Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

STENO:an expert system for medical diagnosis based on graphical models and model search

Lars Rude Andersen [a] , Jens Herman Krebs [a] & Jens Damgaard Andersen [a]
[a] Department of Computer Science , University of Copenhagen
Published online: 28 Jul 2006.

To cite this article: Lars Rude Andersen , Jens Herman Krebs & Jens Damgaard Andersen (1991) STENO:an expert system for medical diagnosis based on graphical models and model search, Journal of Applied Statistics, 18:1, 139-153, DOI: 10.1080/02664769100000012

To link to this article: http://dx.doi.org/10.1080/02664769100000012

STENO: an expert system for medical diagnosis based on graphical models and model search

LARS RUDE ANDERSEN, JENS HERMAN KREBS & JENS DAMGAARD ANDERSEN, *Department of Computer Science, University of Copenhagen*

SUMMARY *Causal probabilistic models have been suggested for representing diagnostic knowledge in expert systems. This paper describes the theoretical basis for and the implementation of an expert system based on causal probabilistic networks. The system includes model search for building the knowledge base, a shell for making the knowledge base available for users in consultation sessions, and a user interface. The system contains facilities for storing knowledge and propagating new knowledge, and mechanisms for building the knowledge base by semi-automated analysis of a large sparse contingency table. The contingency table contains data acquired for patients in the same diagnostic category as the intended application area of the expert system. The knowledge base of the expert system is created by combining expert knowledge and a statistical model search in a model conversion scheme based on a theory developed by Lauritzen & Spiegelhalter and using exact tests as suggested by Kreiner. The system is implemented on a PC and has been used to simulate the diagnostic value of additional clinical information for coronary artery disease patients under consideration for being referred to coronary arteriography.*

1 Introduction

The use of probabilistic and statistical methods to deal with uncertainty in expert systems requires the combination of declarative knowledge (typically formulated by an expert) concerning associations between entities with knowledge generated by statistical analysis of data relating the same entities.

During the design of an expert system used for diagnosis, a desirable objective is an accurate and preferably *verifiable* representation of the diagnostic knowledge. By 'verifiable' it is meant that the structural model on which the expert system is based can be described in a commonly accepted way, e.g. as a statistical model, the assumptions of which can be tested if desired.

The declarative knowledge states dependences between entities; these can be modelled as 'causal' relations or links between entities. Statistical knowledge is derived from analysis of data acquired by collecting data concerning patients in the same diagnostic category as the prospective application area of the expert system. These data include disease manifestations and the final diagnosis.

Designing a verifiable expert system is therefore related to a statistical model search in that (i) some structural relationships between variables are known *a priori* (e.g. from pathophysiological causality considerations) and (ii) data are available which summarize past experience by recording for each case the same variables. The two information sources should be integrated to create an optimal model of the diagnostic knowledge. This process of constructing the expert system is called the *knowledge engineering phase* and is supported by the expert system STENO with tools for modelling data to establish a knowledge base for use in the system.

Furthermore, the system includes an expert system shell which can be used with the modelled data during simulation experiments using data from single patients. Typically some patient data known for a particular patient are entered in the system and the system is then able to calculate the probability for a certain combination of outcomes of specified hypotheses being true or false. This phase is termed the *consultation phase*. The shell includes a mechanism for ranking unsettled questions according to their information value during the consultation phase. The system has been used in simulation experiments for medical decision making and compared favourably with existing alternatives for decision support. A more detailed evaluation of the system is now under way. Although the system is described here as applied to clinical diagnostics, the principles are general and the system can be used in any area containing diagnostic problems.

The primary objective of this paper is to describe the expert system STENO and experiments with the system in clinical decision making. Section 2 explains the knowledge representation scheme in terms of subclasses of log-linear models. Section 3 details the combination of domain structural knowledge and statistical evidence. Furthermore, the control strategy of selecting most significant information in the given context during the consultation phase is described in Section 4. Section 5 describes the current implementation and in Section 6 a clinical decision problem—diagnosis of coronary artery disease and optimum strategy for referral to coronary arteriography—is used to illustrate the system. This section also contains a preliminary performance evaluation of the system.

2 Knowledge represented by recursive causal and graphical models

One way to represent the structure of the diagnostic knowledge of an expert system is to model using recursive causal models. Such models can be depicted graphically by a recursive causal graph or causal network (Wermuth & Lauritzen, 1983). Properties of causal networks have been investigated since the introduction of causal models by Kiiveri *et al.* (1984). In this paper we follow the terminology of the latter reference.

2.1 Terminology

In a causal network each categorical variable is represented by a vertex. To simplify the subsequent exposition, a graph vertex and the variable it represents is denoted by the same letter when no ambiguity can arise. A directed edge from vertex v to vertex

w represents a 'causal' relationship in the broadest sense, i.e. loosely speaking it supplies the information that the value of variable w (the child) depends on the value of variable v (the parent) (a *directed relation* or *dependence relation*). The set of parents for a given variable v is denoted $\Pi(v)$. If the problem area is faithfully modelled, the network reflects the property that no source other than the parents $\Pi(w)$ of a given vertex w influence the probability distribution over the values of the vertex. Variables without parents are called *exogenous* variables (denoted by index x, i.e. V_x) and variables with at least one parent *endogenous* variables (denoted by index n, i.e. V_n). Finally, *explanatory variables* are variables considered as input, typically manifestations, the interaction of which we are not interested in but cannot exclude.

