Evaluating Statistical Tests on OLAP Cubes to Compare Degree of Disease

Carlos Ordonez, Zhibo Chen
University of Houston
Houston, TX 77204, USA

***Abstract*—Statistical tests represent an important technique used to formulate and validate hypotheses on a data set. They are particularly useful in the medical domain, where hypotheses link disease with medical measurements, risk factors and treatment. In this work, we propose to compute parametric statistical tests treating patient records as elements in a multidimensional cube. We introduce a technique that combines dimension lattice traversal and statistical tests to discover significant differences in the degree of disease within pairs of patient groups. In order to understand a cause-effect relationship we focus on patient group pairs differing in one dimension. We introduce several optimizations to prune the search space, to discover significant group pairs, and to summarize results. Our proposal can work on any relational database system since it is based on dynamically generated SQL code. We present experiments showing important medical findings and evaluating scalability with medical data sets.**

I. Introduction

Medical data sets are commonly analyzed with statistical [11], [20] and machine learning [13], [18] techniques. Such techniques range in complexity from computing descriptive statistics such as the mean, variance and histograms, to building complex predictive models. In this work we focus on improving and extending statistical tests. In terms of complexity and computations, statistical tests are more sophisticated than descriptive statistics, but simpler than predictive models. Statistical tests stand out for their wide applicability in the medical domain given their simplicity, flexibility and reliability [10], [20]. In general, many experiments (e.g. clinical trials, survey analysis [20]) are performed to discover an important medical fact, which is validated with a statistical test. The statistical test is based on a hypothesis about the statistical properties of the data set being analyzed. The overall goal is to find a hypothesis having high reliability (confidence). This problem becomes more difficult when the user needs to analyze multiple subsets of a data set, considering different combinations of attribute values (risk factors).

On-Line Analytical Processing (OLAP) [5], [10] is a collection of exploratory database techniques [8], [10]. In an OLAP database, users try to find interesting or unexpected results by analyzing subsets of a data set with aggregations performed at multiple granularity levels. OLAP techniques offer promising application in the medical domain, since they are intuitive, comprehensive and efficient [5], [8], [10]. In OLAP most computations are based on simple aggregations such as sums and counts, which are the basic ingredients used in statistical tests. Therefore, we propose to integrate statistical tests and OLAP techniques in order to formulate and validate medical hypotheses on multiple subsets of a data set. Our proposal extends statistical analysis with database techniques to improve disease understanding and diagnosis. Figure 1 presents an overview of the data mining process.

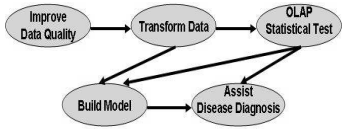

Fig. 1. OLAP statistical tests within data mining process.

Statistical tests [20] exhibit several advantages. They have simple assumptions about the probability distribution behind the data set. Such distribution is in general the normal (Gaussian) distribution, or a closely related probabilistic function. Such assumption is not an issue when there are many observations (e.g. ≥ 30) and when there are few outliers [20]. Most statistical tests use equations that can be efficiently evaluated with SQL queries in a database system, because they do not require matrix manipulation, making them efficient and easy to implement. Statistical tests can produce statistically reliable results with both large and small data sets, whereas statistical and machine learning techniques commonly require larger data sets in order to find significant results; such limitation is particularly important in medicine, where in studies about a specific type of disease or in clinical trials on new drugs or treatments, only small sets of patient records are available. Also, different institutions do not share patient records due to government and privacy regulations. Statistical tests are frequently used as an initial step to building predictive models. Therefore, our proposal is not intended to substitute predictive models, but to complement them. For instance, the statistical test can filter risk factors that can then be used as input for a predictive model. It can help understanding why a predictive model has low accuracy or can validate a model whose reliability is questionable. Statistical tests can enhance a decision support system by providing a mechanism to validate knowledge or

http://doi.ieeecomputersociety.org/10.1109/TITB.2008.926989

intuition about a disease or treatment, by comparing a patient to other patients having a similar medical profile.

Statistical tests have practical disadvantages. They generally require many trial and error runs before a plausible finding is made. Each run requires selecting different variables, varying parameters or selecting subsets of the data set. Nevertheless, we will show both of these limitations can be solved by OLAP techniques. On the other hand, compared to OLAP techniques, statistical tests provide more evidence that a finding is indeed valid, by going beyond simple comparisons or proportions. Our proposal combines statistical tests with OLAP cube exploration techniques in order to automatically generate and test hypotheses for a large collection of subsets from a medical data set. Since the dimension lattice behind the cube represents a combinatorial search space the problem is computationally challenging. We thereby introduce techniques to efficiently evaluate statistical tests on OLAP cubes and to produce a succinct set of significant findings. Our method can work with large data sets, as required in a modern analytic environment. We illustrate our approach with experiments on two medical data sets containing patients with heart disease (stenosis) and thyroid gland abnormality, respectively. The statistical test can uncover combinations of risk factors leading to a greater degree of disease.

The article is organized as follows. Section II introduces basic definitions for OLAP databases and statistical tests. Section III explains how to apply statistical tests on OLAP cubes and introduces several optimizations. Section IV presents an experimental evaluation with medical data sets, showing important medical findings and evaluating optimizations. Related work is discussed in Section V. The conclusions and directions for future work are presented in Section VI.

II. Definitions

We focus on an input table F with n records having d cube dimensions [10], $D = \{D_1, \ldots, D_d\}$ and a set of e measure [10] attributes $A = \{A_1, A_2, \ldots, A_e\}$. The data structure representing all subsets of dimensions and their containment is called the dimension lattice [10]. Due to their simplicity and wide application in the medical domain, we focus on binary dimensions, but our proposal also works for categorical dimensions with higher cardinality. In the case of medical data sets, risk factors for a certain disease are represented by the cube dimensions and the degree of disease are cube measures. Section II-B contains a detailed example.

A. OLAP Processing

In OLAP processing, the basic idea is to compute aggregations (sum(), count()) on measures A_i by subsets of dimensions (i.e. subcubes or cuboids [10]) G s.t. $G \subseteq D$, effectively performing aggregations at different granularity levels. Each output record obtained by the aggregation is called a group. The set of all potential aggregations at a certain level is called a subcube. In our proposal, aggregations are used to derive descriptive statistics such as μ, σ, which in turn are the basic elements in the equations of a parametric statistical test, introduced in Section III.

