual data are messy because of the numerous human factors that occur during the conduct of a trial. Every patient, physician, and medical staff person contributes to data irregularity. Much energy and expense is applied to minimize these factors, but they cannot be avoided. It is virtually impossible to conduct a study with all the data fields completed and all the patient measurements occurring as specified. This creates certain challenges when the time comes to analyze data.

In this paper, we first discuss the challenges of mining clinical trial data, and then propose a novel approach for mining information out of the clinical trial databases. The overall approach involves two major steps: 1) applying a mining algorithm over significant and clean (homogeneous) subsets of data, and

Manuscript received August 30, 2004; revised February 27, 2005, April 20, 2005. This work was supported by Pfizer Inc.

F. Altiparmak and H. Ferhatosmanoglu are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: hakan@cse.ohio-state.edu).

S. Erdal is with the Biophysics Graduate Program, The Ohio State University, Columbus, OH 43210 USA.

D. C. Trost is with Pfizer Inc. Global Research and Development, Groton, CT 06340 USA.

Digital Object Identifier 10.1109/TITB.2005.859885

2) joining the information gathered in the first step by identifying common or distinct patterns found over the mining results. We implemented this approach for heterogeneous and high dimensional time series data in clinical trial databases by utilizing the frequent itemset mining, as well as clustering and declustering techniques with novel distance metrics. By clustering the high dimensional time series data, we find groups of clinical laboratory analytes that are strongly correlated. A slight modification to our algorithm results in an effective declustering of high dimensional time series data, which is then used for "feature selection."

The rest of the paper is organized as follows: first, we define and explain the challenges in analyzing clinical trials data by using the current techniques. We explain our information mining process, and show its application to the grouping of blood analytes and defining a global panel of analytes that represents the human health state. We present results and findings of the proposed algorithms in the context of clinical trial data, which is a typical application involving heterogeneous and high dimensional time series data. Our algorithms are applicable to other data with similar characteristics.

II. Challenges

A. Differences Due to Time Series

Depending on the treatment regimen, patients will have different time intervals between their clinical measurements for many reasons, such as the following.

1) *Patient Qualification Criteria:* Patients are recruited for a study over an extended period of time. It is usually impractical to start everyone on the same day. Therefore, patients may start on a study on different dates, may have been measured a variable number of times prior to starting a treatment, may be exposed to differing environmental conditions, especially seasonal effects, and almost definitely are in a variable state of the disease being treated.
2) *Patient Health Status Change:* Generally, clinical trial protocols require patients to be measured at specified, possibly unequal, time intervals relative to the randomization time. Although all patients are supposed to have the same measurements at the same time points, this tends not to happen in practice. Patient and physician availability cannot always be scheduled according to the protocol. Ethically, the health needs of the patient must override the rigid adherence to the protocol. Adverse events, side effects, and other illnesses happen to patients during the study which may cause the patient to drop out of the study, to have additional therapy, to temporarily or permanently discontinue the experimental treatment, or to have additional measurements taken, some of which are the same as those for the regular protocol times, some of which are different measurements specified by the protocol if adverse events occur, and some of which are outside the protocol and are determined by the treating physician.

For all the preceding reasons and several more, we may have irregular, nonstandard data on patients. This brings us to the problem of looking into data sets of time series data, where the length of observation is neither fixed nor standardized.

B. Differences Due to Data Sources

1) *Multiple Study Sites:* In larger clinical trials it is common to see studies that are handled by multiple sites because a single physician cannot recruit all of the subjects needed for the trial. Even the same physician will use different terms or even units at different times to describe the same clinical condition. This problem is exacerbated when studies are international. under these circumstances.
2) *Multiple Lab Sources:* Many protocols use multiple clinical laboratories for evaluating a patient's health status, and these laboratories may use different methods to quantifying various analytes. Different analytes may also be collected at different frequencies within a protocol, and almost certainly across protocols. For instance, hematology (blood cells) results from a study may be collected from a patient at every scheduled visit, while, on the other hand, clinical chemistry (serum) results may be obtained at other times with greater frequency. Analytes may also be related to treatment response, and the timing could vary for each patient and each measurement type.
3) *Differences in Error Distributions:* Different types of laboratory tests and evaluation techniques will be applied to the patient which could have different frequencies; these techniques will also bring different levels of errors due to different procedures that are involved in the process. Even with the same laboratory test, we may see differences in results based on the technique.
4) *Domain Specific Process:* During a drug development project, it is common that a drug would be tested on several types of patients, both healthy and nonhealthy. For example the same drug's effect might be tested on patients with colon cancer, lung cancer, and malignant melanoma. This is due to the fact that drugs might be targeted at different types of cancers with different cell types. If it were the case, that we were evaluating the quality of life for each disease, we would certainly get different results on each patient type. In such cases the observation and related data that we gather from patients might differ dramatically.

C. Differences Due to Data Mining or Analysis Techniques

A patient can be evaluated from a molecular level to the quality of life level. Such an analysis would be highly multidimensional. While the data types and retrieval processes get broader, the requirements on analysis of such studies get more challenging. It would be hard, if not impossible, to come up with a single distance metric or clustering method to solve all the problems that we have mentioned so far on such a high dimensional data set while achieving minimal or no classification errors during the analysis, and not suffer from the curse of dimensionality, with all of the standardization and normalization errors that we have discussed so far. An additional concern

is that while some metrics and clustering techniques would give better results on certain data types, they may suffer and generate more errors in another data type, resulting in differences in the level of errors associated with them.

