Expert Knowledge and its Role in Learning Bayesian Networks in Medicine

Peter Lucas
Institute for Information and Computing Sciences
University of Nijmegen
Toernooiveld 1, 6525 ED Nijmegen,
The Netherlands
E-mail: lucas@cs.uu.nl

Abstract

A major part of the medical knowledge concerns diseases that are uncommon or even rare. The uncommon nature of these disorders renders it impossible to collect data of a sufficiently large number of patients to develop machine-learning models that faithfully reflect the subtleties of the domain. An alternative is to develop a Bayesian network with the help of clinical experts. Lack of data is then compensated for by eliciting the structure with its associated local probability distributions from the experts. The resulting network can be subsequently evaluated using the available dataset. One may also consider adopting very strong independence assumptions, such as in naive Bayesian models. Normally not all subtleties of the interactions among the variables in the domain are reflected in such models. Yet, a relatively small dataset may suffice to obtain an acceptably accurate model. This paper explores the trade-offs between modelling using expert knowledge, and machine learning using a small clinical dataset in the context of Bayesian networks.

Keywords: Bayesian networks, machine learning, Bayesian modelling, naive Bayesian model, tree-augmented naive (TAN) Bayesian models, forest-augmented (FAN) Bayesian network models.

1 Introduction

There exists considerable experience in medicine in analysing data of diseases with high prevalence, such as lung cancer and myocardial infarction. Most of the evidence underlying current medical practice is based on such analyses. The frequent occurrence of these disorders renders it practically feasible to collect data of large numbers of patients. Typically, datasets collected in the context of clinical studies of these disorders may include many thousands of patient records. Such datasets are quite attractive for evaluating machine-learning techniques, and, in fact, have been used for that purpose by many researchers.

However, for more than 90% of medical disorders, the picture is quite different: these disorders do only occur occasionally or rarely, and, as a consequence, even clinical research datasets may only include data of a hundred to a few hundred patients. Developing decision-support systems that assist clinicians in handling patients with these disorders is thus scientifically challenging, because there may not be sufficient data available. Systems covering such disorders would be practically useful, as many doctors will lack knowledge and experience to deal with patients affected by uncommon disorders effectively.

Machine-learning literature in medicine not only tends to consider only problems for which much data are available, but in addition it focuses on models capable of performing single tasks, such as classifying patients in particular disease or prognostic categories to assist clinicians with the tasks of diagnosis or treatment selection. However, medical management cannot be captured in terms of single, simple tasks. As a consequence there appears to be a mismatch between common task-specific computer-based models and the complexity of the field of medicine. In contrast to task-specific models, Bayesian networks that reflect the nature of a problem domain have the virtue of being *declarative* models; they can, therefore, be reused for different tasks [13]. They can also be employed to look at particular problems from different angles, just by varying the supplied evidence and the questions posed to the model. However, developing such models is challenging, both in terms of the amount of probabilistic information required and modelling effort. When it comes to handling uncommon or rare disorders, where only small volumes of data will be available, it is unclear whether such efforts are worthwhile. Developing simple linear discriminators like naive Bayesian classifiers and logistic regression models seems a more obvious choice, as it is straightforward and requires datasets with relatively few cases [2, 5, 6, 10].

This paper attempts to provide answers to a number of questions related to the issues mentioned above, based on our experience in developing declarative prognostic models of non-Hodgkin lymphoma of the stomach, gastric NHL for short. Gastric NHL is a typical example of an uncommon, although not extremely rare disease. The following partially overlapping questions are addressed:

- What is the contribution of qualitative and quantitative expert knowledge in building Bayesian classifiers?
- Does additional structure enhance the learning of knowledge?
- Can the structure of a Bayesian network be learnt, either completely or partially, from the data of a small dataset?

The remainder of this paper is organised as follows. In the next two sections, a number of different Bayesian-network models of gastric NHL, developed using different methods, are introduced. Section 4 pays attention to their evaluation, whereas in Section 5 some results of checking the structure of a Bayesian network are presented. The paper is rounded off by a discussion of what has been achieved.

2 Preliminaries

A *Bayesian network* $\mathcal{B}$ is defined as a pair $\mathcal{B} = (G, \Pr)$, where G is a directed, acyclic graph $G = (V(G), A(G))$, with a set of vertices $V(G) = \{V_1, \ldots, V_n\}$, representing a set of stochastic variables $\mathcal{V}$, and a set of arcs $A(G) \subseteq V(G) \times V(G)$, representing conditional and unconditional stochastic independences among the variables, modelled by the absence of arcs among vertices [12]. The basic property of a Bayesian network is that any variable corresponding to a vertex in the graph G is conditionally independent of its non-descendants given its parents; this is called the *local Markov property*. On the variables $\mathcal{V}$ is defined a joint probability distribution $\Pr(V_1, \ldots, V_n)$, for which, as a consequence of the local Markov

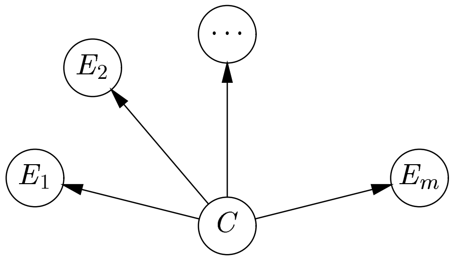

Figure 1: Independent-form Bayesian network.

property, the following decomposition holds:

$$\Pr(V_1, \ldots, V_n) = \prod_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$$

where $\pi(V_i)$ denotes the conjunction of variables corresponding to the parents of V_i, for $i = 1, \ldots, n$. In the following, variables will be denoted by upper-case letters, e.g. V, whereas a variable V which takes on a value v, i.e. $V = v$, will be abbreviated to v. A set (interpreted as a conjunction) of variables to which values are assigned, will be called a *configuration*. When it is not necessary to refer to specific values of variables, we will usually just refer to a variable, which thus stands for any value of the variable.

Bayesian-network models conforming to the topology shown in Fig. 1 correspond to the situation where a distinction is made between *evidence variables* E_i and a *class variable* C, with the evidence variables assumed to be conditionally independent given the class variable. In the following such a network will be called an *independent-form Bayesian network*, by way of analogy with the special form of Bayes' rule, called its independent form, for which the same assumptions hold. The independent form is used to compute the a posteriori probability of a class value c_k given the evidence $\mathcal{E} \subseteq \{e_1, \ldots, e_m\}$ [12]:

$$\Pr(c_k \mid \mathcal{E}) = \frac{\Pr(\mathcal{E} \mid c_k)\Pr(c_k)}{\Pr(\mathcal{E})} = \frac{\prod_{e \in \mathcal{E}} \Pr(e \mid c_k)\Pr(c_k)}{\sum_{j=1}^{q} \prod_{e \in \mathcal{E}} \Pr(e \mid c_j)\Pr(c_j)}$$

where class variable C has q mutually exclusive values c_k, and $\Pr(\mathcal{E}) > 0$. The model is normally used for classification of cases, i.e. as a *classifier*. Note that $\Pr(\mathcal{E} \mid c_k) = \prod_{e \in \mathcal{E}} \Pr(e \mid c_k)$ holds, because of the assumption that the evidence variables E_i are conditionally independent given the class variable C. Furthermore, $\Pr(\mathcal{E}) = \sum_{j=1}^{q} \Pr(\mathcal{E} \mid c_j)\Pr(c_j)$ follows from marginalisation and conditioning.

An independent-form Bayesian network may lack important probabilistic dependence information, hence its nicknames: *naive Bayesian model* and even *idiot's Bayes*. However, the model has the advantage that the assessment of the required probabilities $\Pr(E_j \mid C)$ and $\Pr(C)$ is straightforward, and can be carried out with a relatively small dataset. Determination of the a posteriori probabilities $\Pr(C \mid \mathcal{E})$ is computationally speaking a trivial task under the mentioned assumptions. One would expect that adopting such strong simplifying assumptions may be at the expense of reduced performance. However, Domingos and Pazzani have convincingly shown that the independent form yields a surprisingly powerful classifier, much more robust than previously thought [5]. These features of the independent form of Bayes' rule explain why it is becoming increasingly popular, after having fallen into disgrace two decades ago.