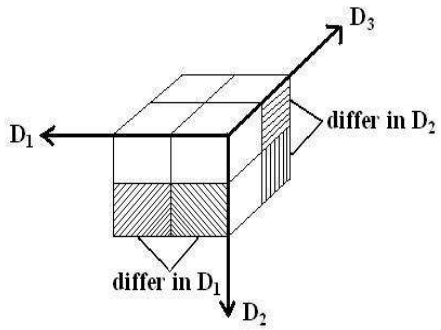

Fig. 2. Finding significant measure differences in the cube.

B. Example

Figure 2 illustrates our approach on a cube with three dimensions D_1, D_2, D_3 and one measure, where each face represents a subcube on two dimensions. We highlight two pairs of groups in two subcubes that differ on one dimension (i.e. some risk factor). The difference in fill pattern means a significant difference was detected on a measure attribute (i.e. degree of disease). Consider, for instance, a medical data set with patients being treated for heart disease. Each dimension D_j can represent a risk factor (e.g. hypertension) or some binned numeric attribute (e.g. cholesterol level). The measure A_h can represent the degree of disease (narrowing) in an artery. In this case, the statistical test is stating that each highlighted pair of patient groups has a significant difference in the degree of disease in an artery (comparing measure mean value for each patient group), which is probably caused by the "discriminating" dimension (i.e. the risk factor that is different across the pair).

III. Statistical Tests and OLAP

This section presents the main contributions of our proposal. We start by motivating and explaining the use of statistical tests on OLAP cubes. Then we introduce an algorithm that explores the cube with different dimension combinations to compare highly similar groups. We discuss several optimizations. Finally, we summarize advantages and disadvantages of our method.

A. Statistical Tests

We propose to use a means comparison parametric test [20] to discover pairs of patient groups with a significant difference in a disease measurement. The basic hypothesis is that a specific risk factor, combined with other risk factors can lead to a high probability of developing disease. Instead of looking for unusual patterns in individual subcubes with simple comparisons and multi-level aggregations like previous work [9], [10], we propose to use the statistical test to compare pairs of groups. The means comparison test offers the following advantages: Two large groups of any size can be compared. In particular, two groups with very different number of elements can be compared (e.g. a large and a

small group). The means comparison takes into account data variance, which measures overlap between the corresponding group subpopulations. This feature is essential to make sound inferences. Measures are assumed to have a normal distribution. The normal distribution assumption works well for medium or large data sets ($n \geq 30$), having few outliers and small skew [20]. That is, a highly skewed or asymmetrical distribution may decrease the reliability of the test.

Parametric Test Statistical Background

We now describe the statistical test in more formal terms. We use a parametric test comparing the means μ_1, μ_2 from two disjoint data subsets (populations), where the size of each data subset is n_1, n_2. Each data subset is assumed to be an independent sample. In this case the null hypothesis H_0 states that $\mu_1 = \mu_2$ and the goal of our proposal is to find group pairs where H_0 can be rejected (deemed false) with high confidence $1-p$, where p generally takes the following thresholds, $p \in \{0.01, 0.05, 0.10\}$. The so-called alternative hypothesis H_1 states $\mu_1 \neq \mu_2$. When H_0 can be rejected the test will return the significance level p; such outcome will allow us to provide strong statistical evidence supporting $H_1 : \mu_1 \neq \mu_2$. We use a two-tailed test which allows finding a significant difference on both tails of the Gaussian distribution in order to compare means in any order ($\mu_1 < \mu_2$ or $\mu_2 < \mu_1$). The statistical test relies on Equation 1 to compute a random variable z with probability distribution function (pdf) $N(0,1)$:

$$z = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}, \tag{1}$$

where μ_i, σ_i correspond to the estimated mean and standard deviation from group i. When both groups are not small (i.e. $n_1 \geq 30, n_2 \geq 30$). the z value just needs to be compared with $z_{p/2}$ in the cumulative probability table for $N(0,1)$ (e.g. $z_{0.495}$). When either group is small the statistical test requires computing the degrees of freedom as $df = n_1 + n_2 - 2$, and then looking up z on the t-student distribution table according to df. This implies that $n_1 > 1$ or $n_2 > 1$ because $df \geq 1$. If one group is much larger than the other one then there exists a row for $df = \infty$. For instance, it is possible to compare a singleton set with a large set of records, although it may be medically irrelevant.

Applying the statistical test on cubes

In general, a statistical analyst needs to perform a batch of statistical tests with different medical measurements and multiple subsets from the data set before a significant finding is made. The medical measurements are in general related to the disease being studied and the subsets are selected based on medical opinion, trying to obtain samples as large as possible. Missing information and irrelevant (unrelated to measurement) dimensions make the problem more difficult. Based on such challenging and time-consuming task, we propose to automatically generate group pairs by applying the statistical test only on pairs differing in a few dimensions. When used in an OLAP cube, the means comparison test has the following advantages: OLAP aggregations can produce many pairs of disjoint groups, which are the required input for the parametric test. Groups produced by the same aggregation query are disjoint. Dimensions can be used to identify highly similar groups, differing in a few dimensions or even one dimension. This can help explaining a cause-effect relationship. From a database perspective, the test represents a natural extension of OLAP computations since it relies on distributive aggregations [9]. That is, we are not introducing a large extra burden on processing. When the cube has low d and small groups are eliminated it is feasible to compute all potential groups pairs traversing the entire dimension lattice.

In the case of a heart disease data set, OLAP aggregations produce disjoint sets of patients having specific risk factors. When groups are compared with each other (forming pairs), the test focuses on groups that have one different risk factor and evaluates the difference in their degree of artery disease. Applying the statistical test has two goals: First, finding significant differences between two groups in a subcube on at least one measure. Finding two or more significant measure differences for the same pair is desirable, but rare. Second, when there exists a significant difference we need to isolate groups that differ in a few dimensions, which can explain cause-effect. With respect to the first goal, when applying a statistical test, a significant difference can only be supported by a small p-value, which takes into account both the means and the standard deviation of the distributions. The smaller the p-value the more likely the difference between both groups is significant. It is expected that many differences will not be significant, making the search problem expensive. Regarding the second goal, we are interested in finding significant differences in highly similar groups because that helps explain which specific dimensions (e.g. risk factors) "trigger" a significant change on the subcube measure (e.g. degree of disease). For instance, finding a significant measure difference, between two dissimilar groups, makes causal explanation difficult, since such difference may by attributed to two or more dimensions. In general, we focus on group pairs differing in exactly one dimension. Groups differing in two or more dimensions are stored on additional tiers. Even though dimensions are considered independent, the aggregation automatically groups records with correlated dimensions together. Therefore, if a high correlation exists in dimensions it will be automatically reflected when patient groups are built by OLAP aggregations.

