35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Prediction of Coronary Atherosclerosis Progression using Dynamic Bayesian Networks

Konstantinos P. Exarchos, Themis P. Exarchos *Member, IEEE*, Christos V. Bourantas, Michail I. Papafaklis, Katerina K. Naka, Lampros K. Michalis, Oberdan Parodi and Dimitrios I. Fotiadis, *Senior Member, IEEE*

Abstract — **In this paper we propose a methodology for predicting the progression of atherosclerosis in coronary arteries using dynamic Bayesian networks. The methodology takes into account patient data collected at the baseline study and the same data collected in the follow-up study. Our aim is to analyze all the different sources of information (Demographic, Clinical, Biochemical profile, Inflammatory markers, Treatment characteristics) in order to predict possible manifestations of the disease; subsequently, our purpose is twofold: i) to identify the key factors that dictate the progression of atherosclerosis and ii) based on these factors to build a model which is able to predict the progression of atherosclerosis for a specific patient, providing at the same time information about the underlying mechanism of the disease.**

I. Introduction

Atherosclerosis is a disease of large vessels involving the build-up of plaque in the atherosclerotic wall. Cardiovascular morbidity and mortality attributed to the development of atherosclerotic lesion is projected to be an epidemic in the next decades [1]. Enhanced prediction of atherosclerosis progression is crucial for improving patient outcomes. Several studies have attempted to reveal significant correlations between patient characteristics and/or examinations (e.g. age, smoking, exercise stress test etc.) and the development of coronary artery disease which is commonly measured using the result of the coronary angiography examination. The majority of these studies focus on computing correlations between single features and the angiography outcome.

Lately, advanced data mining techniques have been presented for the development of "diagnosis and decision support systems", which could have a predictive value for the outcome of coronary angiography or the progression of atherosclerosis. In a recent study, we employed a dataset of 200 patient records and the objective was to classify the patients as having no stenosed vessel or at least one vessel with stenosis, based on the result of coronary angiography [2]. The standard demographic and clinical features were used (age, gender, family history, blood tests, blood pressure, etc.) and additional indices based on non-invasive pulse wave velocity were also employed. Various algorithms for developing the classifiers were used and the best results were reported using a classifier based on decision trees and fuzzy modeling (73% accuracy). Another data mining technique was based on a dataset with 655 patients and 202 features. The dataset contains for each patient a detailed description of the stenosis of each of the four arteries. The authors used decision trees and association rule mining to develop diagnosis systems for predicting the stenosis in every artery [3,4].

Several studies have been presented in the literature, aiming to identify significant correlations between Single Nucleotide Polymorphisms (SNPs) and atherosclerosis progression, either in humans or in animals. Reiner et al. [5] studied European American and African American populations and specifically assessed the correlation between 25 common polymorphisms and the level of P-selectin, as well as the carotid intima-media thickness. In a methodologically different approach Timinskas, et al. [6] applied text mining techniques to the abstract texts of PubMed articles. Szymczak, et al. [7] analyzed a 300 kb segment of the 9p21.3 chromosome, using mutual information-based methods, for potential implications with atherosclerosis, subsequently pinpointing three responsible genes (ADM, FCGR3B and ADORA1). Stangard et al. [8] studied the association between SNPs of the apolipoprotein E gene (APO E) in turn with the levels of HDL-cholesterol, triglycerides, and/or the total cholesterol, using combinatorial partitioning methodologies. Yosef, et al. [9] employed Support Vectors Machines in order to predict the plasma lipid levels. Pan, et al. [10] mined for associations between several SNPs and a series of mice phenotypes using

This work is partly funded by the European Project ARTreat (FP7-224297 - Multi-level patient-specific artery and atherogenesis model for outcome prediction, decision support treatment, and virtual hand-on training).

K. P. Exarchos and T. P. Exarchos are with the Foundation for Research and Technology Hellas, Institute of Molecular Biology and Biotechnology, Department of Biomedical Research, Ioannina, GREECE, GR 45110 (e-mail: exarchos@cc.uoi.gr, kexarcho@cc.uoi.gr).

C. Bournantas is with the Dept. of Academic Cardiology, Castle Hill Hospital, Cottingham, HU 16 5JQ, East Yorkshire, UK (email: cbourantas@gmail.com).

