Handling Uncertainty in a Medical Study of Dietary Intake during Pregnancy

Adele Marshall[1], David Bell[2], and Roy Sterritt[2]

[1] Department of Applied Mathematics and Theoretical Physics
Queen's University of Belfast, Belfast, BT9 5AH, UK
a.h.marshall@qub.ac.uk

[2] School of Information and Software Engineering, Faculty of Informatics
University of Ulster, Jordanstown Campus, Newtownabbey, BT37 0QB
{da.bell, r.sterritt} @ulst.ac.uk

Abstract. This paper is concerned with handling uncertainty as part of the analysis of data from a medical study. The study is investigating connections between the birth weight of babies and the dietary intake of their mothers. Bayesian belief networks were used in the analysis. Their perceived benefits include (i) an ability to represent the evidence emerging from the evolving study, dealing effectively with the inherent uncertainty involved; (ii) providing a way of representing evidence graphically to facilitate analysis and communication with clinicians; (iii) helping in the exploration of the data to reveal undiscovered knowledge; and (iv) providing a means of developing an expert system application.

1 Introduction

This paper is concerned with handling uncertainty as part of the analysis of data from a medical study. The study is recording and analysing details of pregnant women at 28 weeks gestation to examine the relationship between the dietary intake of a mother and the birth weight of her baby. The study is an extension of a major new medical research programme, HAPO (Hyperglycaemia and Adverse Pregnancy Outcome), currently underway at the Royal Victoria Hospital, Belfast.

This paper describes the development of a decision support system (*expert system*) that will deal with (i) uncertainty in the dietary and associated data; (ii) preliminary analysis of the data; and (iii) collection of the data in the medical study itself.

1.1 Background and Previous Work

Consumption of a nutritionally adequate diet is of particular importance during pregnancy and has a considerable influence on birth outcome. Previous research has shown that sudden and severe restriction of energy and protein intake during pregnancy re-

D. Bustard, W. Liu, and R. Sterritt (Eds.): Soft-Ware 2002, LNCS 2311, pp. 206–216, 2002.

duces birth weight by as much as 300g [1]. However, studies have also shown that an adequate energy and protein consumption may not be sufficient without the accompaniment of vitamins, minerals [2] and even fatty acids [3].

There is substantial evidence in the literature to suggest a strong association with low birth weight and increased incidence of neonatal mortality [4] and higher neonatal morbidity [5]. Impairment of foetal growth and low birth weight due to inadequate maternal and foetal nutrition can also increase the risk of chronic diseases in adulthood [6]. Maternal diet not only influences the immediate outcome of pregnancy but also the longer-term health of her offspring [7], possibly influencing susceptibility to diseases such as ischaemic heart disease, hypertension and Type 2 diabetes.

Additionally, there is evidence in the literature to suggest a relationship between high birth weight, caused by excessive foetal growth (macrosomia), and potential risks such as difficulties in childbirth and intensive postnatal care. These problems are most commonly associated with diabetic pregnancy. This has led to numerous attempts to relate complications in pregnancy and birth outcome to the level of maternal glycaemia [8], [9]. However, these studies have not been able to identify a possible threshold level of glycaemia above which there is a high risk of macrosomia.

1.2 The HAPO Medical Study

The US National Institute of Health (NIH) has recently approved an extensive international HAPO (Hyperglycaemia and Adverse Pregnancy Outcome) study in 16 key centres around the world. The Royal Victoria Hospital is one of these centres. Each centre will recruit 1500 pregnant women in the study over a two year period.

Research is primarily concerned with the following hypothesis used to investigate the association between maternal blood glucose control and pregnancy outcome.

Hypothesis: Hyperglycaemia in pregnancy is associated with an increased risk of adverse maternal, foetal and neonatal outcomes.

