Enhancement of Learning by Declarative Expert-based Models

Peter Lucas
Department of Computing Science, University of Aberdeen
Aberdeen, AB24 3UE, Scotland, UK
E-mail: plucas@csd.abdn.ac.uk

Abstract. A major part of the knowledge in the medical field concerns diseases that are uncommon or even rare. Doctors, however, may face severe difficulties in handling such disorders, as they may not have sufficient experience with the management of these disorders in patients. Here there seems to be a clear role for medical decision-support systems. Unfortunately, the uncommon nature of these disorders renders it almost impossible to collect data from a sufficiently large number of patients as required for the development of models that faithfully reflect the subtleties of the domain. Often, one therefore resorts to the development of naive Bayesian models. However, under these unfavourable circumstances, it may still be feasible to design detailed model of the problem domain in collaboration with expert physicians. The advantage of such models, e.g. structured Bayesian-network models, is that they are often suitable for handling more than one task, e.g. not only predicting prognosis but also treatment selection. This raises the question whether such expert-based Bayesian models could incorporate enough structural background knowledge to improve on naive Bayesian models. In this paper, we discuss the development of several different Bayesian-network models of non-Hodgkin lymphoma of the stomach, which is an example of an uncommon disease. It is shown that a declarative, structured model, based on expert knowledge, can indeed outperform naive Bayesian models when supplied with probabilistic information based on data. The handling of missing values, and the checking of the stochastic independence structure are also discussed in the paper, as these are also important issues when dealing with small datasets.

Keywords & Phrases: Bayesian networks, machine learning, background knowledge, medical decision support.

1 Introduction

There is a great deal of experience in the medical field in analysing medical data of diseases with high prevalence, like breast cancer, lung cancer, and myocardial infarction. Most of the evidence underlying current medical practice is based on such analyses. The advantage linked with the frequent occurrence of those disorders is that it is practically feasible to collect data of large numbers of patients, thus making it possible to draw conclusions that are statistically significant. Typically, datasets collected in the context of such studies may include many thousands of patient records. Such datasets are quite attractive for evaluating particular machine-learning techniques, and, in fact, have been used for this 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 patient cases. 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. Furthermore, systems covering such disorders would be practically useful, as many doctors will lack knowledge and experience to deal with patients affected by those disorders effectively. Confronted with this situation when building a decision-support system, there is a clear place for using medical expert knowledge, as background knowledge, to compensate for lack of data.

Machine-learning literature in medicine not only tends to focus on problems for which much data is 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 is more complicated than that, and 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. We believe that it might be worthwhile to consider instead the development of declarative medical models, that can be used to explore different problems, and be reused for different tasks. One of the nice aspects of declarative models is that 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. Admittingly, developing such models will be more challenging, both in terms of required number of variables and probabilistic information, but the extra effort may be out-weighted by their greater potential. Yet, it is presently unclear what the potential and limitations of such declarative models are with respect to capturing knowledge from small datasets.

In this paper, we will try to find answers to a number questions related to the issues mentioned above, based on our experience in developing declarative prognostic models of non-Hodgkin lymphoma of the stomach. Non-Hodgkin lymphoma of the stomach is a typical example of an uncommon, although not extremely rare disease. Here we focus on investigating the consequences of adopting different assumptions underlying Bayesian-network technology, in particular with respect to structure, number of variables included in a model and dealing with missing values. The following issues are addressed:

- Does a declarative model enhance the learning of knowledge?
- Is it worthwhile to handle missing values explicitly?
- 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, the development of different Bayesian-network models of non-Hodgkin of the stomach is discussed. Sections 4 pays attention to their evaluation, whereas in Section 4 some results of checking the structure of a Bayesian network are presented. The paper is rounded-off by a comparison to related work and with 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 independencies among the variables, modelled by absence of arcs among vertices [5, 6, 9]. On the variables $\mathcal{V}$ is defined a joint probability distribution $\Pr(V_1, \ldots, V_n)$, for which the following decomposition property 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. 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.

