Probabilistic Bayesian Network Model Building of Heart Disease[1]

Jayanta K. Ghosh and Marco Valtorta
{ghosh,mgv}@cs.sc.edu
Department of Computer Science, University of South Carolina
Columbia, SC 29208

Introduction

A Bayesian network – also called a belief network or causal probabilistic network- is a graphical representation of probabilistic information: It is a directed acyclic graph in which nodes represent random (stochastic) variables, and links between nodes represent direct probabilistic influences between the variables. In this formalism, propositions are given numerical probability values signifying the degree of belief accorded them, and the values are combined and manipulated according to the rules of probability theory. Typically, the direction of a connection between nodes indicates a causal influence or class-property relationship [1].

Bayesian statistical inference uses probabilities for both prior and future events to estimate the uncertainty that is inevitable with prediction. The fundamental concept in Bayesian networks is that probabilities can be assigned to parameter values, and through

November 30, 1999 (Use mgv@cse.sc.edu for correspondence)
Technical Report TR9911
USCEAST-CSTR-IY99-11
We thank Dr. Kailash Mathur of S.C. State University for letting us use the data collected during the project "Cholesterol, Selected Minerals and Health Status of the Elderly in South Carolina." Both authors acknowledge support from DARPA through the project "Resource Allocation in Uncertain Domains (TargetShare)."

Bayes' theorem, these probabilities can be updated given new data. In Bayesian models the parameter is viewed as a domain variable, with a probability distribution, since the actual value of the parameter is unknown. The causal links between the variables are represented by arrows in the model. The model is strong if the arrows can be interpreted as causal mechanisms. The relationships can, alternatively, be considered an association and this type of model would be viewed more cautiously.

Bayesian networks can express the relationships between diagnoses, physical findings, laboratory test results, and imaging study findings. Physicians can determine the *a priori* ("pre-test") probability of a disease, and then incorporate laboratory and imaging results to calculate the *a posteriori* ("post-test") probability [2]. Bayesian networks can be used to plan diagnostic tests and therapeutic intervention [3]. Efforts are underway to formulate large, general medical decision support systems such as Iliad [4] and QMR [5] into Bayesian networks.

Model Building

Data of 167 elderly subjects used in this model building were randomly collected in our earlier research project entitled "Cholesterol, Selected Minerals and Health Status of the Elderly in South Carolina" at S. C. State University, Orangeburg, SC [6]. The above study took an interdisciplinary look at a variety of factors that affect older adults' well being. The interaction of these variables could yield further information for the

M.V. also acknowledges support from the Office of Naval Research through grant N00014-97-1-0806.

development of nutritional guidelines and exercise recommendations for this population. This exploratory study had four main objectives.

- To assess serum cholesterol values in an elderly population.
- To assess the Calcium, Copper and Magnesium status in an elderly population 65 years old and older.
- To assess the dietary intake of nutrients using 24-Hours Diet Recall.
- An activity questionnaire was completed to assess subject's physical activity level.

The above study used a modified cluster stratified sampling technique, which is an effective method of choosing subjects. Subjects were stratified on four levels of independence since this variable can affect the intake of nutrients, health care and health status of individuals. Seventeen variables were carefully chosen from the above study based on relevant information reported in the literature [6] and also the author's expertise in biochemistry, nutrition and Public health research.

The network was constructed using the Hugin graphic interface [1]. Directed causal links were drawn from the parent variable to its children. The various levels (states) of the variables were also entered. For variables without parents, prior probabilities of the various states were entered. For the variables with one or more parents, the designer of the model assigned conditional probabilities. The network was ready at this point and the prior probabilities of the different states of all the variables were read from the network monitor.

Results and Discussion

Bayesian networks represent a promising technique for clinical decision support and provide a number of powerful capabilities for representing uncertain knowledge. They provide a flexible representation that allows one to specify dependence and independence of variables in a natural way through the network topology. Because dependencies are expressed qualitatively as links between nodes, one can structure the domain knowledge qualitatively before any numeric probabilities need to be assigned.

The schematic diagram (Figure 1) shows the probabilistic model of heart disease. Out of 17 variables, 13 were identified as predecessors and 4 as children of heart disease. The predecessors, Atherosclerosis, High BP, Family History, Serum Selenium and Adverse Medicine were identified as parents, and ECG, Angina Pectoris, Miocardial Infraction and Rapid Heartbeats as children of heart disease. Some of the predecessors do not influence heart disease directly but influence through their respective parent nodes. For example, smoking and alcohol and obesity do not influence heart disease directly but through their common parent node, High BP.

Parent nodes established predictive (or causal) reasoning, i.e. cause to effect, and children nodes established diagnostic (or evidential) reasoning, i.e. effect to cause.

