Kuwait J. Sci. **43 (2)** pp. 151-161, 2016

Estimation of risk factors associated with colorectal cancer: an application of knowledge discovery in databases

Feyza Firat[1], Ahmet K. Arslan[2], Cemil Colak[2], Hakan Harputluoglu[3]

[1]*Inonu University, Faculty of Medicine, Turgut Ozal Medical Centre, Department of Internal Disease, Malatya, Turkey.*

[2]*Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey.*

[3]*Inonu University, Faculty of Medicine, Turgut Ozal Medical Centre, Department of Oncology, Malatya, Turkey.*

Corresponding Author: Email: ahmetkadirarslan@gmail.com

Abstract

Colorectal cancer is one of the first reasons for death due to cancer in the world. The goal of this study is to predict important risk factors of colorectal cancer (CRC) by knowledge discovery in databases (KDD) methods. This study comprised a retrospective CRC data of patients who had been diagnosed with colorectal cancer. The selected records between 1 January 2010 and 1 March 2014 were collected randomly from Turgut Ozal Medical Centre databases. The study included 160 individuals: 80 patients admitted to Department of Oncology and diagnosed with CRC, and 80 control subjects with non-CRC categorization. The groups were matched for age and gender. We mined retrospective CRC data from large integrated health systems with electronic health records. Specific demographical and clinical variables including calcium, hemoglobin, white blood cells, platelets, potassium, sodium, glucose, creatinine and total bilirubin were used in multilayer perceptron (MLP) artificial neural networks (ANN) modeling. In this study, patient and control groups consist of 160 individuals. In each group, 45 of these (56.3%) are male, and 35 (43.7%) are women. Mean age of CRC patients and control groups is 58.6±13.0. While the accuracy was 71.31% in training dataset (n=122), the accuracy was 81.82% in testing dataset. Area under curve (AUC) values of training and testing datasets were 0.73 and 0.81, respectively. The suggested MLP ANN model identified significant factors of calcium, creatinine, potassium, platelets, sodium, hemoglobin and total bilirubin. Taken together, the suggested MLP ANN model might be used for the estimation of risk factors associated with CRC as an application of medical KDD.

Keywords: Artificial neural networks; colorectal cancer; knowledge discovery in databases; risk factors.

1. Introduction

Colorectal cancer (CRC) develops in the colon or the rectum. The colon and rectum are sections of the digestive system. CRC typically forms gradually, over a time of 10 to 20 years (Winaver & Zauber, 2002). CRC is characterized as a carcinoma, generally an adenocarcinoma, in the colon (Fleming *et al.,* 2012). CRC is one of the most well-known and most genuine malignancies around the world (Mogoanta *et al.,* 2014). CRC is commonly observed as a neoplasm in women and men (Durko & Malecka-Panas, 2014). CRC is one of the most common reasons for cancer-related deaths around the world (Chen *et al.,* 2014). The incidence has climbed drastically particularly in the most recent fifty years (Tammana & Laiyemo, 2014). Factors related with CRC are multiple and complicated (Zeng *et al.,* 2014). Cancer is one of the most common causes of death in Turkey (Andsoy & Gul, 2014).

Knowledge discovery in databases (KDD), also referred to as data mining, is the process of discovering relationships and patterns in huge databases (Holsheimer & Siebes, 1994). The whole process of KDD consists of five steps; first, a selection is made to extract a relevant or a target data set from the database. Then, preprocessing is carried out to discard noise and to solve missing data. A suitable transformation technique is applied to dataset. The number of variables is reduced by feature selection methods. An appropriate data mining method is used on the preprocessed data. Ultimately the findings of the data mining is interpreted and evaluated. If the discovered knowledge is not satisfactory, these live steps will be iterated. The discovered knowledge is then applied in decision making (Wong *et al.,* 2000).

In the present study, primary objective is to predict significant risk factors associated with CRC using KDD methods on the selected samples from the related databases, and second objective is to classify CRC, based on the selected significant risk factors.