In addition to the specified HAPO variables, the Royal Victoria Hospital will be collecting information on dietary intake of the pregnant women in the study. Information gathered will include:

- anthropometric measures
- socio-economics
- family history
- metabolic status
- food frequency and diet
- measures of pregnancy outcomes

The details are recorded at 28 weeks gestation during an interview where a food frequency questionnaire is completed. Additional information is gathered about the pregnancy outcome and details recorded on the newborn baby.

The calendar of events for the HAPO study is shown in Figure 1. It shows the initial recruitment and data collection using the food frequency questionnaire followed by the pregnancy outcome variables and home visit follow-up. A follow-up is also scheduled to record information on the general health of the baby when 2-3 years old.

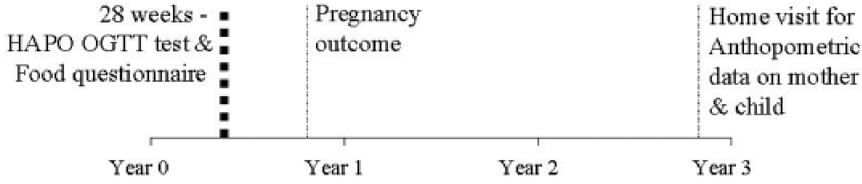

Fig. 1. HAPO event calendar.

The food frequency questionnaire was designed to gather relevant information as the dietary intake of the pregnant mothers. The food frequency questionnaire is responsible for the collection of information on the frequency that pregnant women eat food in each of the various food groups. There is a high level of complexity and uncertainty involved in determining the nutritional value and content of various different foods in addition to capturing the effect of variations between manufacturers and products. After consultation with experts it was advised that the food frequency questionnaire deal with this problem by focusing on groups of food types such as cereals, meat / fish and so on, as summarised in Figure 2.

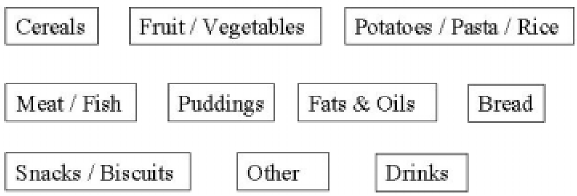

Fig. 2. Food groups analysed in the HAPO study.

The food groups may be associated with different nutritional levels. For example, meat and fish will have a strong association with the nutritional influence of proteins whereas cereals will be strongly associated with folic acid, a nutrient that plays an important role in the prevention of neural tube defects.

There are five parts to the study, evaluating:

1. Dietary intake in pregnancy compared to non-pregnant women of childbearing age.
2. Links between diet and lifestyle/socio-economic factors.
3. Links between diet and maternal/foetal glucose.
4. links between dietary intake and pregnancy outcome.
5. Possible follow-up of links between dietary intake in pregnancy and both maternal and outcome at two years.

Part 1 is a control to assess whether there is a significant difference in the diet of pregnant women in relation to other women. Parts 2 to 5 can be supported with the introduction and development of a decision support (expert) system for assessing the dietary variables for pregnant women.

1.3 Expert Systems

An expert system is a computer program that is designed to solve problems at a level comparable to that of a human expert in a given domain [10]. Rule-based expert systems have achieved particular success in medical applications [11] largely because the decisions that are made can be traced and understood by domain experts. However, their development has suffered due to the difficulty of gathering relevant information, an inability to handle uncertainty adequately, and the maintenance problem of keeping the stored knowledge up-to-date [12].

Another possible AI technique that can be used within an expert system is *neural networks*. Unfortunately, clinicians and healthcare providers have been reluctant to accept this approach because neural networks provide little insight into how they draw their conclusions, leading to a lack of confidence in the decisions made [13].

Bayesian belief networks (BBNs) are another possibility. A key advantage is their ability to represent potentially causal relationships between variables making them ideal for representing causal medical knowledge. The graphical nature of BBNs also makes them attractive for communication between medical and non-medical researchers. BBNs allow clinicians a better insight into the workings of the model thereby improving their confidence in the process and the decisions made.

