ELSEVIER

Artificial Intelligence in Medicine 30 (2004) 215–232

Artificial Intelligence in Medicine

www.intl.elsevierhealth.com/journals/aiim

A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service

Silvia Acid[a], Luis M. de Campos[a,*], Juan M. Fernández-Luna[b], Susana Rodríguez[c], José María Rodríguez[c], José Luis Salcedo[c]

[a]*Departamento de Ciencias de la Computación e I.A., Universidad de Granada, Escuela Técnica Superior de Ingeniería Informática, Avda. de Andalucía 38, Granada E-18071, Spain*

[b]*Departamento de Informática, Universidad de Jaén, Jaén, Spain*

[c]*Hospital Universitario Virgen de las Nieves Granada, Granada, Spain*

Received 2 June 2002; received in revised form 10 November 2002; accepted 23 June 2003

Abstract

Due to the uncertainty of many of the factors that influence the performance of an emergency medical service, we propose using Bayesian networks to model this kind of system. We use different algorithms for learning Bayesian networks in order to build several models, from the hospital manager's point of view, and apply them to the specific case of the emergency service of a Spanish hospital. This first study of a real problem includes preliminary data processing, the experiments carried out, the comparison of the algorithms from different perspectives, and some potential uses of Bayesian networks for management problems in the health service.

Keywords: Bayesian networks; Learning algorithm; Scoring functions; Independence; Emergency medical service; Management decision support in the health service

1. Introduction

Over the past four decades, a lot of effort has been put into developing medical decision support systems. There is a great variety of commercially available programs to assist clinicians with diagnosis, decision-making, pattern recognition, medical reasoning, filtering, etc. both for general and very specialized domain applications. In recent years,

* Corresponding author. Tel.: +34-958-244019; fax: +34-958-243317.
E-mail address: lci@decsai.ugr.es (L.M. de Campos).

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2003.11.002

however, it has become clear that it is not only physicians but health professionals in other areas as well who also need decision support: e.g. nursing, health administration, medical education, patient care, etc. This paper is concerned with management in the health service, but not as an isolated system component, since there is a great deal of interdependence between components (for instance, clinical actions affect the treatment cost per patient; conversely, reorganization and changes in the scheduling administration may change the medical procedures). Moreover, they may have conflicting goals. Thus, from a hospital manager's perspective, a trade-off between quality of service and financial costs with budgetary limitations must be found.

Health-care systems are complex and depend on organizational, economical, and structural factors. The availability of appropriate tools for their representation would allow the interactions between the different elements that determine their behavior to be studied and understood, as well as some alternatives to be analyzed so as to improve their performance. As many of the factors that influence the performance of a health-care system are uncertain, Bayesian networks could play an important role in their study as formal models to represent knowledge and handle uncertainty. We wish to take advantage of their ability to describe the interactions between variables explicitly. An example of the interest for managing resources for geriatric services in a hospital using Bayesian networks can be found in [22], which aims to forecast the duration of stay and destination on discharge of elderly people.

In this paper, we introduce some representation models, based on Bayesian networks, which are applied to the specific case of an emergency medical service. These models have been obtained from real data recorded at the hospital ''Virgen de las Nieves'', by using several algorithms for learning Bayesian networks. Although our long-term objective is to develop a management-oriented decision support system, in this paper we focus on a less ambitious but necessary preliminary aspect: the study of the capabilities of different Bayesian network learning algorithms in order to generate useful models for this problem. We have therefore selected a representative subset of the currently available algorithms for learning Bayesian networks and we have carried out a series of experiments to evaluate their behavior from different perspectives.

The paper is structured as follows: in Section 2 we shall describe the problem to be studied, the available data, and the pre-processing steps (discretization, variable selection, etc.) which are used to obtain a suitable database for the learning algorithms. In Section 3, we comment on the different learning algorithms we have considered for our experiments. Section 4 describes the networks obtained for the different algorithms. In Section 5, we summarize the results of several experiments, which attempt to assess the quality of the networks from different points of view. Finally, Section 6 discusses the conclusions of this work.

2. The problem

As we have already mentioned, we wish to model certain aspects of the health-care system for patients arriving at a hospital's emergency department. Our first aim is simply to better understand the interactions between some of the factors that shape this system, and

Table 1
Variables initially considered

Variable	Possible values
Financing	10
Date of Admission	Date
Time of Admission	0:01–24:00
Cause of admission	8
Pathology	7
P10	2
Identification	6
Date of Discharge	Date
Time of Discharge	0:01–24:00
Cause of Discharge	9
Medical Service	36

obtain a model that describes the nature of the system reasonably well. This model could then be used to make predictions about some of the variables of interest, or even to make decisions about the configuration of the system itself. Our approach is management oriented, and attempts to help the hospital manager in organizational and economical questions (for example, the possible redistribution or reinforcement of personnel and/or infrastructure) rather than clinical problems (although a better use of the available resources would also imply an improvement in the medical care).

2.1. The data set

From the set of variables which are collected when a patient enters the emergency department, the variables displayed in Table 1 were initially selected. Some of these are recorded when the patient arrives, others once the patient has been treated (although no patient clinical data is taken into account). In this table, we also show either the number of possible values or the range for each variable. For the experiments we had at our disposal a database containing 31,937 records (corresponding to all the arrivals to the emergency departments of the hospital ''Virgen de las Nieves'' at Granada, from 01 January–20 February 2001), although we could dispose of a separate data set of 12,291 records, corresponding to the next admissions, occurred from 21 February–10 March 2001. This second dataset will be used as a test set in our experiments.