2. Materials and methods

This study comprised retrospective data of patients, who had been diagnosed with CRC. The selected records between 1 January 2010 and 1 March 2014 were collected randomly from Turgut Ozal Medical Centre databases. The study included 160 individuals; 80 patients admitted to Inonu University Turgut Ozal Center of Medicine, Department of Oncology, diagnosed with CRC and 80 control subjects with non-colorectal cancer categorization. Colorectal cancer was diagnosed using colonoscopy and sigmoidoscopy. Ethical Review Board of Inonu University Faculty of Medicine extrapolated that informed consent of patients was not required for this retrospective analysis and approved the study (Date/Protocol number: 2014/63). All procedures in the present study were conformed to the Declaration of Helsinki.

The obtained data from the databases that were records of CRC patients and

healthy individuals (control) were used to determine significant risk factors associated with colorectal cancer. CRC patients and control groups were matched for age and gender. For this purpose, the studied variables among the risk factors for colorectal cancer: (Djuric *et al.,* 2012, Doubeni *et al.,* 2012, Galas *et al.,* 2013, Lee *et al.,* 2013, Pericleous *et al.,* 2013, Youmans *et al.,* 2012); calcium (mg/dl), hemoglobin (g/dl), white blood cells (10^3/ml), platelets (10^3/ml), potassium (mmol/L), sodium (mmol/L), glucose (mg/dl), creatinine (mg/dl), total bilirubin (mg/dl) were taken from the databases retrospectively.

In the application of KDD, the following steps were implemented (Fayyad *et al.,* 1996): The related steps for KDD are illustrated in Figure 1.

1. Data selection: Data relevant to the analysis task are taken from the database.
2. Data pre-processing: Data pre-processing step is the process of dissolution of noisy data, outliers or extreme values (Barakat *et al.,* 2014).
3. Data transformation: In this stage, the data in the previous step above were cleaned by applying various conversion operations made available to data mining.
4. Data mining: Building descriptive/predictive models.
5. Interpretation/Evaluation: Reaching descriptive/predictive models that best solve the problem (Gervilla Garcia *et al.,* 2009).

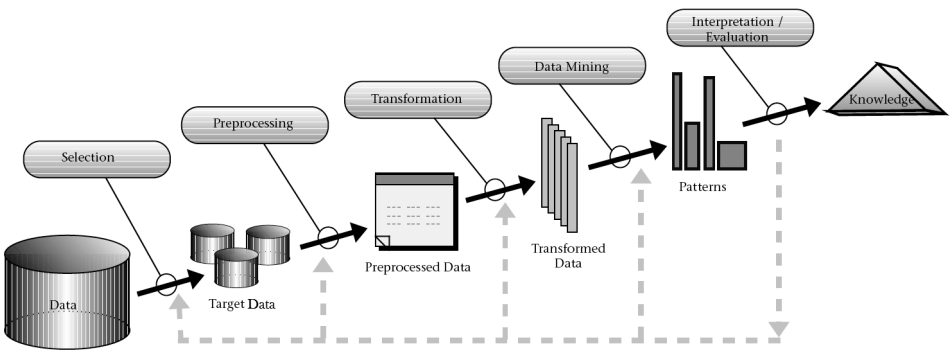

Fig. 1. An Outline of the steps of the KDD process (Fayyad *et al.,* 1996).

KDD methods and IBM SPSS Modeler Professional 16.0 for Windows software package were used for data analysis-modeling and advanced modeling-analysis, respectively. Power analysis suggested that each group included at least 78 individuals regarding average difference of calcium for patient and control groups of 0.34, estimated standard deviation of 0.75, Type I error (alpha) of 0.05 and Type II error (beta) of 0.20 (Minitab 16.2 for Windows). Normally and non-normally distributed data were summarized as Mean±SD (Min-Max) or Median (Min-Max), respectively.

3. Results

In this study, patient and control groups consist of 160 individuals (80 per group): in each group, 45 of these (56.3%) are male and 35 (43.7%) are women. Mean age of colorectal cancer patients and control groups is 58.6±13.0. Descriptive statistics of the variables for the subjects were tabulated in Table 1.

