A computational algorithm for the risk assessment of developing acute coronary syndromes, using online analytical process methodology

Hara Kostakis*
TEI Piraeus Research Centre
Methonis 8-10
106 80, Athens, Greece
Fax: +30-210-3845347 E-mail: kostaki@upatras.gr
*Corresponding author

Basilis Boutsinas
University of Patras Artificial Intelligence Research Center,
University of Patras,
26500 Patras, Greece
E-mail: vutsinas@upatras.gr

Demosthenes B. Panagiotakos
Harokopio University,
17671 Athens, Greece
E-mail: d.b. panagiotakos@usa.net

Leo D. Kounis
PartsInterlink, Spare Part Provider for American Cars,
97 Euterpe Street, Rhodopolis,
Attica 14574, Greece
Fax: +30-210-62-10-707 E-mail: leo_kounis@hotmail.com

Abstract: This paper investigates patterns in cardiovascular risk factors from a large population sample of cardiac patients and their matched controls. Various factors were taken into consideration and were used as inputs to effectively demonstrate online analytical process, OLAP methodology.

OLAP is a new method that is used to explore the role of several risk factors in cardiovascular disease risk assessment. It equally serves as a means to extract knowledge from the investigated factors’ levels.

This paper discusses the application of OLAP-specific procedures in order to explore hidden pathways associated with risk factors among patients and controls. It does so, as the latter proves to be time consuming when classical statistical methods, in particular logistic regression are applied.

Finally, this work builds on earlier findings, with odds ratios converging among the studies. The outcome of this work results in a more accurate risk assessment, as it takes into account variable-interaction.

Copyright © 200x Inderscience Enterprises Ltd.
Keywords: OLAP methodology; online analytical process; multidimensional analysis; acute coronary syndromes.


Biographical notes: Hara Kostakis is based at the TEI Piraeus Research Center, Athens, Greece. She received her BSc in Business Administration from the University of Piraeus, Greece, her MSc in Decision Sciences from the London School of Economics and Political Sciences, UK and her PhD in Healthcare Management from the University of Patras, Greece. Her research interests include the use of data mining and AI techniques in healthcare management.

Basilis Boutsinas received his diploma in Computer Engineering and Informatics in 1991 at the University of Patras, Greece. He has also conducted studies in Electronics Engineering at the Technical Education Institute of Piraeus, Greece and Pedagogics at the Pedagogical Academy of Lamia, Greece. He received his PhD in Knowledge Representation from the University of Patras in 1997. He is an Associate Professor in the Department of Business Administration, University of Patras. His primary research interest includes data mining, business intelligence, knowledge representation techniques, non-monotonic reasoning and parallel AI.

Demosthenes Panagiotakos is a Lecturer in Biostatistics and Epidemiology of Human Nutrition at Harokopio University, Department of Nutrition Science-Dietetics. He received his BSc in Mathematics from the University of Patras, Greece in 1989. He has a PhD in Mathematical Statistics and a Doctorate in Medicine (Epidemiology) from the University of Athens. He has conducted Postgraduate studies in Biostatistics and in Climate Change and Human Health, as well as Post-doctorate studies in Cardiovascular Epidemiology at the Veterans Affair Medical Center, Georgetown University Medical Center, Washington DC. His research interests focus on nutritional epidemiology, risk stratification of cardiovascular disease and multivariate risk analysis.

Leo D. Kounis is a Lecturer in Physics at the American University of Athens and Visiting Professor in Operations and Financial Management and Business Ethics at the Higher Technological Education Institute, ATEI of Chalkis, in Evia, Greece. He acts as Research Assistant on a number of projects of the University of Sheffield in Athens, Greece. He received his BEng (Hons) in Manufacturing Systems Engineering, an MSc in Quality Engineering and a PhD in Systems Reliability from the University of Hertfordshire. His research interest focuses in the area of quality, reliability, sustainable development and renewable energy sources.
1 Introduction

In recent years, the interest in a global risk assessment is becoming more common. New questions concerning the role of several lifestyle risk factors are being raised and innovative approaches to analysis of the data, such as data mining techniques are increasing in importance. These lifestyle risk factors include but are not limited to depression, education and smoking and nutrition habits (Panagiotakos et al., 2002).