The variables and their states, and conditional probabilities are enumerated in

Tables 1- 6. In order to simplify the calculation of conditional probabilities of the variables, we have assumed that the model typically behaves like a *Noisy-OR-gate* with discrete binary nodes. There are two reasons why the noisy-OR model is valuable. The

first reason is that the model can easily be expanded to many diseases. Each new cause of heart disease only requires the specification of one conditional probability to generate an exponentially growing table of conditional probabilities for heart disease. The second reason, which is probably the most important, is that the obvious difficulties involved in providing statistical data for all possible combinations of 13 predecessors that may all or combination of all may cause heart disease. In this case we are better off replacing statistics by some plausible assumption about the interaction of the different causes of heart disease, allowing us to generate plausible conditional probabilities that cannot be supported by statistics from the few conditional probabilities that can be supported by statistics. However, the Noisy-OR-gate assumes that the variables act independently and there are no synergestic or antagonistic effect of variables on heart disease [7]. This is one of the limitations of the model.

The conditional probability tables have been generated (Tables 1 – 6) assuming Noisy-OR gate approximation. For the sake of clarity and understanding, we have given the assumptions and calculations involved in the conditional probability of Atherosclerosis given Serum Triglycerides, Serum LDL, Moderate Exercise and Cholesterol HDL Ratio. There are five events that cause Atherosclerosis.

- Background causes are responsible for Atherosclerosis 20% of the time.
- Serum Triglycerides (High) cause Atherosclerosis with probability 0.60.
- Serum LDL (High) causes Atherosclerosis with probability 0.40.
- Moderate Exercise (No) causes Atherosclerosis with probability 0.10.
- Cholesterol HDL Ratio (High) causes Atherosclerosis with probability 0.10.

The Noisy-OR approximation can be interpreted in the following way. If any one of the causes is present then there would be Atherosclerosis, unless something has prevented it. For example, there is a 40% of chance that some inhibitor prevents Atherosclerosis when Serum Triglycerides are high. We assumed that these preventing factors are independent. So the combined probabilities are easy to calculate as one minus the product of the probabilities for the appropriate inhibitors (Note that the background causes are always a fact). For Example, the conditional probability of Atherosclerosis when Serum Triglycerides, Serum LDL, Cholesterol HDL Ratio are high and Moderate Exercise is absent is given by the following expression:

P (Atherosclerosis | Serum Triglycerides = high, Serum LDL = high,

Cholesterol HDL Ratio = high, Moderate Exercise = No, Background Causes)

$= 1 – (1 – 0.6)(1 – 0.4)(1 – 0.1)(1 – 0.1)(1 – 0.2) = 0.84$

Similarly, P (Atherosclerosis | Serum Triglycerides = low, Serum LDL = low,

Cholesterol HDL Ratio = low, Moderate Exercise = Yes, Background Causes)

$= (1 – 0.8) = 0.20$

The conditional probabilities of other fourteen cases can be calculated in an analogous way. The conditional probabilities of all sixteen cases have shown in Table 3.

We generated seven cases with instantiation for a subset of variables to be sure that the model is working effectively. We ran these cases on the model and came up with the probabilities for each of the variables. In the first case (Figure 2), there was no evidence, and the model calculated the probability of heart disease based on probabilities of the

input variables. We found that our subjects have 64% probability of heart disease, which justified our findings of lipid (cholesterol) profile report that the subjects belong to a moderate risk category of heart disease.

In the second (Figure 3) and third (Figure 4) cases we have introduced evidence of adverse and good health respectively, and the model calculated the probabilities of heart disease. In the case of adverse health scenario, the evidence was that all of the predecessors of heart disease, except for Adverse Medicine and Serum Selenium, were in the adverse state. The model correctly updated the probabilities of every other variable. In particular, the probability of heart disease was updated to 93%, the probability of abnormal Serum Selenium, a parent of heart disease, increased from 46% in the no-evidence case to 70%, and the probability of abnormal ECG, a child of heart disease, increased from 62% in the no-evidence case to 88%. Similarly in case of good health scenario, the evidence was that all of the predecessors of heart disease, except for Adverse Medicine and Serum Selenium, were in the good state. The model correctly updated the probabilities of every other variable. In particular, the probability of heart disease was updated to 17%, the probability of abnormal Serum Selenium, a parent of heart disease, decreased from 46% in the no-evidence case to 30%, and the probability of abnormal ECG, a child of heart disease, decreased from 62% in the no-evidence case to 20%. This small probability of heart disease (17%) (in case of evidence of good health) can be attributed to a leak probability in the *Noisy-OR model* i.e. some other factors may be responsible for heart disease, which the model did not take into account. This is another limitation of the model.