Table 1. Descriptive statistics of the variables for the subjects

Variables	**Groups**	
	Patient (*n*=80)	**Control (*n*=80)**
Age (Year) [Median (Min-Max)]	59 (29 - 84)	61 (29 - 84)
Calcium (mg/dl) [Mean±SD (Min-Max)]	8.965±0.766 (6.9 – 10.5)	9.2±0.776 (6.8 – 11.5)
Hemoglobin (g/dl) [Median (Min-Max)]	12.7 (7.2 – 17.4)	13.6 (0.1 – 16.8)
White Blood Cells (10^3/ml) [Median (Min-Max)]	7.9 (5 – 30.9)	8.6 (3.1 – 53.4)
Platelets (10^3/ml) [Median (Min-Max)]	311.5 (139 - 722)	240 (17 - 560)
Potassium (mmol/L) [Median (Min-Max)]	4.2 (3.1 – 6.9)	4.3 (3.03 - 58)
Sodium (mmol/L) [Median (Min-Max)]	137 (107 – 152.7)	138 (132 - 151)
Glucose (mg/dl) [Median (Min-Max)]	102.5 (61 - 249)	103 (11 - 347)
Creatinine (mg/dl) [Median (Min-Max)]	0.785 (0.4 – 2.35)	0.9 (0.46 – 8.31)
Total Bilirubin (mg/dl) [Median (Min-Max)]	0.515 (0.17 – 8.3)	0.5 (0.11 – 2.79)

The applied steps for KDD are explained below:

1. Data selection: From the database, the target/response variable was absence or presence of CRC, and the predictors were gender, calcium, hemoglobin, white blood cells, platelets, potassium, sodium, glucose, creatinine and total bilirubin.
2. Data pre-processing: Outliers in the data were inspected using T^2test based on the Mahalanobis distance. The identified outliers were removed, and remaining dataset (n=144) was used for following processes.

3. Data transformation: The variables of calcium, hemoglobin, white blood cells, platelets, potassium, sodium, glucose, creatinine and total bilirubin were transformed to standard units (Mean=0, Standard Deviation=1) called Z-transformation. Feature selection was applied by Pearson Chi-square test. The importance value of each variable was determined as (1-*p*) where *p* is the *p* value of the Pearson Chi-square test of association between the predictor and the target variable. Variables having importance greater than 0.80 were selected, based on the findings of feature selection method. Then, platelets, hemoglobin, sodium, creatinine, calcium and total bilirubin were chosen. Variable importance values are given in Table 2.

Table 2. Importance values of the variables

Variables	Importance Values
Platelets*	0.999
Hemoglobin*	0.998
Sodium*	0.997
Creatinine*	0.966
Calcium*	0.939
Total Bilirubin*	0.868
White Blood Cells*	0.537
Potassium*	0.237
Glucose*	0.043

*: Transformed variable (Mean=0, Standard Deviation=1)

4. Data mining: Multilayer perceptron (MLP) artificial neural networks (ANN) model was used for the prediction of absence or presence of CRC based on the selected predictors of platelets, sodium, creatinine, calcium, hemoglobin, potassium and total bilirubin. The general architecture of the MLP ANN was depicted in the Figure 2. MLP ANN is one of the feed-forward ANNs containing 1 input layer, 1 or more hidden layers, and 1 output layer. The MLP model uses a supervised neural network and is trained by a gradient descent method to minimize an error function (Celik *et al.,* 2014). Our MLP ANN had hidden layer with 3 neurons, and hidden layer activation function of hyperbolic tangent and output layer activation function of softmax. Relative predictor importance of the selected variables is given in Table 3. Relative importance for the predictors was calculated using the following formula:

$$VI_i = \frac{-log_{10}(sig_i)}{max_{j\in\Omega}\left(-log_{10}(sig_j)\right)}$$

where Ω defines the predictor set and the fields of evaluation, sig_i is the significance or p value. If sig_i is equal to zero, set $sig_i = MinDouble$, where $MinDouble$ is the minimal double value (IBM SPSS Modeler Professional 15 Algorithms Guide).

Based on the results of ANN model, relevant predictor importance of the selected variables in descending order was ranked as platelets, hemoglobin, sodium, total bilirubin, creatinine and calcium.

