International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 15 (Special Issue), 2016, pp 1-18

Bayesian Networks to Decision Making in Lymph Node Metastasis of Colorectal Cancer

Md. Aminul Hoque[1,2*], Toshifumi Wakai[3] and Kohei Akazawa[2]

[1]Department of Statistics, University of Rajshahi, Rajshahi-6205

[2]Medical Informatics, Niigata University of Medical & Dental Hospital,
Niigata, Japan

[3]Division of Digestive and General Surgery, Niigata University Graduate
School of Medical and Dental Sciences, Niigata, Japan

[*]Correspondence should be addressed to Md. Aminul Hoque
(Email: aminul@ru.ac.bd ; mdaminulh@gmail.com)

Abstract

[Received March 15, 2016; Accepted August 10, 2016]

The quality of medical care has always been a key issue for both practitioners and patients and the highest standards and practice guidelines are expected in all fields of medicine. The diagnosis of cancer metastasis is often difficult because of atypical clinical histories, clinical signs, and the results of laboratory tests. Recently, the Bayesian network (BN) model has been used in a variety of research fields for decision support. We wanted to validate the prognostic ability of BN analysis in comparison with Neural Network (NN) and logistic regression analysis to predict more accurately of lymph node metastasis. This study has been conducted on patients who suffered from colorectal cancer collected from Niigata University Hospital in Japan. A total of 778 patients with colorectal cancer were analyzed in this study; there were 460 (59.1%) men (from 32 to 90 years of age) and 318 women (from 35 to 93 years of age). The optimal structure of the BN model was determined based on R package deal and a network structure was selected based on the Bayesian score and expert knowledge. Through the course of illness, only 88 of 778 (11.3%) of the patients were diagnosed as having regional lymph node metastasis.The network structure showed a complex relationship among the graph nodes, and metastasis node is directly connected with five other nodes. The conditional probabilities of bump (23.9% to 53.9%), permeation of the Lymphatic vessels (PLV) (32.6% to 73.3%), degree of differentiation for SM layer (26.5% to 57.1%), and blood vessel invasion (BVI) (23.9% to 36.6%) have been remakable changed when found the patient with nodal disease. These results justified that these variables are directly influenced by regional lymph node metastasis. Predictive accuracy (82.45%) of BN alaysis is also higher than ANN (81.718%) and logistic regression analysis (76.43%). We constructed a BN model for the diagnosis of cancer metastasis and found BN model provided the best prognostic prediction of colorectal cancer in clinical practice.

Keywords: Bayesian Network, diagnostic prediction, cancer metastasis, network structure, conditional probability, Neural networks, Logistic regression model.

AMS Classification: 62C$_{xx}$.

1. Introduction

Colorectal cancer (CRC) is now a worldwide problem with an annual incidence of approximately 1 million cases and an annual mortality of more than 500,000 (Winawer, 2007, Parkin et al., 2002). Lymph node metastasis colorectal cancer is a common cancer in Japan. In 2005, approximately 115,000 new patients were diagnosed with colorectal cancer in Japan, making it one of the most common types of cancer in the country and during last fifty years the incidence of colorectal cancer has increased significantly. The absolute number of cases will increase over the next 2 decades as a result of aging and expansion of populations in both developed and developing countries. The risk for this cancer varies from country to country and even within countries. The risk also varies among individual people based on diet, lifestyle, and hereditary factors etc (Winawer et al., 2003, Morson, 1974, Voglestein et al., 1988). The known risk factors and important prognosis factors of the lymph node metastasis cancer are the size of the tumor, status mucosa (SM), degree of differentiation for SM layer, depth of differentiation for SM layer, bump of tumor, blood vessel invasion and permeation of the lympthatic vessels (PLV). Many studies have been done on colorectal cancer using data mining techniques. Artificial neural network (ANN), multiple regression analysis and logistic regression model were used for breast cancer and lunch cancer (Lundin et al., 1999; Burke et al., 1997, Choi, 2003; Choi et al., 2009). Very recently Choi et al., introduced a hybrid Bayesian Network model for predicting breast Cancer prognosis and Sakai et al., introduced Bayesian Network Analysis using appendix data. In various types of cancer there are have been proposed prognostic indices which are derived by multiple regression analysis of number of patients, diseases and treatment parameters, but the main problem with multiple regression analysis is that the independent variables considered simultaneously cannot be mutually related i.e. they should be orthogonal (Bucinski et al., 2007; Bayer et al., 1996; Jollife, 1986).Generally ANN shows better predictive index and good estimation power over other existing methods for breast cancer prognosis (Choi et al., 2009; Berner, 2007). But the main disadvantage of neural networks stems from difficulties in their representation of knowledge. Acquired knowledge in the form of nodes and actual links cannot be interpreted easily and system does not explain the results (Berner, 2007). Even error rate of ANN is higher than BN but area under the curve (AUC) of ANN is smaller than BN for appendix data (Sakai et al., 2007). Logistic regression model has also less predictive index than other predictive models (Delen et al., 2005; Choi et al., 2009 and Sakai et al., 2007).

