Association between Colon Polyp and Colorectal Cancer using Text-Mining Analysis

Seong-Mi Moon

Graduate School of Public Health
Yonsei University
Department of Epidemiology
and Health Promotion
Division of Epidemiology

Association between Colon Polyp and Colorectal Cancer using Text-Mining Analysis

A Masters Thesis Submitted to the Department of Epidemiology and Health Promotion, Division of Epidemiology and the Graduate School of Yonsei University in partial fulfillment of the requirements for the degree of Master of Public Health

Seong-Mi Moon

December 2022

This certifies that the Masters Thesis
of Seong-Mi Moon is approved.

_____

Thesis Supervisor: Sun Ha Jee

_____

Thesis Committee Member: Tae Il Kim

_____

Thesis Committee Member: Keum Ji Jung

Graduate School of Public Health
Yonsei University
December 2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Association between Colon Polyp and Colorectal Cancer using Text-Mining Analysis

Seong-Mi Moon
Graduate School
of Public Health
Yonsei University

(Directed by Professor Sun Ha Jee, Ph.D.)

INTRODUCTION

Colorectal cancer (CRC) is the one of the major healthcare problem worldwide with increasing incidence and mortality. To reduce the incidence and mortality of CRC, it is important to detect and remove colon polyp early through colonoscopy. Although there are many studies on the association between colon polyp and CRC based on colonoscopic findings, previous studies had limitations of sample size or information. Therefore, this study aims to identify the association between colon polyp and CRC using text-mining (TM) analysis.

METHODS

We conducted a cross-sectional study using health screening examination data in Korea Medical Institute (KMI). We included all participants who underwent health screening examination with colonoscopy between 2008 and 2019 (N=360,753). We categorized participants in no polyp group, low-risk polyp group, and high-risk polyp group. Information on colon polyp and CRC was extracted by colonoscopic findings using text-mining analysis. To identify factors related to prevalent colon polyp, we conducted multivariable logistic regression adjusted demography, lifestyle, and health screening examination. Then we estimated adjusted prevalence of CRC in each group to estimate the association between colon polyp and CRC. Sensitivity analysis regarding age and sex was also performed.

RESULTS

Among 360,753 participants, 63.0% did not have colon polyp, 33.4% had low-risk polyp, and 3.5% had high-risk polyp. The adjusted odds ratio for prevalent high-risk polyp in female, former smoker, current smoker and alcohol intake were 0.71 (0.66-0.75), 1.39 (1.30-1.48), 2.89 (2.71-3.07), and 1.37 (1.31-1.44), respectively. Similar results were found in sensitivity analysis. Each adjusted prevalence for CRC in no polyp group, low-risk polyp group, and high-risk polyp group was 0.06 (95% CI 0.05-0.07), 0.09 (0.07-0.10), and 0.13 (0.08-0.17), respectively. Adjusted prevalence ratios were 1.22 (0.91-1.53) and 1.49 (0.73-2.25) in low-risk polyp group and high-risk polyp group compared with no polyp group.

CONCLUSION

In our study, factors related to prevalent colon polyp were sex, age, Body Mass Index (BMI), smoking status, alcohol intake and regular physical activities. Our study suggests high-risk polyp is associated with CRC. We suggest male aged 50 or above needs to conduct colonoscopy to prevent CRC. Our results may be used to provide evidence for healthcare policies.

Key words: Colonoscopy, Colon polyp, Colorectal cancer, Health screening examination, Cross-sectional study, Text-mining

Ⅰ. INTRODUCTION

1. Background

Colorectal cancer (CRC) is a major healthcare problem in the world. The global burden of CRC has increased rapidly in recent years. The global burden of CRC more than doubled in 2019 compared to 1990 (GBD 2019 Colorectal Cancer Collaborators, 2022).

The incidence of CRC was estimated to be 1.93 million, and it was the third most diagnosed cancer worldwide in 2020. Also, the mortality of CRC was estimated to be 0.94 million, and it was the second most deaths caused by cancer worldwide in 2020 (Xi Y & Xu P, 2021).

Similar to global trends, the incidence and mortality of CRC are high in Korea as well. A total of 254,718 new cases of CRC was diagnosed in 2019, and it was the fifth most diagnosed cancer in Korea. A total of 81,203 deaths was caused by CRC in 2019, and it was the fourth most deaths caused by cancer in Korea (Kang MJ et al., 2022).

CRC is classified into colon cancer and rectal cancer, according to the location of the malignant tumor. Colon cancer has a higher incidence than rectal cancer. The incidence of colon cancer was 1,15 million, while the incidence of rectal cancer, including anus cancer, was 0.78 million in 2020. Also, colon cancer has a higher mortality than rectal cancer. The mortality of colon cancer was 0.58 million, while the mortality of rectal cancer, including anus cancer, was 0.36 million in 2020 (Sung H et al., 2021).

CRC is a malignant tumor that arises in the large intestine due to the progression of hereditary or acquired pre-malignant lesions. It develops from interactions among modified or unmodified risk factors (Conteduca V et al., 2013). 30% of CRC was reported heredity CRC and the other 70% of CRC was reported as sporadic CRC (Brosens LA et al., 2015).

The sporadic CRC is generated through 2 pathways: adenoma-carcinoma sequence, serrated pathway (Keum N & Giovannucci E, 2019).

The traditional pathway is called the adenoma-carcinoma sequence, which accounts for 85~90% of sporadic CRC. The adenoma-carcinoma sequence was first described by Molson as the mechanism by which adenoma becomes adenocarcinoma (Morson BC, 1974). Adenoma develops in the normal epithelial cells of the large intestine due to genetic mutation and it develops into CRC after 10~20 years. During this process, adenoma acquired the characteristics of CRC such as uncontrolled growth, invasion and destruction of surrounding tissues (Sillars-Hardebol AH at al., 2012). Not all adenomas become CRC, and only 5% of adenoma become CRC (Brenner H et al., 2007). People who have 3 or more adenoma, an adenoma larger than 1cm, have an adenoma with villous feature or high-grade dysplasia (Winawer SJ et al., 2006).

The new pathway is called the serrated pathway, which accounts for 10~15% of sporadic CRC. Sessile serrated polyp (SSP) and sessile serrated adenoma (SSA) were previously known not to be CRC. However, many studies have found that SSP and SSA can be CRC by change in genetic mutation (Sano W et al., 2020). SSP and SSA were reclassified as sessile serrated lesions (SSL) in 2019 (Nagtegaal ID et al., 2020).

To reduce the incidence and mortality of CRC, it is important to detect and remove the polyp early through colonoscopy. The incidence of CRC was higher in the polyp-detected group in the health screening examination compared to the population who did not have health screening examination. Among polyp-detected groups, high-risk polyp group had a higher mortality than low-risk polyp group (Zauber AG et al., 2012)

To detect colon polyp and CRC, colonoscopy, fecal occult blood test (FOBT), including guaiac-fecal occult blood test (G-FOBT) and fecal immunochemical test (FIT), and flexible sigmoidoscopy are mainly performed (Bond JH, 1999). All of the above health screening examinations are effective in preventing the incidence and mortality of CRC by observing colon polyp (Fitzpatrick-Lewis D et al., 2016).

Among theses, colonoscopy is considered the gold standard of screening examination for CRC. Colonoscopy has higher specificity and sensitivity than other examinations, so the detection rates for colon polyp and CRC are high (Quintero E et al., 2012). Also, it has a great advantage in preventing CRC because it can be removed colon polyp immediately (Brenner H et al., 2014).

However, colonoscopy is performed only in high-risk group because of some disadvantages. Patients who have scheduled the colonoscopy need to effort to empty the intestine, and there are restrictions on daily life for several hours after the examination (Wilkins T et al., 2018). There is a large difference in the detection rate depending on the skill of the medical staff, and the procedure by the inexperienced medical staff rather causes side effects (Ishaq S et al., 2017).

Because of these disadvantages, colonoscopy is selectively performed only in high-risk group in most countries (Navarro M et al., 2017).

Most European member states have established screening programs for CRC based on strong evidence from the Council of the Europeans published in 2003 (Cardoso R et al., 2021). The screening program for CRC targets all people between the ages of 50 and 74. A number of European countries initially conducted FOBT, and if the result from FOBT was positive, they performed colonoscopy (Altobelli E et al., 2014).

The Screening program for CRC is also available in Korea. It was introduced in 2004 and became the fifth national screening program for cancer after gastric, breast, cervical and liver cancer. All people who have the National Health Insurance Service (NHIS) aged 50 or above perform FIT. And if the result from FIT was positive, they recommended performing a colonoscopy (Park B et al., 2021)

There are many studies on the association between colon polyp and CRC based on colonoscopy (Duvvuri A et al., 2021). Previous studies had some limitations. Most studies had detailed information but small sample sizes, while a few studies had large sample sizes, but limited information (Bjerrum A et al., 2020; Chiu HM at al., 2014; Chung SJ et al., 2011; Cotte V et al., 2012; Laish I et al., 2015; Lieberman D et al., 2020).

In order to supplement the limitations of previous studies, we included a sufficient number of participants using a health screening examination database performed at multiple check-up centers. Also, we extracted detailed information about colon polyps through text-mining of colonoscopy report.

The Colonoscopy report was written in free-text format by medical staff. It includes demographics and history of disease assessment of risk and comorbidity, procedure indication, procedure, colonoscopic findings, assessment, interventions or unplanned events, follow-up plan and etc (Lieberman D et al., 2007). Among these, detailed information about the colon polyp of the large intestine such as type, location and size was described in the colonoscopic findings.

To extract the detailed information about colon polyp, we used text-mining for colonoscopic findings. text-mining is one of the methods to analyze large unstructured data. It is a process for the extraction of interesting information, where an unstructured text is the source. It includes document clustering and classification, information extraction, information retrieval (IR), name entity recognition (NER), natural language processing (NLP), question-answering (QA), visualization (de Bruijn and B, Martin J, 2002).

However, text-mining is a difficult method for non-experts to use. Therefore, we used a method of cutting the text and extracting the desired words using statistical software. We conducted a study on the association between colon polyp and CRC using structured data and unstructured data from which colonoscopic findings were extracted.

2. Objectives

The purpose of this study was to investigate the association between colon polyp and CRC. The specific objectives of this study are as follows (**Figure 1**).

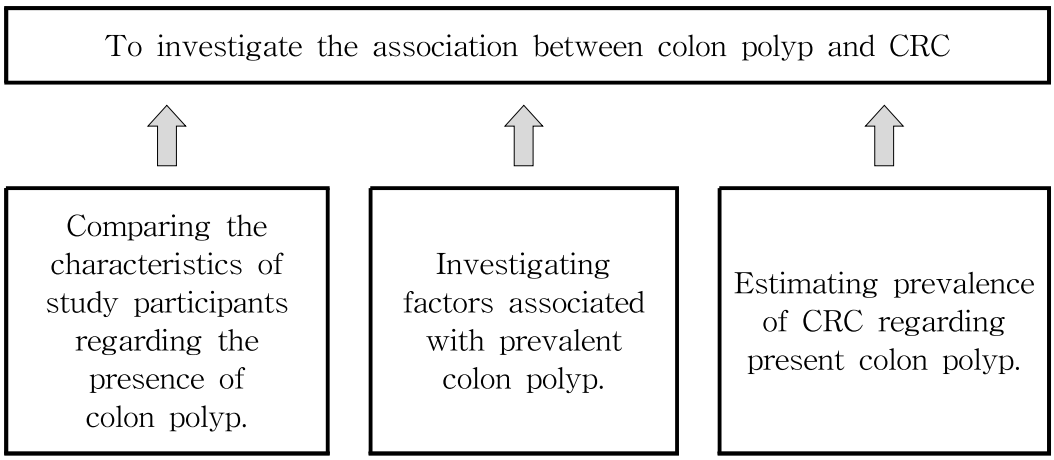

Figure 1. Objective of the study.

Ⅱ. METHODS

1. Study design & Settings

This study is a cross-sectional study using the database of participants who underwent health screening examination, including colonoscopy at Korea Medical Institute (KMI) between 2008 and 2019.

KMI is a representative health check-up center in Korea. KMI was established at Gwanghwamun, Seoul in 1985, and it has 7 health check-up centers across the country, including Seoul, Suwon, Daegu, Busan and Gwangju. KMI provides a variety of comprehensive health screening examination programs for the health of Koreans.

All the health screening examination programs at KMI include basic inspections, urine tests, blood tests, electrocardiography (ECG), chest X-ray, abdominal ultrasonography and gastrointestinal graphy. Additional tests such as magnetic resonance imaging (MRI), computed tomography (CT), carotid ultrasonography, pulmonary function test (PFT) were performed according to the selected health screening examination program.

