eISSN 2255-0569

ORIGINAL

Usefulness of Bayesian networks in epidemiological studies

Utilidad de la redes bayesianas en los estudios epidemiológicos

P. Fuster-Parra, P. Tauler, M. Bennasar, A. Aguiló

Research Group on Evidence, Lifestyles & Health. Research Institute on Health Sciences (IUNICS)
Universitat Illes Balears, Palma de Mallorca, Spain

Corresponding author
Pilar Fuster-Parra
Edifici Anselm Turmeda
Ctra. Valldemossa km 7,5 - 07122 Palma de Mallorca
E-mail: pilar.fuster@uib.es

Recibido: 12 – V – 2014
Aceptado: 23 – VI - 2014

doi: 10.3306/MEDICINABALEAR.29.03.10

Abstract

Introduction: Bayesian networks are a form of statistical modelling, which has been widely used in fields like clinical decision, systems biology, human immunodeficiency virus (HIV) and influenza research, analyses of complex disease systems, interactions between multiple diseases and, also, in diagnostic diseases. The present study aimed to show the usefulness of Bayesian networks (BNs) in epidemiological studies.
Material and Methods: 3,993 subjects (men 1,758, women 2,235) belonging to the public productive sector from the Balearic Islands (Spain), which were active workers, constitute the data set.
Results: A BN was built from a dataset composed of twelve relevant features in cardiovascular disease epidemiology. Furthermore, the structure and parameters were learnt with GeNIe 2.0 tool. Taking into account the main topological properties some features were optimized, obtaining a hypothesized scenario where the likelihoods of the different features were updated and the adequate conclusions were established.
Conclusions: Bayesian networks allow us to obtain a hypothetical scenario where the probabilities of the different features are updated according to the evidence that is introduced. This fact makes Bayesian networks a very attractive tool.

Keywords: Bayesian networks, model averaging, cardiovascular lost years, cardiovascular risk score

Resumen

Introducción: Las redes Bayesianas son una forma de modelización estadística, las cuales han sido ampliamente utilizadas en campos como la decisión clínica, biología de sistemas, virus de inmunodeficiencia humana (VIH) e investigación en influenza, análisis de sistemas de enfermedades complejos, interacciones entre múltiples enfermedades y, también, en enfermedades de diagnóstico. Este estudio tiene como objetivo mostrar la utilidad de las redes Bayesianas en estudios epidemiológicos.
Material y Métodos: 3,993 individuos (hombres 1,758, mujeres 2,235) pertenecientes al sector productivo público de las Islas Baleares (España), los cuales eran trabajadores activos, constituyen la base de datos.
Resultados: Una red Bayesiana se ha obtenido a partir de una base de datos compuesta de doce características relevantes de la epidemiología de la enfermedad cardiovascular. Por otra parte, la estructura y los parámetros se han obtenido con la herramienta Genie 2.0. Teniendo en cuenta las principales propiedades topológicas algunas características fueron optimizadas.
Conclusiones: Las redes Bayesianas permiten obtener un escenario hipotético donde las probabilidades de las diferentes características se van actualizando de acuerdo con la evidencia introducida. Este hecho hace de las redes Bayesianas una herramienta muy atractiva, además permite establecer diversas conclusiones.

Palabras clave: Redes bayesianas, modelo promediado, años cardiovasculares perdidos, escala de riesgo cardiovascular

Introduction

Bayesian networks (BNs)[16, 25] also referred to as causal networks or beliefs networks, are a form of statistical modelling which allow us to obtain a graphical network describing the dependencies and conditional independencies from empirical data. They have proven to be a promising tool for discovering relationships[9], they capture the way an expert understands the relationships among all the features [6] and, even, they have been used in data analysis [8]. The origins of BN modelling lie within the data mining and machine learning literature[5, 13]. BNs are a kind of probabilistic graphical model (PGM)[18], which combine graph theory (to help in the representation and resolution of complex problems) and probability theory (as a way of representing uncertainty). A PGM is defined as a graph where nodes represent random variables[4, 12] and arcs represent dependencies between such variables[11, 24]. A PGM is called a BN when the graph connecting its variables is a directed acyclic graph (DAG). The graphical representation of BNs captures the compositional structure of the relations and the general aspects of all probability distributions factorized according to that structure[12].

BN modelling is widely used in fields like clinical decision[23], systems biology[7, 13], human immunodeficiency virus (HIV) and influenza research[21, 26], analyses of complex disease systems[14, 19, 20], interactions between multiple diseases[17] and, also, in diagnostic diseases[1, 2, 3, 22, 27].

The aim of the present study was to show the usefulness of Bayesian networks (BNs) in epidemiological studies focused in cardiovascular risk factors.

Theoretical Background

Let us consider a probabilistic model **M**, consisting of a set **V** of discrete random variables (features) and a joint probability distribution **P**. Let **D** (it is the graphical structure of the causal network) be a directed acyclic graph (DAG), whose set of vertices has a one to one correspondence with the variables of **M**. Two random variables X and Y in a causal network are *d-separated* if for all the paths between X and Y, there is an intermediate variable Z such that either i) the connection is serial (X→Z→Y or X←Z←Y) or diverging (X←Z→Y) and Z is instantiated or ii) the connection is converging (X→Z←Y), and neither Z nor any of Z's descendants have received evidence[15]. **D** is said to be an I-map of **M** if every d-separation condition in **D** corresponds to a conditional independence relationship in **M**. **D** is a minimal I-map of **M** if none of its arrows can be remove without destroying its –I-mapness. A BN of the probabilistic model is defined as a DAG **D** that is a minimal I-map of **M**. The joint probability distribution factorized as a product of several conditional distributions and denotes the dependency/independency structure by a DAG, which is called the *chain rule for BNs*:

$$P(X_1,\ldots,X_n)=\prod_{I=1}^{n} P(X_i \mid Pa(X_i))$$

The independencies from the graph are easily translated into the probabilistic model.