Since the mid ‘70s, several epidemiological studies have attempted to describe the potential candidate for cardiovascular disease (Kannel et al., 1976). These studies evaluate emerging risk factors (Wood, 2001; Hallman et al., 2001) such as obesity, lack of exercise and depression aiming at assessing the risk of developing heart diseases. Current methodologies include multivariate analysis of variance and logistic regression to name but a few. These methods are implemented for evaluating differences in continuous measurements between study groups and for estimating the relative risks of developing acute coronary syndromes by calculation of the odds ratio (OR) respectively. The corresponding confidence intervals are calculated through multiple conditional logistic regression analysis (Hung et al., 2008). For optimising computational time various methods such as logistic regression analysis or log-linear models combined with correspondence analysis were implemented (Panagiotakos and Pitsavos, 2004). These methods were aimed at reducing the number of tested interaction terms within the final log-linear model.

In order to overcome observed bottlenecks in the health sector this paper places an emphasis in introducing a flexible and effective algorithm to optimise computational time and the risk assessment of developing acute coronary syndromes. In particular, this research work provides the user with the potential of assessing global risk to efficiently address the time constraint observed in classical statistical methods. In view of the above, the aims and objectives of this research work are:

Aim To introduce the implementation of a computational algorithm for the assessment of the risk of developing acute coronary syndromes. The objectives of this paper are as such:

• to review existing literature on the subject of heart disease risk assessment
• to highlight the drawbacks of existing methodologies in the area of heart disease risk assessment
• to apply a computational algorithm for effectively addressing and optimising the computational time and risk assessment.

2 Literature review

Performed research revealed that classical statistical methods have been commonly used as techniques to improve the understanding and prevention of coronary heart disease in various populations (e.g., Vartiainen et al., 2000) and various patient groups (e.g., Bauman and Spungen, 2008). These techniques, which include multivariate analysis of variance and multiple regression models among others, aim at evaluating differences in
measurements in risk factors in study groups, as well as classify high risk patients respectively.

In addition, data mining techniques for analysing heart disease databases have been applied amongst others in specialised medical institutions (Gamberger et al., 2003). Studies in data mining include neural networks supervised classification, data visualisation and discriminant analysis applied in heart disease data, aiming at identifying high risk patients and classifying them accordingly, defining the most important factors in heart disease and establishing a relationship between any two variables (Lee et al., 2000). In 2003, the European Society of Cardiology (ESC) attempted to evaluate global risk of ACS using an additive model, known as the SCORE project (Conroy et al., 2003).

In particular, the SCORE project discusses the use of an additive system developed for use in clinical management across Europe, as it pools data from 12 European cohort studies. This system is based on a risk scoring system and its aim is the direct estimation of fatal cardiovascular risk within the constraints of clinical practice.

However, aforementioned studies are deemed by the authors to be ineffective in addressing the global risk assessment. In an effort to achieve this goal OLAP methodology has been applied, which is based on an additive model, in an acute coronary syndromes database, the CARDIO2000 (Panagiotakos et al., 2002).

Online analytical processing (OLAP) has been initially introduced and implemented by Codd et al. (1993), Pense and Creeth (1995) and Lehner (1998) amongst others. The outcome of aforementioned studies led to industrial and commercial applications, mainly in the financial and marketing area. Indeed, OLAP tools provide a multidimensional conceptual view of the data. Codd et al. (1993) discusses the use of OLAP in gaining direct and useful information from large amounts of data. This process is accomplished by means of pre-aggregation in order to improve the response time to queries. In particular, OLAP supports data analysis through a multi-dimensional data model. This model provides the user with data facts that are presented as points in a multi-dimensional application-related space (data cube). Dimensions are organised into levels that are structured by means of a hierarchy.

Vassiliadis and Sellis (1999) discuss other OLAP models. However, current research work introduces and discusses new and improved problem-solving methods (Boutsinas, 2005). These methods imply the OLAP model in a practical case study. This case study highlighted and emphasised the advantages of OLAP techniques in global risk assessment. In particular, this research work provides the user with the capability of assessing global risk. The latter however, is time consuming for it to be addressed by classical statistical methods. These in turn, cannot incorporate interactions between risk factors (the joint probabilities of variables occurring together) resulting in an underestimation of an event (in this case acute coronary syndrome-ACS).

Finally, the proposed OLAP model is flexible enough to be expanded by using the time dimension. The latter occurs in order to obtain comparative results for different studies concerning the same investigation.

