Journal of Next-generation Convergence Information Services Technology Vol.9, No.3, September (2020), pp.283-290
ISSN 2384-101X(p) 2672-1163(e) *http://dx.doi.org/10.29056/jncist.2020.09.09*

Colon Polyps Risk Correlation Analysis Technology based on Clinical Data

Sung-Jong Eun[1]

Abstract

This paper proposes ways to draw elements having a high correlation to a risk of occurrence of colon polyp by analyzing life habit data and medical treatment data collected from physical examination center for the purpose of preventing colorectal cancer. A control group and a group with colon polyp are compared based on cross analysis and ANOVA to draw risk factors. Drawn risk factors include drinking habit, the average amount of drinking at one time per person, age of starting smoking, past duration of smoking, current duration of smoking, the average amount of smoking a day per person, whether family members were treated, and BMI. Validity of analysis of correlation can be verified by checking cases which P-value is less than 0.05. This paper is to propose service which provides life habit information record in physical examination center, cases of risk factors of colon polyp and colorectal cancer by using relevant analysis technique.

Keyword : colon polyp, colorectal cancer, analysis of variance, cross analysis, life habit

1. Introduction

Occurrence of colorectal cancer worldwide has increased sharply recently. Crude incidence rate of colorectal cancer in South Korea stood at 51.7 cases per every 100 thousand people in 2012 accounting for 12% of occurrence of all cancers. Occurrence of colorectal cancer among male is 15,612 annually ranking second in cancers among male while occurrence of colorectal cancer among female is 10,170 cases annually ranking third in cancers among female [1]. Compared with other cancers, colorectal cancer has clearer risk factors for occurrence and thus lots of studies focusing on prevention of colorectal cancer have been conducted. Since Jackman Mayo [2] published theory of 'adenoma-carcinoma sequence' which develops from adenomatous colon polyp into colorectal cancer in 1951, several studies have been made which presented an evidence of the development. In most cases, the fact that colorectal cancer develops from adenomatous colon polyp has been acknowledged even though in some cases colorectal cancer develops without taking a form of polyp [3]. A report states that in case

1 Computer Science, Gachon University, Incheon, Korea [Researcher]
e-mail: asclephios@naver.com

* This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2018-2-00861, Intelligent SW Technology Development for Medical Data Analysis)

Received(October 7, 2019), Review Result(1st: October 25, 2019, 2nd: November 15, 2019), Accepted(September 4, 2020), Published(September 30, 2020)

adenoma known as prodromal change is found in advance and removed, occurrence of colorectal cancer has decreased by 76~90%. As the report suggests, removal of adenomatous colon polyp is deemed to be important to prevent colorectal cancer. Grasping risk factors of colon polyp is needed to prevent colorectal cancer [4]. In general, for adenomatous colon polyp, risk factors of colorectal cancer includes obesity, age, smoking and drinking. Specifically, central obesity, impaired glucose tolerance, hypertension, lack of high density lipoprotein and hypertriglyceridemia are reported to have a high correlation to occurrence of colorectal cancer [5][6]. Among them, abdominal obesity reflects degree of visceral fat and is highly correlated to formation of colon polyp [7][8]. This paper is to draw elements having a high correlation to a risk of occurrence of colon polyp through multivariate regression analysis with elements having high priority as input, cross analysis and ANOVA to find risk factors of colon polyp for the purpose of predicting a risk of occurrence of colon polyp.

2. Related Work

2.1 Subjects and duration

The subjects of this study are fifteen hundred (1,500) people who had checkups at comprehensive examination center in a hospital from 2015 to 2019. This study is to extract elements (fields) having a high correlation to risk of colon polyp by comparing a difference between patient group and non-patient group at a rate of fifty (50) percent (%).

2.2 Research methods

2.2.1 Basic data of subjects

Basic data used in this study have been extracted from comprehensive examination center in a hospital and treatment EMR DB and consist of general information and detailed information. General information as defined in the following [Table 1] includes patient number, date of birth, gender, treatment date, stature, weight and waist measure. Additional extracted query content is marked along with general information to extract data of relevant field.

