Original Article

J Prev Med Public Health 2023;56:312-318 • https://doi.org/10.3961/jpmph.23.033

pISSN 1975-8375 eISSN 2233-4521

Journal of Preventive Medicine & Public Health

Operational Definitions of Colorectal Cancer in the Korean National Health Insurance Database

Hyeree Park[1,2,3], Yu Rim Kim[4], Yerin Pyun[5]*, Hyundeok Joo[1], Aesun Shin[1,2,3]**

[1]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Korea; [2]Cancer Research Institute, Seoul National University, Seoul, Korea; [3]Interdisciplinary Program in Cancer Biology Major, Seoul National University College of Medicine, Seoul, Korea; [4]College of Medicine, Ewha Womans University, Seoul, Korea; [5]College of Nursing, Seoul National University, Seoul, Korea

Objectives: We reviewed the operational definitions of colorectal cancer (CRC) from studies using the Korean National Health Insurance Service (NHIS) and compared CRC incidence derived from the commonly used operational definitions in the literature with the statistics reported by the Korea Central Cancer Registry (KCCR).

Methods: We searched the MEDLINE and KoreaMed databases to identify studies containing operational definitions of CRC, published until January 15, 2021. All pertinent data concerning the study period, the utilized database, and the outcome variable were extracted. Within the NHIS-National Sample Cohort, age-standardized incidence rates (ASRs) of CRC were calculated for each operational definition found in the literature between 2005 and 2019. These rates were then compared with ASRs from the KCCR.

Results: From the 62 eligible studies, 9 operational definitions for CRC were identified. The most commonly used operational definition was "C18-C20" (n=20), followed by "C18-C20 with claim code for treatment" (n=3) and "C18-C20 with V193 (code for registered cancer patients' payment deduction)" (n=3). The ASRs reported using these operational definitions were lower than the ASRs from KCCR, except for "C18-C20 used as the main diagnosis." The smallest difference in ASRs was observed for "C18-C20," followed by "C18-C20 with V193," and "C18-C20 with claim code for hospitalization or code for treatment."

Conclusions: In defining CRC patients utilizing the NHIS database, the ASR derived through the operational definition of "C18-C20 as the main diagnosis" was comparable to the ASR from the KCCR. Depending on the study hypothesis, operational definitions using treatment codes may be utilized.

Key words: Colorectal neoplasms, Rectal neoplasms, Incidence, National Health Programs

Received: January 20, 2023 Accepted: May 10, 2023

Corresponding author: Aesun Shin
Department of Preventive Medicine, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul 03080, Korea
E-mail: shinaesun@snu.ac.kr

*Current affiliation: Law School, Kyung Hee University, Seoul, Korea.
**Current affiliation: Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA.

INTRODUCTION

Colorectal cancer (CRC) ranks as the fourth most common cancer in terms of incidence and the third leading cause of death in Korea [1,2]. According to the Korea Central Cancer Registry (KCCR), the age-standardized incidence rate (ASR) of CRC was 28.7 per 100 000 in 2019. This rate increased from 1999 to 2010 but has been on a decline since then [3].

The National Health Information Database (NHID), established by the National Health Insurance Service (NHIS), is widely utilized for real-world data in clinical and epidemiological studies [4]. As a claims database, it encompasses healthcare utilization,

socio-demographic variables, and health screening information for approximately 97% of the Korean population. However, since the NHIS gathers data for reimbursement purposes, the main diagnosis code in the NHID may not accurately reflect the true disease status [5]. Consequently, it is essential to establish an appropriate operational definition in the study design. Examples of operational definitions include diagnosis codes based on the 10th International Classification of Diseases, treatment codes, hospitalization codes, and codes for registered cancer patients' payment deductions.

In a prior analysis that compared CRC incidence based on 6 operational definitions and the KCCR, the pattern of CRC incidence was found to be most similar to the registry data when CRC patients were identified using main diagnosis codes [6]. Another study, which compared cancer incidence according to 2 operational definitions—primary diagnosis versus rare and intractable disease (RID) claims—demonstrated that both definitions had over 90% sensitivity in identifying CRC patients [5]. However, no reviews have specifically focused on the utilization of operational definitions for CRC.