To investigate the association between colon polyp and CRC, we used the KMI database of participants who underwent colonoscopy from 2008 to 2019. The framework of this study is as follows (**Figure 2**).

This study was written according to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline (Cuschieri S, 2019).

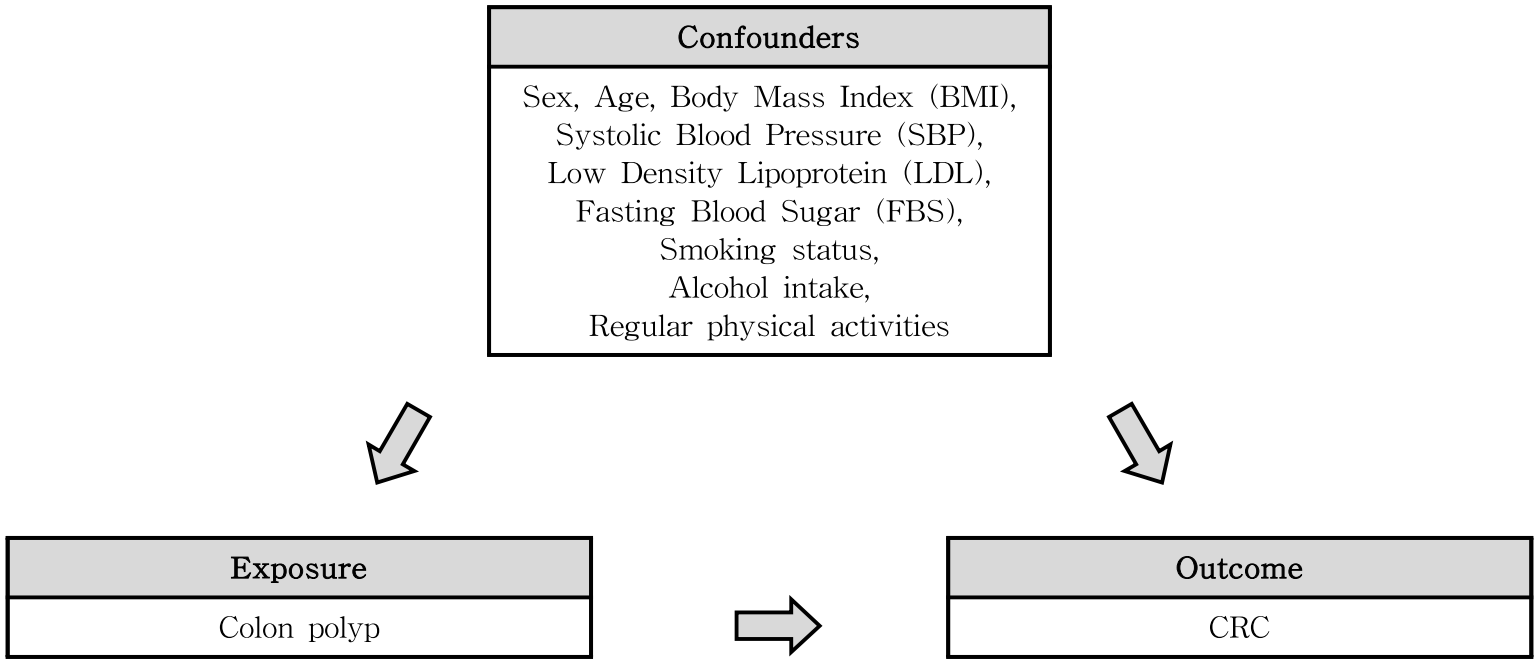

Figure 2. Framework of the study.

2. Data sources

The KMI database consists of 6 datasets. Demography and lifestyle were obtained using the questionnaires. And vital sign, laboratory, anthropometry and reports were obtained from the results of examinations. All data in the database are as follows (**Table 1**).

Table 1. Data elements in the KMI database.

No	Dataset	Variable	Source
1	Demography	Sex, Age	Questionnaire
2	Lifestyle	Smoking status, Pack-year of smoking Alcohol intake, Amount of alcohol, Walking, Moderate physical activity, Vigorous physical activity	Questionnaire
3	Vital sign	Systolic blood pressure (SBP), Diastolic blood pressure (DBP)	Examination
4	Laboratory	High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Thyroglobulin (TG), Fasting Blood Sugar (FBS)	Examination
5	Anthropometry	Height, Weight, Waist circumference	Examination
6	Reports	Date of colonoscopy Colonoscopic finding	Examination

3. Study population

We included participants who underwent health screening examination, including colonoscopy at KMI between 2008 and 2019 (N = 424,887). We excluded participants who had unknown data, such as "see the detailed opinion" (N = 2,058) and who failed to exam for any reasons (N = 51,919) and who had missing data on age or sex (N = 11,212). The final sample size was 360,753 (**Figure 3**).

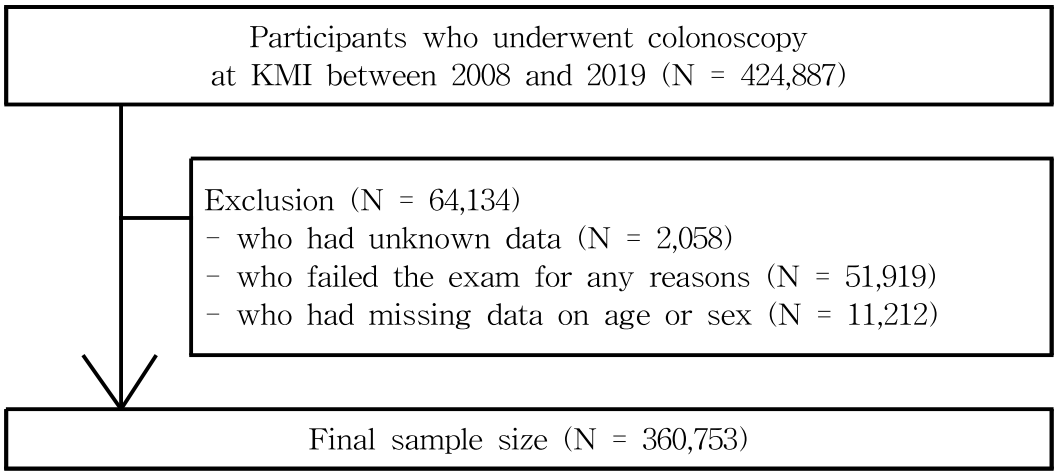

Figure 3. Flowchart of study population.

The Institutional Review Board of Yonsei Severance hospital approved this study and we only used de-identified data (IRB No. 4-2011-0444).

4. Data collection

In this study, we used colonoscopic findings to obtain information about colon polyps with text-mining because colonoscopic findings were unstructured data with free-text format. All information about prevalent colon polyp, including type, location and size, was obtained with the following methods: (1) screening data (2) selecting words (3) extracting words (4) reviewing data (5) re-extracting words.

First, we converted the lowercase alphabet to the uppercase alphabet for the convenience of word extraction. We divided colonoscopic findings into words based on delimiters, and calculated the frequency of words. The most common words are as follows (**Table 2, Figure 4**).

Table 2. Frequency of words in colonoscopic findings.

No	Word	N (%)	No	Word	N (%)
1	정상	245,545 (19.41)	10	부근	14,961 (1.18)
2	대장용종	84,085 (6.65)	11	S상결장	14,661 (1.16)
3	용종	62,543 (4.95)	12	대장게실	14,541 (1.15)
4	대장	48,432 (3.83)	13	발견되어	13,465 (1.06)
5	게실	27,719 (2.19)	14	3MM	13,283 (1.05)
6	상행결장	23,879 (1.89)	15	불량	13,195 (1.04)
7	변찌꺼기	17,476 (1.38)	16	다소불량	12,908 (1.02)
8	직장	15,866 (1.25)	17	0.3CM	12,655 (1)
9	조직검사	15,132 (1.2)			

Figure 4. Word cloud plot in colonoscopic findings.

Second, we excluded data in which the following words appeared from all analyses (**Table 3**).

Table 3. List of Excluded words.

Value	Word
Failure	"불가", "실패", "중단", "거부", "미실시", "미검진", "미촬영", "검사연기", "못함", "안됨", "제한적검사", "-까지관찰", "일부관찰", "정결", "찌꺼기", "잔변", "청결", "불량", "불충분", "불완전", "미흡", "좋지않-", "진입", "과잉행동", "움직임", "협조"
Un-known	"세부소견참조", "종합소견참조", "수검함", "검진함", "검사함", "-인한재검", "재검요망", "대장재검", "재검", "추후", "전원", "TRANSFER", "권유"

5. Measurements

To investigate the association between colon polyp and CRC, We developed a directed acyclic graph (DAG) through literature reviews (Chi Z et al., 2021; He X et at., 2018; Sninsky JA et al., 2022) (**Figure 5**).

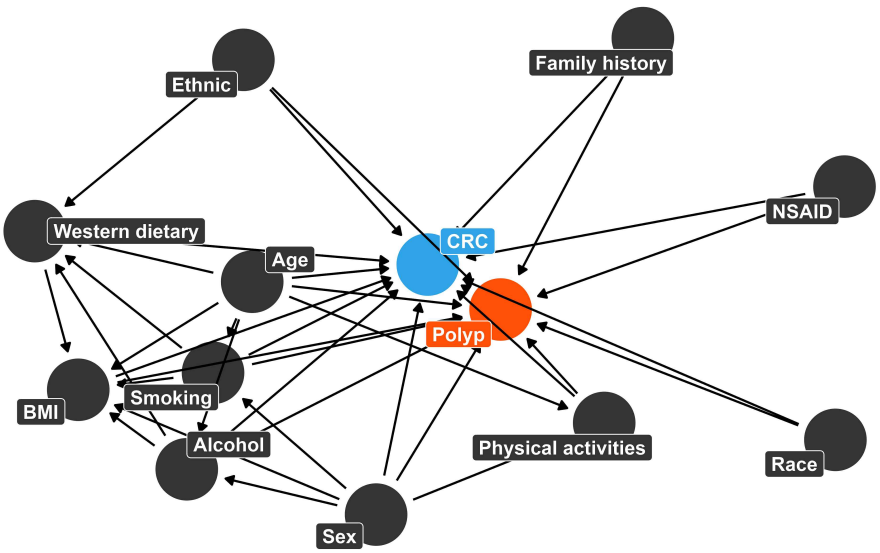

Figure 5. DAG between colon polyp and CRC.

According to the DAG, we defined variables as exposure, confounder, and outcome, and created variables in the KMI database. If the participant underwent health screening examination 2 or more, We integrated all the records of each participant. Continuous variables such as laboratory were averaged. If the variable included missing value, we imputed the mean value of study participants and categorized them as "unknown". The final variables are as follows (**Table 4**).

Table 4. Measurements in the study.

Measurement	Variable	Type
Exposure	Colon polyp	1 = No polyp, 2 = Low-risk polyp, 3 = High-risk polyp
Outcome	CRC	0 = No, 1 = Yes
Confounder	Sex	1 = Male, 2 = Female
	Age	Continuous
	Height	Continuous
	Weight	Continuous
	Waist circumference	Continuous
	BMI	Continuous
	SBP	Continuous
	DBP	Continuous
	HDL	Continuous
	LDL	Continuous
	TG	Continuous
	FBS	Continuous
	Smoking statue	1 = Never, 2 = Former, 3 = Current
	Alcohol intake	0 = No, 1 = Yes
	Walking	Continuous
	Moderate physical activity	Continuous
	Vigorous physical activity	Continuous
	Regular physical activities	0 = No, 1 = Yes
	Colonoscopic finding	Text

5.1. Exposure

The exposure of this study was the presence of colon polyp. Colon polyp was defined as the prevalent colon polyp such as polyp, adenoma, SSL related words in colonoscopic findings. Colon polyp related words are as follows (**Table 5**).

Table 5. List of colon polyp related words.

Value	Word
Polyp	"용종", "폴립", "POLYP"
Adenoma	"선종", "샘종", "관상", "융모", "아데노마", "ADENOMA", "TUBULAR", "VILLOUS", "TUBULOVILLOUS"
SSL	"무경성", "SESSILE", "톱니", "거치상", "SERRATED"

We divided our study participants into 3 groups: no polyp, low-risk polyp and high-risk polyp. No polyp was defined as the absence of any colon polyp related words in colonoscopic findings. Participants with colon polyp were categorized into low-risk polyp group and high-risk polyp group. High-risk polyp group was defined if participants had 3 or more colon polyps, larger than 1cm colon polyp based on Korean guidelines for post-polypectomy colonoscopic surveillance (Hong SN et al., 2012), Otherwise, low-risk polyp group was defined as the rest of the participants with colon polyp, excluding the high-risk polyp group (**Table 6**).

Table 6. Classification of colon polyp group.

