RESEARCH COMMUNICATION

Use of an Artificial Neural Network to Determine Prognostic Factors in Colorectal Cancer Patients

Mahmood Reza Gohari[1], Akbar Biglarian[2]*, Enayatollah Bakhshi[2], Mohammad Amin Pourhoseingholi[3]

Abstract

Background & Objectives: The aim of this study was to determine the prognostic factors of Iranian colorectal cancer (CRC) patients and their importance using an artificial neural network (ANN) model. **Methods**: This study was a historical cohort study and the data gathered from 1,219 registered CRC patients between January 2002 and October 2007 at the Research Center for Gastroenterology and Liver Disease of Shahid Beheshti University of Medical Sciences, Tehran, Iran. For determining the risk factors and survival prediction of patients, neural network (NN) and Cox regression models were used, utilizing R 2.12.0 software. **Results**: One, three and five-year estimated survival probability in colon patients were 0.92, 0.71, and 0.48 and for rectum patients were 0.86, 0.71, and 0.42, respectively. By the ANN model, pathologic distant metastasis, pathologic regional lymph nodes, tumor grade, high risk behavior, pathologic primary tumor, familial history and tumor size variables were determined as ordered important factors for colon cancer. Tumor grade, pathologic stage, age at diagnosis, tumor size, high risk behavior, pathologic distant metastasis and first treatment variables were ordered important factors for rectum cancer. The ANN model lead to more accurate predictions compared to the Cox model (true prediction of 89.0% vs. 78.6% for colon and 82.7% vs. 70.7% for rectum cancer patients). **Conclusion**: This study showed that ANN model is a more powerful tool in survival prediction and influential factors of the CRC patients compared to the Cox regression model. Therefore, this model is recommended for predicting and determining of risk factors of these patients.

Keywords: colorectal cancer - prediction - survival analysis - Cox regression - artificial neural network

Introduction

Cancer is a group of more than 100 different diseases that affect the cells of body. Cancer of the colon and rectum (also referred to as colorectal cancer) varies around the world and is one of the most important causes of death in Western world (Ju et al., 2007; American Cancer Society, 2008; Toyoda et al., 2009). It is common in the Western world and is rare in Asia and Africa (MedicineNet, 2011). It's estimated that worldwide in 2008, 1.23 million new cases of CRC cancer were diagnosed and death of it was more than 600,000 people of the world. More than half of all deaths from the disease occur in the more developed regions of the world (WHO, 2008). New cases of CRC in the United States in 2010 was estimated 102,900 (colon); 39,670 (rectal) and deaths of it was estimated 51,370 (National Cancer Institute, 2010). Based on many different studies, the researchers found: a) patients treated within four weeks of surgery had a better chance of survival compared with people who started chemotherapy more than four weeks post-surgery. b) Monthly treatment delay

increased risk of disease recurrence after surgery and also increased risk of death after. c) Even supposing risk of recurrence went up with every extra one month after operation that a patient didn't start chemotherapy; there was still some benefit to receiving chemotherapy even after three months post-surgery (About.com, 2011). In Iran, the incidence rate of CRC is lower than Western countries (Sadjadi et al., 2005; Ministry of Health and Medical Education, 2006) but incidence rate of disease in younger people is higher than expected (Pahlavan and Jensen, 2005; Ansari et al., 2006, Foroutan et al., 2008) for that reason CRC is a major public health problem in Iran (Safaee et al., 2008).

Many different studies were conducted to determine risk factors of CRC and their importance (Giovannucci, 2002; Bingham et al., 2003; Corrao et al., 2004; Chan et al., 2008; American Cancer Society, 2008; Kelsall et al., 2009). The survival rate of CRC patients was related to tumor grade, tumor size, type of first treatment, body mass index (Boyle and Langman, 2000; Moghimi-Dehkordi et al., 2008; Asghari et al., 2009). The analyzing of CRC data

[1]*Department of Biostatistics & Hospital Management Research Center, Tehran University of Medical Sciences, [2]Department of Biostatistics, University of Social Welfare and Rehabilitation Sciences, [3]Research Center of Gastroenterology and Liver Diseases, Shahid Beheshti Medical University, Tehran, Iran *For correspondence: abiglarian@gmail.com*

and prediction of CRC was made on statistical methods such as survival analysis. In the previous decades, the data analysts have used a various survival methods for determining of prognostic factors and survival rate of patients. One alternate method to prediction of CRC survival is artificial neural network (ANN). ANN models are flexible and nonlinear methods that allow better fit to the data and leads to accurate prediction (Bishop, 1997). Use of ANN method for CRC was reported in few studies (Bottaci et al., 1997; Annand et al., 1999; Grumett et al., 2003; Lee et al., 2004; Kyung-Joong et al., 2004; Bittern et al., 2005; Ahmed wt al., 2005, Alladi et al., 2008). In this study, the ANN model was used to determine the risk factors of CRC patients and then the accuracy prediction of this model was compared to Cox regression model.