Even though independent-form Bayesian networks appear more powerful than previously thought, one can imagine that adding some information regarding dependences among variables may in some cases yield a better classifier. This is indeed the case, as was shown by Friedman et al. [6], building upon work from the late 1960s by Chow and Lui [3]. When the evidence variables are linked together as a directed tree, and these variables are then connected to the class variable as in an independent-form network, the resulting network is called a *tree-augmented naive (TAN) Bayesian network*, or TAN for short (see Fig. 2). Chow and Lui's method uses a minimum-cost spanning tree algorithm [1] to construct the tree based on data [3]. The measure of cost that is used in selecting branches in constructing the minimum-cost spanning tree is the negative value of the mutual information between variables E_i, E_j, $i \neq j$: $I_{\mathrm{Pr}}(E_i, E_j)$, where Pr is the probability distribution corresponding to the data. Friedman et al. suggest using instead mutual information between variables E_i, E_j *conditioned* on the class variable C [6]:

$$I_{\mathrm{Pr}}(E_i, E_j \mid C) = \sum_{E_i, E_j, C} \mathrm{Pr}(E_i, E_j, C) \log \frac{\mathrm{Pr}(E_i, E_j \mid C)}{\mathrm{Pr}(E_i \mid C)\,\mathrm{Pr}(E_j \mid C)}$$

which offers advantages when focusing on building a classifier, as the conditional mutual information takes class influences into account. They demonstrated that TANs often dominate their associated independent form in performance [6].

The algorithm to construct TAN models used in this paper is based on the modification by Friedman et al. of the Chow and Lui algorithm [6]. It consists of the following steps:

(1) The conditional mutual information $I_{\mathrm{Pr}}(E_i, E_j \mid C)$ for all pairs of evidence variables E_i, E_j, $i \neq j$, are computed. We used an algorithm for the generation of all non-equivalent three-vertex Bayesian networks with associated joint probability distribution of the form depicted in Fig. 4 for a given dataset, class variable and a set of evidence variables. There are $\binom{m}{2}$ of such networks, where m is the number of evidence variables. The probability distributions $\mathrm{Pr}(E_k \mid C)$ for evidence variables E_k and class variable C were obtained from the independent-form network shown in Fig. 1, again after training the network with data.

(2) An undirected graph with vertices E_i, $i = 1, \ldots, m$, and C is build, with costs attached to the edges defined by $-I_{\mathrm{Pr}}(E_i, E_j \mid C)$.

(3) A minimum-cost spanning tree for the undirected cost graph is constructed [1].

(4) The undirected tree is transformed into a directed tree by choosing a root vertex, and by adding an outward direction to the branches encountered on the paths from the root to every other vertex.

(5) The directed tree is transformed into a TAN model by adding an arc (directed edge) from the class vertex C to every evidence vertex E_i.

(6) The conditional probability distributions of the TAN model are learnt from the data in the dataset.

In this paper, we compare the various Bayesian-network forms, where the availability of training data versus the availability of expert knowledge constitutes the spectrum of possibilities to be explored. The context of the research is machine learning with a small dataset.

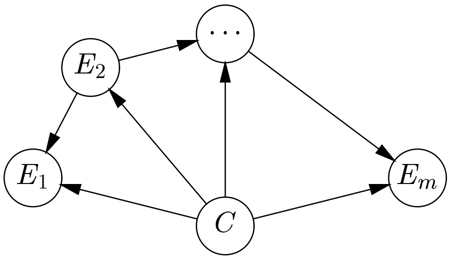

Figure 2: Tree-augmented naive (TAN) Bayesian network.

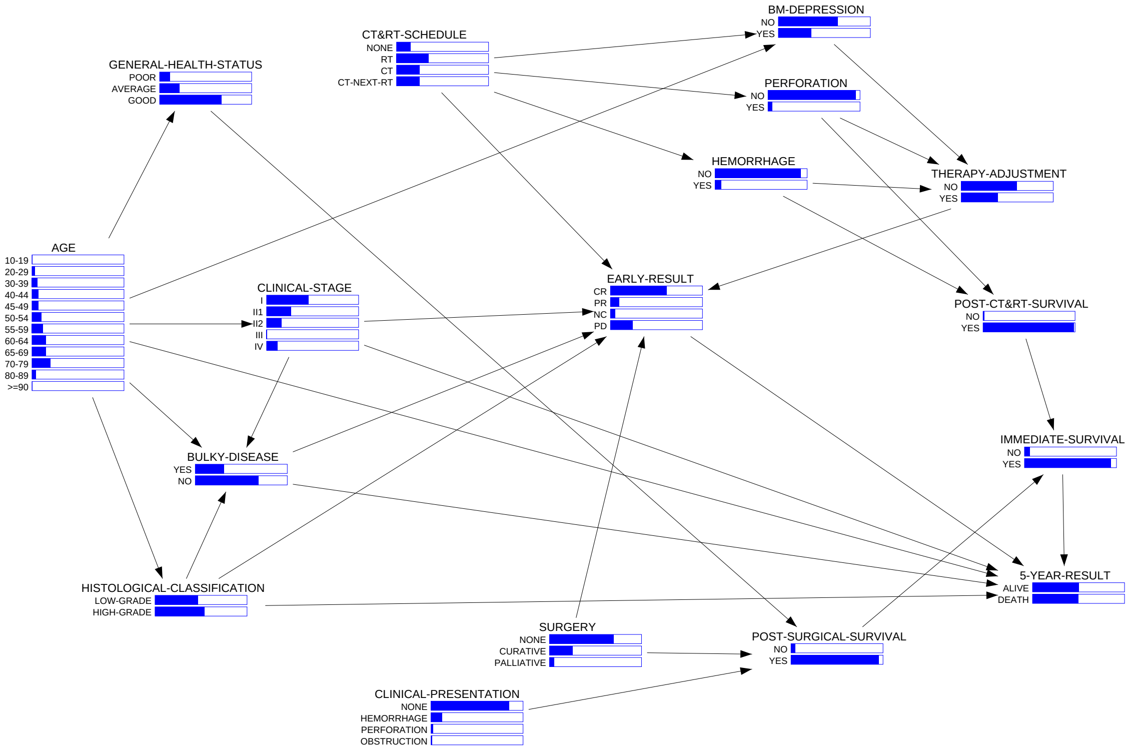

Figure 3: Bayesian-network model as designed with the help of expert clinical oncologists.

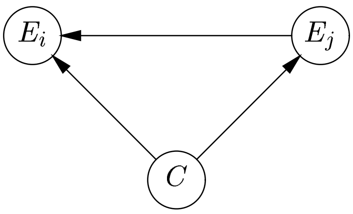

Figure 4: Three-vertex Bayesian network, used in computing the joint probability distributions $\Pr(E_i, E_j, C) = \Pr(E_i \mid E_j, C) \Pr(E_j \mid C) \Pr(C)$ and $\Pr(E_i, E_j \mid C) = \Pr(E_i \mid E_j, C) \Pr(E_j \mid C)$.

