Title: Bayesian Network Modeling: a Case Study of an Epidemiologic System Analysis of Cardiovascular Risk

Author: P. Fuster-Parra P. Tauler M. Bennasar-Veny A. Ligęza A.A. López-González A. Aguiló

PII: S0169-2607(15)30114-0
DOI: http://dx.doi.org/doi:10.1016/j.cmpb.2015.12.010
Reference: COMM 4035

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 18-8-2015
Revised date: 28-11-2015
Accepted date: 11-12-2015

Please cite this article as: P. Fuster-Parra, P. Tauler, M. Bennasar-Veny, A. Ligęza, A.A. López-González, A. Aguiló, Bayesian Network Modeling: a Case Study of an Epidemiologic System Analysis of Cardiovascular Risk, *<![CDATA[Computer Methods and Programs in Biomedicine]]>* (2015), http://dx.doi.org/10.1016/j.cmpb.2015.12.010

Highlights

- An epidemiologic system analysis of cardiovascular risk is presented through a Bayesian network model.
- The Bayesian networkmodel can serve as a generic tool for application oriented activities: explanation, prediction, monitoring and prevention.
- Due to Cardiovascular Disease is multifactorial, the application of this kind of model is of special interest, both from theoretical and practical point of view.
- The induced Bayesian network was used to make inferences taking into account three reasoning patterns: causal reasoning, evidential reasoning, and intercausal reasoning.

Bayesian Network Modeling: a Case Study of an Epidemiologic System Analysis of Cardiovascular Risk

P. Fuster-Parra[a,b], P. Tauler[b], M. Bennasar-Veny[b], A. Ligęza[c], A.A. López-González[d], A. Aguiló[b]

[a]*Department of Mathematics and Computer Science, Universitat Illes Balears, Palma de Mallorca, Baleares, E-07122, Spain. E-mail: pilar.fuster@uib.es*
[b]*Research Group on Evidence, Lifestyles & Health. Research Institute on Health Sciences (IUNICS). Universitat Illes Balears, Palma de Mallorca, Baleares, E-07122, Spain.*
[c]*Department of Applied Computer Science, AGH University of Science and Technology, Kraków, PL-30-059, Poland*
[d]*Prevention of Occupational Risks in Health Services, GESMA, Balearic Islands Health Service, Hospital de Manacor, Manacor, Baleares, E-07500, Spain*

Abstract

An extensive, in-depth study of cardiovascular risk factors (CVRF) seems to be of crucial importance in the research of cardiovascular disease (CVD) in order to prevent (or reduce) the chance of developing or dying from CVD. The main focus of data analysis is on the use of models able to discover and understand the relationships between different CVRF. In this paper a report on applying Bayesian network (BN) modeling to discover the relationships among thirteen relevant epidemiological features of heart age domain in order to analyze *cardiovascular lost years* (CVLY), *cardiovascular risk score* (CVRS), and *metabolic syndrome* (MetS) is presented. Furthermore, the induced BN was used to make inference taking into account three reasoning patterns: *causal* reasoning, *evidential* reasoning, and *intercausal* reasoning. Application of BN tools has led to discovery of several direct and indirect relationships between different CVRF. The BN analysis showed several interesting results, among them: CVLY was highly influenced by smoking being the group of men the one with highest risk in CVLY; MetS was highly influence by physical activity (PA) being again the group of men the one with highest risk in MetS, and smoking did not show any influence. BNs produce an intuitive, transparent, graphical representation of the relationships between different CVRF. The ability of BNs to predict new scenarios when hypothetical information is introduced makes BN modeling an Artificial Intelligence (AI) tool of special interest in epidemiological studies. As CVD is multifactorial the use of

BNs seems to be an adequate modeling tool.

Keywords: Bayesian Networks, model averaging, cardiovascular lost years, cardiovascular risk score, metabolic syndrome, causal dependency discovery

1. Introduction

Bayesian Networks (BNs) [1, 2] also referred to as *Belief Networks* or probabilistic causal networks are an established framework for uncertainty management in Artificial Intelligence (AI). They constitute a tool which combines graph theory and probability theory to represent relationships between variables (nodes in the graph) [3]. Contrary to deterministic understanding of the *causality* phenomenon [4], BN modeling has its origins within data mining and machine learning research [5, 6] and captures probabilistic influences induced out of big data sets. They constitute a powerful knowledge representation and an efficient reasoning tool under conditions of uncertainty [7]. The network structure is a directed acyclic graph (DAG) where each node represents a random variable [8, 9] and the arcs are suitable for representing causality [10].

BNs have been proven to be a strong tool to discover the relationships between variables that attempts to separate out direct and indirect dependencies [11, 12], and can capture the way an expert understands the relationships among all the features [13]. BN modeling is widely used in fields like clinical decision support [14], systems biology [15, 16], human immunodeficiency virus (HIV) and influenza research [17, 18], analyzes of complex disease systems [19, 20, 21], interactions between multiple diseases [22], and also in diagnostic diseases [23, 24, 25, 26, 27].

The metabolic syndrome is a set of risk factors that include abdominal obesity, insulin resistance, dyslipidemia and hypertension leading to increased risk of developing cardiovascular diseases and type 2 diabetes [28, 29, 30, 31]. Cardiovascular disease (CVD) epidemiology is a worldwide public health problem [32]. The economic burden of CVD is already affecting the economies of the world's wealthiest countries. However, in the next decades developing countries will be more affected due to the great increase in CVD prevalence expected in these countries [33]. It is estimated that in 2015, more than 20 million people may die worldwide because of CVD. This number is expected to increase in the upcoming decades, that every 5 seconds in the world a myocardial infarction would occur [34, 35].

CVDs are closely related to the well-known cardiovascular risk factors (CVRF). The concept of CVRF appeared in 1961, when the group of Kanned defined CVRF

as biological traits or behaviors that increased the chance of developing or dying from CVD [36, 37]. The high prevalence of certain risk factors to which we are exposed is the cause of this situation, in which the prevalence of CVD is increased every year. It is necessary to control the factors that influence the development of CVD, such as smoking, hyperlipidemia, hypertension, diabetes, obesity, a diet high in saturated fats, alcohol abuse, a sedentary lifestyle, and stress [38]. In fact, WHO (World Health Organization) estimates that 80% of premature deaths from cardiovascular disease and diabetes could be prevented by efficient controlling these risk factors [39].

There are some scores that numerically quantify cardiovascular risk (CVR). One of the most widely used is Framingham score, with its calibrated form for the Spanish population, the Framingham-REGICOR [35]. This scale estimates the global CVR to 10 years and it is expressed as a percentage. Recently, a new score has been proposed, the so-called Heart Age tool (HA), which is based on Framingham score, and supposes a simple and graphic way to communicate the CVR because it expresses the CVR as an age. If the HA value is older than chronological age the term "lost years", defined as the HA minus the chronological age, could be used. The HA is a novel concept designed specifically to help people to understand their own cardiovascular disease risk and implement changes into their lifestyles to prevent the incidence of CVD [40].

Development and analysis of models to examine the relationships between different CVRF could be not only of theoretical interest, but can serve as a generic tool for application oriented activities: explanation, prediction, monitoring and prevention. It enables both theoretical analysis of the relationships between numerous variables, and having in mind the probabilistic nature of the causal dependencies, BNs seem to be an adequate tool. Moreover, BN models are capable of creating different scenarios based on hypothetical cases when new observations are instantiated.

The paper is organized as follows. Section 2 introduces BNs and some basic concepts for inference flow. Section 3 presents the materials and methods for the epidemiologic study and the process of inducing a BN from a data set. Section 4 shows different reasoning patterns to analyze the BN. Section 5 presents a discussion. Finally, Section 6 concludes the paper.

2. Bayesian networks

A BN consists of [41]: (i) a set of variables and a set of directed edges between these variables, where (ii) each variable has a finite set of mutually exclu-

sive states, and (iii) the variables together with the directed edges form a DAG. BN models estimate the joint probability distribution P over a vector of random variables $\mathbf{X} = (X_1, \ldots, X_n)$. The joint probability distribution factorized as a product of several conditional distributions denotes the dependency/independency structure by a DAG:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i^{\mathcal{G}})) \tag{1}$$

Equation (1) (where $Pa(X_i^{\mathcal{G}})$ denotes the parent nodes of X_i) is the main reason for the formulation of a multivariate distribution by BNs; this equation is also called the *chain rule for Bayesian networks*.

As BNs are used to make inference [8], it is necessary to understand the flow of influence when new information is introduced in a BN. Below we introduce some basic concepts.

Two variables X and Y in a BN are *d-separated* if, for every possible path between X and Y, there is an intermediate variable Z such that either: (i) the connection is serial ($X \rightarrow Z \rightarrow Y$ or $X \leftarrow Z \leftarrow Y$) or diverging ($X \leftarrow Z \rightarrow Y$) and Z is instantiated, or (ii) the connection is converging ($X \rightarrow Z \leftarrow Y$) and neither Z nor any of Z's descendants have received evidence. When influence flows from a node X to another node Y via a node Z, it is said that the trail $X \rightleftharpoons Z \rightleftharpoons Y$ is active. A causal trail $X \rightarrow Z \rightarrow Y$ (serial connection), an evidential trail $X \leftarrow Z \leftarrow Y$ (serial connection) or, a common cause trail $X \leftarrow Z \rightarrow Y$ (diverging connection) is active if and only if Z is not observed. A common effect trail $X \rightarrow Z \leftarrow Y$ (converging connection) is active if and only if either Z or one of Z's descendants is observed.

Let P be a joint probability distribution of the random variables in some set of features $\mathbf{F}$, the set of arcs is denoted by $\mathbf{A}$, and a DAG $\mathbf{G} = (\mathbf{F}, \mathbf{A})$; then $(\mathbf{G}, P)$ satisfies the local Markov condition if for each variable (feature) $X \in \mathbf{F}$, X is conditionally independent of the set of all its non-descendants given the set of all its parents. The global Markov property states that any node X is conditionally independent of any other node given its *Markov blanket*, i.e., $I(X, \textit{non-markov-blanket}(X) \mid \textit{markov-blanket}(X_i))$; the *Markov blanket* of a node includes its parents, its children, and the children's other parents (spouses). Any node in the BN would be *d-separated* of the nodes belonging to the non-Markov blanket given its Markov blanket.