Materials and Methods

In this study, we analyzed the data from 1219 patients with CRC were been collected by cancer registry of the Research Center for Gastroenterology and Liver Disease of Shahid Beheshti University of Medical Sciences, Tehran, Iran (Asghari et al., 2009). At first we dropped out those patients with lower than one-month or higher than six years survival time. Based on this exclusion criterion a total of 1007 patients (786 colon cancer and 204 rectum cancer subjects) were entered in the study. The event (non-censored patients status) was defined as death from CRC occurring within follow-up time beginning from the date of first diagnosis. Other survived patients were considered as censored cases. In this historical cohort study, the required information for each patient including age at diagnosis, sex, marital status, familial history, BMI, high risk behavior (i.e. tobacco smoking, or alcohol history, or opium, or IV drug user, or betel use), weight loss, pathologic tumor stage, tumor grade, tumor size, pathologic primary tumor, pathologic distant metastasis, pathologic regional lymph node, and first treatment were used in the analysis. We also registered the survival time of each patient (in month). Pathologic stage of tumor was defined as I, II, III, and IV according to American Joint Committee on Cancer (AJCC) on TNM staging criterion (1988). In addition, Tumor grade was considered on a three-grade scale (well differentiated, moderately differentiated, and poorly differentiated).

Estimated survival time was reported monthly and was represented as mean (±standard deviation). The survival probabilities were compared in groups by the generalized Wilcoxon test procedure (Kleinbaum and Klein, 2005). The p-values less than 0.05 were considered as significant (it is mentioned that Back-ward stepwise method with significant level entry 0.10 and removal to the model 0.15 was used to model building). The assumptions of proportionality have been tested by Shoenfield residuals ph-test (Kleinbaum and Klein, 2005).

In the ANN modeling process, we randomly divided the data into two subsets: nearly 70% of patients for training the models and the remaining (nearly 30%) for testing and validation the model. After evaluating the model, we applied multiple layer perceptron (MLP) networks to determine important risk factors. In this context, we use independent variable importance analysis by using normalized importance, the risk factors were determined. In addition, the areas under receiver operation characteristic (AUROC) curves were used for comparing the prediction power of the described models. Note that, in fitting ANN model we used a three-layer MLP network with 14 variables in input layer, 5 to 20 nodes in middle layer and one node in output layer. Because of patient's status was bi-state (dead or censored), the logistic transfer function was utilized as the transfer function in middle and output layers. In addition, we have utilized back-propagation learning algorithm with learning rate of 0.01 to 0.40 and momentum of 0.80 to 0.95 for learning net. Concordance index was calculated for two models (ANN and Cox model) as a measure of ability of model predictions. Data were analyzed using R 2.12.0 software.

Results

Of the 1007 CRC patients 216 (21.4%) died (44 due to rectum cancer) and 791 (78.6%) were censored (618 cases in colon patients); 628 patients (62.4%) were men (600 patients with colon and others with rectum cancer) and others were women. The mean±SD of survival times for colon and rectum cancer were 51.4±37.3 and 52.1±35.4 respectively. The mean age at diagnosis in colon and rectum cancer were 53.7 (±14.6) and 52.4(±13.7) respectively. One, three and five-years estimated survival probability in colon patients were 0.92, 0.71, and 0.48 and for rectum patients were 0.86, 0.71, and 0.42 respectively. For analysis of data using ANN method, in the first step, the data was divided in training (nearly 70% of patients) and testing (nearly 30% of patients) groups. The generalized Wilcoxon test showed that the estimated survival curves using the training and testing groups have no significant difference (P=0.268). Afterward, MLP (three-layer) model was fitted to data based on 5 to 20 nodes in middle layer, and with 0.8 to 0.95 momentum and learning rate 0.01 to 0.40. The best model had 12 hidden node, learning rate of 0.15, and momentum of 0.90. For this model, sum of squares error prediction of colon and rectum patients' survival were 11.38 and 44.46 respectively. In addition, average relative error prediction of colon and rectum patients' survival were 0.94 and 1.01 respectively.