Financing represents the type of entity that covers the expenses (Social Security, Insurance Companies, International Agreements, Mutual Health Insurance, etc.). Cause of Admission codifies eight different values (considered as confidential by the hospital staff). Pathology includes Common Disease, Common Accident, Industrial Accident, Traffic Accident, Aggression, Self-inflicted Lesion and Other. P10 represents whether the patient was sent to the emergency medical service by a family doctor. Identification codifies the type of patient's identification document (Identity Card, Social Security Card, Passport, Oral Identification, Unidentified and Other). Cause of discharge represents several reasons (Return to Duty, Death, Hospitalization, Transfer to Another Hospital, Voluntary Discharge, Indeterminate, etc.). Medical Service includes all of the

36 different emergency units at the hospital (Hematology, Intensive Care, Endocrinology, etc.). All of the variables described (those in italics in Table 1) were used just as they were, but for the remaining four variables in Table 1, some additional treatment was necessary.

2.2. Pre-processing of data

We have discretized some variables as follows:

- *Date of Admission*. We discretized this into seven values, corresponding to the days of the week. From now on, we shall call this variable *Day*.
- *Time of Admission*. We discretized this into three values, corresponding to the three different time periods of the day: morning (8:01–15:00), evening (15:01–22:00) and night (22:01–8:00). From now on, we shall call this variable *Shift*.

We also defined any new variables which were considered relevant:

- *Duration*. The length of time (h) that the patient stayed in the emergency department. This value is calculated from the values of Date and Time of Admission and Date and Time of Discharge. In addition, this new variable was discretized into three values (from 0 to 8 h, from 8 to 72 h, and more than 72 h) which were considered meaningful by the physicians. They correspond, respectively, to 'normal', 'complicated' and 'anomalous' cases.
- *Centre*. The hospital has three different emergency departments corresponding to the three centres that comprise it (Maternity Hospital, Orthopedic Surgery, and General Hospital).

The variables Date and Time of Discharge were considered irrelevant for our purposes, since the truly relevant information is the Duration of stay. These two variables were therefore removed. Consequently, we have considered a total of 11 variables, which are showed in Table 2. Note that the size of the space of states for the 11 variables is quite large: 411,505,920 possible configurations.

Table 2
Variables used in our model

Variable	Possible values
Financing	10
Day	7
Shift	3
Cause of admission	8
Pathology	7
P10	2
Identification	6
Duration	3
Cause of discharge	9
Medical service	36
Centre	3

3. The learning algorithms

As we are looking for a representative model for our problem, we used several algorithms for learning the structure of a Bayesian network from the data set containing 31,937 cases. The selected algorithms are driven by different principles and/or metrics, so the resulting models may differ in their results—the relationships they extract. On the one hand, we aim to compare their performance on a real problem; on the other hand, the arcs appearing in all the learned networks could be considered as being the 'core' for this representation model. Any consensus Bayesian network should be built from this shared structure.

Although there are a great many algorithms for learning Bayesian networks from data, they can be subdivided into two general approaches: methods based on *conditional independence tests*, and methods based on a *scoring function* and a *search* procedure. There are also *hybrid* algorithms that use a combination of independence-based and scoring-based methods.

The algorithms based on independence tests perform a qualitative study of the dependence and independence relationships between the variables in the domain, and attempt to find a network that represents these relationships as far as possible. They therefore take a list of conditional independence relationships (obtained from the data by means of conditional independence tests) as the input, and generate a network that represents most of these relationships. Some of the algorithms based on this approach can be found in [10,12,26].

The algorithms based on a scoring function (also called a *metric*) attempt to find a graph that maximizes the selected score; the scoring function is usually defined as a measure of fit between the graph and the data. All use a scoring function combined with a search method in order to measure the goodness of each explored structure from the space of feasible solutions. During the exploration process, the scoring function is applied in order to evaluate the fitness of each candidate structure to the data. Each algorithm is characterized by the specific scoring function and search procedure used. The scoring functions are based on different principles, such as entropy [19], Bayesian approaches [8,14,18], or the Minimum Description Length (MDL) [6,21].

We have used the following algorithms, which are a representative sample of the different approaches for learning Bayesian networks:

- PC [26], an algorithm based on independence tests. It starts by forming the complete undirected graph, which it then thins by removing edges with zero order conditional independence relationships, and then rethins with first order conditional independence relationships, and so on. The set of variables conditioned only needs to be a subset of the set of variables adjacent to one or other of the variables conditioned; this is constantly changing as the algorithm progresses. We used an independence test based on the measure of conditional mutual information [20], with a fixed confidence level equal to 0.99.
- Another algorithm, the BN Power Constructor (BNPC), uses independence tests and mutual information [12]. This algorithm has a three-phase operation: drafting, thickening, and thinning. In the first phase, the algorithm computes mutual information of each pair of nodes as a measure of closeness, and creates a draft based on this information. In the second phase, the algorithm adds arcs when the pairs of nodes are not conditionally

independent on a certain conditioning set. In the third phase, each arc is examined using conditional independence tests and will be removed if the two nodes of the arc are conditionally independent.