3 Materials and methods

In order to optimise and effectively address the constraints mentioned therein, this research work will introduce the advantages of using the proposed methodology in accurately assessing global risk, as well as efficiently resolving time-related constraints.
In particular, the proposed OLAP model successfully handles the dependencies between time and the other dimensions in medical research, such as hypercholesterolemia, hypertension and body mass index (BMI). Its building block is the data hypercube. A data hypercube (DH), is an n-dimensional array: DH[d1, ..., dn] of lists l = (e1, ..., em). Every element of a list is an ordered pair ei = {labeli, valuei} denoting that each element is defined by two attributes: its label that is distinct and its value. All the lists of a data hypercube have the same number of elements, each with the same label in all the lists. To every dimension di there is also assigned a label labeli, denoting that each dimension has a distinct label. A dimension consists of dimension values, which belong to a certain domain. To every dimension value a label is equally assigned. The function label(i) that returns the label of an element or dimension value is also defined. The type of the label of an element or dimension value is always alphanumeric. The type of the value of an element could be either numeric or alphanumeric. The function dom(di) that returns the domain of dimension values of a dimension di is equally defined.

Intuitively, elements (correspond to measures) represent the actual data that are organised around concepts of the real world that need to be analysed. Concepts are represented by dimensions in the proposed OLAP-model. Examples of such concepts are sales, purchases, pricing, customer base, etc. The label of an element denotes the concept that this element represents. The value of an element is a measurement of such a concept. The identification of a certain element depends on the identification of the list it belongs to. In order to identify a list, one must first define the dimensions of the n-dimensional array that represents the hypercube. These dimensions represent the different views through which the data can be accessed. The proposed OLAP-model is defined using the notion of hierarchical data hypercube.

A hierarchical data hypercube is based on a data hypercube, the specialised data hypercube DHs[d1, ..., dn], where ∀ vsi ∈ dom(di) ∃ eij ∈ lij and vsi = value(eij) and lij is a list that belongs to another data hypercube, the generalised data hypercube: DHg[d1, ..., dm].

Intuitively, a hierarchical data hypercube is a specialised form of a data hypercube that corresponds to a fact table. The latter has at least one dimension, which is identified through the elements of another generalised form of a data hypercube, corresponding to a dimension. Thus, the values of at least one dimension of the specialised data hypercube and hence the identification of its lists and elements, depend on the values of all dimensions of the generalised data hypercube through the identification of its lists and elements.

Elements of the lists of generalised data hypercubes are values of dimensions of their specialised data hypercube. There are three kinds of relations between dimension values of generalised and specialised data hypercubes.

The first is a one-to-one relation, where every list of a generalised data hypercube is related to only one dimension value of its specialised data hypercube. The reverse is also true.

The second is a one-to-many relation, where every list of a generalised data hypercube is related to many dimension values of its specialised data hypercube, while every dimension value of the latter is related to only one list of the corresponding generalised data hypercube.

The third is a many-to-many relation, where every list of a generalised data hypercube can be related to many (possibly one) dimension values of its specialised data hypercube.
hypercube and every dimension value of the latter is related to many lists of the corresponding generalised data hypercube.

Therefore, the dependencies between the different data hypercubes form a tree rooted at the base data hypercube. Thus, the dependency tree of the OLAP-model is defined. The structure of a dependency tree introduces levels identifying where the different data hypercubes belong. A hierarchy is defined as a path of the dependency tree starting from the base data hypercube and ending to a leaf. Since a hierarchy starts from the base data hypercube, it is related with one of its dimensions. Actually, every generalised data hypercube that belongs to a hierarchy can be used in the identification of lists of the base data hypercube, through defining the values of this related dimension. In general, there might be more than one hierarchies related to the same dimension of the base data hypercube.

According to the above, one could access more than one list of elements in the base data hypercube. This situation corresponds to the identification of more than one value for some dimension(s). It is, actually, a rule in the case where the accessed elements are identified indirectly through the definition of elements of generalised data hypercube(s). This is because a list of a generalised data hypercube usually consists of many elements that are dimension values of some dimension of its specialised hypercube. Since the different lists of elements are identified indirectly, they must also be viewed indirectly. This is achieved by aggregating the different lists of elements of the base data hypercube. Aggregating is reduced to applying a certain aggregating operator over these lists. Examples of aggregating operators, applying to elements with numerical values are ‘sum’, ‘average’, etc. Examples of aggregating operators applying to elements with alphanumerical values are ‘concatenate’, ‘substring’, etc. Usually, different aggregating operators are applied to different sets of corresponding elements. Moreover, different aggregating operators can be applied to the same set of corresponding elements viewed by different levels of the dependency tree. These aggregating operators are formally defined by attaching to each node of the dependency tree a list of operators \( o = \{op_1, \ldots, op_g\} \), where every aggregating operator \( op_i \) is applied to the \( i \)th element of the different lists of the base data hypercube.