In recent years, there has been a growing demand for epidemiological research on CRC using the NHID. Therefore, we reviewed the operational definitions of CRC from published studies utilizing the NHID and aimed to compare the CRC incidence derived from these commonly used operational definitions with the statistics reported by the KCCR. Furthermore, we endeavored to propose a guideline for choosing the most suitable operational definition based on the research objective.

METHODS

Data Source and Search Strategy

Two authors (YP and HJ) conducted literature searches of the PubMed and KoreaMed databases using the following keywords: ("Health Insurance Review and Assessment" or "HIRA" or "National Health Insurance Service" or "National Health Information database" or "NHIS" or "population-based cohort") and ("colorectal cancer" or "colon cancer" or "rectal cancer") and ("Korea" or "Korean"). For the PubMed search, we utilized Medical Subject Heading (MeSH) terms. The most recent search was carried out on January 8, 2021. We did not apply any language restrictions during the search process. We reviewed papers published before December 8, 2020, and excluded any duplicates.

Study Selection

The inclusion criteria for eligible studies were as follows: (1) the study provided an operational definition of CRC; (2) it was an observational study with a cohort, case-control, or cross-sectional design; and (3) it was based on the NHID, Health Insurance Review and Assessment Service database, or national cancer registry. Studies were excluded if they met any of the following criteria: (1) lacking an operational definition of CRC; (2) not being based on the NHIS database; (3) using data collected in 2002 or earlier; and (4) being secondary literature, such as systematic reviews and meta-analyses.

Data Extraction

The first and second authors (HP and YK) independently screened the titles and abstracts of studies that met the inclusion criteria. The full texts were reviewed by 2 independent reviewers (HP and YK). The third reviewer (AS) compared the results of the 2 reviewers and resolved any disagreements. The first author (HP) extracted pertinent data concerning the operational definition, study period, sample size, database, and outcome variable.

Comparison of Colorectal Cancer Incidence by Operational Definitions

The NHIS provides a population-based sample cohort known as the National Health Insurance Service-National Sample Cohort (NHIS-NSC). The NHIS-NSC comprises a total of 1 137 861 randomly selected NHIS participants from 2002 to 2019, representing approximately 2% of the NHIS insured individuals and Medical Aid beneficiaries. Further details about the cohort have been published elsewhere [7].

We identified the diverse operational definitions of CRC found in the existing literature and calculated the number of CRC cases in the NHIS-NSC based on these prevalent definitions. The date of CRC diagnosis was established as the first instance when an individual received the diagnosis code for CRC, since this information is not provided by the NHIS-NSC. In order to exclude subjects with a prior history of CRC, we excluded those who had any diagnosis code for CRC before January 2005. The codes utilized for identifying CRC diagnosis and treatment can be found in Supplemental Material 1.

The ASRs of CRC were calculated using various operational definitions and subsequently compared with the ASRs obtained from the KCCR. The KCCR-derived ASRs were calculated using crude incidence rates of cancer by site, sex, and age group (in

5-year intervals) from Statistics Korea [8]. The standard population used was the mid-year Korean population in 2010. The payment deduction code for registered cancer patients has been available since 2005, coinciding with the implementation of the RID program. Therefore, we calculated the ASRs from 2005 to 2019. For each operational definition, we computed the absolute mean difference between the ASR obtained using the operational definition and the KCCR's ASR using the formula below.

$$\text{Absolute mean difference} = \frac{\sum |(\text{ASR of KCCR}) - (\text{ASR of operational definition})|}{15 \text{ (years)}}$$

When the ASR based on an operational definition is similar to that of the KCCR, the absolute mean difference will approach 0.

Ethics Statement

The Institutional Review Board of Seoul National University College of Medicine/Seoul National University Hospital approved this study as exempt (approval No. E-2111-115-1273), as all analyses were conducted using publicly available data without any personally identifiable information.