M. Papafaklis is with the Harvard Medical School, Cardiovascular Division, Brigham and Women's Hospital, Boston, MA 02115, USA (e-mail: m.papafaklis@yahoo.com).

L.K. Michalis is with the Michaelideion Cardiac Center, Dept. of Cardiology in Medical School, University of Ioannina, GR 45110 Ioannina, Greece (email lmihalis@cc.uoi.gr).

O. Parodi is with the Institute of Clinical Physiology, National Research Council, Pisa, 56124, Italy (email oberpar@tin.it)

D. I. Fotiadis is with the Unit of Medical Technology and Intelligent Information Systems, University of Ioannina and with the Foundation for Research and Technology Hellas, Institute of Molecular Biology and Biotechnology, Department of Biomedical Research Ioannina, GREECE, GR 45110 (tel.: +302651008803, fotiadis@cc.uoi.gr)

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

a tree-based approach, yielding ten candidate genes correlated with the levels of HDL-C.

In the current report, we have gathered and analyzed a multitude of heterogeneous data for a set of patients under consideration. Specifically, we have assembled a dataset of patients with a large variety of characteristics which have never been previously used for the prediction of atherosclerosis progression. The resulting dataset contains data coming from the patient's medical history, risk factors, inflammatory markers, demographic data as well as additional pieces of information, that are explicitly presented in the sections that follow. Our aim is to analyze all these sources of information in order to frame all possible manifestations of the disease; subsequently, our purpose is twofold: (i) to identify the key factors that dictate the progression of atherosclerosis and (ii) based on these factors build a model that is able to assess the prognosis of a specific patient regarding the cardiovascular risk, and to predict the severity of stenoses in the future, providing at the same time information about the underlying mechanisms of the disease.

II. METHODS

Dynamic Bayesian Networks (DBNs) [11] are specifically tuned to capture temporal dependencies over time, and thus constitute a very appealing choice for modeling atherosclerosis evolution. DBNs have been previously employed for similar purposes [12,13]. The steps followed for this analysis are depicted in Fig. 1. First, the collected data are preprocessed with a twofold objective: (i) to prepare subsets of patients that will formulate the basis of the subsequent analysis in order to make the most out of our dataset, and (ii) to discretize the values from the employed variables into clinically meaningful bins that will facilitate the invocation of the algorithms as well as the interpretation of the results. Next, the features under consideration are fed as input to a DBN aiming to model the progression of the disease. From the Dynamic Bayesian Network we are able to identify the most significant factors affecting the progression of atherosclerosis based on the features retained by the algorithm and additionally extract the dependencies among those variables potentially gaining further insight regarding the atherosclerosis mechanism. Consequently, we are able to conjecture about the progression of the stenosis for any patient and whether he/she belongs to a high or low risk group.

Prior to the utilization of the DBNs, the input data need to be preprocessed accordingly. Specifically, all variables must be discretized in order to be able to invoke the training algorithms of the DBNs, since continuous variables cannot be employed. All variables considered are discretized in a hierarchical manner into four classes; the only exception is the adhesion molecules, i.e. Selectin, ICAM-1, VCAM-1 which have been discretized using clinically meaningful thresholds [14] into three categories, denoting the status of the patient. Afterwards, we aim to remove all redundant and irrelevant features and come down to a more manageable list of features (using Fischer's exact test between each variable and the outcome).

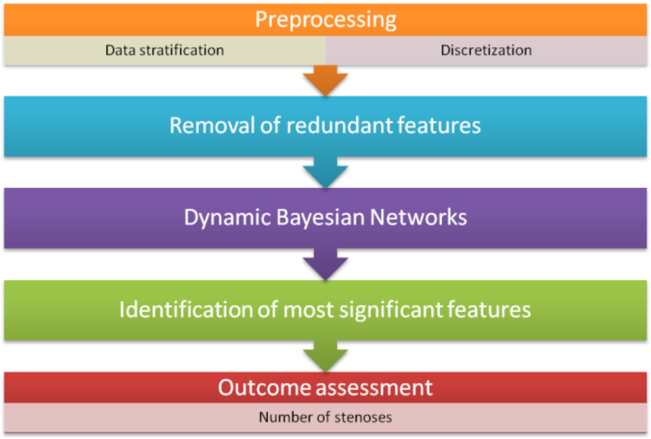
Preprocessing
Data stratification
Discretization
Removal of redundant features
Dynamic Bayesian Networks
Identification of most significant features
Outcome assessment
Number of stenoses

Fig. 1: Flowchart of the Dynamic Bayesian Network analysis.