In cases four and five, the model calculated the marginal probabilities of each variable when heart disease was present and absent, respectively.

In case four (Figure 5), the marginal probabilities of the parents nodes i.e. Atherosclerosis, High BP, Serum Selenium increased considerably as compared to case one. Similarly the marginal probabilities of all the children nodes increased considerably in case four as compared to case one. This is justified by the causal links between heart disease and its parents and heart disease and its children. In case five (Figure 6), our observation was just the opposite of what we observed in case four. One significant observation in this case was the increase in marginal probabilities of normal ECG (95%) and no Miocardial Infraction (95%) when heart disease was absent. This suggests that ECG and Miocardial Infraction are the most specific diagnostic tools for heart disease. Combining cases four and five we conclude that ECG is both a very sensitive and specific diagnostic tool for heart disease, whereas Miocardial Infraction is very specific but not as sensitive as ECG.

Case six (Figure 7) justified the causal links between heart disease and its parents. When Atherosclerosis, High BP and Serum Selenium were forced to be adverse, the probability of heart disease increased significantly (64% to 91%) with P (evidence) = 0.124406. This established the fact that among its parents Atherosclerosis, High BP and Serum Selenium contribute significantly towards heart disease. On the contrary in case seven (Figure 8), when we increased the probabilities of Adverse Medicine and Family History to maximum, keeping the probabilities of other three parent nodes unchanged, the

probability of heart disease increased very little (64% to 72%) with P (evidence) = 0.03. The change in probability of heart disease suggests that Adverse Medicine and Family History do not contribute significantly towards heart disease as compared to Atherosclerosis, High BP and Serum Selenium. But when either High BP or Atherosclerosis was in the adverse state along with Adverse Medicine and Family History, the probability of heart disease increased to 88%. The large difference between the probability of evidence in the two cases (case six and seven) suggests that Adverse Medicine and Family History are not sufficient causal factors for heart disease.

Here are some comments on additional test cases. We noticed that among the children of heart disease, when ECG or Miocardial Infraction was in the adverse state, the probability of heart disease increased very substantially (64% to 97%), but when Angina Pectoris or Rapid Heartbeats was in the adverse state, the probability of heart disease increased only marginally: from 64% to 79% and from 64% to 85% respectively. When High BP was in the adverse state and the other four parents were not observed, the probability of heart disease increased from 64% to 80%, and when Atherosclerosis was in the adverse state and the other four parents were not observed, the probability of heart disease increased from 64% to 81%. The results on these cases also suggested that High BP and Atherosclesosis are important factors for heart disease, and ECG and Miocardial Infraction are important diagnostic tools of heart disease.

Conclusions

- The model should be treated only as a prototype.
- The variables of the model act independently i.e. the model did not take into account any synergistic or antagonistic interactions of the variables that might influence the heart disease [7].
- The model did not take into account other unknown factors that might influence heart disease.
- The available test results show that the model is working well with the available data.
- ECG is both a very sensitive and specific diagnostic tool for heart disease, whereas Miocardial Infraction is very specific but not as sensitive as ECG.
- Adverse Medicine and Family History do not contribute significantly towards heart disease as compared to Atherosclerosis, High BP and Serum Selenium.
- Careful thoughts and consideration should be given to find the causes of heart disease.
- Synergistic and antagonistic interactions of the variables should be taken into consideration that might influence the heart disease.
- More data need to be collected to make sure that the model is fully functional.
- Our model can be improved by using more recent datasets and by addition of new variables or editing the directed causal links. The available data made it impossible to improve the accuracy further.
- The model should be validated using a different dataset than the one used for its construction.

Table 1. Conditional probability of Heart Disease given Atherosclerosis, Serum Selenium, High Blood Pressure, Family History and Adverse Medicine.

Abbreviation used: Atheresclerosis (ATS), Serum Selenium (SSM), High Blood Pressure (HBP), Family History (FH), Adverse Medicine (ADM), and Heart Disease (HD) and Probability (P).

States used: ATS (Yes, No), SSM (Norm, Abnorm), HBP (High, Norm), FH (Yes, No), ADM (Yes, No).