The class of log-linear models for contingency tables contains some subclasses of interest for the present system. One is the subclass of *graphical models* which has the property that each member has a one-to-one correspondence to an undirected graph (see Darroch *et al.*, 1980). Decomposable models are graphical models whose graphs are triangulated, i.e. contain no cycle of length greater than 3 without a chord (Lauritzen, 1989). The *recursive models* are models in which a certain order is imposed on the variables such that a variable v_i, $i \in \{1, \ldots, k\}$ may be considered to be a response variable with respect to some or all other variables v_j, $j \in \{i+1, \ldots, n\}$, but not thought of as being a response to any one of the variables v_h, $h \in \{1, \ldots, i-1\}$.

2.2 *Model transformations*

Recently Lauritzen & Speigelhalter explored local representations for the joint probability distribution corresponding to graphical transformation from a causal network to a Markov random field (the *potential representation* of Lauritzen & Spiegelhalter (1988)). The latter representation is suitable for decomposing complex relationships into clusters of simpler interactions. After decomposition, a recursive causal model can be created without changing the probability distributions. The resulting model may then be used as a simulation tool. Simulations can support decision making, e.g. in medical diagnostics, since the effects of the various alternatives in a given choice situation may be illuminated by 'what if' types of questions.

Figure 1 illustrates the conversion sequence. The graphical models are used for the model search. When completed, the resulting graphical model is converted into a recursive model and into a Markov random field. The Markov random field representations are used as a knowledge base for propagation of probabilities and inference.

There are three major reasons for representing the knowledge of the expert system by means of a recursive causal model. These are as follows:

(1) Two separate knowledge bases, each represented by a recursive causal model, can be combined into one model. This is useful for handling large knowledge bases which may be decomposed, analyzed and then reassembled.

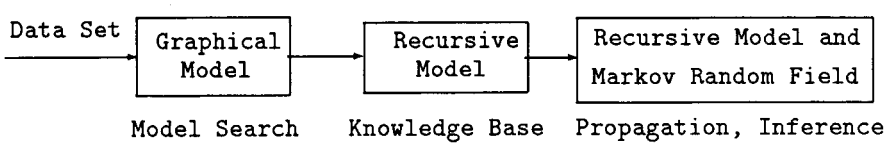


FIG. 1. Model transformation scheme.

(2) The structure of recursive causal models is easy to represent graphically and conceptually acceptable.
(3) The recursive causal structure of the model makes it possible to adapt the marginal probabilities to a new population.

3 The knowledge engineering phase

The aim of the knowledge engineering phase is to construct—under expert guidance—a knowledge base using observations from the application domain. We observe a set of random variables which are assumed to be multinomially distributed. For each random variable we then obtain a set of observations. Subsequently, we form a contingency table which has all the variables in the decision problem as categorical variables, both variables used as input as well as hypotheses needed to be evaluated. This contingency table is usually large and sparse.

Using standard goodness-of-fit tests with asymptotic text probabilities does not suffice for a model search because it is impossible to attain the average cell count necessary for these tests to be reliable (see Kreiner (1989) for a detailed analysis of this problem). We have therefore chosen to use simulated exact tests as suggested by Kreiner (1986), which can be used to compute goodness-of-fit test probabilities for tests where, given sufficient marginals, the distribution of the table is known. This is the case for the test of the removal of a single edge in a graphical model, which is what we have used.

An example of this is testing the model [ABC][BCD] against [ABCD], i.e. testing the removal of the edge AD in the complete graphical model with vertices A, B, C and D (see Fig. 2).

This choice implies that we are testing within the framework of graphical models, which leaves us with two problems. First, which model search strategy should be used, and secondly, how should we convert the graphical model which is the result of the search, to a recursive causal model, which is the knowledge base.

3.1 Model search strategy

The model search strategy which we use here is that described by Kreiner (1986). We will provide a short description of model selection and refer to the thorough discussion in Kreiner (1986, 1987b) for further information.

The model search procedure uses a few auxiliary notions which are as follows.

Let V be a set of vertices and CG(V) the complete graph having V as vertices, and

- $CG(\{V_1, \ldots, V_i, \ldots, V_n\}) = \{CG(V_1), \ldots, CG(V_n)\}$ the set of complete graphs with vertices V_i

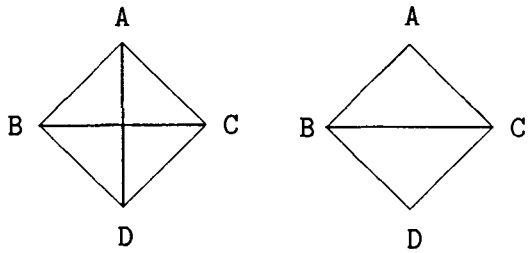


FIG. 2. Testing edge removal in graphical models.