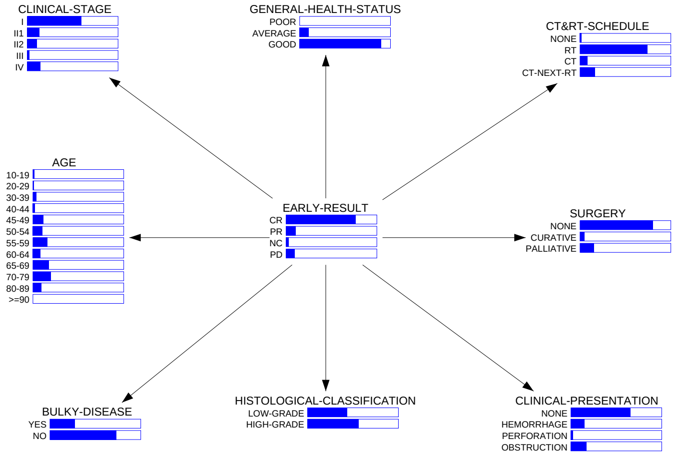

Figure 5: Independent-form models for EARLY-RESULT

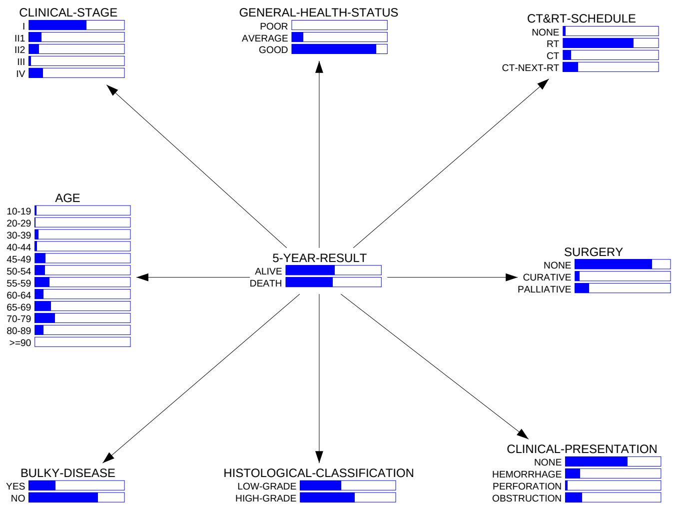

Figure 6: Independent-form models for 5-YEAR-RESULT

3 Bayesian models of gastric non-Hodgkin lymphoma

In the following, a number of different Bayesian models for gastric NHL are described. These models are analysed in Section 4 to shed some light on the issues mentioned above.

3.1 The models

A Bayesian-network model incorporating most factors relevant for the management of gastric NHL was developed in collaboration with clinical experts from the Netherlands Cancer Institute (NKI). The resulting model, shown in Fig. 3, includes variables like age of the patient and clinical stage of the tumour. Some of the included variables concern patient information obtained from diagnostic procedures, which will be available prior to the selection of treatment. We shall refer to this information as *pretreatment* information. Another part of the model variables will only become available following treatment; examples are EARLY-RESULT and

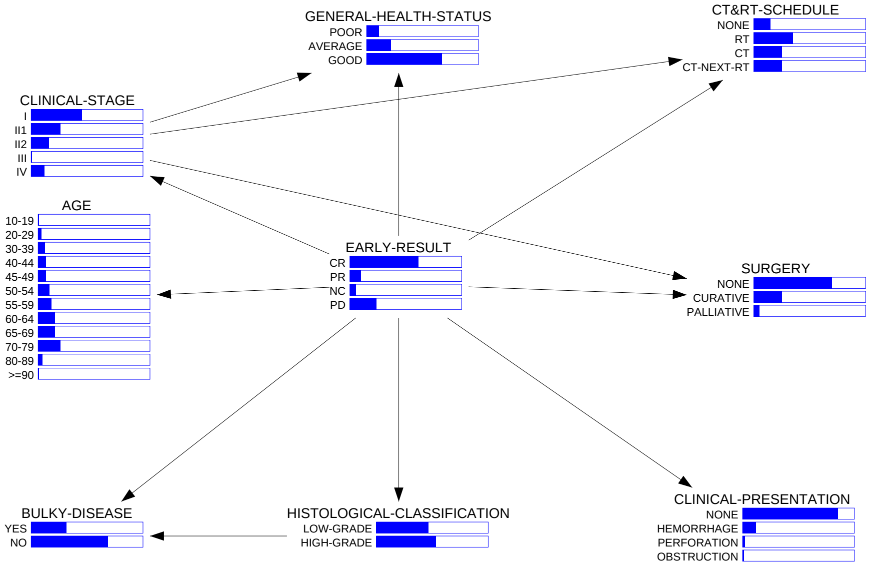

Figure 7: FAN model for EARLY-RESULT, derived from the structured model in Figure 3.

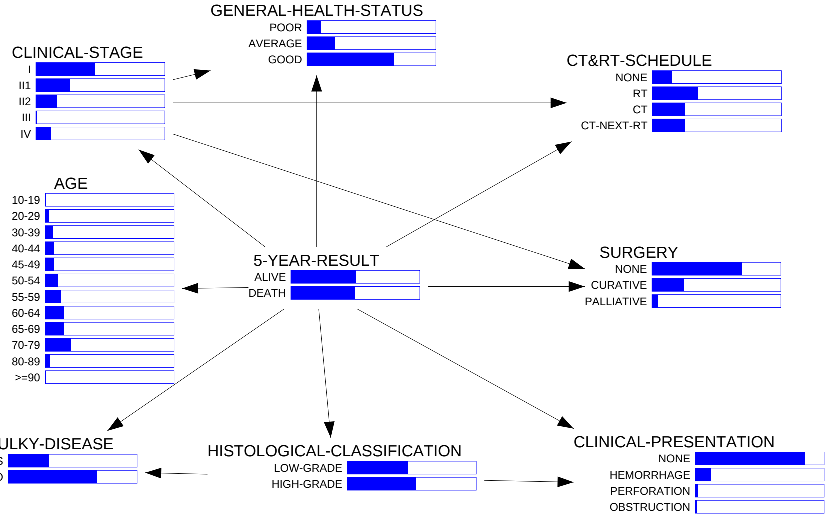

Figure 8: FAN model for 5-YEAR-RESULT, derived from the structured model in Figure 3.

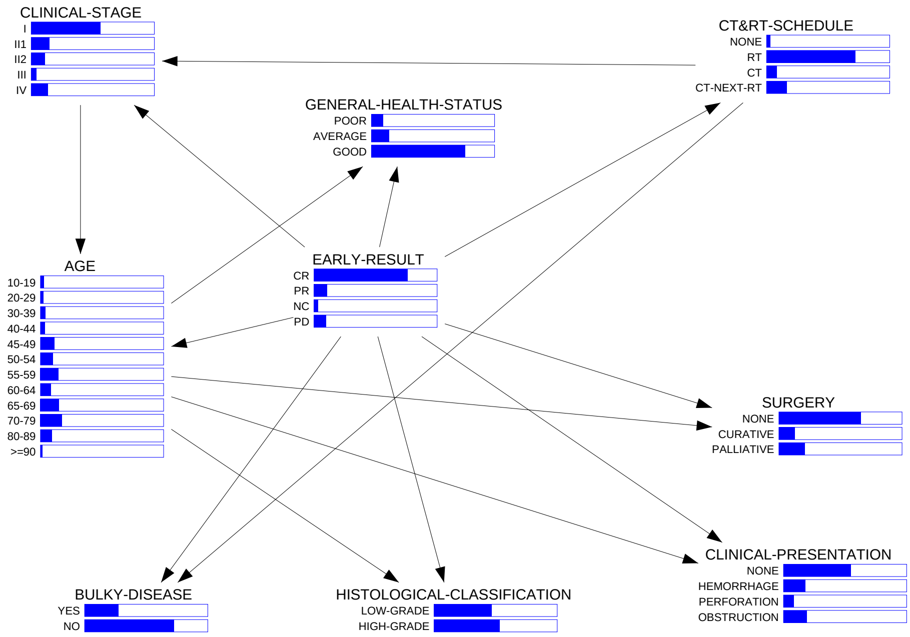

Figure 9: TAN model for EARLY-RESULT learnt from data.

5-YEAR-RESULT. It is called the *posttreatment* part of the model. Finally, the model includes the *treatment* variables SURGERY and CT&RT-SCHEDULE (schedule for treatment by either chemotherapy or radiotherapy, or by chemotherapy followed by radiotherapy). Note that this declarative model of gastric NHL can be deployed for different tasks, such as the prediction of prognosis, treatment selection – possibly using utility information yielding a decision network or influence diagram [19, 9] – and the generation of patient profiles [14]. For example, the model can be used to predict the early result and likelihood of 5-year survival after treatment by supplying patient information concerning age, general health status, clinical stage and histological features of the tumour, whether or not bulky disease (a tumour with maximum diameter more than 10 cm) is present, and whether the patient presented at the hospital with complications due to the tumour such as gastric hemorrhage. On the other hand, a user can extract detailed probabilistic information about the characteristics of patients in a subgroup of patients using the same Bayesian-network model, for example by filling in information about the subgroup of patients who have died before 5 years after treatment while in stage I of the disease, and obtaining radiotherapy as the treatment of choice. Finally, by varying over all possible (combinations) of treatments, a physician can select the one offering the 'best' results, possibly taking side-effects into account. Hence, a declarative model such as the gastric NHL model is in comparison to independent-form Bayesian networks and TAN models really very versatile.