Classification	Definition
No polyp	absence of any colon polyp
High-risk polyp	larger than 1cm or 3 or more adenoma larger than 1cm or 3 or more SSL
Low-risk polyp	Other polyp expect for high-risk polyp

We additionally extracted the location and size of polyp. And we re-classification The words for the location of the colon polyp are as follows (**Table 7**).

Table 7. List of words for location.

Value	Word
Cecum	"맹장", "충수", "CECUM", "APPENDIX",
Ascending colon	"상행", "오름", "ASCENDING"
Hepatic flexure	“간만곡”, “HEPATIC”
Transverse colon	"횡행", "가로", "TRANSVERSE"
Splenic flexure	“비만곡”, “비장만곡”, “SPLENIC”
Descending colon	"하행", "내림", "DESCENDING"
Sigmoid colon	"S상", “S결장”, "에스", “구불”, "SIGMOID"
Rectum	“직장”, “RECTUM”
Anus	“항문”, “ANUS”

And we re-classified location into proximal colon and distal colon relative to the splenic flexure based on the U.S. Preventive Services Task Force guideline (Lin JS et al., 2016) (**Table 8**).

Table 8. Re-classification of location.

Classification	Value
Proximal colon	Cecum, Ascending colon, Hepatic flexure,Transverse colon
Distal colon	Splenic flexure, Descending colon, Sigmoid colon, Rectum, Anus

5.2. Outcome

The main outcome of this study was CRC. CRC was defined as the prevalent CRC related words in colonoscopic findings. CRC related words are as follows (**Table 9**).

Table 9. List of CRC related words.

Value	Word
CRC	"선암종", "샘암종", "흑색종", "선암", "샘암", "대장암", "직장암", "결장암", "항문암", "아데노칼시노마", "아데노카르시노마", "ADENOCARCINOMA", "COLORECTALCANCER", "COLONCANCER", "RECTALCANCER", "MELANOMA"

5.3. Confounders

We considered potential confounders related to prevalent colon polyp and CRC as sex, age at colonoscopy, BMI, waist circumference, SBP, DBP, HDL, LDL, TG, FBS, smoking status, alcohol intake, and regular physical activities.

We calculated BMI as weight in kilograms divided by height in meters squared (kg/m^2). Smoking status was categorized into 3 groups; never, former, current. Considering the missing values in smoking status, we did not consider the amount of smoke such as pack year. alcohol intake was categorized into 2 groups; yes, no. Similar to smoking status, we did not include the amount of alcohol intake. Regular physical activities were defined using either of the following criteria, based on the International Physical Activiy Questionnaire (IPAQ) Scoring Protocol: vigorous intensity activity 3 or more days per week, or moderate intensity activity 5 or more days per week (Lee PH et al., 2011).

To reduce potential bias and multi-collinearity in the statistical model, we investigate the correlation between confounders (**Figure 6**). We set the threshold to exclude from the statistical model as 0.4. As a result, we excluded waist circumference, DBP, HDL, and TG. Although the correlation between sex and smoking status was higher than 0.4, we did not exclude smoking status in confounders due to clinical importance.

Figure 6. Correlation in confounders.

Finally, we confirmed confounders as sex, age, BMI, SBP, LDL, FBS, smoking status, alcohol intake, and regular physical activities (**Table 10**).

Table 10. Confounders in the study.

Variable	Value
Sex	1 = Male, 2 = Female
Age	Continuous
BMI	Continuous
SBP	Continuous
LDL	Continuous
FBS	Continuous
Smoking Status	1 = Never, 2 = Former, 3 = Current
Alcohol intake	0 = No, 1 = Yes
Regular Physical Activities	0 = No, 1 = Yes

6. Statistical analysis

To compare characteristics between participants with colon polyp and those without colon polyp, we conducted chi-square test and Student's t-test on categorical and continuous variables, respectively. We additionally described type, number, size, and location of colon polyp.

We calculated odds ratios (ORs) with 95% confidence interval (CI) for prevalent colon polyp using a logistic regression. We included pre-defined potential risk factors into the model, then estimated effect of each variable. Additionally, we conducted sensitivity analysis regarding age and sex.

To evaluate the association between colon polyp and CRC, We conducted multivariable logistic regression and regression standardization adjusting for sex, age, BMI, SBP, LDL, FBS, smoking status, alcohol intake and regular physical activities. We estimated adjusted prevalence, risk difference, adjusted prevalence ratio and 95% CI of each estimation for prevalent CRC. Sensitivity analysis was performed on those aged less than 50 and those aged 50 or above.

All analyses were performed using SAS 9.4 (SAS Institute Inc., NC, USA) and R version 4.1.0 (R Foundation for statistical computing, Vienna, Austria).

Ⅲ. RESULTS

1. Baseline characteristics

Among 360,753 participants, 63.0% did not have any colon polyp, 33.4% had low-risk polyp, and 3.5% had high-risk polyp. About 63.6% of the participants were male, and 36.4% of the participants were female. The mean age was 46.0 (10.6). Of those, participants who aged less than 50 were 64.6% and aged 50 or above were 35.4%. The mean BMI was 24.2 (3.31), the mean SBP was 119 (12.4), the mean LDL was 117 (34.9) and the mean FBS was 98.1 (19.8). 168 (0.1), 113 (0.1), 34 (0.1) were prevalent CRC in no polyp, low-risk polyp, and high-risk polyp, respectively. 19.3% were current smokers, 38.9% had alcohol intake and 39.7% worked out regularly.

The low-risk polyp group was more likely to be male, older, obese, current smoker, drinking alcohol, and less regular physical activities compared with the no polyp group. Also, high-risk polyp group was more likely to be male, older, obese, current smoker, drinking alcohol, and less regular physical activities compared with the no polyp group ($P < 0.001$) (**Table 11**).

Table 11. Baseline characteristics in all participants.

	Total (N=360,753)	No polyp (N=227,334)	Low-risk polyp (N=120,658)	High-risk polyp (N=12,761)	P-value
Sex					
Male	229,526 (63.6)	131,553 (57.9)	88,265 (73.2)	9,708 (76.1)	<0.001
Female	131,227 (36.4)	95,781 (42.1)	32,393 (26.8)	3,053 (23.9)	
Age, years	46.0 (10.6)	43.9 (10.2)	49.3 (10.3)	51.3 (10.6)	<0.001
aged less than 50	233,193 (64.6)	164,168 (72.2)	63,250 (52.4)	5,775 (45.3)	<0.001
aged 50 or above	127,560 (35.4)	63,166 (27.8)	57,408 (47.6)	6,986 (54.7)	
BMI, kg/m^2	24.2 (3.31)	23.9 (3.32)	24.7 (3.23)	24.9 (3.22)	<0.001
SBP, mmHg	119 (12.4)	118 (12.5)	120 (12.0)	122 (12.6)	<0.001
LDL, mg/dL	117 (34.9)	116 (34.7)	119 (35.1)	118 (35.9)	<0.001
FBS, mg/dL	98.1 (19.8)	96.1 (18.2)	101 (21.7)	104 (24.0)	<0.001
Prevalent CRC	315 (0.1)	168 (0.1)	113 (0.1)	34 (0.3)	<0.001

* Data are shown as number (percent) for categorical and mean (standard deviation) for continuous.

Table 11. Baseline characteristics in all participants (Continued).

	Total (N=360,753)	No polyp (N=227,334)	Low-risk polyp (N=120,658)	High-risk polyp (N=12,761)	P-value
Smoking Status					
Never	141,328 (39.2)	98,713 (43.4)	39,092 (32.4)	3,523 (27.6)	<0.001
Former	69,406 (19.2)	39,124 (17.2)	27,345 (22.7)	2,937 (23.0)	
Current	69,528 (19.3)	35,424 (15.6)	30,338 (25.1)	3,766 (29.5)	
Unknown	80,491 (22.3)	54,073 (23.8)	23,883 (19.8)	2,535 (19.9)	
Alcohol intake					
Yes	140,316 (38.9)	81,444 (35.8)	53,660 (44.5)	5,212 (40.8)	<0.001
No	29,616 (8.2)	16,997 (7.5)	11,513 (9.5)	1,106 (8.7)	
Unknown	190,821 (52.9)	128,893 (56.7)	55,485 (46.0)	6,443 (50.5)	
Regular Physical Activities					
Yes	143,303 (39.7)	85,147 (37.5)	52,823 (43.8)	5,333 (41.8)	<0.001
No	197,657 (54.8)	128,144 (56.4)	62,607 (51.9)	6,906 (54.1)	
Unknown	19,793 (5.5)	14,043 (6.2)	5,228 (4.3)	522 (4.1)	

* Data are shown as number (percent) for categorical and mean (standard deviation) for continuous.

Among 229,526 participants who were male, 57.3% did not have any colon polyp, 38.5% had low-risk polyp, and 4.2% had high-risk polyp. The mean age was 45.7 (10.3). Of those, participants who aged less than 50 were 66.4% and aged 50 or above were 33.6%. The mean BMI was 24.9 (3.04), the mean SBP was 121 (11.6), the mean LDL was 117 (34.9) and the mean FBS was 100.1 (20.8). 66 (0.1), 64 (0.1), 23 (0.2) were prevalent CRC in no polyp, low-risk polyp, and high-risk polyp, respectively. 28.8% were current smokers, 43.5% had alcohol intake and 52.5% worked out regularly.

The low-risk polyp group was more likely to be older, obese, current smoker, drinking alcohol, and less regular physical activities compared with the no polyp group. Also, high-risk polyp group was more likely to be older, obese, current smoker, drinking alcohol, and less regular physical activities compared with no polyp group. There were significant differences among all groups ($P < 0.001$) (**Table 12**).

Table 12. Baseline characteristics in male.

	Total (N=229,526)	No polyp (N=131,553)	Low-risk polyp (N=88,265)	High-risk polyp (N=9,708)	P-value
Age, year	45.7 (10.3)	43.3 (9.8)	48.7 (10.1)	50.9 (10.5)	<0.001
aged less than 50	152,383 (66.4)	99,112 (75.3)	48,713 (55.2)	4,558 (47)	<0.001
aged 50 or above	77,143 (33.6)	32,441 (24.7)	39,552 (44.8)	5,150 (53)	
BMI, kg/m^2	24.9 (3.04)	24.8 (3.04)	25.1 (3.03)	25.2 (3.07)	<0.001
SBP, mmHg	121 (11.6)	120.7 (11.7)	121.2 (11.5)	122.9 (12.0)	<0.001
LDL, mg/dL	119.4 (34.6)	119.2 (34.4)	119.8 (34.7)	117.9 (35.6)	<0.001
FBS, mg/dL	100.1 (20.8)	98.1 (19.4)	102.5 (22.2)	104.8 (24.1)	<0.001
Prevalent CRC	153 (0.1)	66 (0.1)	64 (0.1)	23 (0.2)	<0.001

* Data are shown as number (percent) for categorical and mean (standard deviation) for continuous.

Table 12. Baseline characteristics in male (Continued).

	Total (N=229,526)	No polyp (N=131,553)	Low-risk polyp (N=88,265)	High-risk polyp (N=9,708)	P-value
Smoking Status					<0.001
Never	49,595 (21.6)	32,348 (24.6)	15,899 (18)	1,348 (13.9)	
Former	65,640 (28.6)	36,251 (27.6)	26,522 (30)	2,867 (29.5)	
Current	66,090 (28.8)	33,045 (25.1)	29,403 (33.3)	3,642 (37.5)	
Unknown	48,201 (21)	29,909 (22.7)	16,441 (18.6)	1,851 (19.1)	
Alcohol intake					0.06
Yes	99,747 (43.5)	52,447 (39.9)	42,979 (48.7)	4,321 (44.5)	
No	11,236 (4.9)	5,533 (4.2)	5,147 (5.8)	556 (5.7)	
Unknown	118,543 (51.6)	73,573 (55.9)	40,139 (45.5)	4,831 (49.8)	
Regular Physical Activities					<0.001
Yes	120,418 (52.5)	70,902 (53.9)	44,351 (50.2)	5,165 (53.2)	
No	97,389 (42.4)	52,933 (40.2)	40,263 (45.6)	4,193 (43.2)	
Unknown	11,719 (5.1)	7,718 (5.9)	3,651 (4.1)	350 (3.6)	

* Data are shown as number (percent) for categorical and mean (standard deviation) for continuous.

Among 131,227 participants who were female, 73.0% did not have any colon polyp, 24.7% had low-risk polyp, and 2.3% had high-risk polyp. The mean age was 46.4 (11.1). Of those, participants who aged less than 50 were 61.6% and aged 50 or above were 38.4%. The mean BMI was 22.8 (3.34), the mean SBP was 114.3 (12.5), the mean LDL was 112 (35.0) and the mean FBS was 94.7 (17.4). 102 (0.1), 49 (0.2), 11 (0.4) were prevalent CRC in no polyp, low-risk polyp, and high-risk polyp, respectively. 2.6% were current smokers, 30.9% had alcohol intake and 58.9% worked out regularly.