RESULTS

Search Results and Characteristics

Figure 1 illustrates the study selection process. Initially, 310 papers were retrieved, with 300 from MEDLINE and 10 from KoreaMed. We removed 10 duplicate studies, leaving 300 studies. Two reviewers independently screened the titles and abstracts of these studies, resulting in the exclusion of 226 studies for the following reasons: not based on the NHIS database (n=194); utilization of data collected in 2002 or earlier (n=12); and secondary literature such as meta-analyses (n=20). Subsequently, we reviewed the full texts of the remaining 74 studies and excluded 12 articles due to the following reasons: lack of a definition for CRC (n=4); not based on the NHIS database (n=6); and utilization of data collected in 2002 or earlier (n=2). The characteristics of the included studies can be found in Supplemental Material 2.

Summary of Operational Definitions

The operational definitions utilized in studies with CRC as the outcome variable can be found in Table 1. A total of nine operational definitions for CRC were identified from the 62 eligible studies. The most frequently used operational definition was "C18-C20" (n=20), followed by "C18-C20 with claim code for treatment" (n=3) and "C18-C20 with V193 (code for registered cancer patients' payment deduction)" (n=3). Eight papers did not provide a specific definition of CRC. For colon cancer, 5 types of operational definitions were discovered in 11 studies. The most commonly used operational definition was "C18 including patients who underwent regional colectomy" (n=3), followed by "C18-C20", "C18-C20 and code for surgery," "C18-C20 and V193," and "C18-C21," each of which was

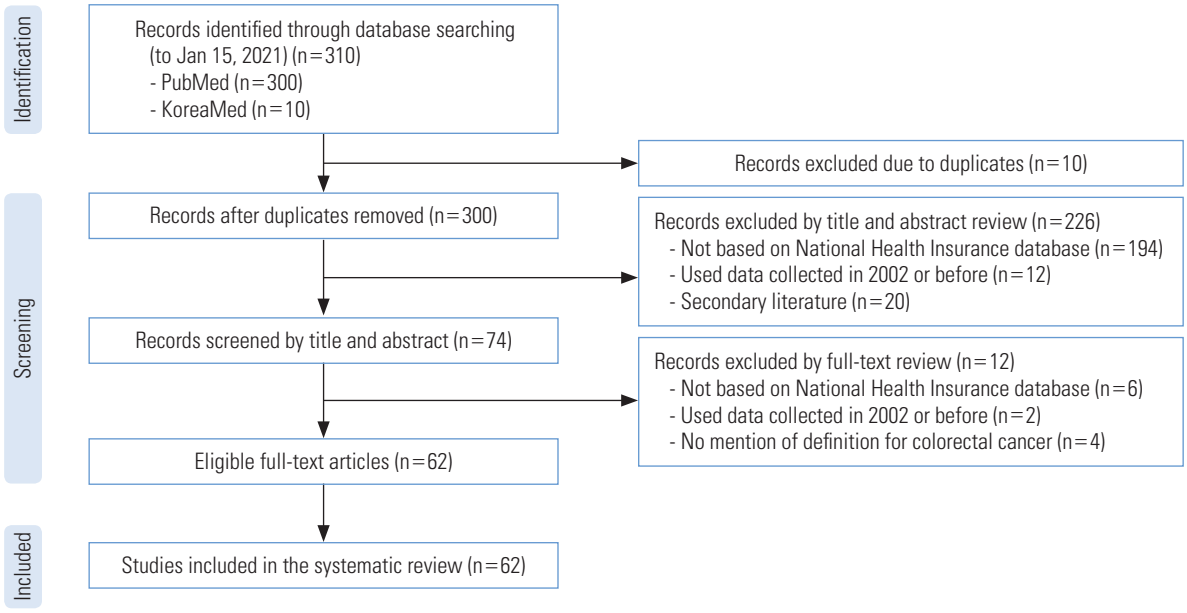

Figure 1. Flow chart of study selection for the review.