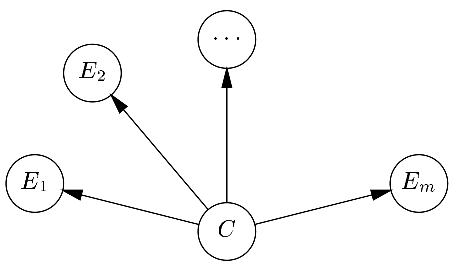

Figure 1. Independent-form Bayesian network.

The Bayesian-network models conforming to the topology shown in Figure 1 have been particularly popular in the statistical and machine-learning communities [8]. It corresponds 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. This form of Bayes' rule is also known as the *naive Bayes' rule*. The independent form of Bayes' rule is used to compute the a posteriori probability of a class value c_k given the evidence $\mathcal{E}$ [6]:

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

where class variable C has q mutually exclusive values, and $\Pr(\mathcal{E}) > 0$. 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,

$$\Pr(\mathcal{E}) = \sum_{j=1}^{q} \Pr(\mathcal{E} \mid c_j)\Pr(c_j)$$

using marginalisation and conditioning.

Although an independent-form Bayesian network may ignore important probabilistic dependence information, it has the virtue that assessment of the required probabilities $\Pr(E_j|C)$ and $\Pr(C)$ is rather straightforward, and can be carried out with a relatively small dataset. Determination of the a posteriori probabilities $\Pr(C|\mathcal{E})$, where $\mathcal{E} \subseteq \{E_1, \ldots, E_m\}$, is computationally speaking a trivial task under the mentioned assumptions. Furthermore, it is quite straightforward to handle missing values with the independent form. For example, M. Ramoni and P. Sebastiani have developed an interval-based method, and implementation of it as the ROC system (Robust Bayesian Classifier), that is capable of dealing with missing values in a mathematically sound way [13]. These features of the independent form of Bayes' rule probably explain why it is again increasingly popular, having fallen into disgrace two decades ago.

An independent-form Bayesian network is especially suited for the classification of cases based on a set of known features $\mathcal{E}$. It is not meant for providing a description of the knowledge in a particular problem domain. In most cases, only those variables considered relevant for the classification task are included. Since relationships among features are also omitted, the result is relatively naive from the perspective of domain modelling, hence its nickname 'naive' Bayesian model. On the other hand, building a Bayesian network that models the problem domain more accurately would almost certainly require the elicitation of domain knowledge from human experts. The development of such expert-based models can be quite time-consuming. Hence, developing such models would only be worthwhile when collecting expert knowledge would counterbalance the lack of data in small datasets. The role of encoded human expertise in machine learning is a very relevant issue.