However, Bayesian networks (BN) was seldom used for predicting colorectal cancer prognosis. It is a probabilistic model that consists of dependency structure and local probability. It uses prior probability in the prediction of dependent variables. Bayesian networks (BN) have been introduced in the 1980s as a formalism for representing and reasoning with models of problems involving uncertainty, adopting probability theory as a basic framework (Pearl, 1988). Since the beginning of the 1990s researchers are exploring its possibilities for developing medical application (Acid et al., 2003; Berzuini et al., 1992; Cheng et al., 2002; Cooper and Herakovits 1992). Research in Bayesian networks with clinical data is undergoing very interesting. The BN formalism offers a natural way to represent the uncertainties involved in medicine when dealing with diagnosis, treatment selection, planning, and prediction of prognosis (Lucas et al, 1998, Alvarez et al., 2006, Burnside et al., 2006, Kline et al., 2005). In this study Bayesian Network application is discussed on the decision making in clinical practice**.** We present a BN approach to predicting factors influence to the lymph node metastasis cancer clinical data. Consider a situation in which one feature on an entity has a direct effect on another feature of that entity. For example, the presence or absence of a disease in a human being has a direct effect on whether a test for that disease turns out positive or negative (Nealpolitan 2004, Jensen and Nielsen, 2007, Oliver et al., 2008). Bayes theorem has been used to perform probabilistic inference since last few decades to identify the factors of disease influence. Here in this study, we will use this theorem to compute the conditional probability of an individual having a disease when a test for the disease came back positive.

We would want to determine, for example, the conditional probabilities for cancer disease when it is known that an individual smokes or drinks and both. For simple example with few related variables it would be easy to compute the conditional probabilities but it would be very complex to determine such probabilities if there are many variables by the conventional computer programs. To meet these difficulties Bayesian networks may be very useful. By exploiting conditional independencies, we are able to represent a large instance in a Bayesian network using little space and we are able to perform probabilistic inference among the features in a shorter times. In addition, the graphical nature of Bayesian networks gives us a much better grasp of the interrelationship among the variables.

A Bayesian network is a way of representing the joint probability distribution of a set of random variables that exploits the conditional independence relationships among the variables, often greatly reducing the number of parameters needed to represent the full joint probability distribution (Parkin et al., 2002; Silvia et al., 2004; Voglestein et al., 1988; Cheng et al., 2002; Cooper and Herakovits 1992). Also, Bayes nets give a powerful and natural way to represent the dependencies that do exist. Bayesian networks allow for risk probabilities to be specified based on subjective assessments (“expert opinion”), empirical evidence, or a combination of both. Incorporating expert opinion is important because, often, there are insufficient data to learn and model relationships between risk factors and outcomes using population based data (“machine learning”). When there is a paucity of data, domain expert opinion can be used to create Bayesian networks. Expert-derived probabilities can be improved over time with observational data from multiple sources, obviating the need for a single data repository that contains all known risk factors. Risk predictions for an individual draw on: (i) relationships observed in populations studied; (ii) expert opinion; and (iii) the individual`s risk factors. BNs are gaining an increasing popularity as modeling tools for complex problems involving probabilistic reasoning under certainty (Neapolitan, 2004; Wainawer et al., 2006, Oliver et at, 2002; Rodin and Boerwinkle, 2005; Sakellaropoulos and Nikiforidis, 1999; Jensen and Nielsen, 2007; Wang et al., 1999).

Probabilistic graphical models are suitable for network analysis for many reasons. They provide a concise language for describing probability distributions over the observation and also show a complete structure of the interdependencies among the variables. In addition, the literature on graphical models provides approaches to learning from data that are derived from the basic well-understood principles of statistics (Friedman, 2004).Many research have been done with colorectal cancer but very few research has been done to actual predict with lymph node metastasis of colorectal cancer. However, we will try to shed light actual probability predictions using Bayesian network (BN) analysis in this study.

2. Materials and Methods

2.1 Data Sources

In this study we used clinical data for lymph node metastasis colorectal cancer obtained from different Japanese Hospitals and health centers.

2.2 Colorectal Cancer Data

This study has been conducted on patients who suffered from Cancer Metastasis collected from6 institutions in Japan. From The Fukuoka university Tsukushi Hospital 111, from the Osaka University 104, National cancer center 155, Medical Science University, Japan 128, Cancer Laboratory, Japan 95 and from Niigata University 215. The data from a total of 808 sample patients were taken initially. Some factors have many missing values and we excluded all the missing values from the data. Finally we analyzed 778 patients with full information in this study; A total of 778 patients with colorectal cancer were analyzed in this study; there were 460 (59.1%) men (from 32 to 90 years of age) and 318 women (from 35 to 93 years of age).

