JOURNAL OF CANCER PREVENTION
VOL. 23, NO. 4, DECEMBER 2018

https://doi.org/10.15430/JCP.2018.23.4.183
pISSN 2288-3649 · eISSN 2288-3657
www.jcpjournal.org

ORIGINAL ARTICLE

Validation of Administrative Big Database for Colorectal Cancer Searched by International Classification of Disease 10th Codes in Korean: A Retrospective Big-cohort Study

Young-Jae Hwang[1], Nayoung Kim[1,4,5], Chang Yong Yun[1], Hyuk Yoon[1], Cheol Min Shin[1], Young Soo Park[1], Il Tae Son[2], Heung-Kwon Oh[2], Duck-Woo Kim[2], Sung-Bum Kang[2], Hye Seung Lee[3], Seon Mee Park[6], Dong Ho Lee[1,4,5]

Departments of [1]Internal Medicine, [2]Surgery, and [3]Pathology, Seoul National University Bundang Hospital, Seongnam, [4]Department of Internal Medicine and Institute of Liver Research, Seoul National University College of Medicine, [5]Tumor Microenvironment Global Core Research Center, Seoul National University, Seoul, [6]Department of Internal Medicine, Chungbuk National University College of Medicine and Medical Research Institute, Cheongju, Korea

Background: As the number of big-cohort studies increases, validation becomes increasingly more important. We aimed to validate administrative database categorized as colorectal cancer (CRC) by the International Classification of Disease (ICD) 10th code.
Methods: Big-cohort was collected from Clinical Data Warehouse using ICD 10th codes from May 1, 2003 to November 30, 2016 at Seoul National University Bundang Hospital. The patients in the study group had been diagnosed with cancer and were recorded in the ICD 10th code of CRC by the National Health Insurance Service. Subjects with codes of inflammatory bowel disease or tuberculosis colitis were selected for the control group. For the accuracy of registered CRC codes (C18-21), the chart, imaging results, and pathologic findings were examined by two reviewers. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for CRC were calculated.
Results: A total of 6,780 subjects with CRC and 1,899 control subjects were enrolled. Of these patients, 22 subjects did not have evidence of CRC by colonoscopy, computed tomography, magnetic resonance imaging, or positron emission tomography. The sensitivity and specificity of hospitalization data for identifying CRC were 100.00% and 98.86%, respectively. PPV and NPV were 99.68% and 100.00%, respectively.
Conclusions: The big-cohort database using the ICD 10th code for CRC appears to be accurate.
(J Cancer Prev 2018;23:183-190)

Key Words: Database, Validation, Colorectal neoplasms, International Classification of Disease

INTRODUCTION

Recently, medical technology has rapidly developed in various areas, including imaging and surgical technology and medical database technology. With the development in medical technology for the use, storage, and analysis of patients' information, administrative databases that include personal information, diagnosis, laboratory results, and treatment modalities of healthcare centers have also rapidly increased worldwide. Most population-based administrative big-cohort data are valuable for research and disease surveillance.[1,2] Consequently, physicians are showing considerable interest in the use of administrative data to identify the burden of comorbidities in patients with various diseases, such as cancer,[3,4] infection,[5-7] and rare conditions including peptic ulcer disease complications[8,9] and connective tissue disease.[6] Diagnosis of a patient's disease is usually stored as diagnosis codes according to the International Classification of Diseases, 10th Revision edition (ICD 10th code) published by the

Received December 2, 2018, **Revised** December 14, 2018, **Accepted** December 18, 2018

Correspondence to: Nayoung Kim
Department of Internal Medicine, Seoul National University Bundang Hospital, 82 Gumi-ro 173beon-gil, Bundang-gu, Seongnam 13620, Korea
Tel: +82-31-787-7008, Fax: +82-31-787-4051, E-mail: nayoungkim49@empas.com, ORCID: Nayoung Kim, https://orcid.org/0000-0002-9397-0406

World Health Organization[10,11] to provide an easy way to access information for practical research. The Korea National Health Insurance System (NHIS) contains a complete set of health information pertaining to 50 million Koreans, including an eligibility database (age, sex, socioeconomic variables, type of eligibility, income level, etc), a medical treatment database (based on medical bills claimed by medical service providers for their medical expense claims), a health examination database (results of general health examinations and questionnaires on lifestyle and behavior), and a medical care institution database (type of medical care institution, location, equipment, and number of physicians). The source population of the NHIS is the Health Insurance Review and Assessment database, which includes all insurance claims information of approximately 97% of the Korean population.

