Vol. 1, No. 1 (2013) 44-48

JOURNAL of COMPUTER SCIENCE
journals.sjp.ac.lk/jcs

Coronary Heart Event Analysis with Association Rule Mining

B. Gomathy[1*], S.M. Ramesh[2] and A. Shanmugam[3]

[1]*Department of CSE, Bannari Amman Institute of Technology, Sathyamangalam, India*

[2,3]*Department of ECE, Bannari Amman Institute of Technology, Sathyamangalam, India*

Email: [1]bgomramesh@gmail.com, [2]drsmramesh@gmail.com

Abstract

Coronary heart disease (CHD) is one of the major causes of disability in adults as well as one of the main causes of death in the developed countries. Although significant progress has been made in the diagnosis and treatment of CHD, further investigation is still needed. The objective of this study was to develop the assessment of heart event-risk factors targeting in the reduction of CHD events using Association Rule Mining. The risk factors investigated were: 1) before the event: a) non modifiable—age, sex, and family history for premature CHD, b) modifiable—smoking before the event, history of hypertension, and history of diabetes; and 2) after the event: modifiable—smoking after the event, systolic blood pressure, diastolic blood pressure, total cholesterol, high density lipoprotein, low-density lipoprotein, triglycerides, and glucose. The events investigated were: myocardial infarction (MI), percutaneous coronary intervention (PCI), and coronary artery bypass graft surgery (CABG).Data-mining analysis was carried out using the Association Rule Mining for the afore mentioned three events using five different splitting criteria for larger datasets. The most important risk factors, as extracted from the classification rules analysis were: 1) for MI, age, smoking, and history of hypertension; 2) for PCI, family history, history of hypertension, and history of diabetes; and 3) for CABG, age, history of hypertension, and smoking. It is anticipated that data mining could help in the identification of high and low risk subgroups of subjects, a decisive factor for the selection of therapy, i.e., medical or surgical.

Keywords: Association rule mining, Coronary heart disease (CHD), Data mining, Risk factors

1. INTRODUCTION

Coronary heart disease (CHD) is the single most common cause of death in Europe, responsible for nearly two million deaths a year. Advances in the field of medicine over the past few decades enabled the identification of risk factors that may contribute toward the development of CHD. However, this knowledge has not yet helped in the significant reduction of CHD incidence. There are several factors that contribute to the development of a coronary heart event. These risk factors may be classified into two categories, not modifiable and modifiable. The first category includes factors that cannot be altered by intervention such as age, gender, operations, family history, and genetic attributes. Modifiable risk factors are those for which either treatment is available or in which alternations in behavior can reduce the proportion of the population exposed. Established, modifiable risk factors for CHD currently include smoking, hypertension, diabetes, cholesterol, high-density lipoprotein, low-density lipoprotein, triglyceride

Data-mining analysis was carried out using the association rule mining using five different splitting criteria for extracting rules based on the aforementioned risk factors. Preliminary results of this study were previously published. In order to improve the patient safety and avoid the underreporting bias,

this paper aims at automatically discovering CHDs that occurred in inpatients. This will be done by identifying situations at risk of CHD by data mining of routinely collected data of past hospitalizations. In those data, the CHDs are not explicitly flagged as no preliminary review is performed. Outpatients' CHDs leading to hospitalization will not be studied.

A list of outcomes will first be defined, and the link between those outcomes and prior drug administration or discontinuations will be studied by means of association rule mining techniques applied on a training set. Rules will be obtained, in which an outcome is explained by a set of drugs in combination with a clinical background, in the form of ADE detection rules (e.g., drug_A & background_B → outcome_C). Then those rules will be applied onto past hospital stays of an evaluation set to get contextualized statistics such as the confidence (e.g., probability of outcome_C when drug_A and background_B are present). Regarding data mining techniques, two issues have to be solved:

1) the temporal constraints have to be taken into account; and
2) we have to use supervised rule-induction methods, although the ADEs are not explicitly flagged in the routinely collected data, which are usually required in the classical rule induction method

TABLE 1: Description of the Hospitals and Stays Used

Hospital	No of stays	Age in yrs	Men proportion	Duration in days
French#1	50072	52.8	29.2%	5.48
French#2	1367	71.4	42.1%	11.4
Danish#1	26245	45.4	51.6%	10.7
Bulgarian	6880	49.4	26.4%	6.96