3. Data and Methodological Issues

This section presents some methodological issues concerning data acquisition. Reliability of data was assured due to standard medical procedures. A brief description follows.

3.1. Participants

All participants were workers from the public sector of the Balearic Islands (Spain). Subjects in the study were invited to participate during their annual work health assessment. Any worker attending the work health assessment could be included in the study. 4300 workers were invited to participate. Among them, 3993 subjects (Men = 1758, Women = 2235) agreed to participate. Participants signed informed consent prior to enrollment. After acceptance, a complete family and personal medical history was recorded. The project of the study was in accordance with the Declaration of Helsinki and received approval from the Balearic Islands Clinical Research Ethical Committee.

3.2. Instruments

3.2.1. Determining variables

All anthropometric measurements were made in the morning after an overnight fast, and according to the recommendations of the International Standards for Anthropometric Assessment [42]. Body weight (electronic scale Seca 700; Seca, Hamburg, Germany), height (stadiometer Seca 220 cm), and abdominal waist circumference (Lufkin Executive Thinline W606, precision 1 mm) were determined according to recommended techniques mentioned above. Body mass index (BMI) was calculated as weight (kg) divided by height (m) squared. BMI values were categorized following the criteria from WHO [39].

Blood samples were collected during the same session and in the same place after an overnight fast of 12 hours. Serum was obtained and total cholesterol, HDL cholesterol, glucose, and triglycerides were measured using an automated analyzer (Technicon DAX system). Blood pressure was measured with a calibrated automatic sphygmomanometer (Omron M3). Measurements were repeated three times with a pause of 1 min between measurements and the average value was recorded. To calculate physical activity practice, self-reported number of sessions of physical activity per week was obtained.

3.2.2. Determining cardiovascular risk variables

The presence of metabolic syndrome (MS) was ascertained by using the criterion suggested by the National Cholesterol Educational Program Adult Treatment Panel III (NCEP ATP III). The Framingham equation calibrated for the Spanish population (Framingham-REGICOR) was used to determine the cardiovascular risk at 10 years (software tool *calcumedplus*, available at http://www.fisterra.com). Classification of the participants in the study according to cardiovascular disease (CVD) risk was the Framingham-REGICOR guidelines:$>$ 10% High risk CVD, $5-9.9\%$ Moderate risk CVD, $<5\%$ Low risk CVD [43].

The heart age was calculated using the Heart Age Calculator, available at http://www.heartage.me. Cardiovascular lost years (CVLY) is defined as the difference between the heart age and the chronological age [44]. CVLY takes the values: First Quartile $[-20,-4]$, Second Quartile $[-3,3]$, Third Quartile $[4,12]$, and Fourth Quatile $[13,20]$.

With slight differences between them, the parameters required for calculating the Framingham-REGICOR score and the heart age are: age, sex, height (in centimeters), weight (in kilograms), waist circumference (in centimeters), familiar history of cardiovascular diseases, the presence or absence of diabetes, smoking habit, total cholesterol and HDL-cholesterol levels, and systolic pressure or antihypertensive treatment [45].

3.3. Learning Bayesian networks

To obtain a BN, it is necessary to determine a structure (defined by a DAG) and the conditional probabilities assigned to each node of the DAG. Therefore, to learn a BN implies two tasks: (i) *structural learning*, that is, the identification of the topology of the BN, and (ii) *parametric learning*, that is the estimation of numerical parameters (conditional probabilities) given a network topology.

3.3.1. Structural learning

The problem of discovering the causal structure increases with the number of variables [46, 47, 48]. Table 1 shows a description of the variables considered.

We are interested in obtaining a DAG, so only three possible connections are considered. The number of different structures, f(n), grows more than exponentially in the number of nodes, in [49] the following efficiently computable recursive function is given in Equation (2):

$$f(n)=\sum_{i=1}^{n}(-1)^{i+1}\frac{n!}{(n-1)!n!}2^{i(n-1)}f(n-1) \tag{2}$$

Table 1: Description of 13 data set features used to learn the structure.

Variable name	Description	Values
Gender	Male and Female	Men, Women
Age	Age in years	35-44, 45-54, 55-64
Smoking	Never smoker, Former smoker and Current smoker	Never smoker, Former smoker, Current smoker
PA	Physical Activity (three or more times/week during 1 hour)	No practice, Practice
BMI	Body Mass Index (kg/m^2)	Underweight, Normal weight, Overweight GI, Overweight GII, Obesity TI, Obesity TII, Obesity TIII
WC	Waist Circumference (cm)	High, Normal, Very High
BP	Blood Pressure (mmHg)	Normal, Optimal, Normal High Mild, Moderate, Serious
HDL	HDL-cholesterol (mg/dl)	Normal, Low, High
CVLY	Cardiovascular Lost Years	First Quartile, Second Quartile Third Quartile, Fourth Quartile
Glucose	Fashing blood glucose (mg/dl)	High, Normal
TG	Triglycerides (mg/dl)	Normal, Limit, Hyper
CVRS	Framingham-REGICOR Score	Low, Moderate, High
MetS	Metabolic Syndrome	Yes, No

There are two approaches to structure learning that could basically be considered [50]: (i) *search-and-score* structure learning, and (ii) *constraint-based* structure learning; combination of both gives a *hybrid* learning framework. Search-and-score search algorithms assigns a number (score) to each BN structure, and then the structure model with the highest score is chosen. Constraint-based search algorithms establish a set of conditional independence analysis on the data [51]. Using this analysis an undirected graph could be generated. Taking into account additional independence test, the network is transformed into a BN. *Hybrid algorithms* combine aspects of both constraint-based and score-based algorithms, they use conditional independence test to reduce the search space and network score to find the optimal network in the reduced space.

In order to obtain the DAG, we used the *bnlearn* package [52, 53] of R language [54]. As there are many structures that are consistent with the same set of independencies, prior knowledge of the system under study was taken into account in model selection process; to choose a structure that reflects the causal order and dependencies, that is those causes are parents of the effects, are considered structures that tend to work well [1], causal graphs tend to be sparser. Causality would be in the world, not in the inference process.

We included our prior knowledge of the system under study into the model selection process, thus variables were divided into four blocks: 1) *background variables* = {*Gender*, *Age*}, 2) *conditional variables* = {*Smoking*, *PA*}, 3) *intermediate variables* = {*BMI*, *TG*, *WC*, *HDL*, *BP*, *Glucose*}, and, 4) *diagnostic variables* = {*CVLY*, *CVRS*, *MetS*}. We restricted the model selection process by blacklisting arrows that point from a later to an earlier block [55]. To obtain the structure, two options either select a single *best* model or obtain some *average* model, which is known as *model averaging* [56]. Our model was learnt by *hill-climbing* (*hc*) algorithm. The final model was obtained repeating several times structure learning, a large number of network structures were explored (500 BNs) to reduce the impact of locally optimal (but globally suboptimal) networks on learning. The networks learned were averaged to obtain a more robust model. The averaged network structure was obtained using the arcs present in at least 85% of the networks, which gives a measure of the strength of each arc and establishes its *significance* given a *threshold* (85%) (see Figure 1).

3.3.2. Parametric learning

Parameters were obtained again with the *bnlearn* package in R language by performing a Bayesian parameter estimation using the Dirichlet distribution [57].

A conditional probability distribution is obtained for each node. In Table 2 an

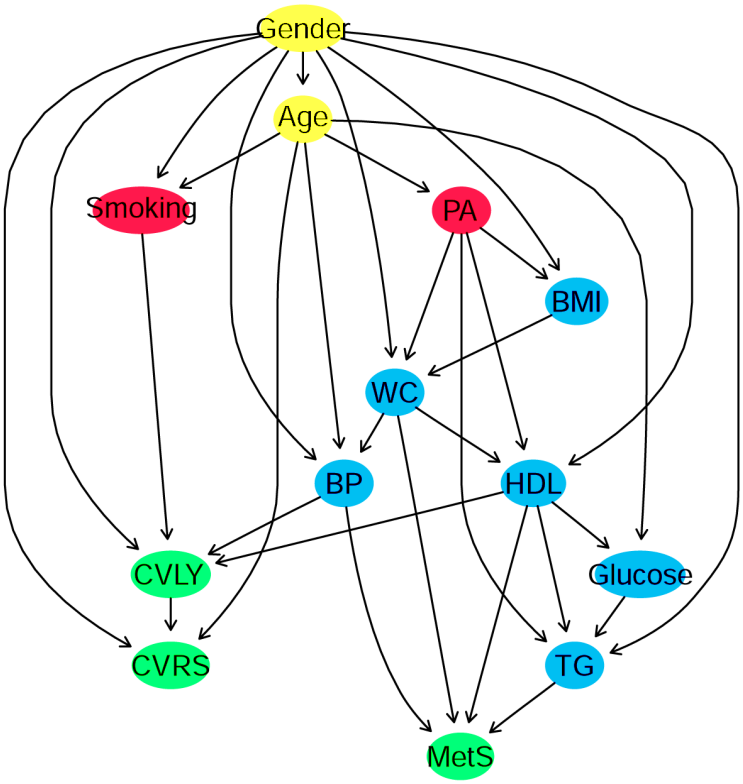

Figure 1: Structure obtained by *model averaging* over 500 networks. It was built with the hill climbing learning algorithm *hc* from *bnlearn* package in R language using a threshold = 0.85. In model selection process we included prior knowledge, thus variables were divided into four blocks: 1) *background variables* = {*Gender*, *Age*}, 2) *conditional variables* = {*Smoking*, *PA*}, 3) *intermediate variables* = {*BMI*, *TG*, *WC*, *HDL*, *BP*, *Glucose*}, and, 4) *diagnostic variables* = {*CVLY*, *CVRS*, *MetS*}.