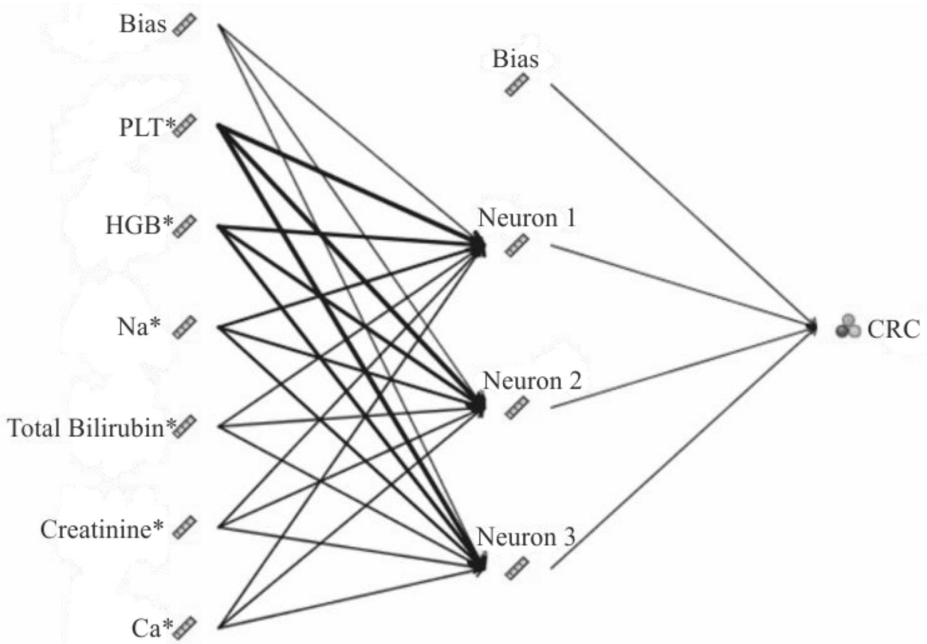

Fig. 2. General structure of the ANN process (*: Transformed variable (Mean=0, Standard Deviation=1))

Table 3. Relative predictor importance of the selected variables

Variables	Relative Predictor Importance
Platelets*	0.38
Hemoglobin*	0.26
Sodium*	0.16
Total Bilirubin*	0.08
Creatinine*	0.06
Calcium*	0.06

*: Transformed variable (Mean=0, Standard Deviation=1)

5. Interpretation/Evaluation: The MLP ANN model generated the accuracy of 71.31% in training dataset (n=122) and the accuracy of 81.82% in testing dataset. In addition, area under curve (AUC) values of training and testing datasets were 0.73 and 0.81, respectively.

4. Conclusions

CRC is one of the first reasons for death due to cancer in the world. In this study, some significant risk factors of CRC were predicted based on KDD. CRC is associated with many novel and conventional risk factors such as platelets, calcium, vitamin D, BMI, red meat, fat, alcohol, dietary fiber, APC (*adenomatous polyposis coli*) mutation, microRNA, smoking, unhealthy diet, physical inactivity, age, education, vitamin C, dietary iron, ulcerative colitis, acromegaly, diabetes mellitus, ischemic heart disease (Doubeni *et al.,* 2012, Galas *et al.*, 2013, Kajzrlikova *et al.*, 2014, Pericleous *et al.*, 2013, Youmans *et al.*, 2012). In the present study, calcium, hemoglobin, white blood cells, platelets, potassium, sodium, glucose, creatinine and total bilirubin were examined to associate with CRC.

Based on the analyses of KDD, the most significant risk factor associated with CRC was platelets. CRC is significantly associated with platelets. Pretreatment thrombocytosis may be accepted as beneficial prognostic markers in CRC patients (Al-Saeed *et al.*, 2014). It was reported that an advanced platelets/lymphocyte ratio has been connected to prognosis in various malignancies (Farazi, 2014, Templeton *et al.*, 2014). Therefore, platelets may be a significant risk factor in CRC patients. Additionally, the study identified preoperative thrombocytosis to be a negative prognostication variable in different solid tumors and examined, whether thrombocytosis was one of the risk factors for CRC (Baranyai *et al.*, 2014, Erlinger *et al.*, 2004). On the other hand, a significant difference in platelets was determined between CRC patients and control groups. Platelets activation was determined in inflammatory illness and intestinal tumorigenesis (Dovizio *et al.*, 2014).

From the results of this study, hemoglobin was found to be another significant factor for CRC. Pretreatment hemoglobin was contrarily related with essential tumor size and nodal condition. Right-sided CRC had fundamentally low pretreatment hemoglobin. Curiously, pretreatment thrombocytosis was observed merely in right-sided CRC (Al-Saeed *et al.*, 2014).