The low-risk polyp group was more likely to be older, obese, current smoker, alcohol drinking, and less regular physical activities compared with the no polyp group. Also, high-risk polyp group was more likely to be older, obese, current smoker, and less regular physical activities compared with no polyp group. There were significant differences among all groups ($P < 0.001$) (**Table 13**).

Table 13. Baseline characteristics in female.

	Total (N=131,227)	No polyp (N=95,781)	Low-risk polyp (N=32,393)	High-risk polyp (N=3,053)	P-value
Age, year	46.4 (11.1)	44.8 (10.7)	50.9 (10.9)	52.3 (11)	<0.001
aged less than 50	80,810 (61.6)	65,056 (67.9)	14,537 (44.9)	1,217 (39.9)	<0.001
aged 50 or above	50,417 (38.4)	30,725 (32.1)	17,856 (55.1)	1,836 (60.1)	
BMI, kg/m^2	22.8 (3.34)	22.6 (3.25)	23.5 (3.44)	23.9 (3.48)	<0.001
SBP, mmHg	114.3 (12.5)	113.4 (12.3)	116.5 (12.6)	118.3 (13.6)	<0.001
LDL, mg/dL	112.4 (35.0)	110.5 (34.4)	117.1 (36.1)	119.3 (36.7)	<0.001
FBS, mg/dL	94.7 (17.4)	93.3 (16.1)	98 (19.8)	100.7 (23.5)	<0.001
Prevalent cancer	162 (0.1)	102 (0.1)	49 (0.2)	11 (0.4)	<0.001

* Data are shown as number (percent) for categorical and mean (standard deviation) for continuous.

Table 13. Baseline characteristics in female (Continued).

	Total (N=131,227)	No polyp (N=95,781)	Low-risk polyp (N=32,393)	High-risk polyp (N=3,053)	P-value
Smoking Status					<0.001
Never	91,733 (69.9)	66,365 (69.3)	23,193 (71.6)	2,175 (71.2)	
Former	3,766 (2.9)	2,873 (3)	823 (2.5)	70 (2.3)	
Current	3,438 (2.6)	2,379 (2.5)	935 (2.9)	124 (4.1)	
Unknown	32,290 (24.6)	24,164 (25.2)	7,442 (23)	684 (22.4)	
Alcohol intake					0.11
Yes	40,569 (30.9)	28,997 (30.3)	10,681 (33)	891 (29.2)	
No	18,380 (14)	11,464 (12)	6,366 (19.7)	550 (18)	
Unknown	72,278 (55.1)	55,320 (57.8)	15,346 (47.4)	1,612 (52.8)	
Regular Physical Activities					<0.001
Yes	77,239 (58.9)	57,242 (59.8)	18,256 (56.4)	1,741 (57)	
No	45,914 (35)	32,214 (33.6)	12,560 (38.8)	1,140 (37.3)	
Unknown	8,074 (6.2)	6,325 (6.6)	1,577 (4.9)	172 (5.6)	

* Data are shown as number (percent) for categorical and mean (standard deviation) for continuous.

In the low-risk polyp group, the mean number of colon polyp was 1.39 (0.959). Of these, 58.8% of cases were not described the location, 24.3% occurred in the proximal colon and 27.8% occurred in the distal colon. Among the proximal colons, the most common prevalent site was the ascending colon, which accounted for 10.4%. Otherwise, among the distal colons, the most common prevalent site was the sigmoid colon, which accounted for 13.5%. The max size of the colon polyp was 4.10 (2.13) and the mean size was 3.93 (1.77).

In the high-risk polyp group, the mean number of colon polyp was 3.14 (2.14). Of these, 10.9% of cases were not described the location, 55.6% occurred in the proximal colon and 72.6% occurred in the distal colon. Among the proximal colons, the most common prevalent site was the ascending colon, which accounted for 31.4%. Otherwise, among the distal colons, the most common prevalent site was the rectum, which accounted for 43.4%. The max size of the colon was 8.50 (5.80) and the mean size was 8.38 (5.73) (**Table 14**).

Table 14. Characteristics of colon polyp group.

Low-risk polyp (N=120,658)		High-risk polyp (N=12,761)	
Characteristics	N (%)	Characteristics	N (%)
Number	1.39 (± 0.959)	**Number**	3.14 (± 2.14)
Location		**Location**	
Proximal colon	32,792 (24.3 %)	Proximal colon	9,269 (55.6 %)
Cecum	4,965 (4.1 %)	Cecum	1,562 (12.2 %)
Ascending	12,534 (10.4 %)	Ascending	4,007 (31.4 %)
Hepatic flexure	4,133 (3.4 %)	Hepatic flexure	1,521 (11.9 %)
Transverse	12,553 (10.4 %)	Transverse	3,793 (29.7 %)
Distal colon	37,490 (27.8 %)	Distal colon	12,103 (72.6 %)
Splenic flexure	1,113 (0.9 %)	Splenic flexure	469 (3.7 %)
Descending	6,107 (5.1 %)	Descending	2,158 (16.9 %)
Sigmoid	16,292 (13.5 %)	Sigmoid	5,067 (39.7 %)
Rectum	11,662 (9.7 %)	Rectum	5,542 (43.4 %)
Anus	4,989 (4.1 %)	Anus	884 (6.9 %)
Unknown	70,902 (58.8 %)	Unknown	1,396 (10.9 %)
Size, mm		**Size, mm**	
max	4.10 (± 2.13)	max	8.50 (± 5.80)
mean	3.93 (± 1.77)	mean	8.38 (± 5.73)

2. Factors associated with prevalent polyp

For prevalent low-risk polyp, sex, age, smoking status, alcohol intake and regular physical activities were significantly associated in the adjusted model. Female was less likely to have low-risk polyp compared with male (OR 0.69, 95% CI 0.67-0.70). Older Participants were more likely to have low-risk polyp than younger participants (OR 1.06, 95% CI 1.06-1.06). Participants with higher BMI were more likely to have low-risk polyp than participants with and lower BMI (OR 1.05, 95% CI 1.04-1.05). Participants who ever smoked were more likely to have low-risk polyp than participants who never smoked (former OR 1.17, 95% CI 1.14-1.20; current OR 1.86, 95% CI 1.81-1.90). Participants who had consumed alcohol were more likely to have low-risk polyp than participants who did not consume alcohol (OR 1.64, 95% CI 1.61-1.67). Participants who had physical activities regularly were more likely to have low-risk polyp than those who did not (OR 1.14, 95% CI 1.12-1.16).

For prevalent high-risk polyp, sex, age, smoking status, and alcohol intake were significantly associated in the adjusted model. Female was less likely to have high-risk polyp compared with male (OR 0.71, 95% CI 0.66-0.75). Older participants were more likely to have high-risk polyp than younger participants (OR 1.08, 95% CI 1.07-1.08). Participants with higher BMI were more likely to have high-risk polyp than participants with lower BMI (OR 1.05, 95% CI 1.04-1.06). Participants who ever smoked were more likely to have high-risk polyp than participants who never smoked (former OR 1.39, 95% CI 1.30-1.48; current OR 2.89, 95% CI 2.71-3.07). Participants who had consumed alcohol were more likely to have high-risk polyp than participants who did not consume alcohol (OR 1.37, 95% CI 1.31-1.44) (**Table 15, Figure 7**).

Table 15. Odds ratio for prevalent polyp.

	Adjusted odds ratio (95% CI)	
	Low-risk polyp	High-risk polyp
Sex		
Male	reference	reference
Female	0.69 (0.67-0.70)	0.71 (0.66-0.75)
Age, year	1.06 (1.06-1.06)	1.08 (1.07-1.08)
BMI, kg/m^2	1.05 (1.04-1.05)	1.05 (1.04-1.06)
SBP, mmHg	1.00 (1.00-1.00)	1.01 (1.01-1.01)
LDL, mg/dL	1.00 (1.00-1.00)	1.00 (1.00-1.00)
FBS, mg/dL	1.00 (1.00-1.00)	1.01 (1.00-1.01)
Smoking Status		
Never	reference	reference
Former	1.17 (1.14-1.20)	1.39 (1.30-1.48)
Current	1.86 (1.81-1.90)	2.89 (2.71-3.07)
Alcohol intake		
No	reference	reference
Yes	1.64 (1.61-1.67)	1.37 (1.31-1.44)
Regular Physical Activities		
No	reference	reference
Yes	1.14 (1.12-1.16)	1.04 (1.00-1.09)

* Adjusted for age, sex (male or female), BMI, SBP, LDL, FBS, smoking status (yes or no), alcohol intake (yes or no), and regular physical activities (yes or no).

In participants who were male, age, BMI, smoking status, alcohol intake and regular physical activities were significantly associated with prevalent low-risk polyp in the adjusted model. Older Participants were more likely to have low-risk polyp than younger participants (OR 1.07, 95% CI 1.06-1.07). Participants with higher BMI were more likely to have low-risk polyp than participants with and lower BMI (OR 1.05, 95% CI 1.05-1.06). Participants who ever smoked were more likely to have low-risk polyp than participants who never smoked (former OR 1.18, 95% CI 1.13-1.23; current OR 1.91, 95% CI 1.84-1.99). Participants who had consumed alcohol were more likely to have low-risk polyp than participants who did not consume alcohol (OR 1.14, 95% CI 1.08-1.20). Participants who had physical activities regularly were more likely to have low-risk polyp than those who did not (OR 1.25, 95% CI 1.21-1.29).

In participants who were male, age, BMI, smoking status, and regular physical activities were significantly associated with prevalent high-risk polyp in the adjusted model. Older participants were more likely to have high-risk polyp than younger participants (OR 1.07, 95% CI 1.07-1.08). Participants with higher BMI were more likely to have high-risk polyp than participants with lower BMI (OR 1.06, 95% CI 1.05-1.07). Participants who ever smoked were more likely to have high-risk polyp than participants who never smoked (former OR 1.41, 95% CI 1.27-1.56; current OR 2.97, 95% CI 2.68-3.29). Participants who had physical activities regularly were more likely to have low-risk polyp than those who did not (OR 1.09, 95% CI 1.01-1.17) (**Table 16, Figure 7**).

Table 16. Odds ratio for prevalent polyp in male.

	Adjusted odds ratio (95% CI)	
	Low-risk polyp	High-risk polyp
Age, year	1.07 (1.06-1.07)	1.07 (1.07-1.08)
BMI, kg/m^2	1.05 (1.05-1.06)	1.06 (1.05-1.07)
SBP, mmHg	1.00 (1.00-1.00)	1.00 (1.00-1.01)
LDL, mg/dL	1.00 (1.00-1.00)	1.00 (1.00-1.00)
FBS, mg/dL	1.00 (1.00-1.00)	1.00 (1.00-1.00)
Smoking Status		
Never	Reference	Reference
Former	1.18 (1.13-1.23)	1.41 (1.27-1.56)
Current	1.91 (1.84-1.99)	2.97 (2.68-3.29)
Alcohol intake		
No	Reference	Reference
Yes	1.14 (1.08-1.2)	1.11 (0.99-1.25)
Regular Physical Activities		
No	Reference	Reference
Yes	1.25 (1.21-1.29)	1.09 (1.01-1.17)

* Adjusted for age, BMI, SBP, LDL, FBS, smoking status (yes or no), alcohol intake (yes or no), and regular physical activities (yes or no).

In participants who were female, age, BMI, smoking status, and regular physical activities were significantly associated with prevalent low-risk polyp in the adjusted model. Older Participants were more likely to have low-risk polyp than younger participants (OR 1.05, 95% CI 1.05-1.05). Participants with higher BMI were more likely to have low-risk polyp than participants with and lower BMI (OR 1.04, 95% CI 1.03-1.05). Participants who were current smoker were more likely to have low-risk polyp than participants who never smoked (current OR 1.48, 95% CI 1.30-1.68). Participants who had physical activities regularly were more likely to have low-risk polyp than those who did not (OR 1.19, 95% CI 1.13-1.24).

In participants who were female, age, BMI, and smoking status were significantly associated with prevalent high-risk polyp in the adjusted model. Older participants were more likely to have high-risk polyp than younger participants (OR 1.06, 95% CI 1.05-1.07). Participants with higher BMI were more likely to have high-risk polyp than participants with lower BMI (OR 1.06, 95% CI 1.04-1.08). Participants who were current smoker were more likely to have high-risk polyp than participants who never smoked (current OR 2.34, 95% CI 1.70-3.14). (**Table 17, Figure 7**).

Table 17. Odds ratio for prevalent polyp in female.