Table 1. Frequency of operational definitions used in studies

Outcome variables	Operational definition	Frequency
Colorectal cancer	C18-C20	20
	C18-C20 and code for treatment (surgery, chemotherapy, radiation therapy)	3
	C18-C20 and V193	3
	C18-C20 and code for surgery	2
	C18-C20 and code for radiation therapy	2
	C18-C20 and code for treatment or code for hospitalization	2
	C18-C20 and code for hospitalization	2
	C18-C20 or D010-D012 and code for colonoscopy or sigmoidoscopy	2
	C18-C21	2
	Others	3
	No specific definition	8
Total		49
Colon cancer	C18 including patients who went through regional colectomy	3
	C18-C20	1
	C18-C20 and code for surgery	1
	C18-C20 and V193	1
	C18-C21	1
	Others	2
	No specific definition	2
Total		11
Rectal cancer	C20	4
	No specific definition	1
Total		5

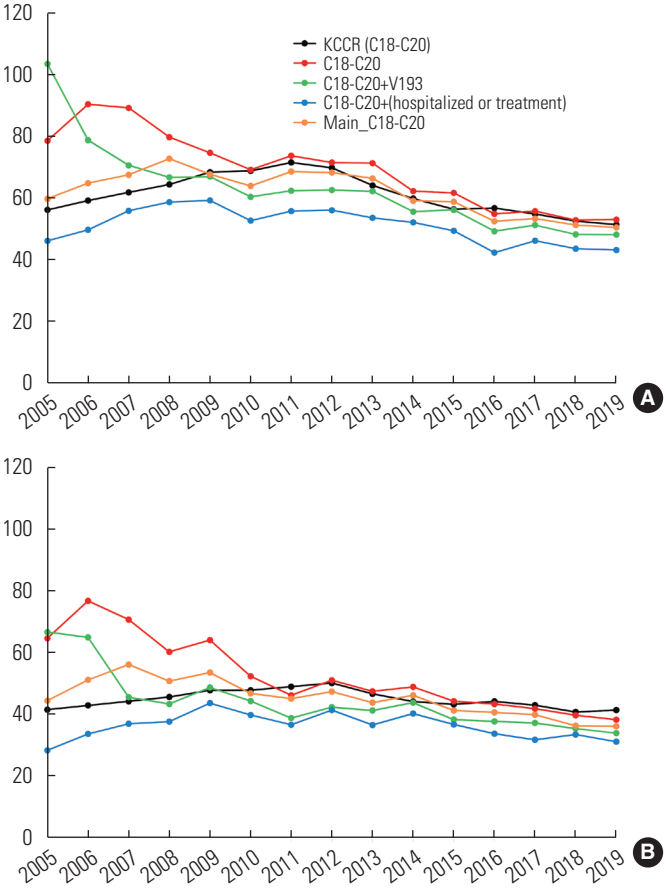

Figure 2. Age-standardized incidence rate of colorectal cancer according to four operational definitions and Korea Central Cancer Registry (KCCR) data. (A) Male. (B) Female.

used once. Lastly, "C20" (n=4) was the sole operational definition used in studies focusing on rectal cancer.

Comparison With Age-standardized Incidence Rates From the Korea Central Cancer Registry

The ASRs from the KCCR were generally higher than the ASRs based on operational definitions, with the exception of "C18-C20" (Figure 2 and Table 2). For both males and females, the smallest difference in ASRs was observed for "C18-C20 used as main diagnosis," followed by "C18-C20," "C18-C20 with V193," and "C18-C20 with claim code for hospitalization or code for treatment." The smallest absolute mean difference, obtained using the ASR based on "C18-C20 used as main diagnosis" and the KCCR's ASR, was 3.1/100 000 for males and 4.3/100 000 for females. However, the ranking of operational definitions that yielded the smallest absolute mean difference for males differed from that observed for females. The next top 3 operational definitions with the smallest absolute mean differences were "C18-C20" (8.4/100 000), "C18-C20 with V193" (8.6/100 000), and "C18-C20 with claim code for hospitalization or code for treatment" (10.1/100 000) for males, and "C18-C20 with V193" (7.2/100 000), "C18-C20 with claim code for hospitalization or code for treatment" (8.7/100 000), and "C18-C20" (9.0/100 000) for females.

