Building a Bayesian Network Model of Heart Disease (Extended Abstract)

Jayanta K. Ghosh and Marco Valtorta

Department of Computer Science and Engineering
University of South Carolina
Columbia, South Carolina USA 29208
{ghosh,mgv}@cs.sc.edu

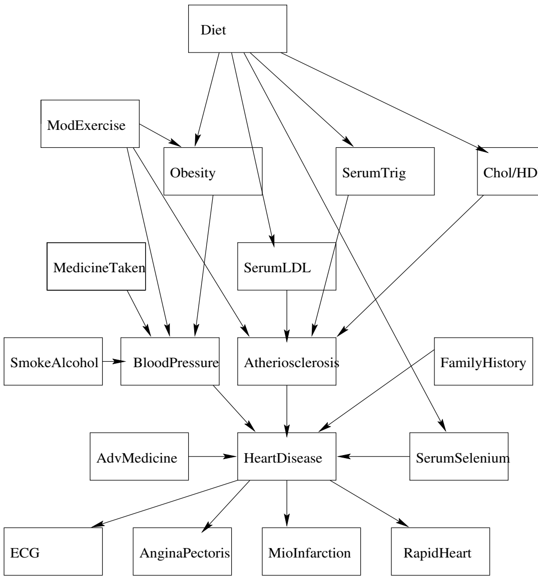

Figure 1: Bayesian Network Structure

Bayesian networks [2] represent a promising technique for clinical decision support and provide powerful capabilities for representing uncertain knowledge, including a flexible representation of probability distributions 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.

We report here on a Bayesian network model of heart disease constructed on the basis of information collected in an earlier study [3]. Figure 1 shows the Bayesian network structure of the model. Please refer to [1] for the numerical specification of the model and additional detail.

In order to simplify the calculation of conditional probabilities of the variables, we 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. For example, each new parent (direct cause, in a causal interpretation) 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 there are obvious difficulties involved in providing statistical data for each family of variables (i.e., a variable and it parents) in our Heart Disease model. 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 synergistic or antagonistic effect of variables on Heart Disease [4].

For the sake of clarity and understanding, we present below the assumptions and calculations involved in building the conditional probability table 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, $P(Atherosclerosis|Serum\ Triglycerides = low, Serum\ LDL = low, Cholesterol\ HDL\ Ratio = low, ModerateExercise = Yes, BackgroundCauses) = (1-0.8) = 0.20$. The conditional probabilities of the other fourteen cases can be calculated in an analogous way.

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 (using Hugin [2]) and came up with the probabilities for each of the variables. In the first case, 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 and third cases we introduced evidence of adverse and good health respectively, and the model calculated the probabilities of Heart Disease. In the 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 the 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.

Here are some comments on the remaining test cases. Atherosclerosis, High BP, and Serum Selenium are the most significant contributors to Heart Disease. In particular, when these three variables were observed to be adverse, the probability of Heart Disease increased from 64% to 91%. We noticed that among the children of Heart Disease, when ECG or Miocardial Infarction 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. Overall, we conclude that ECG is both a very sensitive and specific diagnostic tool for Heart Disease, while Miocardial Infarction is very specific, but not as sensitive as ECG.

Here are some conclusions and suggestions for future work: The model should be treated only as a prototype (and, of course, not be used in medical practice!); the variables of the model act independently, i.e. the model does not take into account any synergistic or antagonistic interactions of the variables that might influence Heart Disease [4]; the model does 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 Infarction 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; 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 ones used for its construction.

References

[1] J.K. Ghosh and M. Valtorta. "Probabilistic Model Building of Heart Disease." Tech Report TR9911, Department of Computer Science, University of South Carolina.

[2] F.V. Jensen. *An Introduction to Bayesian Networks*. New York: Springer, 1997.

[3] K. Mathur, B. Frishberg, and B. Underwood. "Cholesterol, Selected Minerals and Health Status of the Elderly in South Carolina," A Research Bulletin Reported to the U.S. Dept. of Agriculture, South Carolina State University, March 1996.

[4] M. Wellman and M. Henrion. "Explaining Explaining Away." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 3 (March 1993), 287–292.