III. Foundations

A. Clinical Trial Data

1) *Nature of the Data:* We will illustrate our algorithms and results using a sample set of industrial clinical trial data. The data set is a good example demonstrating the common problems mentioned previously. The main division of the data was based on the drugs that were studied for marketing. For each drug, a sample of patients from different regions with different genders and different ages were selected. The data set at hand contained more than 28,000 patients for different drugs. Each patient was in only one study, and was measured at a limited number of unequally spaced time points. For each visit (time point), a patient's blood and urine samples were taken; these samples are called *analytes*. Here we define an analyte as any substance inside the blood or urine that we can measure, such as hemoglobin, calcium, or phosphate. Each study has a set of required analytes that needs to be observed for every patient in the sample. For a given drug, a set of evaluated analytes might differ across studies. Furthermore, within a drug study, patients have differences in number of observations and intervals of observations; for example, different patients may have differences in the total number of visits, and they may have differences in time length between visits.
2) *Analyzing Analyte Relationships:* For each analyte, the range for a normal healthy person is published by numerous sources, such as textbooks and the NIH. However, a patient who has an observation for an analyte outside the range of normal cannot be considered to be unhealthy right away, due to the fact that the values for the range of normal are decided based on cutoff values from a probability distribution. However, if we have used a set of analytes instead of a single one, we might have a better answer regarding the patient's state. In this paper, we aim to find subsets of analytes that are related, and to identify a global panel of analytes representing the overall health state of a patient. A practical outcome of this panel is feature selection. In the data set described in Section III, there are 43 analytes (dimensions), which is too high for many statistical analysis techniques to be effective. Current dimensionality reduction techniques define new dimensions that do contain representations of existing dimensions within them. However, these new dimensions are not actual analytes anymore, and, as such, are not medically interpretable and tend to require all the measurements anyway; hence, they do not provide a good way to reduce the number of tests necessary for diagnosis.

B. Preprocessing of Data

1) *Selection of Appropriate Subset:* The subset of data taken as the input to the algorithm contains the patients that have at least k observations for each member of a set of analytes, which is determined as follows. First, for each analyte, the total number of patients that has at least k observations is calculated. Second, based on the numbers found in step one, a threshold is decided, and each analyte that passes the threshold test is selected as a member of analyte set. Patients that have at least k observations for each member of the previously selected analyte set are chosen. k was set to four after a set of tries in our experiments. The result of these tries validated two facts: 1) while k increases, the total number of analytes and patients in the subset decreases, and 2) while k decreases, the lengths of series become inadequate to analyze and compare.
2) *Separation of Data Sources:* Time series data of each analyte for a single patient, by nature, has a homogeneous format which then leads to a more accurate analysis. Hence, we start the proposed *information mining* process by mining the data within its atomic source (i.e., data for each patient). Separating the sources is a natural choice, due also to the difficulties when comparing time series from different sources in clinical data.

C. Brute-Force Solution

A straightforward data mining process would apply the algorithms globally over the whole data set. For example, to identify strongly related groups of analytes, one can apply a standard clustering algorithm over the whole data set where the analytes are the data objects to be clustered, and the distance between each analyte is defined by the distance between the corresponding high-dimensional vectors (values) of analytes. One can improve this process by separating each patient's records to reduce dimensionality. It is intuitive, after separating the data sources, that the distance between two analytes can be described as the sum of distances for each data source. If the total number of patients (data sources) is p, then to find the distance between analyte_1 and analyte_2, the distance between these two analytes in all (p) patients needs to be summed. Then, according to this global distance, analytes can be clustered and the clusters, can be assigned as biological groups. We implemented this approach to test the level of its effectiveness. Clustering results in different sizes of clusters, each of which has different levels of correlations. The results had little or no meaning, partially because of the obvious problems caused by the heterogeneity and incompleteness of the data, and partially because of the difficulty of interpreting the output of such an analysis.

IV. Information Mining Algorithm

The proposed information mining approach (summarized in Table I) consists of a simple preprocessing followed by two major general steps. In step 1, significant, clean, and homogeneous subsets of data are identified and analyzed using a data mining algorithm. For the clinical trials case study proposed in

TABLE I
ALGORITHM

tNP = total number of patients
tNA = total number of analytes
ClusterMatrix[tNP][tNA] [tNA]
Step-1: Subset-mining over homogeneous chunks of data (e.g., analyte-clustering for each patient)
  Cluster analytes for *each patient*
    K-Medoid Clustering with one of the metrics to compute distance between analytes: {*Correlation coefficient, Euclidian, Qualitative, DTW*}

  Save pairing analytes into ClusterMatrix
    if for patient$_i$, analyte$_j$ and analyte$_k$ co-occur in the same cluster,
      set ClusterMatrix[i][j][k] to 1
    else set ClusterMatrix[i][j][k] to 0

Step-2: Mine the *information* gathered in the first step
  Input: { Support and Confidence Limits, ClusterMatrix}
  Output1:{ Groups of strongly related analytes}
  Output2:{ A global panel of analytes}
    Run frequent item-set mining over ClusterMatrix
      Biological groups= Frequent analyte-sets
      Global Panel= Least frequently co-occured group of analytes

this paper, this step corresponds to the preprocessing, followed by the clustering of analytes for each patient. Step 2 is to join the information gathered in the first step by identifying common (or distinct) patterns over the results of mining of the subsets. For our case study, this corresponds to finding strongly related groups of analytes (or a set of representative analytes with minimal cross-correlations) by mining common patterns over the clusters generated in the first step. The algorithm has two distinct fold results: 1) groups of analytes that are strongly correlated and 2) a global panel of analytes that can model the human health state. In this section, we will describe how we achieve the first result. A minor change in the process results in the second outcome, which will be described in Section VII.