B. Optimizations

Our algorithm was implemented by dynamically generating SQL queries to search the d-dimensional cube, traverse the lattice, form pairs and compute the parametric test. We found several technical issues in trying to optimize processing. (1) Searching the dimension lattice is the most time-consuming stage. (2) Very small patient groups may not produce medically reliable findings; in particular inferences involving groups with one patient should be avoided. (3) It is not possible to pre-define a general-purpose index for the cube because d may vary and the corresponding columns will be different, given different fact tables F. Instead the

cube table has either a simple primary key to uniquely identify groups (cube cells) or a primary index on all the dimensions. (4) OLAP cubes combined with statistical tests cannot be solved as an "association rule" problem, although there are similarities. Statistical tests on cubes represent a different technique from association rules [10] because binary dimensions are not equivalent to items [1]. In general, items represent only the value 1 of some binary dimension, ignoring the value 0. On the other hand, in the statistical test both 1 and 0 are considered. Second, the test is applied on a pair of groups, which would correspond to comparing the records behind two highly similar itemsets. Finally, the test compares numeric attributes, which has no counterpart in association rules because association rules are defined over binary dimensions (called items); numeric attributes are binned and represented as items.

Based on the issues above we present two sets of optimizations: mathematical optimizations, applicable from a general perspective and database optimizations, useful to generate efficient SQL code in a relational database system.

We present the following mathematical optimizations for efficient statistical test computation: (1) We introduce a frequency threshold ψ, which is somewhat similar to the "support" threshold in association rules. This ψ threshold has the primary purpose of discarding small patient groups, but also applying the test on pairs of large groups with the $N(0,1)$ pdf ($n_1 \geq 30$ and $n_2 \geq 30$). This threshold significantly prunes the search space in the dimension lattice. (2) Sufficient statistics are computed on each cell so that univariate statistics can be derived in one pass. Such statistics include count(*), sum(A) and sum(A^2) for a measure column A. In formal terms, N_i is the count patient records for group i and: $L_i = \sum A$, $Q_i = \sum A^2$ for measure A. Then $\mu_1, \mu_2, \sigma_1, \sigma_2$ can be easily derived from sufficient statistics ($\mu_i = L_i/N_i$, $\sigma_i^2 = Q_i/N_i - \mu_i^2$). (3) Depending on input parameters, the SQL code will compare only groups having up to δ dimension differences, being $\delta = 1$ the default. This means groups in a pair must differ in up to δ dimensions in order to be compared. Dissimilar groups are pruned out.

We introduce the following database optimizations, which are particular to using a relational database system. (1) Search for a specific cell is indexed. The secondary index allows efficient retrieval of cell pairs in one indexed search per group (i.e. two indexed searches per pair). (2) All aggregations are stored on the same table. This table contains all subcubes at different aggregation granularities and has an index on all dimensions, (3) F is aggregated at the finest granular level producing a cube table C and further coarser-grained aggregations are computed from C. This step is important to eliminate empty groups. (4) In general, for large groups it is required to compare the test statistic against a given z value using the normal $N(0,1)$ distribution, which generally requires looking up a value in a small table. We introduce a simple optimization, finding the significance of the test statistic without visiting such table with a CASE statement (if-then in a common programming language) based on the specific p-values commonly used in the medical domain ($p \in \{0.01, 0.05, 0.10\}$). This optimization can be applied because we do not need to know the exact p-value, just the range (i.e. $p < 0.05$ or $p < 0.01$) that the pair falls into. That is, finding $z_{p/2}$ for the two-tailed test is done in main memory. On the other hand, when the group is small an indexed search is performed on the t-student table using df as the search key.

Optimization properties: We summarize the improvement that the proposed optimizations may achieve.

Property 1: When $\psi = 30$ the algorithm reduces to a parametric test with the standard normal distribution for every pair. When $\psi = 10$ the algorithm evaluates the t-student distribution for a pair with a group having $10 \leq n_i < 30$, and discards smaller groups. When $\psi = 0$ the algorithm also evaluates the t-student on pairs with small groups, perhaps having one patient in one group. A high ψ threshold produces mostly general results, whereas a low ψ threshold also includes specific results.

Property 2: Assuming dimensions are binary, an aggregation on k dimensions may produce up to 2^k different groups. Clearly, this may take exponential space. When ψ is used to prune small groups the final number of groups (dimension combinations) will be $< 2^k$. As k grows the number of small groups increases and therefore more groups are pruned out.

Property 3: Suppose an aggregation on k binary dimensions produces m groups. Then there are $\binom{m}{2}$ potential pairs. By property 2 in the worst case there are $2^k(2^k - 1)$ potential pairs. Focusing on pairs with $\delta = 1$ different dimensions there are only $k2^{k-1}$ potential pairs. In general $k2^{k-1} << 2^k(2^k - 1)$; this is a significant reduction in the number of pairs and processing time.

Property 4: For a pair having both groups $N_i \geq \psi$ every subpair with some dimension subset is also above ψ. Conversely, if one group with dimensions X in a pair is below ψ then every group whose dimensions are a superset of X is also below ψ. This is analogous to the well-known downward closure property of association rules [10], [16]. This property is used to prune the exponential search space.

C. Algorithm

We introduce an algorithm that integrates statistical tests with cube exploration and the optimizations introduced above. Our algorithm has the following goals: (1) Searching the cube for groups with a minimum number of patients. (2) Building group pairs differing in δ cube dimensions. (3) Computing the statistical test for every pair. (4) Producing a succinct output with a few significant group pairs.

The algorithm assumes a low d, which can be used to analyze a user-selected set of binary dimensions. In general, the low d assumption is reasonable because there may be medical knowledge on the most important factors for degree of disease. Also, for high d, highly correlated risk factors can be discarded, by using a representative one using dimensionality reduction (e.g. PCA [11]). The algorithm searches the cube for groups above ψ and then applies statistical tests for every pair. We apply a bottom-up traversal on the dimension lattice, working level-wise, building higher-d cubes at each level. The algorithm input and output is as follows:

- Input parameters: a maximum p-value threshold, δ threshold of maximum number of different dimensions (e.g. $\delta = 1$) and ψ, a minimum frequency threshold.
- Output: a table C containing group pairs with a significant measure difference, passing the test at a small p-value, with δ different dimensions.

