DOI: 10.15171/mejdd.2019.148

Predictive Factors of Advanced Colonic Adenomas and Cancer Using Data Mining

Atieh Sadat Fatemian [1], Neda Abdolvand [1,*], Hamideh Salimzadeh [2], Alireza Delavari [2]

1. Faculty of Economics and Social Sciences, Alzahra University, Tehran, Iran
2. Digestive Disease Research Institute, Shariati Hospital, Tehran, Iran

* **Corresponding Author:**
Neda Abdolvand, PHD
Associate Professor of Department of IT Management, Faculty of Economic and Social Sciences, Alzahra University, Sheikh Bahaei Sq., Tehran, Iran
Tel: + 98 2185692369
Fax: + 98 2188047862
Email: N.abdolvand@alzahra.ac.ir

Received: 07 May. 2019
Accepted: 09 Sep. 2019

ABSTRACT

BACKGROUND

Colorectal cancer is the third common cancer in Iran. In this study we aimed to identify factors associated with the prevalence of advanced colonic neoplasms among a high-risk population.

METHODS

Participants were 474 first degree relatives of patients with colon cancer who underwent a screening colonoscopy at Digestive Disease Research Institute, Shariati Hospital affiliated to Tehran University of Medical Sciences. Features examined in this study were age, sex, body mass index, Aspirin use, smoking, and relationship type with patients with cancer in family. Also, patient's age at the time of cancer diagnosis, number and sex of the patients with colon cancer in the family were assessed. Data analysis was performed by data mining methods using K-Medoid clustering and decision tree C4.5.

RESULTS

Results showed that female sex of the patients with colon cancer and their young age (< 60 years old) at the time of cancer diagnosis were important predictive factors for the prevalence of colorectal advanced neoplasms among their family members.

CONCLUSION

Data mining methods were found to be applicable in recognizing predictive factors of colorectal advanced neoplasms in each cluster and tree.

KEYWORDS:
Colorectal Cancer, Data Mining, Clustering, Decision Tree, Crisp Methodology

Please cite this paper as:
Sadat Fatemian A, Abdolvand N, Salimzadeh H, Delavari AR. Predictive Factors of Advanced Colonic Adenomas and Cancer Using Data Mining. *Middle East J Dig Dis* 2019;**11**:192-198. doi:10.15171/mejdd.2019.148.

INTRODUCTION

In 2018, among all cancers, colorectal cancer (CRC) was the 3rd most common cancer and the second leading cause of death worldwide. [1] According to the GOLOBOCAN 2018, CRC is also the 3rd common neoplasm with mostly 10000 new cases per year in Iran.[2] Early diagnosis of CRC is very important for better survival and effective treatment of the disease.[3] Family history of CRC has been reported to be an important risk factor for CRC development.[4] It is recommended that first degree relatives (FDRs) of patients with CRC undergo screening colonoscopy starting at age of 40 or 10 years before the diagnosis

age of the index patient in their family.[5,6]

Several factors are known to be associated with CRC such as genetics[7] and environmental elements e.g., air pollution, nutrition,[8] smoking,[9] and obesity.[10] As the number of potential risk factors of CRC increases, the number of all possible variable combinations that might explain the prevalence of colorectal advanced neoplasms will increase significantly. Therefore, it would be challenging to comprise all of these candidate combinations in a single statistical model. This is particularly right when the research question exhausts theory and becomes non-hypothesis driven; establishing a statistical model for testing becomes time consuming. However, strategies have appeared that may help identification of item sets, collections of variables that are relevant for specific populations and are easily used in practice settings.[11]