- A scoring-based algorithm, that uses local search (LS) in the space of directed acyclic graphs (DAGs) [18]. This kind of method starts from an initial DAG and, at each step, performs the local change (operator) yielding the maximal gain, until a local maximum of the scoring function is reached. In our case, the local search used is based on the classical operators of arc addition, deletion and reversal (and an initial empty graph). The (Bayesian) scoring function considered is BDeu[1] [18]. We used BDeu since it is by far the most popular scoring function in recent Bayesian network learning literature.

  We also carried out experiments with scoring-based algorithms using more powerful search heuristics than a simple local search: *Tabu Search* (TS) [7] and *Variable Neighborhood Search* (VNS) [11]. However, we obtained exactly the same results as those of LS (and therefore, we do not report them separately).
- A version of the BENEDICT[2] (BE) algorithm [5]. This algorithm, which searches in the space of equivalence classes of DAGs, is based on a hybrid methodology [1] (other versions of BENEDICT, that search in the space of DAGs with a given ordering of the variables, and use a slightly different metric, can be found in [3,4]). In contrast to other approaches [15,28] that maintain the independence-based and scoring-based algorithms as separate processes, combined in some way, in this case the hybridization is based on the development of a scoring function that quantifies the discrepancies between the independences displayed by the candidate network and the database, and the search process is limited by the results of some independence tests. The basic idea of this algorithm is to measure the discrepancies between the conditional independences represented in any given candidate network G and those displayed by the database. The smaller these discrepancies are, the better the network fits the data. The aggregation of all these local discrepancies results in a measure of global discrepancy between the network and the database (this is the scoring function to be minimized). The local discrepancies are measured using the conditional mutual information between pairs of non-adjacent variables in the candidate graph G, given a d-separating set of minimum size [2]. The main search process is greedy and only addition of arcs is permitted, although a final refining process (reinsertion of discarded arcs and pruning of inserted arcs) mitigates the irrevocable character of the whole search method.

The experiments we shall describe have been performed using our own implementations for the cases of PC, LS, and BE. The first two algorithms are integrated in the Elvira[3] software package available at http://leo.ugr.es/∼elvira. For BNPC, we used the software package available at http://www.cs.ualberta.ca/∼jcheng/bnsoft.htm.

In order to compute the conditional (or marginal) probability distributions stored at each node in the network, thus obtaining a complete Bayesian network, we used a maximum likelihood estimator (frequency counts) in all the cases.

[1] With the value of the equivalent sample size parameter set to 1 and a uniform structure prior.

[2] Acronym for BElief NEtwork DIscovery using Cut-set Techniques.

[3] An environment for the edition, evaluation and learning of Bayesian networks and influence diagrams, developed as a research project in our department, in collaboration with other Spanish universities.

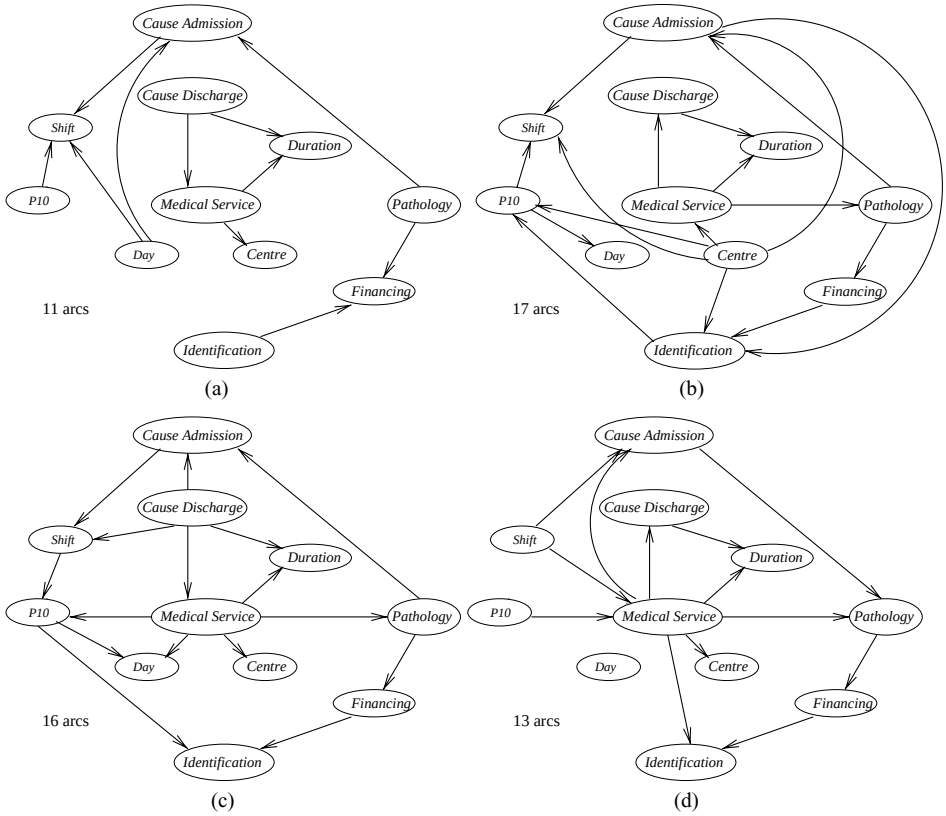

Fig. 1. The different structures recovered by the selected algorithms: (a) PC, (b) LS, (c) BE and (d) BNPC.