The trends in ASRs of CRC over time were as follows: the ASR from the KCCR consistently increased until 2011, after which it began to decrease. Following a local peak in 2016, the decline continued up to the present time. The temporal trend in ASR, based on each of the 4 operational definitions, exhibited an initial increase followed by a subsequent decrease since 2011, with minor fluctuations observed from year to year. Among these definitions, the ASR according to "C18-C20 used as the main diagnosis" demonstrated a similar pattern for both sexes as that of the KCCR.

Table 2. Age-standardized incidence rates of colorectal cancer according to operational definitions and KCCR data, per 100 000 persons

Operational definition	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	Absolute mean difference
Male																
KCCR (C18-C20)	56.4	59.4	62.0	64.6	68.6	69.0	71.7	70.0	64.3	60.0	56.6	57.0	55.0	52.8	51.6	0.0
C18-C20: main diagnosis	60.0	65.0	67.7	73.0	67.9	64.1	68.8	68.5	66.6	59.3	59.0	52.7	53.6	51.5	50.7	3.1
C18-C20	78.8	90.6	89.4	79.9	74.9	69.4	73.9	71.7	71.5	62.5	61.9	55.1	56.0	53.0	53.3	8.4
C18-C20+V193	103.7	79.0	70.8	66.9	67.2	60.6	62.6	62.8	62.4	55.8	56.4	49.5	51.4	48.4	48.4	8.6
C18-C20+(hospitalization or treatment)	46.4	49.9	56.1	58.9	59.4	52.9	56.0	56.3	53.8	52.3	49.6	42.5	46.4	43.8	43.4	10.1
Female																
KCCR (C18-C20)	41.6	42.9	44.2	45.6	47.8	47.8	49.0	50.1	46.7	44.2	43.3	44.2	43.0	40.8	41.4	0.0
C18-C20: main diagnosis	44.5	51.2	56.1	50.8	53.5	46.8	45.1	47.4	43.8	46.2	41.3	40.7	39.9	36.4	36.2	4.3
C18-C20	64.5	76.6	70.6	60.2	64.0	52.3	46.2	51.1	47.5	48.9	44.3	43.4	41.9	39.8	38.3	9.0
C18-C20+V193	66.6	64.8	45.6	43.4	48.7	44.3	38.8	42.4	41.3	43.8	38.4	37.8	37.3	35.5	34.0	7.2
C18-C20+(hospitalization or treatment)	28.5	33.8	37.0	37.7	43.7	39.9	36.7	41.4	36.6	40.3	36.8	33.8	31.9	33.6	31.3	8.7

KCCR, Korea Central Cancer Registry.

DISCUSSION

In this study, we observed that the ASRs from the KCCR were generally higher than those based on operational definitions, with the exception of the C18-C20 code for main and subdiagnoses. For both males and females, the incidence rate for the main diagnosis with the C18-C20 code was most similar to that of the KCCR data, which is considered the gold standard in cancer incidence statistics. Relying solely on diagnosis codes for analysis may lead to an overestimation of cancer incidence; however, the degree of overestimation was significantly reduced when using only inpatient data to estimate ASRs.

Prior to 2010, there was a larger discrepancy between the ASRs derived from operational definitions and the ASR from the KCCR than in the period after 2010 (Figure 2). It is worth noting that significant events, such as the expansion of the RID program in 2005 and the establishment of the hospital accreditation program by the Korea Institute for Healthcare Accreditation in 2010, may have impacted the claims process in the NHIS [5,9]. Consequently, the incidence of false-positive CRC cases decreased over time, resulting in a convergence of ASRs.