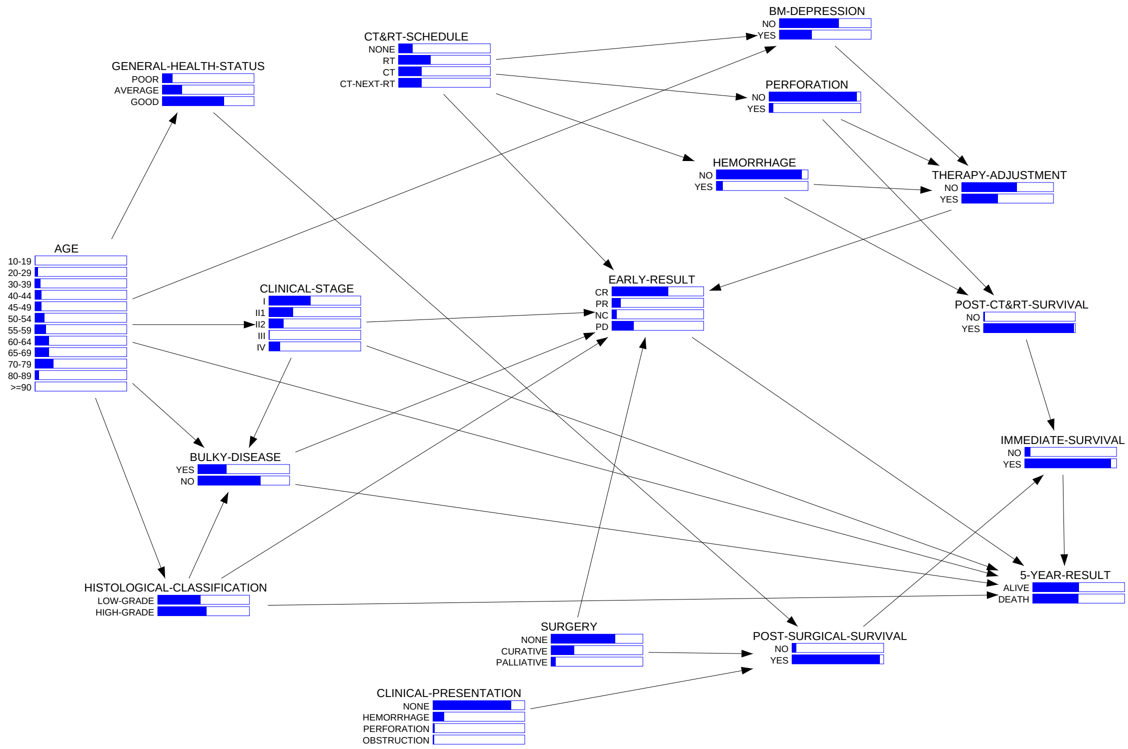

Figure 2. Bayesian-network model as designed with the help of medical experts.

In the following, we shall discuss a number of declarative and independent-form Bayesian models for non-Hodgkin lymphoma of the stomach, which sheds some light on the issues mentioned above.

3 Bayesian models of non-Hodgkin lymphoma of the stomach

A Bayesian model incorporating most factors relevant for the management of non-Hodgkin lymphoma of the stomach was developed in collaboration with clinical experts from the Netherlands Cancer Institute (NKI). The resulting model, shown in Figure 2, includes variables like age of the patient and clinical stage of the tumour (stage I is generally associated with good prognosis, whereas stage IV is generally associated with very poor prognosis). Some of the included variables concern patient information obtained from diagnostic procedures, which will be available prior to treatment selection. 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 5-YEAR-RESULT. It is called the *posttreatment* part of the model. Finally, the model includes the variables SURGERY and CT&RT-SCHEDULE, which represent *treatment* variables with possible values: 'none', 'curative', 'palliative' for SURGERY, and 'none', 'chemotherapy', 'radiotherapy' or 'combination therapy' for CT&RT-SCHEDULE. Note that the declarative model of non-Hodgkin lymphoma of the stomach can indeed be used in the context of different tasks, such as prediction of prognosis, treatment selection – by comparing the likely outcomes of alternative treatment choices, possibly using preference information expressed as utilities – and generation of patient profiles [7].

The independent-form Bayesian network corresponding to the network shown in Figure 2 is depicted in Figure 3. Note that this model includes a single class variable: the posttreatment variable 5-YEAR-RESULT; in addition, all pretreatment and treatment variables are incorporated. The remainder of the posttreatment variables have been left out, since as these variables are always unknown for a patient, they are not immediately relevant for the prediction of 5-year survival in patients. This illustrates one difference between a declarative Bayesian model, as shown in Figure 2, and a task-specific model, as shown in Figure 3.