Example of relevant work in this area include

- Research into the graphical *conditional independence* representation of BBNs as an expert system for medical applications [14].
- The application of BBNs in the treatment of kidney stones [15].
- The diagnosis of patients suffering from *dyspnoea* (shortness of breath) [16].
- The representation of medical knowledge [17].
- Assistance in the assessment of breast cancer [21].
- Assessing the likelihood of a baby suffering from congenital heart disease [22].

In addition, various other applications have been developed in biology for prediction and diagnosis [13].

2 Research and Methodology: Handling Uncertainty

2.1 Bayesian Belief Networks

Bayesian belief networks (BBNs) are probabilistic models which represent potential causal relationships in the form of a graphical structure [18], [19]. There has been much discussion and concern over the use of the term *causality*. In the context of BBNs, causal is used is in the representation of relationships which have the potential to be causal in nature—when one variable directly influences another. [20] explain this reasoning by stating that it is rare that firm conclusions about causality can be drawn from one study but rather the objective is to provide representations of data that are potentially causal—those that are consistent with, or suggestive of causal interpretation. For the potential casual relationships to be considered truly causal, external intervention is required to attach an overall understanding to the whole process of generating the data and the deeper concepts involved.

The BBNs are a set of variables and a set of relationships between them, where each variable is a set of mutually exclusive events. Relationships are represented in graphical form. The graphical models use Bayes' Theorem, the probability of an event occurring given that some previous event has already taken place. Uncertainty in the data can therefore be handled appropriately by the probabilistic nature of the BBN.

The structure of the BBN is formed by nodes and arrows that represent the variables and causal relationships presented as a directed acyclic graph. An arrow or directed arc, in which two nodes are connected by an edge, indicates that one directly influences the other. Attached to these nodes are the probabilities of various events occurring. This ability to capture potential causality supported by the probabilities of Bayes' Theorem makes the BBN appealing to use.

2.2 Development Process of the BBN

There are many ways of constructing a BBN according to the amount of expert knowledge and data available. If there is a significant uncertainty in the expert knowledge available, learning of the BBNs may only occur through induction from the data. Alternatively, a high contribution of expert advice may lead to the initial development of the BBN structure originating entirely from the experts, with probabilities attached from the data.

A purely expert-driven approach is unattractive because of the basic difficulty in acquiring knowledge from experts. Such a problem can be alleviated by supplying experts with data induced relationships. A pure data approach would often seem ideal, providing the opportunity of discovering new knowledge along with the potential for full automation. However, the data approach will not always capture every possible circumstance and still demands the attention of expert or human assistance for interpretation purposes—a discovery only becomes knowledge with human interpretation. It is between these two extremes that most developments lie, but that in turn makes the development process unclear.

A BBN combines a mixture of expert consultation and learning from data. The objective is to achieve the best of both worlds and minimise the disadvantages of each source, thus reducing the uncertainty. Generally, the approach consists of three main stages:

HUMAN: Probabilities and structure defined by consultation with experts and literature.

SYSTEM: Learning of structure and probabilities from data.

COMBINE: Knowledge base amended with discoveries and probabilities obtained from the data, or the structure induced from data is adjusted to include expert reasoning.

As the intention is to investigate the nature of BBNs as a research development technique, one possible approach is to run the *system* and *human* stages in parallel without any collaboration andcompare the outcomes in the *combine* stage. Alternatively, BBNs may be developed using any combination of the three development components on the first data set. Then, as the data collection continues the process of developing BBNs can also continue, with various BBNs developed for the growing data set. This exercise will indicate the benefits of using BBNs as a research development technique as more and more discoveries on the data are obtained.

The focus of this paper is the *system* component of the development process in which the structure and probabilities are derived from analysing the study questionnaires for implicit causal relationships. This will be followed by a second stage derivation of a BBN from the experts involved in the study, with the final outcome being a comparison of the resulting BBNs.