- Let MCS (G, v, w) be the set of minimal vertex cut sets between the vertices v and w in the graph G. A cut set is a set of vertices that cuts off all paths between v and w. A minimal cut set is a cut set with minimal cardinality.
- For sets $V_1, \ldots, V_n$ we define $\{V_1, \ldots, V_n\} \dot{\cup} V$ as $\{V_1 \cup V, \ldots, V_n \cup V\}$.

Then the model search procedure is given the set of vertices V of the graphical model as input. It produces a resulting model R, initially consisting of all vertices in V, but no edges. The procedure works as follows.

(1) For each pair of vertices $\{v, w\} \in V \times V$ we test the model $CG(V) \setminus \{v, w\}$ against CG(V). Those edges that cannot be removed are added to R.
(2) Add an edge $\{v, w\}$ to R that satisfies the following criteria.

NoRemoval: Removal of $\{v, w\}$ is not accepted in at least one of the graphs $CG(MCS)(R, v, w) \dot{\cup} \{v, w\})$.

Continue with step (2) until no further edges satisfy (*NoRemoval*).

The idea of the procedure is to check the initial selection of edges in step (1) in smaller graphs, corresponding to the observation of fewer variables.

Example

Given vertices ABCDE. We start in a saturated model testing for single edge removal.

(1) Suppose that edges AB, AC, BC, BD, CE, DE have been rejected. This yields the model shown in Fig. 3.
(2) Test for inclusion of edge AD.

$$MCS([ABC,BD,CE,DE],A,D) = \{\{B, C\}, \{B, E\}\}$$

and

$$CG(\{\{B, C\}, \{B, E\}\} \dot{\cup} \{A, D\}) = \{[ABCD], [ABDE]\}$$

If removal of AD is rejected in any of the two models, AD is added to R. Note that we have illustrated one iteration of step (2) only.

Once a new model has been found, the whole procedure is repeated. During these iterated searches some edges persist ('strong edges') while other edges usually disappear in every repetition of the procedure ('weak edges'). We use this information in the final execution of the procedure, trying to add 'strong edges' early. This method allows us to exploit the full information content of the data.

The reason for this is that in the early stages of the search, a considerable degree of freedom of choice is left to the implementation, since many edges generally satisfy

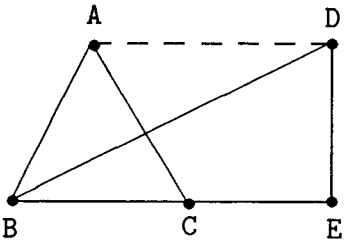


FIG. 3. The edge removal procedure.

the second criterion (*NoRemoval*). Since only some of the edges persist in being rejected, these can be added first, possibly leaving out some other edges.

Obviously, this method is too crude for a thorough analysis of the data. One reason for this is that the graphical models themselves cannot express all relations (see for example Andersen & Krebs, 1989; Kreiner, 1987a, 1988). However the objective is an interaction knowledge base rather than statistical testing. Furthermore, the procedure for forming the expert system knowledge base is reproducible from input data with few assitional guidelines on the iterative performance of tests.

3.2 Using a recursive model for a knowledge base

This section describes how to convert the graphical model into a recursive causal model, thus making it applicable as a knowledge base. Since graphical models and recursive causal models express different relations, we must specify what we mean by 'converting' the model. To do this, we first turn our attention to a simpler case, namely where the graphical model is decomposable. In this case, there exists a corresponding recursive causal model. The proof of this fact, along with background material, can be found in Wermuth & Lauritzen (1983).

3.2.1 Converting decomposable models

We will now describe an algorithm for converting a decomposable, graphical model (V, E, p) with vertices V, edges E and distribution p to a recursive causal model with the same probability distribution. Since the recursive causal model is built on a directed graph, we need an ordering of the vertices, $\sigma: V \rightarrow \mathbb{N}$. This can be produced by a maximum cardinality search (see Tarjan & Yannakakis, 1984). Having found the structure of the recursive causal model it is easy to compute the necessary numerical assignments for the probability distribution. What we need are the conditional probabilities

$$p(v|\Pi(v)) = \frac{p(\{v\} \cup \Pi(v))}{p(\Pi(v))}$$

Using the definition of conditional probabilities and the fact that $\Pi(v) \cup \{v\}$ is contained in a single clique, the probabilities can be computed starting from the top of the graph, i.e. from $\sigma^{-1}(1)$.

As an example, consider the conditional probability $p(c|a)$ given by the graph [ba][ac], where $\sigma(b) = 1$, $\sigma(a) = 2$ and $\sigma(c) = 3$. This is given by

$$p(c|a) = \frac{p(a, c)}{p(a)}$$

The ease of computation derives from the fact that all the probabilities needed are contained in a single clique; this is a property of the decomposable model. This follows from the fact that what we are actually computing is the maximum likelihood estimate under the assumption of the hypothesis of the graphical model [ab][ac] (see Lauritzen, 1989).