The *local Markov condition* and the global *Markov property* are important characteristics of the network topology when the BN is used to make inferences (that is to predict new scenarios when new information is introduced). The *local Markov condition* establishes that each feature 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 is conditionally independent of any other node given its *Markov blanket* (which includes its parents, its children, and the other children's parents (spouses). Any node in the BN would be *d-separated* from the nodes belonging to the *non-Markov blanket* given its *Markov blanket*.

Learning Bayesian networks

Learning BNs from a dataset has become an increasingly active area of research. Although, sometimes experts can create good BNs from their own experience, it can be a very tedious task for domains with large knowledge bases. Many methods (learning algorithms) from machine learning community have been developed to automate the creation of BNs using cases collected from past experience.

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 involves 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) for a network topology.

Bayesian network structure learning

Basically, there are two approaches to structure learning: I) *search-and-score* structure learning, and *constraint-based* structure learning. Search and score search algorithms assigns a number (score) to each Bayesian network structure, and we look for the model structure with the highest score. Constraint based search algorithms establish a set of conditional independence analysis on the data. Based on this analysis an undirected graph is generated. Using additional independence test, the network is converted into a BN.

Bayesian network parameter learning

In a BN the conditional probability distributions are called the parameters, obtaining a conditional probability distribution for each node of the network topology.

Materials and Methods

Participants

Participants were active workers belonging to the public productive sector from the Balearic Islands (Spain). Subjects were invited to participate in the study during their annual work health assessment. Any worker attending the work health assessment could be included in the study. 4,300 workers were invited to participate. Among them, 3,993 subjects (men 1,758, women 2,235) agreed to participate. Participants signed informed consent prior to enrolment in the study. After acceptance, a complete family and personal medical history was recorded. The study design was in accordance with the Declaration of Helsinki and received approval from the Balearic Islands Clinical Research Ethical Committee.

Epidemiological model

From the dataset, and using GeNIe 2.0 tool [10], a BN structure was inferred. A Bayesian search algorithm was selected to obtain a DAG, which is a search-and-score algorithm. **Figure 1** shows the obtained structure.

Once the structure is obtained the parameters could be calculated. The EM (expectation-maximization) algorithm, which is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, was used to determine the parameters. A distribution probability was obtained for each node (feature) in the DAG. The resulting BN is shown in **Figure 2**.

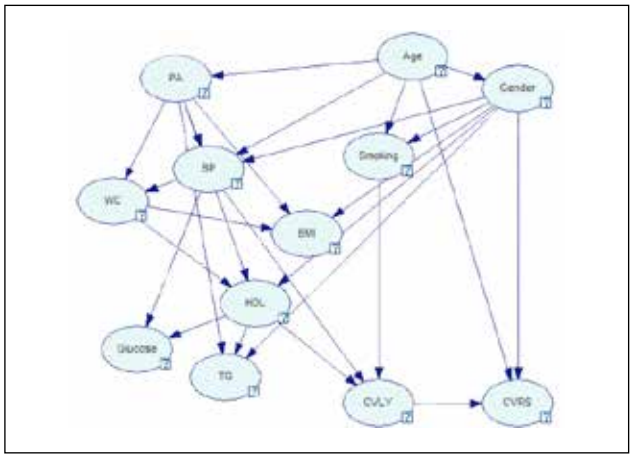


Figure 1: Structure of a BN obtained performing a Bayesian search algorithm. From the DAG it can be observed how the different features are connected.

Results and Discussion

BNs are used to make inferences [4], that is, to obtain new likelihoods of features when new information is introduced. To show it, two patterns of reasoning were selected: *causal reasoning* (from top to bottom), and *intercausal* reasoning which is close to human reasoning. In this last case the concept of *Markov blanket* was considered.

BN Inference: Causal reasoning pattern

We use this pattern when we reason from top to bottom. To illustrate this sort of reasoning we have selected four examples comparing the likelihood variations in men and women groups.

In **Figure 3** Physical Activity feature is instantiated to the "no practice" value (PA = no practice), Smoking feature is instantiated to the "yes value" (Smoking = yes), and Gender feature is instantiated to the "men value". Then, it can be observed how the likelihoods of the different features change. Likelihood of Blood Pressure (BP) feature in optimal state decreased from 0.47 to 0.18; and increased states such as HTA severe, mild and moderate, from 0.01, 0.16, and 0.05 to 0.03, 0.28, and 0.11 respectively. BMI feature in Obesity TI and Overweight GII states increased its likelihoods from 0.13 and 0.19 to 0.29 and 0.32 respectively.