In the next step, based on validation set, the NN model was used to determine the risk factors. Based on importance analysis in ANN strategy, pathologic distant metastasis, pathologic regional lymph nodes, tumor grade, high risk behavior, pathologic primary tumor, familial history and tumor size variables were determined as ordered important variables for colon cancer. Tumor grade, pathologic stage, age at diagnosis, tumor size, high risk behavior, pathologic distant metastasis and first treatment variables were determined as ordered important factors for rectum cancer (Table1)

For comparing the accuracy of the models' prediction, we used accuracy classification table for the testing subset. The results were presented in Table 2. The ANN model lead to more accurate predictions compared to the Cox model (true prediction of 89.0% vs. 78.6% for colon and 82.7% vs. 70.7% for rectum cancer patients). The areas

Table 1. Cox and NN Modeling Results for Determining the Effects of Prognostic Factors on CRC Patients' Survival

	ANN model				Cox regression			
	Colon		Rectum		Colon		Rectum	
Ordered variables	Normalized importance*	Ordered variables	Normalized importance*	Ordered variables	P_value	Ordered variables	P_value	
Distant metastasis	100.0	Tumor grade	100.0	High risk behavior	0.042	Age at diagnosis	0.022	
Regional lymph nodes	94.8	Pathologic stage	90.5	First treatment	0.051	Tumor grade	0.031	
Tumor grade	83.8	Age at diagnosis	87.7	Age at diagnosis	0.078	Pathologic stage	0.049	
High risk behavior	77.8	Tumor size	62.6	Tumor grade	0.095	High risk behavior	0.063	
Pathologic primary tumor	65.6	High risk behavior	58.0	Pathologic stage	0.141	BMI	0.078	
Age at diagnosis	65.6	Distant metastasis	58.4	Primary tumor	0.272	First treatment	0.121	
Familial history	63.7	First treatment	52.8	BMI	0.324	Marital status	0.215	
Tumor size	56.8	Sex	47.8	Regional nodes	0.451	Familial history	0.328	
Pathologic stage	47.0	Primary tumor	37.5	Tumor size	0.459	Primary tumor	0.435	
First treatment	40.2	Familial history	22.2	Distant metastasis	0.589	Tumor size	0.455	
Marital status	35.9	Regional nodes	20.9	Familial history	0.628	Distant metastasis	0.651	
BMI	23.4	BMI	25.3	Sex	0.715	Regional nodes	0.665	
Sex	23.4	Marital status	18.9	Marital status	0.755	Sex	0.704	

* gain value

Table 2. Classification Accuracy of ANN and Cox Models in Testing Subsets of CRC patients

Cancer	Status	Observed	True prediction	
			ANN	Cox
Colon	Dead	38	24 (63.2)	17 (44.7)
	Survived	172	163 (94.8)	148 (86.0)
	Total	210	187 (89.0)	165 (78.6)
Rectum	Dead	12	8 (66.7)	5 (41.7)
	Survived	46	40 (86.9)	36 (78.3)
	Total	58	48 (82.7)	41 (70.7)

under ROC curves, calculated from testing colon and rectum data subset, for ANN model were 0.73, and 0.68 and for Cox model, 0.62 and 0.59.

Discussion

Because of increasing rate of CRC in Iran especially in youth through recent decades (Hosseini et al., 2004, Pahlavan and Jensen, 2005), this may be a major public health problem. Therefore prediction of CRC survival probability and determining of risk factors of CRC patients is important. Published studies have reported marital status, race and education site-specifically (Charles and Thomas, 1992; Tavani et al., 1999; Troisi et al., 1999; Cheng et al., 2001; Wu et al., 2004; Li et al., 2007), BMI (LeMarchand et al., 1992; Asghari-Jafarabadi, 2009), high risk behavior (Cho et al., 2004; Erhardt et al., 2002; Giovannucci et al., 1995;, Mizoue et al., 2006; Asghari-Jafarabadi, 2009), Tumor size (Li et al., 2007), Tumor grade (Li et al., 2007) are as risk factors. Another study with large sample data (72,214 patients) showed that marital status, tumor size, stage, gender, race, site, grade, age, and income variables have had significant relation to colon cancer survival (Lai and Stotler, 2010). All of these researcher have used statistical survival analysis methods and few studies have been used other techniques such as ANN methods (Bottaci et al., 1997; Annand et al., 1999; Grumett et al., 2003; Lee et al., 2004; Kyung-Joong et al., 2004; Bittern et al., 2005; Ahmed wt al., 2005, Alladi et al., 2008). They concluded that NN approach is superior to popular techniques. Only one study was reported important variables in prediction, based on gain values (Kassem-Fathy, 2011). He used 19 covariates as input variables and reported primary tumor size, historic stage A and largest tumor size have highest gain values.