2.3 Variables Used in the Analyses for cancer data

The age and sex of patients, position, endoscope, size, organizational diagnosis (OD), status of mucosa (SM), degree of differentiation for the SM layer (DSM), depth of cancer spread in SM layer (DSML), extent of cancer spread in SM layer (ESML), infiltrative growth pattern (IGP), permeation of the lymphatic vessels (PLV), blood vessel invasion (BVI), Bump, LM and vertical margin (VM) were used to predict the factors of cancer metastasis. Some of these variables may cause of cancer metastasis and some are the outcomes of cancer metastasis. Variable LM and VM are identically similar and we excluded these variables to avoid singularity and difficulty. Variables OD, and IPG have also many missing values and excluded this variable from our analysis.

2.4 Methods of analyses

We mainly interested in Bayesian Networks (BN) analysis in this study. But to elucidate its superiority over other methods ANN, and logistic regression model we constructed Accuracy and AUC for these techniques too. R comprehensive package deal was used to construct Bayesian net structure. R packages ‘amore’ and ‘nnet’ were used ANN and ‘mlogit’ was used for and logistic regression. We also used SPSS software (SPSS Inc., Chicago, IL) and MATLAB Neural Network Toolbox, Statistics Toolbox (Math Works, Inc., Natrick, MA), infor our purpose (Sakai et al., 2007). NETICA (Version 3.19; Norsys Software Corp, Vancouver Canada) was used for Bayesian network construction and performance evaluation. The differences in the means of the AUC between models were analyzed by using Krusal-Walish tests and Wilcoxon’s rank sum tests. Bonferoni’s method was used to adjust for multiple comparisons between the constructed models.

3. Results

Performance of the predictive models was evaluated for several iteration sequence of the validation process (Table 1). The BN model achieved the highest prediction accuracy of 82.45% where as 81.78% for ANN and 76.43% for logistic regression model. For AUCs, the BN model achieved highest 0.882, whereas the corresponding values of ANN and logistic model were 0.869 and 0.794 respectively.

Table 1: Comparison results of performance evaluation for the models used

Models	Accuracy (SD)	AUC (SD)
Logistic	76.43% (0.026%)	0.794 (0.056)
BN	82.45% (0.031%)	0.882 (0.044)
ANN	81.78% (0.018%)	0.869 (0.047)

The difficult task of Bayesian network analysis is to establish the interrelationship between the existing variables without prior information or expert knowledge. For the large number of variables it will remain very complex to find exact relationship between the variables. Fortunately Bottcher and Dethlefsen (2003) have established a program into R that will give us guide line to find a way of relationships between the variables. Next but the main problem is to find the conditional probabilities of the related variables that direct or indirect relation with the target variable cancer. We met this problem by using Netica software.

Figure 1 shows the structure of Bayesian network derived from existing variables directly or indirectly related with cancer metastasis.

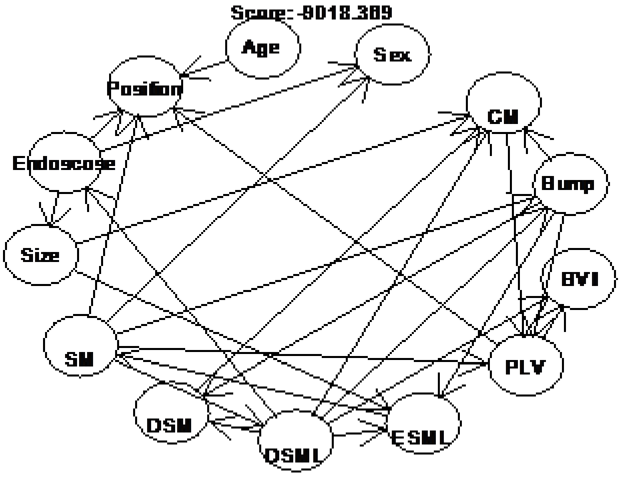


Figure1: Structure of Bayesian Netwrks for colorectal cancer data. The labelled circles represent nodes (variables) and the arcs reprent probabilistic dependencies. SM: Status of Mucosa, DSM: Degree of Differentiation for the SM Layer, DSML:Depth of Cancer spread in SM Layer, EMSL: Extent of Cancer spread in SM Layer, Size, PLV: Permeation of the Lymphatic Vessels, BVI: Blood Vessel Invasion.

Probability structure of whole systems for colorectal cancer was presented in Figures 2. From this figure we got the real picture of the cancer condition from the given clinical data. Through the course of illness, 11.3% of the patients were diagnosed as having a cancer metastsis. Predicted conditional probabilities given cancer metastasis were shown in Figure 3 and we found some intersting results for few variables,particularly, SM, DSM,PLB, BVI, and Bump.The structure of the network for the cancer metastasis showed very complex relationaship among the graph nodes. Cancer matastasis node is directly connected with seven of the other nodes. Predicted conditional probabilities for the variables given cancer metastasis were presented in Figure 4. From this figure we found that bump, PLV, DSM, Size, SM, and BVI have significant impact on metastasis. Other variables have moderate influence on cancer metastasis except the variables age and sex (Appendix Table 2).