Recently new techniques such as data mining have discovered the hidden knowledge from massive datasets. Data mining is a knowledge discovery tool and is used in information analysis. This approach might be useful for anticipation, classification, and discerning relationships between factors.[12] These extracted data would be helpful for prevention, treatment, or improving patients' care.[3] Depending on data types, different data mining techniques have been used. In this area, data mining has been used for many studies, such as identifying CRC contributing factors just by considering lifestyle without family history consideration that used integrative data mining and regression.[11] Also, another study demonstrated the use of classification for CRC diagnosis by genes.[7] Furthermore, by using the Bayesian approach and regression, one study showed the relation between living in a polluted and industrial environment and CRC death rate.[13] Another important factor is nutrition that a study evaluated the impact of food on CRC treatment by SVM (Support Vector Machines) and KNN (K Nearest Neighbors).[8] In this study, we used K-Medoid, which is one of the clustering methods, to identify different clusters and then applied C4.5 tree on each cluster to extract the important factors associated with the prevalence of colorectal advanced neoplasms.

MATERIALS AND METHODS

In this study we used the data of a screening study conducted in the Digestive Disease Research Institute (DDRI) affiliated to Tehran University of Medical Sciences. Data were collected from 474 participants who underwent a screening colonoscopy in Shariati Hospital and had at least one patient with CRC among their immediate family members. Our dataset consisted of age, sex, body mass index (BMI), Aspirin usage, smoking, relationship with patient in the family (parent or sibling), number of patients in the family, age of the patient with CRC in the family at the time of diagnosis, and sex of the patient with CRC in the family. Outcome variable was the prevalence of colorectal advanced neoplasms among the family members.

Data mining is a knowledge discovery tool.[12] Specifically, in this study we used the Crisp methodology to implement data mining techniques on the dataset of the study. Crisp-DM is a cross-industry standard process for data mining, which is popular among industry members. This cycle process contains six steps: business understanding, data understanding, data preparation, data modeling, evaluation, and deployment. These steps are not identical for every process and any step may be non-essential for empirical analysis in a scientific study. Crisp-DM has provided a good framework for data mining.[14] Business understanding in this study means the necessity of using data mining in this area. By using data mining, higher accuracy is achievable for recognition of factors related to the incidence of illnesses. In the data understanding phase, the features of data are examined and it is necessary to understand the range and type of data presented. In this study, some features such as BMI, number of patients with CRC in family, participant's age, and patient's age at the time of diagnosis of CRC were numerical and other attributes i.e., sex, smoking, Aspirin usage, type of relationship with the patients with CRC, and the detection of colonic cancers or neoplasms were nominal variables.

In the preparation data phase, data are changed to be suitable for data processing so the numerical data are discretized and an empty record filled with 'no data', which were just in the BMI attribute. Sex has had an important role to play in this disease,[11] therefore, data of female and male participants were analyzed separately. In the data modeling phase, data were first clustered using K-Medoid. Prominent features of clusters should be described since details of each cluster is notable for better

results understanding of implemented trees on each cluster in the next step. This technique is an unsupervised method and puts similar data in a cluster. Data differences among clusters must be great.[15]

After this, decision tree C4.5 was implemented on each cluster to show the effective factors of cancer incidence in each cluster separately. Decision tree is a classification method. Classification is a supervised method, which means learning a function to put data in predefined class.[16] One of the most popular classification methods is decision tree and in every decision tree a sample is examined from the first node to the last one. Sample in each node is examined in regards to the separation rule and if it is true in a rule directed to the next node. This approach is continued until no node left.[17]

In the evaluation step, the model must be examined. Davies-Bouldin is an evaluation method for clusters and its aim is measuring inner cluster similarities and between clusters dissimilarities. If μi is the mean of a cluster:

$$\mu_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$$

And if $\sigma_{\mu i}$ is the average deviation of a cluster mean:

$$\sigma_{\mu_i} = \sqrt{\frac{\sum_{\mathbf{x}_j \in C_i} \delta(\mathbf{x}_j, \mu_i)^2}{n_i}} = \sqrt{var(C_i)}$$

Then Davies-Bouldin measure for the cluster Ci and Cj would be like this:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \{DB_{ij}\}$$