2. MATERIALS AND METHODS

Data Collection, Cleaning, and Coding Data from 1500 consecutive CHD subjects were collected between the years 2003–2006 and 2009 (300 subjects each year) according to a prespecified protocol, under the supervision of the participating cardiologist (Dr.J. Moutiris, second author of this paper) at the Department of Cardiology, a the Paphos General Hospital in Cyprus. Subjects had at least one of the following criteria on enrollment, history of MI, or percutaneous coronary intervention (PCI), or coronary artery bypass graft surgery (CABG). Data for each subject were collected as given in Table 1: 1) risk factors before the event, a) non modifiable— age, sex, and family history (FH); 2) modifiable— smoking before the event (SMBEF), history of hypertension (HxHTN), and history of diabetes (HxDM); and 2) risk factors after the event, modifiable—smoking after the event (SMAFT), systolic blood pressure (SBP) in mmHg, diastolic blood pressure (DBP) in mmHg, total cholesterol (TC) in mg/dL, high-density lipoprotein (HDL) in mg/dL, low-density lipoprotein (LDL) in mg/dL, triglycerides (TG) in mg/dL, and glucose (GLU) in mg/dL.

To clean the data, the fields were identified, duplications were extracted, missing values were filled, and the data were coded as given in Table 2. After data cleaning, the number of cases was reduced as given in Table 3, mainly due to unavailability of biochemical results.

TABLE 2: Coding of Risk Factors

Risk factors before the event : non modifiable

Risk factor	Code1	Code 2	Code3	Code4
AGE	1:34-5		2:51-6	3:61-7
SEX	M:male	F:female		
FH	Y:yes	N:no		

Risk factors before the event: modifiable

SMBEF	Y:yes	N:no
HxHTN	Y:yes	N:no
HxDM	Y:yes	N:no

Table 3: No. of Cases Per Set of Rules/Models Investigated

	Model	MI	PCI	CABG
Event	Yes	378/75/75	72/36/36	86/43/43
	No	150/75/75	274/36/36	307/43/43

2.1 Aggregation of the Complex Data of the Stays Into Simple Events General Principles

The data described in the data repository are characterized by a complex data scheme, very numerous classes (about 17 000 codes for ICD10, about 5400 codes for the ATC, etc.) and repeated measurements throughout the hospitalization (e.g., laboratory parameters and drug administrations). Those characteristics make those data too complex to be mined using statistical methods. The aim of the data-to-event aggregation process is to automatically get a simpler representation of data for data mining purposes. Aggregation engines are developed in order to transform the available data into information described as sets of events. For each kind of data (administrative information, diagnoses, drugs, and laboratory results), a specific aggregation engine is developed and fed with a mapping. Each mapping is described by means of extensible markup language (XML) files outside the engine. The aggregation engines enable to describe the events in terms of binary variables complemented by start and stop dates. Those engines are not static and can be adapted with respect to the context.

2.2 Identification of the Outcomes in Relation with CHDs

As described, a list of outcomes is extracted from the summaries of product characteristics. The outcomes are traced in the data essentially by screening the laboratory results and administered drugs; this is possible through different ways depending on the category of outcome. For instance, the occurrence of a hyperkalemia (laboratory-related outcome) is directly traced using the potassium level in the blood. The occurrence of a hemorrhage under vitamin K antagonists (VKA) can be traced through different ways: 1) an increase of the international normalized ratio (INR), a laboratory parameter that rises up in case of VKA overdose; and 2) the vitamin K administration, an antidote which is prescribed in case of hemorrhage under VKA. The structured SPC database describes 228 different kinds of outcomes. 83 (37%) of those outcomes are traceable in this paper, due to the available data.

Duplicate entries are then removed; for instance, in the initial list, “hyperbilirubinemia” is also

described using two synonyms, "bilirubinemia higher than twice the normal upper bound" and "jaundice." As a consequence, those 83 outcomes are traced through 56 different variables. Those outcomes correspond to life-threatening CHDs, such as hyperkalemia of hemorrhage hazard. Unfortunately, some outcomes cannot be traced in the data. This is the case especially for minor clinical incidents such as nausea or gastric pain cannot be traced. Those outcomes could correspond to ICD10 codes but in most hospitals, such codes are not flagged with a date.