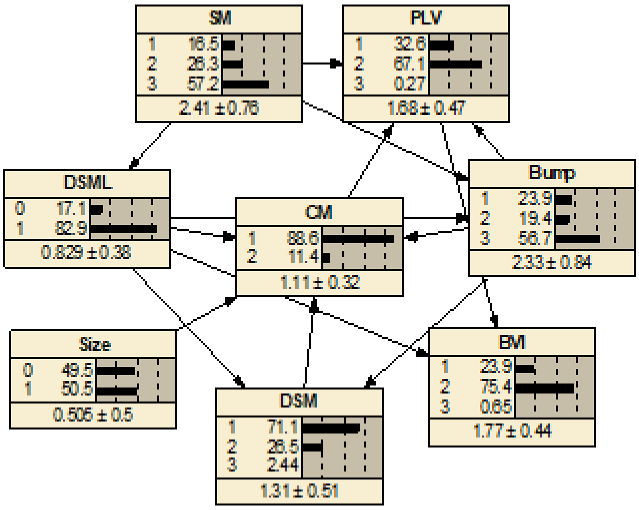


Figure 2: Probability structure of Bayesian Networks for colorectal cancer

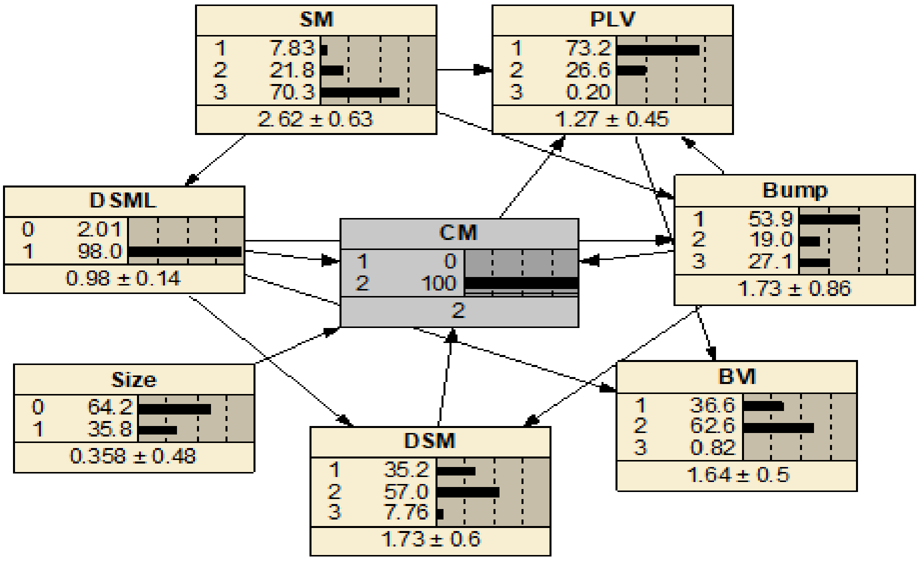


Figure 3: Conditional probability structure of Bayesian Networks for colorectal cancer data

It was also established that the error rate is the lowest in the Bayesian network model (error rate= 0.270)compared with artificial neural network (error rate=0.304), naïve Bayes model (error rate = 0.325), and logistic regrassion analysis (error rate =0.337) (Sakai et al., 2007; Linder et al., 2006).

4. Discussions

A common problem in BN approach is that the choice of the prior distribution is too subjective and this problem is related to the fact that, in some cases, the posterior distribution is very sensitive to the choice of prior. There are many ways to find approximations of the conditional probabilities in a Bayesian network. Which way is the best depends on the exact nature of the network. We constructed the Bayesian network structures for clinical triallymph node cancer metastasis based on the combination of expert knowledge and R programme. Although the graph structure of the constructed BN model did not necessarily represent the diagnostic process (Fig. 1), which showed the complex relationaships among the variables, the BN model was the best and most accurate for the diagnosis of thelymph node cancer metastasis Table 1). The possible reason for this result is for the learning processes. The BN model has two learning processes (such as structure and parameter learing), whereas the other models have only one learing process (Sakai et al., 2007; Anderson 2004; Cowell et al., 1999; Friedman, 2004).The conditional probabilities of bump (23.9% to 53.9%), permeation of the Lymphatic vessels (PLV) (32.6% to 73.3%), degree of differentiation for SM layer (26.5% to 57.1%), and blood vessel invasion (BVI) (23.9% to 36.6%) have been remakable changed when found the patient with nodal disease (Figures 2-3 and Table 3) for cancer metastasis to be happened which implied that these variables have direct effect on cancer disease of lymph node metastasis. These results justified that these variables are directly influenced by cancer metastasis. In this study, we combined two methods for improving the diagnostic accuracy of BN model. First we used R pacakge ‘deal’ then verified by expert to contruct the BN structures. Our proposed Bayesian predictive model is very reader friendly because of its simplicity. People, even very unknow of statistics and medicine can understand and predict his probabily to be cancer patient if he has clinical results in hand.