Smaller amount of Davies-Bouldin would be preferable because the lower amount of Davies-Bouldin measure means that there is more similarities in each cluster and more dissimilarities between two different clusters.[18] For decision tree evaluation, F-measure has been used. This method is the combination of recall and precision. Precision is positive predictive value that is the fraction of relevant instances among the retrieved instances. Recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. F-measure is a better measure than precision and recall and only has upper value when both of them are high.[19]

f-measure = 2 ((prec × rec) / (prec + rec))

Deployment will be depending on the result, if the results were useful, it would be used in health care for precaution.

RESULTS

table 1 illustrates information about clusters in both sexes. Each row represents factors that have been processed and each column demonstrates different values of each factor. First cluster of men contains more than 40% of all male participants and first cluster of women covers more than 50% of female members. More precise information about clusters will be presented in the following.

After clustering, according to the lowest values of Davies-Bouldin measure, the best number of clusters for male participants was 4 and this measure was 0.066 in this respect. Moreover, the best number of clusters for female participants was 3 and Davies-Bouldin measure was 0.061.

Repeated pattern was seen in some clusters. Factors such as having young patient with CRC (less than 60 years old at the time of diagnosis of CRC) and female patient with CRC in family were associated with the prevalence of colorectal advanced neoplasms. This pattern was also observed in the second and fourth male clusters and the first female cluster of the participants. One of the eminent features of these clusters was the younger age of participants (mostly under 50 years old). Having young patient with CRC (< 60 years old) also was an important factor for the prevalence of colorectal advanced neoplasms among participants in the second and third female's clusters. Besides, having one female patient with CRC in the family was associated with the prevalence of colorectal advanced neoplasms in the third male participant's cluster who were mostly under 50 years old.

Having young patient with CRC (under 60 years old at the time of diagnosis of CRC) and female patient with CRC in family were associated with the prevalence of colorectal advanced neoplasms in some clusters that they are described below. In the second male's cluster, 91% were younger than 50 years old, 87% had a male patient with CRC in their family, 65% had a young patient with CRC in family (under 60 years old at the time of diagnosis of CRC). F-measure for decision tree in this cluster was 89.66%.

Table 1: Characteristics of the clusters data

				Male (n = 224)			Female (n = 250)		
Attributes	**Value**	**Total, n (%)**	**cluster.1**	**cluster 2**	**cluster 3**	**cluster 4**	**cluster 1**	**cluster 2**	**cluster**
Screened family members		474 (100)	90	23	64	47	123	67	60
Age of screened participants	age < 50	290 (61)	28	21	32	47	101	58	3
	age >= 50	184 (39)	62	2	32	0	22	9	57
Smoking	No	399 (84)	76	15	31	37	116	66	58
	Yes	75 (16)	14	8	33	10	7	1	2
Aspirin use	No	426 (90)	73	19	57	46	114	64	53
	Yes	48 (10)	17	4	7	1	9	3	7
No. of patients with CRC in family	1	367 (79)	70	6	57	39	107	45	43
	2	52 (10)	16	4	3	4	5	8	12
	3	38 (8)	2	7	3	4	7	12	3
	4	17 (3)	2	6	1	0	4	2	2
Sex of patients with CRC	Female	286 (60)	31	3	43	38	58	55	58
	Male	188 (40)	59	20	21	9	65	12	2
Age at the time of diagnosis of CRC	age >= 60	141 (30)	10	8	51	9	28	20	15
	age < 60	333 (70)	80	15	13	38	95	47	45
Type of relationship with the CRC patients	Parenting	202 (43)	10	13	42	33	42	50	12
	Sibling	272 (57)	80	10	22	14	81	17	48
Body mass index	BMI >= 25	259 (55)	37	12	55	12	41	57	45
	BMI < 25	128 (27)	21	10	9	30	52	5	1
	no data	87 (18)	32	1	0	5	30	5	14
Colonic advanced neoplasms in screened members		63 (13)	19	10	7	4	9	4	10

CRC: Colorectal cancer

In the fourth male's cluster, 83% of the participants reported having one patient with CRC, 100% were younger than 50 years old, 81% had a female patient with CRC in their family, 81% has a young patient with CRC in family (under 60 years old at the time of diagnosis of CRC), and 70% had ill parent. F-measure for decision tree in this cluster was 84.85%.