In reality, the process, *human-system-combine* or *system-human-combine* is similar to an evolutionary development process since a re-examination or re-learning of the system will be required throughout the study at significant stages to include new data. This process should facilitate fine-tuning in the development of the causal network.

2.3 Expected BBN Formulated from Literature

The research literature reveals various potential causal relationships that can be represented in a BBN structure. For example, one simple causal relationship [1], is the direct influence of the level of proteins in the diet of the mother on the final birth weight of the baby. This may be represented as part of a BBN model as shown in Figure 3.

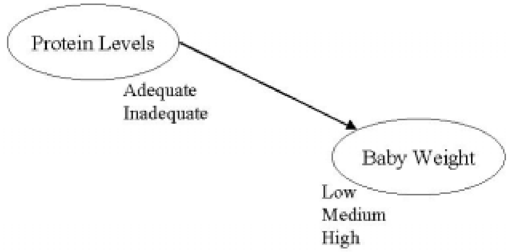

Fig. 3. BBN representation of a causal relationship.

The hypothetical BBN in Figure 4 represents potential causal relationships induced from consultation of the literature and the study domain. The model considers the foods classified into their basic food type. The arrows indicate direct influences. For example, in this model, the food groups proteins, vitamins and minerals and fatty acids will all have potential causal influences on the birth weight of the baby.

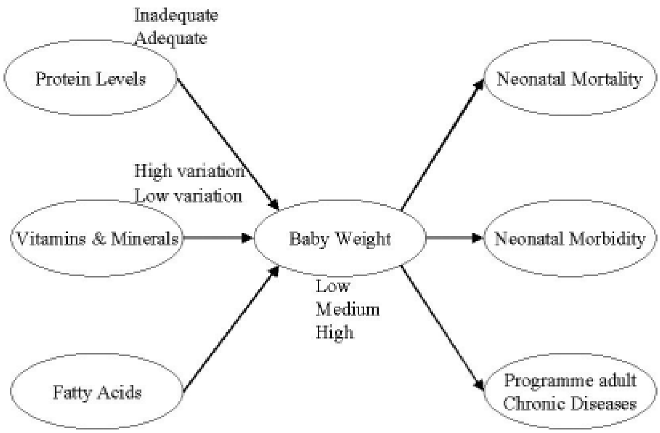

Fig. 4. A hypothetical BBN for the HAPO study.

The birth weight of the baby will in turn have a direct influence on its survival. For example a baby with a very low birth weight may be more likely to die. Also taken into account in the BBN structure is the influence on baby weight on variables obtained from the follow-up study—for example, the occurrence of adult chronic diseases, such as ischaemic heart disease. At the current stage in the data collection, it is only possible to hypothesise about such relationships as follow-up variables will not be available to the research team until at least 2003 and throughout the babies development into children and adulthood.

3 Preliminary Results

The medical study has now started with the first set of participants' details being recorded. Currently, the study has recruited and interviewed 294 women, 108 of whom have had their babies. Preliminary statistical analysis has been carried out on this initial data set. The pregnancy outcomes include measures recorded for the new-born baby along with additional information such as delivery type and mothers condition. Baby information includes the baby weight, length, head circumference and sex. In this sample of data, the birth weights of the 108 babies ranged from 1.9kg to 4.93kg (4.18-10.85 lbs.) with an average of 3.51kg (7.7 lbs.). Variables identified as having a direct influence on baby weight include the frequencies at which pasta, bread and potatoes are consumed. Other potential influencing factors are the number of cigarettes smoked, the number of children that the mother has already, and whether there are any relatives who have diabetes. The statistical analysis performed on the data set has indicated some useful observations on the variables that have a possible influence on the baby outcomes. To investigate these relationships further, it would be useful to construct a BBN.