3.2.2 Converting non-decomposable graphical models

We now turn our attention to the case where the graphical model given from the statistical analysis is not decomposable.

The first step in this case is to add edges to the graphical model to make the graph decomposable. This produces a graphical model, which also can be accepted as a description of the data, since the graph contains the graph of the accepted model as a subset. It would be possible to keep the graph decomposable during the model search. This, however, is not desirable since it adds irrelevant criteria to the model search.

Our approach is therefore to make the graph decomposable but to compute the numerical assignments of the probability distribution based on the original graph. In this way we can add edges to the graph without changing the joint probability. We will term such edges 'dummy edges'. The recursive causal model that corresponds to a graphical model may therefore have more edges, but the joint probability distribution is unchanged since we simply disregard the dummy edges when computing the joint probability distribution.

Making the graph decomposable is easily done with an extension of the maximum cardinality search (see Tarjan & Yannakakis, 1984 or Lauritzen & Spiegelhalter, 1988). The probabilities are once again the maximum likelihood estimates. These, however, cannot be computed using simple formulae like those above. In general, we have to compute the joint probability as the solution to the set of equations given by

$$p(i_c) = n(i_c)/n, \quad c \in \mathrm{C}$$

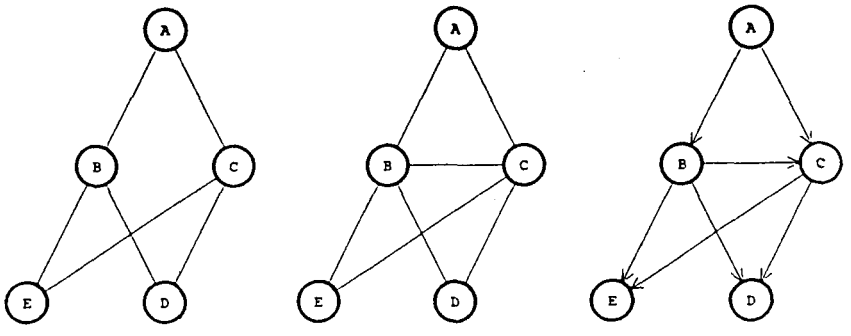
where C is the set of cliques and i_c is a member of the subspace R^{i_c} of R^i, corresponding to the clique c, and n is the total number of observations. The formula, as well as an introduction to the terminology, can be found in Lauritzen (1989).

As an example, consider the graph in Fig. 4(a). This corresponds to a graphical model which is not decomposable. The graph in Fig. 4(b) is decomposable and has been produced from the graph in Fig. 4(a) by filling in the edge (B, C), and Fig. 4(c) is the recursive version corresponding to both Figs. 4(a) and 4(b), produced by means of the algorithm.

To illustrate the algorithm, we compute a factorization of the joint distribution for the graph in Fig. 4(a) using Bayes' theorem and the Markov property of the probability:

$$\begin{aligned} p(abcde) &= p(\mathrm{e}|\mathrm{abcd})p(\mathrm{d}|\mathrm{abc})p(\mathrm{c}|\mathrm{ab})p(\mathrm{b}|\mathrm{a})p(\mathrm{a}) \\ &= p(\mathrm{e}|\mathrm{bc})p(\mathrm{d}|\mathrm{bc})p(\mathrm{c}|\mathrm{a})p(\mathrm{b}|\mathrm{a})p(\mathrm{a}) \end{aligned}$$

By letting $p(\mathrm{c}|\mathrm{ab}) = p(\mathrm{c}|\mathrm{a})$, we obtain that this factorization can be used for the graph in Fig. 4(c).



(*a*) (*b*) (*c*)

FIG. 4. Creation of a decomposable graphical model and conversion to a recursive model: (a) the original graphical model; (b) a decomposable model derived from it; (c) the resulting recursive model.

3.3 Explanatory variables

The explanatory variables are used as input in the consultation phase. This has the important practical implication that the variables usually are easy to observe in any population, and they are therefore suitable for adjusting probabilities.

As noted in Section 2, we are not interested in the interaction between the explanatory variables. We wish to keep the explanatory variables exogenous in the recursive model resulting from the knowledge engineering phase. The reason is that this makes it possible to adjust probabilities (cf. Section 3.4) and combine models (cf. Section 3.5).

There are some technical details in keeping the explanatory variables exogenous, which are as follows.

(1) During the model search the explanatory variables are kept in a single clique since we are not interested in their interaction.
(2) When the recursive causal model is generated, the numbering of vertices starts from explanatory variables and proceed to other variables only when all explanatory variables have been numbered.
(3) The final model is generated with 'dummy' edges between the explanatory variables.

This means that the probability distribution is adjusted by the methods given in Section 3.2 for both the dummy edges needed for making the model decomposable and for making the explanatory variables exogenous.

In the model actually generated there is only one 'true' exogenous variable, i.e. $\sigma^{-1}(1)$. Because of the special structure of the probability distribution, i.e.

$$p(e|\Pi(e)) = p(e)$$

for an explanatory variable e, all the explanatory variables can be regarded as exogenous.