ATS	SSM	HBP	FH	ADM	P(HD)	P(¬HD)
Yes	Abnorm	High	Yes	Yes	0.94	0.06
Yes	Abnorm	High	Yes	No	0.93	0.07
Yes	Abnorm	High	No	Yes	0.92	0.08
Yes	Abnorm	High	No	No	0.91	0.09
Yes	Abnorm	Norm	Yes	No	0.84	0.16
Yes	Abnorm	Norm	Yes	No	0.82	0.18
Yes	Abnorm	Norm	No	Yes	0.80	0.20
Yes	Abnorm	Norm	No	No	0.78	0.22
Yes	Norm	High	Yes	Yes	0.91	0.09
Yes	Norm	High	Yes	No	0.91	0.09
Yes	Norm	High	No	Yes	0.90	0.10
Yes	Norm	High	No	No	0.89	0.11
Yes	Norm	Norm	Yes	Yes	0.80	0.20
Yes	Norm	Norm	Yes	No	0.78	0.22
Yes	Norm	Norm	No	Yes	0.76	0.24
Yes	Norm	Norm	No	No	0.73	0.27
No	Abnorm	High	Yes	Yes	0.79	0.21
No	Abnorm	High	Yes	No	0.77	0.23
No	Abnorm	High	No	Yes	0.74	0.26
No	Abnorm	High	No	No	0.71	0.29
No	Abnorm	Norm	Yes	Yes	0.48	0.52
No	Abnorm	Norm	Yes	No	0.42	0.58
No	Abnorm	Norm	No	Yes	0.35	0.65
No	Abnorm	Norm	No	No	0.28	0.72
No	Norm	High	Yes	Yes	0.74	0.26
No	Norm	High	Yes	No	0.71	0.29
No	Norm	High	No	Yes	0.68	0.32
No	Norm	High	No	No	0.64	0.36
No	Norm	Norm	Yes	Yes	0.35	0.65
No	Norm	Norm	Yes	No	0.28	0.72
No	Norm	Norm	No	Yes	0.19	0.81
No	Norm	Norm	No	No	0.10	0.90

Table 2. Conditional probability of High Blood Pressure given Medicine Taken, Smoking and Alcohol, Moderate Exercise and Obesity.

Abbreviation Used: Medicine Taken (MED), Smoking and Alcohol (SMALC), Moderate Exercise (MEXC), Obesity (OBT), High Blood Pressure (HBP), Probability (P).

States Used: MED (Yes, No), SMALC (Yes, No), MEXS (Yes, No), OBT (Yes, No).

MED	SMALC	MEXS	OBT	P(HBP)	P(¬HBP)
No	Yes	No	Yes	0.92	0.08
No	Yes	No	No	0.87	0.13
No	Yes	Yes	Yes	0.85	0.15
No	Yes	Yes	No	0.79	0.21
No	No	No	Yes	0.76	0.24
No	No	No	No	0.70	0.30
No	No	Yes	Yes	0.58	0.42
No	No	Yes	No	0.52	0.48
Yes	Yes	No	Yes	0.58	0.42
Yes	Yes	No	No	0.52	0.48
Yes	Yes	Yes	Yes	0.40	0.60
Yes	Yes	Yes	No	0.34	0.66
Yes	No	No	Yes	0.31	0.69
Yes	No	No	No	0.25	0.75
Yes	No	Yes	Yes	0.13	0.87
Yes	No	Yes	No	0.07	0.93

Table 3. Conditional probability of Atherosclerosis given Serum LDL, Serum Triglyceride, Moderate Exercise and Choresterol HDL Ratio.

Abbreviation Used: Serum LDL (SLDL), Serum Triglyceride(STRIG), Moderate Exercise (MEXC), Cholesterol HDH Ratio (CHDLR), Atherosclerosis (ATS), Probability (P).

States Used: SLDL (High, Low), STRIG (High, Low), MEXS (Yes, No), CHDLR (High, Low).

SLDL	STRIG	MEXC	CHDLR	Pr (ATS)	Pr (¬ATS)
High	High	No	High	0.84	0.16
High	High	No	Low	0.83	0.17
High	High	Yes	High	0.83	0.17
High	High	Yes	Low	0.81	0.19
High	Low	No	High	0.61	0.39
High	Low	No	Low	0.57	0.43
High	Low	Yes	High	0.57	0.43
High	Low	Yes	Low	0.52	0.48
Low	High	No	High	0.74	0.26
Low	High	No	Low	0.71	0.29
Low	High	Yes	High	0.71	0.29
Low	High	Yes	Low	0.68	0.32
Low	Low	No	High	0.35	0.65
Low	Low	No	Low	0.28	0.72
Low	Low	Yes	High	0.28	0.72
Low	Low	Yes	Low	0.20	0.80

Table 4. Conditional probabilities of ECG, Angina Pectoris, Rapid Heartbeats and Miocardial Infraction as Children, given Heart Disease.

Abbreviation Used: Angina Pectoris (ANGP), Miocardial Infraction (MIOCAR), Rapid Heartbeats (RHB), Heart Disease (HD), Probability (P).

States Used: ECG **(**Norm, Abnorm), ANGP (Yes, No), MIOCAR (Yes, No), RHB (Yes, No).