Our strategy is to identify those variables that exhibit a high p-value. The threshold posed to assess the statistical significance of each test is 0.1. Variables with a p-value higher than this threshold are discarded, whereas the retained variables (with p-value<0.1) serve as the input for the subsequent training of the DBN. This step is rather imperative, since the number of employed features is relatively large considering the number of patients, thus, hindering the training of the DBN; besides, the employment of a more refined set of features allows for more effective and efficient training of the algorithm. There was also the case that certain features remained unaltered (constant) for all patients; these features were also removed since they did not contribute to the discrimination process.

After these two steps, DBNs are implemented. DBNs constitute temporal extensions of the standard Bayesian Networks. A Bayesian Network is actually a directed acyclic graph, where each node of the network corresponds to one of the employed features. Specifically for a provisional network described as *B=(G,P),* where *G* is a directed acyclic graph, $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$, is a set of random variables, and *P* is the joint probability distribution of variables in X, factorizing as follows:

$$P(\mathbf{X}) = \prod_{i=1}^{N} P(x_i \mid \pi_G(x_i)), \tag{1}$$

where $\pi_G(x)$ denotes the parents of x in G. A DBN is defined as a pair $DB=(B_0,B_{trans})$, where B_0 is a BN, defining the prior $P(X_0)$ and B_{trans} is a two-slice temporal BN (2TBN) which defines $P(X_t|X_{t-1})$. The semantics of a DBN can be defined by “unrolling” the 2TBN until we have T time-slices. The resulting joint distribution is given by:

$$P(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(x_i^t \mid \pi_G(x_i^t)). \tag{2}$$

We employ the training data in order to define the architecture of the network, i.e. specify the dependencies among the variables within one time slice and across consecutive time slices; the former dependencies are called intra-slice whereas the latter inter-slice dependencies. For this purpose, we utilize two algorithms, namely the Bayesian Search and the PC algorithm [15], to search across the feature space and pinpoint the network architecture that yields the best performance. After the dependencies among all variables of the network have been defined, we may provide new evidence to the trained model and estimations about the value of each variable in the network in any of the succeeding time slices. Due to the transparent architecture of the DBNs, we are able to deduce information regarding the interactions of the variables under consideration and consequently the mechanism of the atherosclerosis procedure. Moreover, DBNs represent temporal causalities among the variables and therefore depict a better approximation of the actual procedure rather than a static snapshot that is possibly lacking important temporal information.

During the search across all possible network architectures, we come down to a list of the most promising ones, i.e. the ones achieving the highest performance. Among those architectures, there are features that appear more often in the best performing architectures; thus, we are able to identify the factors that play a key role in the progression of atherosclerosis.

III. Dataset

In the current work, a set of 39 patients, enrolled in ARTreat project at the Institute of Clinical Physiology, National Research Center of Italy (IFC CNR), was used; several features were included. All patients were diagnosed with coronary atherosclerosis and have been monitored for a long follow-up period, ranging from 3 up to 53 months (median follow-up: 36 months), after the baseline assessment. The available data cover aspects and manifestations of the disease from the cellular level, to the organ level and up to the whole organism level. In addition, it should be noted that all data have been recorded at two time points, once at the time of the first diagnosis (i.e. the baseline period) and again during the follow-up period, thereby capturing the serial nature of the disease. This is particularly important, since atherosclerosis is a chronic evolving condition, rather than a static one-time manifestation. The types of features/variables as well as the specific features that have been measured are shown in Table I.

All types of variables except the demographic ones weren measured at baseline as well as during the follow-up. All these features were used in order to assess the status and the progression of the coronary atherosclerosis. For each patient under consideration, assessment of the severity of coronary stenoses was carried out by analysis of coronary Computed Tomography Angiography (CTA), using dedicated software for lesion characterization (General Electric Healthcare) after visual analysis of the atherosclerotic segments. The percentage of lumen diameter reduction, as compared with immediately proximal and distal reference segments, has been measured at the time of baseline CTA as well in images obtained at follow-up. The outcome that is predicted by the DBN is the number of stenoses existing at follow-up.