The administrative database used in these studies requires high sensitivity or a positive predictive value (PPV). Especially, the reliability of the registered ICD 10th code varies among healthcare centers. Thus, adequate validation of administrative data about diagnosis is important. The validity of the claims database has been questioned due to inaccurate or incomplete coding.[12] Reliability of administrative database was important as studies using administrative database were increasing. However, the numbers of papers or specific research methods for evaluating the validation of these administrative databases are insufficient.[13-15] Direct validation is nearly impossible due to the Personal Information Protection Actin Korea. A similar situation can also be found in other countries. However, validation can be performed at the hospital at which the diagnosis of each disease is made and reported to the Health Insurance Review and Assessment database for insurance claims. Adequate design is important for accurate validation of certain diseases. The disease to be validated should also have high prevalence and incidence rate. In addition, a simple diagnostic method of this disease is preferred. Colorectal cancers (CRCs) account for approximately 9.4% of the total worldwide cancer cases, with about 1 million new cases diagnosed annually.[16] The incidence of CRC shows wide geographical variation, with higher rates observed in Australia, North America, Europe, Japan, and Korea. Actually, CRC develops as a consequence of genetics, diet, or obesity. These factors reflect the development of each country[3] and cause morbidity and mortality.[17-19] In addition, gender-specific medicine is used for this disease.[20,21] Also CRC could be diagnosed in primary clinics and secondary hospitals in Korea where the use of colonoscopy is very popular for screening and surveillance purposes. Simple and common diagnostic method of CRC could be strength as an object of validation study. But there is no validation study about diagnosis of CRC. So we planned to identify validation of CRC diagnosis.

The hospital at which the validation is to be performed also requires two conditions: high burden of CRC patients and well established Clinical Data Warehouse (CDW) system. Seoul National University Bundang Hospital (SNUBH) might be an appropriate hospital because it has been utilizing in-house developed comprehensive Electronic Medical Record (EMR) since 2003 and is the first full digitized paperless hospital in South Korea. It was also accredited as a stage 7 recipient hospital by Healthcare Information and Management Systems Society (HIMSS) in 2010, which was the first case outside the United States. This HIMSS system has been exported to several countries such as Saudi Arabia, United Arab Emirates, and United States owing to its integrity and excellence.[22]

With such background information, the aims of this study are to estimate the sensitivity and specificity of CRC case ascertainment based on the administrative database. This study would support the validity of other studies using administrative database.

MATERIALS AND METHODS

1. Data source

We conducted a retrospective hospital-based study using SNUBH CDW from May 1st, 2003 to December 31st, 2016. It is hard for Administrative database of SNUBH to represent the Korean population. However, this kind of study is rather hard to be performed in case of hospitals where the CDW is not so well arranged. In addition, SNUBH is a tertiary hospital located near Seoul in South Korea to which primary or secondary hospitals refer cancer patients to confirm a diagnosis of cancer and obtain treatment. And subjects from various regions visit SNUBH for cancer screening evaluation with colonoscopy. Maybe this kind of analysis will be helpful for the general hospitals in Korea. From this reason we chose the SNUBH database even there is a limitation to represent the Korean population.

Medical records of all patients were collected using SNUBH's CDW[23] and EMR including the visiting hospital department, the principal diagnosis, and the surgical and diagnostic procedures for each patient. In addition, pathologic finding and imaging modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) were included. And other hospital medical charts were reviewed through up-loaded data file.