The first step of the algorithm is to cluster the analytes for each data source. Any clustering technique can be utilized for this purpose. We chose the K-Medoid clustering algorithm, since it has many advantages over others such as the widely-used hierarchical clustering, including its robustness, the ability to specify the number of clusters (K), the relative simplicity of the algorithm, and its minimal use of the computing resources [2]. For each patient, the pairs of analytes that come together in the same cluster are saved as the output of the first step. The output can be considered as a transaction data (e.g., market basket data [3]) that includes information about which sets of analytes appear together. In other words, each cluster is mapped to a single transaction containing analytes. While clustering the analytes for each patient, short time series, still potentially with unequal intervals, are compared to each other. Distance metrics for comparison of this kind of time series are needed. The metrics utilized by our algorithm will be discussed in Section V.

In the second step of the algorithm, a frequent item set mining algorithm [3] is run to find strongly related analyte groups, most of which are shown to have high biological correlation. Any frequent item set mining algorithm can be used for this purpose. Frequent item sets are defined as the sets of items that cooccurred often enough to pass a given threshold called the support limit. This output itself is usually not very useful, since, typically, a large number of overlaps and redundancies (such as subsets) exist. We apply a second level of pruning and select a set of item sets using the confidence values of their corresponding association rules. Confidence for an association rule $X \Rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X [3]. In our context, the confidence for an association between a set and its subset is described as the number of times members of set cooccurred in the same cluster/number of times members of the subset co-occurred in the same cluster. While finding a group, not only the total number of cooccurrences of the group members compared with a threshold (support limit), but also, the ratio between this total to that of each one-item-less subset is compared with another threshold (confidence limit). A group is announced if it can pass both tests. Our algorithm finds groups of size greater than or equal to three. Therefore, the initial support limit (sp) and the initial confidence limit (cl) given to the algorithm are used to get the frequent items of size three. While the size of the set increases, these thresholds are transformed according to group size (details of this transformation can be found in the following).

As an example, we examine the rules of being reported as a group for the $\{\text{analyte}_1, \text{analyte}_2, \text{analyte}_3\}$ set. Assume that they come together in the same cluster for 50 data sources (patients). Thus, 50 must be greater than sp. Then the ratios of 50 to the total number of cooccurrences (support) of each subset of size two are compared to cl. This means that if the support of $(\{\text{analyte}_1, \text{analyte}_2\})$ is 60, then that of $(\{\text{analyte}_1, \text{analyte}_3\})$ is 67, and that of $(\{\text{analyte}_2, \text{analyte}_3\})$ is 73, then each of $50/60, 50/67$, and $50/73$ should be greater than or equal to cl. If all conditions met, $\{\text{analyte}_1, \text{analyte}_2, \text{analyte}_3\}$ can be reported as a group; otherwise, not.

As mentioned previously while the size of the set increases, different support and confidence limits are used by the algorithm. These new limits are obtained from the sp and the cl by using a special transformation factor for each limit. The one used for the support limit is called tsp, and the other is tcl. For a group of size greater than three, the algorithm uses sp-(group size-3)*tsp as the support limit and $\text{cl} + (\text{group size} - 3) * \text{tcl}$ as the confidence limit. So, for the set of four, the support limit is $\text{sp} - \text{tsp}$ and the confidence limit is $\text{cl} + \text{tcl}$. Whereas, for the set of five; the support limit is $\text{sp} - 2 * \text{tsp}$ and the confidence limit is $\text{cl} + 2 * \text{tcl}$, and so on. The confidence limit is increased until 0.9. Our experiments showed that five is the best value for tsp and 0.25 is the best one for tcl. As a result of this process, strongly related item sets of reasonable sizes are the output of the algorithm.

An additional measure over support and confidence, called Lift (Correlations), has been proposed [4]–[7]. It is defined as:

$$\text{Lift}(A => B) = \frac{P(AUB)}{P(A) * P(B)}$$

In our case, the probability of a set $S(P(S))$ is: Support(S)/total, where *total* is total # of transactions. Thus:

$$\text{Lift}(A => B) = \frac{(\text{Support}(AUB)/\text{total})}{[(\text{Support}(A)/\text{total}) * (\text{Support}(B)/\text{total})]}$$

$$\text{Lift}(A => B) = \frac{(\text{Support}(AUB))}{(\text{Support}(A) * [(\text{Support}(B)/\text{total})}$$

and Lift for a member of a set and remaining can be defined as:

$$\text{Lift}(A - \{x\} => \{x\}) = \frac{(\text{Support}(A))}{(\text{Support}(A - \{x\}) * [(\text{Support}(\{x\})/\text{total})]}$$

While clustering analytes for each patient, an analyte shows up in one of the clusters. Thus, support of an analyte is equal to the number of patients. We used the K-Medoid algorithm, so for each patient, there are K transactions (clusters). Therefore, the total number of transactions is $K*$ number of patients. As a result, for an analyte a, Support(a)/total $= 1/K$. In our experiment, we compare Support(A)/Support$(A - \{a\})$ to confidence limit which is more than $1/K$, hence the value of Lift is greater than 1. Eventually, for each resulting strong group, we can say with a high degree of assurance that each subset of the group has a positive correlation greater than $\text{cl} * K$ with the remainder of the group.