Table 2: Expected values of probabilities for *Smoking* feature conditional on combinations of its parent values, in this case conditional on *Gender* and *Age* features.

Gender	Age	Smoking = *Former*	Smoking = *Current*	Smoking = *Never*
Men	35-44	0.0668	0.3636	0.5695
Men	45-54	0.0845	0.3825	0.5329
Men	55-64	0.1122	0.2852	0.6026
Women	35-44	0.1139	0.3231	0.5630
Women	45-54	0.1415	0.3371	0.5206
Women	55-64	0.1348	0.1311	0.7341

example of conditional probability distribution is shown.

3.4. Cardiovascular risk model

Although the *bnlearn* package in R allows us to make inference, in order to have a clear graphical representation from the structure and parameters obtained with *bnlearn* in R language the BN was implemented in Netica [58]. The compiled network is represented in Figure 2. The joint probability distribution of the BN in Figure 2 requires the specification of 13 conditional probability tables, one for each variable conditioned to its parents' set.

As we can observe in Figure 2, *CVLY* and *CVRS* variables have a direct connection, and both are connected to *MetS* variable through different trails, e.g., *MetS* variable is connected to *CVLY* variable through *BP* variable (*BP* is a common cause), once *BP* is instantiated the connection via this trail is broken), and *MetS* variable is also connected to *CVRS* variable through *TG* variable (it is also a common cause, once *TG* variable is instantiated the connection via this trail is broken), however there are other possible trails such as: *MetS* $\leftarrow$ *HDL* $\leftarrow$ *Gender* $\rightarrow$ *CVLY*, *MetS* $\leftarrow$ *WC* $\leftarrow$ *Gender* $\rightarrow$ *CVLY* $\rightarrow$ *CVRS*, etc.

The final BN obtained from the data set shows a *High* likelihood in *Low* value of *CVRS* variable, a *High* likelihood in *No* value of *MetS* variable, a *High* likelihood in *normal* value of *Glucose* variable, a *High* likelihood in *Normal* value of *TG* variable, a *high* likelihood in *Normal* value of *BP* variable, a *high* likelihood in *Normal* value of *HDL* variable, a *high* likelihood in *Normal weight* values of *BMI* variable, a *high* likelihood in *Normal* value of *WC* variable, a *high* likelihood in *Never* value of *Smoking* variable, and similar likelihoods in the different labels of *CVLY* and *PA* variables.

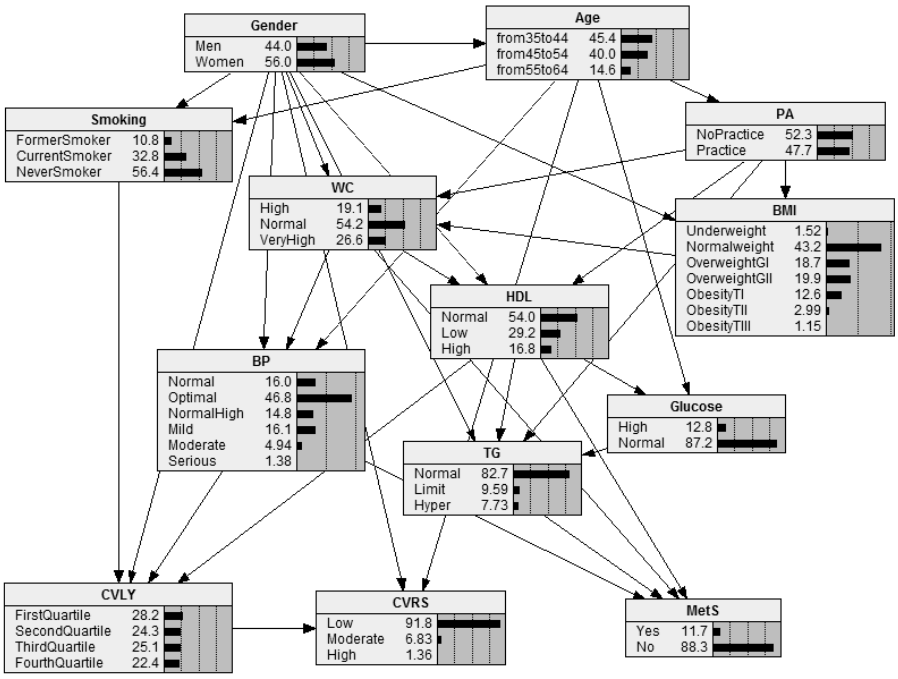

Figure 2: BN for the study of features relationships to evaluate *CVLY*, *CVRS* and *MetS* features. The BN shows an optimal (46.8%) blood pressure (*BP*), normal (82.7%) triglycerides (*TG*), normal (87.2%) *Glucose*, normal weight (43.2%) (*BMI*), and practice physical activity (*PA*) (47.7%) and no practice physical activity (PA) (52.3%). It also shows low levels of Framingham-REGICOR score (*CVRS*) (91.8%), no metabolic syndrome (MetS)(88.3%) and similar cardiovascular lost years (*CVLY*) in the four quartiles.

3.5. Validation of the BN

The BN was validated using a 10-fold cross-validation for BN, using a log-likelihood loss function, obtaining an expected loss of 9.3895. In Table 3, the area under the ROC curve (AUC), and the percentage correctly classified for the different features is shown.

3.6. Performance comparison

In order to provide reference benchmarks about how our BN classifies, we also report other classification performances (see Table 4) obtained by the widely used Naïve Bayes (NB), Tree Augmented Naïve Bayes (TAN), Multilayer Perceptron (MLP), and the C4.5 decision tree algorithm integrated in WEKA [59]. Only the diagnostic features (CVLY, CVRS, MetS) were considered as a compar-

Table 3: AUCs and percentage correctly classified for the different features.

Variable name	State	AUC	Accuracy
Gender	Men	0.9048	82.1938
Gender	Women	0.9047	82.1938
Age	35-44	0.6756	53.4435
Age	45-54	0.6088	53.4435
Age	55-64	0.7273	53.4435
Smoking	Former Smoker	0.6864	73.3534
Smoking	Current Smoker	0.8772	73.3534
Smoking	Never Smoker	0.8117	73.3534
PA	No Practice	0.8763	78.6126
PA	Practice	0.8773	78.6126
BMI	Underweight	0.8242	55.1966
BMI	Normal weight	0.8460	55.1966
BMI	Overweight GI	0.7110	55.1966
BMI	Overweight GII	0.7338	55.1966
BMI	Obesity TI	0.8654	55.1966
BMI	Obesity TII	0.8905	55.1966
BMI	Obesity TIII	0.8638	55.1966
WC	High	0.7487	73.0278
WC	Normal	0.8677	73.0278
WC	Very High	0.9150	73.0278
BP	Normal	0.7384	59.2787
BP	Optimal	0.8902	59.2787
BP	Normal High	0.7505	59.2787
BP	Mild	0.8453	59.2787
BP	Moderate	0.8805	59.2787
BP	Serious	0.9408	59.2787
HDL	Normal	0.7639	69.4465
HDL	Low	0.8762	69.4465
HDL	High	0.8806	69.4465
CVLY	First Quartile	0.9188	63.9600
CVLY	Second Quartile	0.7926	63.9600
CVLY	Third Quartile	0.8238	63.9600
CVLY	Fourth Quartile	0.9335	63.9600
Glucose	High	0.7274	87.1525
Glucose	Normal	0.7277	87.1525
TG	Normal	0.8523	84.5980
TG	Limit	0.7953	84.5980
TG	Hyper	0.8636	84.5980
CVRS	Low	0.8095	91.2597
CVRS	Moderate	0.8201	91.2597
CVRS	High	0.8067	91.2597
MetS	Yes	0.9836	96.4438
MetS	No	0.9835	96.4438

Table 4: Performance for CVLY, CVRS, and MetS features comparing our BN and using a 10-fold cross validation experiments with the corresponding algorithms.

Algorithms	CVLY	CVRS	MetS
	Accuracy		
Bayesian network	63.9600	91.2597	96.4438
Naïve Bayes	59.0033	90.4833	95.4921
Tree Augmented Naïve Bayes	63.8900	91.2580	96.0690
Multilayer Perceptron	61.9835	91.2596	96.2434
Trees C4.5	62.1337	91.2597	95.4420
	Sensitivity		
Bayesian network	0.6392	0.9131	0.9901
Naïve Bayes	0.5901	0.9050	0.9550
Tree Augmented Naïve Bayes	0.6389	0.9126	0.9900
Multilayer Perceptron	0.6200	0.9130	0.9620
Trees C4.5	0.6210	0.9130	0.9544
	Specificity		
Bayesian network	0.8785	0.2874	0.7967
Naïve Bayes	0.8610	0.2790	0.7920
Tree Augmented Naïve Bayes	0.8784	0.0874	0.7565
Multilayer Perceptron	0.8720	0.0870	0.7670
Trees C4.5	0.8740	0.0870	0.7460

ative example. Performance of each classification model is evaluated using three statistical measures: accuracy, sensitivity and specificity.

Learning a BN from data is a form of *unsupervised* learning, in the sense that the learner does not distinguish the class variable from the attribute variables in the data [60]. We compare our BN with several *supervised* learning algorithms: NB, TAN, MLP, and the C4.5 decision tree.

NB and TAN classifiers are special types of BN, where a supervised learn is performed. NB is a probabilistic graphical classifier based on Bayes theorem which uses very strong assumptions on the independence between the predictor variables. The NB model assumes that instances fall into one of a number of mutually exclusive classes, and it is the simplest BN classifier, where the predictive variables are assumed to be conditionally independent given the class. The performance of NB is surprising, since this assumption is unrealistic. The TAN classifier [60] extends the NB model with a tree-shape graph across the predictor variables.

TAN model is similar to NB except that each predictor variable is allowed to depend on other predictor variable in addition to the class. This model provides more information than the NB model as it is included information about the relationship among all predictor variables. MLP is a feedforward artificial neural network model which consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. C4.5 algorithm is a decision tree induction method develop by Quinlan [61].