Our initial goal was to apply the BNs using clinical data for decision making and prediction and to attain that goal we successfully applied BN in this regards.

One of the limitations of this study is that we excluded four variables due to many missing values.To draw a conclusion, though few limitations, still BN is the most accurate model for the disgnosis of cancer metastasis over the other models. BN may be applicable to the all other clinical data with different diseases. Further studies are needed to support our study in different clinical practice using BN by Netica.

References

[1] Acid, S., de Campos, L. M., Fernandez-Luna, J. M., Huete, J. F. (2003). An information retrieval model based on simple Bayesian networks. International Journal of Intelligent Systems, 18, 251-265.

[2] Adam, B., Tomasz, B., Jerzy, K., Renata, S., Jerzy, Z. (2007). Clinical data analysis using artificial neural networks (ANN) and principal component analysis (PCA) of patients with breast cancer after mastectomy. Rep Pract Oncol Radiother, 12(1),9-17.

[3] Alvarez, S. M., Poelstra, B. A., Burd, R. S. (2006). Evaluation of a Bayesian decision network for diagnosing pyloric stenosis. J Pediatr Surg, 41 (1), 155-161.

[4] Andersson, R. E. (2004). Meta-analysis of the clinical and laboratory diagnosis of appendicitis. Br J Surg, 91, 28–37.

[5] Berzuini, C., Bellazzi, R., Quaglini, S., Spiegelhalter, D. J. (1992). Bayesian Networks for patient monitoring. Artificial Intelligence in Medicine, 4, 243-260.

[6] Bottcher, S. G., and Dethlefsen, C. (2003). Learing Bayesian Networks with R. Proceedings of the 3rd International Workshop on Distributed Statistical Computing, March 20-22, Vienna, Austria.

[7] Beyer, J., Kramar, A., Mandanas, R. et al. (1996). High-dose chemotherapy as salvage treatment in germ cell tumours: a multivariate analysis of prognostic variables. J Clin Oncol, 14, 2638–45.

[8] Burnside, E. S., Rubin, D. L., Fine, J. P., Shachter, R. D., Sisney, G. A., Leung, W. K. (2006). Bayesian network to predict breast cancer risk of mammographic micro calcifications and reduce number of benign biopsy results: initial experience. Radiology, 240 (3), 666-673.

[9] Cheng, J., Greiner, R., Kelly, J., Bell, D., Liu, W. (2002). Learning Bayesian networks from data: an information-theory based approach. Artificial Intelligence, 137, 43-90.

[10] Choi, J. P., Ham, T. H., Park, R. W. (2009). A Hybride Bayesian Network Model for Predicting Breast Cancer Prognosis. J Kor Soc Med Informatics, 15(1), 49-57.

[11] Cooper, G. F., and Herakovits, E. (19920. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9(4), 309-348.

[12] Cowell, R. G., Dawid, A. P., Lauritzen, S. L., Spiegelhalter, D. J. (1999). Probabilistic Networks and Expert Systems. Springer, New York.

[13] Friedman, N. (2004). Inferring Cellular Networks Using Probabilistic graphical Models. Science, 303,799-805.

[14] Herskovits, E. H., Cooper, G. H. (1991). Algorithms forBayesian belief-network precomputation. MethodsInf Med, 30 (2), 81-89.

[15] Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., De Moor, B. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics, 22 (14), e184-190.

[16] Imperiale, T. F., Wagner, D. R., Lin, C. Y. Y. et al. (2000). Risk of advanced proximal neoplasm in asymptomatic adults according to the distal colorectal findings. N Engl J Med, 343, 169–174.

[17] Jensen, F. V., and Nielsen, T. D. (2007). Bayesian Networks and Decision Graphs. 2nd ed. Springer.

[18] Kharbanda, A. B., Taylor, G. A., Fishman, S. J., Bachur, R. G. (2005). A clinical decision rule to identify children at low risk for appendicitis. Paediatrics, 116, 709-716.

[19] Jollife, I. T. (1986). Principal Component Analysis, Springer, New York.

[20] Kline, J. A., Novobilski, A. J., Kabrhel, C., Richman, P. B., Courtney, D. M. (2005). Derivation and validation of a Bayesian network to predict pretest

probability of venous thromboembolism. Ann Emerg Med, 45 (3), 282-290.

[21] Lieberman, D. A., and Weiss, D. G. (2001). Veterans Affairs Cooperative Study Group 380.One-time screening for colorectal cancer with combined fecal occult-blood testing and examination of the distal colon. N Eng J Med, 345, 555–560.