An example of a dependency tree is shown in Figure 1.

To this end, a database, the CARDIO 2000 study was applied to assess global risk of cardiovascular disease. The CARDIO 2000 is a multicenter case-control study (Panagiotakos et al., 2002) that explores the association between several demographic, nutritional, psychological, lifestyle and medical risk factors. Some of these factors standing either alone, or in combination are of predominant importance as they may lead to developing non-fatal acute coronary syndromes.

In order to reduce the unbalanced distribution of several measured or unmeasured confounders, both patients and controls were randomly selected. From January 2000 to August 2001, a sample of 956 individuals was selected. These individuals had just entered targeted hospitals showing first symptoms of ACS (stable angina was excluded from the analysis). Furthermore, out of these 956 individuals, 848 agreed to be enrolled into the study and they were categorised as cases.

The inclusion criteria for cardiac cases included the following conditions:

- diagnosis of first acute myocardial infarction (MI)
- diagnosis of unstable angina.
MI was defined by the following features:

- electrocardiographic changes
- compatible clinical symptoms
- specific diagnostic enzyme elevations.

Figure 1 An example of a dependency tree modelling an ITU showing tree variables (dimensions) and one hierarchy over the time dimension.
Unstable angina was defined as one or more angina episodes at rest within the preceding 48 hours. These episodes corresponded to class III of the Braunwald classification. The latter discusses the conditions of unstable angina, such as the clinical history, the presence or absence of ECG changes and the intensity of anti-ischemic therapy (Hamm and Braunwald, 2000).

Of the 848 patients, 49% of these had myocardial infarction and 51% of the patients had unstable angina.

After the selection of the cardiac patients, cardiovascular disease-free controls were randomly chosen among 1300 subjects. Of these 1078 (83% response rate) agreed to be enrolled into the study. The controls were matched to the patients according to age distribution (±3 years), their sex and the same Greek region. Controls are defined as individuals who visited the outpatient departments of the same hospital and at the same period with the coronary patients, for minor surgical operations, such as orthopaedic, etc. The controls were subjects without any clinical symptoms, signs or suspicion of cardiovascular disease in their medical history.

An example of an additive model by using influencing variables is described. These influencing variables include:

- body mass index, BMI
- family history
- physical activity
- hypertension
- hypercholesterolemia
- diabetes.

Obesity was defined as BMI \( \text{weight/(height)}^2 \) greater than 29.9 kg/m\(^2\). Obese individuals showing BMI greater than 29.9 kg/m\(^2\) were assigned the number ‘1’ to indicate the equivalent status (BMI = 1). Individuals having BMI less than 29.9 kg/m\(^2\) were assigned the number ‘0’ to indicate non-obesity.

Family history of premature ACS was any coronary episode in first-degree relatives (<45 years old in men and <55 years old in women). Presence of family history was indicated by the number ‘1’ (FAM_HIST=1), while absence of family history was identified by the number ‘0’.

Physical activity was defined as any type of non-occupational physical exercise that occurred at least once per week during the past year. The number ‘1’ denoted such individuals, who were defined as physically active. The remainder of the individuals were defined as physically inactive and were assigned the number ‘0’ to indicate such physical inactivity (EXERC = 0). This physical activity was quantified by intensity (Kcal/min) and duration (in minutes); sedentary, < 4 kcal/day, 4–7 kcal/day, ≥ 7 kcal/day.

The previous medical information, as well as the patients’ and controls’ clinical records, enabled the characterisation of the participants as hypertensive, hypercholesterolemic and diabetics, respectively. Thus, in keeping with the long-standing classification criteria used in several population-based studies (Panagiotakos et al., 2002), individuals whose mean blood pressure levels were greater than or equal to 140/90 mmHg or were under antihypertensive medication were classified as hypertensive.
and were assigned the number ‘1’ to indicate accordingly (HTN = 1). The remainder of the individuals were assigned the number ‘0’ to show absence of hypertension.

Fasting total cholesterol levels were measured in serum. Hypercholesterolemia was defined as:

- cholesterol levels greater than 220 mg/dl
- cholesterol levels greater than 200 mg/dl when two other risk factors for ACS exist
- the use of special hypo-lipidemic treatment.