The major advantage of BN is the ability to represent and hence understand knowledge. Our BN model gives the best classification performances. Furthermore their graphical representation is very informative.

4. Reasoning patterns

BNs are used to calculate new probabilities when new information is obtained [8]. Given the evidence **E** = e, our goal is to find the most likely assignment to the variables in $\mathbf{U} = complementary(\mathbf{E})$, see Equation (3):

$$MAP(\mathbf{U} \mid e) = argmax_u P(u, e) \tag{3}$$

There are two main types of queries: 1) in a *probability query*, we try to find the most likely assignment to a single variable, i.e. to compute $P(\mathbf{X} \mid e)$; 2) in a MAP query, we find the most likely joint assignment to the variables in **U**. In order to introduce evidence in the network we have selected three reasoning patterns: *causal* reasoning, *evidential* reasoning, and *intercausal* reasoning.

4.1. Causal reasoning

Causal reasoning takes place when we predict effects from causes (an so we proceed from top to bottom). We instantiate one variable at each a single step. In step 1 *Gender* variable is instantiated either to *Men* or *Women*, in step 2 *Smoking* variable is instantiated to *Current Smoker* or *Never Smoker*, in step 3 physical activity (*PA*) variable is instantiated to *Practice* or *No Practice*, in step 4 *Age* variable is instantiated to *35-44*, in step 5 *Age* variable is instantiated to *45-54*, and in step 6 *Age* variable is instantiated to *55-64*.

4.1.1. Analysis of cardiovascular lost years CVLY variable.

In Figure 3, a summary about how the different quartiles of *CVLY* variable changes at each step is shown. Taking into account the conditional variables (Smoking and physical activity) the one with greatest influence on cardiovascular lost years *CVLY* is the smoking habit, obtaining two clear patterns: 1)

When *Smoking* is in the *Never* state, Figure 3 shows that the highest probability is achieved for first quartile *Women* follow by the second and third quartiles in *Men*. Adding physical activity *PA* variable in the *Practice* state shows a decrease in the probability for fourth quartile in Men and Women, showing slower values in Women; and also, an increase in the probability for first quartile in Women and for second quartile in Men; and 2) When *Smoking* is in the *Current* state, Figure 3 shows that the highest probability is achieved for fourth quartile *Men* followed by the third quartile *Women*. Figure 3 also shows an improvement of the situation when physical activity is instantiated to *Practice* and if the group of the youngest population is considered, being the group of Men with the highest risk.

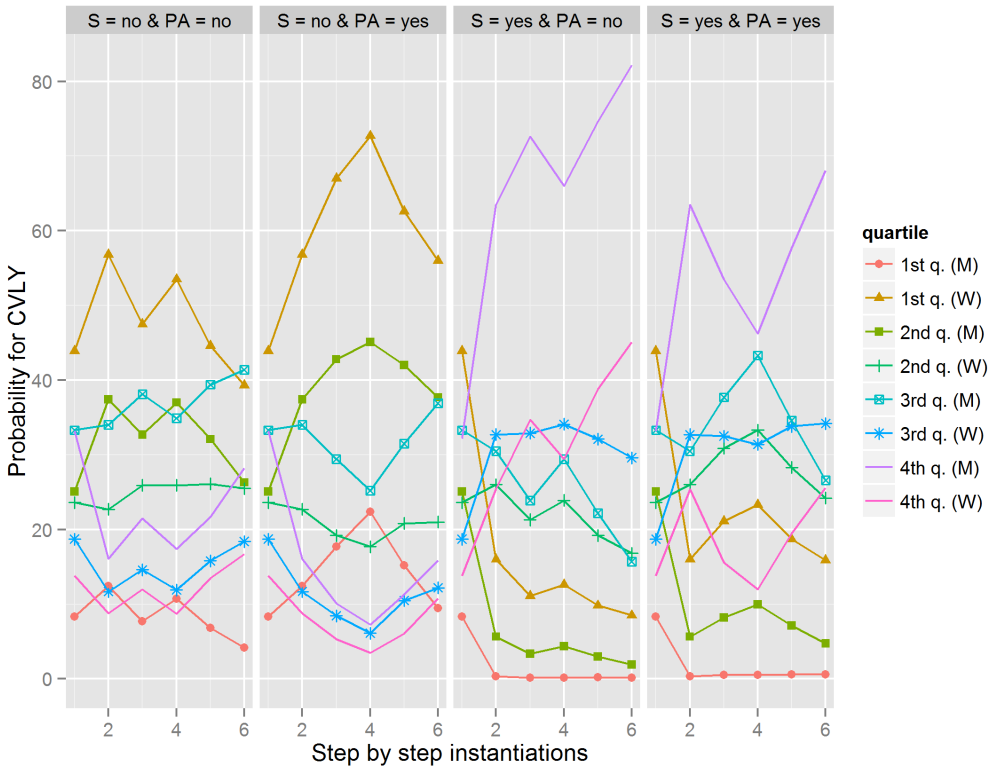

Figure 3: Step by step instantiations. The different steps: step 1 = Gender, step 2 = Smoking, step 3 = PA, step 4 = Age = 35-44, step 5 = Age = 45-50, and, step 6 = Age = 55-64 to evaluate *CVLY*. Where S = Smoking, and PA = Physical Activity. The different steps are represented in the horizontal axis. The estimated probability for *CVLY* variable expressed as a percentage at the different quartiles is showed in the vertical axis: M: Men, and W: Women.

4.1.2. Studying metabolic syndrome MetS.

From Figure 4, we can differentiate two patterns taking into account whether the subjects practice physical activity or not (*PA* variable). When physical activity (*PA* variable) is instantiated to *Practice* we obtain the highest probability in the *No* state for Metabolic Syndrome (*MetS* variable), showing that Smoking variable does not have any influence and the group of Women were the most privileged (with the highest probability for Mets variable in the *No* state). However, when Physical Activity (*PA* variable) is instantiated to *No Practice* we observe that for Metabolic Syndrome (*MetS* variable) in the *Yes* state the probability increases, showing that the *Smoking* variable does not have any influence again. The group with the highest risk of getting MetS = Yes is the group of Men.

4.1.3. Studying cardiovascular risk score CVRS.

From Figure 5 when physical activity is instantiated to *Practice* we obtain similar probabilities for *CVRS* variable independently of whether the subject smokes or not. Similarly when physical activity is instantiated to *No Practice*.

4.2. Evidential reasoning

Queries, where we reason from effects to cases (from bottom to up), are instances of *evidential reasoning* or *explanation*.

Mets and *CVRS* variables are instantiated to values *Yes* and *High* respectively. We observe how the probability of the different variables changes. *CVLY* variable increases its Fourth quartile value from 22.4% to 90%. *BP* variable increases its *Mild* value from 16.1% to 44.5%, and decreases its *Optimal* value from 46.8% to 3.28%. *TG* variable achieve similar likelihoods for all its values: *Normal*, *Limit* and *Hyper*. *HDL* variable increases its *High* value from 16.8% to 69.1%. *WC* variable increases its *Very High* value from 26.6% to 68.8%. *BMI* variable decreases its *Normal Weight* value from 43.2% to 11.1%, and increases the probability of *Overweight GII* from 19.9% to 28.4%, and the probability of *Obesity T1* from 12.6% to 33.7%. The *PA* variable increases the probability of the No Practice value from 52.3% to 89.7%. The *Glucose* variable increases its High value from 12.8% to 31.5%. The *Gender* variable increases its *Men* value from 44.0% to 78.5%. The *Age* variable increases its 55-64 value from 14.6% to 50.7%. Figure 6 shows the probability variations.

4.3. Conditional Entropy

In Shannon [62] theory, entropy of X is the lower bound on the average number of bits that are needed to encode values of X. Another way of viewing the entropy

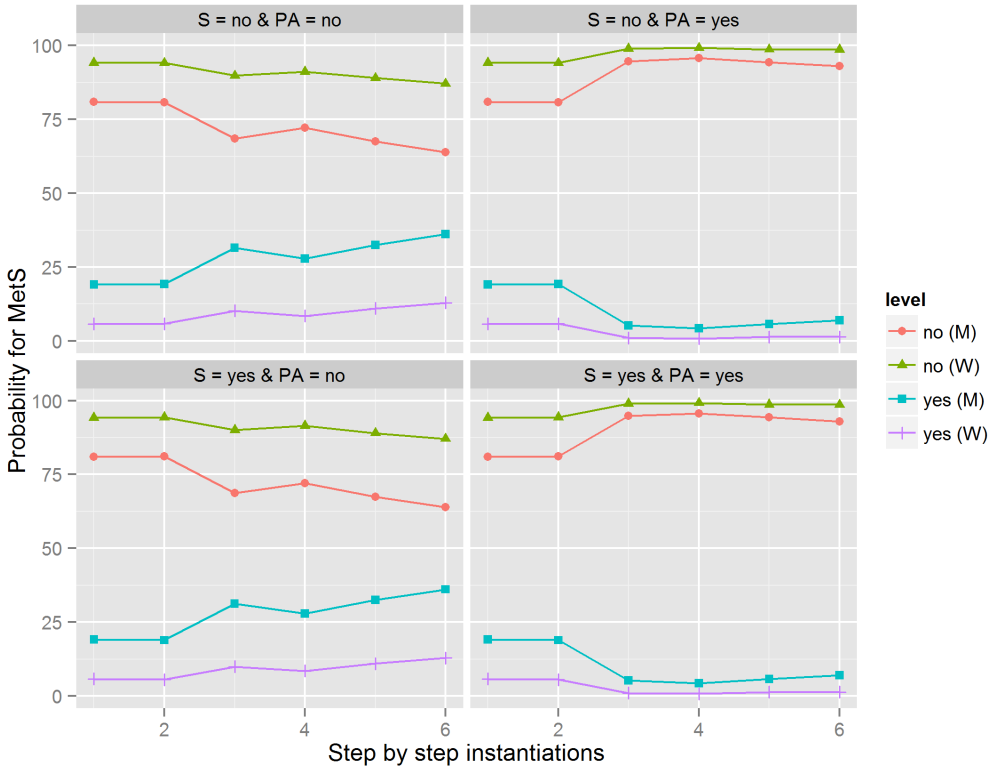

Figure 4: Step by step instantiations. The different steps: step 1 = Gender, step 2 = Smoking, step 3 = PA, step 4 = (Age = 35-44), step 5 = (Age = 45-50), and, step 5 = (Age = 55-64) to evaluate *MetS* feature. Where S = Smoking, and PA = *PA* = No Practice. The different steps are represented in the horizontal axis. The estimated probability for *MetS* variable expressed as a percentage at the different values (*yes*, *no*) is showed in the vertical axis: M: Men, and W: Women.

is as a measure of our uncertainty about the value of X, i.e., little uncertainty about X will produce a low entropy value.