3.4 Adjusting marginal probabilities

We will now describe how marginal probabilities can be adjusted in a recursive causal model. This serves two purposes. First, it is a necessary basis for the combination of two recursive causal models, and secondly, it may be reasonable to adjust the marginal probabilities of a model constructed based on a limited data sample to those found in the environment where we intend to use the system.

We will assume that the conditional probabilities of the data reflect those of the total population. If this is not the case, the sample data does not have the same causal properties as the total population, which gives reason to scrutinize the data further.

The marginal probabilities of the probability distribution estimated by the data set used for knowledge engineering may differ from the corresponding marginal probabilities m estimated for the total population. In this case it is possible to adjust p to m. Let $(\mathrm{V}, \mathrm{E}, p)$ be a recursive causal model and m the distribution to which we wish to adjust. If for vertices $v \in \mathrm{V}_n$

$$m(v|\Pi(v)) = p(v|\Pi(v))$$

then the distribution defined by

$$p_1(\mathrm{V}) = \prod_{v \in V_x} m(v) \prod_{v \in V_n} p(v|\Pi(v))$$

will provide the same marginal probabilities in p_1 and m because

$$\begin{aligned} p_1(\mathrm{V}) &= \prod_{v \in V_x} m(v) \prod_{v \in V_n} p(v|\Pi(v)) \\ &= \prod_{v \in V_x} m(v) \prod_{v \in V_n} m(v|\Pi(v)) \\ &= m(V). \end{aligned}$$

This shows us that it is sufficient to adjust the marginal variables, provided that the conditional probabilities are correct.

3.5 Combining models

In this section we will discuss the combination of two recursive causal models. The motivation for this combination is the infeasibility of constructing a comprehensive knowledge base using a single statistical analysis. This is caused by the fact that the computational complexity of the calculation is exponential, and that the amount of data needed would be enormous. Our approach is therefore to use *a priori* expert knowledge of the domain to perform several separate analyses each concerned with its own subset of the variables.

Assume two recursive causal models, $(\mathrm{V}, \mathrm{E}, p)$ and $(\mathrm{W}, \mathrm{F}, q)$ with a set of common variables, $\mathrm{M} = \mathrm{V} \cap \mathrm{W}$. If M only consists of exogenous variables from the two models, they can be combined using the simple adjustment of the marginal probabilities shown above.

If the models have similar recursive structure, intuition tells us that it should be possible to combine them. This can be made more precise by the notion of *combinable models*.

Definition. Two recursive causal models $(\mathrm{V}, \mathrm{E}\, p)$ and $(\mathrm{W}, \mathrm{F}, q)$ can be *combined* if the identity mapping id: $\mathrm{M}_v \rightarrow \mathrm{M}_w$ is an isomorphism with respect to the relations E and F.

We will in the following construction also assume that the marginal probabilities for common exogenous variables are identical in the two models. If this is not the case, we may use the methods given in Section 3.4 to adjust them. The construction is the combination $\oplus$ of two recursive models, $(\mathrm{X}, \mathrm{H}, r) = (\mathrm{V}, \mathrm{E}, p) \oplus (\mathrm{W}, \mathrm{F}, q)$, defined as follows.

$$(\mathrm{X}, \mathrm{H}) = (\mathrm{W}, \mathrm{F}) \cup (\mathrm{V}, \mathrm{E})$$

$$r(v|\Pi_x(v)) = p(v|\Pi_v(v)) \qquad \text{for } v \in \mathrm{V} \setminus \mathrm{W}$$

$$r(v|\Pi_x(v)) = q(v|\Pi_w(v)) \qquad \text{for } v \in \mathrm{W} \setminus \mathrm{V}$$

$$r(v|\Pi_x(v)) = p(v) \qquad \text{for } v \in \mathrm{M} \text{ exogenous}$$

$$r(v|\Pi_x(v)) = [p(v|\Pi_v(v)) + q(v|\Pi_w(v))]/2 \qquad \text{otherwise}$$

Π_x, Π_v and Π_w are parents defined on models X, V and W respectively. The combination satisfies the property that if $(\mathrm{V}, \mathrm{E}, p)$ and $(\mathrm{W}, \mathrm{F}, q)$ are recursive causal

models and if we compute their combination $(X, H, r) = (V, E, p) \oplus (W, F, q)$ then for any variable v appearing in both the original models

$$r(v) = \frac{p(v) + q(v)}{2}$$

i.e. the marginal probabilities are given by the average of the marginal probabilities of the original distributions (see for example, Anderson & Krebs, 1989).

This result is an indication that the combination $\oplus$ is reasonable since the average is what one would expect if the two original models were weighted equally. It is also possible to assign weight to models according to the size of their data sets.

While this solves the problems with combining models regardless of the differences in conditional probabilities, major discrepancies for the common variables is in our opinion an indication of incompatible data sets.

4 The control strategy of the consultation phase

This section describes the control strategy of the shell; the control strategy is in this context an algorithm that defines the hypotheses and questions which are currently being investigated.