Two independent-form Bayesian networks, depicted in figures 5 and 6, were derived from the network shown in Fig. 3. In one of these, the variable 5-YEAR-RESULT was taken as a class variable; the other includes EARLY-RESULT as a class variable. Their probability distributions were also derived from that of the complete Bayesian network. In both models, all pretreatment and treatment variables, and a single posttreatment variable were included. As the other posttreatment variables would typically be unknown, it would not make sense to

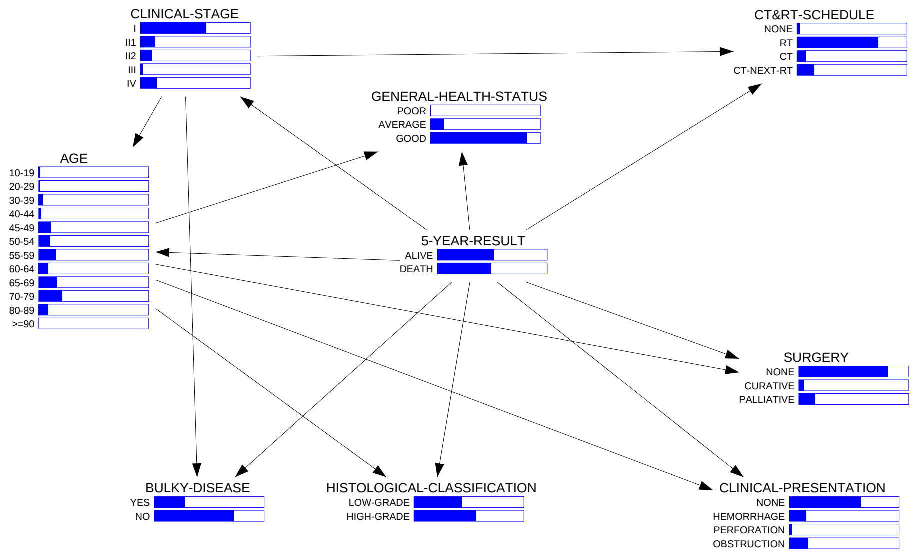

Figure 10: TAN model for 5-YEAR-RESULT learnt from data.

include them as well in networks typically used for classifying patients into groups defined by the values of the class variable. This illustrates one additional difference between a declarative Bayesian model, as shown in Fig. 3, and task-specific models, as shown in figures 5 and 6.

Two TAN-like models were derived as well by taking all evidence variables, and linking two variables together when they were assumed to be dependent according to medical literature. Causality was used as a heuristic to decide on the direction of arcs. The topology of the resulting graphs was restricted to forests of trees. Since not all evidence variables were therefore linked together, the resulting networks are not TANs in the sense of Chow and Lui [3], but nonetheless closely related. We call them *forest-augmented naive Bayesian* (FAN) networks. One was developed by taking 5-YEAR-RESULT as a class variable; the resulting Bayesian network is shown in Fig. 8. The other concerned the class variable EARLY-RESULT, and is shown in Fig. 7. Note the two networks have slightly different forests.

Finally, two TAN models (one regarding 5-YEAR-RESULT and the other one concerning EARLY-RESULT) were learnt using the TAN-construction algorithm described in the previous section. For this purpose data of 137 gastric NHL patients with known 5-year result were employed; for 133 of these patients the early result was known as well. A more detailed description of the data and the learning method is given in Section 3.2. The learnt TAN models are shown in figures 10 and 9, respectively.

A brief description of the individual models is given in Table 1; they are ordered from top to bottom in accordance with decreasing complexity. Models with the letter S in their name are declarative or structured models; models with the letter I in their name are independent-form Bayesian networks. The models with the letter F are forest-augmented network models, whereas the models with the letter T are tree-augmented network models. The superscript *er* indicates that EARLY-RESULT was the class variable; the superscript *5y* is used to denote models with 5-YEAR-RESULT as a class variable. The subscript E indicates that a model's underlying joint probability distribution is based on a subjective probability assessment by an expert.

Table 1: Bayesian models.

Model	Description	Topology	Missing Data
S_E	expert-assessed model	declarative	—
S	learnt model	declarative	ignored
S_u	learnt model	declarative	uniform distribution
T^{er}	learnt model	TAN form	ignored
T_u^{er}	learnt model	TAN form	uniform distribution
T^{5y}	learnt model	TAN form	ignored
T_u^{5y}	learnt model	TAN form	uniform distribution
F_E^{er}	derived from S_E	FAN form	—
F^{er}	learnt model	FAN form	ignored
F_u^{er}	learnt model	FAN form	uniform distribution
F_E^{5y}	derived from S_E	FAN form	—
F^{5y}	learnt model	FAN form	ignored
F_u^{5y}	learnt model	FAN form	uniform distribution
I_E^{er}	derived from S_E	independent form	—
I^{er}	learnt model	independent form	ignored
I_u^{er}	learnt model	independent form	uniform distribution
I_E^{5y}	derived from S_E	independent form	—
I^{5y}	learnt model	independent form	ignored
I_u^{5y}	learnt model	independent form	uniform distribution

3.2 Patient data and learning strategy

Taking the topologies of the seven networks (S, I^{5y}, I^{er}, F^{5y}, F^{er}, T^{5y}, T^{er}) as fixed, the probability distributions of these models were learnt. Learning and evaluation took place using a dataset with patient data from the Netherlands Cancer Institute, comprising 137 cases, with some missing data. Since in 4 patients the early result of treatment was unknown, we used only 133 cases when studying the performance of the models concerning the class variable EARLY-RESULT. See Table 2 for a description of the characteristics of the dataset. The joint probability distributions of the various Bayesian networks were learnt using Bayesian updating with Dirichlet priors [2, 8, 10, 11]. Thus, the conditional probability distribution for each variable V_i was computed as the weighted average of a probability estimate and the Dirichlet prior, as follows:

$$\Pr(V_i \mid \pi(V_i), D) = \frac{N}{N+N_0}\widehat{\Pr}_D(V_i \mid \pi(V_i)) + \frac{N_0}{N+N_0}\Theta_i$$

where $\widehat{\Pr}_D$ is the probability distribution estimate based on a given dataset D, and Θ_i is the Dirichlet prior. We choose Θ_i to be a uniform probability distribution. Furthermore, N_0 is equal to the number of past cases on which the contribution of Θ_i is based, and $N = |D| = 137$ is the size of the dataset. When there were no cases at all in the dataset for any configuration of the variable V_i given a configuration of its parents $\pi(V_i)$, a uniform probability distribution was assumed.

The small size of the gastric NHL dataset made it necessary to assign a very low weight to the prior probability distribution Θ_i. After some experimentation, the results of which are shown in Fig. 11, we took as a value $N_0 = 1$. This value does not yield the best performance for the models, but any value larger than 10 would have been difficult to justify given the size of the dataset and the nature of the domain. The optimal choice of the weight is a function of both the structure of the model and the size of the dataset. This may give rise to different Bayesian updating schemes for different models, rendering it more difficult to explain differences among results. This also explains why the same value of the parameter N_0 was used for different models.

Missing data were either ignored or distributed uniformly among the values of the variable for which a value was missing, indicated by the subscript u in the name of the models described in Table 1.

4 Evaluation

Two evaluation procedures were performed. In the first one, when probability distributions were learnt from the dataset (the training set), the same set was used for testing purposes. As this normally yields results which are too optimistic, cross validation was carried out as well. An advantage of this two-way approach is that it yields insights into the risks of overfitting for particular network topologies.