Briefly, the algorithm clusters analytes for each patient in the first step. In the second step, the analytes that occurred together more than a user-defined support limit and supported by each "an-item-less subgroup" more than a confidence limit are reported as groups. These groups, which can be used as separate panels for clinical trials, are expected to be biologically meaningful. We will analyze the algorithm, its extensions, its results, and how it is relevant to current practice in the Section VI. A brief description of the algorithm can be found in Table I.

1) *Generality of the Algorithm:* Our algorithm can be applied to high dimensional and heterogeneous data sets such as other pharmaceutical and or microarray data sets. Pharmaceutical data are generally analyzed patient by patient. In the first step of the algorithm, data sources are evaluated separately so that a natural path of analyzing such data is followed. Therefore, our algorithm can be used for any pharmaceutical data that has the same properties as our dataset: variety of data resources, high dimensionality, and a series of observations for each attribute. In applications of microarray data analysis, expression levels of thousands of genes are compared. Evaluations are made at the gene level where a series of observations for each gene are collected. Having these properties makes microarray data sets a suitable input for our algorithm. Many researchers are currently combining microarray experiments with patient datasets as well [8].

V. Distance Metrics for Short Time Series

We utilize several well-established distance functions to measure similarity between time series data: dynamic time warping (DTW), Euclidean, correlation coefficient, and qualitative distance. An important goal is to show the strong relationships between patterns of analytes regardless of the metric used to compare them. Therefore, we use metrics that are frequently used in different disciplines, and that have different capabilities and drawbacks. Among these metrics, DTW and qualitative use a local distance metric inside. To improve the quality of these similarity distances, we propose two novel distance metrics: mean-wise comparison (MWC) and Slope-wise Comparison (SWC). A concise description of each metric will be given in the following sections. The time series compared by the metrics are called "series$_1$" and "series$_2$" in all sections.

A. Correlation Coefficient and Euclidean

A correlation coefficient finds the positive correlation between two length n time series. We used $\sqrt{1-r}$ as a distance metric, where r is the Pearson correlation coefficient between two time series. This metric is very good at detecting linear relationships between entities; however, it cannot be used to detect nonlinear or nonmonotonous relationships [9], [10], and is poorly estimated for short time series [11].

Euclidian distance metric and correlation coefficient are the most commonly used metrics for time series analysis. Euclidean distance defines the distance between two length n time series, by first finding the distance between ith entries and combining these n distances. The main drawback of this metric is that it mainly captures the difference in the scale and baseline.

B. Qualitative Approach

This metric compares movements between all possible (i, j) pairs; this means that for a series of length n, it compares movements from first entry to each one of remaining $n-1$ entries. Since the movement from first entry to second entry is compared for the first one, movements from second entry to the remaining $n-2$ entries are compared for the second one, and so on. The sum of distances of all pairs is divided into the total number of pairs which is $n*(n-1)/2$, and is returned as the qualitative distance.

The qualitative metric is shown to be effective for short time series [11]. It captures the similarity between patterns of changes of time series regardless of whether the nature of the dependence between them is linear or nonlinear [11]. Qualitative distance uses a local distance function to compare the relationship between ith and jth entries of series$_1$ to the relationship between same entries of series$_2$. So, the local metric must be capable of comparing the relationship between the movements. We have proposed slope-wise comparison (SWC), which is discussed in Section V-D, to make this comparison.

C. Dynamic Time Warping (DTW)

The alignment of temporal patterns by DTW has been extensively used in speech recognition [12] and time series studies

TABLE II
RULES OF THE SWC METRIC

$(y1,y2)$ \ $(x1,x2)$	*Increasing*			*Decreasing*		
Increasing		$AX < pt$	$AX >= pt$		$AX >= nt$	$AX < nt$
	$AY < pt$	0	0.25	$AY < pt$	0.5	0.75
	$AY >= pt$	0.25	0	$AY >= pt$	0.75	1
Decreasing		$AX < pt$	$AX >= pt$		$AX >= nt$	$AX < nt$
	$AY >= nt$	0.5	0.75	$AY >= nt$	0	0.25
	$AY < nt$	0.75	1	$AY < nt$	0.25	0

[13], [14]. DTW uses another metric, which is also called the local distance metric, to compare point i of series$_1$ to point j of series$_2$, where i and j do not necessarily need to be equal. Our algorithm uses two different local metrics: square of series$_1(i)$ − series$_2(j)$, where series$_p(r)$ is the rth entry of series$_p$, and MWC. Since Euclidian distance uses the same method to compare ith entries of series, the first local metric was accepted as Euclidian based, and details about MWC can be found in Section V-E. Due to space limitations, we are not covering this metric in detail; the details can be found in [13], [14].

D. Slope-Wise Comparison (SWC)

The SWC metric takes four inputs, x_1, x_2, y_1, and y_2, and compares the relationships between x_1, x_2, and y_1, y_2. There are five possible distances which can be assigned: $0, 0.25, 0.5, 0.75$, and 1. The method is called slope-wise comparison, but instead of the duration, the sum of absolute values of x_1, x_2 is used to find an artificial slope. Let AX be the artificial slope between x_1 and x_2, and AY the slope between y_1 and y_2.

$$AX = \frac{x_2 - x_1}{|x_2| + |x_1|} \quad \text{and} \quad AY = \frac{y_2 - y_1}{|y_2| + |y_1|}.$$

These artificial slopes are compared to positive threshold (pt) and negative threshold (nt) in order to determine the distance between these two pairs. The rules of SWC are defined in Table II.