4. Results

After running the learning algorithms, we obtained four different networks, and these are displayed in Fig. 1. We do not assume a causal interpretation of the arcs in the networks, although in some cases this might be reasonable (other approaches that explicitly try to detect causal influences are discussed in [17,24]). Instead, we interpret the arcs as direct dependence relationships between the linked variables, and the absence of arcs means the existence of conditional independence relationships.

In order to summarize the differences and resemblances between models, Table 3 shows the two numbers l/a for each pair of algorithms, where l is the number of common edges (in either direction), and a the number of common arcs[4] between the networks learned by these algorithms. The main diagonal in this table represents the number of arcs contained in each network. Fig. 2 displays the edges in common to all the networks: three arcs and five

[4] Taking into account the fact that the direction of some arcs is not relevant, i.e. if we change the direction of these *reversible* arcs, we obtain an equivalent model [25].

Table 3
Number of common links and arcs, l/a, between pairs of learned networks

	PC	LS	BE	BNPC
PC	11/11	9/8	9/7	8/5
LS	–	17/17	12/10	9/7
BE	–	–	16/16	10/7
BNPC	–	–	–	13/13

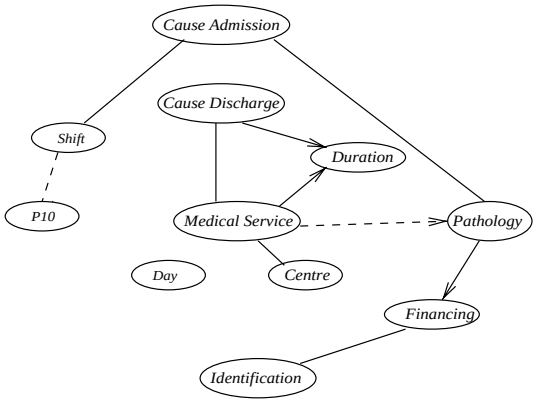

Fig. 2. The incomplete structure shared by all the networks (solid lines). Dashed lines represent edges shared by three of the four networks.

undirected edges[5]. We also display two additional edges in this figure that are supported by all the networks except one. Note that the number of possible edges in this domain is 55, and only a total of 26 different edges appear in these models. The four models therefore agree in the presence of 8 edges and the absence of 29 edges, i.e. the existence of 8 direct dependence and 29 conditional independence assertions between pairs of variables.

The direct dependence relationships that are common to all models may be explained in the following way: the reason for the strong relation between Pathology and Financing[6] is due to the fact that different entities cover the expenses depending on the type of pathology (Traffic Accident, Industrial Accident, etc.). Financing also depends on Identification (obviously the expenses will only be covered by a particular entity or company if the patient can be identified as belonging to this entity). The connection between Pathology and Cause of Admission is obvious. The relation between Cause of Admission and Shift may be due to the fact that the reason for going to the emergency department varies according to the arrival time. The connection between Medical Service and Centre is justified because Centre is a variable functionally dependent on Medical Service (each

[5] In the last ones, there was some disagreement in the directionality of the edges: in Fig. 2, an undirected edge A–B means that each network contains either the arc $A \rightarrow B$ or the arc $A \leftarrow B$.

[6] This link was introduced into the graph at a very early stage by the different algorithms when their respective models were constructed.

Centre has its own emergency medical units). The Duration of the stay at the emergency department essentially depends only on the medical unit (Medical Service) that treated the patient, and the Cause of Discharge (the seriousness of the diseases and the degree of congestion of the service, which are strongly related with the duration of the stay, probably vary from one unit to another). In turn, these two variables are highly correlated: for example, the cause of discharge being death is much more unlikely for some medical units than others. Some of these relationships are more or less obvious, but others, although they may not be particularly remarkable, may be useful for management purposes: the edge connecting Cause of Admission and Shift may suggest a reinforcement of some Services for some Shifts. Similarly, the fact that only Cause of Discharge and Medical Service directly influence the Duration of the stay (all the remaining variables being conditionally independent of Duration) suggests the need for a detailed study of these three variables in order to better understand why some medical units require a longer stay than others.

Each network, in addition to the eight direct dependence relationships described above, represents other connections. For example, three of the four models establish an edge linking Medical Service and Pathology, which is, in our opinion, quite plausible. Three of the networks also find a direct connection between Shift and P10, which may indicate that the arrival pattern is different according to whether the patients have a P10 document or not. Two of the models establish a (probably weak) connection between the existence of a P10 document and the three variables Day, Medical Service, and Identification. Finally, there are several edges which are supported by only one network model.

Apart from these dependence relations, by using the graphical criterion of independence called *d*-separation [23], we also can obtain a number of conditional independence relationships, some of which might contribute useful information. For example, all the models indicate that Pathology and Cause of Discharge are independent once Medical Service is known; in addition, Financing and Duration are conditionally independent given Pathology (and given Pathology together with any other subset of variables).

With respect to the algorithm running times, it is not very useful in this case to make time comparisons, since the algorithms proceed from different sources (except PC and LS) and they were run over different platforms. In any case, our implementations of PC, LS, and BE were quite fast: they required 63, 41, and 30 s, respectively, in order to learn the structure of the corresponding networks.

5. Experiments

When we have a set of different algorithms for performing a task (or the only algorithm available may run with different parameters), and the obtained results (the network models) are different, it is useful to provide some criteria in order to select a preferred model.