The algorithm steps are the following:

1) Precompute cube with d dimensions $D_1, \ldots, D_d$ on all e measures getting groups at the finest granularity level.
2) Compute groups having $N_i \geq \psi$ by traversing the dimension lattice bottom up, computing groups with 1 dimension, 2 dimensions and so on. Prune out supersets of groups below ψ. Compute sufficient statistics N, L, Q for every measure $A_1, \ldots, A_e$ for each group.
3) Compute pdf parameters μ, σ, based on N, L, Q for each group. This is done for each measure $A_1, \ldots, A_e$.
4) Create group pairs with groups differing in at most δ dimensions (in general, $\delta = 1$).
5) Compute a statistical test for every group pair and every measure using Equation 1.
6) Filter pairs having a significant difference ($z > z_{p/2}$ using $N(0,1)$ or t-student pdf, depending on $N_1 \geq 30, N_2 \geq 30$). This is done for each measure. A pair is eliminated if there is no significant difference in any measure.
7) Optionally eliminate redundant pairs. If a pair X is significant then every pair Y, s.t. $X \subset Y$, is eliminated.

Table F is first aggregated at the finest granular level as given by $D_1, \ldots, D_d$ and then the cube exploration proceeds with a bottom-up search having $2, 3, \ldots,$ dimensions. Given one level of aggregation the algorithm performs a pair-wise test-based comparison of all group pairs. Such comparison is made on each of the measure attributes. The algorithm tries a list of increasing p-values, in order to find the smallest p value at which H_0 can be rejected. When $p > 0.1$ the parametric test shows there is no evidence $\mu_1 \neq \mu_2$. When $\delta > 1$ we categorize results into tiers having 1,2,3 dimensions differences, up to δ. Notice we eliminate redundant pairs containing the same discriminating dimension combined with other matching dimensions. Only when a discriminating dimension appears alone all pair supersets are discarded. Indexing with null values prevents efficient search. Therefore, a search for a specific cell requires handling nulls and "All" separately. Nulls represent missing information and "All" means the corresponding dimension was not used as a grouping column in SQL. In particular, they were coded as negative integer values (i.e. codes different from 0 and 1) in order to allow for expansion to categorical (nominal) values (dimensions with more than just 0 and 1). The summary table has an index on all dimensions, coding "All" and "null" separately.

D. Advantages and Disadvantages

Advantages include the following. Our method automates a time-consuming task to try to discover which specific combinations of risk factors are associated with significant differences in the degree of heart disease. Since the method is based on a comprehensive parametric statistical test it can handle and compare large data sets (thousands of records) or small samples (no more than one hundred records), which are common in clinical trial studies. Our method allows the detection of risk factors triggering disease in two or more arteries, which can be interpreted as a complex multitarget predictive task. Compared to machine learning or similar techniques, our proposed method performs feature (variable) selection for each group pair. It does not require to create binary targets (predicted class); it is not sensitive to imbalanced populations; and it finds many patient group pairs instead of building a single model. Disadvantages include the following. The statistical test assumes measures (degree of disease) follow a normal distribution, which may decrease its reliability when distributions are highly skewed. However, the statistical test considers the variance (Equation 1) to compare populations, which mitigates issues with skewed or asymmetrical distributions. The method incorporates several optimizations and pruning strategies to make processing faster, but the search space is still combinatorial. Therefore, the method is applicable on data sets of any size, but analyzing no more than 20 risk factors at one time.

TABLE I
ATTRIBUTES OF HEART DATA SET.

Attribute	Description
Dimensions:	
OLDYN	Age ≥ 60
SEX	Gender
HTA	Hyper-tension Y/N
DIAB	Diabetes Y/N
HYPLD	Hyperlipidemia Y/N
FHCAD	Family history of disease
SMOKE	Patient smokes Y/N
CLAUDI	Claudication Y/N
PANGIO	Previous angina Y/N
PSTROKE	Prior stroke Y/N
PCARSUR	Prior carot surg Y/N
HIGHCHOL	High Cholesterol
Measures:	
LM	Left Main
LAD	Left Anterior Desc.
LCX	Left Circumflex
RCA	Right Coronary

IV. EXPERIMENTAL EVALUATION

Our experiments were performed on the Microsoft SQL Server DBMS running on a computer with a CPU at 3.2GHz, 4GB of memory and 600GB on disk. The SQL code generator computing the parametric test on the OLAP cube was developed in the Java language and queries were submitted through the Java DataBase Connectivity (JDBC) interface. Times are measured in seconds.

A. Data Sets

For a medical doctor it is difficult to judge how sick or healthy a patient is and which are the specific causes for the disease. The following experimental evaluation is based on two medical data sets, that contain detailed medical profiles for diseased patients with numeric attributes, that measure a degree of disease and binary dimensions, which represent

TABLE II
ATTRIBUTES OF THYROID DATA SET.

Attribute	Description
Dimensions:	
OLD	Age ≥ 60
GENDER	Male/Female
THYROXINE	On Thyroxine Y/N
ANTITHYROID	On Antithyroid Y/N
SICK	Sick Y/N
PREGNANT	Pregnant Y/N
SURGERY	Prior Thyroid Surgery Y/N
TUMOR	Tumor Y/N
I131TREATMENT	On I131 Y/N
LITHIUM	On Lithium Y/N
GOITRE	Goitre Y/N
HYPOPITUITARY	Hypopituitary Y/N
HYPOTHYROID	Hypothyroid Y/N
HYPERTHYROID	Hyperthyroid Y/N
Measures:	
TSH	Thyroid Stimulating Hormone
TT4	Total Thyroxine
T4U	Thyroxine Utilization Rate
FTI	Free Thyroxine Index

risk factors or demographic information on each patient. The OLAP statistical test will attempt to uncover highly similar patient groups (with only one different risk factor) with a substantially different degree of disease (their means are significantly different). The first medical data set contains profiles of $n = 655$ patients and has 25 attributes containing categorical, numeric and perfusion image data. This data set, collected over a period of five years [7], [6], was obtained from the Emory University Hospital and we call it the "Heart" data set. There were medical measurements such as weight, heart rate, blood pressure and pre-existence of related diseases. Finally, the data set contains the degree of artery narrowing (stenosis) for the four heart arteries. We converted attributes to binary dimensions using medically-recommended thresholds [15]. Artery narrowing measurements were treated as measurements for OLAP processing. There were $d = 12$ binary dimensions (e.g. gender, hypertension Y/N) and $e = 4$ measures (artery disease measurement) Perfusion image attributes were not used. The Heart data set is summarized in Table I. To provide a more complete and reproducible experimental evaluation we also analyze a publically available data set. This second data set was obtained from the UCI Machine Learning repository [2] and we call it "Thyroid" and is summarized in Table II. The Thyroid data set contained the profiles of $n = 9,172$ patients. We transformed this data set for OLAP processing to have $d = 14$ binary dimensions (common risk factors) and $e = 4$ measures (medical measurements of chemicals on Thyroid gland); there were no image attributes. We discarded two measures (TBG and T3), that had a significant amount of missing values (more than 80%). Both data sets were treated as table F, defined in Section II.