It was reported that sodium intake cannot improve the risk of cancer (Takachi *et al.*, 2010). Conversely, salted nourishment intake might augment the risk of cancer (Tsugane, 2005). Also, it was demonstrated that sodium restrained the growth of different cancers (Tailor *et al.*, 2014). From the present study, sodium was found to be related to CRC as a significant factor.

The bile pigment bilirubin is an antioxidant and is related with the preservation from cancer. In addition, unconjugated bilirubin could have helped against genotoxic potential by protecting oxidative damage to DNA (Wallner *et al.*, 2013). In our study, total bilirubin was found to be associated with CRC among the studied factors.

Several risk factors such as body mass index, family history of CRC, genetic syndromes, alcohol use, smoking, obesity and so on, associated with CRC were not evaluated in the current study owing to not available in the studied databases. For this reason, the accuracy of the suggested MLP ANN model was not quite high as desired. If the mentioned other risk factors not evaluated in the current study are used to predict CRC, the suggested MLP ANN model will generate much higher accuracy. A study employed ANNs to construct a model for predicting CRC and reported that ANNs can recognize patterns in CRC data sets (Bottaci *et al.*, 1997). Another limitation may be a retrospective study design constraining the information obtained from the databases and the sample size used in the present study. Therefore, prospective studies are more useful for further evaluation of the analysis of risk factors. In addition, much larger sample sizes of individuals will be able to increase the classification accuracy and other KDD methods apart from MLP ANN can be evaluated for predicting CRC in the future studies.

Taken together, the suggested MLP ANN model might be used for the estimation of risk factors associated with CRC as an application of medical KDD.

References

Al-Saeed, E.F., Tunio, M.A., Al-Obaid, O., Abdulla, M. & Al-Anazi, A. *et al.* (2014) Correlation of pretreatment hemoglobin and platelet counts with clinicopathological features in colorectal cancer in Saudi population, Saudi Journal of Gastroenterology, **20**(2):134-138.

Andsoy, II. & Gul, A. (2014) Breast, cervix and colorectal cancer knowledge among nurses in Turkey, Asian Pacific Organization for Cancer Prevention, **15**(5):2267-2272.

Barakat, H., Nigm, E. & Khaled, O. (2014) Statistical modeling of extremes under linear and power normalizations with applications to air pollution, Kuwait Journal of Science, **41** (1):1-19.

Baranyai, Z., Krzystanek, M., Josa, V., Dede, K. & Agoston, E. *et al.* (2014) The comparison of thrombocytosis and platelet-lymphocyte ratio as potential prognostic markers in colorectal cancer, Thrombosis and Haemostasis, **111**(3):483-490.

Bottaci, L., Drew, P.J., Hartley, J.E., Hadfield, M.B. & Farouk, R. *et al.* (1997) Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions, The Lancet, **350**(9076):469-472.

Celik, G., Baykan, O.K., Kara, Y. & Tireli, H. (2014) Predicting 10-day mortality in patients with strokes using neural networks and multivariate statistical methods, Journal of Stroke and Cerebrovascular Diseases, **23**(6):1506-1512.

Chen, D., Huang, J.F., Liu, K., Zhang, L.Q. & Yang, Z. *et al.* (2014) BRAFV600E mutation and its association with clinicopathological features of colorectal cancer: a systematic review and meta-analysis, PLoS One, **9**(3):e90607.

Djuric, Z., Ruffin, M.T.T., Rapai, M.E., Cornellier, M.L. & Ren, J. *et al.* (2012) A Mediterranean dietary intervention in persons at high risk of colon cancer: recruitment and retention to an intensive study requiring biopsies, Contemporary Clinical Trials, **33**(5):881-888.

Doubeni, C.A., Major, J.M., Laiyemo, A.O., Schootman, M. & Zauber, A.G. *et al.* (2012) Contribution of behavioral risk factors and obesity to socioeconomic differences in colorectal cancer incidence, Journal of the National Cancer Institute, **104**(18):1353-1362.

Dovizio, M., Alberti, S., Guillem-Llobat, P. & Patrignani, P. (2014) Role of platelets in inflammation and cancer: novel therapeutic strategies, Basic & Clinical Pharmacology & Toxicology, **114**(1):118-127.