Variable	State	P (HD)	P (¬HD)
ECG	Abnorm	0.95	0.05
	Norm	0.05	0.95
ANGP	Yes	0.85	0.15
	No	0.40	0.60
MIOCAR	Yes	0.90	0.10
	No	0.05	0.95
RHB	Yes	0.99	0.01
	No	0.30	0.70

Table 5. Conditional probabilities of Serum Selenium, Serum LDL, Cholesterol HDL Ratio, and Serum Triglyceride given Diet.

Abbreviation Used: Serum LDL (SLDL), Serum Triglyceride(STRIG), Cholesterol HDH Ratio (CHDLR), Serum Selenium (SSM), Probability (P).

States Used: SLDL (High, Low), STRIG (High, Low), SSM (Abnorm, Norm), CHDLR (High, Low), Diet (Good, Bad).

Variable	State	P(Diet) = Good	P(Diet) = Bad
SLDL	High	0.25	0.75
	Low	0.75	0.25
STRIG	High	0.30	0.70
	Low	0.80	0.20
CHDLR	High	0.25	0.75
	Low	0.75	0.25
SSM	Abnorm	0.30	0.70
	Norm	0.70	0.30

Table 6. Conditional probability of Obesity given Moderate Exercise and Diet.

Abbreviation Used: Obesity (OB), Moderate Exercise (MEXC), Diet (DIET) , Probability (P).

States Used: MEXC(No, Yes), DIET (Bad, Good).

MEXC	DIET	P(OB)	P(¬OB)
No	Bad	0.6	0.4
No	Good	0.1	0.9
Yes	Bad	0.1	0.9
Yes	Good	0.05	0.95