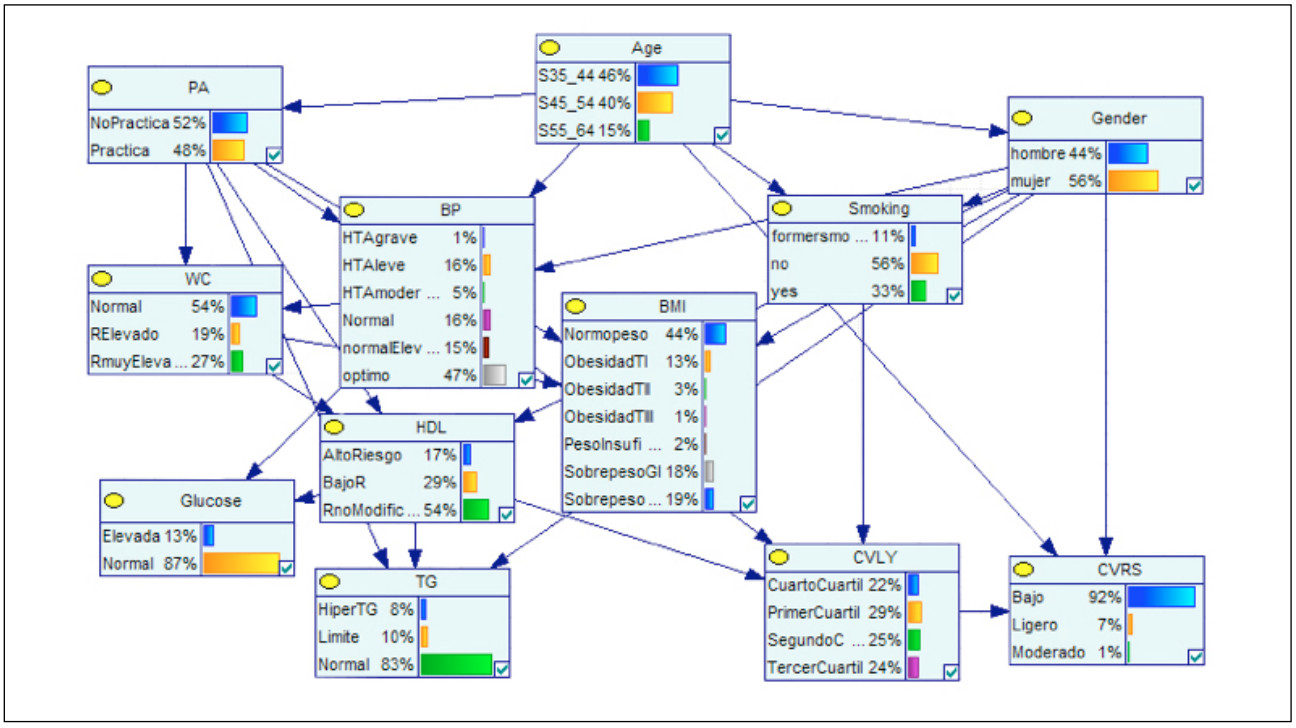


Figure 2: BN model obtained for cardiovascular disease risk factors. From the BN likelihoods this sample shows normal triglycerides (TG), normal glucose, normal cardiovascular risk score (CVRS) and normal waist circumference (WC).