Taking the two topologies as a starting point, the probability distributions of the two models were learnt, adopting a number of different assumptions. The resulting models that will be discussed in this paper are briefly described in Table 1. 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. Learning took place using a dataset with patient data from the Netherlands Cancer Institute, comprising 137 cases, with some missing data. Missing data were either ignored (models S and I), distributed uniformly among the values of the variable for which a value was missing, or probability intervals instead of point probabilities were determined. The last approach thus views missing data as relaxing the constraints on a probability distribution. Model $\mathrm{I}_{\mathrm{ic}1}$ was learnt starting with an initial uniform proba-

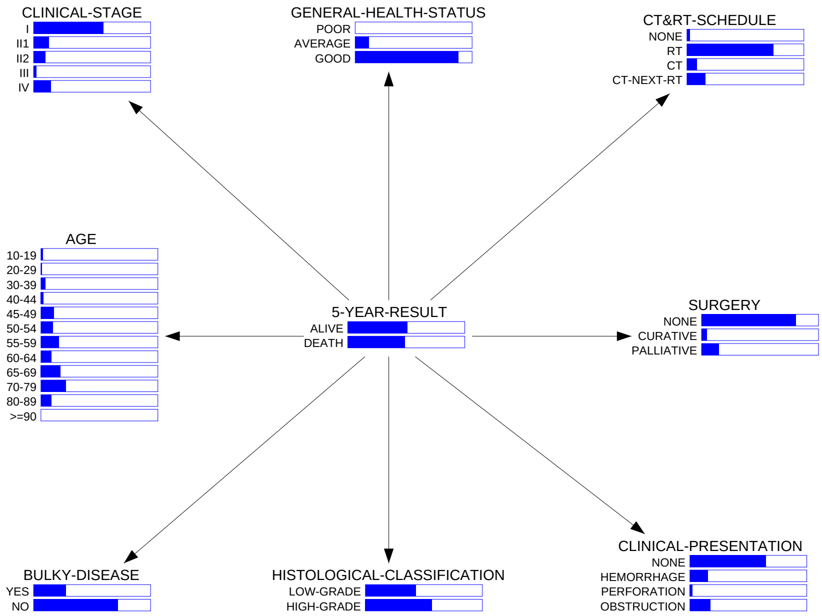

Figure 3. Independent-form Bayesian model.

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
I	learnt model	independent form	ignored
$\mathrm{I}_{\mathrm{ic}1}$	learnt model, sample size 22	independent form	interval calculus
$\mathrm{I}_{\mathrm{ic}2}$	learnt model, sample size 0	independent form	interval calculus
I_u	learnt model	independent form	uniform distribution

bility distribution, which was given an equivalence sample size of 22 cases. This means that the initial uniform probability distribution had a weight as if it was based on a sample of 22 cases. For model $\mathrm{I}_{\mathrm{ic}2}$, it was assumed that there were no such cases at all, i.e. no initial uniform distribution was assumed.

4 Evaluation and comparison

The Bayesian models were evaluated in a number of different ways. First, the a posteriori probability $\Pr(5\text{-YEAR-RESULT} \mid \mathcal{E})$ was computed, where $\mathcal{E}$ was the available evidence for each of the 137 patients with non-Hodgkin lymphoma of the stomach, restricted to values of all pretreatment and treatment variables. When $\Pr(5\text{-YEAR-RESULT} = \text{alive} \mid \mathcal{E}) > 0.5$, and the patient was known to have been alive more than 5 years following treatment, the model's prediction was considered correct; otherwise, it was classified as being incorrect. A similar decision procedure was used when the most likely prediction was that the patient would die within 5 years. No use was made of receiver-operator-characteristic curves [15], because this would not be entirely in line with the declarative nature of structured Bayesian networks.

For the two independent-form Bayesian networks learnt using the ROC system, the interval a posteriori probability distributions were interpreted using the *stochastic-dominance criterion* [13], i.e. it is assumed that the model predicts that class c is most likely if the minimum probability associated with this class value, i.e. $\Pr_{\min}(c \mid \mathcal{E})$, is larger than the max-

Table 2. Results for different Bayesian models. Percentages were computed by dividing the number of correct conclusions by the number of classified cases.