In the first female's cluster, 87% of the participants reported having one patient with CRC, 82% were younger than 50 years old, 77% had a young patient with CRC in family (under 60 years old at the time of diagnosis of CRC), and 66% of the relationships with patient in the family were being sibling. F-measure for decision tree in this cluster was 70.03%. Having one female patient with CRC in the family was associated with the prevalence of colorectal advanced neoplasms in the third male participant's cluster.

In this cluster, 89% of the participants reported having one patient with CRC in the family, 67% had a female patient with CRC in their family, 80% has an old patient with CRC in the family (over 60 years old at the time of diagnosis of CRC), 66% of relationships with patients in the family were parenting, and 85% had high BMI, which was more than 25. F-measure for decision tree in this cluster was 88.24%.

Having young CRC patient (< 60 years old) also was an important factor for the prevalence of colorectal advanced neoplasms among the participants in the 2nd and 3rd female clusters. In the second female's cluster, 67% had one patient with CRC, 87% were younger than 50 years old, 80% had a female patient with CRC in their family, 70% had a young patient with CRC in the family (under 60 years old at the time of diagnosis of CRC), 75% had ill parent, and 86% was overweight. F-measure for decision tree in this cluster was 85.45%. In the third female's cluster,

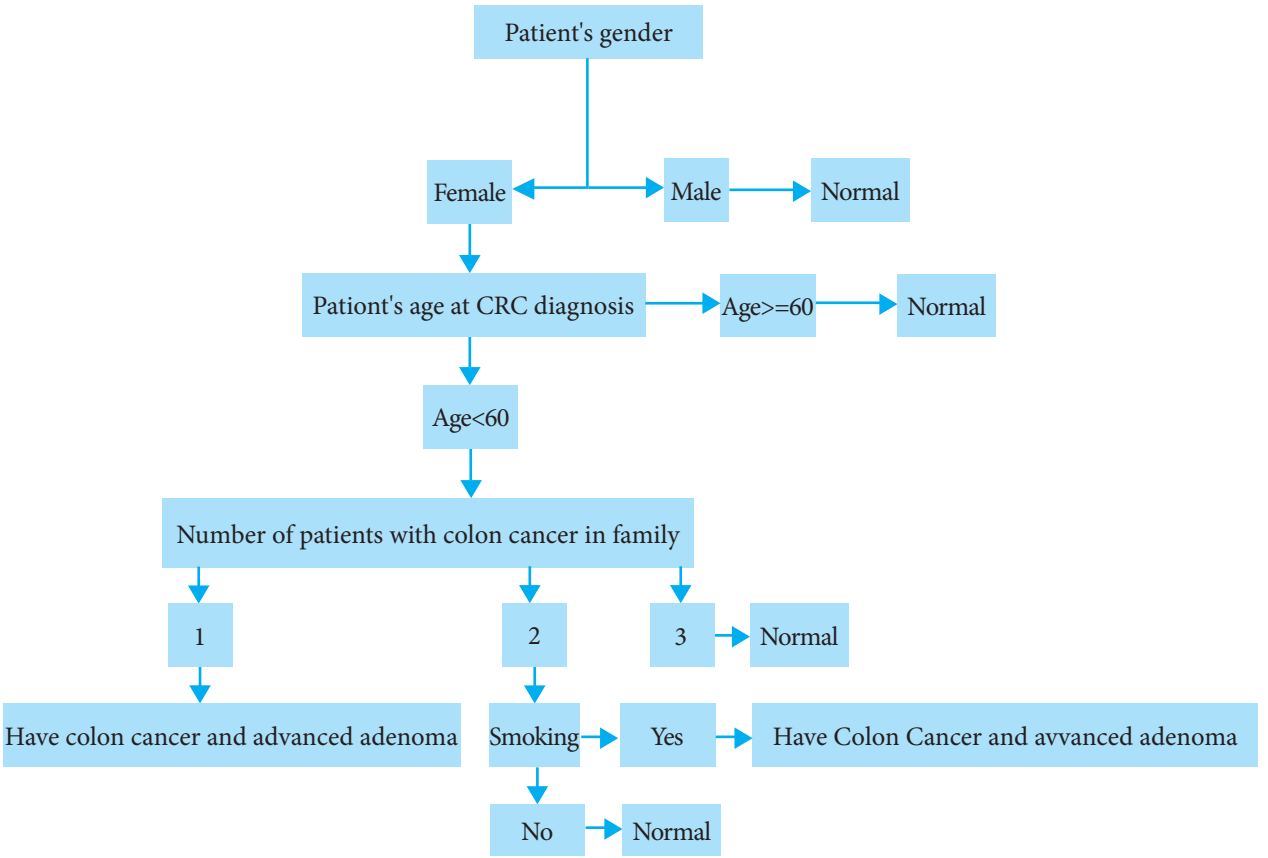

Fig.1: Decision tree of the fourth cluster of male participants

72% of the participants reported having one patient with CRC, 95% were older than 50 years old, 97% had a female patient with CRC in their family, 75% had a young patient with CRC in the family (under 60 years old at the time of diagnosis of CRC), 80% had ill sibling, and 75% was overweight. F-measure for decision tree in this cluster was 62.34%.

Overall, having a young patient with CRC in the family (under 60 years old at the time of diagnosis of CRC) was seen in two cluster. Having a young patient with CRC in the family and a female patient with CRC in the family of study participants concurrently were observed in three clusters. That is, having a young patient with CRC in the family was common in five out of the seven clusters and also having a female patient with CRC was recorded in four out of the seven clusters.