[22] Linder, R., Konig, I. R., Weimar, C., Diener, H. C., Poppl, S. J., Ziegler, A. (2006). Two Models for Outcome Prediction: A comparison of Logistic Regression and Neural Networks. Methods Inf Med, 45, 536-40.

[23] Lucas, P. J. F., Boot, H., Taal, B. G. (1998). Decision-theoretic network approach to treatment management and prognosis. Knowledge-based systems, 11, 321-330.

[24] Maestu, I., Pastor, M., Gomez-Codina, J. et al. (1997). Pretreatment prognostic factors for survival insmall-cell lung cancer: a new prognostic index and validation of three known prognostic indices on 34 patients. Ann Oncol, 547–553.

[25] Morson, B. (1974). The polyp-cancer sequence in the large bowel. Proc R Soc Med, 67(6), 451–457.

[26] Prentice Hall, N. J., Oliver, P., Patrick, N., Bruce, M. (2008).. Bayesian Networks. A Practical Guide to Applications. Wiley.

[27] Peter, L. (2004). Bayesian Analysis, Pattern Analysis and Data Mining in Health Care. Current Opinion in Critical Care, 10, 399-403.

[28] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems. Morgan Kaufman, San Mateo, California..

[29] Parkin, D. M., Whelan, S. L., Ferlay, J. et al. (2002). Cancer incidence in five continents. IARC Scientific Publications, 8(155), Lyon, IARC.

[30] Rodin, A. S., and Boerwinkle, E. (2006). Mining genetic epidemiology data with Bayesian networks,

[31] Andrei, S. Rodin, Eric, B., (2005). Bayesian networks and example application (plasma apoE levels). Bioinformatics, 21(15): 3273-3278.

[32] Sakai, S., Kobayashi, K., Nakamura, J., Toyabe, S., Akazawa, K. (2007). Accuracy in the Diagnostic Prediction of Acute Appendicitis Based on the Bayesian Network Model. Methods Inf Med, 6, 723-726.

[33] Sakellaropoulos, G. C., Nikiforidis, G. C. (1999). Development of a Bayesian network for the prognosis of head injuries using graphical model selection techniques. Methods Inf Med, 38 (1), 37-42.

[34] Silvia, A., de campos, L. M. et al. (2004). A comparison of learning algorithoms for Bayesian networks: a case study based on data from an emergency medical service. Artificial Intelligence in Medicine, 30, 215-232.

[35] Voglestein, B., Fearon, E. R., Hamilton, S. R. et al. (1988). Genetic alterations during colorectal-tumor development. N Eng J Med, 391, 525–532.

[36] Wang, X. H., Zheng, B., Good, W. F., King, J. L., Chang. Y. H. (1999). Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. Int JMed Inform, 54 (2), 115-126.

[37] Winawer, S. J. (2007). Colorectal cancer screening. Best Practice and Research Clinical Gastroenterology 21(6), 1031-1048.

[38] Winawer, S. J., Zauber, A. G., Fletcher, R. H. et al (2006). Guidelines for colonoscopy surveillance after polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer the American CancerSociety. Gastroenterology, 130(6), 1872–1885.

[39] Winawer, S., Fletcher, R., Rex D. et al. (2003). Colorectal cancer screening and surveillance: clinical guidelines and rationale-update based on new evidence. Gastroenterology, 124, 544–560.

Appendix

A Simple Example of Bayesian Network

A Bayesian Network is a model that reflects the states of some real part of a world that is being modeled and it describes how those states are related by probabilities. All the possible states of the modeled represent all the possible worlds that can exist, that is, all the possible ways that the parts or states can be configured. Here is a simple example of artificial cancer data (Table 1) that represent a simple but good Bayes net. In this example there are three variables Smoke, Drink and Cancer. It is also assumed that Smoke and Drink are independent each other but may have direct cause of cancer, that is cancer is dependent on Smoke and Drink. All three variables have two states (Yes=1 and No=0) and are connected as Bayes nets (Figure 1A). When actual probabilities are settled into this net that reflect the reality of real cancer condition. Such a net can be made to answer a number of useful questions, like, “if the patient is with cancer disease, what are the chances it was caused by Smoke or by Drink (Here Drink means alcohol drink), and, if the chance of drink increase, how does that affect the patient to be cancer patient. The important thing to note about this example is that the causal connections are not absolute.