A natural question is what is the cost of encoding X if we are already encoding Y. The conditional entropy of X given Y is

$$H_P(X|Y) = E_P\left[\log \frac{1}{P(X|Y)}\right] = \sum P(X|Y) \cdot \log \frac{1}{P(X|Y)} \tag{4}$$

which captures the additional cost (in terms of bits) of encoding X when we are already encoding Y. Note that the maximum value of probability in $P(X|Y)$ implies the lowest entropy value.

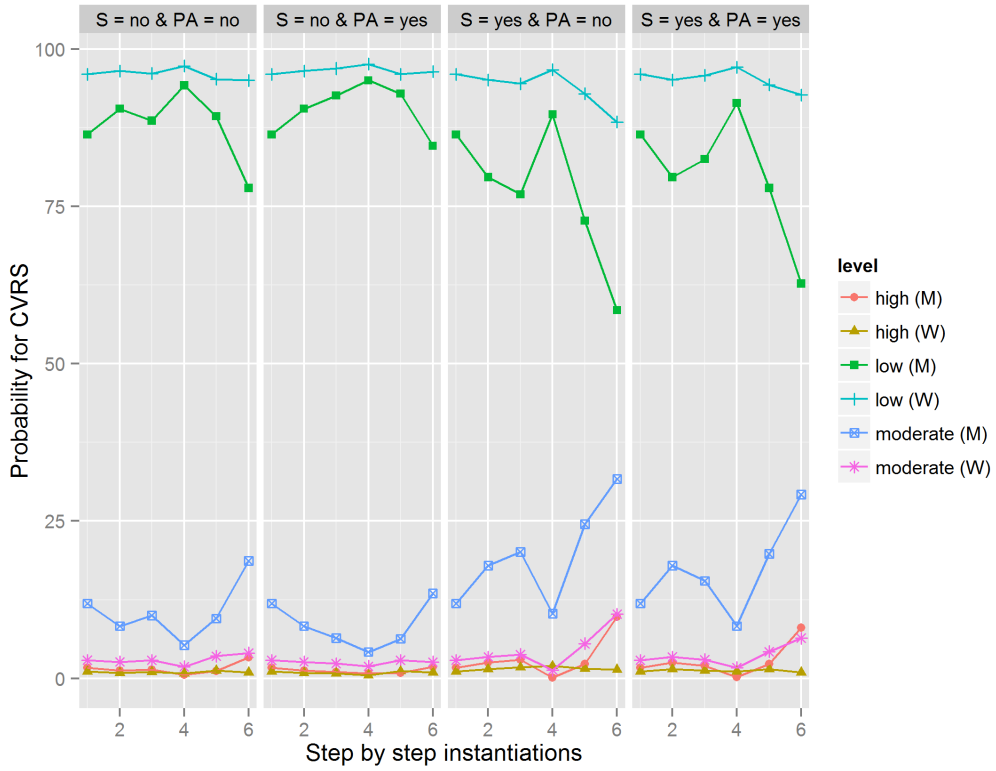

Figure 5: Step by step instantiations to evaluate *CVRS* feature. In step 1 = Gender, step 2 = Smoking, step 3 = Physical Activity, step 4 = Age = 35-44, step 5 = Age = 45-50, and, Age = 55-64

For *MetS*, *CVLY*, and *CVRS* features we are interested in determining and ordering the state values for conditioned features such as we obtain the maximum probability value in some states, which will lead to achieve the minimum conditioned entropy.

4.4. Intercausal reasoning

When different causes of the same effect can interact we talk of *intercausal reasoning*, which constitutes a very common pattern in human reasoning.

Furthermore, BNs are able to produce probability estimates, in this sense we are interested in knowing the features with highest influence in maximizing *MetS*, *CVLY*, and *CVRS* in some of their states.

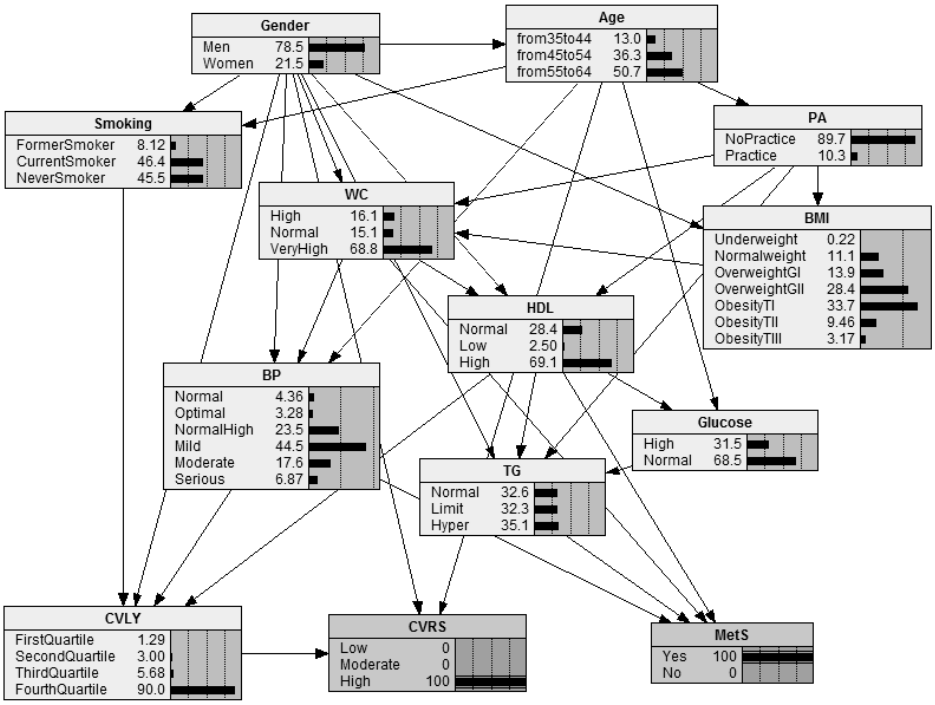

Figure 6: Evidential reasoning. Metabolic syndrome *MetS* variable is instantiated to *No* value and *CVRS* variable is instantiated to *High* value.

4.4.1. Minimizing conditioned entropy for MetS

We maximize *MetS* feature probability in a *Yes* state. To achieve it, we consider the Markov blanket of *MetS* variable, it is composed of the four following variables: *WC*, *HDL*, *BP* and *TG*. We choose from each step the variable and the state that induces the greatest increase in the conditional probability of *MetS* variable in a *Yes* state. A summary is shown in Table 5 and Figure 7. Given the Markov blanket of the *MetS* variable, the global Markov property states that the *MetS* variable is conditionally independent of any other variable.

Again, we maximize the MetS variable probability in a *No* state. We choose at each state the variable and the state that most increases the probability of the MetS variable in a *No* state. A summary is shown in Table 6 and Figure 8.

4.4.2. Minimizing conditioned entropy for CVLY

We maximize *CVLY* feature probability in *First Quartile* state. To achieve it, we consider the Markov blanket of *CVLY* variable, it is composed of the six

Table 5: Step-by-step instantiations leading to maximization of the probability of the MetS variable, where in the initial BN without evidence *MetS = Yes* reached a probability of 11.7%. The different values: *Serious*, *Moderate*, *Mild* and *NormalH* for *BP* variable gave the same probability value for the *MetS* variable.

Step	instantiated variable		value	MetS = Yes
1	*TG*	=	Hyper	48.7%
2	*WC*	=	Very High	85.4%
3	*HDL*	=	Low	100%
4	*BP*	=	NormalH	100%
4'	*BP*	=	Mild	100%
4"	*BP*	=	Moderate	100%
4"'	*BP*	=	Serious	100%

Table 6: Step-by-step instantiations leading to maximization of the probability of the *MetS* variable, where in the initial BN without evidence *MetS = No* reached a probability of 88.3%. The maximum probability for *MetS* feature in state NO is achieved when *BP* = Normal, *WC* = Normal and *TG* = Normal.