4.1 Methods

The a posteriori probability distributions $\Pr(\textsc{early-result} \mid \mathcal{E})$ and $\Pr(5\textsc{-year-result} \mid \mathcal{E})$ were computed, where $\mathcal{E}$ was the available evidence for each of the 133 and 137, respectively, patients with gastric NHL, restricted to values of all pretreatment and treatment variables. We did not carry out ROC (receiver-operating characteristics) analyses to determine the optimal classification thresholds, as part of our models were expert-based, which would not allow for fitting thresholds to training data and then investigating the results for a test set. The small size of the dataset also renders this approach less useful. Instead, the value c^* of the class variable C with largest probability was taken as the resulting classification:

$$c^* = \text{argmax}_c\{\Pr(C = v \mid \mathcal{E})\}$$

However, a disadvantage of this straightforward method of comparing the quality of the Bayesian models is that the actual a posteriori probabilities are ignored. A more precise indication of the behaviour of the Bayesian models is obtained when the resulting probabilities are taken into account as well. For example, if a patient Q is known to have survived more than 5 years following radiotherapy, and a Bayesian model has predicted this event with

Table 2: Characteristics of the datasets.

Class	**Patients (n)**	**Class**	**Patients (n)**
5-YEAR-RESULT		EARLY-RESULT	
alive	70	CR	102
death	67	PR	14
		NC	4
		PD	13
Total	137	Total	133

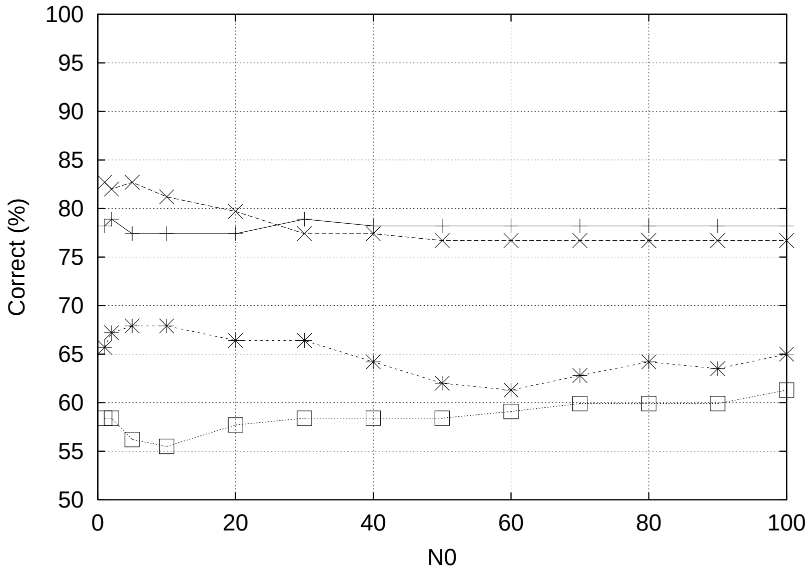

Figure 11: Relationship between values of the parameter N_0 and cross-validation performance results using the leave-one-out method; $+$: results for the independent-form network regarding EARLY-RESULT; $\times$: results for the TAN network regarding EARLY-RESULT: $\star$: results for the independent-form network regarding 5-YEAR-RESULT; $\Box$: results for the TAN network regarding 5-YEAR-RESULT.

probability 0.8, this conclusion would seem intuitively better than the conclusion of another model which predicted this event with a probability of 0.6. One of the simplest scoring rules that measures this effect is the *logarithmic scoring rule* defined below [4].

Let D be a dataset, $|D| = p$, $p \geq 0$. With each prediction generated by a Bayesian model for case $r_k \in D$, with actual class value c_k, we associated a score which can be interpreted formally as the entropy [2]:

$$S_k = -\ln \Pr(c_k \mid \mathcal{E})$$

which has the informal meaning of a penalty: when the probability $\Pr(c_k \mid \mathcal{E}) = 1$, then $S_k = 0$ (actually observing c_k generates no information); otherwise, $S_k > 0$. The total score for dataset D is now defined as the sum of the individual scores:

$$S = \sum_{k=1}^{p} S_k$$

Since S is a stochastic quantity, it can be characterised further by means of central moments, such as the average E_k:

$$E_k = -\sum_{i=1}^{q} \Pr(c_i \mid \mathcal{E}) \ln \Pr(c_i \mid \mathcal{E})$$

where q is the number of different values for class variable C, and total average

$$E = \sum_{k=1}^{p} E_k$$

and the variance V_k:

$$V_k = -\sum_{i=1}^{q} \Pr(c_i \mid \mathcal{E})(\ln \Pr(c_i \mid \mathcal{E}))^2 - E_k^2$$

with total variance

$$V = \sum_{k=1}^{p} V_k$$

A limitation of the logarithmic scoring rule is that the quantities S_k, E_k and V_k are undefined for $\Pr(c_k \mid \mathcal{E}) = 0$.

Three additive components make up the error in classifying: (1) the intrinsic error due to noise in the data, (2) the statistical bias in the model, and (3) the variance (model's sensitivity to the characteristics of the dataset) [2, 5, 10]. One would expect a large statistical bias for the naive Bayesian classifier, as its independence assumptions are almost always unjustified, and a somewhat lower one for FAN and TAN models. On the other hand, experimental evidence shows that the naive Bayesian classifier has a low variance [5]. Methods with a greater representational power have a greater ability to respond to the dataset, i.e. they have a large *information-storage capacity*, and tend to have a lower bias and higher variance [2]. These properties one would expect to hold for the declarative Bayesian network and somewhat less so for the FAN and TAN models [6].

4.2 Upper bound to performance and lower bound to logarithmic scores

As an initial experiment, the performance of the various models was investigated by determining predictions for the patient data, either for the expert-based models (S_E, I_E^{er}, I_E^{5y}, F_E^{er} and F_E^{5y}) without learning, or after learning the joint probability distribution from the data. The results obtained for the nineteen models (three S models, six I models, six F models and four T models, with for the I, F and T models separate models for the class variables EARLY-RESULT and 5-YEAR-RESULT) are shown in Table 3. Note that for the trained models, training and performance evaluation was accomplished using the same dataset, with the exclusion of four patient cases when evaluating the results for the variable EARLY-RESULT, as mentioned above.

Prediction of 5-year survival is medically speaking more difficult than predicting the early result of treatment. Note that according to Table 2, the two classes for 5-YEAR-RESULT are approximately uniformly distributed, i.e. the prior probability of correct classification is about 50%, with a lower bound of about 0%. The distribution of EARLY-RESULT, however, is skewed towards CR (complete remission) with an upper bound of 76.7% and a lower bound of 53.4% when using a random classification scheme that respects the frequency of occurrence of classes in the dataset. This is reflected in the results obtained by the models, in the sense that the percentage of correctly classified cases is higher for the class variable EARLY-RESULT than for the variable 5-YEAR-RESULT. The performance of the declarative model S_E, with its associated derived independent Bayesian models I_E^{er} and I_E^{5y} and FAN models F_E^{er} and F_E^{5y}, which incorporated expert-assessed probabilities, were lowest. As can be seen, when deriving probability distributions from expert-based assessments, keeping some of the dependence information from a declarative model in a FAN model gives rise to performance

Table 3: Results for different Bayesian models for EARLY-RESULT and 5-YEAR-RESULT. Percentages were computed by dividing the number of correct conclusions by the number of classified cases. The underlying probabilistic information in models was either based on expert knowledge, or learnt from data. In the latter case, the training and test set were the same (except that 4 patients were removed from the test set for EARLY-RESULT).