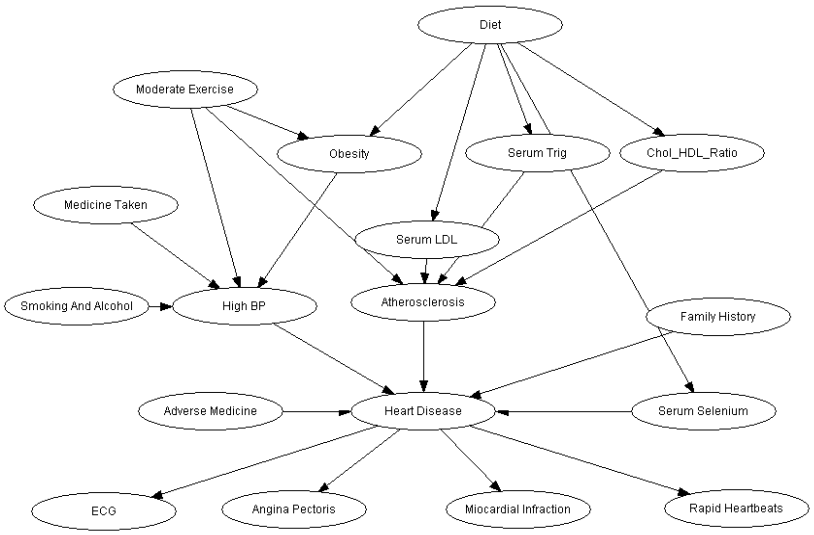

Figure 1: Probabilistic Model of heart disease

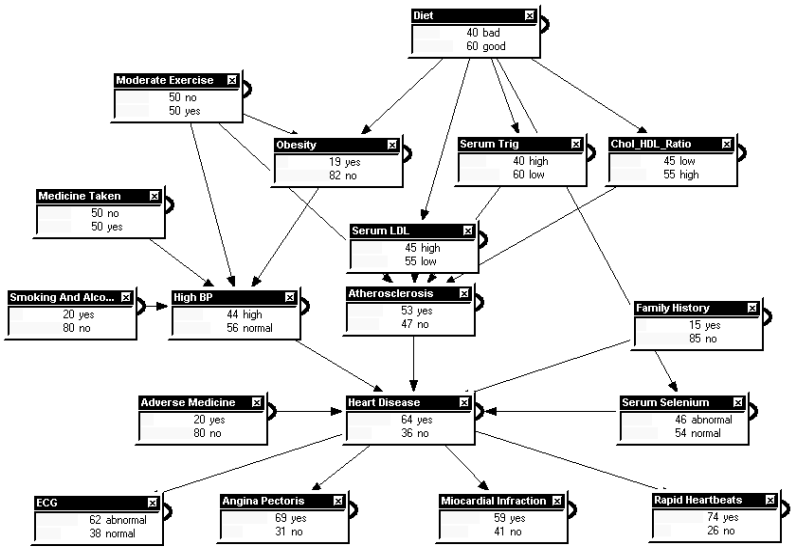

Figure 2: Probability of heart disease with no evidence

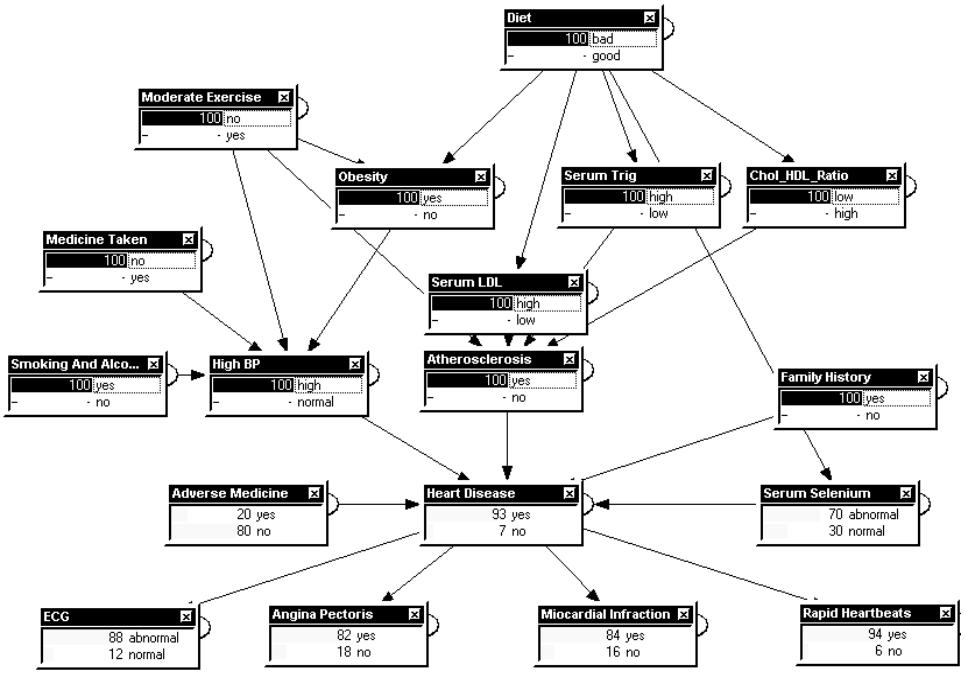

Figure 3: Probability of heart disease with evidence of adverse health

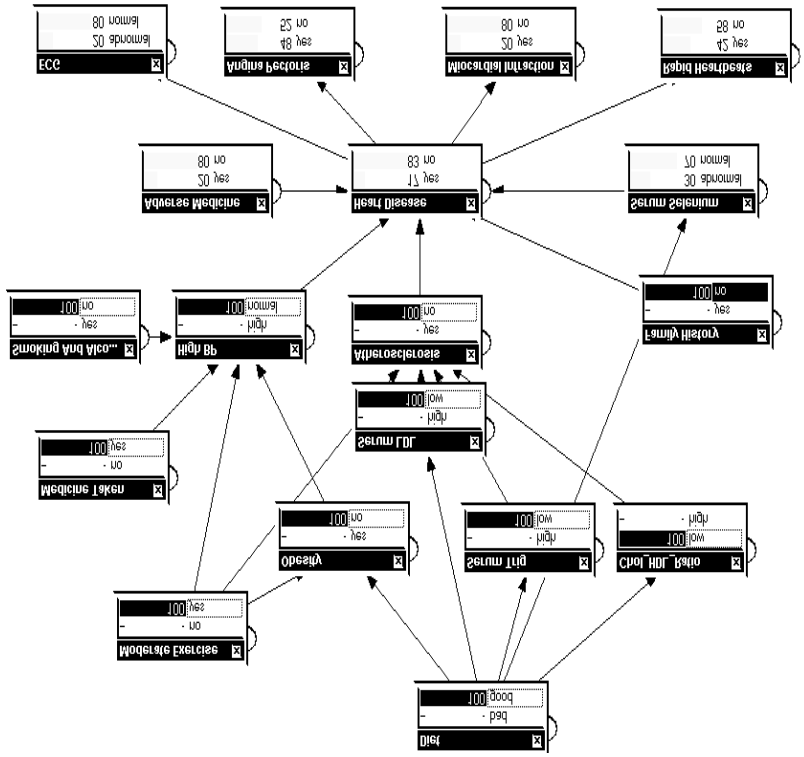

Figure 4: Probability of heart disease with evidence of good health

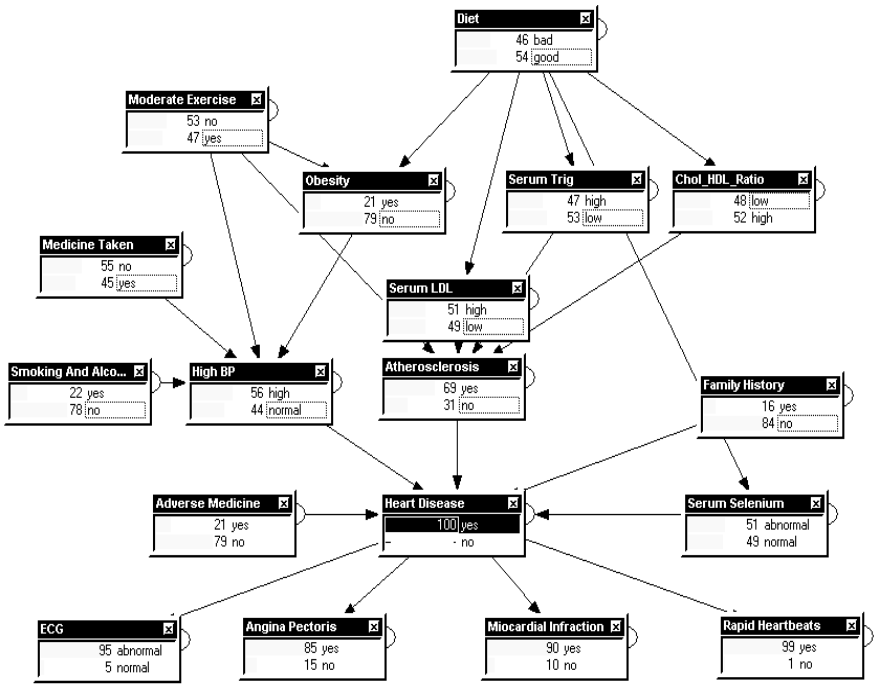

Figure 5: Marginal probability of each variable when heart disease is present

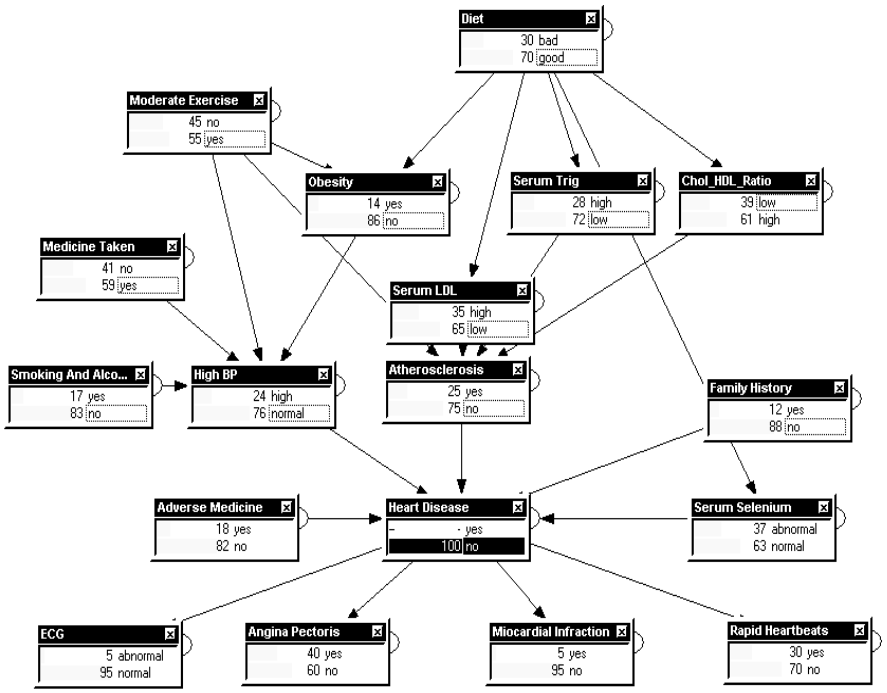

Figure 6: Marginal probability of each variable when heart disease is absent

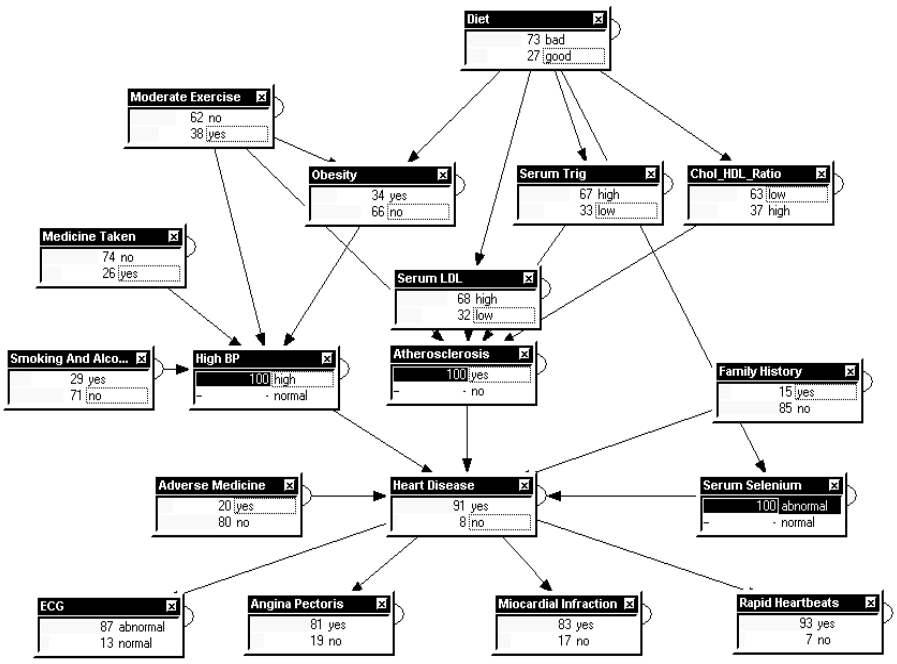