Step	instantiated variable		value	MetS = No
1	*BP*	=	Normal	95.4%
2	*WC*	=	Normal	99.1%
3	*TG*	=	Normal	100%
4	*HDL*	=	Normal	100%
4'	*HDL*	=	Low	100%
4"	*HDL*	=	High	100%

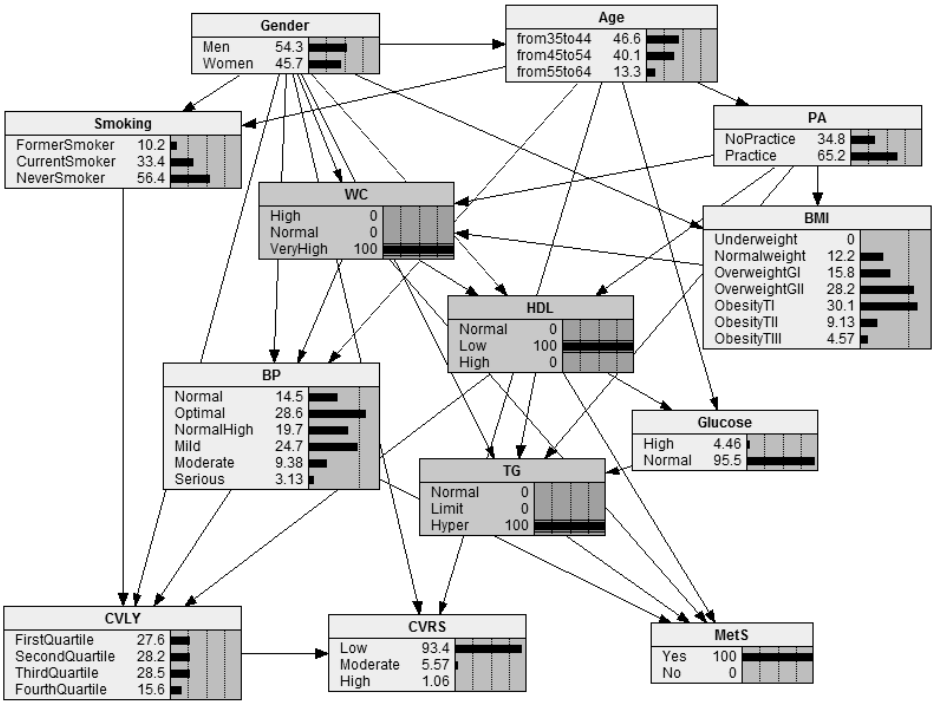

Figure 7: Intercausal reasoning: maximizing *MetS* feature in the *Yes* state. We try to obtain the highest probability for *MetS* = *Yes* after introducing the following evidence: *TG* = *Hyper*, *WC* = *Very High*, *HDL* = *Low*.

following variables: *CVRS*, *Gender*, *Age*, *Smoking*, *BP*, and *HDL*. We choose from each step the variable and the state that induces the greatest increase in the conditional probability of *CVLY* variable in *First Quartile* state. *Age* feature has not been included, because it does not increase the probability of *CVLY* in *First Quartile* once *BP*, *HDL*, *Smoking*, *Gender* and *CVRS* features are instantiated. A summary is shown in Table 7. Given the Markov blanket of the *CVLY* variable, the global Markov property states that the *CVLY* variable is conditionally independent of any other variable.

Again, we maximize *CVLY* feature probability in a *Second Quartile* state. The order of features is: *Smoking*, *BP*, *HDL*, *Gender*, *CVRS*, and *Age*. Achieving a maximum probability value of 68% for *CVLY* feature in *Second Quartile* value. A summary is shown in Table 8.

Again, we maximize *CVLY* feature probability in a *Third Quartile* state. The order of features is: *BP*, *HDL*, *Smoking*, *Gender*, *CVRS*, and *Age*. Achieving a

Table 7: Step-by-step instantiations leading to maximization of the probability of the CVLY variable, where in the initial BN without evidence *CVLY = First Quartile* reached a probability of 28.2%.

Step	instantiated variable		value	CVLY = First Quartile
1	*BP*	=	Optimal	51.2%
2	*HDL*	=	Low	70.7%
3	*Smoking*	=	Never Smoker	90.8%
4	*Gender*	=	Women	91.6%
5	*CVRS*	=	Low	91.7%

Table 8: Step-by-step instantiations leading to maximization of the probability of the CVLY variable, where in the initial BN without evidence *CVLY = Second Quartile* reached a probability of 24.3%.

Step	instantiated variable		value	CVLY = Second Quartile
1	*Smoking*	=	Never Smoker	29.2%
2	*BP*	=	Normal	43.3%
3	*HDL*	=	Normal	58.3%
4	*Gender*	=	Men	64.2%
5	*CVRS*	=	Low	65.1%
6	*Age*	=	55-64	68.0%

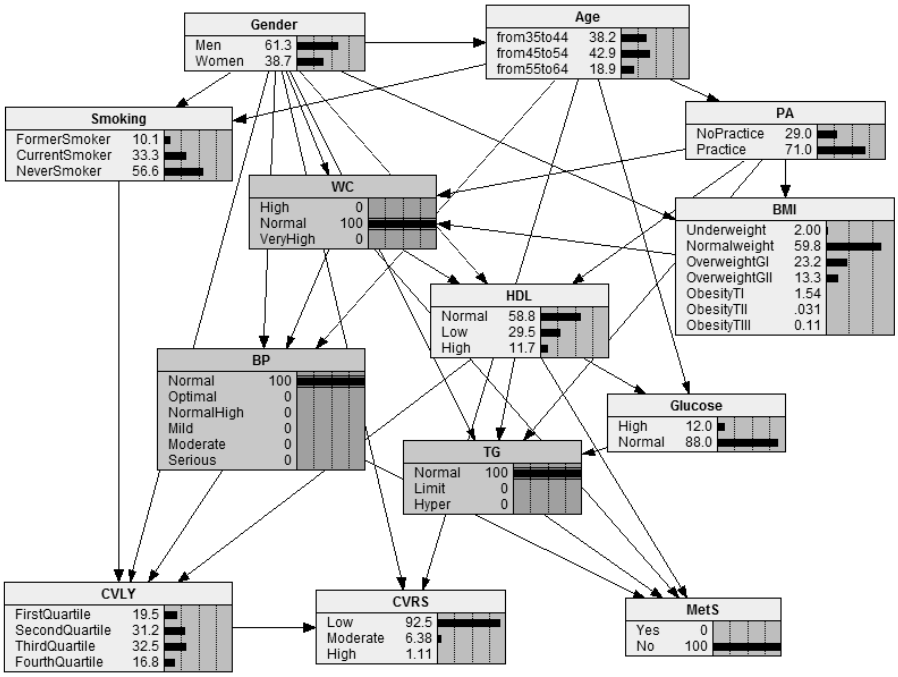

Figure 8: Intercausal reasoning: maximizing *MetS* feature in the *No* state. We try to obtain the highest probability for *MetS* = *No* after introducing the following evidence: *BP* = *Normal*, *TG* = *Normal*, *WC* = *Normal*.

maximum probability value of 79% for *CVLY* feature in *Third Quartile* value. A summary is shown in Table 9.

Finally, we maximize *CVLY* feature probability in a *Fourth Quartile* state. The order of features is: *BP*, and *Smoking*. Achieving a maximum probability value of 100% for *CVLY* feature in *Fourth Quartile* value. A summary is shown in Table 10.

4.4.3. Minimizing conditioned entropy for CVRS

We maximize *CVRS* feature probability in a *Low* state. To achieve it, we consider the Markov blanket of *CVRS* variable, it is composed of the three following variables: *CVLY*, *Age*, and *Gender*. We choose from each step the variable and the state that induces the greatest increase in the conditional probability of *CVRS* variable in a *Low* state. The order of features is: *CVLY*, *Age* and, *Gender*. Achieving a maximum probability value of 100% for *CRVS* feature in *Low* value. A summary is shown in Table 11. Given the Markov blanket of the *CVLY* variable, the global

Table 9: Step-by-step instantiations leading to maximization of the probability of the CVLY variable, where in the initial BN without evidence *CVLY = Third Quartile* reached a probability of 25.1%.

Step	instantiated variable		value	CVLY = Third Quartile
1	*BP*	=	Normal	34.1%
2	*HDL*	=	High	49.2%
3	*Smoking*	=	Never Smoker	70.1%
4	*Gender*	=	Men	75.5%
5	*CVRS*	=	Low	77.6%
6	*Age*	=	45-54	79.0%

Table 10: Step-by-step instantiations leading to maximization of the probability of the CVLY variable, where in the initial BN without evidence *CVLY = Fourth Quartile* reached a probability of 22.4%.

Step	instantiated variable		value	CVLY = Fourth Quartile
1	*BP*	=	Serious	79.6%
2	*Smoking*	=	Current Smoker	100%

Markov property states that the *CVRS* variable is conditionally independent of any other variable.

Again, we maximize *CVRS* feature probability in a *Moderate* state. The order of features is the same that the case before: *CVLY*, *Age* and, *Gender*. Achieving a maximum probability value of 34.4% for *CRVS* feature in *Moderate* value. A summary is shown in Table 12.

Finally, we maximize *CVRS* feature probability in a *Low* state. The order of features is the same that the cases before: *CVLY*, *Age* and, *Gender*. Achieving a maximum probability value of 11.9% for *CRVS* feature in *High* value. A summary is shown in Table 13.

Table 11: Step-by-step instantiations leading to maximization of the probability of the *CVRS* variable in *Low* state, where in the initial BN without evidence *CVRS = Low* reached a probability of 91.8%. The maximum conditioned probability for *CVRS* feature in state *Low* is achieved when *CVLY* = First Quartile, *Age* = 55-64 and *Gender* = Men.

Step	instantiated variable		value	CVRS = Low
1	*CVLY*	=	First Quartile	97.5%
2	*Age*	=	55-64	99.0%
3	*Gender*	=	Men	100%

Table 12: Step-by-step instantiations leading to maximization of the probability of the *CVRS* variable in *Moderate* state, where in the initial BN without evidence *CVRS* = *Moderate* reached a probability of 6.83%. The maximum conditioned probability for *CVRS* feature in state *Moderate* is achieved when *CVLY* = Fourth Quartile, *Age* = 55-64 and *Gender* = Men.

Step	instantiated variable		value	CVRS = Moderate
1	*CVLY*	=	Fourth Quartile	19.4%
2	*Age*	=	55-64	30.5%
3	*Gender*	=	Men	34.4%

Table 13: Step-by-step instantiations leading to maximization of the probability of the *CVRS* variable in *High* state, where in the initial BN without evidence *CVRS* = *High* reached a probability of 1.36%. The maximum conditioned probability for *CVRS* feature in state *High* is achieved when *CVLY* = Fourth Quartile, *Age* = 55-64 and *Gender* = Men.

Step	instantiated variable		value	CVRS = High
1	*CVLY*	=	Fourth Quartile	3.77%
2	*Age*	=	55-64	9.41%
3	*Gender*	=	Men	11.9%

5. Discussion

This study demonstrates the feasibility of BNs in epidemiological studies, particularly when data from cardiovascular risk factors is considered. BNs can be used for answering clinical questions based on unobserved evidence since the probability distributions can be automatically updated when new patient information is added in an appealing way.