3.1 Resulting BBNs

The focus of this paper is the development of an initial BBN in which the structure and probabilities are derived from analysing the data collected from the study questionnaires. It is hoped that this will be followed by a second stage derivation of a BBN from the experts involved in the study.

The BBNs are constructed using the PowerConstructor package [23]. PowerConstructor takes advantage of Chow and Lui's algorithm [24] which uses mutual information for learning causal relationships and enhances the method with the addition of further procedures to form a three-stage process of structure learning from the data. The first phase (drafting) of the PowerContructer software utilises the Chow-Liu algorithm for identifying strong dependencies between variables by the calculation of mutual information. The second stage (thickening) performs conditional independence (CI) tests on pairs of nodes that were not included in the first stage. Stage 3 (thinning) then performs further CI tests to ensure that all edges that have been added are necessary. This three-stage approach manages to keep to one CI test per decision on an edge throughout each stage and has a favourable time complexity of $O(N^2)$ unlike many of its competitors which have exponential complexity.

Preliminary analysis was carried out using the PowerConstructor package on the food frequency variables for the first 108 pregnant mothers along with the outcome variable, the baby's birth weight. The BBN in Figure 5 was induced from the data.

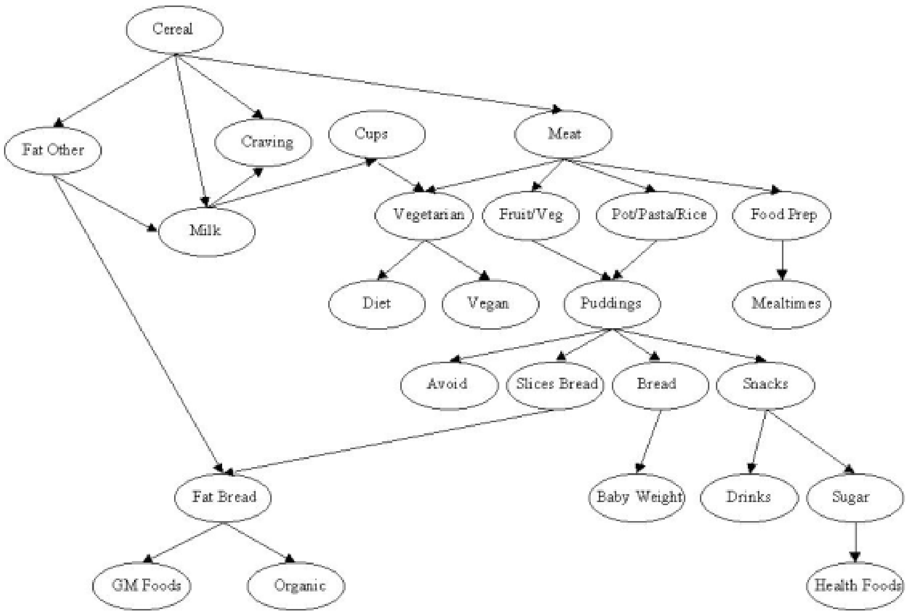

Fig. 5. Initial BBN for the birth weight outcome and food frequency variables using [23].

As before, the variables are represented by ovals in the structure while the edges between the ovals represent potential causal relationships between the variables. From inspection of the BBN, it is apparent that there are many inter-relationships between the dietary intake variables, but already at this early stage in the data collection a relationship is emerging for the baby's birth weight. This may be made clear by removing some of the less significant variables and repeating the induction of the BBN. The resulting BBN is shown in Figure 6.