Model	Total	Classified (n)			Missing Data
		Incorrect	Correct	(%)	
		EARLY-RESULT			
S_E	133	31	102	(76.7)	—
S	133	23	110	(82.7)	ignored
S_u	133	17	116	(87.7)	distributed among values
T^{er}	133	13	120	(90.2)	ignored
T^{er}_u	133	13	120	(90.2)	distributed among values
F^{er}_E	133	31	102	(76.7)	—
F^{er}	133	19	114	(85.7)	ignored
F^{er}_u	133	19	114	(85.7)	distributed among values
I^{er}_E	133	35	98	(73.7)	—
I^{er}	133	24	109	(82.0)	ignored
I^{er}_u	133	24	109	(82.0)	distributed among values
		5-YEAR-RESULT			
S_E	137	43	94	(68.6)	—
S	137	22	115	(83.9)	ignored
S_u	137	20	117	(85.4)	distributed among values
T^{5y}	137	26	111	(81.0)	ignored
T^{5y}_u	137	25	111	(81.0)	distributed among values
F^{5y}_E	137	44	93	(67.9)	—
F^{5y}	137	35	102	(74.5)	ignored
F^{5y}_u	137	31	106	(77.4)	distributed among values
I^{5y}_E	137	40	97	(70.8)	—
I^{5y}	137	38	99	(72.3)	ignored
I^{5y}_u	137	39	98	(71.5)	distributed among values

Table 4: Lower bound to logarithmic scores of evaluation results: EARLY-RESULT (a), and 5-YEAR-RESULT (b); (a) Models S_E and F_E have no scores for the variable EARLY-RESULT, as for some cases the probability of the class value was equal to 0.

Model	S	E	V
	EARLY-RESULT		
S_E	–	–	–
S	82.69	120.58	78.36
S_u	75.23	128.06	83.65
T^{er}	35.08	35.16	51.35
T_u^{er}	37.63	41.98	60.50
F_E^{er}	–	–	–
F^{er}	50.57	44.77	63.33
F_u^{er}	52.75	51.57	68.93
I_E^{er}	92.66	114.29	85.75
I^{er}	59.49	53.02	64.75
I_u^{er}	60.71	60.68	72.74

(a)

Model	S	E	V
	5-YEAR-RESULT		
S_E	81.28	71.47	30.65
S	69.30	85.65	14.62
S_u	66.70	84.17	17.11
T^{5y}	47.77	46.25	30.89
T_u^{5y}	48.26	46.15	30.94
F_E^{5y}	80.28	72.14	31.48
F^{5y}	58.02	55.16	34.00
F_u^{5y}	59.80	56.56	35.02
I_E^{5y}	75.43	73.91	29.91
I^{5y}	67.56	52.91	31.71
I_u^{5y}	69.43	55.02	33.35

(b)

results that are closer to those for the declarative model than those for the independent-form networks.

For the class variable EARLY-RESULT, the performance of the F and T models were better, for the T models even much better, than for the declarative model. The results for the I models were close to those for the S models. It appears that the T models, and less so the F models, can easily adapt to the skewed nature of the EARLY-RESULT data; they have more difficulty in dealing with class data that are uniformly distributed, and then the results of the S models are the best. Thus, the Bayesian networks S and S_u, which have the same topology as S_E, but with joint probability distributions learnt from data, yielded the best results for the class variable 5-YEAR-RESULT. As might be expected, the expert-based models were inferior to the other models due to the subjective nature of their underlying probability distributions. However, at least part of the good results for the other models can be explained by overfitting, as will be discussed in Sections 4.3 and 4.4.

The various independent-form Bayesian networks yielded results that were always inferior to those of the two trained declarative models S and S_u, as well as those of the trained FAN and TAN models. The Bayesian models in which missing values were uniformly distributed among values yielded sometimes superior results than the other models, but sometimes inferior results as well. It only resulted in a systematic improvement for the declarative model. The logarithmic scores shown in Table 4(a)-(b) offer more detailed insight into this phenomenon, as for the variable EARLY-RESULT the S scores were always lower for the models in which missing data were handled, with the exception of the results for the models I and I_u. However, the scores of models in which missing data are handled and those in which no missing-data mechanism was in place were usually quite close. The TAN models clearly fitted the data best, although for 5-YEAR-RESULT the S model yielded a slightly better performance.

Finally note that the performance of the declarative model was for 5-YEAR-RESULT sometimes only slightly less than for EARLY-RESULT. This may be taken as evidence that lack of

Table 5: Cross-validation results for different Bayesian models for EARLY-RESULT and 5-YEAR-RESULT using the leave-one-out method. Only correctness percentages are given.

Model	Total	Classified (n) Incorrect	Correct	(%)	Missing Data
		EARLY-RESULT			
S	133	27	106	(79.7)	ignored
S_u	133	21	112	(84.2)	distributed among values
T^{er}	133	23	110	(82.7)	ignored
T_u^{er}	133	24	109	(82.0)	distributed among values
F^{er}	133	25	108	(81.2)	ignored
F_u^{er}	133	24	109	(82.0)	distributed among values
I^{er}	133	29	104	(78.2)	ignored
I_u^{er}	133	30	103	(77.4)	distributed among values
		5-YEAR-RESULT			
S	137	54	83	(60.6)	ignored
S_u	137	52	85	(62.0)	distributed among values
T^{5y}	137	57	80	(58.4)	ignored
T_u^{5y}	137	54	83	(60.6)	distributed among values
F^{5y}	137	48	89	(65.0)	ignored
F_u^{5y}	137	45	92	(67.2)	distributed among values
I^{5y}	137	47	90	(65.5)	ignored
I_u^{5y}	137	48	89	(65.0)	distributed among values

data in relationship to the density of the model gave rise to networks with almost identical joint probability distribution. This made the model rather insensitive to change when using another evaluation scheme, such as cross validation, based on the same dataset.

4.3 Cross validation

After excluding the five expert-based models, for fourteen of the models cross validation was carried out using the leave-one-out method [2, 10]. The results are given in tables 5 and 6.

The most significant result was a drop in performance of the declarative Bayesian networks, in comparison to the upper-bound results discussed in the previous section, between 4-28%, whereas the drop in performance of the independent form models was only between 5-9%. The drop in performance of the FAN models was between 4-13%. Finally, the drop in performance of the TAN models was also large, between 8-28%. It seems that the declarative and TAN models have difficulty in dealing with uniformly distributed data.

Finally, note that by adding structure to an independent-form model, the S score decreased for the variable EARLY-RESULT, but the S score increased for the cross-validation results for variable 5-YEAR-RESULT. It appears that the latter effect is caused by overfitting in producing the upper-bound results. Dealing with the missing data had again only effect on the performance of the declarative model. Finally, note that the performance of the expert-based models S_E, I_E^{er} and I_E^{5y} is again inferior to that of the other models for the variable EARLY-RESULT, but for the variable 5-YEAR-RESULT it is between the upper- and lower-bound and cross-validation results. This is a strong indication that the form of the prior probability distribution for each class variables had a major effect on the results.

4.4 Comparison

The performance results are summarised in figures 12, 13, 14 and 15. Based on these figures, the following conclusions can be drawn. When comparing the results when the training and test datasets were the same with the cross-valitation results, it appears that there were various amounts of overfitting. In particular, there was significant overfitting for the TAN networks, irrespective of the class variable. Rather surprisingly, overfitting was not as much of a problem for the declarative S models for the variable EARLY-RESULT. However, for the S models only a small part of the probability tables is actually based on gastric NHL data. This is also reflected in the closeness of the logarithmic scores. However, the value CR for the variable 5-YEAR-RESULT is so dominantly present in the dataset that this allowed the declarative model to learn the data adequately. The leave-one-out method was thus unable to discover overfitting in this case, as merely leaving out one case was not sufficient to change the model's classification behaviour. There was, however, much overfitting when there is no skewness in the data, which is the case for the variable 5-YEAR-RESULT.

The balance between the density of the topology of a Bayesian network, the skewness of the underlying probability distribution for a class variable, and the size of the dataset seems to be the most important factor determining the performance. The independent-form and FAN networks are sparse and thus do not need much data. Independent-form networks are known to yield robust classifiers [5, 18], and given the fact that the FAN models are only slightly more complex, one expects at least a similar and sometimes a better performance.

The large drop of performance of the structured models in comparison to the independent-form models may be explained by the larger information-storage capacity of these structured models. This phenomenon is in particular significant when training and test sets are the same. The performance of the FAN models, although dropped more than that of the independent-form models, was still comparable or better than that of the independent-form networks. Hence, adding a little bit of structure appears to help, whereas adding much structure may have negative effect when only a small dataset is given.