The BNs allow us to establish the relationships between features through the relationships of dependency and conditional independency. Given *Gender*, *Age*, *BP* and *HDL* then *CVLY* and *CVRS* features are d-separated of *MetS* feature, any active trail connecting them was found. However, considering the local Markov property of a node, e.g., given the parents of *CVLY* feature, which is composed of *Gender*, *Smoking*, *BP* (blood pressure) and *HDL* (cholesterol) and taking into account the local Markov condition *CVLY* feature remains independent of its non-descendants, *CVLY* feature is independent of all other variables except of CVRS feature, in particular independent of the *MetS* feature. Similarly, given *BP*, *WC*, *HDL*, *TG* features, then *MetS* feature is independent of the remaining features; in this case, as the MetS feature does not have any descendants, *BP*, *WC*, *HDL*, *TG* features constitute its Markov blanket, and the global Markov property states that the *MetS* feature is conditionally independent of any other feature given its

Markov blanket.

Given the structure of a BN, the use of the global Markov property on each feature allows us to establish the set of features (which will be constituted by the Markov blanket of this specific feature) with the strongest influence on that feature; furthermore, the Markov blanket of a particular feature (node in the DAG) can be used to find the combination of the different states that allow to maximize or minimize a particular state of such feature. In this study we focus mainly on *CVLY*, *CVRS*, and *MetS* features. However, using the BNs a characterization of the whole set of variables could be given; e.g., the Markov blanket for *BMI* feature is given by physical activity (*PA*), gender (*Gender*), and waist circumference (*WC*), given these three features, *BMI* feature is independent of the remaining ones; furthermore it could be used to find the combination of states which maximize or minimize a specific state of *BMI* feature.

In our BN model Gender and BMI are connected. An association, or a link, between gender and BMI has been widely shown in the literature. However, reasons for BMI gender difference are unclear. Differences in anatomic, physiologic, metabolic and sex hormonal status between genders could contribute to these differences. In [63] and [64] from a Swedish and Canadian population respectively Gender and BMI appear related. In [65] from a dataset for epidemiological research of Korean population the authors build a BN for predicting metabolic syndrome, Gender appears completely isolated, it is neither related to BMI nor related to any other variable (WC, Age, HDL, Cholesterol, etc.).

The BN model included BMI and WC features. The most commonly method used for classifying an individual as overweight or obese is the body mass index (BMI). The BMI is defined as the body mass divided by the square of the body height, and is universally expressed in units of kg/m2, resulting from mass in kilograms and height in metres. However, the BMI has limitations and can lead to the misclassification of certain individuals such as those with increased muscle mass or the elderly. Waist circumference (WC) may be a better indicator of health risk than BMI alone, especially when used in combination with BMI. WC is particularly useful for individuals with a BMI of 25-34. For individuals with a BMI less than 35, WC adds little predictive power on the disease risk classification of BMI. Results obtained in recent studies reported that correlations between WC, waist-to-hip ratio (WHR) and waist-to-height ratio (WHtR) and cardiovascular risk factors are better than BMI (see for instance [66, 67]).

Reasons for the sex difference in CVRS are not fully understood. Differences in major cardiovascular risk factors, particularly in HDL cholesterol level, obesity and smoking rate, explained a substantial part of the sex difference in cardiovas-

cular risk [68, 69].

The main difference with respect other cardiovascular risk studies in the literature [65, 70, 71, 72, 73, 74] is that we include three diagnostic features: CVLY, CVRS, and MetS. This fact helps to determine those features with the greatest influence in each of the diagnostic features.

In summary, BNs are a graph-based structure of a joint multivariate probability distribution which capture the way an expert establishes the relationships between variables. Furthermore, BNs are a powerful tool for modeling the decision-making process under uncertainty, which combine a qualitative and quantitative representation at the same time. Due to similar knowledge pattern, a BN network (a modeling tool) can serve as an informal basis for development of a framework of Decision Support System (DSS) in the form of tabular rule-based system [75] for medical recommendations (a DSS tool).

6. Conclusions

BNs have been chosen in order to produce an intuitive, transparent, graphical representation of the investigated interdependencies. The obtained model helps us to easily identify the relationships of *probabilistic causal dependencies* and conditional *independencies* between features. As a result, we can then visualize the relationships between 13 features in the domain of cardiovascular risk. In this case, due to CVD is multifactorial, the application of this kind of networks is of special interest, both from theoretical and practical point of view.

Furthermore, the implemented BN was used to make inferences i.e., to predict new scenarios when hypothetical information was introduced. Adding evidence like different CVRF values in the implemented BN may be of great interest in epidemiological studies . To make a BN analysis three reasoning patterns were considered: causal, evidential and intercausal reasoning. Combining the reasoning patterns together with local and global Markov properties and the concept of Markov blanket some features were optimized.

Acknowledgements

This research was funded by the Spanish Ministry of Science and Innovation (PI13/01477).

References

[1] D. Koller, N. Friedman, Probabilistic graphical models. Principles and techniques, The MIT Press, Cambridge, Massachusetts, London, England 2010.

[2] J. Pearl, Causality. Models, reasoning and inference, Cambridge university press, Cambridge, 2000.

[3] P. Larrañaga, S. Moral, Probabilistic graphical models in artificial intelligence, Appl Soft Comput 11 (2011) 1511-1528.

[4] A. Ligęza, P. Fuster-Parra, AND/OR/NOT causal graphs – A model for diagnostic reasoning. Int. J. Appl Math Comput Sci 7 (1997), 185-203.

[5] G.F. Cooper, E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, Mach Learn 9 (1992) 309-347.

[6] D. Heckerman, D. Geiger, D.M. Chickering, Learning Bayesian networks: the combination of knowledge and statistical data, Mach Learn 20 (1995) 197-243.

[7] F Liang, J Zhang, Learning Bayesian networks for discrete data, Comput Stat Data Anal 53 (2009) 865-876.

[8] C.J. Butz, S. Hua, J. Chen, H. Yao, A simple graphical approach for understanding probabilistic inference in Bayesian networks. Inform Sci 179 (2009) 699-716.

[9] C. Glymour, R. Scheines, P. Spirtes, K. Kelly, Discovering causal structure, Technical report CMU-PHIL-1, 1986.

[10] P. Spirtes, C. Glymour, R. Scheines, Causation, prediction and search, Adaptive Computation and machine learning (2nd ed.), The MIT Press, 2001.

[11] P. Fuster-Parra, A. García-Mas, F.J. Ponseti, P. Palou, J. Cruz, A Bayesian network to discover relationships between negative features in sport: a case study of teen players, Qual Quant 48 (2014) 1473-1491. DOI: 10.1007/s11135-013-9848-y.

[12] P. Fuster-Parra P, A. García-Mas, F.J. Ponseti, F.M. Leo, Team performance and collective efficacy in the dynamic psychology of competitive team: A Bayesian network analysis, Hum Mov Sci 40 (2015) 98-118. DOI:http://dx.doi.org/10.1016/j.humov.2014.12.005.

[13] J. DeFelipe, P.L. López-Cruz, R. Benavides-Piccione, C. Bielza, P. Larrañaga et al., New insights into the classification and nomenclature of cortical GABAergic interneurons, Nature Review Neuroscience 14 (2013) 202-216.

[14] M.B. Sesen, A.E. Nicholson, R. Banares-Alcantara, T. Kadir, M. Brady, Bayesian networks for clinical decision support in Lung Cancer Care, Plos One 8 (2013) e82349. DOI: 10.1371/journal.pone.0082349.

[15] A. Djebbari, J. Quackenbush, Seeded Bayesian networks: constructing genetic networks from microarray data, BMC Syst Biol (2008) 2-57, DOI:10.1186/1752-0509-2-57.

[16] C.J. Needham, J.R. Bradford, A.J. Bulpitt, et al., A primer on learning in Bayesian networks for computational biology, PLoS Comput Biol 3 (2007), (doi:10.1371/journal.pcbi.0030129).

[17] S.J. Lycett, M.J. Ward, F.I. Lewis, et al., Detection of mammalian virulence determinants in highly pathogenic avian influenza H5N1 viruses: multivariate analysis of published data, J Virol 83(19) (2009) 9901-9910.

[18] A.F. Poon, F.I. Lewis, S.L. Pond, et al., Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope, PLoS Comput Biol 3(1) (2007), DOI:10.1371/journal.pcbi.0030011.

[19] R. Jansen, H. Yu, D. Greenbaum, et al., A Bayesian networks approach for predicting protein-protein interactions from genomic data, Science 302(5644) (2003) 449-453.

[20] F.I. Lewis, F. Brälisauer, G.J. Gunn, Structure discovery in Bayesian networks: an analytical tool for analysing complex animal health data, Prev Veterin Med 100(2) (2011) 109-115.

[21] F.I. Lewis, B.J. McCormick, Revealing the Complexity of Health Determinants in Resource-poor Settings, Am J Epidemiol 176(11) (2012) 1051-1059.

[22] M. Lappenschaar, A. Hommerson, P.J.F. Lucas, J. Lagro, S. Visscher, Multilevel Bayesian networks for the analysis of hierarchical health care data, Artif Intell Med 57 (2013) 171-183.

[23] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, B.D. Moor, Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection, Artif Intell Med 29 (2003) 29-60.

[24] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, B.D. Moor, Using literature and data to learn Bayesian networks as clinical models of ovarian tumors, Artif Intell Med 30 (2004) 257-281.

[25] T. Charitos, L.C. Gaag, S. Visscher, K.A.M. Schurink, P.J.F. Lucas, A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients, Expert Systems with Applications 36 (2009) 1249-1258.

[26] S.M. Maskery, H. Hu, J. Hooke, C.D. Shriver, M.N. Liebman, A Bayesian derived network of breast pathology co-occurrence, J Biomed Inform 41 (2008) 242-250.

[27] X.H. Wang, B. Zheng, W.F. Good, J.L. King, Y.H. Chang, Computer assisted diagnosis of breast cancer using a data-driven Bayesian belief network, Int J Med Inform 54 (1999) 115-126.