Detailed information consists of sixty five columns and fields. Detailed information is primarily composed of information for predicting and preventing colon polyp. Detailed information is mainly comprised of information on drinking habit, whether to smoke, disease history(hypertension, cardiac disorder, stroke, diabetes mellitus, hyperlipidemia, thyroid gland, kidney disease, pulmonary tuberculosis,

asthma, cancer etc.), drug-taking, family history, exercise and endoscopic examination result. Relevant data were extracted from hospital EMR data. [Table 2] shows data definition sample.

[Table 1] Data Definition of General Information

Section 1: Common			
ID	**Column**	**Extract Query**	**Comment**
1	Patient ID		
2	Date of birth	SELECT substring(ptnt_prsn_no,1,4) FROM h1ptntinfo WHERE ptnt_no = 21234533	
3	Gender	SELECT CASE substring(ptnt_prsn_no,9,1) WHEN '1' THEN 'm' WHEN '3' THEN 'm' WHEN '2' THEN 'f' WHEN '4' THEN 'f' ELSE NULL END FROM h1ptntinfo WHERE ptnt_no = 21234533	
4	Date of clinic		
5	Height	select recept_no from clinic_reservation where ptnt_no=12345678 select res_value from [dbo].[Clinic_Result02] where recept_no='2025141722' and gum_code='TP01'	
6	Weight	select res_value from [dbo].[Clinic_Result02] where recept_no='2025141722' and gum_code='TP02'	

[Table 2] Data Definition Sample of Detail Information

Section 2: Common			
ID	**Column**	**Extract Query**	**Comment**
1	Pulmonary tuberculosis	SELECT ptnt_no, diag_cd, 8, clinic_ymd FROM h2ptnt_diag WHERE ptnt_no IN (ptnt_no list) AND diag_cd IN (SELECT diag_cd FROM h2diag WHERE diag_hnm LIKE '%Pulmonary tuberculosis%')	
2	Asthma	SELECT ptnt_no, diag_cd, 9, clinic_ymd FROM h2ptnt_diag WHERE ptnt_no IN (ptnt_no list) AND diag_cd IN (SELECT diag_cd FROM h2diag WHERE diag_hnm LIKE '%Asthma%')	
3	Cancer	SELECT ptnt_no, diag_cd, 10, clinic_ymd FROM h2ptnt_diag WHERE ptnt_no IN (ptnt_no list) AND diag_cd IN (SELECT diag_cd FROM h2diag WHERE (diag_hnm LIKE '%Cancer%' AND diag_hnm NOT LIKE '%Malignant%') OR diag_hnm LIKE '%Malignant%')	

2.2.2 Data processing and statistical analysis

This study used Matlab2016b version program to analyze statistics. Frequency analysis was used for general characteristics. Cross analysis and ANOVA were conducted to analyze correlation to risk of

colon polyp. Elements having a high correlation to risk of colon polyp through multivariate logistic regression analysis [9-12] using statistically significant variables. Statistical significance was set at $p<0.05$.

3. Study Result

3.1 Logistic regression analysis with colon polyp as dependent variable

This study built logistic regression analysis model with significant variable from univariate analysis as independent variable to conduct analysis. Colon polyp was set as dependent variable. When analysis was conducted based on information on drinking habit, the average amount of drinking at one time per person, age of starting smoking, past duration of smoking, current duration of smoking, the average amount of smoking a day per person, whether family members were treated, BMI and waist measure, it was found that they have a high drinking correlation to occurrence of colorectal cancer. Each P-value is .001, for drinking habits.005, smoking related .001~030, family history .001, BMI .008, waist circumference .003. The results showed that all the items were highly correlated. Among them, the items related to drinking, smoking, and family cancer treatment were the highest. This proves that drinking, smoking and family history are important factors, as discussed in existing relevant clinical studies. [Fig. 1] and [Table 3] show a result.