Durko, L. & Malecka-Panas, E. (2014) Lifestyle modifications and colorectal cancer, Current Colorectal Cancer Reports, **10**:45-54.

Erlinger, T.P., Muntner, P. & Helzlsouer, K.J. (2004) WBC count and the risk of cancer mortality in a national sample of U.S. adults: results from the Second National Health and Nutrition Examination Survey mortality study, Cancer Epidemiology, Biomarkers & Prevention, **13**(6):1052-1056.

Farazi, P.A. (2014) Cancer trends and risk factors in Cyprus, Ecancermedicalscience, **8**:389.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996) From data mining to knowledge discovery in databases, Artificial Intelligence Magazine, **17**(3):37.

Fleming, M., Ravula, S., Tatishchev, S.F. & Wang, H.L. (2012) Colorectal carcinoma: Pathologic aspects, Journal of Gastrointestinal Oncology, **3**(3):153-173.

Galas, A., Augustyniak, M. & Sochacka-Tatara, E. (2013) Does dietary calcium interact with dietary fiber against colorectal cancer? A case-control study in Central Europe, Nutrition Journal, **12**:134.

Gervilla Garcia, E., Jimenez Lopez, R., Montano Moreno, J.J., Sese Abad, A. & Cajal Blasco, B. *et al.* (2009) The methodology of Data Mining. An application to alcohol consumption in teenagers, Adicciones, **21**(1):65-80.

Holsheimer, M. & Siebes, A. (1994) Data mining: The search for knowledge in databases. CWI Amsterdam.

Kajzrlikova, I.M., Vitek, P., Chalupa, J. & Dite, P. (2014) Dietary habits of colorectal neoplasia patients in comparison to their first-degree relatives, World Journal of Gastroenterology, **20**(17):5025-5030.

Lee, C.K., Kim, Y.W., Shim, J.J. & Jang, J.Y. (2013) Prevalence of proximal serrated polyps and conventional adenomas in an asymptomatic average-risk screening population, Gut and Liver, **7**(5):524-531.

Mogoanta, S.S., Vasile, I., Totolici, B., Neamtu, C. & Streba, L. *et al.* (2014) Colorectal cancer - clinical and morphological aspects, Romanian Journal of Morphology and Embryology, **55**(1):103-110.

Pericleous, M., Mandair, D. & Caplin, M.E. (2013) Diet and supplements and their impact on colorectal cancer, Journal of Gastrointestinal Oncology, **4**(4):409-423.

Tailor, D., Hahm, E.R., Kale, R.K., Singh, S.V. & Singh, R.P. (2014) Sodium butyrate induces DRP1-mediated mitochondrial fusion and apoptosis in human colorectal cancer cells, Mitochondrion, **16**:55-64.

Takachi, R., Inoue, M., Shimazu, T., Sasazuki, S. & Ishihara, J. *et al.* (2010) Consumption of sodium and salted foods in relation to cancer and cardiovascular disease: the Japan Public Health Center-based Prospective Study, The American Journal of Clinical Nutrition, **91**(2):456-464.

Tammana, V.S. & Laiyemo, A.O. (2014) Colorectal cancer disparities: issues, controversies and solutions, World Journal of Gastroenterology, **20**(4):869-876.

Templeton, A.J., Ace, O., McNamara, M.G., Al-Mubarak, M. & Vera-Badillo, F.E. *et al.* (2014) Prognostic role of platelet to lymphocyte ratio in solid tumors: a systematic review and meta-analysis, Cancer Epidemiology, Biomarkers & Prevention, **23**(7):1204-1212.

Tsugane, S. 2005. Salt, salted food intake, and risk of gastric cancer: epidemiologic evidence, Cancer Science, **96**(1):1-6.

Wallner, M., Antl, N., Rittmannsberger, B., Schreidl, S. & Najafi, K. *et al.* (2013) Anti-genotoxic potential of bilirubin in vivo: damage to DNA in hyperbilirubinemic human and animal models, Cancer Prevention Research (Philadelphia, Pa.), **6**(10):1056-1063.

Winawer, S.J. & Zauber, A.G. (2002) The advanced adenoma as the primary target of screening, Gastrointestinal Endoscopy Clinics of North America, **12**(1):1-9, v.