Model	Total	Classified (n) Incorrect	Correct	(%)	Unclassified (n)	Missing Data
S_E	137	43	94	(68.6)	—	—
S	137	22	115	(83.9)	—	—
S_u	137	20	117	(85.4)	—	—
I	137	40	97	(70.8)	—	—
I_{ic1}	137	28	98	(77.8)	11	stochastic dominance
	137	35	102	(74.5)	0	weak dominance
I_{ic2}	137	28	100	(78.1)	9	stochastic dominance
	137	32	105	(76.6)	0	weak dominance
I_u	137	39	98	(71.5)	—	—

imum probability associated with the other class values, i.e.

$$\Pr_{\min}(c \mid \mathcal{E}) > \max\{\Pr_{\max}(c' \mid \mathcal{E}) \mid c' \neq c\}$$

A disadvantage of the stochastic-dominance criterion is that often some cases remain unclassified. Therefore, a weaker criterion, called *weak dominance*, was used as well. This criterion associates a score $s_u(c \mid \mathcal{E})$ with each class value c; when C has two mutually exclusive class values, the score is defined as follows [13]:

$$s_u(c \mid \mathcal{E}) = \frac{\Pr_{\min}(c \mid \mathcal{E}) + \Pr_{\max}(c \mid \mathcal{E})}{2}$$

i.e. the interval midpoint is chosen. The class value with the highest score is selected. The results for the Bayesian-network models, as defined in Table 1, are given in Table 2.

A disadvantage of the straightforward method for comparing the quality of the Bayesian models, as described above, is that the actual a posteriori probabilities are ignored. A more precise impression of the behaviour of the Bayesian models would have been obtained if the resulting probabilities had been taken into account as well. For example, if patient Q was known to have survived more than 5 years following radiotherapy, and a Bayesian model had predicted this event with 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. A number of different scoring rules have been designed in the field of statistics that measure exactly this effect. One of the simplest scoring rules that can be given a statistical interpretation is the *logarithmic scoring rule* [2]. We shall briefly discuss this rule in the following.

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:

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

which has the informal meaning of a penalty: when the probability $\Pr(c_k|\mathcal{E}) = 1$, then $S_k = 0$; otherwise, the score becomes rapidly larger than 0. The total score for an entire database D is now defined as follows:

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

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

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

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

The results obtained for five of the models are shown in Table 3.

Table 3. Logarithmic scores.

Name	S	E	V
S_E	81.28	71.47	30.65
S	57.05	65.93	25.65
S_u	48.92	58.88	25.91
I	67.50	51.31	28.20
I_u	69.50	54.07	31.99

The performance of the declarative model S_E, which incorporated expert-assessed probabilities, was lowest; the Bayesian networks S and S_u with the same topology as S_E, but with probabilities learnt from data, yielded the best results. It was not really surprising that model S_E was inferior to the other models, as its probability distribution was assessed taking into account recent changes in treatment policy as well as experience at other hospitals, as reported in the literature. In addition, some deviation of subjective probabilities from relative frequency information may always be expected to exist. In a sense, it was surprising that the performance of S_E was still near to that of model I.

The various independent-form Bayesian networks yielded results that were always below those of the two trained declarative models S and S_u. However, the results for the two independent-form models, where missing values were handled by means of interval calculus, were in turn better than those for the other two independent-form models. These results are consistent with previous results by M. Ramoni and P. Sebastiani [14], who also showed that using interval calculus may improve performance, although only to a slight extent when using the weak-dominance criterion. Stochastic dominance may leave difficult cases unclassified, which explains the