E. Mean-Wise Comparison (MWC)

MWC takes four inputs, X_i, Mean_X, Y_i, and Mean_Y, where X_i and Y_i are ith points of the series X and Y, and Mean_X and Mean_Y are the means of these series. If both X_i and Y_i are more than or less than the mean of their own series, then distance is set to 0; otherwise, distance is set to 1. There is also a fuzzy region inserted into algorithm if $|(X_i/\text{Mean}_X) - (Y_i/\text{Mean}_Y)|$ is less than a threshold; then distance is also set to 0. As the functionality of this metric depicts, it is a way of discretizing each series according to mean of the series itself.

VI. EXPERIMENTAL RESULTS

A. Experiment Setup

A sample of data is taken as input to the software we developed. Details of how this sample is selected were given in Section IV. The selected sample for the experiments has 26 analytes and 152 patients and k equals 4; i.e., each patient has more than four observations for each of the 26 analytes. As a result, for each patient there are 26 short time series to analyze. As shown in Table I, inputs for the tool we developed are the total number of patients (152), total number of analytes (26), distance metric to compare short time series, confidence limit, and support limit.

The first output of the algorithm is to find the most intra-related groups. These groups will be listed for each distance metric for Sections VI-B to F. The only difference between the experiments in these sections and the experiments in Section VI-G is value of the support limit, which was discussed in the algorithm. For each distance metric of each setup, the output of the algorithm is compared with a list of clinical panels provided by the experts.[1] New groups will be reported, and the interpretation of experts will be given for each of these new groups. For the results indicated in the sections except VI-G, the initial support limit (sp) is 45, and the initial confidence limit (cl) is 0.4.

B. Distance Metric = Correlation Coefficient

Group 1 and Group 5 in Table III are the groups which were not provided by our experts; however, they were found by the algorithm. The main function of WBC is to fight infection, and neutrophils are the main defenders of the body against acute infection. WBC count is the total number of leukocytes (white blood cells) per unit volume of blood. Neutrophils (%) is the proportion of leukocytes that are neutrophils. Neutrophils (abs) are total number of neutrophils in a unit volume of blood. Thus Group 1 can be referred to as Acute Infection Group. Hemoglobin is the main transporter of oxygen in the blood. Moreover, the main function of red blood cells (RBC) is to carry the hemoglobin to the tissues. This process is possible through the RBC containing hemoglobin, which combines easily with oxygen and releases it at the tissue sites. Hematocrit is the measurement of the volume percentage of red blood cells in whole blood. Albumin transports drugs, hormones, calcium, and many other components of the blood. Hence, Group 5 can be identified as the Transporter Group. Total Protein, Albumin and Globulin is the Serum Protein Group. Our algorithm adds calcium to the first two members of this group in Group 2 because albumin is a major transporter of calcium in the serum. The same element (calcium) is added to total protein and serum globulin (Group 3) by the algorithm for the same reason. Since total protein is albumin plus globulin, Group 2 and Group 3 are equivalent in some sense.

It seems that the Albumin and Globulin pair does not have sufficient support to appear in the same group according to the algorithm. This is reasonable, because albumin is synthesized

[1]Donald C. Trost, M.D., Ph.D., Pfizer Inc. Andrej Rotter, Ph.D., Department of Pharmacology at The Ohio State University (OSU).

TABLE III
CORRELATION COEFFICIENT, SP $= 45$, CL $= 0.4$

Group 1	WBC Count, Neutrophils(%), Neutrophils(abs)
Group 2	Total Protein, Albumin, Calcium
Group 3	Total Protein, Globulin, Calcium
Group 4	SGOT(AST), SGPT(ALT), LDH
Group 5	Hemoglobin, Hematocrit, RBC Count, Albumin

TABLE IV
EUCLIDEAN, SP $= 45$, CL $= 0.4$

Group 1	Hemoglobin, Hematocrit, RBC Count
Group 2	Hemoglobin, Hematocrit, Total Bilirubin
Group 3	WBC Count, Neutrophils(%), Neutrophils(abs)

only in the liver, while other proteins come from many other sites. SGOT, SGPT, and LDH (Group 4) are enzymes which leak from the liver and are given as three members of the Liver Group. Clinically, alkaline phosphate is usually considered as the fourth member of this group However, it does not show up as a member of the Group 4. Since it is prominent in both liver and bone, and leaves the liver by another route, this is not surprising. This implies that according to the algorithm, the strength of the connection between the other three members of *Liver Group* and alkaline phosphate is not adequate.

C. Distance Metric = Euclidean

Group 1 and Group 3 (Table IV) are effectively the same as Group 1 and Group 5 of Section VI-B. Total bilirubin was added to Hemoglobin, Hematocrit pair to form Group 2 by the algorithm. This group was consistent, because bilirubin is a waste product of hemoglobin. However, it was not represented in Table IX, because this group was only reported by this metric. Total bilirubin in the blood in also a function of liver metabolism.

D. Distance Metric = Qualitative

After changing the distance metric to qualitative metric, the total number of significant groups is also 5 like the previous case. We already know Group 1, Group 3, Group 4, and Group 5 (Table V) from Section VI-B. Platelets and WBC count with RBC count together forms the Bone Marrow Group. However, the results (Group 2) indicate that RBC count is not strongly related to this group according to this distance metric. This suggests that Group 2 is reflecting granulocyte production in the bone marrow, rather than RBC production or oxygen transport. If a new metric could be defined according to the series of corresponding analytes, then RBC count could show up with the other members of this group.