	Adjusted odds ratio (95% CI)	
	Low-risk polyp	High-risk polyp
Age, year	1.05 (1.05-1.05)	1.06 (1.05-1.07)
BMI, kg/m^2	1.04 (1.03-1.05)	1.06 (1.04-1.08)
SBP, mmHg	1.00 (1.00-1.00)	1.00 (1.00-1.01)
LDL, mg/dL	1.00 (1.00-1.00)	1.00 (1.00-1.00)
FBS, mg/dL	1.00 (1.00-1.01)	1.01 (1.00-1.01)
Smoking Status		
Never	Reference	Reference
Former	1.1 (0.97-1.25)	1.29 (0.88-1.83)
Current	1.48 (1.3-1.68)	2.34 (1.7-3.14)
Alcohol intake		
No	Reference	Reference
Yes	1.04 (0.98-1.09)	1.06 (0.91-1.23)
Regular Physical Activities		
No	Reference	Reference
Yes	1.19 (1.13-1.24)	1.09 (0.95-1.24)

* Adjusted for age, BMI, SBP, LDL, FBS, smoking status (yes or no), alcohol intake (yes or no), and regular physical activities (yes or no).

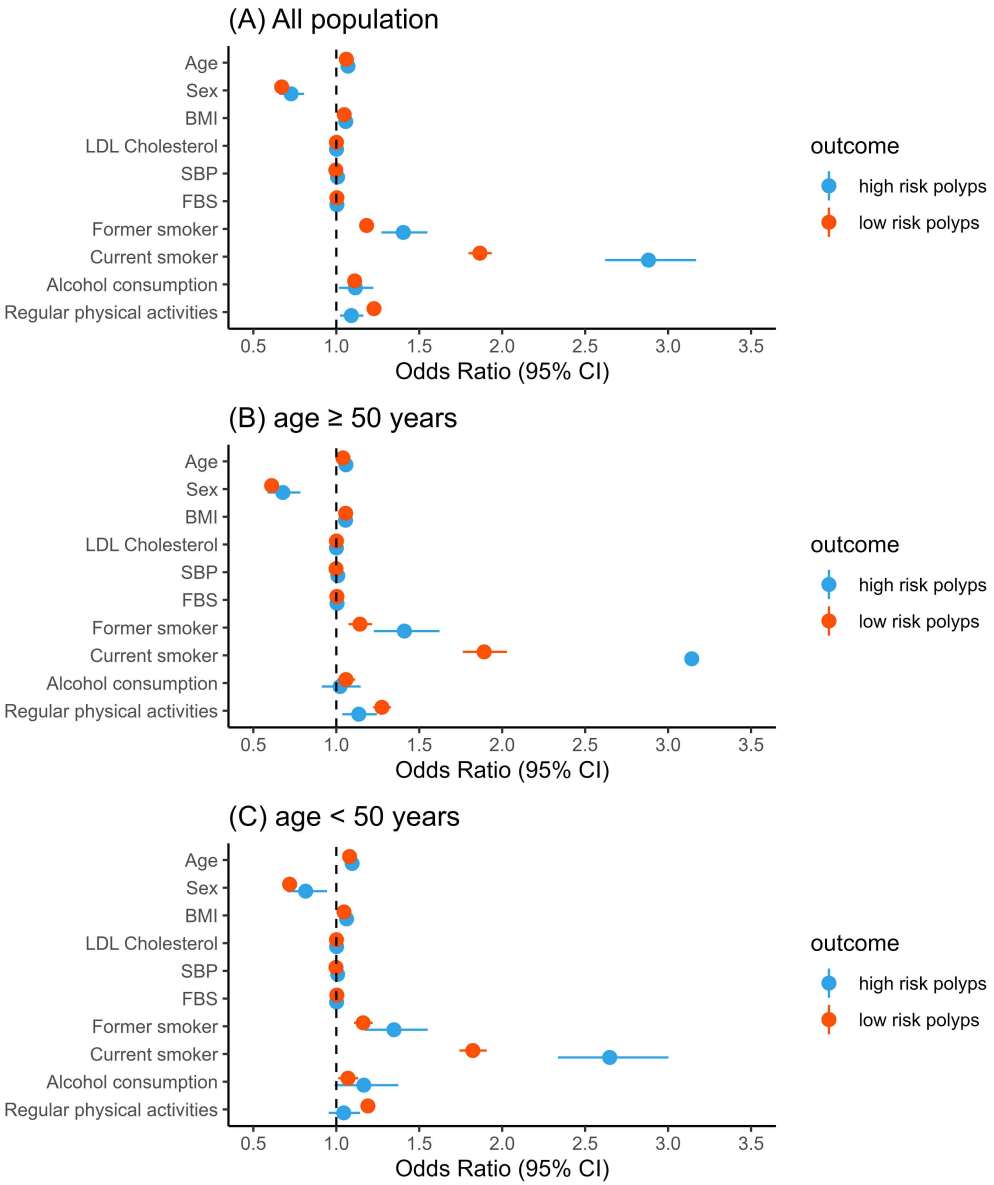

Figure 7. Odds ratio plots for prevalent polyp.

In participants who were men aged less than 50, age, BMI, smoking status, alcohol intake and regular physical activities were significantly associated with prevalent low-risk polyp. Older participants were more likely to have low-risk polyp than younger participants (OR 1.08, 95% CI 1.07-1.08). Participants with higher BMI were more likely to have low-risk polyp than participants with lower BMI (OR 1.05, 95% CI 1.04-1.05). Participants who ever smoked were more likely to have low-risk polyp than participants who never smoked (former OR 1.16, 95% CI 1.12-1.20; current OR 1.84, 95% CI 1.78-1.90). Participants who had consumed alcohol were more likely to have low-risk polyp than participants who did not (OR 1.60, 95% CI 1.55-1.64). Participants who had physical activities regularly were more likely to have low-risk polyp than those who did not (OR 1.13, 95% CI 1.10-1.16).

Also, age, BMI, smoking status, alcohol intake and regular physical activities were significantly associated with prevalent high-risk polyp. Older participants were more likely to have high-risk polyp than younger participants (OR 1.09, 95% CI 1.09-1.10). Participants with higher BMI were more likely to have high-risk polyp than participants with lower BMI (OR 1.05, 95% CI 1.04-1.06). Participants who ever smoked were more likely to have high-risk polyp than participants who never smoked (former OR 1.49, 95% CI 1.34-1.65; current OR 2.86, 95% CI 2.60-3.14). Participants who had consumed alcohol were more likely to have high-risk polyp than participants who did not (OR 1.52, 95% CI 1.42-1.63). Participants who had physical activities regularly were more likely to have low-risk polyp than those who did not (OR 1.09, 95% CI 1.02-1.17).

In participants who were women aged less than 50, age, BMI, smoking status, alcohol intake and regular physical activities were significantly associated with prevalent low-risk polyp. Older participants were more likely to have low-risk polyp than younger participants (OR 1.06, 95% CI 1.05-1.06). Participants with BMI were more likely to have low-risk polyp than participants with lower BMI (OR 1.04, 95% CI 1.03-1.04). Participants who ever smoked were more likely to have low-risk polyp than participants who never smoked (former OR 1.20, 95% CI 1.09-1.31; current OR 1.55, 95% CI 1.41-1.70). Participants who had consumed alcohol were more likely to have low-risk polyp than participants who did not (OR 1.60, 95% CI 1.53-1.67). Participants who had physical activities regularly were more likely to have low-risk polyp than those who did not (OR 1.09, 95% CI 1.04-1.14).

Otherwise, age, BMI, smoking status, and alcohol intake were significantly associated with prevalent high-risk polyp. Older participants were more likely to have high-risk polyp than younger participants (OR 1.07, 95% CI 1.06-1.09). Participants with higher BMI were more likely to have high-risk polyp than participants lower BMI (OR 1.06, 95% CI 1.04-1.08). Participants who were current smoker were more likely to have high-risk polyp than participants who never smoked (OR 1.91, 95% CI 1.45-2.47). Participants who had consumed alcohol were more likely to have high-risk polyp than participants who did not (OR 1.46, 95% CI 1.27-1.68) (**Table 18**).

Table 18. Odds ratio for prevalent polyp aged less than 50.

	Men aged less than 50 (N=152,378)		Women aged less than 50 (N=77,143)	
	Low-risk polyp	High-risk polyp	Low-risk polyp	High-risk polyp
Age, year	1.08 (1.07-1.08)	1.09 (1.09-1.10)	1.06 (1.05-1.06)	1.07 (1.06-1.09)
BMI, kg/m^2	1.05 (1.04-1.05)	1.05 (1.04-1.06)	1.04 (1.03-1.04)	1.06 (1.04-1.08)
SBP, mmHg	1.00 (1.00-1.00)	1.01 (1.01-1.01)	1.00 (1.00-1.00)	1.01 (1.00-1.01)
LDL, mg/dL	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
FBS, mg/dL	1.00 (1.00-1.00)	1.00 (1.00-1.01)	1.00 (1.00-1.00)	1.00 (1.00-1.01)
Smoking Status				
Never	reference	reference	reference	reference
Former	1.16 (1.12-1.20)	1.49 (1.34-1.65)	1.20 (1.09-1.31)	1.11 (0.80-1.49)
Current	1.84 (1.78-1.90)	2.86 (2.60-3.14)	1.55 (1.41-1.70)	1.91 (1.45-2.47)
Alcohol intake				
No	reference	reference	reference	reference
Yes	1.60 (1.55-1.64)	1.52 (1.42-1.63)	1.60 (1.53-1.67)	1.46 (1.27-1.68)
Regular Physical Activities				
No	reference	reference	reference	reference
Yes	1.13 (1.10-1.16)	1.09 (1.02-1.17)	1.09 (1.04-1.14)	1.05 (0.92-1.21)

* Adjusted for age, BMI, SBP, LDL, FBS, smoking status (yes or no), alcohol intake (yes or no), and regular physical activities (yes or no).

In participants who were men aged 50 or above, age, BMI, smoking status, alcohol intake and regular physical activities were significantly associated with prevalent low-risk polyp. Older participants were more likely to have low-risk polyp than younger participants with (OR 1.04, 95% CI 1.04-1.04). Participants with higher BMI were more likely to have low-risk polyp than participants with lower BMI (OR 1.06, 95% CI 1.05-1.06). Participants who ever smoked were more likely to have low-risk polyp than participants who never smoked (former OR 1.11, 95% CI 1.07-1.16; current OR 1.93, 95% CI 1.84-2.03). Participants who had consumed alcohol were more likely to have low-risk polyp than participants who did not (OR 1.74, 95% CI 1.68-1.80). Participants who had physical activities regularly were more likely to have low-risk polyp than those who did not (OR 1.16, 95% CI 1.12-1.20).

Otherwise, age, BMI, smoking status, and alcohol intake were significantly associated with prevalent high-risk polyp. Older Participants were more likely to have high-risk polyp than younger participants (OR 1.07, 95% CI 1.06-1.07). Participants with higher BMI were more likely to have high-risk polyp than participants with lower BMI (OR 1.06, 95% CI 1.05-1.06). Participants who ever smoked were more likely to have high-risk polyp than participants who never smoked (former OR 1.32, 95% CI 1.20-1.45; current OR 3.17, 95% CI 2.87-3.49). Participants who had consumed alcohol were more likely to have high-risk polyp than participants who did not (OR 1.20, 95% CI 1.12-1.29).

In participants who were women aged 50 or above, age, BMI, smoking status, alcohol intake and regular physical activities were significantly associated with prevalent low-risk polyp. Older participants were more likely to have low-risk polyp than younger participants (OR 1.04, 95% CI 1.04-1.05). Participants with higher BMI were more likely to have low-risk polyp than participants with lower BMI (OR 1.05, 95% CI 1.04-1.05). Participants who were current smoker were more likely to have low-risk polyp than participants who never smoked (OR 1.62, 95% CI 1.37-1.90). Participants who had consumed alcohol were more likely to have low-risk polyp than participants who did not consume alcohol (OR 1.52, 95% CI 1.44-1.61). Participants who had physical activities regularly were more likely to have low-risk polyp than those who did not (OR 1.14, 95% CI 1.09-1.19).

Otherwise, age, BMI, smoking status, and alcohol intake were significantly associated with prevalent high-risk polyp. Older participants were more likely to have high-risk polyp than younger participants (OR 1.05, 95% CI 1.04-1.05). Participants with higher BMI were more likely to have high-risk polyp than participants with lower BMI (OR 1.05, 95% CI 1.03-1.07). Participants who were current smoker were more likely to have high-risk polyp than participants who never smoked (OR 3.31, 95% CI 2.46-4.38). Participants who had consumed alcohol were more likely to have high-risk polyp than participants who did not (OR 1.22, 95% CI 1.05-1.40) (**Table 19**).