Wong, M.L., Lam, W., Leung, K.S., Ngan, P.S. & Cheng, J.C. (2000) Discovering knowledge from medical databases using evolutionary algorithms, IEEE Engineering in Medicine and Biology Magazine, **19**(4):45-55.

Youmans, L., Taylor, C., Shin, E., Harrell, A. & Ellis, A.E. *et al.* (2012) Frequent alteration of the tumor suppressor gene APC in sporadic canine colorectal tumors, PLoS One, **7**(12):e50813.

Zeng, H., Lazarova, D.L. & Bordonaro, M. (2014) Mechanisms linking dietary fiber, gut microbiota and colon cancer prevention, World Journal of Gastrointestinal Oncology, **6**(2):41-51.

Submitted **:** 31/12/2014
Revised **:** 14/06/2015
Accepted **:** 24/06/2015

تقدير عوامل الخطر المرتبطة بسرطان القولون والمستقيم: تطبيق اكتشاف المعرفة في قواعد البيانات

[1]فايزة فرات، [2]أحمد أرسلان، [2]سيميل كولاك، [3]هاكان هاربوتلوغلو

[1]جامعة إينونو – كلية الطب – مركز ترغوت أوزال الطبي – قسم الامراض الداخلية – مالاتيا – تركيا.

[2]جامعة إينونو – كلية الطب – قسم الاحصاء الحيوي والمعلومات الطبية – مالاتيا – تركيا.

[3]جامعة إينونو – كلية الطب – قسم علم الأورام – مالاتيا – تركيا.

المؤلف: ahmetkadirarslan@gmail.com

خلاصة

سرطان القولون والمستقيم هو واحد من الأسباب الأولى للوفاة بسبب السرطان في العالم. الهدف من هذه الدراسة هو التنبؤ بعوامل الخطر الهامة بالنسبة إلى سرطان القولون والمستقيم (CRC) من خلال أساليب اكتشاف المعرفة في قواعد البيانات(KDD). تضمنت هذه الدراسة بيانات CRC بأثر رجعي من المرضى الذين تم لهم تشخيص سرطان القولون والمستقيم. تم جمع السجلات المحددة ما بين 1 يناير 2010 و 1 مارس 2014 عشوائيا من قواعد بيانات المركز الطبي تورغوت أوزال. وشملت الدراسة 160 شخصا: 80 المرضى الذين تم إدخالهم إلى قسم الأورام وتشخيصها CRC، و 80 مع عدم التصنيف CRC. وجرى الربط بين المجموعات بالنسبة للعمروالجنس. تم تلغيم بيانات CRC بأثررجعي من النظم الصحية المتكاملة مع السجلات الصحية الإلكترونية. استخدمت المتغيرات الديموغرافية والسريرية المحددة بما في ذلك الكالسيوم، والهيموجلوبين، وخلايا الدم البيضاء والصفائح الدموية والبوتاسيوم والصوديوم والجلوكوز، والكرياتينين، والبيليروبين الكلي في المستقبلات متعددة الطبقات (MLP) وفي نماذج في الشبكات العصبية الاصطناعية (ANN). وفي هذه الدراسة تكونت مجموعة المرضى ومجموعات المراقبة من 160 فردا. في كل مجموعة: 45 من هؤلاء (56.3٪) هم من الذكور،و 35 (43.7٪) من النساء. متوسط العمر لمرضى CRC ومجموعات المراقبة بلغ 58.6 ± 13.0. بينما كانت الدقة 71.31٪ في مجموعة البيانات التدريبية (ن = 122)،وكانت الدقة 81.82٪ في اختبار البيانات. وكانت المنطقة تحت المنحنى (AUC) لقيم تدريب واختبار قواعد البيانات 0.73 و 0.81 على التوالي. نموذج MLP ANN المقترح حدد عوامل هامة هي الكالسيوم، والكرياتينين، والبوتاسيوم، والصفائح الدموية، والصوديوم، والهيموجلوبين والبيليروبين الكلي. وبأخذها مجتمعة فقد يتم استخدام نموذج MLP ANN المقترح لتقديرعوامل الخطر المرتبطة بال CRC كتطبيق لاكتشاف المعرفة في قواعد البيانات KDD الطبية.

الكلمات المفتاحية:

الشبكات العصبية الاصطناعية، سرطان القولون، اكتشاف المعرفة في قواعد البيانات، عوامل الخطر