In the present study, the ANN model was used to determine the important risk factors of CRC patients and also the results of this method was compared to Cox regression model. Our findings indicate that the ANN strategy was an appropriate technique for prediction and determining of risk factors and this method was efficient than Cox regression. Therefore, the ANN model is suggested to determine the important risk factors of survival probability of CRC patients.

Acknowledgements

References

About.com, Colon cancer (2011). Available at: http://coloncancer.about.com/

Ahmed FE (2005), Artificial neural networks for diagnosis and survival prediction in colon cancer. Mol cancer, Aug 6; 4:29. Available at: http://creativecommons.org/licenses/by/2.0

Alladi SM, Santosh S, Ravi V, et al (2008). Colon cancer prediction with genetic profiles using intelligent techniques. *Bioinformation*, **3**, 130-3

American Cancer Society (2008). Cancer Facts and Figures 1999–2008.

American Joint Committee on Cancer (1988). American Joint Committee on Cancer: AJCC Cancer Staging Manual (ed 3). Available at: http://www.cancerstaging.org/products/ajccproducts.html, Accessed June 20, 2006.

Anand SS, Smith AE, Hamilton PW, et al (1999). An evaluation of intelligent prognostic systems for colorectal cancer. *Artif Intell Med*, **15**, 193-214.

Asghari-Jafarabadi M, Hajizadeh E, Kazemnejad A, et al (2009). Site-specific evaluation of prognostic factors on survival in

Iranian colorectal cancer patients: A competing risks survival analysis. *Asian Pac J Cancer Prev*, **1**, 815-22.

Ansari R, Mahdavinia M, Sadjadi A, et al (2006). Incidence and age distribution of colorectal cancer in Iran: results of a population-based cancer registry. *Cancer Lett*, **18**, 143-7.

Bittern R, Cuschieri A, Dolgobrodov SD, et al(2005). An artificial neural network for analyzing the survival of patients with colorectal cancer. ESANN'2005 proceeding-Uropean symposium on artificial neural networks, Bruges (Belgium), 27-29 April 2005, ISBN 2-930307-05-6.

Bingham S, Day N, Luben R, et al (2003). Dietary fiber in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC): an observational study. *Lancet*, **361**, 1496–501.

Bishop C.M., Neural Networks for pattern recognition, Oxford University Press, New York, 1997.

Boyle P, Langman JS (2000). ABC of colorectal cancer. *BMJ*, **321**, 805-8.

Chan J, Meyerhardt J, Niedzwiekci D, et al (2008). Family History of colorectal cancer: A new survival predictor of colon cancer. *JAMA*, **299**, 2515–23.

Charles R, Thomas J (1992). Racial differences in the anatomical distribution of colon cancer. *Arch Surg*, **127**, 1241-5.

Cheng X, Chen VW, Steele B, et al (2001). Subsite-specific incidence rate and stage of disease in colorectal cancer by race, gender, and age group in the United States,1992-1997. *Cancer*, **92**, 2547-54.

Cho E, Smith-Warner SA, Ritz J, et al (2004). Alcohol intake and colorectal cancer: a pooled analysis of 8 cohort studies. *Ann Intern Med*, **140**, 603-13.

Corrao G, Bagnardi V, Zambon A, et al (2004). A meta-analysis of alcohol consumption and the risk of 15 diseases. *Prev Med*, **38**, 613-9.

Erhardt JG., Kreichgauer HP, Meisner C, et al(2002). Alcohol, cigarette smoking, dietary factors and the risk of colorectal adenomas and hyperplastic polyps. *Eur J Nutr*, **41**, 35-43.

Foroutan M, Rahimi N, Tabatabaeifar M, et al (2008) Clinical features of colorectal cancer in Iran: A 15-year review. *J Dig Dis*, **9**, 225-7.