Table 19. Odds ratio for prevalent polyp aged 50 or above.

	Men aged 50 or above (N=80,809)		Women aged 50 or above (N=50,416)	
	Low-risk polyp	High-risk polyp	Low-risk polyp	High-risk polyp
Age, year	1.04 (1.04-1.04)	1.07 (1.06-1.07)	1.04 (1.04-1.05)	1.05 (1.04-1.06)
BMI, kg/m^2	1.06 (1.05-1.06)	1.06 (1.05-1.08)	1.05 (1.04-1.05)	1.05 (1.03-1.07)
SBP, mmHg	1.00 (1.00-1.00)	1.01 (1.01-1.01)	1.00 (1.00-1.00)	1.01 (1.00-1.01)
LDL, mg/dL	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
FBS, mg/dL	1.00 (1.00-1.00)	1.00 (1.00-1.01)	1.00 (1.00-1.01)	1.01 (1.00-1.01)
Smoking Status				
Never	reference	reference	reference	reference
Former	1.11 (1.07-1.16)	1.32 (1.20-1.45)	0.96 (0.81-1.15)	1.12 (0.72-1.66)
Current	1.93 (1.84-2.03)	3.17 (2.87-3.49)	1.62 (1.37-1.90)	3.31 (2.46-4.38)
Alcohol intake				
No	reference	reference	reference	reference
Yes	1.74 (1.68-1.80)	1.20 (1.12-1.29)	1.52 (1.44-1.61)	1.22 (1.05-1.40)
Regular Physical Activities				
No	reference	reference	reference	reference
Yes	1.16 (1.12-1.20)	0.98 (0.92-1.05)	1.14 (1.09-1.19)	1.04 (0.93-1.16)

* Adjusted for age, BMI, SBP, LDL, FBS, smoking status (yes or no), alcohol intake (yes or no), and regular physical activities (yes or no).

3. Prevalent CRC regarding colon polyp group

The adjusted prevalence for CRC in no polyp, low-risk polyp, and high-risk polyp were 0.06 (95% CI 0.05-0.07), 0.09 (0.07-0.10), and 0.13 (0.08-0.17), respectively. Compared to no polyp, adjusted risk differences in low-risk polyp and high-risk polyp were 0.03 (0.01-0.04) and 0.06 (0.01-0.11), respectively. The adjusted prevalence ratio were 1.22 (0.91-1.53) and 1.49 (0.73-2.25) in low-risk polyp and high-risk polyp compared with no polyp group.

Among those aged less than 50, the adjusted prevalence for CRC in no polyp, low-risk polyp, and high-risk polyp were 0.03 (0.03-0.04), 0.06 (0.04-0.07), and 0.09 (0.04-0.14), respectively. Compared to no polyp, adjusted risk differences in low-risk polyp and high-risk polyp were 0.02 (0.01-0.04) and 0.06 (0.00-0.11), respectively. The adjusted prevalence ratio were 1.17 (0.93-1.42) and 1.38 (0.81-1.95) in low-risk polyp and high-risk polyp compared with no polyp group.

Among those aged 50 or above, the adjusted prevalence for CRC in no polyp, low-risk polyp, and high-risk polyp group were 0.12 (0.10-0.14), 0.15 (0.14-0.17), and 0.20 (0.15-0.25), respectively. Compared to no polyp, adjusted risk differences in low-risk polyp and high-risk polyp were 0.04 (0.01-0.06) and 0.08 (0.02-0.15), respectively. The adjusted prevalence ratio were 1.26 (1.09-1.43) and 1.59 (1.15-2.02) in low-risk polyp and high-risk polyp compared with no polyp (**Table 20, Figure 8**)

Table 20. Prevalent CRC regarding polyp groups.

	Adjusted prevalence (95%CI)	Adjusted risk difference (95%CI)	Adjusted prevalence ratio (95% CI)
All population			
No polyp	0.06 (0.05-0.07)	Reference	Reference
Low-risk polyp	0.09 (0.07-0.10)	0.03 (0.01-0.04)	1.22 (0.91-1.53)
High-risk polyp	0.13 (0.08-0.17)	0.06 (0.01-0.11)	1.49 (0.73-2.25)
aged less than 50			
No polyp	0.03 (0.03-0.04)	Reference	Reference
Low-risk polyp	0.06 (0.04-0.07)	0.02 (0.01-0.04)	1.17 (0.93-1.42)
High-risk polyp	0.09 (0.04-0.14)	0.06 (0.00-0.11)	1.38 (0.81-1.95)
aged 50 or above			
No polyp	0.12 (0.10-0.14)	Reference	Reference
Low-risk polyp	0.15 (0.14-0.17)	0.04 (0.01-0.06)	1.26 (1.09-1.43)
High-risk polyp	0.20 (0.15-0.25)	0.08 (0.02-0.15)	1.59 (1.15-2.02)

* Adjusted for age, sex (male or female), BMI, SBP, LDL, FBS, smoking status (yes or no), alcohol intake (yes or no), and regular physical activities (yes or no).

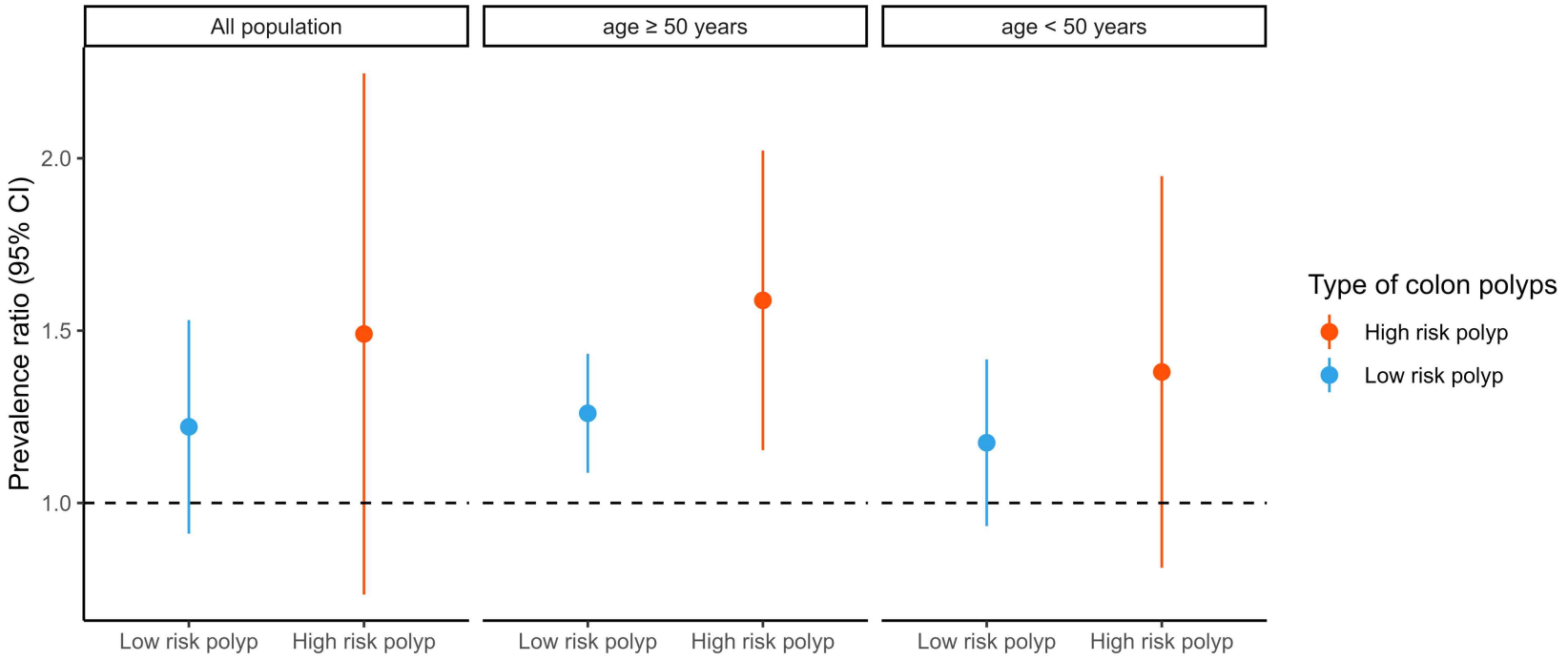

Figure 8. Prevalence ratio plot for CRC.

Ⅳ. DISCUSSION

1. Summary of findings

In our study, participants with high-risk polyp were more likely to be male, older, obese, smoker and drinker compared to participants with no polyp or low-risk polyp. Sex, age, BMI, smoking status and alcohol intake were associated with prevalent high-risk polyp. Participants who had high-risk polyp aged 50 or above had a higher prevalence of CRC compared to participants with no polyp and low-risk polyp aged 50 or above. It suggests that people aged 50 or above with high-risk polyp are at high-risk for CRC. Therefore, lifestyle modification is needed in high-risk group to prevent CRC.

2. Characteristics of study participants

The characteristics of participants are slightly different from previous studies because the setting of our study is based on health check-up center. 37.0% of participants had colon polyp when we used text-mining to detect the prevalence of colon polyp in our study. Also, 3.5% of participants had high-risk polyp. In a previous study, however, the detection rates of colon polyp and high-risk polyp was 49.4% and 5.8%, respectively, which were slightly higher than those of our study (Choi SY et al, 2014).

The previous study conducted by the hospital. Examinations conducted in hospital tend to target those who have symptom or have who have resulted in a positive fecal occult blood test. However, our study was conducted on participants who were examined at a health examination center. Participants in our study tend to visit regularly for examination according to government or company. Therefore, participants in our study may have included healthier people than in previous studies.

Health examinations conducted in hospitals tend to perform colonoscopy for high-risk groups with positive stool tests. However, since KMI performs colonoscopy as a screening test for companies or individuals, it targets relatively healthy people. Therefore, our study might include more healthier participants due to settings. That is, the difference of characteristics may reflect the difference of settings which include hospital and health examination center.

3. Factor associated with prevalent polyp

In our study, Factors related prevalent colon polyp were sex, age, BMI, smoking status, alcohol intake and regular physical activities. The results of our study were similar to previous studies.

Sex and age are well-known unmodified factors associated with prevalent colon polyp. Men have a greater risk of colon polyp than women. The risk of detecting colon polyp through the colonoscopy increases with age (McCashland TM et al., 2001) Also, men tend to have colon polyp at an earlier age than women (Grahn SW and Varma MG, 2008).

Although the mechanism which of the association between increased BMI and colon polyp is unknown, many studies have suggested positive association between BMI and colon polyp (Kitahara CM et al, 2013; Bailie L et al, 2017).

Smoking is the modifiable risk factor for colon polyp (Botteri E et al, 2007; Bailie L et al, 2017). Tobacco contains carcinogens that are thought to create irreversible genetic damage to the colorectal mucosa, initiating the formation of colon polyp (Botteri E et al, 2008; Liang PS et al, 2009). Even though smoking status, duration and intensity are associated with increased risk of colon polyp (Øines M et al, 2017).

Also, Alcohol is the major modifiable risk factor for colon polyp. Alcohol intake is a probable risk factor for colon polyp, but the mechanism by which it may affect the risk of prevent colon polyp risk is not known (Shrubsole MJ et al, 2008).

A meta-analysis of 15 studies that assessed alcohol intake and risk of serrated polyp showed a pooled estimate of 33% increased risk in individuals with a high versus a low alcohol intake. However, the authors found a significant risk of publication bias, and the effect estimate of the three largest studies in the meta-analysis all showed no statistical significant effect (Botteri E et al, 2008).

There is convincing evidence that a higher level of physical activity can reduce risk for prevalent CRC (Wiseman M, 2008). However, our study showed that regular physical activity increased risk for prevalent colon polyp. It might be the results of reverse causation due to study design. In our study, participants with regular physical activity were more likely to be men, older, and higher BMI. Those might be to manage their health.

4. Association between colon polyp and CRC

Our study suggests that high-risk polyp is associated with CRC. Traditionally, the polyp to cancer progression sequence was proposed a step that initiates the formation of adenoma and SSL, followed by a step that promotes the progression to more histologically advanced neoplasms, and then a step that transforms the tumors to invasive carcinoma. Today, molecular pathogenesis of CRC has advanced considerably and led to numerous revisions of the traditional models. (Vogelstein B, 1988). It is now recognized that SSL also have the potential for malignant transformation (Goldstein NS, 2006; Jass JR, 2004). In this study, we defined high-risk polyp as multiple or large adenoma or SSL Through the results of previous studies, we can suggest that our definition of polyp is reliable.