[28] J.J. Cabre, F. Martin, B. Costa, J.L. Pinol, J.L. Llor, Y. Ortega et al., Metabolic syndrome as a cardiovascular disease risk factor: Patients evaluated in primary care, BMC Public Health 8 (2008) 251. DOI:10.1186/1471-2458-8-251

[29] S.M. Grundy, J.I. Cleeman, S.R. Daniels, K.A. Donato, R.H. Eckel, B.A. Franklin, D.J. Gordon, R.M. Krauss, P.J. Savage, S.C. Smith Jr, J.A. Spertus, F. Costa, Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement, Circulation 112 (2005) 2735-2752.

[30] J.G. Lee, S. Lee, Y.J. Kim, H.K. Jin, B.M. Cho, Y.J. Kim et al., Multiple biomarkers and their relative contributions to identifying metabolic syndrome, Clin Chim Acta 408 (2009) 50-55.

[31] P. Tauler, M. Bennasar-Veny, J.M. Morales-Asencio, A.A. Lopez-Gonzalez, T. Vicente-Herrero, J. De Pedro-Gomez, V. Royo, J. Pericas-Beltran, A. Aguilo, Prevalance of Premorbid metabolic syndrome in Spanish adult workers using IDF and ATPIII diagnostic criteria: relationships with cardiovascular risk factors, PLoS One 9(2) (2014). DOI: 10.1371/journal.pone.0089281. eCollection.

[32] B. Van Steenkiste, T. Van der Weijden, H.E. Stoffers, A.D. Kester, D.R. Timmermans, R. Grol, Improving cardiovascular risk management: a randomized, controlled trial on the effect of a decision support tool for patients and physicians, Eur J Cardiov Prev Rehabil 14(1) (2007) 44-50.

[33] P.D. Sorlie, D.E. Bild, M.S. Lauer, Cardiovascular epidemiology in a changing world-challenges to investigators and the National Heart, Lung, and Blood Institute, Am J Epidem 175(7) (2012) 597-601.

[34] M. Franco, U. Bilal, E. Guallar, G. Sanz, A.F. Gómez, V. Fuster, R. Cooper, Systematic review of three decades of Spanish cardiovascular epidemiology: improving translation for a future of prevention, Eur J Prev Cardiol (2012). DOI: 10.1177/2047487312455314.

[35] J. Marrugat, R. Elosua, H. Marti, Epidemiology of ischaemic heart disease in Spain: estimation of the number of cases and trends from 1997 to 2005, Rev Esp Cardiol 55(4) (2002) 337-346.

[36] A. Willis, M. Davies, T. Yates, K. Khunti, Primary prevention of cardiovascular disease using validated risk scores: a systematic review, J Roy Soc Med 105(8) (2012) 348-356.

[37] F.H. Zimmerman, Cardiovascular disease and risk factors in law enforcement personnel: a comprehensive review, Cardiol Rev 20(4) (2012) 159-166.

[38] R.B. D'Agostino, R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, W.B. Kannel, General cardiovascular risk profile for use in primary care: the Framingham heart study, Circulation, 117(6) (2008) 743-753.

[39] World Health Organization, Obesity: Preventing and managing the global epidemic, WHO, Geneva, 1998.

[40] A.A. Lopez-Gonzalez, A. Aguilo, M. Frontera, M. Bennasar-Veny, I. Campos, T. Vicente-Herrero, M. Tomas-Salva, J. De Pedro-Gomez, P. Tauler, Effectiveness of the Heart Age tool for improving modifiable cardiovascular risk factors in a Southern European population: a randomized trial, Eur J Prev Cardiol 22(3) (2015) 389-396. doi: 10.1177/2047487313518479.

[41] F.V. Jensen, T.D. Nielsen, Bayesian networks and decision graphs, Information Science & Statistics, Springer, 2007.

[42] M. Marfell-Jones, T. Olds, A. Stewart, L. Carter, International standards for anthropometric assessment, International Society for the Advancement of Kinanthropometry, Potchefstroom: South Africa, 2006.

[43] F. Buitrago, L. Canon-Barroso, N. Diaz-Herrera, E. Cruces-Muro, M. Escobar-Fernandez, J.M. Serrano-Arias, Comparison of the REGICOR and SCORE function charts for classifying cardiovascular risk and for selecting patients for hypolipidemic or antihypertensive treatment, Rev Esp Cardiol 60 (2007) 139-147.

[44] M.R. Cobain, Assessment heart age, http://www.heartagecalculator.com, 2011.

[45] A. Soureti, R. Hurling, P. Murray, W. van Mechelen, M. Cobain, Evaluation of a cardiovascular disease risk assessment tool for the promotion of healthier lifestyles, Eur J Cardiovasc Prev Rehabil 17 (2010) 519-523.

[46] W. Buntine, A guide to the literature on learning probabilistic networks from data, IEEE T Knowl Data Eng 8(2) (1996) 195-210. DOI: 10.1109/69.494161.

[47] J. Cheng, R. Greiner, J. Kelly, D. Bell, W. Liu, Learning Bayesian networks from data: an information-theory based approach, Artif Intell 137 (2002) 43-90.

[48] L.E. Sucar, M Martínez-Arroyo, Interactive structural learning of Bayesian networks, Expert Syst Appl 15 (1998) 325-332.

[49] R.W. Robinson, Counting unlabeled acyclic digraph. In: Little CHC, editor, Lecture notes in mathematics, 622, Combinatorial mathematics V Springer-Verlag New York, (1977) 28-43.

[50] R. Daly, Q. Shen, S. Aitken, Learning Bayesian networks: approaches and issues, Knowl Eng Rev 26(2) (2011) 99-157.

[51] D. Margaritis, Learning Bayesian network model structure from data. PhD Thesis of CMU-CS-03-153 (2003).

[52] R. Nagarajan, M. Scutari, S. Lèbre, Bayesian networks in R: with applications in systems biology, Springer, 2013.

[53] M. Scurati, Learning Bayesian networks with the bnlearn R package, J Stat Soft 35(3) (2010) 1-22.

[54] R Development Core Team, R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/. 2012.

[55] S. Hojsgaard, D. Edwards, S. Lauritzen, Graphical models with R, Springer, New York, 2012.

[56] G. Claeskens, N.L. Hjort, Model selection and model averaging, Cambridge University Press, Cambridge, 2008.

[57] R.E. Neapolitan, Learning Bayesian networks, Prentice Hall, Inc Upper Saddle River, NJ, USA, 2003.

[58] Norsys Software Corporation, Netica is a trademarks of Norsys Software Corporation, Retrieved from http://www.norsys.com. Copyright 1995-2012, 2012.

[59] Weka 3.6.9: Waikato Environment for knowledge Analysis, The University of Waikato, Hamilton, New Zealand, 2013.

[60] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach Learn 29 (1997) 131-163.

[61] J.R. Quinlan, C4.5: Programs for Machine Learning, San Francisco, C.A: Morgan Kaufman, 1993.

[62] C.E. Shannon, A mathematical theory of communication. Bell Labs Tech J 27 (1948) 379-423. DOI: 10.1002/j.1538-7305.1948.tb01338.x

[63] C. Li, G. Engström, B. Hedblad, S. Calling, G. Berglund, L. Janzon, Sex differences in the relationships between BMI, WHR and incidence of cardiovascular disease: a population-based cohort study Int J Obesity 30 (2006) 1775-1781. DOI: 10.1038/sj.ijo.0803339.

[64] D.R. McCreary. Gender and age differences in the relationships between Body Mass Index and Perceived Weight: exploring the paradox, Int J Men's Health 1(1) (2002) 31-42.

[65] H.S. Park, S.B. Cho. Evolutionary attribute ordering in Bayesian networks for predicting the metabolic syndrome. Expert Systems with Applications 39 (2012) 4240-4249.

[66] M. Bennasar-Veny, A.A. Lopez-Gonzalez, P. Tauler, M.L. Cespedes, T. Vicente-Herrero, et al., Body Adiposity Index and Cardiovascular Health Risk Factors in Caucasians: A Comparison with the Body Mass Index and Others. PLoS ONE 8(5) (2013) e63999.

[67] M.B. Snijder, M. Nicolaou, I.G. van Valkengoed, L.M. Brewster, K. Stronks, Newly proposed body adiposity index (bai) by Bergman et al. is not strongly related to cardiovascular health risk, Obesity (Silver Spring) 20 (2012) 1138-1139.

[68] R.G. Baeza, V. Neira, C. Neira, M. Acevedo, Gender differences in cardiovascular risk by two different scores: a five years follow up analysis of a 1500-patient database, J Am Coll Cardiol 65(10) (2015) A1502.

[69] A. Lopez-Gonzalez, et al. Desigualdades socioeconómicas y diferencias según sexo y edad en los factores de riesgo cardiovascular. Gaceta Sanitaria 29 (2015) 27-36

[70] J. Vila-Francés, J. Sanchís, E. Soria-Olivas, A.J. Serrano, Expert system for predicting unstable angina based on Bayesian networks, Expert Syst Appl 40 (2013) 5004-5010

[71] V.G. Almeida, J. Borba, H.C. Pereira, T. Pereira, C. Correia, M. Pêgo, J. Cardoso, Cardiovascular risk analysis by means of pulsemorphology and clustering methodologies, Comput Meth Prog Bio 117 (2014) 257-266.

[72] Ch.R. Twardy, A.E. Nicholson, K.B. Korb, J. Mcneil, Epidemiological data mining cardiovascular Bayesian networks, e-J Health Informatics 1(1) (2006).

[73] S. Paredes, T. Rocha, P. de Carvalho, J. Henriques, M. Harris, J. Morais, Long term cardiovascular risk models' combination. A new approach. Comput Meth Prog Bio 101(3) (2009) 231-242.

[74] A. Elsayad, M. Fakr, Diagnosis of cardiovascular diseases with Bayesian classifiers, J Comput Sci 11(2) (2015) 274-282. DOI: 10.3844/jcssp.2015.274.282.

[75] A. Ligęza, G.J. Nalepa, A study of methodological issues in design and development of rule-based systems: proposal of a new approach. Wires Data Min Knowl, 1(2) (2011) 117-137.