Figure 1 illustrates the decision tree implemented on the fourth cluster of male participants, as an example. We can see the impact of having a female patient with CRC and younger CRC patient, and having one or two patients with CRC in the family, and smoking on the prevalence of colorectal advanced neoplasms among screened participants. However, other features shown in these trees that were implemented on clusters such as smoking, high BMI (the BMI that was more than 25) and having more than two patients with CRC in the family were not common in most of the trees.

DISCUSSION

While it is vital to discover a wide-ranging of variables that are theoretically linked to CRC, it is also essential to identify a small number of critical variables that can be effectively addressed while advising patients, who have familial history, about CRC precaution. The current findings confirm the importance of some issues currently assessed and examined in the empirical literature.[20,21] Having more than one patient with CRC in family, having under 60 years old patient in family, [21,22] and female

sex of CRC patients in family are associated with the prevalence of advanced colonic neoplasms in participants. Smoking, also, was recognized as an important factor for advanced colonic neoplasms in male clusters and this result also admitted in study of Lee et al and Li et al. [23,24] Furthermore, obesity was cited by Liu et al. [25] as a risk factor and this study achieved reported this factor as well. The effect of obesity on the incidence of CRC was mentioned in study results of Bardou et al. [10] Many people who anticipated to have cancer were under 50 years old and it seems that age of presenting cancer is declining, and this result admitted in paper of Smith et al. [26] However, it has been shown in few clusters that old people are on the verge of cancer. People who are older than 50 are at the risk of this disease, which is also declared in paper of Smith et al. [26]

In addition, using data mining techniques showed more than 80% accuracy, which was so similar to other data mining studies [27,28] and it should be noticed that this measures are based on the data has been use. [29]

The novelty of this study is using data mining techniques in this area in Iran with high accuracy, and introducing the impact of patient's sex in family on cancer incidence. Research is needed to determine the impact of other factors like life style and behavioral factors, which also might be important on CRC. Therefore, studying nutrition, cancer screening test attendance, and physical activity are suggested for future studies. Moreover, additional research is required to determine the importance of having a second degree family members with CRC to understand the effect of having patient in wider range in family in Iran like what that was done in paper of Chau et al. [30] Finally, researchers should attempt to identify specific health education for precaution of this disease.