IV. Results

To evaluate the methodology, we employed the leave-one-patient out technique which is specifically suited for limited datasets, as it exploits the largest possible patient cohort for training and also uses all patients for testing. Moreover, for the case of DBNs, this notion is further extended to the time-slices that are taken into consideration in the model. In detail, we may feed evidence to the model for the first time-slice and predict the values of all variables in the succeeding time slices; then we may feed evidence to the trained model for the first and second time-slice aiming to predict the values of the variables in remaining time-slices. As expected, he more evidence is presented to the model, the outcomes are further refined yielding more accurate results. From this point on, the variables maintained are fed as input to the training algorithms of the DBNs. For the number of stenoses as outcome, the features Diabetes, Hypercholesterolemia, Cholesterol/HDL, Triglycerides and Glucose qualified to the next step.

Table I: Variables used in the current analysis, sorted according to the source of data.

Demographic				
Age	Sex	Height		
Clinical				
Weight	BMI	Diabetes	Family History	Hypercholesterolemia
Hypertension	Smoke	Framingham	Angina	Infarct
Infarct site	LVEF (Left Ventricular Ejection Fraction)			
Biochemical analytes				
Cholesterol	HDL	Cholesterol/HDL	LDL	Triglycerides
Creatinine	Creatinine clearance	Glucose	HCT	MCV
Monocytes	WBC			
Adhesion molecules				
E-Selectin	ICAM-1	VCAM-1		
Monocyte markers				
HLA DR %	HLA DR RFU	CX3CR1 %	CX3CR1 RFU	CCR2 %
CCR2 RFU	CD11b %	CD11b RFU		
Therapy				
ACE/ARB	Aspirin	Beta-blockers	Ca Antagonists	Statins

Using the aforementioned features, we employ the Bayesian Search algorithm which yields the network shown in Fig. 2. Table II shows the results achieved with both algorithms (Bayesian Search and PC), towards the prediction of the number of stenoses. All figures contained in the table refer to accuracy metrics. The numbers below the column "Baseline" denote the accuracy achieved towards the prediction of the number of stenoses using solely information from the baseline visit, whereas, the number under the column entitled "Follow-up #1" denotes the accuracy achieved when additional information from the first baseline visit becomes available. The last row in the bottom of the table contains the overall accuracy achieved by the model for both cases.

Table II: Results yielded by the Dynamic Bayesian Network for the outcome: number of stenoses.

	Bayesian Search		PC	
	Baseline (Acc. %)	Follow-up #1 (Acc %)	Baseline (Acc %)	Follow-up #1 (Acc %)
	72	87	72	87
Overall accuracy	79		79	

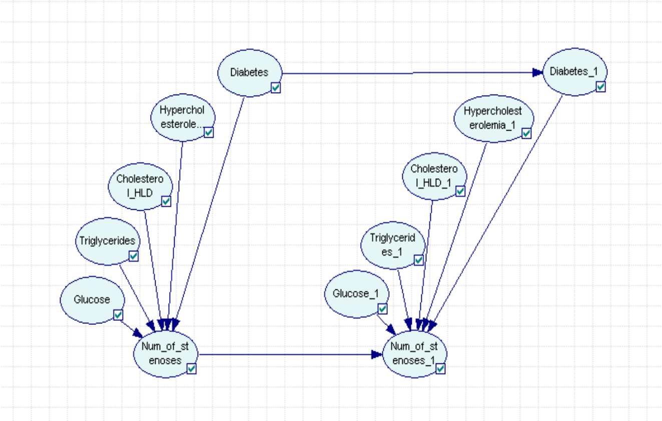

Fig. 2: Dynamic Bayesian Network built with the Bayesian Search algorithm.

As shown in Fig.2 the employed DBN architecture contains 5 features that have been found to affect more prominently the progression of atherosclerosis, as expressed by the number of stenoses. The resulting set constitutes a manageable feature subset, compared to the initial set of 43 attributes, achieving at the same time quite promising results. The contribution and implication of the retained features have been well documented in the literature, thus, supporting the results of our analysis.