B. Default Settings

We analyzed all d dimensions, exploring the entire lattice. The default settings for parameters were: $p = 0.01$, $\delta = 1$ and $\psi = 30$. In words, we wanted to find significant measure differences, with 99% confidence, on all group pairs differing in one dimension with groups having at least 30 patients. A group pair in the cube can have up to d dimensions, out of which one will be different. It is medically important that a group pair has significant differences in two or more measures, as revealed by the statistical test. The lower settings for ψ trigger the t-student pdf and the higher settings for the p-value provide additional, but less significant, group pairs.

C. Significant Medical Findings

Table III shows significant findings on the Heart data set, based on discussions with a medical doctor. Dimensions (risk factors) are on the left part and measures (degree of disease) are on the right part. Each row represents the comparison between two patient groups, differing in one dimension indicated by "0/1", meaning the discriminating risk factor is absent (equal to 0) in the first group and it is present (equal to 1) on the second one. For a given group pair, each matching dimension will be 0, indicating absence of that risk factor in both groups, or 1 indicating presence of a risk factor in both groups. The * symbol indicates there was a significant difference detected in the corresponding measure (artery disease). Finally, a blank space indicates such risk factor was not considered in the corresponding aggregation (i.e. including"All" records). When $p < 0.01$, the test indicates two similar groups of patients, differing in one risk factor have a high likelihood of having a different degree of disease.

The following discussion focuses on the Heart data set, whose results were discussed with a medical doctor. Table III shows some of the most important results at the significance level $p = 0.01$. To make the table more succinct we do not show N_1, N_2, μ_1, μ_2. The program produced a total of 67 group pairs, involving at most five dimensions; all of them represent valuable medical knowledge. There were additional, less significant pairs, whose p-value was in [0.01-0.05], which are excluded from these results. Cube dimensions, on the left part, are well-known risk factors for heart disease including family history of heart disease, diabetes, gender, high cholesterol and high blood pressure (hypertension). Cube measures representing artery stenosis (narrowing) are on the right. Based on the domain expert opinion we selected some of the most significant pairs, based on these criteria: (1) Identifying key risk factors which can discriminate between healthy and sick patients. (2) Isolating risk factor combinations linked to disease in two or more arteries. (3) Identifying pairs involving a large fraction of the patients and a few (or possibly one) risk factors. (4) Finding pairs that confirm risk factor combinations for heart disease, which are well known in medicine (DIAB and HTA). (5) Finding group pairs involving several risk factors, which uncover specific risk factor combinations that trigger disease in some artery. All pairs shown in Table III fall into one of these categories.

Table IV provides a summary of medically important results for the Heart data set. According to our method only CLAUDI, DIAB, HTA, HYPLPD, OLDYN, SEX, SMOKE can discriminate patients based on the degree of disease. From these risk factors the three most important, found in most pairs, were OLDYN, SEX and SMOKE. On the other hand, it was interesting FHCAD, HIGHCHOL, PANGIO,

TABLE III
HEART DISEASE DATA SET: GROUP PAIRS WITH SIGNIFICANT MEASURE DIFFERENCES AT p=0.01.

D_1 CLAUDI	D_2 DIAB	D_3 FHCAD	D_4 HIGHCHOL	D_5 HTA	D_6 OLDYN	D_7 PCARSUR	D_8 SEX	D_9 SMOKES	A_1 LAD	A_2 RCA	A_3 LCX	A_4 LM
	0	0			0/1			1	*	*		*
					0/1			0	*	*	*	
				0	0/1	0			*	*		*
			0				0/1		*	*	*	
		0			0		0/1			*	*	*
					0/1				*	*		
	0							0/1		*	*	
	0/1				0				*		*	
0		0		0/1	1			0	*			
	0			0	1			0/1	*			

TABLE IV
HEART DISEASE DATA SET: SUMMARY OF MEDICAL RESULTS AT $p = 0.01$.

Discriminating risk factor	Combined risk factors	Arteries
CLAUDI		RCA
DIAB	OLDYN PANGIO	LCX LAD
HTA	CLAUDI DIAB HYPLPD SMOKE	LAD LCX
HYPLPD	DIAB FHCAD HTA OLDYN	LAD RCA
OLDYN	DIAB FHCAD HTA PCARSUR SEX SMOKE	LAD RCA LCX LM
SEX	CLAUDI DIAB FHCAD HIGHCHOL OLDYN PANGIO SMOKE	LAD RCA LCX LM
SMOKE	CLAUDI DIAB HTA HYPLPD OLDYN	LAD RCA LCX

TABLE V
THYROID DATA SET: PAIRS WITH SIGNIFICANT MEASURE DIFFERENCES.

D_1 GENDER	D_2 OLD	D_3 TUMOR	D_4 SURG.	A_1 FTI	A_2 TSH
0/1				$p < 0.01$	$p < 0.01$
	0/1	0			$p < 0.01$
		0/1		$p < 0.01$	$p < 0.01$
		0	0/1	$p < 0.01$	$p < 0.05$

PCARSUR and PSTROKE could not discriminate healthy from sick patients. The medical doctor noted, however, that three of these variables refer to medical history of the patient which was commonly short because the patient was being treated for the first time. The arteries that were diseased most frequently are LAD and RCA; LCX can also be diseased in combination with those two. But LM never appeared alone in a pair, meaning it always appeared in pairs having some other diseased artery. Such finding made medical sense because the other three arteries branch from LM: high stenosis in LM would interrupt blood flow to the other three arteries. The medical doctor stated our tool can be useful to explore a data set to find plausible hypotheses, to identify discriminating risk factors for specific diseased arteries, to isolate patients groups for further analysis and to identify which specific arteries are more likely to be diseased, given a patient profile.