C18-C20 and treatment codes were utilized in 7 studies to define CRC patients. These studies used detailed definitions to identify patients and enhance accuracy in evaluating the association between exposures and clinical outcomes of CRC. Three studies used operational definitions based on surgery, chemotherapy, and radiation therapy, which were derived from the NHIS sample cohort or customized database [10-12]. Two studies used subcodes of surgery only, while 2 others focused on the utilization of radiation therapy in cancer patients [13-16]. In contrast, 2 studies used C18-C20 in combination with claim codes for hospitalization or treatment codes as operational definitions of CRC; these studies used tailored definitions to capture patients who received cancer treatment or were prescribed specific medications [17,18]. Lee et al. [11] defined CRC cases as subjects who simultaneously had both the main diagnosis code of C18-C20 and the claim code for its treatment, using the NHIS-NSC of 2002-2015 period in their analysis. They demonstrated that the ASR of CRC from this approach was lower than that of the KCCR, but the trends over time were similar. Although the ASRs obtained from the operational definitions used in these studies were lower than those from the KCCR, it is important to consider that these studies primarily aimed to construct a patient cohort and identify po-

tential prognostic factors associated with CRC. Therefore, those operational definitions may be considered optimal in achieving the research goal, despite the studies reporting overall lower ASRs compared to the ASR from the KCCR.

There are several limitations to this study. First, unlike a previous study conducted in collaboration between KCCR and NHIS, individual CRC cases identified in the NHIS-NSC were not matched to the actual cancer patients recorded in the national cancer registry. However, this study still offers a meaningful analysis by comparing the ASR obtained from the NHIS to the gold standard of ASR from the KCCR. Second, 92% of the studies (n=57) included in our analysis did not specify the diagnosis code used or were unclear about whether it pertained to inpatient or outpatient data. Nevertheless, we obtained diagnosis codes from both inpatient and outpatient cases and defined hospitalized patients using a specific code in our current study. Finally, since the inception of the RID program in 2005, cancer patients were retroactively registered, resulting in the low accuracy of the "C18-C20 with V193" during 2005-2006. However, an analysis from 2007 onwards demonstrated that the ASR of "C18-20 with V193" was comparable to the ASR of other operational definitions.

The main strength of our study is that we conducted an extensive review of published CRC studies and calculated the ASR of CRC based on operational definitions found in the literature. Additionally, we utilized a representative nationwide database for the analysis of ASR. While the NHIS-NSC may not reflect real-time cancer incidence, it is widely accepted that its data represent population-wide statistics, as the cohort was generated through random sampling of the general Korean population [6]. In this context, our study could provide valuable guidance for researchers working with operational data in the NHIS database.

In conclusion, the ASRs for CRC defined by the operational definition "C18-C20 used as the main diagnosis" in the NHIS database were most closely aligned with those obtained from the KCCR. Operational definitions that rely on hospitalization or treatment details may underestimate the incidence of CRC compared to using only diagnosis codes and could be used depending on the study hypothesis.

SUPPLEMENTAL MATERIALS

Supplemental materials are available at https://doi.org/10.3961/jpmph.23.033.

CONFLICT OF INTEREST

The authors have no conflicts of interest associated with the material presented in this paper.

FUNDING

This research was supported by a grant of the MD-PhD/Medical Scientist Training Program through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea. This research was also supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A2C1004608).

ACKNOWLEDGEMENTS

None.

AUTHOR CONTRIBUTIONS

Conceptualization: Park H, Kim YR, Pyun Y, Shin A. Data curation: Park H, Kim YR, Pyun Y, Joo H, Shin A. Formal analysis: Park H, Kim YR. Funding acquisition: Park H, Shin A. Methodology: Park H, Kim YR, Pyun Y, Joo H, Shin A. Visualization: Park H. Writing – original draft: Park H. Writing – review & editing: Park H, Kim YR, Pyun Y, Joo H, Shin A.