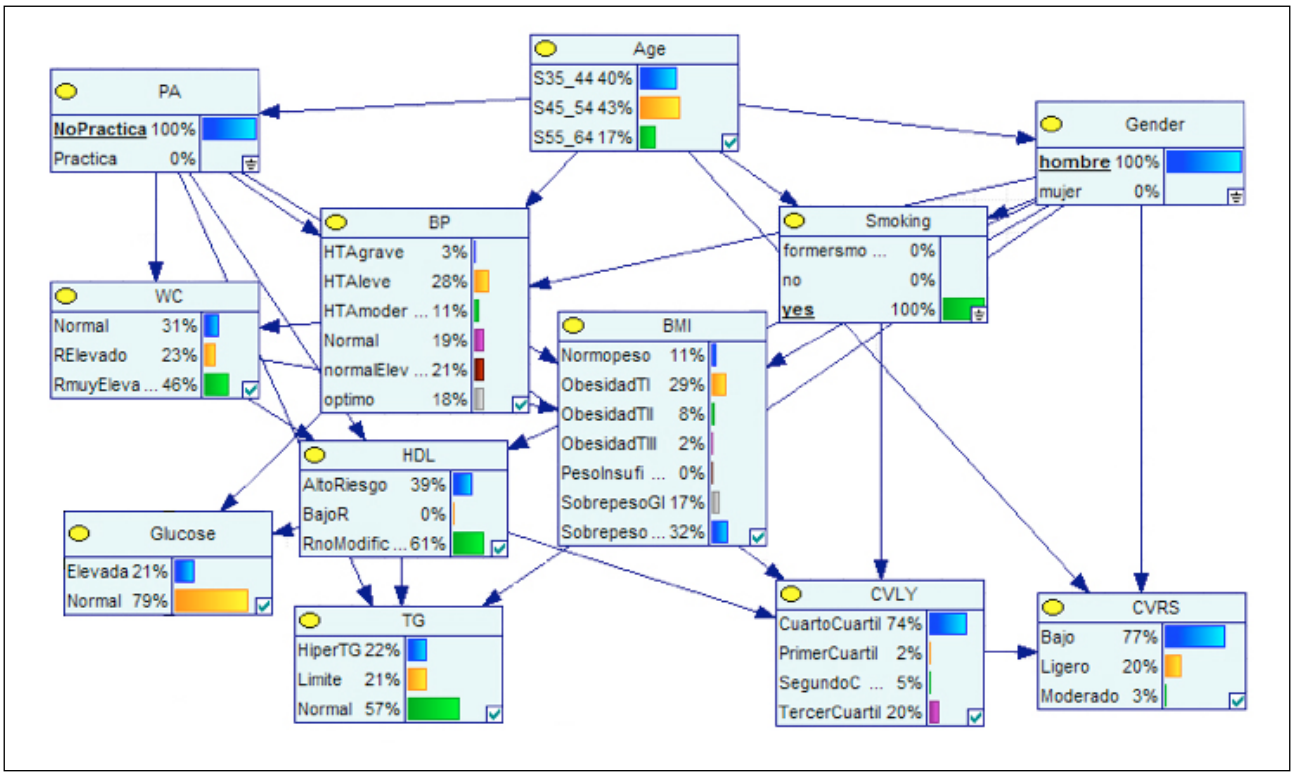


Figure 3: BN where the following evidence is introduced: physical activity (PA) = no practice, Smoking = yes, and Gender = men.

Triglycerides (TG) in normal state decreased its likelihood from 0.83 to 0.57. Glucose in normal state decreased its likelihood from 0.87 to 0.79. Cardiovascular lost years feature (CVLY) increased its fourth quartile from 0.22 to 0.74.

To compare the likelihoods variations within the group of women, we selected women state in the Gender feature. The likelihood variations are shown in **Figure 4**.

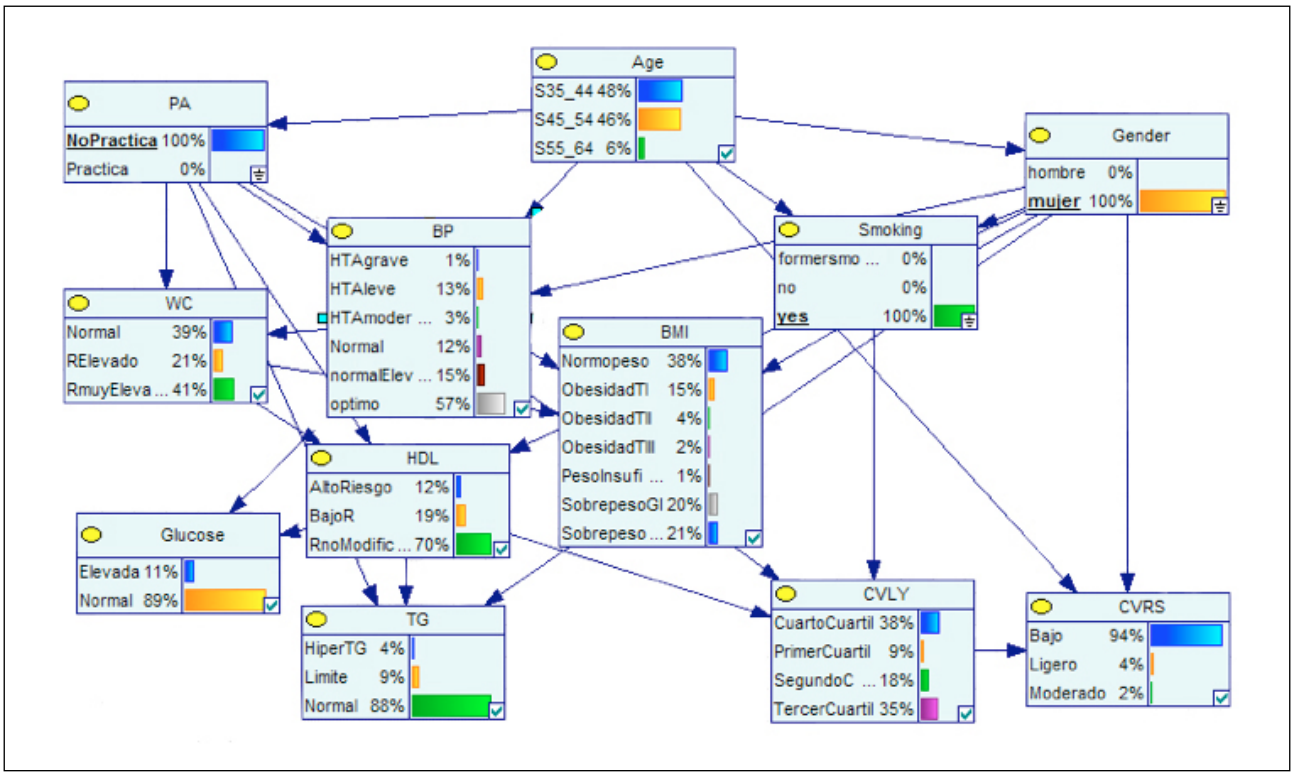


Figure 4: BN considering the following evidence is introduced: physical activity PA = no practice, Smoking = yes, and Gender =women.