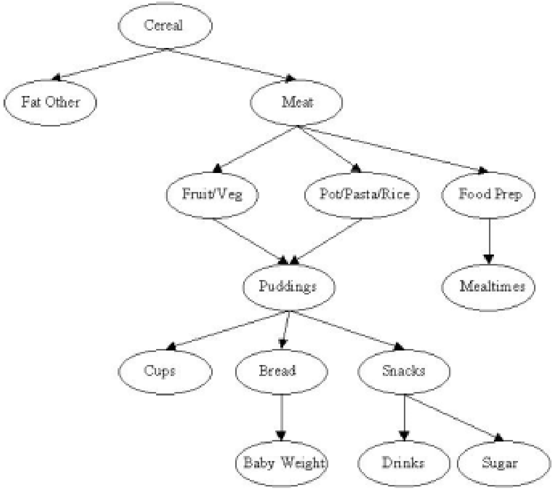

Fig. 6. BBN representing the birth weight outcome along with food type variables using [23].

The relationship emerging from the BBN is that the baby's birth weight seems to be directly influenced by the variety of bread the mother consumes during pregnancy. In fact if the probabilities are considered it is evident that the greater the variety of bread consumed, the greater the probability of a larger baby. In addition to this, the variable bread is in turn influenced by the consumption of puddings which is in turn influenced by the consumption of fruit and vegetables, potatoes, pasta, rice which are influenced by meat and cereal.

The BBN in Figure 6 captures some of the significant relationships on baby weight. However, as the data set grows, it is expected that the number of edges in the BBN will also increase. It is hoped that the BBNs will not only be a tool for the representation of the evolving model but also as a research development technique to aid discussion as the study progresses, helping to identify further causal relationships.

4 Conclusions and Observations

The paper has discussed the use of Bayesian belief networks (BBNs) for handling uncertainty in a medical study. The study is concerned with modelling the influencing factors of a mother's dietary intake during pregnancy on the final birth outcomes of the baby. In particular, the baby's birth weight is important as this may cause further complications for both the mother and baby.

A food frequency questionnaire designed to record the frequencies of consumption of various different foods is being used to collect dietary information. There is a high level of complexity and uncertainty involved in determining the nutritional value and content of various different foods and in capturing the effect of variations between manufacturers and products. In addition to developing a system that can handle such uncertainty, the development project itself is a research project with undiscovered knowledge and unproven hypotheses.

The objectives are to develop a system that can represent relationships between dietary and outcome variables, to be able to handle uncertainty, while also producing results in such a way that can be understood by clinicians. BBNs seem to be an appropriate technique for such a challenge. Advantages of using them include their ability to represent potentially causal relationships, their visual graphical representation and their capability of dealing with uncertainty using probability theory.

The systems or software aspect of this project is to engineer an intelligent system. Ideally this would involve acquiring or learning from an environment with proven hypotheses; however, the medical research in this project is running in parallel so there is uncertainty in the expert knowledge available. Thus, the process to engineer the system should also assist in expressing the evidence contained in the evolving study to the medical experts and the system designers as well as exploring the study data as it is gathered for undiscovered knowledge.

The study is currently at the data collection stage and further evolution of the BBNs will follow. It is hoped that an analysis of this evolution will also provide interesting insights into the BBN development process.

Acknowledgements. The authors are greatly indebted to our collaborators Mrs. A. Hill (Nutrition and Dietetics, Royal Hospital Trust), Professor P.R. Flatt and Dr. M.J. Eaton-Evans (Faculty of Science, University of Ulster, Coleraine), Dr. D. McCance and Professor D.R. Hadden (Metabolic Unit, Royal Victoria Hospital).

References

1. Stein, Z., Fusser, M.: The Dutch Famine, 1944-45, and the Reproductive Process. *Pediatric Research* 9 (1975) 70-76
2. Doyle, W. Crawford, M.A., Wynn, A.H.A., Wynn, S.W.: The Association Between Maternal Diet and Birth Dimensions. *Journal of Nutrition and Medicine* 1 (1990) 9-17
3. Olsen, S.F., Olsen, J., Frische, G.: Does Fish Consumption During Pregnancy Increase Fetal Growth? *International Journal of Epidemiology* 19 (1990) 971-977
4. Bakketeig, L.S., Hoffman, H.J., Titmuss Oakley, A.R.: Perinatal Mortality. In: Bracken M.B. (ed.): *Perinatal Epidemiology*, Oxford University Press, New York, Oxford (1984) 99-151
5. Walther, F.J., Raemaekers, L.H.J.: Neonatal Morbidity of SGA Infants in Relation to their Nutritional Status at Birth. *Acta Paediatric Research Scandia* 71 (1982) 437-440