In order to assess the quality of the different network models, one of the most commonly used criteria is the percentage of classification success. However, while it is important to stress the representation power of the Bayesian networks for a given problem, there is not a unique classification variable: from a manager's perspective, the duration of the stay, the involved medical unit or even the shift might be of interest. Consequently, other additional evaluation methods are therefore necessary. We need some measures that assess the degree

of discrepancy or the fitness of a network to the available data, for example the probability that the data have been generated by a given network model.

We have collected the following performance measures about the networks obtained with the different learning algorithms:

- The Kullback–Leibler (KL) distance (cross-entropy) between the probability distribution, P_D, associated to the database D (the empirical frequency distribution), and the probability distribution associated to the learned network, P_G. In this way, we attempt to assess the performance of the algorithm from the perspective of how closely the probability distribution learned approximates the empirical frequency distribution. We have in fact calculated a decreasing monotonic transformation of the Kullback–Leibler distance, since this has exponential complexity and the transformation may be computed very efficiently [9]. The interpretation of our transformation of the Kullback–Leibler distance is: the higher this value, the better the network fits the data. However, this measure should be handled cautiously, because a high KL value may also indicate overfitting (a network with many edges will probably have a high KL value).
- The values (in log version) of the K2 [14], the BDeu [18], and the BIC [27] metrics for the learned networks. These measures can offer an idea of the quality of the networks from different points of view. BDeu and K2 are Bayesian metrics, and both measure the marginal likelihood $P(D|G)$ (which, together with a uniform structure prior, $P(G)$, enables us to compute $P(G, D)$). The difference between BDeu and K2 lies in the choice of the priors for the conditional Dirichlet distributions of the network parameters given a fixed structure[7]. The Bayesian Information Criterion (BIC) metric is a penalized version of the likelihood $P(D|\hat{G})$ (with the parameters associated to the network structure estimated using maximum likelihood), and contains an explicit penalty term for network complexity. It should be noted that the BIC metric can also be seen as an MDL metric. In all three cases, the higher the value of the metric, the better the network.

The values of the different metrics for all the networks considered are showed in Table 4. We have also computed the performance measures corresponding to the empty network ($\emptyset_{\text{em}}$), which is obviously a rather poor model (with no interaction between the variables), but its corresponding values may serve as a kind of base line. In the table, the numbers in brackets represent, for each metric, the relative merit of each algorithm (with (1) corresponding to the best value, and (5) to the worst one). We also show the number of arcs included in each network[8].

In the light of the resulting values, we can conclude that the LS algorithm performs quite well with respect to all the metrics. Moreover, LS is the algorithm that obtains the densest network (17 arcs). On the other hand, PC produces the sparsest network (11 arcs) and obtains bad KL, K2, and BDeu values. The BE and BNPC algorithms obtain quite balanced networks with respect to all the metrics and an intermediate number of arcs. It should be

[7] BDeu uses a uniform joint distribution whereas K2 uses a distribution that is locally but not globally uniform.

[8] This number may be of assistance when selecting simpler networks, according to Occam's razor, if other measures do not discriminate between models.

Table 4
Performance measures for the different learned networks, with respect to the training set

Algorithm	Metrics				Number of arcs
	KL	K2	BIC	BDeu	
BE	2.447 (3)	−101016 (2)	−243420 (2)	−233339 (3)	16
PC	2.152 (4)	−104834 (4)	−249509 (3)	−240611 (4)	11
LS	2.490 (1)	−100241 (1)	−243243 (1)	−229728 (1)	17
BNPC	2.485 (2)	−101308 (3)	−258123 (4)	−231768 (2)	13
$\emptyset_{em}$	0.000 (5)	−133315 (5)	−306937 (5)	−306874 (5)	0

noted that we are using a logarithmic version of the metrics, so the differences are much greater in a non-logarithmic scale.

In order to test a possible overfitting of the networks to the data, we have also computed the same performance measures but using a test set which differs from the training set used to learn the networks (the data set containing 12,291 cases). The results are showed in Table 5.

We can see that the results in Table 5 are similar to the ones obtained in Table 4. We can therefore conclude that from the point of view of the selected performance measures, the best algorithm for this domain is LS, the worst is PC, whereas BE and BNPC obtain intermediate results.

However, an important question is whether the differences between the networks in terms of these metrics also lead to differences in terms of the usefulness of these networks for specific situations.

As we mentioned above, the networks learned can also be used with predictive purposes, by using the inference methods (propagation of evidence) available for Bayesian networks. More precisely, from the perspective of a classification problem, we want to use the networks in order to predict the most probable values of any variable of interest given some evidence, and compare the predictions obtained with the true values of this variable, thus obtaining the corresponding percentages of success. For this purpose, we have considered three different situations:

(a) Predicting the values of Duration, given evidence about the values of all the other variables, except Cause of Discharge. In this way, we attempt to determine the most

Table 5
Performance measures for the different learned networks, with respect to the test set

Algorithm	Metrics			
	KL	K2	BIC	BDeu
BE	2.35 (3)	−38740 (2)	−99483 (1)	−89643 (3)
PC	2.07 (4)	−39972 (4)	−99816 (2)	−91238 (4)
LS	2.40 (1)	−38324 (1)	−99896 (3)	−87279 (1)
BNPC	2.40 (1)	−39054 (3)	−113297 (4)	−88776 (2)
$\emptyset_{em}$	0.00 (5)	−49969 (5)	−115032 (5)	−114974 (5)

Table 6
Success percentages of classification for Duration and Medical Service, using the test set

Algorithm	Duration (%)	Medical Service (%)
BE	91.6 (2)	75.0 (2)
PC	91.6 (2)	76.1 (1)
LS	91.6 (2)	74.5 (3)
BNPC	91.6 (2)	71.5 (4)
$\emptyset_{em}$	96.1 (1)	31.9 (5)

probable duration of the stay at the emergency department before the patient is effectively discharged.