ORCID

Hyeree Park *https://orcid.org/0000-0003-1551-260X*
Yu Rim Kim *https://orcid.org/0000-0002-8547-7902*
Yerin Pyun *https://orcid.org/0000-0002-9672-2061*
Hyundeok Joo *https://orcid.org/0000-0002-6134-8815*
Aesun Shin *https://orcid.org/0000-0002-6426-1969*

REFERENCES

1. Korea Central Cancer Registry. Annual report of cancer statistics in Korea in 2019. Sejong: Ministry of Health and Welfare; 2021 (Korean).
2. Vital Statistics Division. Annual report on the causes of death statistics in 2021. Daejeon: Statistics Korea; 2022 (Korean).
3. Kang MJ, Won YJ, Lee JJ, Jung KW, Kim HJ, Kong HJ, et al. Cancer statistics in Korea: incidence, mortality, survival, and prev-

alence in 2019. Cancer Res Treat 2022;54(2):330-344.
4. Seong SC, Kim YY, Khang YH, Park JH, Kang HJ, Lee H, et al. Data resource profile: the National Health Information Database of the National Health Insurance Service in South Korea. Int J Epidemiol 2017;46(3):799-800.
5. Yang MS, Park M, Back JH, Lee GH, Shin JH, Kim K, et al. Validation of cancer diagnosis based on the National Health Insurance Service Database versus the National Cancer Registry Database in Korea. Cancer Res Treat 2022;54(2):352-361.
6. Kim DW, Lee SM, Lim HS, Choi JK, Park HY, Yuk TM, et al. A study on the manipulative definition of disease based on National Healthcare Insurance claims. Goyang: National Healthcare Insurance Service Ilsan Hospital; 2017 (Korean).
7. Lee J, Lee JS, Park SH, Shin SA, Kim K. Cohort profile: the National Health Insurance Service-National Sample Cohort (NHIS-NSC), South Korea. Int J Epidemiol 2017;46(2):e15.
8. Statistics Korea. Cancer incident cases and incidence rates by 61 sites, sex and 5-year age group [cited 2022 Aug 6] Available from: https://kosis.kr/statHtml/statHtml.do?orgId=117&tblId=DT_117N_A0024&conn_path=I3 (Korean).
9. Lee S. Healthcare accreditation in Korea: the current status and challenges ahead. Health Policy Manag 2018;28(3):251-256 (Korean).
10. Choe S, Lee J, Park JW, Jeong SY, Cho YM, Park BJ, et al. Prognosis of patients with colorectal cancer with diabetes according to medication adherence: a population-based cohort study. Cancer Epidemiol Biomarkers Prev 2020;29(6):1120-1127.
11. Lee J, Choe S, Park JW, Jeong SY, Shin A. The risk of colorectal cancer after cholecystectomy or appendectomy: a population-based cohort study in Korea. J Prev Med Public Health 2018;51(6):281-288.
12. Jang D, Choe S, Park JW, Jeong SY, Shin A. Smoking status before and after colorectal cancer diagnosis and mortality in Korean men: a population-based cohort study. Cancer Med 2020;9(24):9641-9648.
13. Kang JK, Kim MS, Jang WI, Seo YS, Kim HJ, Cho CK, et al. The clinical utilization of radiation therapy in Korea between 2009 and 2013. Radiat Oncol J 2016;34(2):88-95.
14. Kim E, Jang WI, Kim MS, Paik EK, Kim HJ, Yoo HJ, et al. Clinical utilization of radiation therapy in Korea, 2016. J Radiat Res 2020;61(2):249-256.
15. Rim CH, Kim CY, Yang DS, Yoon WS. Clinical significance of gender and body mass index in Asian patients with colorectal cancer. J Cancer 2019;10(3):682-688.
16. Song N, Huang D, Jang D, Kim MJ, Jeong SY, Shin A, et al. Optimal body mass index cut-off point for predicting colorectal cancer survival in an Asian population: a national health information database analysis. Cancers (Basel) 2020;12(4):830.
17. Kim YA, Lee YR, Park J, Oh IH, Kim H, Yoon SJ, et al. Socioeconomic burden of cancer in Korea from 2011 to 2015. Cancer Res Treat 2020;52(3):896-906.
18. Hwang IC, Chang J, Park SM. Emerging hazard effects of proton pump inhibitor on the risk of colorectal cancer in low-risk populations: a Korean nationwide prospective cohort study. PLoS One 2017;12(12):e0189114.