E. Distance Metric = DTW (Euclidean)

As discussed in Section V-C, DTW uses a local distance metric to compare point i of series 1 to point j of series 2. For the experiment in this section, Euclidean is selected as this local metric. The resulting groups (Table VI) of this combination are also reported earlier. Group 1 corresponds to the Group 1 of Section VI-D and Group 2 corresponds to Group 3 of Section VI-D.

TABLE V
QUALITATIVE(SWC), SP $= 45$, CL $= 0.4$

Group 1	Hemoglobin, Hematocrit, RBC Count
Group 2	Platelets, WBC Count, Netrophils(abs)
Group 3	WBC Count, Neutrophils(%), Neutrophils(abs)
Group 4	Total Protein, Albumin, Globulin
Group 5	SGOT(AST), SGPT(ALT), LDH

TABLE VI
DTW (EUCLIDEAN), SP $= 45$, CL $= 0.4$

Group 1	Hemoglobin, Hematocrit, RBC Count
Group 2	WBC Count, Neutrophils(%), Neutrophils(abs)

TABLE VII
DTW (MWC), SP $= 45$, CL $= 0.4$

Group 1	Hemoglobin, Hematocrit, RBC Count
Group 2	Hemoglobin, RBC Count, Albumin
Group 3	Hematocrit, RBC Count, Total Protein
Group 4	Hematocrit, RBC Count, Albumin
Group 5	Hematocrit, RBC Count, Calcium
Group 6	WBC Count, Neutrophils(%),Neutrophils(abs)

F. Distance Metric = DTW (MWC)

Group 1 and Group 6 (Table VII) are equal to Group 1 and Group 2 of Section VI-E. In addition to hemoglobin and RBC count (from Group 1), the algorithm adds the albumin member of "serum protein group" and forms Group 2. The backbone of Groups 3 to 5 is hematocrit, and RBC count. In Group 3, total protein was added to this backbone, while albumin was added in Group 4 and calcium in Group 5 by the algorithm. Members of Groups 2 to 5 are subsets of the union of Group 3 and Group 1 of Table IX. These four groups may show that the inner relationship between these two groups is too high.

G. Metric-Independent Results

1) *Common Groups:* Although each distance metric captures certain types of similarities, when the support limit is sufficiently high, the proposed algorithm consistently gives the same set of groups independent of the underlying distance metrics. When the support limit is set to 60 with a confidence of 0.4, there are two groups left for all distance metrics, as presented in Table VIII. This clearly states that the intrarelationships for these groups are significantly higher than intrarelationships of others. After evaluating the results of experiments with five different distance metrics, our algorithm concludes that the strength of the inner relationships within these two groups does not depend on the distance metric, but on the nature of the analytes in these sets. This result confirms the existence and the strength of these groups, which were not given by the experts, but found by the algorithm.
2) *Ensemble Algorithm:* Each similarity metric results in different sets of clusters. We further consider the set of clusters by each metric as a different data source, and give it as an input to our algorithm. As a result, the first step of our algorithm produces five times more sets of clusters than when a single distance is used. Table IX shows the result of this algorithm on our data set. Four groups of analytes

TABLE VIII
ALL METRICS SP $= 60$, CL $= 0.4$

Group 1	Hemoglobin, Hematocrit, RBC Count
Group 2	WBC Count, Neutrophils(%), Neutrophils(abs)

TABLE IX
ENSEMBLE GROUPS

Group 1	Hemoglobin, Hematocrit, RBC Count
Group 2	WBC Count, Neutrophils(%),Neutrophils(abs)
Group 3	Total Protein, Albumin, Calcium
Group 4	SGOT(AST), SGPT(ALT), LDH

are identified. Not surprisingly, the two common groups still show up with two more groups shown in Table IX.

VII. FEATURE SELECTION ALGORITHM

Utilizing the proposed algorithm, we now propose a method for feature selection which identifies a global panel that models human health. Identifying such a panel will not only reduce the number of analytes to a manageable size; it will also be a key to define what the normal is. A small change to our algorithm leads us to find such global panels whose elements have minimal intrarelationships. After identifying such panels of analytes, we will also calculate how often the biological groups found in the previous section (shown in Table IX) are represented in these panels.

Appearing in the same cluster is the complement of not being in the same cluster. By the same token, if the algorithm can determine which analytes occurred in the same cluster more than the support limit, it can also determine which of them did not come together more than the difference between the number of patients (data sources) and the support limit. We define the new support limit as "complement support limit."

Complement support limit, which is utilized to find a group with most unrelated analytes, is calculated by subtracting the support limit for most related groups (45) from the total number of patients (data sources). Since increasing support limit does mean decreasing complement support limit, the results for "60" will grow and managing this output will become more difficult when compared to the one for "45." In order to calculate how many times the set of analytes dispersed, the first algorithm calculates how many times they came together, and takes the complement of this number (the total number of patients minus this number). Based on these values sets, most unrelated groups are formed.

Our algorithm gave two analyte-panels of size eight (largest set for this metric) for the correlation coefficient metric. These panels are as follows:

1) *Selected Features set-1:Hematocrit, Neutrophils(%), Total Bilirubin, Globulin, SGOT(AST), BUN, Creatinine, Phosphorus*; and
2) *Selected Features set-2:Hematocrit, Total Bilirubin, Globulin, SGOT(AST), BUN, Creatinine, Phosphorus, Neutrophils(abs).*

Intersection of these two groups is seven, and only neutrophils(%) and Neutrophils(abs) are different for these two selected features sets. They are members of the same group (Group 3 of Table IX). This results can be interpreted such that one of them is enough to represent whole group and replacing one with another also forms a set of selected features. This is reasonable, because Neutrophil(%) $=$ Neutrophil(abs)/WBC.