(b) Predicting the values of Medical Service, given evidence relative to all the remaining variables, except Pathology, Cause of Discharge, and Duration, which would be unknown at the time the patient arrives. If accurate, this prediction could serve to direct the arriving patient to the appropriate emergency unit.

(c) Predicting the value of each of the 11 variables, given evidence about all the 10 remaining variables. In this way, we attempt to test the behavior of the network models for different problems. This experiment could serve to assess the robustness of the networks as general classifiers (as opposed to having to manage a different model to classify each variable of interest).

For all the classification problems, we used the previously learned networks and the success percentages were calculated using the independent test set containing 12,291 cases.

Table 6 displays the percentages of success of the different networks for the first two classification problems considered. In the case of predicting the duration of the stay, all the learned networks perform equally well, whereas in the other situation, PC and BE obtain the best results. With respect to predicting the duration of the stay, it should be noted that the results are worse than the ones obtained by the empty network. The reason is that the distribution of the duration of the stay is rather biased towards its first value (from 0 to 8 h), and therefore the default rule, which assigns the 'a priori' most probable class to all the cases, obtains a high percentage of correct classifications[9]. For the problem of predicting the medical service involved, the results remarkably outperform the prediction of the empty network.

Table 7 displays the percentages of success of the different networks for the other 11 classification problems. The results are somewhat surprising, because the supposedly best algorithm, LS, performs rather poorly, whereas BE and BNPC obtain the best results.

In the light of the poor result obtained by LS from a classificatory point of view, we raise the following question: Is this result due to the specific metric (BDeu) being considered? In other words, could an LS algorithm equipped with another scoring metric outperform the results obtained by BE and BNPC (which are algorithms based on independence tests instead on scoring metrics)? In order to answer this question, we have considered two

[9] A finer discretization of the variable Duration would probably lead to much better results.

Table 7
Success percentages of classification for the 11 variables, using the test set

	BE	PC	LS	BNPC	$\emptyset_{em}$
CoA%	91.7 (2)	91.6 (3)	88.1 (5)	91.8 (1)	91.4 (4)
CoD%	75.6 (1)	74.3 (2)	74.3 (2)	74.3 (2)	62.4 (5)
Cen%	100 (1)	100 (1)	94.3 (4)	100 (1)	39.6 (5)
Day%	13.8 (1)	13.2 (2)	12.6 (3)	12.6 (3)	12.6 (3)
Dur%	91.4 (2)	91.4 (2)	91.4 (2)	91.4 (2)	96.1 (1)
Fin%	93.7 (2)	93.7 (2)	93.6 (3)	93.6 (3)	93.8 (1)
Ide%	81.9 (2)	79.1 (3)	79.1 (3)	79.1 (3)	82.2 (1)
MS%	81.8 (2)	81.6 (3)	81.6 (3)	83.0 (1)	31.9 (5)
P10%	95.2 (1)	95.2 (1)	95.2 (1)	95.2 (1)	95.2 (1)
Pat%	83.3 (1)	81.9 (4)	83.3 (1)	83.3 (1)	80.7 (5)
Shi%	46.2 (2)	45.1 (5)	45.9 (3)	46.6 (1)	45.8 (4)

Table 8
Number of common links and arcs, l/a, between pairs of learned networks, using the LS algorithm and different metrics

	LS + BDeu	LS + BIC	LS + K2
LS + BDeu	17/17	12/10	14/11
LS + BIC	–	13/13	12/11
LS + K2	–	–	27/27

Table 9
Performance measures for the different learned networks, with respect to the training set, using the LS algorithm and different metrics

	KL	K2	BIC	BDeu
LS + K2	2.59 (1)	−99679 (1)	−12855896 (3)	−242665 (3)
LS + BIC	2.43 (3)	−100530 (3)	−233672 (1)	−230545 (2)
LS + BDeu	2.49 (2)	−100241 (2)	−243243 (2)	−229728 (1)

Table 10
Success percentages of classification for Duration and Medical Service, using the LS algorithm and different metrics

Algorithm	Duration (%)	Medical Service (%)
LS + BDeu	91.6	74.5
LS + BIC	91.6	76.1
LS + K2	91.6	76.1

Table 11
Success percentages of classification for the 11 variables, using the LS algorithm and different metrics

	LS + BDeu	LS + BIC	LS + K2
CoA%	88.1	91.9	87.9
CoD%	74.3	74.6	74.3
Cen%	94.3	99.9	94.1
Day%	12.6	12.6	11.9
Dur%	91.4	91.6	91.4
Fin%	93.6	93.7	93.6
Ide%	79.1	82.1	79.0
MS%	81.6	81.8	81.5
P10%	95.2	95.2	95.2
Pat%	83.3	81.9	80.5
Shi%	45.9	45.4	42.3

additional scoring metrics, K2 and BIC, and we have used them within an LS algorithm. We have then carried out the same experiments. The results are showed in Tables 8–11.