The control strategy should be restricted to the data available in the knowledge base, since the idea of building a verifiable knowledge base would be lost if further data had to be added to specify the control strategy. We have therefore refrained from defining a control program that will guide the user through the questions. In more practical terms, we will leave part of the control strategy to the user of the system, specifically when to stop the consultation and which hypotheses to consider.

Having 'solved' this part of the problem, we are left with the problem of ranking questions according to their information value during a consultation session with the user. Thus the purpose of the control program is basically to compute an ordering of the questions according to their 'relevance'. It is then left to the user to choose which of the questions to answer. Accordingly, the problem of the control strategy is to define a weight function for the questions.

With a medical application in mind, the purpose of the system can be defined to be of assistance to the doctor during consultation. Consequently a good question is one of a 'best' sequence of questions leading to the correct diagnosis. We have based the ranking of questions on their expected information contribution. Questions thought to supply most information are ranked first. When calculating the content of information, we have decided to use the mutual information $I(—)$ as the basic measure (Shore & Johnson, 1980). Let us define N as the set of vertices in the knowledge base, $H \subseteq N$ as the set of hypotheses and $Q \subseteq N$ as the set of questions. Note that Q and H may have common vertices.

The first suggestion of a weight function could be the *mutual information* $I(q\|H)$

$$w_0(q) = I(q\|H), \quad q \in Q$$

but for computational reasons this weight function is not practical. When computing the mutual information $I(q\|H)$ it is necessary to compute $p(q|H)$ for each configuration of the hypotheses. For large data bases this will be impossible, so we propose the weight function

$$w_1(q) = \sum_{h \in H} I(q\|h), \quad q \in Q$$

It can be shown that when all hypotheses are independent given q, then $w_0(q) = w_1(q)$. Obviously this rarely will be the case, but it gives us information about the behaviour of w_1. Furthermore, the difference between w_1 and w_0 can be described as follows: w_0 focuses on questions yielding the maximum decrease in the total entropy of the hypotheses, while w_1 focuses on questions yielding the maximum decrease in the sum of the entropy of the hypotheses.

The weight function described so far considers all hypotheses as being equally interesting. From a purely statistical point of view this may be satisfactory, but we believe that a mechanism for taking the users' interest in each hypothesis into account may be relevant. This idea of decomposing the diagnostic goal into subgoals using various diagnostic strategies has been advocated by Patil *et al.* (1982) among others. They classify diagnostic strategies into Confirm, Rule-Out, Differentiate, Group-and-Differentiate, Refine, and Explore. The choice of an appropriate strategy for a given situation may be based on the number of hypotheses in the primary goal, and their strength.

Therefore, we have extended w_1 with a mechanism to scale the influence of the hypotheses according to their current interest. Let $i(h)$ be a measure of the interest of hypothesis h, then

$$w_2(q) = \sum_{h \in \mathrm{H}} i(h) I(q \| h), \quad q \in \mathrm{Q}$$

Let Ω_h be the set of values of h, and $\Gamma_h \subset \Omega_h$ the set of *goal values*, i.e. values that are of current interest, e.g. a particular outcome of a hypothesis. Now define

$$p(\Gamma_h) = \sum_{x \in \Gamma_h} p(h = x)$$

and

$$i(h) = \begin{cases} p(\Gamma_h)^{\alpha}, & \alpha \in [0, \infty) \\ 1 - p(\Gamma_h)^{-\alpha}, & \alpha \in (-\infty, 0] \end{cases}$$

The interest function has been parameterized to enable user control. Note that if $\alpha = 0$ then $i(h) \equiv 1$ and hence $w_2 \equiv w_1$. The function is exponential in α which means that differentiation between the hypotheses is non-linear. During a consultation session various diagnostic strategies can be modelled by adjusting α appropriately.

We suggest the following four strategies.

Initial stage. The aim at the start of a consultation is to gather as much information as possible and no hypotheses should have priority. Therefore a value of 0 for α is preferable.

Intermediate stage. When the initial information has been gathered, but no hypotheses are definitely dominating, α should have a value between 0 and 1.

Final stage. Now one or more hypotheses have a goal probability significantly greater than the remaining, and the aim is to gather information about these hypotheses so that they may be confirmed or ruled out. A value greater than 1 should be given to α.

Rewind stage. This situation occurs when the user has reason to believe that a wrong hypothesis is pursued. When a value less than 0 is given to α hypotheses with low goal probability will be considered.

In this section we have defined the control strategy of the decision support program. We have proposed a weight function for the questions depending on the hypotheses

and a user-controllable parameter α. Note that although the control strategy is based upon a heuristic function, the probabilities and the conclusions will still be based on mathematically rigorous and reproducible methods.

5 Implementation

Our implementation of the principles described above runs on an IBM PC and consists of two programs.

(1) The model search program implements the procedure of Section 3.1 with the limitation that the program is capable of computing the distribution for decomposable models only.
(2) The consultation program contains the features described in Section 4, as well as a number of functions useful for simulation.

- Capability of observing and retracting evidence, with interactive computation of the resulting probabilities.
- The program contains links to the original data used for building the knowledge base. This makes it possible to verify the conclusions of the system, since the number of patients in the original data set that matches evidence entered can be computed interactively.