These individuals were assigned the number ‘1’ (HCHOL = 1). The rest of the individuals were assigned the number ‘0’, thus identifying absence of hypercholesterolemia.

Finally, diabetics were those individuals, whose fasting blood glucose was greater than 125 mg/dl or who were under a special diet or treatment. The number ‘1’ was also assigned to such individuals (DM = 1). The remainder were characterised by the number ‘0’.

In the presented study for the assessment of acute coronary syndromes all risk factors are represented as dimensions. Moreover, a dummy measure was used that equals to one in all the records of the base data hypercube (fact table). As an aggregating operator the count of instances was selected. This technique provides the user with the number of individuals online with certain user-defined risk factors.

In view of the above, BMI, FAM_HIST, EXERC, HTN, HCHOL and DM were measured taking into account SEX (male and female) and GROUP (coronary patients and controls) and were used as dimensions. The presented results in the following section reflect only one study, and as such the TIME dimension was inactivated. The data hypercube was implemented using Microsoft Data Analyzer 2002.

### 4 Results

The output produced is different multidimensional views of the data, as the one shown in Figure 2.

Figure 2  The multidimensional view that supports the presented results (step 1)

<table>
<thead>
<tr>
<th>TIME=ALL</th>
<th>BMI=30</th>
<th>FAM_HIST=ALL</th>
<th>EXERC=ALL</th>
<th>HTN=ALL</th>
<th>HCHOL=ALL</th>
<th>DM=ALL</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 0 (controls)</td>
<td>0.198</td>
</tr>
<tr>
<td>Group 1 (patients)</td>
<td>0.210</td>
</tr>
</tbody>
</table>

Initially, the analysis takes account of the first dimension independently. In the first step (shown in Figure 2), the probability of a person having BMI = 1 (corresponding to BMI values greater than 29.9 kg/m²) in the entire database is calculated, both for males and
females in each group (controls and patients), while all the other factors are not taken into account. Thus, for example the odds of a male patient having BMI = 1 is 0.210, while for a male control is 0.198. Consequently the OR of BMI = 1 between patients and controls is 0.210/0.198 = 1.06. In the second step the joint probability of a person having BMI = 1 and a family history of ACS (FAM_HIST) is calculated. The third step calculates the joint probability of a person having BMI = 1, a family history of ACS and physical inactivity. The fourth step includes the case where the joint probability of a person showing BMI = 1, a family history of coronary syndrome, physical inactivity and hypertension is calculated. The fifth step incorporates the case where the joint probability of a person having BMI = 1, a family history of coronary syndrome, physical inactivity, hypertension and hypercholesterolemia is evaluated. The last step includes the case where the joint probability of a person having BMI = 1, a family history of coronary syndrome, physical inactivity, hypertension, hypercholesterolemia and diabetes is calculated. Following, once all the queries have been answered (corresponding to the last step), one might obtain an aggregated table with results for all dimensions, as shown in Table 1.

**Table 1** Results from the additive OLAP model that was developed for the evaluation of the investigated parameters on the risk of having ACS in males and females

<table>
<thead>
<tr>
<th>Factors</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds of having ACS (patient/control)</td>
<td>Odds of having ACS (patients/control)</td>
</tr>
<tr>
<td>Model 1 BMI &gt; 29.9 kg/m²</td>
<td>0.210 / 0.198</td>
<td>0.219/0.200</td>
</tr>
<tr>
<td>Model 2 Model 1 + family history of ACS</td>
<td>0.085 / 0.041</td>
<td>0.096/0.043</td>
</tr>
<tr>
<td>Model 3 Model 2 + physical inactivity</td>
<td>0.063 / 0.018</td>
<td>0.075/0.012</td>
</tr>
<tr>
<td>Model 4 Model 3 + hypertension</td>
<td>0.037 /0.010</td>
<td>0.048/0.009</td>
</tr>
<tr>
<td>Model 5 Model 4 + hypercholesterolemia</td>
<td>0.026/0.006</td>
<td>0.027/0.006</td>
</tr>
<tr>
<td>Model 6 Model 5 + diabetes</td>
<td>0.010/0.004</td>
<td>0.014/0.0001</td>
</tr>
</tbody>
</table>