From Table 8, we can see that LS + BIC produces a rather sparse network, as expected (see Fig. 3), whereas LS + K2 obtains an extremely dense network.

In Table 9, we can see that each algorithm obtains its best score with the metric used to guide the search process. Globally, LS + BDeu seems to be slightly more robust than the other metrics.

With respect to the classification problems, we should remark that, due to the great complexity of the network obtained by LS + K2, we were not able to perform the propagations.[10] For this reason, the results displayed for LS + K2 refer to a pruned network containing only the first 18 arcs of the original network. The results in Tables 10 and 11 show that the simpler network built using LS + BIC performs much better than LS + BDeu. Nevertheless, BE and BNPC are still preferable for classification purposes.

Observing the results of the experiments, it is quite surprising that the combination of an algorithm such as LS (and it should be remembered that we obtained the same results as LS using more powerful search methods) and the metric BDeu, both of which are very common in the literature, does not obtain good results on this problem. The use of a non-Bayesian metric such as BIC, within the score + search paradigm, improves the results. However, other algorithms, based on independences, such as BNPC (even the classic PC), or the hybrid BENEDICT, perform better. A possible explanation might be based on the following observation: as we mentioned above, there is a variable, Centre, which is functionally dependent on Medical Service in our domain; therefore, Centre is conditionally independent of any other variable given Medical Service. Despite this, all the score + search algorithms include several edges linking Centre with several variables other than Medical Service (whereas in the independence-based algorithms these arcs connect Medical Service with exactly the same variables). Perhaps the problem arises because Medical Service is a variable with 36 cases, whereas Centre only has 3 cases. A conditional independence test (or the independence-based metric used by BENEDICT) can easily detect this independence, but

[10] In order to give an idea of the complexity of this network, the amount of disk space required to store it was 54 Mb, while LS + BDeu required only 27 Kb.

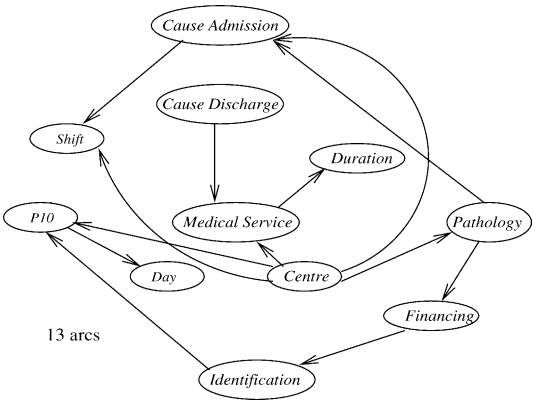

Fig. 3. Structure recovered by the LS + BIC algorithm.

all the metrics considered appear to be rather sensitive to the number of cases of the variables, and penalize edges involving variables with a high number of cases.

In addition to the ability of Bayesian networks to represent available information intelligibly and to make predictions when new data is received, they can also be useful tools for performing specific inference tasks such as those requested by a hospital manager: a network model can be used to compute the posterior probability of any variable in different contexts. In the following experiment, in order to illustrate this possibility, we have calculated the posterior probability distribution of Shift given P10 and Day for all the possible values of these two variables (using the network learned by BE). Table 12 summarizes the results.

It is interesting to note how the pattern of arrival at the emergency medical services is homogeneous for the different days (including the weekend), but this pattern is different according to whether the patient has a P10 document or not (as expected, patients with a P10 document arrive more frequently in the morning); this would allow patient categories to be defined.

With the same process, the manager could study anomalous cases, for instance, the duration of stays longer than 72 h (theses cases represent almost 1% of the database). Given that the variable Duration reaches its greatest value, we have computed the posterior probability distributions of the variables Centre and Medical Service. The Centre involved is almost always the same (with a probability greater than 0.99), and there are basically only two medical units involved (with probabilities of 0.87 and 0.11). Another sign which reveals some kind of anomaly is that in these cases the variable Cause of Discharge takes the value ''indeterminate'' with a probability of 0.88.

Table 12
Posterior distribution of Shift given P10 and Day

Configuration	Morning	Evening	Night
P10 = no for all Day	0.34	0.47	0.18
P10 = yes, Day = Weekday	0.47	0.36	0.17
P10 = yes, Day = Weekend	0.44	0.37	0.18

6. Concluding remarks

Due to the complexity of health-care systems, they should be represented, studied, and optimized with the appropriate tools. Bayesian networks offer a very attractive formalism for representing uncertain knowledge (resulting from the synergy of statistical methods for data analysis and Artificial Intelligence tools) and have successfully been applied in different fields. However, although Bayesian networks have so far only been used in medicine essentially to assist in the diagnosis of disorders and to predict the natural course of disease after treatment (prognosis), we believe that Bayesian networks can also be applied to other management-oriented, medical problems.

What we have presented in this paper is by no means a conclusive document that introduces a mature management decision support system ready to be implemented, but rather a first prototype that would have to be considerably extended and refined in the future. Nevertheless, we believe that our work illustrates the usefulness of Bayesian networks and their technologies for non-diagnostic medical problems.