The remaining part of the theory, i.e. computing the probability distributions for non-decomposable graphical models and combinations of models, is being explored.

6 An illustrative example

The expert system has been used to simulate the diagnostic value of additional clinical information for coronary artery disease patients who might be considered for referral to coronary arteriography.

Patients suspected of having coronary artery disease are referred to coronary arteriography if their presenting features, disease manifestations and non-invasive clinical tests are indicative of coronary artery disease. Referral is based on demographic data (sex, age), risk factors (smoking, hypercholesterolaemia, hereditary dispositions, verified previous myocardial infarction) and results of laboratory tests (resting and exercise electrocardiogram (ECG)). It is desirable to achieve a sort of optimization in such a way that most patients with coronary artery disease (CAD) are referred to coronary arteriography while few without CAD are subjected to this invasive procedure. Table 1 lists the variables considered in the analysis.

Figure 5 (a) shows the complete graphical model obtained by the model search procedure and Fig. 5 (b) a simplified causal model obtained by lumping vertices a, t, c and h of Fig. 5 (a). From the model search procedure they appear to be strongly connected and are combined in one 'supernode' to simplify the graphical presentation of the model.

6.1 Evaluation of the system

The system was evaluated (i) by reclassifying the data set originally used for formulating the model (303 patients) and (ii) for classifying additional 50 patients not part of the original data set, for which data were obtained in a prospective study.

TABLE 1. Clinical variables for the coronary artery disease example

Node	Variable	Values
s	sex	male, female
a	angina pectoris	none, typical, atypical
A	previous AMI	not certain, definite
Q	Q wave in ECG	no, yes
q	q wave informative	no, yes
t	ST segment informative	yes, no
T	ST segment shift	no, yes
w	maximum workload	no, yes
h	left ventricular hypertrophy	no, yes
H	hypercholesterolaemia	no, yes
S	smoker	no, yes
I	hereditary predisposition	no, yes
K	congenital heart or valve disease	no. yes
c	coronary artery disease	no, yes

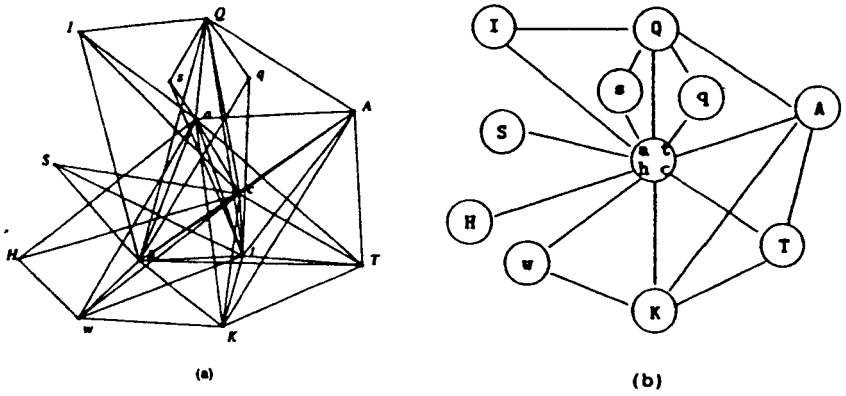


FIG. 5. (a) The graphical model obtained by the model search procedure, and (b) a simplified causal model obtained by combining four tightly connected vertices.

TABLE 2. Evaluation of STENO on the training set and a test set

p(CAD)	303 pts reclassified	50 new pts
0	0/88	0/6
1–10	3/46	0/7
11–50	10/37	7/13
51–89	18/27	8/8
90–99	18/21	3/3
100	84/84	8/13
Total	133/303	26/50

Patients are grouped according to the probability of CAD attributed by the expert system. For each group the fraction of patients actually having CAD (found by coronary arteriography) for the learning set and the test set are listed.

There were fewer variables recorded in the latter study. The information available in the second study was as follows.

(1) Presence or absence of coronary artery disease judged from coronary arteriography.
(2) Whether the patient had typical angina. There was no information on atypical angina, so the absence of typical angina is interpreted as no angina.
(3) Classification of the ECG with respect to ST segment changes, the categories being (i) ST changes in rest or exercise ECG, (ii) no ST changes in rest or exercise ECG and (iii) missing classification within this system. ST changes excluded hypertrophy and no ST change implied that maximum workload had been used during the exercise test. Missing classification was due to either submaximal workload in the exercise test or factors rendering the ST segment information useless (not informative).
(4) Q wave and previous verified acute myocardial infarction (AMI). These variables were merged into one variable. If this variable was positive it was taken as previous AMI and Q wave present, if negative no previous AMI and no Q wave present. This merging caused loss of information compared with the separated pieces of information used in the knowledge engineering phase.

Table 2 shows the results of reclassification of the original data set and results obtained on classification of the additional 50 patients who were not part of the original data set.