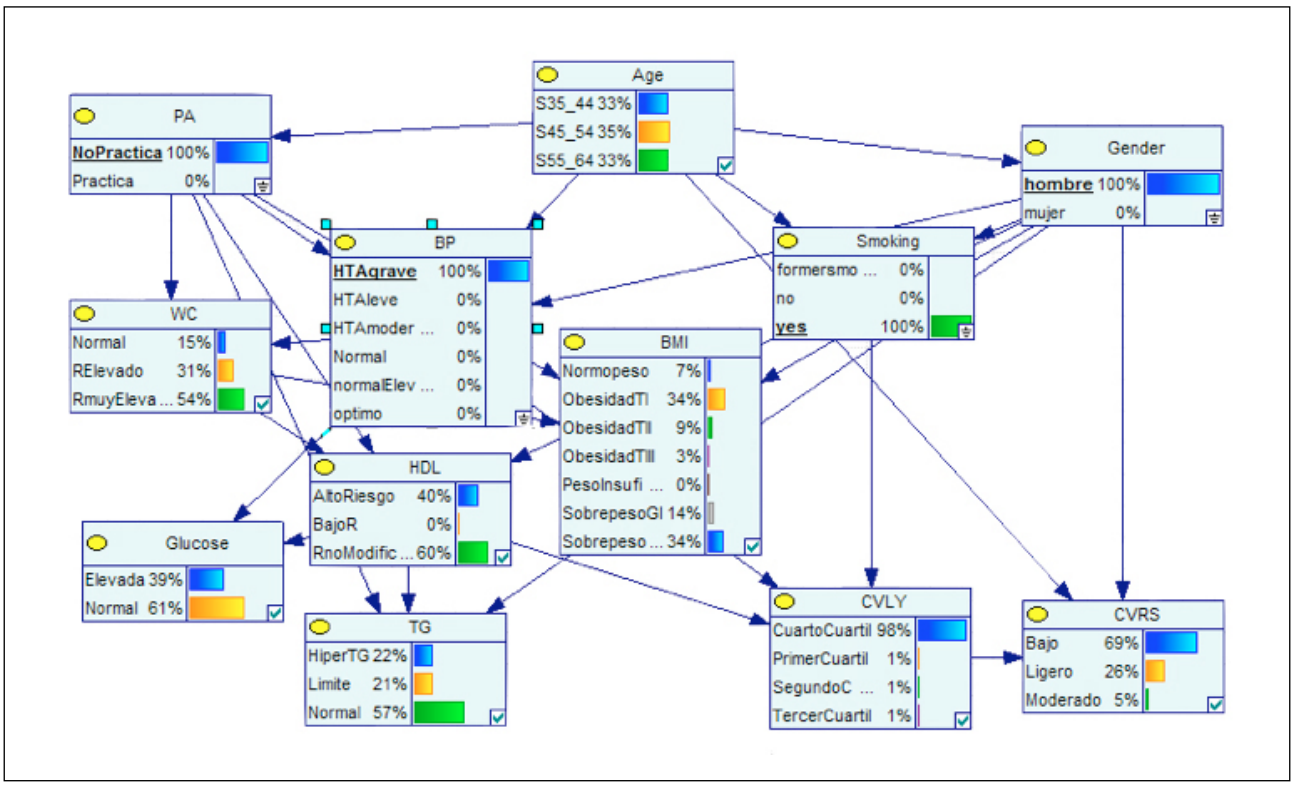


Figure 5: BN considering the following evidence: physical activity PA = no practice, Smoking = yes, Gender = men, and BP = severe.

Likelihood of Blood Pressure (BP) feature in optimal state increased from 0.47 to 0.57; and decreased states such as HTA mild and moderate, from 0.16 and 0.05 to 0.13 and 0.03 respectively; but also normal state decreased from 0.16 to 0.12. BMI feature in Obesity TI and Overweight GII states increased its likelihoods from 0.13 and 0.19 to 0.15 and 0.21 respectively, being higher in men than in women.

Triglycerides (TG) in normal state increased its likelihood from 0.83 to 0.88. Glucose in normal state decreased its likelihood from 0.87 to 0.89. Cardiovascular lost years feature (CVLY) increased its fourth quartile from 0.22 to 0.36; and its third quartile from 0.24 to 0.35. Results indicated that men were at increased cardiovascular disease risk compared to women under conditions of smoking and no practice of physical activity.
We also considered another hypothetical situation characterized by a severe hypertension (BP). The likelihood changes are shown in **Figure 5**.

Cardiovascular lost years feature (CVLY) increased its likelihood in fourth quartile from 0.22 to 0.96 and Cardiovascular risk score (CVRS) decreased its low state from 0.92 to 0.69. On the other hand, results obtained in women when blood pressure was instantiated to the highest value are shown in **Figure 6**.

In this last situation, Cardiovascular lost years feature (CVLY) increased its likelihood in fourth quartile from 0.22 to 0.83. And Cardiovascular risk score (CVRS) preserved its low state in 0.92. Taking into account these results, men were revealed again as the gender with higher cardiovascular disease risk.

BN Inference: Intercausal reasoning pattern

We refer to intercausal reasoning, which constitutes a very common pattern in human reasoning, when different causes of the same effect can interact. Using this reasoning pattern, the following two examples were considered: maximizing CVLY in first quartile state and maximizing CVLY in fourth quartile state.

To illustrate this way of reasoning the following features from the Markov blanket of CVLY were considered: CVRS, BP, HDL and Smoking. **Figure 7** shows the likelihood variation when Smoking feature is instantiated to non-smoker, Cardiovascular risk score (CVRS) is instantiated to low value, Blood pressure (BP) is instantiated to optimal value and HDL is instantiated to low value.

Under these instantiations the following changes can be observed: Gender in women state increases from 0.56 to 0.90, showing that under this situation the likelihood of being a woman is higher. Age in 35-44 state increased its likelihood from 0.46 to 0.59. The likelihood of practising physical activity increased from 0.48 to 0.83. BMI in normal weight state increased from 0.44 to 0.68.