Table V has some important pairs obtained with the Thyroid data set, focusing on two key thyroid gland abnormality measures, at several p-values. These results were not interpreted by any domain expert, but are included because Thyroid is a publically available data set and therefore it can be used as an example to reproduce our experimental results. First, gender alone is an important risk factor for disease given the fact that two measurements are significantly different on two large groups of patients, as evidenced by the first pair. The absence of a tumor combined with increased age or having a surgery lead to significant change in the amount of the Thyroid stimulating hormone (TSH), as shown by the second and third pairs. The fourth pair shows a link between a gland surgery and disease, despite not having a tumor detected. Since this data set is larger than Heart and we used a comparatively lower ψ there are many more pairs found with more dimensions.

D. Impact of Optimizations

Parameters vary as follows: $\psi \in \{0, 10, 30\}$ and $p \in \{0.01, 0.05, 0.10\}$. We first analyze the importance of ψ to reduce the number of generated groups and the number of pairs. Table VI shows the impact of the ψ threshold on pruning the search space. We later analyze the impact of the p-value. In this case we are considering all groups created during lattice exploration and all pairs created by the algorithm before passing/failing the parametric test. This table measures the decrease in the analyzed number of pairs and the amount of work saved. The number of groups and pairs significantly decreases for the Heart data set, where the number of groups and pairs is about two orders of magnitude smaller. On the other hand, the decrease in number of groups is more modest for the Thyroid data set, but still important; the number of pairs created goes down to one third. This is explained by the fact that Thyroid is a much larger data set and then the likelihood that there are groups above ψ is much higher. The last column shows the percentage of pairs that are above the ψ threshold. By Property 1, discussed in Section III-B, and uses only Equation 1 assuming normal distributions. By Property 2, $\psi = 30$ clearly produces significant pruning, compared to

TABLE VI
IMPACT OF ψ ON PRUNING.

Data set	ψ	Groups	%	All pairs	%
Heart	0	604584	100%	205065	100%
	10	113803	19%	21068	10%
	30	42003	7%	5078	2%
Thyroid	0	3319794	100%	1152000	100%
	10	1738546	52%	434048	38%
	30	1197138	36%	241792	21%

TABLE VII
IMPACT OF p-VALUE ON PRUNING (NUMBER OF PAIRS).

Data set	ψ	$p = 0.10$	$p = 0.05$	$p = 0.01$	% All
Heart	0	153265	138561	113159	55%
	10	11518	8383	4111	2%
	30	3551	2842	1605	1%
Thyroid	0	948153	850192	687087	60%
	10	362970	324828	256171	22%
	30	227416	214413	179203	16%

$\psi = 0$.

We now analyze the impact of the p-value, combined with ψ. Table VII shows the number of pairs passing the test at different significance levels. By Property 3, introduced in Section III-B, the number of pairs differing in one dimension is much smaller than the total number of potential pairs; such larger numbers are not shown since those pairs are not generated. The two rightmost columns show the number and percentage of the most significant pairs at $p = 0.01$. Reading the table horizontally from left to right, we can gauge the amount of pruning by the p-value. Reading the table from top to bottom at each p-value we can assess the pruning impact of ψ. From these results we conclude ψ provides a higher impact on pruning than the p-value. Finally, we measure the combined impact of ψ and the p-value on pruning with the quotient (not shown in table) between the number of pairs at $\psi = 30$ and $p = 0.01$ and the total number of pairs from Table VI without any pruning: for the Heart data set the number of pairs goes down to 0.5% and for the Thyroid data set to 17%. In summary, pruning is essential to obtain a manageable number of significant pairs, especially for the Heart data set.

Table VIII gives a breakdown of the total execution time to explore the entire dimension lattice, generate pairs and apply the parametric test. These times include the JDBC overhead to submit queries and retrieve results. As we can see most computations take significant time, despite F being a small data set. Traversing the dimension lattice is the slowest operation for Heart; this step requires creating an output group for each dimension combination and it is I/O bound since it requires visiting every F row. This step exploits Property 4, introduced in Section III-B, to traverse the lattice more efficiently. Filtering and saving pairs is the slowest operation for Thyroid: time to append records depends more on the table size than the number of pairs being appended.

Table IX shows the impact of optimizations. To compare time to traverse the dimension lattice we use $\psi = 0$ and $\psi = 30$. We use $d = 12$ on both data sets. Precomputing the d cube and computing N, L, Q is compared with directly computing N, L, Q from F. That is, we want to understand if it is worth it to compress the (small) data set by precomputing the cube at the finest granularity level. As we can see there is an important performance improvement for the Thyroid data set. An index on all dimensions is compared with a simple primary key for each group (i.e. a group id), to understand the reduction on time to search each group efficiently. The index on dimensions has indeed a significant impact on performance for Thyroid. Finally, $\psi = 30$ is an essential optimization since it reduces time to 10% for Heart and to less than one third for Thyroid.

TABLE VIII
PROFILE OF CUBE EXPLORATION; $\psi = 30$, $d = 12$ (TIMES IN SECS).

Data set	Step	Time	%
Heart	Compute N, L, Q on lattice	151	31%
	Compute μ, σ from N, L, Q	70	14%
	Create group pairs using δ	97	20%
	Compute test statistic each pair	91	18%
	Filter & save pairs based on p-value	87	18%
Thyroid	Compute N, L, Q on lattice	6162	32%
	Compute μ, σ from N, L, Q	684	4%
	Create group pairs using δ	747	4%
	Compute test statistic each pair	651	3%
	Filter & save pairs based on p-value	10923	57%

TABLE IX
OPTIMIZATIONS: COMPUTING NLQ (TIME IN SECS).

Step	Data set	N	Y
Precompute cube & NLQ lattice	Heart	670	650
Index on dimensions	Heart	690	650
Prune at $\psi = 30$	Heart	6128	650
Precompute cube & NLQ lattice	Thyroid	2459	1917
Index on dimensions	Thyroid	2598	1917
Prune at $\psi = 30$	Thyroid	6145	1917

E. Scalability

Table X shows scalability of our method with large data sets. We replicated each medical data set 100, 1000, and 10000 times, storing records in a random order. The algorithm was run with all dimensions from each data set, $p = 0.1$ and $\psi = 0.05$. The significant pairs obtained from the large data sets were exactly the same as those from the small data sets. In order to understand the impact of n on time we include two time measurements. The first column (lattice) shows the time needed to produce the groups and their statistics NLQ in order to apply the statistical test; such time measurement considers computing the cube at the finest aggregation level from F and

TABLE X
SCALABILITY VARYING n (TIMES IN SECS).