TABLE X
ANALYTES AND THEIR GROUPS

GROUP	**ANALYTE(s)**
Liver Functions	SGOT(AST), Total Bilirubin
Liver Enzymes	SGOT(AST)
Proteins	Serum Globulin
Renal-related	BUN, Creatinine
Hematology	Hematocrit
Electrolytes	Phosphorus
Immune System	Neutrophils

TABLE XI
PERCENTAGE

	Group 1	Group 2	Group 3	Group 4
	$1, 2, 3$	$7, 8, 608$	$24, 25, 26, 58$	$28, 30, 32$
$CorrCoeff$	100.00	86.67	80.00	93.33
$Qualitative$	97.06	100.00	69.12	100.00
DTW_{EUC}	100.00	100.00	100.00	100.00
DTW_{SWC}	38.46	100.00	61.54	100.00
$Euclidean$	68.25	100.00	97.63	58.77

Common laboratory tests in liver diseases are divided into the following groups: liver-related, hematology, electrolytes, proteins, lipids, renal-related, thyroid, and immune system [15]. Our data set contains seven groups: liver function, liver enzymes, hematology, electrolytes, proteins, renal-related, and immune system. A global panel should include at least one representative from each of these seven groups. The proposed algorithm has this property. The list of analytes in the global panel, and all major biological groups and their representatives in this panel, is given in Table X.

Members of each group shown in Table IX are strongly correlated. A member of these panels is expected to represent the behavior of the whole group. Hence, one member from each group is expected to appear in the global panel. For each group given in Table IX, the proportion of being represented in the global panels given by the algorithm is calculated and presented in Table XI for each distance metric. Note that the average appearance of a representative from each group is higher than 80%.

VIII. RELATED WORK

Sequence data mining and clustering [16] have been extensively studied from the perspective of efficiently searching or extracting rules from them [17]. Das *et al.* [18] first cluster all fixed length subsequences of each one-dimensional stock time series; then, based on the labels of the clusters, a sequence mining algorithm is executed to find interesting rules. A similar approach can be developed for clinical trials data to see if the drug has an obvious effect on a single analyte; however, it won't capture multivariate relationships between analytes.

There have been several studies directly applying association rule mining for medical data. Ordonez *et al.* [6] map patient data into transactional data, and run an association rule mining algorithm to find interesting rules. Similarly, Doddi *et al.* [19]

obtained association rules that show relationships between medical procedures and the corresponding diagnosis. These algorithms consider only one observation for each attribute, and are not applicable to multidimensional time-series data. Ohsaki *et al*. [20] propose a discretization technique based on sequence extraction and conversion into patterns by clustering. Since only a single series per patient is considered, the relationship between different series of the same patient are not taken into account. Also, time series are required to have sufficient length for the algorithm to be applicable.

IX. Conclusion

We have proposed a novel approach for mining heterogeneous and high dimensional time series data that are commonly found in a large number of applications. We focused on mining clinical trials data and illustrated our findings on an example industry-sponsored data set. Global mining of a clinical trials data set is infeasible because of the high heterogeneity and dimensionality of the data. Therefore, the proposed technique included two steps: first, homogeneous subsets of the data are identified and locally analyzed, and then information gathered in this step is mined to identify global patterns. This two-step process minimizes the differences among time series caused by source variation and, as a result, local groups of time series become equal in length, and equal interval. Therefore, well known distance metrics, such as Euclidean, correlation coefficient, qualitative, and DTW, that are generally suitable for equal series, become eligible to find the distances between heterogeneous series. We have proposed two new similarity distance metrics that are suitable to the nature of the clinical trial data. Using these distance metrics, we cluster the series of attributes in each data source. In the second step, the information gathered in the first step; i.e., the clusters, is further mined to find groups of attributes that occur in all subsets. Thus, we refer to our approach as information mining rather than data mining; the data are too dirty to be mined as is. Our results have verified two biological panels (of blood analytes) already well-known in the medical field. Besides the well-known groups, we have also identified biological groups that are not commonly used.

A slight change to our algorithm leads to a novel technique for feature selection for high dimensional pharmaceutical clinical trials data. It identifies a global panel of analytes that effectively represents the normal state of health. The panel consists of eight analytes that are found to be most unrelated. The panel successfully includes at least one instance from each of the seven major analyte groups present in the human body.

The proposed algorithm is general and can be applied to pharmaceutical and clinical data, as well as other high dimensional and heterogeneous data sets. It illustrates the interactions between blood analytes that is crucial to understand the human body. Our colleagues involved in this industry-sponsored project are already using the results of the proposed technique. The statisticians use the strongly related groups of analytes and the global panel of analytes in developing accurate data distribution tests (such as multivariate normality) for high dimensional clinical trial data sets. The mathematicians involved in the project use the results in modeling the behavior of the human body in response to drug treatments. We expect several other uses of these results by researchers in pharmacology, biomedicine, and biology. One such use is prediction of health state, value of an analyte, and diagnosis. The proposed technique can also be used to significantly reduce the number of blood analytes needed to be tested in pharmaceutical and medical studies.

Acknowledgment

The data sets for this project were provided by Pfizer Inc. The authors are grateful to the OSU Pfizer group: A. Campbell, A. Friedman, P. March, J. Ostroff, S. Parthasarathy, D. Pearl, and A. Rotter for their help in this project.

References

[1] FDA, *Innovation or stagnation: Challenge and opportunity on the critical path to new medical products*,, Mar. 2004

[2] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data*. New York: Wiley, 1990.