[Table 3] The Final Result of P-value

Data field	P-value
Drinking habit (number of times)	.001
The average amount of drinking at one time per person	.005
Age of starting smoking	.002
Past duration of smoking	.001
Current duration of smoking	.030
The average amount of smoking a day per person	.005
Whether family members were treated	.001
BMI	.008
Waist measure	.003

The analysis showed that drinking habit, the average amount of drinking at one time per person, age of starting smoking, past duration of smoking, current duration of smoking, the average amount of

smoking a day per person, whether family members were treated, BMI and waist measure among data are highly correlated to occurrence of colon polyp. In [Fig. 1], fields with high correlation were marked in red. [Table 3] shows items whose p-value is less than P-value.

person_id	biono	birth_year	gender	clinic_date	heihgt	weight	bmi	음주여부	음주습관	1회음주량	음주 주종	음주 병	음주 잔	흡연상태	과거 흡연:담배 시작	금연 기간	흡연기간	일흡연량	암 가족력
1624		1956	m	20140929	NULL	NULL	0	3		0		0	0	3	0		0	0	0
1625		1973	m	20120426	175.8	74	23.94385	0		0		0	0	3	0		0	0	0
1626		1958	m	20150429	169.3	74.3	25.92238	3		0		0	0	2	0		0	0	0
1627		1942	m	20121119	165.6	44.9	16.3729	0		0		0	0	1	0		0	0	1
1628		1964	m	20161220	172.2	72.7	24.51704	0		0		0	0	1	0		0	0	0
1629		1964	f	20140819	157.7	57.7	23.20131	0		0		0	0	1	0		0	0	0
1630		1933	m	20111209	166.5	63.4	22.86972	3	2	4	4	1	0	1	0		0	0	0
1631		1958	m	20131218	170.9	73.7	25.23384	3	1	5	2	0	5	3	0		0	0	0
1632		1958	m	20141210	174	93.4	30.84952	1		0		0	0	1	0		0	0	0
1633		1965	m	20120724	176.2	82.1	26.44426	3	3	21	2	3	0	3	0		0	0	0
1634		1958	f	20170217	153.2	53.5	22.79482	0		0		0	0	1	0		0	0	0
1635		1965	m	20140915	166.2	67	24.25564	3	2	0	2	0	0	2	0		0	0	0
1635		1965	m	20160620	166.9	65.7	23.58591	3	2	14	2	2	0	2	0	10	0	10	0
1636		1961	m	20160310	168.2	66.3	23.43482	3	5	7	2	1	0	1	0		0	0	0
1637		1963	f	20160513	160.3	50.3	19.57496	3		0		0	0	3	0		0	0	0
1639		1959	f	20121204	153.6	65.3	27.67775	1		0		0	0	1	0		0	0	0
1640		1961	m	20091021	165.1	56.3	20.65448	3		0		0	0	3	0		0	0	0
1641		1962	m	20151120	170.3	70.1	24.17067	3	1	1	1	0	1	2	0	10	0	10	0
1642		1963	m	20120918	165.3	71.7	26.24058	3	1	7	2	1	0	3	0		0	0	0
1643		1976	f	20170214	158.3	48.5	19.35441	1		1	1	0	1	1	0		0	0	0
1644		1959	f	20131227	155.1	50.6	21.03424	1		2	2	0	2	1	0		0	0	1

[Fig. 1] Results of Correlation Analysis about Colorectal Polyp Risk

4. Study and conclusion

This paper proposes ways to draw elements having a high correlation to occurrence of occurrence of colon polyp by analyzing life habit data collected from treatment and checkup center for the purpose of preventing colorectal cancer. Based on 1,500 data extracted from hospital checkup center and EMR treatment DB, risk factors of colon polyp were drawn by comparing general control group and patient

group with colon polyp by conducting cross analysis and ANOVA. Drawn risk factors include drinking habit, the average amount of drinking at one time per person, age of starting smoking, past duration of smoking, current duration of smoking, the average amount of smoking a day per person, whether family members were treated, BMI, and waist measure. Validity of correlation analysis technique was verified by cases whose P-value was less than 0.05. We will develop a customized system of predicting colorectal cancer by making good use of information on life habit and treatment data collected from checkup center based on relevant analysis technique.