Each number in the cells in Table 1 represents probabilities. For example, 0.210 is the probability of a male patient to show a value of BMI>29.9 kg/m². On the other hand, the ratios shown in the OR columns represent odds (the risk of an event, in this case developing ACS). Thus, 1.06 corresponds to the OR, i.e., the likelihood of having ACS according to BMI status (> 29.9 kg/m²). The number 2.07 shown in the first OR column corresponds to the OR of a male person having the disease according to BMI and the presence of a family history of coronary syndrome. Likewise, the value of 2.5 indicated in the last line of the first OR column corresponds to the OR of a male person having the disease according to the additive model that includes BMI > 29.9 kg/m², family history of coronary syndrome, physical inactivity, hypertension, hypercholesterolemia and diabetes. In the above model, lifestyle factors have been incorporated as well as prevalence indices. In theory, one can add an infinite number of risk factors and obtain the corresponding odds ratios.
5 Discussions

During the past years the interest in a global risk assessment is increasing as new methods for analysing data are emerging. These methods include, among others, neural networks supervised classification (Azuaje et al., 1997), a methodology very similar to logistic regression. These studies, which use neural networks attempt to train algorithms with risk factors to distinguish coronary heart disease (CHD) patients and thus classify high-risk individuals accordingly (Shen et al., 1994).

Moreover, classification algorithms have been used in risk assessment of CHD. They have also been compared with logistic regression for the task of classifying cardiac patients in a large database. The odds ratios converge among the results, while the classification algorithms have the added advantage of being free of assumptions (Kostakis et al., 2007). However, current bibliography on risk assessment of ACS is mainly based on classical statistical methods, such as multiple logistic models (Thiru et al., 2003), which are in turn based on the assumption of absence of multicolinearity. Nevertheless, it is hard to say whether this assumption holds true in most epidemiological studies, as for example age cannot be fully independent of the presence of various clinical characteristics, such as hypertension and diabetes.

Various other studies have used risk functions in order to associate factors with cardiovascular disease. Many studies have been based on the Framingham prediction scores and functions, which however have proved to overestimate CHD risk in various populations (Hense et al., 2003). In 2003, the European Society of Cardiology attempted to evaluate global risk using an additive model, known as the SCORE project (Conroy et al., 2003). This study aims at producing risk charts derived from risk functions using a Weibull proportional hazards model. The charts produced incorporated 400 combinations of risk factors. However, the limitation of the model is that it is capable of including only five independent variables (factors) as predictors of the dependent variable (CHD).

A model for estimating the global risk of developing ACS by using OLAP methodology is presented in this study. More specifically, an OLAP model that was developed at the University of Patras was used, in order to extract knowledge from the investigated factors’ levels. The applied procedures explored hidden pathways concerning the risk factors among patients and controls that it is extremely time consuming to be addressed by the classical statistical methods. Moreover, the proposed methodology is free of assumptions, while statistical models are usually based on a set of assumptions.

The odds ratios calculated by using the proposed OLAP methodology converge the results published in earlier studies (Panagiotakos et al., 2002), with the advantage of incorporating the interaction between the variables (the joint probabilities of factors occurring together), which the classical statistical models (i.e., the multinomial Logit parametric association model) (Panagiotakos and Pitsavos, 2004) did not effectively address. These models resulted in losing useful information in the data and led to the underestimation of the risk of an event (ACS).

OLAP is defined (Codd et al., 1993) as to enable end users to gain an insight into data through fast, consistent, interactive access to a wide variety of possible views of information. In the presented study for the assessment of ACS, cardiologists and physicians are regarded as the end users of the system. Due to the easiness of using the methodology, a physician has the advantage of easily identifying high-risk patients by...
simply indicating their personal data in the model. Data are defined as risk factors, such as demographic, nutritional, psychological, lifestyle features, as well as biochemical indices. Thus, of great importance is the user-friendly structure of this methodology and hence it can be applied by users who do not possess certain specialised knowledge of the analysis procedure itself. Moreover, this methodology is fast enough in obtaining results. The accuracy of the results, which is translated into the correct probability assessment of a person developing ACS, is based on the availability of large databases. Therefore, one can add a very large number of variables and calculate the corresponding odds ratios.

OLAP methodology is an exploratory technique in real time. It is not used in order to develop models that fit the data or vice versa (Greenacre, 1984). It rather supports a visualisation of the data in order to study their ‘structure’. Since the visualisation is based on pre-aggregations, it can efficiently include all the dimensions, i.e., risk factors of a problem. Thus, it is safer to generalise conclusions about the data at the population level than it is for other methods, which are based on dimension reduction. On the other hand, a user is able to exclude certain dimensions (risk factors) from presentation by summarising them via aggregation to the highest level and thus be able to analyse the data in certain sub-spaces.