Dataset	n	Lattice	Total
Heart	65500	143	490
	655000	147	498
	6550000	168	539
Thyroid	91720	1093	3310
	917200	1104	3335
	9172000	1142	3433

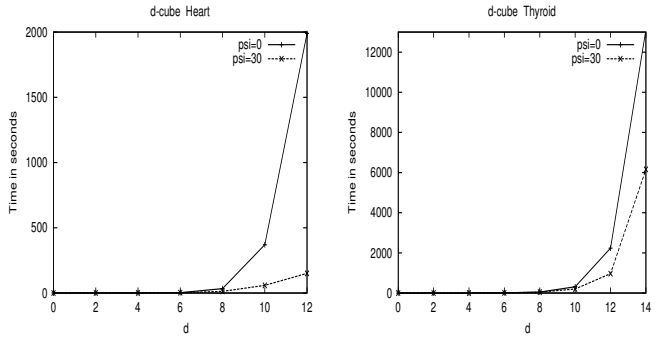

Fig. 3. Time to compute cube varying d.

TABLE XI
ENTIRE PROCESS (TIMES IN SECS).

	Heart		Thyroid	
d	$\psi = 0$	$\psi = 30$	$\psi = 0$	$\psi = 30$
2	5	1	2	1
4	6	2	7	4
6	27	9	33	17
8	139	43	156	75
10	913	200	838	388
12	6128	670	6145	2459
14	-	-	74995	28215

computing NLQ by traversing the lattice. The second column (total) shows time for the entire process. We can see that n has a modest impact on the time to traverse the lattice on both data sets. The overall impact on total time is even smaller: for the Heart data set increasing size 100-fold increases time only by 5% and for the Thyroid data set a 100-fold growth produces merely a 4% growth. These results prove our method can analyze large data sets.

Figure 3 illustrates scalability as we vary the number of dimensions d to traverse the lattice and create groups. The first graph graph corresponds to the Heart data set and the second one refers to the Thyroid data set. The impact from ψ is important for the Heart data set, highlighting the reduction in the number of pairs and time, especially at the highest d. The performance gap is smaller for Thyroid, but still above 50%. This is an indication ψ needs to be increased to get larger groups and fewer pairs. These experiments show the challenge for our problem is d and not n.

Table XI illustrates time scalability for the entire process. These times include the JDBC overhead to submit queries and retrieve results. Time complexity grows exponentially as d increases, highlighting the expensive search process for significant pairs. Our pruning strategy using ψ significantly accelerates processing for the Heart data set reducing time to almost 10% and it produces over 50% improvement for the Thyroid data set. These trends indicate pruning with ψ is essential to make the problem tractable.

V. RELATED WORK

To the best of our knowledge, we are the first to propose the integration of statistical tests and OLAP cubes. Statistical tests have been extensively applied in the medical literature [20], but they have not been combined with exploratory database techniques, as proposed in this work. On the other hand, OLAP cubes have been studied in the database literature for efficient ad-hoc analysis [5], [10], [9], [19], but have never been extended with statistical tests. Our approach goes a step beyond by comparing pairs of groups with a statistical test obtained with automated cube exploration.

This work is a continuation of previous studies using data mining, statistical and machine learning techniques on medical data sets to improve heart disease prediction [6], [7], [17], [15]. In [3], neural networks were used to predict heart response based on exercise stress and heart muscle thickening images. A basic set of search constraints for association rules was also introduced in [17], and experimental results stress their importance. Reference [15] introduces techniques to learn predictive association rules for disease prediction by testing and filtering them on independent subsets of a data set. Our work exhibits similarities with [17], [15] in the sense that we explore a large search space to find interesting results. However, there are important differences: one association rule refers to one group of patients, whereas a pair relates two; association rules require binned (discretized) numeric attributes; the p-value is different from confidence. OLAP techniques allow exploration of binary dimensions, but the the parametric test allows comparing numeric attributes (degree of disease in this case). We stress parametric tests are not a substitute for a predictive model, but a complementary technique to analyze medical data sets. Also, statistical tests have wider applicability given their more general assumptions.

Related work on improving disease diagnosis with data mining techniques includes the following. Reference [4] looked at two different datasets to predict radiation pneumonitis based on superior-to-inferior tumor position, maximum dose, and D35; while a predictive statistical model based on one of the datasets did not perform well, when both datasets were used, the result was greatly improved. In a different study, multivariate analysis was used in [14] to determine that in addition to only using PSA (prostate specific antigen) and DRE (direct rectal examination) to screen patients for PC (prostate cancer), one should also look into other factors, such as age and family history. In these two studies, the first looked at a set number of factors while the second analyzed different sets of factors. In terms of analysis results, the last reference most closely matches our proposal. However, our method takes subset analysis even further: our method can find all significant subsets from a given set of factors without user participation.

We close this section discussing related work on OLAP from the database field. OLAP and a taxonomy of aggregations originates in the seminal paper [9]. In database research most approaches avoid exploring the entire cube but since medical data sets are small, the cube dimensionality is lower and several pruning strategies are used, it becomes feasible to search for significant pairs on the entire cube. Reference [10] summarizes several techniques to accelerate OLAP cube computations. Key differences are: (1) our summary tables contain sufficient statistics for the parametric test; (2) we focus on multiple pairs of groups instead of analyzing single groups; (3) mean comparisons are made with the statistical test and not

comparing averages directly. Similar to our proposal, OLAP operations are used as a mechanism to find interesting patterns in [12]. However, instead of mining collections of similar rules referring to single groups in order to gain knowledge that cannot be obtained from individual rules, we analyze pairs of groups within subcubes to detect significant differences. In [19] the authors explore techniques to guide the user to interesting cube regions in order to highlight anomalous behavior, by identifying exceptions. That is, they identify values in cells of a data cube that are significantly different from some expected value, based on a regression model. In contrast, we propose to use statistical tests to do pairwise comparison of neighboring cells in subcubes to discover significant metric differences between highly similar groups. Our optimizations generalize downward closure properties of association rules [10] to multidimensional group pairs and avoid generating many unnecessary patterns.