Table 2 shows for each classification of the system (p(CAD)) the fractions of patients in each class having coronary artery disease. Choosing 50% as the decision threshold, the number of false negative classified patients in the learning set is $FN_1 = 13/171 = 0.08$ and the number of false positively classified is $FP_1 = 12/132 = 0.09$. For the test set the corresponding figures are $FN_2 = 7/26 = 0.27$ and $FP_2 = 5/24 = 0.21$.

If patients in the interval $11 \leqslant p(\text{CAD}) \leqslant 89$ are considered 'not classified' (no definitive diagnosis) then $3/134 = 0.02$ were false negative and $3/105 = 0.03$ false positive in the learning set. The system could not classify $64/303 = 0.21$. In the test data set the corresponding figures are $0/13 = 0.00$ false negative and $5/16 = 0.31$ false positive, and the fraction of unclassifiable patients were $21/50 = 0.42$. We believe that the high false positive figure for the test set can be explained by the fact that the Q wave and AMI were merged in the test set.

In a corresponding analysis based on clinical classification by the combination of criteria of the same data set, Hansen (1982) found that out of 234 patients in a learning set, 94 could not be classified definitely, i.e. $94/234 = 0.40$. The fraction of false positive was $2/64 = 0.03$ and of false negative $1/76 = 0.01$. In a test set consisting of 63 there were $1/15 = 0.07$ false positive, $1/18 = 0.06$ false positive, and $30/63 = 0.48$ could not be classified. The high proportion of unclassified patients in the test set could possibly be explained by the fewer number of observed variables in this set. These figures suggest that system simulation can give valuable contribution in complex clinical diagnostic problems.

7 Conclusions

Our approach has been to use existing techniques for statistical model search for graphical models combined with expert *a priori* knowledge on relations between variables. The main advantage of this procedure is twofold. First, most of the

formulation of the expert system rests on experimentally obtained data sets and reproducible, generally acknowledged statistical techniques. Secondly, the model obtained may be communicated either as a statistical model or graphically as a recursive causal graph, the causal probabilistic interpretation of which is comparatively easy to understand, even for non-statistically orientated users.

The preliminary evaluation indicates that this approach may provide useful support in simulating the value of performing additional clinical tests, given that a certain knowledge on the patient is already available.

The programs described may be obtained by contacting Jens Damgaard Andersen (whose electronic mail address is jda@diku.dk).

Acknowledgements

The authors wish to thank Steffen L. Lauritzen and Svend Kreiner for stimulating discussions and valuable comments during the completion of this paper. We are also indebted to Dr. J. Fischer Hansen for inviting us to investigate the problem, for helpful conversations and for providing data sets.

Correspondence: Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen Ø, Denmark.

REFERENCES

ANDERSEN, L. R. & KREBS, J. H. (1989) *Design and Implementation of a Fully Verifiable Expert System Shell* (Department of Computer Science, University of Copenhagen).

DARROCH, J. N., LAURITZEN, S. L. & SPEED, T. P., (1980) Markov fields and log-linear interaction models for contingency tables, *Ann. Statist.*, 8, pp. 522–39.

HANSEN, J. F. (1982) *Diagnosen af koronararteriesygdom* (diagnosis of coronary artery disease) (in Danish) (Copenhagen, FADL Forlag).

KIIVERI, H., SPEED, T. P. & CARLIN, J. B. (1984) Recursive causal models, *Austral. Math. Soc. (Series A)*, 36, pp. 30–52.

KREINER, S. (1986) Computerized exploratory screening of large-dimensional contingency tables, in *Compstat 1986* (Heidelberg, Physica).

KREINER, S. (1987a) *Collapsibility of Multidimensional Contingency Tables: Theorems, algorithms, and programs* (The Danish Institute for Educational Research).

KREINER, S. (1987b) Analysis of multidimensional contingency tables by exact conditional tests: techniques and strategies, *Scand. J. Statist.*, 14, pp. 97–112.

KREINER, S. (1988) *Noter til analyse af kontingenstabeller* (The Danish Institute for Educational Research).

KREINER, S. (1989) On tests of conditional independence, *Research Report 89/14* (Statistical Research Unit, University of Copenhagen).

LAURITZEN, S. L. (1989) *Lectures on contingency tables,* 3rd edn (Aalborg, Aalborg University Press).

LAURITZEN, S. L. & SPIEGELHALTER, D. (1988) Local computations with probabilities on graphical structures and their applications to expert systems. *J. R. Stat. Soc. B,* 50, 2, pp. 157–224.

PATIL, R. S., SZOLOVITS, P. & Schwartz, W. B. (1982) Modeling knowledge of the patient in acid-base and electrolyte disorders, in: P. Szolovitz (ed.), *Artificial Intelligence in Medicine. AAS Selected Symposia No. 51* (Boulder, CO, Westview Press).

SHORE, J. E. & JOHNSON, R. W. (1980) Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inform. Theory,* IT-26, pp. 26–37.

TARJAN, R. E. & YANNAKAKIS, M. (1984) Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM J. Comput.,* 13, 3, pp. 566–579.

WERMUTH, N. & LAURITZEN, S. L. (1983) Graphical and recursive models for contingency tables, *Biometrika,* 70, pp. 537–552.