2.3 Expert Validation and Reorganization of the Rules :

It is mandatory to filter, validate, and organize the rules that are obtained from the data mining: as the rules have to be used by physicians, they must provide simple, validated, and unquestionable knowledge. Several meetings are organized with external experts (physicians, pharmacologists, pharmacists, and statisticians) to filter and reorganize the set of rules. The rules are examined and validated against the SPCs and scientific references. During the review, the experts may ask for complementary queries on the potential CHD cases. At this step, the experts may manually add a few rules that are considered as mandatory although they were not discovered by the data mining process, for instance because the conditions of the rules never occur (e.g., absolute contraindication) or because the conditions occur but do not lead to any outcome. In every rule, there is a set of conditions; the experts are asked to characterize each condition according to one of the following types.

1) Segmentation: Conditions are conditions that do not explain why an outcome occurs, but deeply change its probability. This kind of condition enables us to reduce over alerting.
2) Subgroup: Conditions are fixed when, for some medical reasons, it does not make sense to consider Rules are stored in a rule repository. A machine evaluation automatically computes various statistics (occurrences) of the rules in every medical department the sample before computing the statistics. The following subgroups are systematically defined.
    - The INR deviations or vitamin K administrations are only explored for VKA-treated patients.
    - The increase of activated partial thromboplastin time is only explored for heparin-treated patients.
    - The hyperkalemia is explored separately for patients suffering from renal insufficiency or not.
3) Basic: Conditions group together all the other conditions.

3. RESULTS AND DISCUSSION

In this paper, 56 different outcomes enable to trace the potential consequences of CHDs. The supervised rule induction generates rules that predict each outcome. The rules are always filtered, validated, and tuned by the expert committee. 236 validated rules are obtained. The experts also add some rules that appear to be important in the academic knowledge and are not discovered by the data mining (e.g., the conditions never occur, or occur but not lead to the outcome). Over the 56 outcomes, we have the following.

- Twenty-seven kinds of outcomes are observed and enable to discover CHD detection rules.
- Ten outcomes are never or too rarely observed in the data, so that no rule is discovered. Data mining will be performed on larger datasets to get results.
- Eighteen outcomes are observed but cannot be explained by the use of drugs in the available dataset: the medical background of the patient is a sufficient explanation, so that no rule is discovered.

4. CONCLUSION & FUTURE WORK

This paper brings innovative and semi automated solutions for CHD detection. The method is quite generic and could be applied to other kinds of data as soon as they are available in the EHR, such as structured results of electrocardiograms. The results of the method used here bring an important

contribution to CHD knowledge. The rules that are obtained are versatile and can be used either as detection rules on past hospital stays, or as prevention rules in a CDSS context. Those rules are already loaded in several prototypes that are developed in the frame of the PSIP Project. CDSS is the one embedded in a computerized physician order entry, another embedded in an EHR, and a prescription simulation tool that is available even without any Hospital information system.

5. REFERENCES

[1] Euro aspire study group, "A European Society of Cardiology survey of secondary prevention of coronary heart disease: Principal results," Eur. Heart J., vol. 18, pp. 1569–1582, 1997.
[2] Euro aspire II Study Group, "Lifestyle and risk factor management and use of drug therapies in coronary patients from 15 countries," Eur. Heart J., vol. 22, pp. 554–572, 2002.
[3] Euro aspire study group, "Euro aspire III: A survey on the lifestyle, risk factors and use of cardio protective drug therapies in coronary patients from 22 European countries," Eur. J. Cardiovasc. Prev. Rehabil., vol. 16, no. 2, pp. 121–137, 2009.
[4] T. Marshall, "Identification of patients for clinical risk assessment by prediction of cardiovascular risk using default risk factor values," Br. Med. Assoc. Public Health, vol. 8, pp. 25, 2008.
[5] W. B. Kannel, "Contributions of the Framingham Study to the conquest of coronary artery disease," Amer. J. Cardiol., vol. 62, pp. 1109–1112, 1988.
[6] M. Karaolis, J. A.Moutiris, and C. S. Pattichis, "Assessment of the risk of coronary heart event based on data mining," in Proc. 8th IEEE Int. Conf. Bioinformatics Bio eng., pp. 1–5, 2008.