[3] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in very large databases," *Proc. ACM SIGMOD Conf. Management of Data*, pp. 207–216, 1993.

[4] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur "Dynamic itemset counting and implication rules for market basket data," SIGMOD, Record 6., pp. 255–264, 1997.

[5] R. J. Bayardo Jr., R. Agrawal, and D. Gunopulos, "Constraint-based mining in large, dense databases," in *Proc. 15th Int. IEEE Computer Soc. Conf. Data Engineering*, Sydney, Austrialia, Mar. 1999, pp. 23–26.

[6] C. Ordonez, C. A. Santana, and L. Braal, "Discovering interesting association rules in medical data," *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 78–85, 2000.

[7] P. Soldacki and G. Protaziuk, "Discovering interesting rules from financial data," in *Conf. Data Mining and Warehouses SiKDD*, 2002, pp. 109–119.

[8] D. Beer *et al*., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Med*., vol. 9, p. 816, 2002.

[9] M. G. Walker, "Pharmaceutical target identification by gene expression analysis," *Mini Reviews in Medicinal Chemistry*, vol. 1, pp. 197–205, 2001.

[10] M. G. Walker, "Introduction to statistics for bioinformatics, with applications to expression microarrays," *Computational Systems Bioinformatics (CSB), Bioinformatics Conf. Tutorial Sessions*, Stanford, CA, 2003.

[11] L. Todorovski, B. Cestnik, and M. Kline, "Qualitative clustering of short time series: A case study of firms reputation data," *Information Society, Proc. Int. Multi-Conf*., vol. A, pp. 143–146, 2002.

[12] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process*., vol. 26, no. 1, 1978.

[13] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping to massive datasets," in *Proc. 3rd European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD'99)*, 1999, pp. 1–11.

[14] T. M. Rath and R. Manmatha, "Lower-bounding of dynamic time warping distances for multivariate time series," Center for Intelligent Information Retrieval, Univ. Massachusetts, Amherst, Tech. Rep. MM-40, 2003.

[15] J. J. Jaeger and H. Hedegaard, *"Liver function tests and blood tests"*, Feb. 2002. [Online]. Available: http://home3.inet.tele.dk/omni/alttest.htm

[16] K. J. Cios and L. A. Kurgan, *"Trends in Data Mining and Knowledge Discovery," in Knowledge Discovery in Advanced Information Systems, Springer, 2002*.

[17] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," *J. Knowledge and Information Systems*, vol. 3, no. 3, 2001.

[18] G. Das, K. Lin, H. Mannila, G. Renganathan, and P. Smyth, "Rule discovery from time series," in *Proc. 4th Int. Conf. Knowledge Discovery and Data Mining*, New York City, 1998, pp. 16–22

[19] S. Doddi, A. Marathe, S. S. Ravi, and D. C. Torney, "Discovery of association rules in medical data," *Med. Inform. Internet Med.*, vol. 26, no. 1, pp. 25–33, 2001.

[20] M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi, "A rule discovery support system for sequential medical data—In the case study of a chronic hepatitis dataset," in *Proc. IEEE Int. Conf. Data Mining ICDM, Int. Workshop Active Mining*, Maebashi City, Japan, Dec. 2002, pp. 97–102.

Fatih Altiparmak received the B.S. degree in computer engineering from Bilkent University, Turkey, in 2002.

Since 2003, he has been a Graduate Research Assistant in the Database Group at The Ohio State University, Columbus. His research interests include analyzing clinical trials databases, time-series analysis, distance/similarity metrics and data integration.

Hakan Ferhatosmanoglu received the Ph.D. degree from the Computer Science Department, University of California, Santa Barbara, in 2001.

He is an Assistant Professor of computer science and engineering at The Ohio State University (OSU), Columbus. Before joining OSU, he worked as an intern at AT&T Research Labs. During his Ph.D. studies, he proposed several techniques for efficient retrieval and scalable storage of large-scale multidimensional data. His current research interest is to develop data management systems and applications for physical, medical, and biological sciences. He leads projects on microarray and clinical trial databases, online compression and analysis of multiple data streams, and high performance databases for multidimensional data repositories.

Dr. Ferhatosmanoglu is a recipient of the Early Career Principal Investigator Award from the U.S. Department of Energy.

Selnur Erdal received the Dental degree from the College of Dentistry, Ege University, Izmir, Turkey, in 1997, and the M.S. degree in computational biology and bioinformatics from the biophysics program at The Ohio State University, Columbus, in 2005.

Since 2005, he has been a Graduate Research Assistant in the Database Group at The Ohio State University. His research interests include time-series analysis, multidimensional indexing techniques and scientific visualization.

Donald C. Trost received the M.D. and M.S. degrees in preventive medicine and environmental health from the University of Iowa, Iowa City, in 1978, and the Ph.D. degree in biostatistics from the University of North Carolina, Chapel Hill.

He is currently Senior Director and Global Head of Mathematical Medicine in Clinical R&D at Pfizer Inc., Groton, CT. He served an NHLBI postdoctoral fellowship in cardiovascular biostatistics with the Lipid Research Clinics Program at the University of North Carolina, Chapel Hill and a residency in Clinical Pathology with an emphasis in medical informatics at the University of Florida, where he became a member of the faculty. His current interests include the application of mathematics and computer algorithms to quantifying patterns in dynamic biomedical data, especially in the area of drug safety. He has held clinical, computational, statistical, and management positions in several medical software and pharmaceutical companies.