V. Conclusions

In conclusion, we developed a set of models that capture and predict the progression of atherosclerosis, taking into account a large set of parameters; this multitude of input variables encompassing clinical, biochemical and treatment characteristics as well as additional factors constitutes a crucial novelty of the current work. In order to further verify and validate our results, it is important to perform similar analyses in larger populations of patients that can support more concrete conclusions.

During all the aforementioned analyses, certain features have been repeatedly identified as significant, either by being maintained by feature selection algorithms, learning algorithms or statistical testing. Among the most prominent features were the following: Diabetes, Cholesterol, Cholesterol/HDL, VCAM-1, hypercholesterolemia and family history of coronary disease. Consequently, these features need to be carefully investigated by medical experts but also to be further verified using larger patient datasets.

References

[1] World Health Organization. Cardiovascular Diseases, 2009.

[2] M.G. Tsipouras, T.P. Exarchos, D.I. Fotiadis, A.P. Kotsia, K.V. Vakalis, K.K. Naka, L.K. Michalis, "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling", IEEE Trans. on Biomedical Engineering, vol. 12, Issue 4, pp. 447-458, 2008.

[3] C. Ordonez, "Comparing Association Rules and Decision Trees for Disease Prediction", Proc. ACM HIKM'06, Arlington, 2006.

[4] C. Ordonez, N. Ezquerra and C. Santana, "Constraining and Summarizing Association Rules in Medical Data", Knowledge and Information Systems, vol. 9, Issue 3, pp. 259-283, 2006.

[5] A. Reiner, C. Carlson, B. Thyagarajan, M. Rieder, J. Polak, D. Siscovick , D. Nickerson, D. Jacobs Jr, and M. Gross. "Soluble P-Selectin, SELP Polymorphisms, and Atherosclerotic Risk in European-American and African-African Young Adults", in Arteriosclerosis, Thrombosis and Vascular Biology, August 2008.

[6] A. Timinskas, Z. Kucinskiene, and V. Kucinskas. "Atherosclerosis: alterations in cell communication", in ACTA MEDICA LITUANICA, vol. 14, Issue 1. P. 24–29, 2007

[7] S. Szymczak, B.W. Igl, and A. Ziegler. "Detecting SNP-expression associations: A comparison of mutual information and median test with standard statistical approaches" in Statistics in Medicine, vol. 28, pp. 3581–3596, 2009.

[8] J. Stangard, S. Kardia, S. Hmon, R. Schmidt, A. Tybjaerg-Hansen, V. Salomaa, E. Boerwinkle, and C. Sing. "Contribution of regulatory and structural variations in APOE to predicting dyslipidemia", in The Journal of Lipid Research, vol. 47, pp. 318-328, 2006.

[9] N. Yosef, J. Gramm, Q. Wang, W. Noble, R. Karp, and R. Sharan."Prediction Of Phenotype Information From Genotype Data", in Communications In Information And Systems, vol. 10, Issue 2, pp. 99-114, 2010.

[10] F. Pan, L. McMilan, F. Pardo-Manuel De Villena, D. Threadgill, and W. Wang."TreeQA: Quantitative Genome Wide Association Mapping Using Local Perfect Phylogeny Trees", in Pac Symp Biocomputing, pp. 415-426, 2009.

[11] Z. Xiang., R. M. Minter, et al. "miniTUBA: medical inference by network integration of temporal data using Bayesian analysis." Bioinformatics vol. 23(18), pp. 2423-2432, 2007.

[12] M. A. van Gerven, B. G. Taal, et al. "Dynamic Bayesian networks as prognostic models for clinical patient management." J. Biomed. Inform. vol.41, pp. 515-529, 2008.

[13] A. H. Marshall, L. A. Hill, et al. "Continuous Dynamic Bayesian networks for predicting survival of ischaemic heart disease patients", IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS), 2010

[14] P. Tretjakovs, A. Jurka, et al. "Circulating adhesion molecules, matrix metalloproteinase-9, plasminogen activator inhibitor-1, and myeloperoxidase in coronary artery disease patients with stable and unstable angina." Clin Chim Acta, vol. 413, pp. 25-29, 2012.

[15] P. Spirtes, C. Glymour, and R. Sheines, "Causation, Prediction and Search". Cambridge, Massachusetts. U.S.A.: MIT Press, 2000.