Table 2: Artificial Cancer data

Case No.	Smoke	Drink	Cancer
1	0	0	0
2	0	0	1
3	0	0	0
4	0	1	1
5	0	1	1
6	1	0	0
7	1	0	1
8	1	1	0
9	1	1	1
10	1	1	1

Smoke (S)	Probability
Yes	0.5
No	0.5

Drink (D)	Probability
Yes	0.5
No	0.5

Smoke	Drink	Cancer (C)	
		Yes	No
Yes	Yes	0.67	0.33
Yes	No	0.5	0.5
No	Yes	1.0	0
No	No	0.33	0.67

The joint probability density function (pdf) is given by,

$P(C, S, D) = P(C/\ S, D)P(S)P(D)$ (1)

By using Bayes theorem on conditional probability and summing over all variables

$P(S=$ Yes/ C=Yes) = $\frac{P(S=Yes, C=Yes)}{P(C=Yes)}$ (2)

Where P(S=Yes, C=Yes) = $\sum_{D\in\{Yes,No\}} P(S = Yes, D, C = Yes)$ (3)

And $P(C = Yes) = \sum_{S,D\{Yes,No\}} P(C = Yes, S, D)$ (4)

Now from (3)

$\sum_{D\in\{Yes,No\}} P(S = Yes, D, C = Yes)$ = P(S=Yes, D=Yes, C=Yes) + P(S=Yes, D=No, C=Yes)

= P(S=Yes) P(D=Yes)P(C=Yes/S=Yes, D=Yes) +
P(S=Yes)P(D=No)P(A=Yes/S=Yes, D=Yes)

=0.5x0.5x0.67 + 0.5x0.5x0.5 = 0.1675 + 0.125 = 0.2925 (5)

Again from (4)

$\sum_{S,D\epsilon\{Yes,No\}} \sum_{\_}) = [P(C = Yes, S, D)]$ = P(C=Yes/ S=Yes, D=Yes)P(S=Yes)P(D=Yes) + P(C=Yes/S=Yes, D=No)P(S=Yes)P(D=No) + P(C=Yes/S=No, D=Yes)P(S=No)P(D=Yes) + P(C=Yes/S=No, D=No)P(S=No)P(D=No)

= 0.67x0.5x0.5+0.5x0.5x0.5+1.0x0.5x0.5+0.33x0.5x0.5

= 0.1675+0.125+0.25+0.0825=0.625 (6)

Now combining (2) to (6) and we have

$P(S =$ Yes/ C=Yes) = $0.2925 \div 0.626 = 0.467$ (7)

Which means about 46.8% effect of smoking to be caused of cancer.

And

P(D=Yes/C=Yes) = $\frac{\sum_{S\epsilon\{Yes,No\}} P(C=Yes,S,D=Yes)}{\sum_{S,D\in\{Yes,No\}} P(C=Yes)}$ (8)

From equation (6) we already got the value of denominator

$P(C = Yes) =$ 0.625, now to obtain numerator

$\sum_{S\epsilon\{Yes,No\}} P(C = Yes, S, D = Yes) =$ P(S=Yes, D=Yes, C=Yes) + P(S=No, D=Yes, C=Yes)

= P(S=Yes)P(D=Yes)P(C=Yes/S=Yes, D=Yes) +
P(S=No)P(D=Yes)P(A=Yes/S=No, D=Yes)

=0.5x0.5x0.67+0.5x0.5x1.0=0.1675+0.25 = 0.4175 (9)

Combining (6), (8) and (9) we have,

P(D=Yes/C=Yes) = 0.667 (10)

Which implies that 66.7% positive effect of drinking to be cancer. Our results are also supported by the Bayes net (Figure 1B) obtained by Netica.