The results of our study suggest early colonoscopy for prevention CRC should conduct in aged 50 or above population. Previous study, which conducted in China reported higher prevalence of colon polyp in aged 50 or above participants. (Liu et al, 2005). This results were similar to Western guideline which describe the incidence of colorectal polyps and CRC increases aged 50 or above (Ahmed M, 2020; Hossain MS, 2022). That is, our results can interpret that population with aged 50 or above has increased risk of high-risk polyp, therefore the prevalence of CRC was significantly higher.

5. Limitation and strength

Our study has several limitations. First, measurement errors related to exposure and outcome might have influenced our results. Our data was collected by multiple check-up centers, the variability might be presented. Moreover, potential information bias about CRC might be presented, there was limited information of participants with CRC to conduct this study. Second, confounders including smoking status, alcohol intake and regular physical examination may induce recall bias and misclassification. Moreover, we did not consider unmeasured confounders such as dietary. As a result, our findings may have limitations to generalize in other settings.

Despite these limitations, our study has strengths in terms of sample size and quantity of information. First, we used above 300,000 samples with multiple check-up centers. Second, through unstructured data analysis, we used enormous data for colon polyp.

V. CONCLUSION

This study is a cross-sectional study using health screening examination data of more than 300,000 participants. Our study suggests that the association between high-risk polyp and CRC was high and participants who are male, aged 50 or above higher BMI, ever smoked, and consumed alcohol were high-risk group of prevalent colon polyp. To prevent CRC, we suggest male aged 50 or above need to conduct colonoscopy. Our results may be used to provide evidence for healthcare policies.

REFERENCE

Ahmed M. Colon Cancer: A Clinician's Perspective in 2019. Gastroenterology Res. 2020 Feb;13(1):1-10. doi: 10.14740/gr1239. Epub 2020 Feb 1. PMID: 32095167; PMCID: PMC7011914.

Altobelli E, Lattanzi A, Paduano R, Varassi G, di Orio F. Colorectal cancer prevention in Europe: burden of disease and status of screening programs. Prev Med. 2014 May;62:132-41. doi: 10.1016/j.ypmed.2014.02.010. Epub 2014 Feb 14. PMID: 24530610.

Bailie L, Loughrey MB, Coleman HG. Lifestyle Risk Factors for Serrated Colorectal Polyps: A Systematic Review and Meta-analysis. Gastroenterology. 2017 Jan;152(1):92-104. doi: 10.1053/j.gastro.2016.09.003. Epub 2016 Sep 14. PMID: 27639804.

Bjerrum A, Lindebjerg J, Andersen O, Fischer A, Lynge E. Long-term risk of colorectal cancer after screen-detected adenoma: Experiences from a Danish gFOBT-positive screening cohort. Int J Cancer. 2020 Aug 15;147(4):940-947. doi: 10.1002/ijc.32850. Epub 2020 Jan 25. PMID: 31894860.

Bond JH. Screening guidelines for colorectal cancer. Am J Med. 1999 Jan 25;106(1A):7S-10S. doi: 10.1016/s0002-9343(98)00339-8. PMID:

10089107.

Botteri E, Iodice S, Raimondi S, Maisonneuve P, Lowenfels AB. Cigarette smoking and adenomatous polyps: a meta-analysis. Gastroenterology. 2008 Feb;134(2):388-95. doi: 10.1053/j.gastro.2007.11.007. Epub 2007 Nov 4. PMID: 18242207.

Brenner H, Chang-Claude J, Jansen L, Knebel P, Stock C, Hoffmeister M. Reduced risk of colorectal cancer up to 10 years after screening, surveillance, or diagnostic colonoscopy. Gastroenterology. 2014 Mar;146(3):709-17. doi: 10.1053/j.gastro.2013.09.001. Epub 2013 Sep 5. PMID: 24012982.

Brenner H, Hoffmeister M, Stegmaier C, Brenner G, Altenhofen L, Haug U. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840,149 screening colonoscopies. Gut. 2007 Nov;56(11):1585-9. doi: 10.1136/gut.2007.122739. Epub 2007 Jun 25. PMID: 17591622; PMCID: PMC2095643.

Brosens LA, Offerhaus GJ, Giardiello FM. Hereditary Colorectal Cancer: Genetics and Screening. Surg Clin North Am. 2015 Oct;95(5):1067-80. doi: 10.1016/j.suc.2015.05.004. Epub 2015 Jun 16. PMID: 26315524; PMCID: PMC4555838.

Cardoso R, Guo F, Heisser T, Hackl M, Ihle P, De Schutter H, Van

Damme N, Valerianova Z, Atanasov T, Májek O, Mužík J, Nilbert MC, Tybjerg AJ, Innos K, Mägi M, Malila N, Bouvier AM, Bouvier V, Launoy G, Woronoff AS, Cariou M, Robaszkiewicz M, Delafosse P, Poncet F, Katalinic A, Walsh PM, Senore C, Rosso S, Vincerževskienė I, Lemmens VEPP, Elferink MAG, Johannesen TB, Kørner H, Pfeffer F, Bento MJ, Rodrigues J, Alves da Costa F, Miranda A, Zadnik V, Žagar T, Lopez de Munain Marques A, Marcos-Gragera R, Puigdemont M, Galceran J, Carulla M, Chirlaque MD, Ballesta M, Sundquist K, Sundquist J, Weber M, Jordan A, Herrmann C, Mousavi M, Ryzhov A, Hoffmeister M, Brenner H. Colorectal cancer incidence, mortality, and stage distribution in European countries in the colorectal cancer screening era: an international population-based study. Lancet Oncol. 2021 Jul;22(7):1002-1013. doi: 10.1016/S1470-2045(21)00199-6. Epub 2021 May 25. PMID: 34048685.

Chiu HM, Lee YC, Tu CH, Chang LC, Hsu WF, Chou CK, Tsai KF, Liang JT, Shun CT, Wu MS. Effects of metabolic syndrome and findings from baseline colonoscopies on occurrence of colorectal neoplasms. Clin Gastroenterol Hepatol. 2015 Jun;13(6):1134-42.e8. doi: 10.1016/j.cgh.2014.10.022. Epub 2014 Oct 29. PMID: 25445768.

Chi Z, Lin Y, Huang J, Lv MY, Chen J, Chen X, Zhang B, Chen Y, Hu J, He X, Lan P. Risk factors for recurrence of colorectal conventional adenoma and serrated polyp. Gastroenterol Rep (Oxf). 2021 Sep

16;10:goab038. doi: 10.1093/gastro/goab038. PMID: 35382162; PMCID: PMC8972988

Choi SY, Park DI, Lee CK, Cha JM, Lee SH, Whangbo Y, Eun CS, Han DS, Lee BI, Shin JE. [Usefulness of polyp and adenoma detection rate in the proximal and distal colon]. Korean J Gastroenterol. 2014 Jan 25;63(1):11-7. Korean. doi: 10.4166/kjg.2014.63.1.11. PMID: 24463283.

Chung SJ, Kim YS, Yang SY, Song JH, Kim D, Park MJ, Kim SG, Song IS, Kim JS. Five-year risk for advanced colorectal neoplasia after initial colonoscopy according to the baseline risk stratification: a prospective study in 2452 asymptomatic Koreans. Gut. 2011 Nov;60(11):1537-43. doi: 10.1136/gut.2010.232876. Epub 2011 Mar 22. PMID: 21427200.

Conteduca V, Sansonno D, Russi S, Dammacco F. Precancerous colorectal lesions (Review). Int J Oncol. 2013 Oct;43(4):973-84. doi: 10.3892/ijo.2013.2041. Epub 2013 Jul 29. PMID: 23900573.

Cottet V, Jooste V, Fournel I, Bouvier AM, Faivre J, Bonithon-Kopp C. Long-term risk of colorectal cancer after adenoma removal: a population-based cohort study. Gut. 2012 Aug;61(8):1180-6. doi: 10.1136/gutjnl-2011-300295. Epub 2011 Nov 22. PMID: 22110052.

Cuschieri S. The STROBE guidelines. Saudi J Anaesth. 2019 Apr;13(Suppl 1):S31-S34. doi: 10.4103/sja.SJA_543_18. PMID: 30930717; PMCID: PMC6398292.

de Bruijn B, Martin J. Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inform. 2002 Dec 4;67(1-3):7-18. doi: 10.1016/s1386-5056(02)00050-3. PMID: 12460628.

Duvvuri A, Chandrasekar VT, Srinivasan S, Narimiti A, Dasari C, Nutalapati V, Kennedy KF, Spadaccini M, Antonelli G, Desai M, Vennalaganti P, Kohli D, Kaminski MF, Repici A, Hassan C, Sharma P. Risk of Colorectal Cancer and Cancer Related Mortality After Detection of Low-risk or High-risk Adenomas, Compared With No Adenoma, at Index Colonoscopy: A Systematic Review and Meta-analysis. Gastroenterology. 2021 May;160(6):1986-1996.e3. doi: 10.1053/j.gastro.2021.01.214. Epub 2021 Jan 29. Erratum in: Gastroenterology. 2022 Aug;163(2):536. PMID: 33524401.

Fitzpatrick-Lewis D, Ali MU, Warren R, Kenny M, Sherifali D, Raina P. Screening for Colorectal Cancer: A Systematic Review and Meta-Analysis. Clin Colorectal Cancer. 2016 Dec;15(4):298-313. doi: 10.1016/j.clcc.2016.03.003. Epub 2016 Mar 31. PMID: 27133893.

GBD 2019 Colorectal Cancer Collaborators. Global, regional, and national burden of colorectal cancer and its risk factors, 1990-2019: a

systematic analysis for the Global Burden of Disease Study 2019. Lancet Gastroenterol Hepatol. 2022 Jul;7(7):627-647. doi: 10.1016/S2468-1253(22)00044-9. Epub 2022 Apr 7. Erratum in: Lancet Gastroenterol Hepatol. 2022 Aug;7(8):704. PMID: 35397795; PMCID: PMC9192760.

Goldstein NS. Serrated pathway and APC (conventional)-type colorectal polyps: molecular-morphologic correlations, genetic pathways, and implications for classification. Am J Clin Pathol. 2006 Jan;125(1):146-53. PMID: 16483003.

Grahn SW, Varma MG. Factors that increase risk of colon polyps. Clin Colon Rectal Surg. 2008 Nov;21(4):247-55. doi: 10.1055/s-0028-1089939. PMID: 20011435; PMCID: PMC2780253.

He X, Wu K, Ogino S, Giovannucci EL, Chan AT, Song M. Association Between Risk Factors for Colorectal Cancer and Risk of Serrated Polyps and Conventional Adenomas. Gastroenterology. 2018 Aug;155(2):355-373.e18. doi: 10.1053/j.gastro.2018.04.019. Epub 2018 Apr 24. PMID: 29702117; PMCID: PMC6067965.

Hong SN, Yang DH, Kim YH, Hong SP, Shin SJ, Kim SE, Lee BI, Lee SH, Park DI, Kim HS, Yang SK, Kim HJ, Kim SH, Kim HJ; Multi-Society Task Force forDevelopment of Guidelines for Colorectal Polyp Screening, Surveillance and Management. [Korean guidelines

for post-polypectomy colonoscopic surveillance]. Korean J Gastroenterol. 2012 Feb;59(2):99-117. Korean. doi: 10.4166/kjg.2012.59.2.99. PMID: 22387835.

Hossain MS, Karuniawati H, Jairoun AA, Urbi Z, Ooi J, John A, Lim YC, Kibria KMK, Mohiuddin AKM, Ming LC, Goh KW, Hadi MA. Colorectal Cancer: A Review of Carcinogenesis, Global Epidemiology, Current Challenges, Risk Factors, Preventive and Treatment Strategies. Cancers (Basel). 2022 Mar 29;14(7):1732. doi: 10.3390/cancers14071732. PMID: 35406504; PMCID: PMC8996939.

Ishaq S, Siau K, Harrison E, Tontini GE, Hoffman A, Gross S, Kiesslich R, Neumann H. Technological advances for improving adenoma detection rates: The changing face of colonoscopy. Dig Liver Dis. 2017 Jul;49(7):721-727. doi: 10.1016/j.dld.2017.03.030. Epub 2017 Apr 9. PMID: 28454854.

Jang HW, Park SJ, Hong SP, Cheon JH, Kim WH, Kim TI. Risk Factors for Recurrent High-Risk Polyps after the Removal of High-Risk Polyps at Initial Colonoscopy. Yonsei Med J. 2015 Nov;56(6):1559-65. doi: 10.3349/ymj.2015.56.6.1559. PMID: 26446637; PMCID: PMC4630043.

Jass JR. Hyperplastic polyps and colorectal cancer: is there a link? Clin Gastroenterol Hepatol. 2004 Jan;2(1):1-8. doi: 10.1016/s1542-3565(03)00284-2. PMID: 15017625.