References

[1] Y. H. Park, J. H. Yu, T. Y Lee, "An Analysis on Risk Factors of Colon Polyps with Health Examination Examinees", Journal of the Korea Academia-Industrial cooperation Society, vol. 15, no. 3, March 2014, pp. 1641-1649, doi: 10.5762/KAIS.2014.15.3.1641.

[2] M. E. Martínez, R. Sampliner, J. R. Marshall, A. K. Bhattacharyya, M. E. Reid, D. S. Alberts, "Adenoma characteristics as risk factors for recurrence of advanced adenomas", Gastroenterology, vol. 120, no. 5, April 2001, pp. 1077-1083, doi: 10.1053/gast.2001.23247.

[3] J. V. Selby, G. D. Friedman, C. P. Quesenberry Jr, N. S. Weiss, "A case-control study of screening sigmoidoscopy and mortality from colorectal cancer", New England Journal of Medicine, vol. 326, no. 10, March 1992, pp. 653-657, doi: 10.1056/NEJM199203053261001.

[4] L. M. Morimoto, P. A. Newcomb, C. M. Ulrich, R. M. Bostick, C. J. Lais, and J. D. Potter, "Risk factors for hyper-plastic and adenomatous polyps: evidence for malignant potential?", Cancer Epidemiol Biomarkers Prev, vol. 11, no. 10, October 2002, pp. 1012-1018.

[5] T. Morita, S. Tabata, M. Mineshita, T. Mizoue, M. A. Moore, S. Kono, "The metabolic syndrome is associated with increased risk of colorectal adenoma development: the Self-Defense Forces health study", Asian Pacific journal of cancer prevention, vol. 6, no. 4, October, 2005, pp. 485-489.

[6] M. C. Kim, D. H. Kim, T. H. Jeong, "Risk Factors of Colorectal Polyps in Korean Adult", Journal of the Korean Academy of Family Medicine, vol. 23, no. 7, July 2002, pp. 890-896.

[7] E. Giovannucci, A. Ascherio, E. B. Rimm, G. A Colditz, M. J Stampfer, W. C Willett, "Physical activity, obesity, and risk for colon cancer and adenoma in men", Annals of internal medicine, vol. 122, no. 5, March 1995, pp. 327-334, doi: 10.7326/0003-4819-122-5-199503010-00002.

[8] Y. H. Park, J. H. Yu, T. Y. Lee, "An Analysis on Risk Factors of Colon Polyps with Health Examination Examinees", Journal of the Korea Academia-Industrial cooperation Society, vol. 15, no.3, March 2014, pp. 1641-1649, doi: 10.5762/KAIS.2014.15.3.1641.

[9] C. Y. Joanne Peng, K. L. Lee, G. M. Ingersoll, "An introduction to logistic regression analysis and reporting", The journal of educational research, vol. 96, no. 1, April 2010, pp. 3-14, doi: 10.1080/00220670209598786.

[10] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, A. R. Feinstein, "A simulation study of the number of

events per variable in logistic regression analysis", Journal of clinical epidemiology, vol. 49, no. 12, December 1996, pp. 1373-1379, doi: 10.1016/s0895-4356(96)00236-3.

[11] E. W. Steyerberg, F. E. Harrell Jr, G. J. Borsboom, M. J. Eijkemans, Y. Vergouwe, J. D. Habbema, "Internal validation of predictive models: efficiency of some procedures for logistic regression analysis", Journal of clinical epidemiology, vol. 54, no. 8, August 2001, pp. 774-781, doi: 10.1016/S0895-4356(01)00341-9.

[12] S. Sperandei, "Understanding logistic regression analysis", Biochemia medica: Biochemia medica, vol. 24, no. 1, February 2014, pp. 12-18, doi: 10.11613/BM.2014.003.