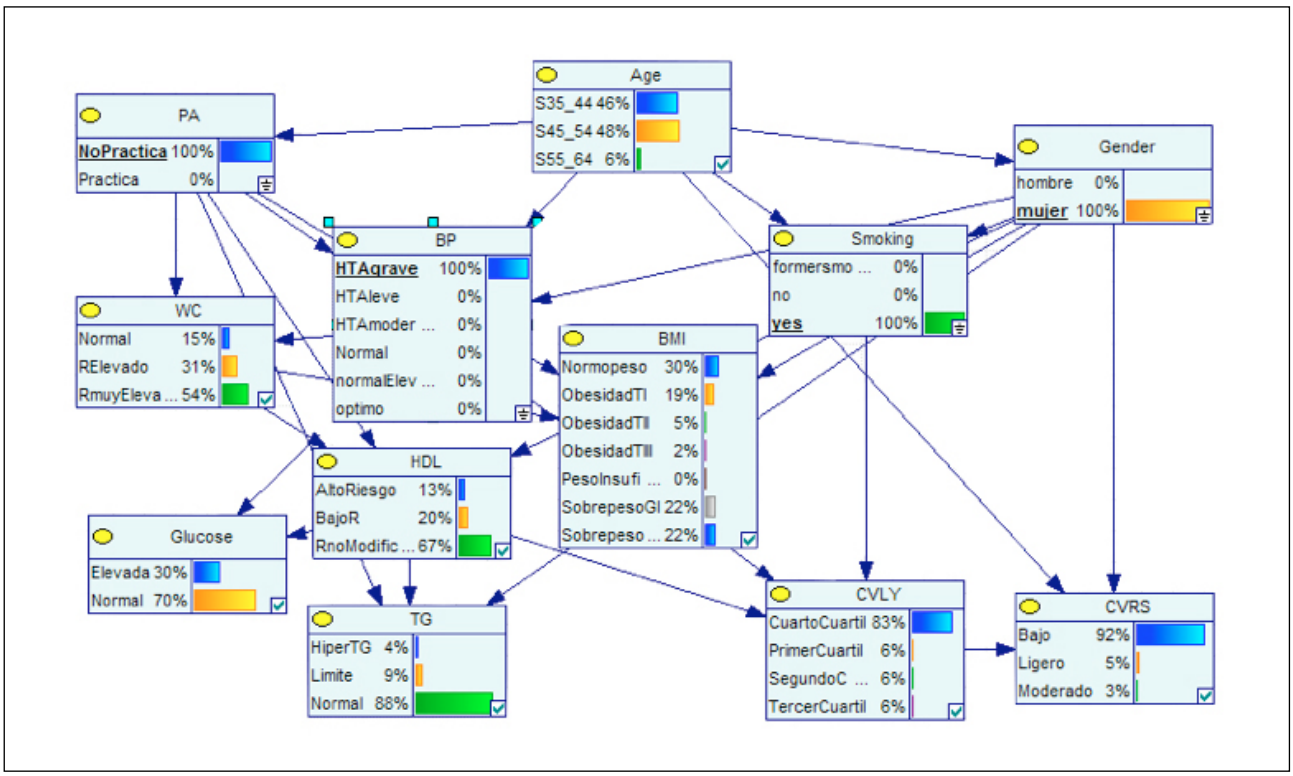
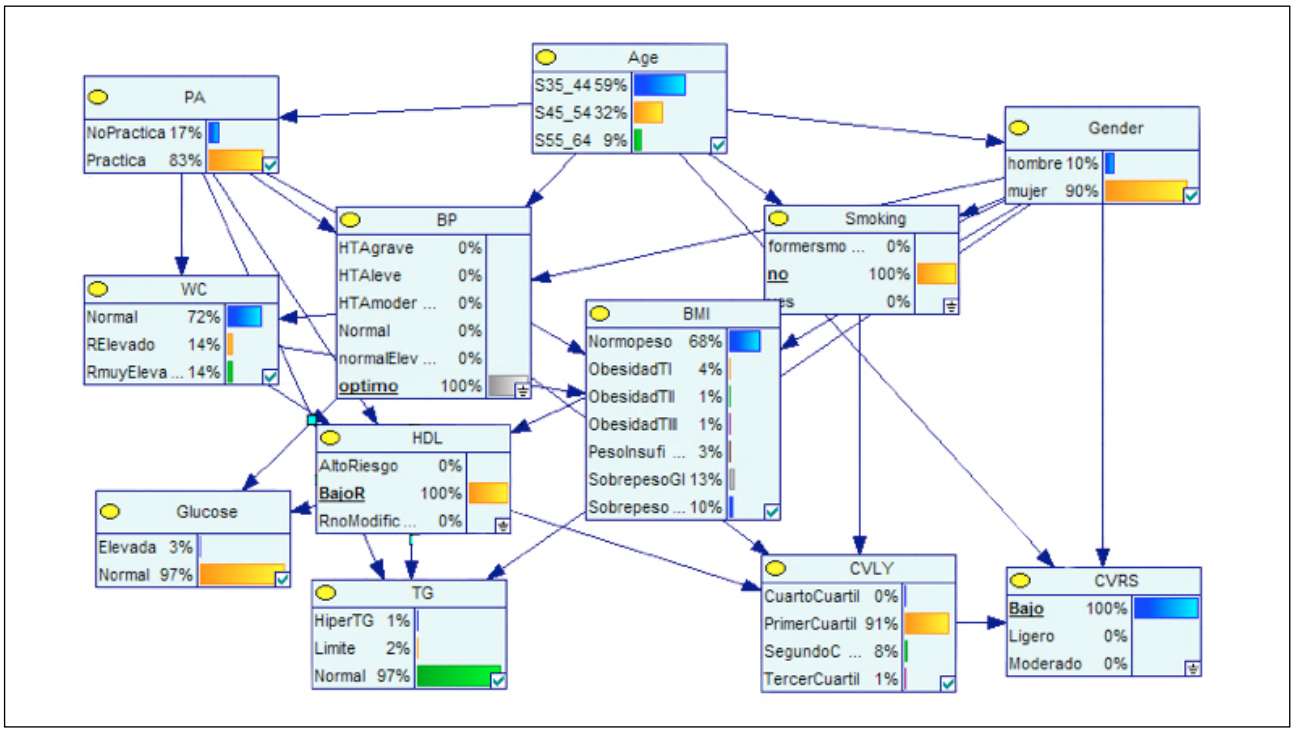