ETHICAL APPROVAL

There is nothing to be declared.

CONFLICT OF INTEREST

The authors declare no conflict of interest related to this work.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;**68**:394–424. doi: 10.3322/caac.21492.
2. Globocan 2018. Statistics on cancers in Iran [Internet]. International agency of cancer research. http://gco.iarc.fr/today/data/factsheets/populations/364-iran-islamic-republic-of-fact-sheets.pdf. 2019 Jan 1
3. Hoogendoorn M, Moons LMG, Numans ME, Sips RJ. Utilizing data mining for predictive modeling of colorectal cancer using electronic medical records. *Int Conf Brain Informatics Heal* 2014;**8609**:132–41. doi: 10.1007/978-3-319-09891-3_13.
4. Samadder NJ, Curtin K, Tuohy TMF, Rowe KG, Mineau GP, Smith KR, et al. Increased risk of colorectal neoplasia among family members of patients with colorectal cancer: a population-based study in Utah. *Gastroenterology* 2014;**147**:814–21. doi: 10.1053/j.gastro.2014.07.006.
5. Dominic OG, McGarrity T, Dignan M, Lengerich EJ. American College of Gastroenterology guidelines for colorectal cancer screening 2008. *Am J Gastroenterol* 2009;**104**:2626-7. doi: 10.1038/ajg.2009.419.
6. Sung JJ, Lau JY, Young GP, Sano Y, Chiu HM, Byeon JS, et al. Asia Pacific consensus recommendations for colorectal cancer screening. *Gut* 2008;**57**:1166–76. doi: 10.1136/gut.2007.146316.
7. Kai S, Lin G, Bing W. An integrated network motif based approach to identify colorectal cancer related genes. *34th Chinese Control Conference* 2015;**8573**–8. doi: 10.1109/ChiCC.2015.7260997.
8. Li T, Zheng C, Zhang L, Zhou Z, Li R. Exploring the Risk Dietary Factors for the Colorectal Cancer. *IEEE* 2015;**570**–3. doi: 10.1109/PIC.2015.7489912.
9. Costa E, Soares AL, de Sousa JP. Information, knowledge and collaboration management in the internationalisation of SMEs: A systematic literature review. *Int J Inf Manage* 2016;**36**:557–69. doi: 10.1016/j.ijinfomgt.2016.03.007.
10. Bardou M, Barkun AN, Martel M. Obesity and colorectal cancer. *Gut* 2013;**62**:933–47. doi: 10.1136/gutjnl-2013-304701.
11. Thompson VL, Lander S, Xu S, Shyu CR. Identifying key variables in African American adherence to colorectal cancer screening: the application of data mining. *BMC Public Health* 2014;**14**:1173. doi: 10.1186/1471-2458-14-1173.
12. Kotsiantis S, Kanellopoulos D. Association rules mining: A recent overview. *GESTS Int Trans Comput Sci Eng* 2006;**32**:71–82.
13. López-Abente G, García-Pérez J, Fernández-Navarro P, Boldo E, Ramis R. Colorectal cancer mortality and industrial pollution in Spain. *BMC Public Health* 2012;**12**:589. doi: 10.1186/1471-2458-12-589.
14. Chauhan D, Jaiswal V. An efficient data mining classification approach for detecting lung cancer disease. *Commun Electron Syst* 2016;**23**:1–8. doi: 10.1109/CESYS.2016.7889872.

15. Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data. *Data Knowl Eng* 2007;**63**:503–27. doi: 10.1016/j.datak.2007.03.016.

16. Brachman RJ, Anand T. The process of knowledge discovery in databases. *Advance Knowledge Discovery data Mining (KDD)*. 1996;**11**:37–57.

17. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. 3rd ed. *Burlington: Morgan Kaufmann*; 2016. 201-203(I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, 2000?????).

18. Zaki MJ, Meira Jr W. Data mining and analysis: fundamental concepts and algorithms. *Cambridge University Press* 2014;**404**-13. doi: 10.1017/CBO9780511810114.

19. Romero C, Ventura S. Educational data mining: A survey from 1995 to 2005. *Expert Syst Appl* 2007;**33**:135–46. doi: 10.1016/j.eswa.2006.04.005.

20. Slattery ML, Kerber RA. Family history of cancer and colon cancer risk: the Utah Population Database. *J Natl Cancer Inst* 1994;**86**:1618–26. doi:10.1093/jnci/86.21.1618.

21. Boardman LA, Morlan BW, Rabe KG, Petersen GM, Lindor NM, Nigon SK, et al. Colorectal cancer risks in relatives of young-onset cases: is risk the same across all first-degree relatives? *Clin Gastroenterol Hepatol* 2007;**5**:1195–8. doi: 10.1016/j.cgh.2007.06.001.

22. Moghimi-Dehkordi B, Safaee A, Zali MR. Prognostic factors in 1,138 Iranian colorectal cancer patients. *Int J Colorectal Dis* 2008;**23**:683–8. doi: 10.1007/s00384-008-0463-7.

23. Lee S, Woo H, Lee J, Oh J-H, Kim J, Shin A. Cigarette smoking, alcohol consumption, and risk of colorectal cancer in South Korea: *A case-control study. Alcohol* 2019;**76**:15–21. doi: 10.1016/j.alcohol.2018.06.004.

24. Li J, Adilmagambetov A, Jabbar MSM, Zaïane OR, Osornio-Vargas A, Wine O. On discovering co-location patterns in datasets: a case study of pollutants and child cancers. *Geoinformatica* 2016;**20**:651–92. doi: 10.1007/s10707-016-0254-1.

25. Liu PH, Wu K, Zauber AG, Fuchs CS, Ogino S, Chan AT, et al. 283-Obesity is Associated with an Increased Risk of Young-Onset Colorectal Cancer. *Gastroenterology* 2018;**154**:S-70-1. doi: 10.1016/S0016-5085(18)30690-5.

26. Smith RA, Andrews KS, Brooks D, Fedewa SA, Manassaram-Baptiste D, Saslow D, et al. Cancer screening in the United States, 2018: a review of current American Cancer Society guidelines and current issues in cancer screening. *CA Cancer J Clin* 2018;**68**:297–316. doi: 10.3322/caac.21446.

27. Yang H, Chen YP. Data mining in lung cancer pathologic staging diagnosis: correlation between clinical and pathology information. *Expert Syst Appl* 2015;**42**:6168–76. doi: 10.1016/j.eswa.2015.03.019.

28. Tseng WT, Chiang WF, Liu SY, Roan J, Lin CN. The application of data mining techniques to oral cancer prognosis. *J Med Syst* 2015;**39**:59. doi: 10.1007/s10916-015-0241-3.

29. Galdi P, Tagliaferri R. Data Mining: Accuracy and Error Measures for Classification and Prediction. in Reference Module in Life Sciences, no. January, Elsevier, 2018, pp. 1–14.

30. Chau R, Jenkins MA, Buchanan DD, Ait Ouakrim D, Giles GG, Casey G, et al. Determining the familial risk distribution of colorectal cancer: a data mining approach. *Fam Cancer* 2016;**15**:241–51. doi: 10.1007/s10689-015-9860-6.