Giovannucci E (2002). Modifiable risk factors for colon cancer. *Gastroenterol Clin North Am*, **31**, 925-43.

Giovannucci E, Rimm E, Ascherio A, et al (1995). Alcohol, lowmethionine--low-folate diets, and risk of colon cancer in men. *J Natl Cancer Inst*, **87**, 265-73.

Grumett S, Snow P, Kerr D (2003). Neural networks in the prediction of survival in patients with colorectal cancer. *Clin Colorectal Cancer*, **2**, 239-44.

Hosseini S, Izadpanah A, Yarmohammadi H (2004). Epidemiological changes in colorectal cancer in Shiraz, Iran: 1980–2000. *ANZ J Surg*, **74**, 547–9.

Ju J-H, Chang S-C, Wang H-S, et al (2007). Changes in disease pattern and treatment outcome of colorectal cancer: a review of 5,474 cases in 20 years. *Int J Colorectal Dis*, **22**, 855–62.

Kassem Fathy S (2011). A predication survival model for colorectal cancer. published in: proceeding AMERICAN-MATH'11/CEA'11 Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications, 36-42.

Kelsall HL, Laura B, David M, et al (2009). The effect of socioeconomic status on survival from colorectal cancer in the Melbourne Collaborative Cohort Study. *Soc Sci Med*, **68**, 290-7.

Kyung-Joong Kim, Sung-Bae Cho (2004). Prediction of colon cancer using an evolutionary neural network. *Neurocomputing*, **61**, 361-79.

Lai KC, Stotler BA (2010). Marital status and colon cancer stage at diagnosis. *Open Colorectal Cancer J*, **3**, 5-11

Lee SM, Kang JO, Suh YM (2004). Comparison of hospital charge prediction models for colorectal cancer patients: neural network vs. decision tree models. *J Korean Med Sci*, **19**, 677-81.

Le Marchand L, Wilkens L, Mi MP (1992) Obesity in youth and middle age and risk of colorectal cancer in men. *Cancer Causes Control*, **3**, 349–54.

Li M, Li JY, Zhao AL, et al (2007). Colorectal cancer or colon and rectal cancer? Clinicopathological comparison between colonic and rectal carcinomas. *Oncology*, **73**, 52-7.

MedicineNet, Inc (1996-2011). Available at: http://www.medicinenet.com/colon_cancer/article.htm

Mizoue T, Tanaka K, Tsuji I, et al (2006). Alcohol drinking and colorectal cancer risk: an evaluation based on a systematic review of epidemiologic evidence among the Japanese population. *Jpn J Clin Oncol*, **36**, 582–97.

Moghimi-Dehkordi B, Safaee A, Zali MR (2008). Prognostic factors in 1,138 Iranian colorectal cancer patients. *Int J Colorectal Dis*, **4**, 683-8.

National Cancer Institute (2010). Available at: http://www.cancer.gov/cancertopics/types/colon-and-rectal

Pahlavan PS, Jensen K (2005) .A short impact of epidemiological features of colorectal cancer in Iran. *Tumori*, **91**, 291-4.

Sadjadi A, Nouraie M, Malekzadeh R (2005). Cancer occurrence in Iran in 2002, an international perspective. *Asian Pac J Cancer Prev*, **6**, 359-60.

Safaee A, Moghimi-Dehkordi B, Fatemi SR, et al (2008) Colorectal cancer in Iran: an epidemiological study. *Asian Pac J Cancer Prev*, **9**, 123-6.

Tavani A, Fioretti F, Franceschi S, et al (1999). Education, Socieconomic status and risk of cancer of the colon and rectum. *Int J Epidemiol*, **28**, 380-5.

Toyoda Y, Nakayama T, Ito Y, et al(2009). Trends in Colorectal Cancer Incidence by Subsite in Osaka, Japan. *Jpn J Clin Oncol*, **39**, 189–91.

Troisi RJ, Freedman AN, Devesa SS (1999). Incidence of colorectal carcinoma in the U.S.: an update of trends by gender, race, age, subsite, and stage, 1975-1994. *Cancer*, **85**, 1670-6.

WHO (2008). http://globocan.iarc.fr/

Wu X, Chen VW, Martin J, et al (2004). Subsite- specific colorectal cancer incidence rates and stage distributions mong Asians and Pacific Islanders in the United States:1995 to 1999. *Cancer Epidemiol Biomarkers Prev*, **13**, 1215-22.