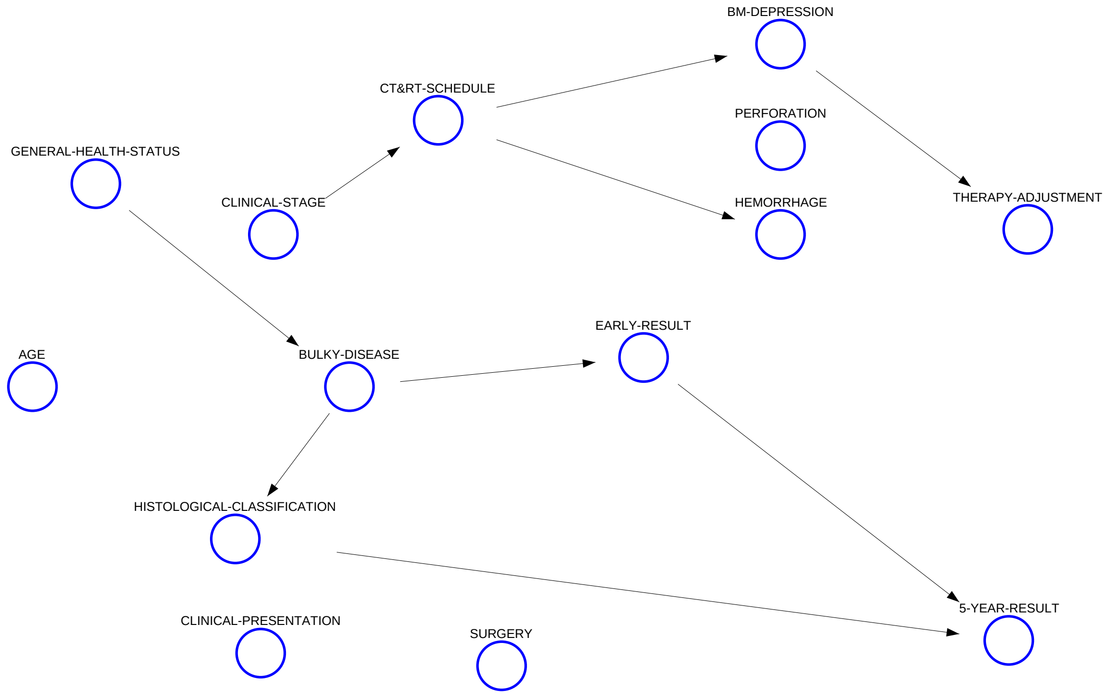

Figure 4. Bayesian-network structure as learnt using BKD.

better performance. However, this feature renders the latter criterion practically speaking less useful.

The Bayesian models in which missing values were uniformly distributed among values generally yielded slightly better results than the other models due to the artificial increase in sample size. The logarithmic scores shown in Table 3, however, indicate that this results in a slightly decreased accuracy of the a posterior probability for the independent-form Bayesian model; in contrast, the accuracy of the a posterior probabilities of the declarative model was improved. The latter effect is likely due to a reinforcement of the influence of the topology of the graph on the outcome; as the topology of the graph reflects clinical expertise, the quality of the results improves.

5 Checking the topology

Even though it is in principle possible to learn the topology of a Bayesian network from data, the dataset we had at our disposal on this occasion (a typical uncommon-disease dataset as mentioned in the Introduction) was just too small for this purpose. Nevertheless, experiments with a model (structure) selection algorithm were carried out in order to obtain more insight in how far one gets with such algorithms, given a small dataset. Use was made of the BKD (Bayesian Knowledge Discoverer) system, which implements a heuristic search method for finding the Bayesian-network structure S that best fits the data D according to the ratio $\Pr(D,S)/\Pr(D,S')$, where S' is an alternative structure [1, 11]. Furthermore, the system includes an algorithm for dealing with missing values, called *bound-and-collapse* [12]. This method first determines the bounds of a probability value using the data that is available, and finally collapses the set of values to a single probability using a convex combination of the extreme values of the set.

The resulting Bayesian network is shown in Figure 4. Only 6 of the 30 arcs in the expert-assessed network were predicted 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 left out in the original model, because it is assumed that therapy is always selected.

It is surprising that the algorithm was not able to find dependencies between the variable AGE and the other variables, as it is well known that many of the variables in the model are influenced by age. A similar observation holds for the variable CLINICAL-STAGE.