Figure 6: BN where the following evidence is introduced: physical activity PA = no practice, Smoking = yes, Gender =women, and BP = severe.



Triglycerides in normal value increased its likelihood from 0.83 to 0.97. Glucose in normal value increased its likelihood from 0.87 to 0.97. Waist circumference in normal value increased its likelihood from 0.54 to 0.72, and Cardiovascular lost years in first quartile value increased its likelihood from 0.29 to 0.91.

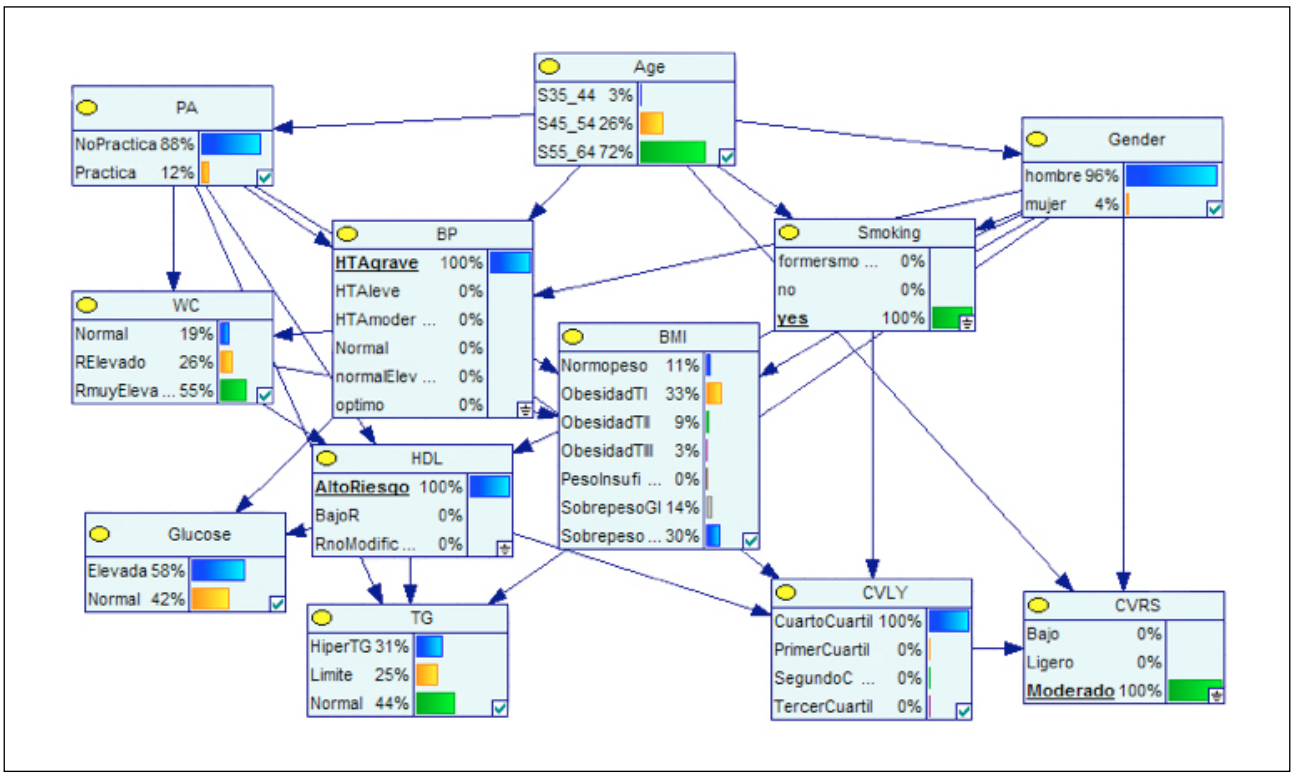


Figure 8: BN with the following evidences: Smoking = yes, CVRS = Moderate, BP = Severe, and HDL = High.

Then, to compare with this last example, Smoking feature was instantiated to the yes value, Cardiovascular risk score (CVRS) was instantiated to moderate value, Blood pressure (BP) was instantiated to severe value and HDL was instantiated to high value. The likelihood variations are shown in **Figure 8**.

Under these instantiations the likelihoods changes are as follows: Gender in men state increased from 0.44 to 0.96, showing that under this situation the likelihood of being a man is higher. Age in 56-64 state increased its likelihood from 0.15 to 0.72. Physical activity in no practice state increased its likelihood from 0.44 to 0.96. BMI in obesity TI state increased its value from 0.13 to 0.33; and in overweight GII increased its likelihood from 0.19 to 0.30. Triglycerides in normal value decreased its likelihood from 0.83 to 0.44. Glucose in high value increased its likelihood from 0.13 to 0.58. Waist circumference in very high value increased its likelihood from 0.27 to 0.55, and Cardiovascular lost years in fourth quartile value increased its likelihood from 0.22 to 1.00.