The proposed methodology is capable of exploiting hierarchy-specific drill-down and roll-up functions. This offers the user the added advantage to analyse data by using different groupings of the values (levels) of the risk factors (dimensions). Moreover, the proposed OLAP model supports the organisation of these groupings of the dimension values into non-monotonic networks (Boutsinas and Vrahatis, 2001). In view of the above, the user is able to analyse data by using the groupings of the values of the risk factors that represent certain concepts of the real world.

The predominant advantage of implementing this specific OLAP-model is its ability to expand the developed hypercube. The latter offers the added flexibility that it can be expanded to using the time dimension, in order for it to be able to obtain comparative results for different studies concerning the same investigation. In particular, the advantage of the proposed OLAP model is the correct treatment of the dependencies between the time dimension and the other dimensions.

In medical research, dependencies between dimensions (risk factors) are common. For example, in a medical database, statistical analysis could show that the presence of hypertension is dependent on smoking, as well as the presence of hypercholesterolemia is dependent on a person’s diet. The most common of such dependencies concerns time dimension, i.e., the likelihood of developing cardiovascular disease based on the current hypertension status that is dependent on the increasing age of the person. This type of dependency is called ‘versioning the dimension structures’ (Lehner, 1998).

For instance, consider obesity, which is discretised by comparing it to the threshold $(t) 29.9 \text{ kg/m}^2$ of BMI. Then dimension BMI has two dimension values:

1. \( \text{BMI} = 0 \) for input data values less than 29.9 kg/m$^2$
2. \( \text{BMI} = 1 \) for input data values greater than 29.9 kg/m$^2$.

Assuming that in future guidelines for cardiac disease prevention threshold $(t)$ changes to 32.9 kg/m$^2$, this action may lead to changes. These in turn may cause a problem in storing both studies within the same data hypercube. The problem could be solved by defining a hierarchy over the BMI dimension.
In the first level, all the different intervals defined by the two thresholds values are considered as dimension values, namely BMI=I1 for (0,29.9), BMI=I2 for (29.9,32.9) and BMI = I3 for (32.9,\infty).

In the second level a generalised data hypercube is defined of DHg[BMIg] and the one to many relation formed by the mapping: BMIg = 0 for BMI = I1 and BMI = 30 for BMI = I2 or BMI = I3. However, such a mapping holds true for the first study, as it does not alter the thresholds. The right mapping for the second study is: BMIg = 0 for BMI = I1 or BMI = I2 and BMI = 30 for BMI = I3.

The proposed OLAP model handles this problem by adding a second dimension (e.g., YEAR_FOR_STUDY) to the generalised data hypercube (DHg[BMIg, YEAR_FOR_STUDY]), which represents time that might be independent from an ordinary time dimension. The used OLAP model represents a capable method of effectively dealing with this problem. In all the available commercial systems every generalised data hypercube has only one dimension. The resulting dependency tree is shown in Figure 3.

**Figure 3** Proposed dependency tree in the case of more than one studies

![Dependency Tree](image)

### 6 Conclusions

This paper has investigated patterns observed in cardiovascular risk factors that arise from a large population sample of cardiac patients and their matched controls. It has taken into account a number of influencing parameters, such as demographic, clinical and life-style factors.

This research work has evaluated current literature and has highlighted the existing drawbacks regarding implied methodologies in the area of heart disease risk assessment. In particular, the implication of existing methods increases the time complexity of developing effective and efficient outcomes.

For overcoming bottlenecks in the health sector this paper proposes an efficient methodology originating from the field of artificial intelligence (AI). The proposed OLAP methodology enables the user to accurately estimate cardiac risk, since it is based on an additive model and takes into account the interaction between the variables through the joint possibilities of factors occurring together. By using classical statistical approaches one might either overestimate or underestimate the associated synergistic effects of global risk factors.
In view of the above, this research work has introduced a computational algorithm for effectively addressing and optimising computational time and risk assessment. Moreover, the advantage of the particular OLAP model is the correct treatment of the dependencies between the time dimension and the other dimensions that are included in a model. The outcome of this study will benefit the health sector, as it will contribute to better understanding and consequently better prevention of cardiovascular disease.

References