VI. Conclusions

In this article, we proposed integrating OLAP cube exploration and statistical tests to analyze medical data sets. Parametric statistical tests are applied on pairs of similar groups in order to find significant measure differences caused by some discriminating dimension. Cube dimensions are typically represented by risk factors and cube measures represent degree of disease or health. Our approach can produce statistically reliable results with both large and small subsets. We introduce several optimizations. We use a frequency threshold used to eliminate small groups, reduce the number of pairs and accelerate dimension lattice exploration. Sufficient statistics are used to directly derive the mean and standard deviation from the cube in one pass. Pair generation is optimized to avoid generating unnecessary pairs. The cube incorporates an index on all dimensions for efficient pair generation. Automatic cube exploration, the parametric statistical test and optimizations are assembled into one efficient and comprehensive algorithm. We discussed interesting experimental results on two medical data sets, for heart and thyroid disease respectively, illustrating the applicability of our approach to contrast similar groups of patients having significant differences in degree of disease. In the case of heart disease we identified combinations of risk factors (absent and present) which can lead to disease (stenosis) in the four heart arteries. Group pairs with several dimensions provide specific profiles of patients who have developed disease. Significant pairs can help identifying important risk factors that trigger health issues in any of the four arteries, isolating patient groups for further analysis and improving disease diagnosis with new patients. From a scalability perspective, experiments show it is feasible to search an entire cube with low dimensionality for significant group pairs by applying our pruning and filtering techniques. Experiments also show there is a significant reduction in the number of groups and pairs when using the frequency threshold and a strong significance p-value, thereby producing only medically relevant results and improving their interpretation.

There are many research issues for future work. We want to understand the interrelationship of degree of disease for groups having two or more significant measure differences. The findings of the parametric test can also be used as a preprocessing step for building predictive models. Our method will be part of a data mining tool incorporating predictive models, data mining and data transformation to be used in medical research. We plan to apply our tool on other medical data sets for heart disease and cancer, to be obtained from other medical institutions. We need to use imputation techniques to improve data quality of attributes with a significant fraction of missing information. Medically significant pairs can be validated with alternative statistical techniques. We want to introduce further optimizations to prune insignificant or redundant patient group pairs. There exists a big family of alternative statistical tests that can also be applied in OLAP cubes.

Acknowledgment

We would like to thank the Emory University Hospital for providing the Heart data set. We also thank Dr. Cesar A. Santana from the Emory University Hospital for his valuable comments to interpret and validate medical findings. The authors thank the anonymous reviewers for their suggestions.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Conference*, pages 207–216, 1993.

[2] A. Asuncion and D.J. Newman. *UCI Machine Learning Repository*. University of California, Irvine. School of Inf. and Comp. Sci., 2007.

[3] L. Braal, N. Ezquerra, E. Schwartz, and Ernest V. Garcia. Analyzing and predicting images through a neural network approach. In *Proc. of Visualization in Biomedical Computing*, pages 253–258, 1996.

[4] J.D. Bradley, A. Hope, I. El Naqa, A. Apte, P.E. Lindsay, W. Bosch, J. Matthews, W. Sause, M.V. Graham, and J.O. Deasy. A nomogram to predict radiation pneumonitis, derived from a combined analysis of rtog 9311 and institutional data. *Int J Radiat Oncol Biol Phys*, 2007.

[5] S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26(1):65–74, 1997.

[6] D. Cooke, C. Ordonez, E.V. Garcia, E. Omiecinski, E. Krawczynska, R. Folks, C. Santana, L. de Braal, and N. Ezquerra. Data mining of large myocardial perfusion SPECT (MPS) databases to improve diagnostic decision making. *Journal of Nuclear Medicine*, 40(5), 1999.

[7] D. Cooke, Cesar Santana, Tahia Morris, Levien de Braal, C. Ordonez, E. Omiecinski, N. Ezquerra, and Ernest V. Garcia. Validating expert system rule confidences using data mining of myocardial perfusion SPECT databases. In *Computers in Cardiology Conference*, pages 116–119, 2000.

[8] R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems*. Addison/Wesley, Redwood City, California, 3rd edition, 2000.

[9] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-total. In *ICDE Conference*, pages 152–159, 1996.

[10] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 1st edition, 2001.

[11] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer, New York, 1st edition, 2001.

[12] B. Liu, K. Zhao, J. Benkler, and W. Xiao. Rule interestingness analysis using olap operations. In *ACM KDD*, pages 297–306, 2006.

[13] T.M. Mitchell. *Machine Learning*. Mac-Graw Hill, New York, 1997.

[14] R.K. Nam, A. Toi, L.H. Klotz, J. Trachtenberg, M.A. Jewett, S. Appu, D.A. Loblaw, L. Sugar, S.A. Narod, and M.W. Kattan. Assessing individual risk for prostate cancer. *J Clin OnCol.*, 25(24):3582–8, 2007.

[15] C. Ordonez. Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine (TITB)*, 10(2):334–343, 2006.

[16] C. Ordonez, N. Ezquerra, and C.A. Santana. Constraining and summarizing association rules in medical data. *Knowledge and Information Systems (KAIS)*, 9(3):259–283, 2006.

[17] C. Ordonez, E. Omiecinski, Levien de Braal, Cesar Santana, and N. Ezquerra. Mining constrained association rules to predict heart disease. In *IEEE ICDM Conference*, pages 433–440, 2001.
[18] J.F. Roddick, P. Fule, and W.J. Graco. Exploratory medical knowledge discovery: Experiences and issues. *SIGKDD Explorations*, 5(1):94–99, 2003.
[19] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of OLAP data cubes. In *EDBT*, pages 168–182. Springer-Verlag, 1998.
[20] N.A. Weiss. *Elementary Statistics*. Addison-Wesley, 2005.

Carlos Ordonez received a degree in applied mathematics and an M.S. degree in computer science, from UNAM University, Mexico, in 1992 and 1996, respectively. He got a Ph.D. degree in Computer Science from the Georgia Institute of Technology, in 2000. Dr Ordonez worked six years extending the Teradata DBMS with data mining algorithms. He is currently is an Assistant Professor at the University of Houston. His research is centered on the integration of statistical and data mining techniques into database systems and their application to scientific problems.

Zhibo Chen received the B.S. degree in electrical engineering and computer science in 2005 from the University of California, Berkeley, and the M.S. degree in computer science in 2008 from the University of Houston, where he is currently working toward a Ph.D. degree in computer science. His research focuses on optimizing OLAP processing in a database system.