Figure7: Probability of heart disease when Atheroscelorosis, High BP and Serum Selenium are high

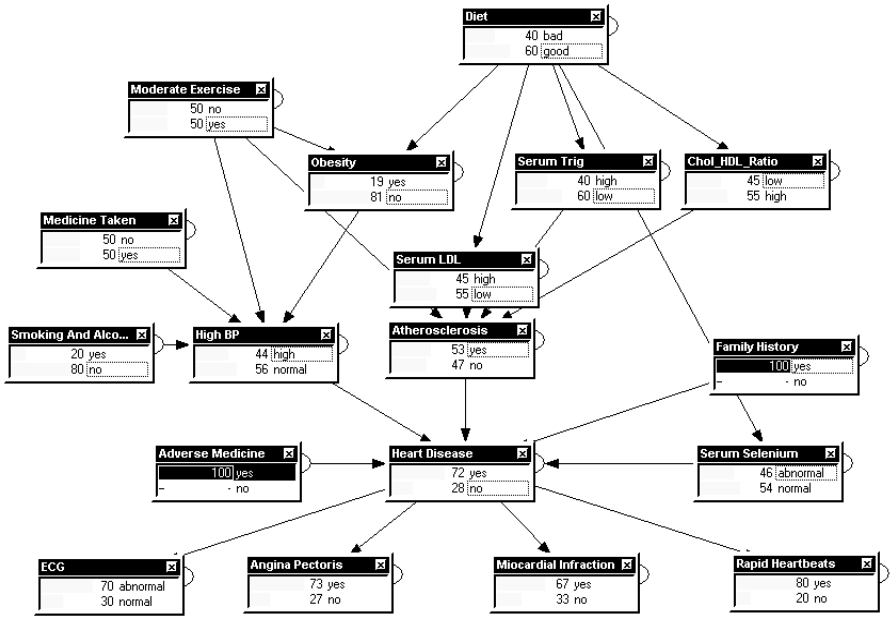

Figure 8: Probability of heart disease when Adverse Medicine and Family History are high

References

[1] Finn V. Jensen, *An Introduction to Bayesian Networks*, Springer-Verlag New York, Inc. 1997, ISBN 0-387-91502-8

[2] Andreassen, S., Jensen, F.V. and Olesen, K.G, "Medical expert systems based on causal probabilistic networks", *Int. J. Biomed. Comput.*, vol. 28, 1, (1991).

[3] Andreassen, S, "Planning of therapy and tests in causal probabilistic networks", *Art. Intelli, Med.*, vol. 4, pp.227, 1992.

[4] Li, Y. C., Haug, P. J., and Warner H. R, "Automated transformation of probabilistic knowledge for a medical diagnostic system", Ozbolt, J.P (Ed.), *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care* (Hanley and Belfus, Philadelphia), pp. 765, 1994.

[5] Shwe, M.A. et al, "Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base, I. The probabilistic model and inference algorithms", *Meth, Inform, Med*., vol. 30, 241 1991.

[6] Mathur, Kailash, Frishberg Barry and Underwood Brannon, "Cholesterol, Selected Minerals and Health Status of the Elderly in South Carolina. "
A Research Bulletin reported to United States Department of Agriculture,
South Carolina State University, Orangeburg, South Carolina, March' 1996.

[7] Wellman, Michael P., and Henrion, Max., "Explaining 'Explaining Away' ", *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 15, No. 3, March 1993.