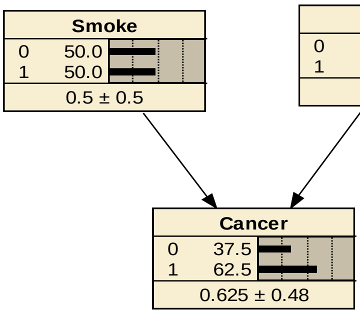


Figure 4A: Simple Bayesian Network

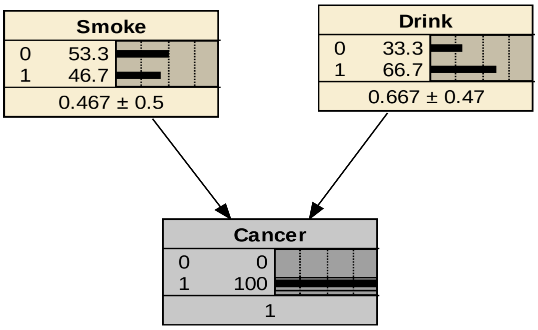
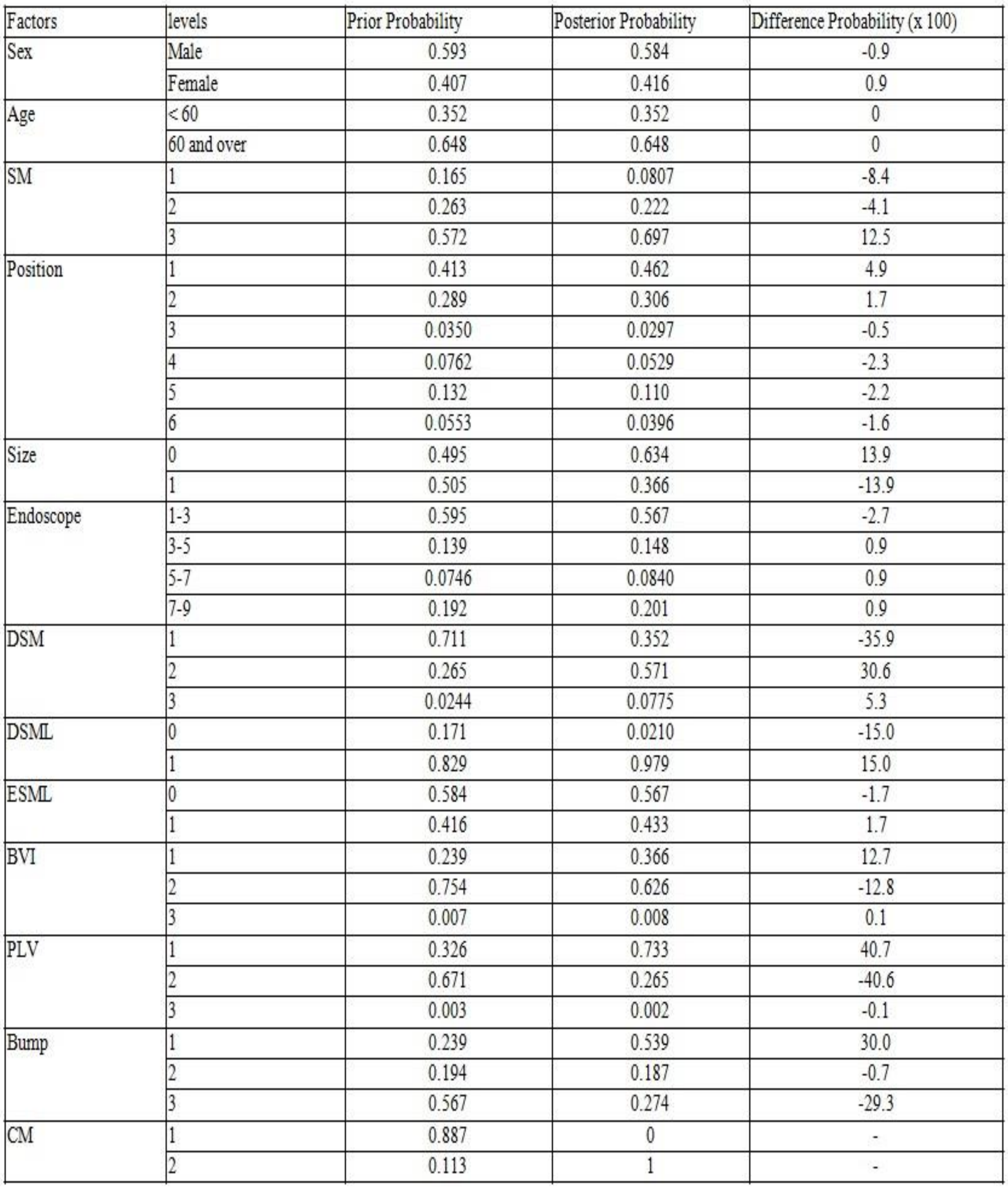


Figure 4B: Bayesian Network analysis

Table 3: Change of nodes value from prior to conditional posterior probabilities from BN analysis with lymph node cancer metastasis.



Factors	levels	Prior Probability	Posterior Probability	Difference Probability (x 100)
Sex	Male	0.593	0.584	-0.9
	Female	0.407	0.416	0.9
Age	< 60	0.352	0.352	0
	60 and over	0.648	0.648	0
SM	1	0.165	0.0807	-8.4
	2	0.263	0.222	-4.1
	3	0.572	0.697	12.5
Position	1	0.413	0.462	4.9
	2	0.289	0.306	1.7
	3	0.0350	0.0297	-0.5
	4	0.0762	0.0529	-2.3
	5	0.132	0.110	-2.2
	6	0.0553	0.0396	-1.6
Size	0	0.495	0.634	13.9
	1	0.505	0.366	-13.9
Endoscope	1-3	0.595	0.567	-2.7
	3-5	0.139	0.148	0.9
	5-7	0.0746	0.0840	0.9
	7-9	0.192	0.201	0.9
DSM	1	0.711	0.352	-35.9
	2	0.265	0.571	30.6
	3	0.0244	0.0775	5.3
DSML	0	0.171	0.0210	-15.0
	1	0.829	0.979	15.0
ESML	0	0.584	0.567	-1.7
	1	0.416	0.433	1.7
BVI	1	0.239	0.366	12.7
	2	0.754	0.626	-12.8
	3	0.007	0.008	0.1
PLV	1	0.326	0.733	40.7
	2	0.671	0.265	-40.6
	3	0.003	0.002	-0.1
Bump	1	0.239	0.539	30.0
	2	0.194	0.187	-0.7
	3	0.567	0.274	-29.3
CM	1	0.887	0	-
	2	0.113	1	-