Kang MJ, Won YJ, Lee JJ, Jung KW, Kim HJ, Kong HJ, Im JS, Seo HG; Community of Population-Based Regional Cancer Registries. Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2019. Cancer Res Treat. 2022 Apr;54(2):330-344. doi: 10.4143/crt.2022.128. Epub 2022 Mar 16. PMID: 35313102; PMCID: PMC9016309.

Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol. 2019 Dec;16(12):713-732. doi: 10.1038/s41575-019-0189-8. Epub 2019 Aug 27. PMID: 31455888.

Kitahara CM, Berndt SI, de González AB, Coleman HG, Schoen RE, Hayes RB, Huang WY. Prospective investigation of body mass index, colorectal adenoma, and colorectal cancer in the prostate, lung, colorectal, and ovarian cancer screening trial. J Clin Oncol. 2013 Jul 1;31(19):2450-9. doi: 10.1200/JCO.2012.48.4691. Epub 2013 May 28. PMID: 23715565; PMCID: PMC3691360.

Laish I, Blechman I, Feingelernt H, Konikoff FM. Yield of second surveillance colonoscopy to predict adenomas with high-risk characteristics. Dig Liver Dis. 2015 Sep;47(9):805-10. doi: 10.1016/j.dld.2015.05.005. Epub 2015 May 19. PMID: 26048253.

Lee PH, Macfarlane DJ, Lam TH, Stewart SM. Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): a systematic review. Int J Behav Nutr Phys Act. 2011 Oct 21;8:115. doi: 10.1186/1479-5868-8-115. PMID: 22018588; PMCID: PMC32148244.

Liang PS, Chen TY, Giovannucci E. Cigarette smoking and colorectal cancer incidence and mortality: systematic review and meta-analysis. Int J Cancer. 2009 May 15;124(10):2406-15. doi: 10.1002/ijc.24191. PMID: 19142968.

Lieberman D, Nadel M, Smith RA, Atkin W, Duggirala SB, Fletcher R, Glick SN, Johnson CD, Levin TR, Pope JB, Potter MB, Ransohoff D, Rex D, Schoen R, Schroy P, Winawer S. Standardized colonoscopy reporting and data system: report of the Quality Assurance Task Group of the National Colorectal Cancer Roundtable. Gastrointest Endosc. 2007 May;65(6):757-66. doi: 10.1016/j.gie.2006.12.055. PMID: 17466195.

Lieberman D, Sullivan BA, Hauser ER, Qin X, Musselwhite LW, O'Leary MC, Redding TS 4th, Madison AN, Bullard AJ, Thomas R, Sims KJ, Williams CD, Hyslop T, Weiss D, Gupta S, Gellad ZF, Robertson DJ, Provenzale D. Baseline Colonoscopy Findings Associated With 10-Year Outcomes in a Screening Cohort Undergoing Colonoscopy Surveillance. Gastroenterology. 2020 Mar;158(4):862-874.e8. doi: 10.1053/j.gastro.2019.07.052. Epub 2019 Jul 31. PMID: 31376388.

Lin JS, Piper MA, Perdue LA, Rutter C, Webber EM, O'Connor E, Smith N, Whitlock EP. Screening for Colorectal Cancer: A Systematic Review for the U.S. Preventive Services Task Force [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2016 Jun. Report No.: 14-05203-EF-1. PMID: 27441328.

Liu HH, Wu MC, Peng Y, Wu MS. Prevalence of advanced colonic polyps in asymptomatic Chinese. World J Gastroenterol. 2005 Aug 14;11(30):4731-4. doi: 10.3748/wjg.v11.i30.4731. PMID: 16094719; PMCID: PMC4615420.

McCashland TM, Brand R, Lyden E, de Garmo P; CORI Research Project. Gender differences in colorectal polyps and tumors. Am J Gastroenterol. 2001 Mar;96(3):882-6. doi: 10.1111/j.1572-0241.2001.3638_a.x. PMID: 11280569.

Morson BC. Evolution of cancer of the colon and rectum. Cancer. 1974 Sep;34(3):suppl:845-9. doi: 10.1002/1097-0142(197409)34:3+<845::aid-cncr2820340710>3.0.co;2-h. PMID: 4851945.

Nagtegaal ID, Odze RD, Klimstra D, Paradis V, Rugge M, Schirmacher P, Washington KM, Carneiro F, Cree IA; WHO Classification of Tumours Editorial Board. The 2019 WHO classification of tumours of

the digestive system. Histopathology. 2020 Jan;76(2):182-188. doi: 10.1111/his.13975. Epub 2019 Nov 13. PMID: 31433515; PMCID: PMC7003895.

Navarro M, Nicolas A, Ferrandez A, Lanas A. Colorectal cancer population screening programs worldwide in 2016: An update. World J Gastroenterol. 2017 May 28;23(20):3632-3642. doi: 10.3748/wjg.v23.i20.3632. PMID: 28611516; PMCID: PMC5449420.

Øines M, Helsingen LM, Bretthauer M, Emilsson L. Epidemiology and risk factors of colorectal polyps. Best Pract Res Clin Gastroenterol. 2017 Aug;31(4):419-424. doi: 10.1016/j.bpg.2017.06.004. Epub 2017 Jun 28. PMID: 28842051.

Park B, Jun JK, Kim BC, Choi KS, Suh M; Expert Advisory Committee; Monitoring Committee; Center for Korean Colonoscopy Screening Pilot Study; Research Team on the Protocol Development of Pilot study. Korean colonoscopy screening pilot study (K-cospi) for screening colorectal cancer: study protocol for the multicenter, community-based clinical trial. BMC Gastroenterol. 2021 Jan 26;21(1):36. doi: 10.1186/s12876-021-01610-1. PMID: 33499810; PMCID: PMC7836193.

Quintero E, Castells A, Bujanda L, Cubiella J, Salas D, Lanas Á, Andreu M, Carballo F, Morillas JD, Hernández C, Jover R, Montalvo I,

Arenas J, Laredo E, Hernández V, Iglesias F, Cid E, Zubizarreta R, Sala T, Ponce M, Andrés M, Teruel G, Peris A, Roncales MP, Polo-Tomás M, Bessa X, Ferrer-Armengou O, Grau J, Serradesanferm A, Ono A, Cruzado J, Pérez-Riquelme F, Alonso-Abreu I, de la Vega-Prieto M, Reyes-Melian JM, Cacho G, Díaz-Tasende J, Herreros-de-Tejada A, Poves C, Santander C, González-Navarro A; COLONPREV Study Investigators. Colonoscopy versus fecal immunochemical testing in colorectal-cancer screening. N Engl J Med. 2012 Feb 23;366(8):697-706. doi: 10.1056/NEJMoa1108895. Erratum in: N Engl J Med. 2016 May 12;374(19):1898. PMID: 22356323.

Sano W, Hirata D, Teramoto A, Iwatate M, Hattori S, Fujita M, Sano Y. Serrated polyps of the colon and rectum: Remove or not? World J Gastroenterol. 2020 May 21;26(19):2276-2285. doi: 10.3748/wjg.v26.i19.2276. PMID: 32476792; PMCID: PMC7243646.

Shrubsole MJ, Wu H, Ness RM, Shyr Y, Smalley WE, Zheng W. Alcohol drinking, cigarette smoking, and risk of colorectal adenomatous and hyperplastic polyps. Am J Epidemiol. 2008 May 1;167(9):1050-8. doi: 10.1093/aje/kwm400. Epub 2008 Feb 27. PMID: 18304959.

Sillars-Hardebol AH, Carvalho B, van Engeland M, Fijneman RJ, Meijer GA. The adenoma hunt in colorectal cancer screening: defining the target. J Pathol. 2012 Jan;226(1):1-6. doi: 10.1002/path.3012. Epub 2011

Nov 14. PMID: 21984228.

Sninsky JA, Shore BM, Lupu GV, Crockett SD. Risk Factors for Colorectal Polyps and Cancer. Gastrointest Endosc Clin N Am. 2022 Apr;32(2):195-213. doi: 10.1016/j.giec.2021.12.008. Epub 2022 Feb 22. PMID: 35361331.

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209-249. doi: 10.3322/caac.21660. Epub 2021 Feb 4. PMID: 33538338.

Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL. Genetic alterations during colorectal-tumor development. N Engl J Med. 1988 Sep 1;319(9):525-32. doi: 10.1056/NEJM198809013190901. PMID: 2841597.

Wilkins T, McMechan D, Talukder A. Colorectal Cancer Screening and Prevention. Am Fam Physician. 2018 May 15;97(10):658-665. PMID: 29763272.

Winawer SJ, Zauber AG, Fletcher RH, Stillman JS, O'Brien MJ, Levin B, Smith RA, Lieberman DA, Burt RW, Levin TR, Bond JH, Brooks D, Byers T, Hyman N, Kirk L, Thorson A, Simmang C, Johnson D,

Rex DK; US Multi-Society Task Force on Colorectal Cancer; American Cancer Society. Guidelines for colonoscopy surveillance after polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer and the American Cancer Society. Gastroenterology. 2006 May;130(6):1872-85. doi: 10.1053/j.gastro.2006.03.012. PMID: 16697750.

Wiseman M. The second World Cancer Research Fund/American Institute for Cancer Research expert report. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. Proc Nutr Soc. 2008 Aug;67(3):253-6. doi: 10.1017/S002966510800712X. Epub 2008 May 1. PMID: 18452640.

Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. Transl Oncol. 2021 Oct;14(10):101174. doi: 10.1016/j.tranon.2021.101174. Epub 2021 Jul 6. PMID: 34243011; PMCID: PMC8273208.

Zauber AG, Winawer SJ, O'Brien MJ, Lansdorp-Vogelaar I, van Ballegooijen M, Hankey BF, Shi W, Bond JH, Schapiro M, Panish JF, Stewart ET, Waye JD. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N Engl J Med. 2012 Feb 23;366(8):687-96. doi: 10.1056/NEJMoa1100370. PMID: 22356322; PMCID: PMC3322371.

국 문 요 약

텍스트 마이닝 기법을 활용한
대장용종과 대장암의 연관성 분석 연구

연구 배경

대장암은 세계적으로 발생률과 사망률이 높아지며, 보건사회적 문제로 대두되고 있다. 대장암을 예방하기 위해서는 대장암의 주요한 위험 요인으로 알려진 대장용종을 조기에 발견하고 제거하는 것이 중요하다. 대장용종과 대장암의 연관성에 대해서는 많은 선행 연구가 진행되었지만, 연구 대상자 수 또는 용종의 제한적인 정보 등에 한계가 존재한다. 따라서 본 연구에서는 전국 단위의 건강검진 정보와 비정형 데이터을 활용하여 대장용종을 고위험군 및 저위험군으로 분류하고, 대장암과의 연관성을 분석하는 연구를 수행하고자 한다.

연구 방법

본 연구는 한국의학연구소의 건강검진 자료를 활용한 단면 연구로, 2008년부터 2019년까지 1회 이상 대장내시경을 수검한 수검자를 대상으로 한다 (N = 360,753). 대장용종은 아형, 개수, 크기를 고려하여 저위험 용종과 고위험 용종으로 구분하였다. 대장용종과 대장암에 대한 정보는 대장내시경 판독문을 텍스트마이닝하여 추출하였다. 대장용종의 위험요인 파악을 위해, 인구사회학적 요인, 생활습관 요인, 검진 결과를 보정하여 로지스틱 회귀분석을 수행하였다. 대장용종과 대장암의 연관성을 파악하기 위해 회귀표준화방법을 수행하였다. 추가적으로, 성별과 연령을 고려하여 민감도 분석을 수행하였다.

연구 결과

총 360,753명 중 63.0%는 대장용종이 없었고 33.4%는 저위험 용종군, 3.5%는 고위험 용종군에 포함되었다. 대장용종 발생에 영향을 미치는 요인은 성별, 연령, BMI, 과거 및 현재 흡연, 음주였다. 보정된 유병률은 대장용종이 없는 군에서 0.06%, 저위험군에서 0.09%, 고위험군에서 0.13%이었다. 보정된 유병률의 비는 대장용종이 없는 군 대비 저위험군에서 1.22 (95% 신뢰구간, 0.91-1.53), 고위험군에서 1.49 (95% 신뢰구간, 0.73-2.25) 였다.

결론

대장용종과 연관이 있는 주요한 인자는 성별, 연령, BMI, 흡연, 음주 및 신체활동이었다. 본 연구 결과는 50세 이상의 남성에서 대장암 예방을 위한 대장내시경이 필요함을 제시한다. 이 연구 결과는 향후 대장내시경과 관련된 건강증진 정책의 근거로써 활용될 것으로 사료된다.

핵심어: 대장내시경, 대장용종, 대장암, 건강검진, 단면 연구, 텍스트마이닝