Our comparative study of several algorithms for learning Bayesian networks using the emergencies dataset has revealed some interesting facts: (1) the widespread belief about the superiority of the scoring-based approach over the independence-based approach is questionable; in our case, the opposite turned to be true; (2) high values of the usual, non-specialized scoring functions do not necessarily result in useful network structures; (3) some non-Bayesian metrics, such as BIC (or the independence-based metric used by BENEDICT) may direct the search process towards network structures that behave more robustly than those obtained by some Bayesian metrics. Although these assertions cannot be generalized without extensive experimentation using many different datasets, previous work on Bayesian network classifiers [13,16] does not contradict our results.

In the future, we plan to extend and refine our model using consensus networks, to include more variables (e.g. seasonal variables or significant temporal periods[11], some additional clinical information (i.e. diagnostic information, number of tests performed on the patients, specific variables for financial control)), to validate it by taking expert knowledge into account, and to use it as a tool to help the hospital manager balance resource allocation. We also plan to apply Bayesian networks to other management medical problems, such as waiting lists (which nowadays are an indicator of the performance of the health service).

Acknowledgements

This work has been supported by the Spanish 'Ministerio de Ciencia y Tecnología' and 'Fondo de Investigación Sanitaria' under projects TIC2001-2973-C05-01 and FIS-PI021147, respectively.

[11] Which may influence the number of arrivals due to traffic accidents, for instance.

References

[1] S. Acid, Métodos de Aprendizaje de Redes de Creencia. Aplicación a la Clasificación. Ph.D. Thesis. Universidad de Granada, Spain, 1999 (in Spanish).

[2] Acid S, de Campos LM. An algorithm for finding minimum *d*-separating sets in belief networks. In: Horvitz E, Jensen F, editors. Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence. San Mateo: Morgan Kaufmann, 1996. p. 3–10.

[3] Acid S, de Campos LM. Benedict: an algorithm for learning probabilistic belief networks. In: Proceedings of the Sixth International Conference on Information Processing and Management of Uncertainty in Knowledge Based Systems, 1996, p. 979–84.

[4] Acid S, de Campos LM. A hybrid methodology for learning belief networks: Benedict. Int J Approx Reason 2001;27:235–62.

[5] Acid S, de Campos LM. An algorithm for learning probabilistic belief networks using minimum *d*-separating sets. DECSAI Technical Report no. 01-02-22. Department of Computer Science and Artificial Intelligence, University of Granada, 2001.

[6] Bouckaert RR. Belief networks construction using the minimum description length principle. In: Clarke M, Kruse R, Moral S, editors. Symbolic and Quantitative Approaches to Reasoning and Uncertainty, Lecture Notes in Computer Science 747. Berlin: Springer-Verlag, 1993. p. 41–8.

[7] Bouckaert RR. Bayesian belief networks: from construction to inference. Ph.D. Thesis. University of Utrecht, 1995.

[8] Buntine W. Theory refinement of Bayesian networks. In: Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence, 1991. p. 52–60.

[9] de Campos LM. Independency relationships and learning algorithms for singly connected networks. J Exp Theor Artif Intell 1998;10:511–49.

[10] de Campos LM, Huete JF. A new approach for learning belief networks using independence criteria. Int J Approx Reason 2000;24:11–37.

[11] de Campos LM, Puerta JM. Stochastic local and distributed search algorithms for learning belief networks. In: Proceedings of the III International Symposium on Adaptive Systems: Evolutionary Computation and Probabilistic Graphical Model, 2001. p. 109–15.

[12] Cheng J, Bell DA, Liu W. An algorithm for Bayesian belief network construction from data. In: Proceedings of AI and STAT'97, 1997. p. 83–90.

[13] Cheng J, Greiner R. Comparing Bayesian network classifiers. In: Laskey K, Prade H, editors. Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. San Mateo: Morgan Kaufmann, 1999. p. 101–7.

[14] Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992;9:309–48.

[15] Dash D, Druzdzel M. A hybrid anytime algorithm for the construction of causal models from sparse data. In: Laskey K, Prade H, editors. Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. San Mateo: Morgan Kaufmann, 1999. p. 142–9.

[16] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997;29:131–61.

[17] Glymour C, Cooper G, editors. Computation, Causation and Discovery. The AAAI Press, 1999.

[18] Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn 1995;20:197–243.

[19] Herskovits E, Cooper GF. Kutató: an entropy-driven system for the construction of probabilistic expert systems from databases. In: Bonissone P, editor. Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence, Cambridge, 1990. p. 54–62.

[20] Kullback S. Information Theory and Statistics. New York: Dover, 1968.

[21] Lam W, Bacchus F. Learning Bayesian belief networks. An approach based on the MDL principle. Comput Intell 1994;10:269–93.

[22] Marshall AH, McClean SI, Shapcott CM, Hastie IR, Millard PH. Developing a Bayesian belief network for the management of geriatric hospital care. Health Care Manage Sci 2001;4:25–30.

[23] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo: Morgan Kaufmann, 1988.

[24] Pearl J. Causality: Models, Reasoning and Inference. Cambridge: Cambridge University Press, 2000.
[25] Pearl J, Verma TS. Equivalence and synthesis of causal models. In: Bonissone P, editor. Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence, Cambridge, 1990. p. 220–27.
[26] Spirtes P, Glymour C, Scheines R. Causation, Prediction and Search, Lecture Notes in Statistics 81. New York: Springer-Verlag, 1993.
[27] Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6:461–4.
[28] Singh M, Valtorta M. Construction of Bayesian network structures from data: a brief survey and an efficient algorithm. Int J Approx Reason 1995;12:111–31.