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

Feyza Firat¹, Ahmet K. Arslan², Cemil Colak², Hakan Harputluoglu³

¹Inonu University, Faculty of Medicine, Turgut Ozal Medical Centre, Department of Internal Disease, Malatya, Turkey.
²Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey.
³Inonu 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.
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).](image)

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>
<thead>
<tr>
<th>Variables</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patient (n=80)</td>
</tr>
<tr>
<td>Age (Year) [Median (Min-Max)]</td>
<td>59 (29 - 84)</td>
</tr>
<tr>
<td>Calcium (mg/dl) [Mean±SD (Min-Max)]</td>
<td>8.965±0.766 (6.9 – 10.5)</td>
</tr>
<tr>
<td>Hemoglobin (g/dl) [Median (Min-Max)]</td>
<td>12.7 (7.2 – 17.4)</td>
</tr>
<tr>
<td>White Blood Cells (10³/ml) [Median (Min-Max)]</td>
<td>7.9 (5 – 30.9)</td>
</tr>
<tr>
<td>Platelets (10³/ml) [Median (Min-Max)]</td>
<td>311.5 (139 - 722)</td>
</tr>
<tr>
<td>Potassium (mmol/L) [Median (Min-Max)]</td>
<td>4.2 (3.1 – 6.9)</td>
</tr>
<tr>
<td>Sodium (mmol/L) [Median (Min-Max)]</td>
<td>137 (107 – 152.7)</td>
</tr>
<tr>
<td>Glucose (mg/dl) [Median (Min-Max)]</td>
<td>102.5 (61 - 249)</td>
</tr>
<tr>
<td>Creatinine (mg/dl) [Median (Min-Max)]</td>
<td>0.785 (0.4 – 2.35)</td>
</tr>
<tr>
<td>Total Bilirubin (mg/dl) [Median (Min-Max)]</td>
<td>0.515 (0.17 – 8.3)</td>
</tr>
</tbody>
</table>

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 $T^2$ test 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

<table>
<thead>
<tr>
<th>Variables</th>
<th>Importance Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platelets*</td>
<td>0.999</td>
</tr>
<tr>
<td>Hemoglobin*</td>
<td>0.998</td>
</tr>
<tr>
<td>Sodium*</td>
<td>0.997</td>
</tr>
<tr>
<td>Creatinine*</td>
<td>0.966</td>
</tr>
<tr>
<td>Calcium*</td>
<td>0.939</td>
</tr>
<tr>
<td>Total Bilirubin*</td>
<td>0.868</td>
</tr>
<tr>
<td>White Blood Cells*</td>
<td>0.537</td>
</tr>
<tr>
<td>Potassium*</td>
<td>0.237</td>
</tr>
<tr>
<td>Glucose*</td>
<td>0.043</td>
</tr>
</tbody>
</table>

*: 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:
where $\Omega$ 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 $sig_i = \text{MinDouble}$, where $\text{MinDouble}$ 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.

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

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

**Table 3.** Relative predictor importance of the selected variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Relative Predictor Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platelets*</td>
<td>0.38</td>
</tr>
<tr>
<td>Hemoglobin*</td>
<td>0.26</td>
</tr>
<tr>
<td>Sodium*</td>
<td>0.16</td>
</tr>
<tr>
<td>Total Bilirubin*</td>
<td>0.08</td>
</tr>
<tr>
<td>Creatinine*</td>
<td>0.06</td>
</tr>
<tr>
<td>Calcium*</td>
<td>0.06</td>
</tr>
</tbody>
</table>

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


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

فايزة فرات, 2 أحمد أرسلان, 2 سميث كولاك, 3 هاكان هاربولغلو

1 جامعة إسطنبول - كلية الطب - مركز ترثوت أوزال الطبي - قسم الأمراض الداخلية - مالاتيا - تركيا.
2 جامعة إسطنبول - كلية الطب - قسم الامراض الخبيثة والمعلومات الطبية - مالاتيا - تركيا.
3 جامعة إسطنبول - كلية الطب - قسم علم الأورام - مالاتيا - تركيا.

ahmetkadirarslan@gmail.com

خلاصة

سرطان القولون والمستقيم هو واحد من الأسباب الأولي للوفاة بسبب السرطان في العالم. الهدف من هذه الدراسة هو التنبؤ بعوامل الخطر الحالية بالنسبة إلى سرطان القولون والمستقيم من CRC (KDD). تضمنت هذه الدراسة بيانات BIRCH خلال أسابيع اكتشاف المعرفة في قواعد البيانات BIRCH. تم جمع البيانات المحدثة ما بين 1 يناير 2010 و 1 مارس 2014 عشوائيا من قواعد بيانات المركز الطبي ترثوت أوزال. وشملت الدراسة 160 شخصا: 80 المرضى الذين تم إدخالهم إلى قسم الأورام وتشخيصهما و 80 بعد التصنيف. W و CRC, و 80 مع عدم التصنيف. و تأثر بربح بين المجموعات بالنسبة للعمروالجنس. تم تلميح بيانات BIRCH كإشارات من النظم الصحية المتكاملة مع السجلات الصحية الإلكترونية. استخدمت المنتطورات الدموغرافية والسريرية المحدثة بما في ذلك الكالسيوم، والهيموجلوبين، وخلايا الدم البيضاء، والصفائح الدموية، والبوتاسيوم والجلوكوز، والكرياتين، والبيوبيلين الكلي للمستقبلات متعددة القضايا (MLP) في ماذج في الشبكات العصبية الاصطناعية (ANN). و في هذه الدراسة تكونت المجموعات المتعلقة بـ ANN، وأفراد في كل مجموعة 45 من هؤلاء (56.3%) هم من الذكور، و35 (24.3%) من النساء. المعدل العمر المرضي في CRC متوسط المجموعات المراقبة بلغ 58.6 ± 13.0. بينما كانت الدقة 71.3% في مجموعة البيانات التشريحي (N = 122), وكانت الدقة 81.8% في اختبار البيانات. وكانت المنطقة تحت المتكشفة (AUC) لقيم تدريب وأختبار قواعد البيانات 0.73 إلى 0.81 على التوالي. نموذج الإفراز المثير حدد عوامل هامة هي الكالسيوم، والكرياتينين، والبوتاسيوم، والصفائح الدموية، والهيموجلوبين والبيوبيلين الكلي. وبالإضافة إلى ذلك، لم يتم استخدام الاتصالات المتغيرة مع CRC كتنبيه لاكتشاف المعرفة في KDD قواعد البيانات.

الكلمات المفتاحية:

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