A surprising result is that the networks with joint probability distribution based on subjective assessments performed generally well for the class variable 5-YEAR-RESULT. This was not the case for the class variable EARLY-RESULT. Even though the marginal probability distribution of the variable EARLY-RESULT in the S_E model offered a more accurate estimate of the clinical situation, the major impact of the many uniform conditional probability distributions in the S and S_u models made it easier to diagnose other values for EARLY-RESULT than CR (complete remission). So, it appears that an inaccurate prior probability distribution can actually yield good results when in real life the frequencies of occurrence of certain events are skewed.

5 Checking the topology

Experiments with a network topology selection algorithm were carried out in order to obtain more insight into how far one gets with such algorithms, given a small uncommon disease dataset. Use was made of the Bayesian Knowledge Discoverer system (BKD), which implements a heuristic search method for finding the Bayesian-network structure G that best fits the data D according to the ratio $\Pr(D,G)/\Pr(D,G')$, where G' is an alternative structure

Table 6: Cross-validation results: logarithmic scores.

Model	S	E	V
	EARLY-RESULT		
S	92.73	122.76	75.10
S_u	88.43	131.70	78.13
T^{er}	70.85	36.38	47.60
T^{er}_u	66.98	43.39	56.17
F^{er}	71.18	46.76	59.58
F^{er}_u	70.41	51.91	63.99
I^{er}	72.92	52.81	61.32
I^{er}_u	73.05	60.55	69.06

(b)

Model	S	E	V
	5-YEAR-RESULT		
S	90.91	86.96	12.52
S_u	89.38	86.25	13.82
T^{5y}	116.29	53.96	33.09
T^{5y}_u	118.64	51.95	34.04
F^{5y}	94.55	58.34	35.37
F^{5y}_u	95.83	60.07	34.88
I^{5y}	86.85	54.96	31.84
I^{5y}_u	88.89	57.43	33.01

(a)

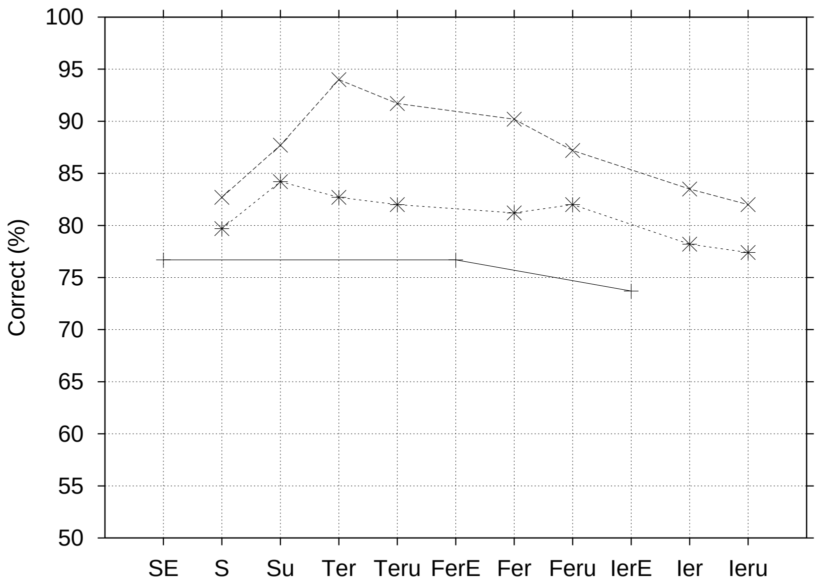

Figure 12: Comparison of upper-bound and cross-validation results for the variable EARLY-RESULT; $\times$: upper-bound results for networks with learnt joint probability distribution; $+$: results for networks with underlying subjective joint probability distribution; $\star$: cross-validation results.

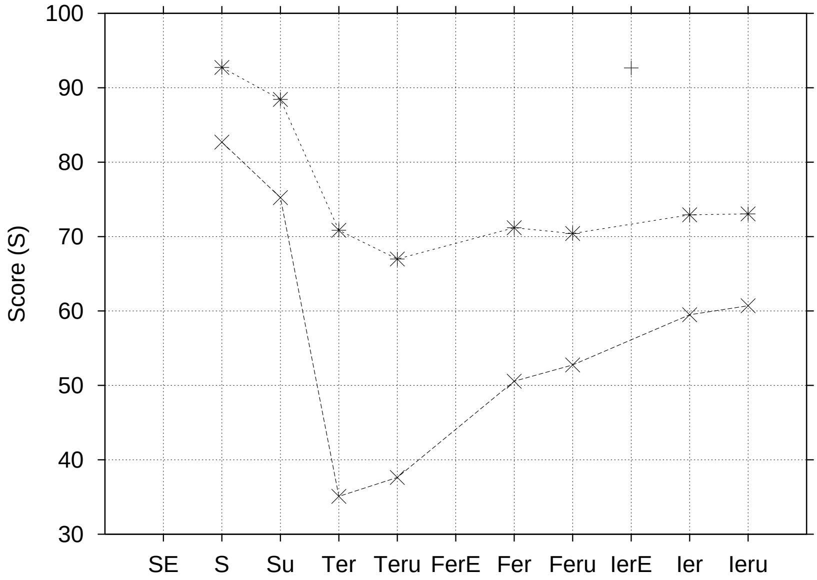

Figure 13: Comparison of lower-bound and cross-validation logarithmic scores S for the variable EARLY-RESULT; $\times$: lower-bound scores for networks with learnt joint probability distribution; $+$: results for networks with underlying subjective joint probability distribution (only one point is included in the plot, as the scores were undefined for the other models); $\star$: cross-validation scores.

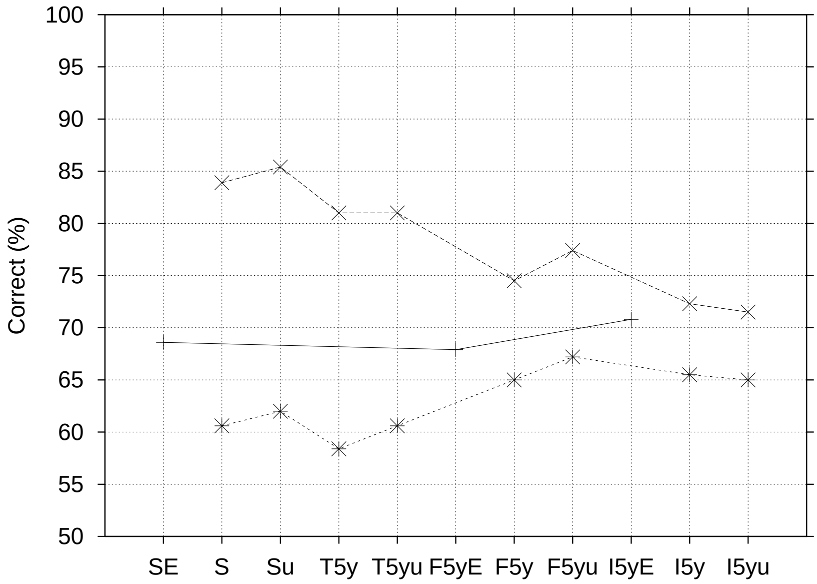

Figure 14: Comparison of upper-bound and cross-validation results for the variable 5-YEAR-RESULT; $\times$: upper-bound results for networks with learnt joint probability distribution; $+$: results for networks with underlying subjective joint probability distribution; $\star$: cross-validation results.