6 Discussion

The present paper is certainly not the first paper showing that human background knowledge may enhance machine learning. Earlier work in the areas of inductive logic programming (e.g. [4], but there are many other papers), and Bayesian networks [3], has demonstrated this. However, other researchers have primarily focused on the gathering of sufficient pieces of knowledge in order to enhance the learning of knowledge from data. In contrast, the present paper investigates the usefulness of extensive declarative models that were not developed for the purpose of learning in the first place. It appears that the effect of the topology of a Bayesian network as assessed by human experts, may be so strong that, even though a small proportion of its probability tables are filled with probabilities obtained from data, a structured model still outperforms independent-form Bayesian networks. Note, furthermore, that is not true, as suggested by Pradhan et al. [10] that once given the structure of a Bayesian network, the incorporated probabilities are not relevant at all, as is demonstrated in this paper by the low performance of the expert-assessed network.

From this study, one can also conclude that dealing with missing values in a typical clinical research dataset may offer some advantages. In particular, it seems worthwhile to consider using interval calculus when standard data imputation techniques are expected to be unreliable. A clinical research database as used in this study will typically have relatively few missing values. Under these circumstances, the improvement resulting from dealing with missing values will only be moderate.

A limitation of the present study is that the resulting networks have not been evaluated using cross validation, which would be computationally quite intensive for the structured models. Given the techniques used – independent Bayes and a structured Bayesian network – it is clear that the declarative model is capable of better capturing the logic implicit in the data. It seems likely that this conclusion would stand further scrutiny.

Finally, although learning the structure of a Bayesian network from a small dataset is not feasible, checking its structure using one of the structure-learning algorithm might offer some insight. Unfortunately, when differences between the expert-derived and predicted topology cannot be explained clinically, there is no alternative than to stick to the original structure.

The results of this paper suggest that it may be worthwhile devoting even more time to the gathering of background knowledge, and to designing extensive domain models, than is usually done in the area of machine learning.

Acknowledgments. I am grateful to Derek Sleeman, who offered some useful comments to the original version of this paper.

REFERENCES

[1] G.F. Cooper, E. Herskovitz. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 1992; 9: 309–347.

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

[3] D. Heckerman, D. Geiger, D. Clickering. Learning Bayesian networks: the combination of knowledge and data. *Machine Learning* 1995; 20: 197–243.

[4] T. Horváth, G. Turván. Learning logic programs with structured background knowledge. In: L. De Raedt (ed.), *Advances in Inductive Logic Programming*, IOS Press, Amsterdam, 1996, pp. 172–191.

[5] S.L. Lauritzen, D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society (Series B)* 1987; 50: 157–224.

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

[7] P.J.F. Lucas, H. Boot, B.G. Taal. Computer-based decision-support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine* 1998; 37: 206–219.

[8] T.M. Mitchell. *Machine Learning.* McGraw-Hill, New-York, 1997.

[9] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufman, San Mateo, California, 1988.

[10] M. Pradhan, M. Henrion, G. Provan, B. Del Favero, K. Huang. The sensitivity of belief networks to imprecise probabilities: an experimental investigation. *Artificial Intelligence* 1996; 84(1-2): 363–397.

[11] M. Ramoni, P. Sebastiani. *Efficient Parameter Learning in Bayesian Networks from Incomplete Data.* Report KMi-TR-41, Knowledge Media Institute (KMI), Open University, 1997.

[12] M. Ramoni, P. Sebastiani. *Bayesian Knowledge Discoverer: reference manual.* Knowledge Media Institute (KMI), Open University, 1997.

[13] M. Ramoni, P. Sebastiani. *An introduction to the Robust Bayesian Classifier.* Report KMi-TR-79, Knowledge Media Institute (KMI), Open University, 1999.

[14] M. Ramoni, P. Sebastiani, R. Dybowski. Robust outcome prediction for intensive-care patients. In: A. Abu-Hanna and P.J.F. Lucas (eds.), Prognostic Models in Medicine: Artificial Intelligence and Decision-analytic approaches. *Workshop Notes AIMDM'99*, Aalborg, 1999.

[15] H.C. Sox, M.A. Blatt, M.C. Higgins, K.I. Marton. *Medical Decision Making.* Butterworths, Boston, 1988.