6. Barker, D.J.P.: The Fetal and Infant Origins of Adult Disease. *British Medical Journal*, London (1992)
7. Mathews, F., Neil, H.A.W.: Nutrient Intakes During Pregnancy in a Cohort of Nulliparous Women. *Journal of Human Nutrition and Dietetics* 11(1998) 151-161
8. Pettitt, D.J., Knowler, W.C., Baird, H.R., Bennett, P.H.: Gestational Diabetes: Infant and Maternal Complications in Relation to 3rd Trimester Glucose Tolerance in Pima Indians. *Diabetes Care*, 3 (1980) 458-464
9. Sermer, M., Naylor, C.D., Gare, D.J.: Impact of Increasing Carbohydrate Intolerance on Maternal-fetal Intrauterine Growth Retardation. *Human Nutrition and Clinical Nutrition* 41C (1995) 193-197
10. Cooper, G.F.: Current Research Directions in the Development of Expert Systems Based on Belief Networks. *Applied Stochastic Models and Data Analysis* 5 (1989) 39-52
11. Millar, R.A.: Medical Diagnostic DSS - Past, Present and Future. *JAMIA* 1 (1994) 8-27
12. Bratko, I., Muggleton, S.: Applications of Inductive Logic Programming. *Comm. ACM* 38(11) (1995) 65-70
13. Lisboa, P.J.G., Ifeachor, E.C., Szczepaniak P.S. (eds.): *Artificial Neural Networks in Biomedicine*. Springer (2000)
14. Andersen, L.R., Krebs, J.H., Damgaard, J.: STENO: An Expert System for Medical Diagnosis Based on Graphical Models and Model Search. *Journal of Applied Statistics* 18 (1991) 139-153
15. Madigan, D.: Bayesian Graphical Models for Discrete Data. Technical Report, *University of Washington*, Seattle (1993)
16. Lauritzen, S.L., Spiegelhalter, D.J.: Local Comparisons with Probabilities on Graphical Structures and their Application to Expert Systems. *Journal Royal Statistical Society* B 50(2) (1988) 157- 224
17. Korver, M., Lucas, P.J.F.: Converting a Rule-Based Expert System into a Belief Network. *Med. Inform.* 18(3) (1993) 219-241
18. Buntine, W.: Graphical Models for Discovering Knowledge. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusay, R. (eds.): *Advances in Knowledge Discovery*, AAAI Press / The MITPress (1996) 59-82
19. Ramoni, M., Sebastiani, P.: Bayesian Methods for Intelligent Data Analysis. In: Berthold, M., Hand, D.J. (eds.): *Intelligent Data Analysis: An Introduction*, Springer, New York (1999)
20. Cox, D.R., Wermuth N.: Multivariate Dependencies, Chapman and Hall (1996)
21. Hojsgaard, S., Thiesson, B., Greve, J., Skjoth, F.: BIROST - A Program for Inducing Block Recursive Models from a Complete Database, Institute of Electronic Systems, Department of Mathematics and Computer Science, *Aalborg University*, Denmark (1992)
22. Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., Cowell, R.G.: Bayesian Analysis in Expert Systems, *Statistical Science* 8(3) (1993) 219-283
23. Cheng, J., Bell, D.A., Liu, W.: An Algorithm for Bayesian Network Construction from Data. *6th International Workshop on AI and Stats* (1997) 83-90
24. Chow, C. J. K., Liu, C. N.: Approximating Discrete Probability Distributions with Dependence Trees, *IEEE Trans. Information Theory*, Vol. 14(3), (1968) 462-467