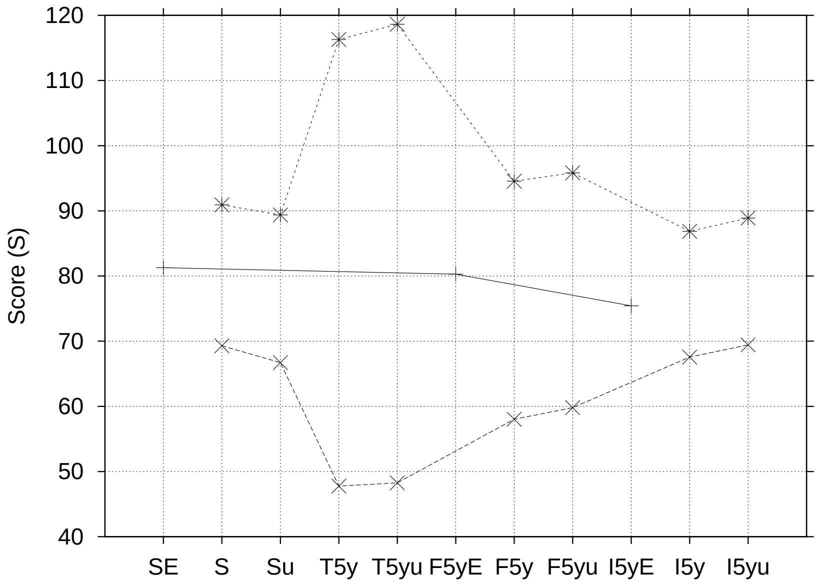

Figure 15: Comparison of lower-bound and cross-validation logarithmic scores S for the variable 5-YEAR-RESULT; $\times$: lower-bound scores for networks with learnt joint probability distribution; $+$: results for networks with underlying subjective joint probability distribution; $\star$: cross-validation scores.

[7, 17]. The system incorporates an interval midpoint algorithm for dealing with missing values [16].

The resulting Bayesian network is shown in Fig. 16. Only 6 of the 30 arcs as suggested by the experts were identified correctly; the arcs between the vertex GENERAL-HEALTH-STATUS and BULKY-DISEASE is clinically incorrect. The arc between HISTOLOGICAL-CLASSIFICATION and BULKY-DISEASE was reversed. Finally, the arc between the vertex CLINICAL-STAGE and CT&RT-SCHEDULE is correct, but omitted by the expert, because therapy is always selected.

6 Discussion

In this paper, we have investigated the relative merits of machine-learning approaches versus knowledge acquisition and modelling in the clinical context of relatively uncommon or rare disorders. In addition, the pros and cons of structured versus unstructured Bayesian networks were studied in order to gather experimental evidence that the development of structured Bayesian networks is worthwhile. The study was carried out against the background of a rising popularity of the naive Bayesian classifier and its variants, such as the tree-augmented naive Bayesian network [5, 6, 15, 18]. The research results reported in this paper extends these results previously reported in the machine-learning literature.

Based on this research, our hypothesis was that independent-form Bayesian networks would outperform declarative Bayesian networks, whereas the addition of some information on interactions as done in TAN Bayesian networks would improve these results somewhat further. What we found was that independent-form Bayesian networks met our expectations, but, it was also established that TAN networks were not particularly robust for small datasets.

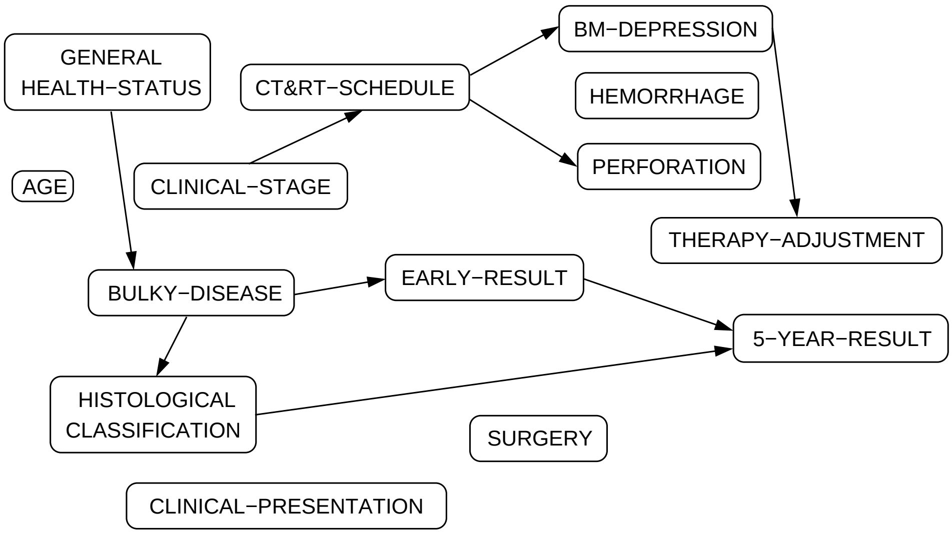

Figure 16: Bayesian-network structure as learnt using BKD.

TAN models appear to suffer from a similar amount of overfitting as the declarative Bayesian model. For the latter type of model, lack of data has in fact a positive effect on their performance when dealing with skewed data distributions.

The naive Bayesian classifier and FAN networks offer the best choice when learning from a small dataset, as they are not very sensitive to variation of the characteristics of the data, such as the presence of skewness. However, assessing subjective probability distributions for declarative networks apparently also offers a reasonably good foundation for the development of classifiers. Despite the fact that subjectively assessed probabilities will not be very accurate, this is being compensated by the *amount* of probabilistic information that is supplied, which easily exceeds the amount of probabilistic information that can be extracted from a small dataset. The rather unsatisfactory conclusion that more accurate probabilistic information may not always yield a better classifier seems to complement similar remarks made in the literature with regard to the naive Bayesian classifier [5].

Finally, although learning the structure of a Bayesian network from a small dataset is not feasible, checking its structure might offer some insight.

References

[1] A.V. Aho, J.E. Hopcroft, J.D. Ullman. *The Design and Analysis of Computer Algorithms.* Addison-Wesley, Reading, MA, 1974.

[2] M. Berthold and D.J. Hand (Eds.). *Intelligent Data Analysis: an Introduction.* Springer, Berlin, 1999.

[3] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory* 14 (1968) 462–467.

[4] R.G. Cowell, A.P. Dawid, D. Spiegelhalter. Sequential model criticism in probabilistic expert systems. *PAMI* 15(3) (1993) 209–219.

[5] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29 (1997) 103–130.

[6] N.I.R. Friedman, D. Geiger and M. Goldszmidt. Bayesian network classifiers. *Machine Learning* 29 (1997) 131-163.

[7] C. Glymour and G.F. Cooper. *Computation, Causation & Discovery.* MIT Press, Menlo Park, CA, 1999.

[8] D. Heckerman. *A Tutorial on Learning with Bayesian Networks.* Technical Report MSR-TR-95-06. Microsoft Corporation, Redmond, 1996.

[9] F.V. Jensen. *An Introduction to Bayesian Networks.* UCL Press, London, 1996.

[10] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2001.

[11] P. Krause. Learning probabilistic networks. *Knowldge Engineering Review* 13(4) (1998) 321–351.

[12] P.J.F. Lucas and L.C. van der Gaag. *Principles of Expert Systems.* Addison-Wesley, Wokingham, 1991.

[13] P.J.F. Lucas, H. Boot, B.G. Taal. Computer-based decision-support in the management of primary gastric non-Hodgkin lymphoma. *Meth Inf Med* 37 (1998) 206–219.

[14] P.J.F. Lucas, H. Boot and B.G. Taal. A decision-theoretic network approach to treatment management and prognosis. *Knowledge-based Systems* 11 (1998) 321–330.

[15] S. Monti, G.F. Cooper. A Bayesian network classifier that combined a finite mixture model and a naive Bayes model. In: K. Blackmond Laskey and H. Prade (Eds.), *Proceeding of UAI'98*, Morgan Kaufmann, San Francisco, CA, 1999, pp. 447–456.

[16] M. Ramoni and P. Sebastiani. *Efficient Parameter Learning in Bayesian Networks from Incomplete Data.* Report KMi-TR-41, KMI, Open University, 1997.

[17] M. Ramoni and P. Sebastiani. *Bayesian Knowledge Discoverer: reference manual.* KMI, Open University, 1997.

[18] M. Ramoni and P. Sebastiani. Robust learning with missing data. *Machine Learning* 45 (2001) 147–170.

[19] R.D. Shachter. Evaluating influence diagrams. *Operation Research* 34(6) (1986) 871–882.