Conclusions

Using a BN model the relationships between features can be determined in a graphical and appealing way. The BN structure allows us to differentiate between direct and indirect relationships. Furthermore, the BN was used to make inferences, i.e., to predict new scenarios when hypothetically new information was introduced. Two reasoning patterns were considered: causal and inter-causal reasoning to show the likelihood variations. Because cardiovascular diseases are multi-factorial, application of this Bayesian networks could be of special interest.

References

1. Antal P, Fannes G, Timmerman D, Moreau Y, Moor B D. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artificial Intelligence in Medicine* 2003; 29: 29-60.

2. Antal P, Fannes G, Timmerman D, Moreau Y, Moor B D. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine* 2004; 30: 257-281.

3. Charitos T, Gaag L C, Visscher S, Schurink KAM, Lucas PJF (2009). A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients. *Expert Systems with Applications* 2009; 36: 12491258.

4. Butz CJ, Hua S, Chen J, Yao H. A simple graphical approach for understanding probabilistic inference in Bayesian networks. *Information Sciences* 2009; 179:699–716.

5. Cooper GF, Herskovits E. . A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*. 1992, 9(4): 309-347.

6. DeFelipe J, López-Cruz PL, Benavides-Piccione R, Bielza C, Larrañaga P et al. New insights into the classification and nomenclature of cortical GABAergic interneurons. *Nature Review Neuroscience* 2013; 14(3): 202–216.

7. Djebbari A, Quackenbush J. Seeded Bayesian networks: constructing genetic networks from microarray data. *BMC Systems Biology* 2008;2:p57. DOI:10.1186/1752-0509-2-57.

8.Friedman N, Goldszmidt M, Wyner A (1999) Data analysis with Bayesian networks: a bootstrap approach. In: Laskey K, Prade H, editors. Proceedings of the 15th annual conference on uncertainty in artificial intelligence. pp. 196–200.

9. Fuster-Parra P, García-Mas A, Ponseti FJ, Palou P, Cruz J. A Bayesian network to discover relationships between negative features in sport: a case study of teen players. *Quality & Quantity* 2013; DOI: 10.1007/s11135-013-9848-y.

10. GeNle 2.0 (Graphical Network Interface). Retrieved from http://genie.sis.pitt.edu. Copyright 1996-2003 by Decision Systems Laboratory, University of Pittsburgh. Accessed 2 Dec 2013.

11. Glymour, C., Scheines, R., Spirtes, P. & Kelly, K. Discovering causal structure. Technical report CMU-PHIL-1; 1986.

12. Glymour, C.: The mind's arrows: Bayes nets and graphical causal models in psychology. MIT Press, New York (2003)

13. Heckerman, D, Geiger, D, Chickering, DM. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* 1995; 20: 197–243.

14. Jansen R, Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003;302(5644):449453.

15. Jensen FV, Nielsen TD. Bayesian networks and decision graphs. *Information Science & Statistics*. Springer; 2007.

16. Koller D, Friedman N. Probabilistic graphical models. Principles and techniques. Cambridge, Massachusetts, London, England: The MIT Press; 2010.

17. Lappenschaar M, Hommerson A, Lucas PJF, Lagro J, Visscher S. Multilevel Bayesian networks for the analysis of hierarchical health care data. *Artificial Intelligence in Medicine* 2013; 57: 171–183.

18. Larrañaga, P. & Moral, S. Probabilistic graphical models in artificial intelligence. *Applied Soft Computing* 2011; 1511-1528.

19. Lewis FI, Brälisauer F, Gunn GJ. Structure discovery in Bayesian networks: an analytical tool for complex animal health data. *Preventive Medicine* 2011; 100(2):109-115.

20. Lewis FI, McCormick BJ. Revealing the Complexity of Health Determinants in Resource-poor Settings. *American Journal of Epidemiology*. 2012; 176(11):1051–1059.

21. Lycett SJ, Ward MJ, Lewis FI, et al. Detection of mammalian virulence determinants in highly pathogenic avian influenza H5N1 viruses: multivariate analysis of published data. *Journal of Virology* 2009;83(19):99019910.

22. Maskery SM, Hu H, Hooke J, Shriver CD, Liebman M N. A Bayesian derived network of breast pathology co-occurrence. *Journal of Biomedical Informatics* 2008; 41: 242-250.

23. Sesen MB, Nicholson AE, Banares-Alcantara R, Kadir T and Brady M. Bayesian networks for clinical decision support in Lung Cancer Care. *Plos One* 2013; 8(12): e82349. DOI: 10.1371/journal.pone.0082349.

24. Spirtes P, Glymour C, Scheines R. Causation, prediction and search, 2nd ed., Adaptive Computation and machine learning. The MIT Press; 2001.

25. Pearl J. Causality. Models, reasoning and inference. Cambridge: Cambridge university press; 2000.

26. Poon AF, Lewis FI, Pond SL, et al. Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. *PLoS Computational Biology* 2007; 3(1):pe11. DOI:10.1371/journal. pcbi.0030011.

27. Wang XH, Zheng B, Good WF, King JL, Chang YH. Computer assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics* 1999; 54: 115126.