2. Study population

After approval of this study protocol by the Ethics Committee at SNUBH (IRB No. B-1701/378-105), the CRC patients' list was requested to CDW using the following ICD 10th code as the primary diagnosis: (1) C18, (2) C18.0-18.8, (3) C19, (4) C20, and (5) C21 (Supplementary Table 1).[11] In addition, the searched cases were confirmed to be registered as V code by the NHIS.[3,24,25] V code is the code system specific to South Korea. It is issued for cancer patients diagnosed with a cancer and recorded in the ICD 10th code of cancer. In South Korea, cancer patients registered with the NHIS receive V codes. With such codes, 95% of the medical cost is supported by the government for 5 years. Since the V code system was established by the Korean Ministry of Health and Welfare in 2008, the accuracy of cancer diagnosis using both the ICD 10th code and V code is very high. To increase the early detection rate of five important cancers including CRC, NHIS recommends a semi-compulsive cancer screening checkup to be undertaken biennially without charge. Thus, NHIS has plentiful resources. Several reports have been published using this NHIS database.[3,4,8,9,24,25] However, in the absence of the V code system, confirmation of CRC should be carefully made as undertaken in the present study.

To calculate the sensitivity and specificity of CRC, the control group comprised of patients with cancers according to the following ICD 10th code as primary diagnosis: (1) K50-51 (inflammatory bowel disease, IBD) and (2) A1831 (tuberculosis colitis) (Supplementary Table 2).[11] Patients in the control group were not registered with the following ICD 10th codes: (1) C18, (2) C18.0-C18.8, (3) C19, (4) C20, or (5) C21. Sometimes, CRC patients are miss-diagnosed as having IBD and the work up pathway for these diseases is similar to that of CRC. IBD and tuberculosis colitis are confirmed through diagnostic pathways including colonoscopy, CT, MRI, and surgical procedure, similar to CRC. Supplementary Table 1 and 2 show the ICD 10th code for each group.

3. Study algorithm and method of identifying validation by two reviewers

Validation of the ICD 10th code with CRC was proved as follows: 1) detection of neoplastic lesions with colon, rectum or anus and documented with endoscopic findings or image (CT, MRI, or PET) readings, and 2) histologic examination for adenocarcinoma or squamous cell carcinoma from primary or metastatic sites. Specimens for histologic diagnosis were mainly obtained through endoscopy, surgical specimen, or fine needle aspiration.

We analyzed medical charts of cases from the administrative database and obtained the chart number, names of subjects, dates of hospital visits, diagnostic procedures, and types of surgery from medical charts to find coded cases with false CRC (false positive) and uncoded cases with true CRC (false negative).

We planned that two reviewers analyzed all the medical charts of the study control group to validate the accuracy of the data registered in the ICD 10th code. Reviewers were two gastroenterology fellows who had received Board of Internal Medicine and at the trainee course at the division of Gastroenterology. They determined the presence or absence of CRC evidence by considering the following: (1) documentation of imaging (CT or MRI) or endoscopy of the mass or nodule lesions in the colon and rectum, and (2) documentation of histological finding about malignancy (adenocarcinoma and squamous cell carcinoma) from the colon, rectum, or metastatic organ. If there was difference decision of reviewers, they analyzed again that case, discussed and made a final decision. All cases of absence of CRC were reanalyzed until consensus was reached. After reviewing the chart and grouping of each subject, the sensitivity, specificity, PPV, and negative predictive value (NPV) (with 95% CI) were calculated.

RESULTS

1. Proposed study algorithm for the inclusion and classification of colorectal cancer patients

We collected subjects of study group carefully as conformingenroll criteria. Two reviewers examined the following criteria: 1) the registered number of ICD 10th codes with CRC and V code, 2) diagnostic record including endoscopy, imaging and histologic examination, or 3) other hospital medical charts. We enrolled subjects who fulfilled these criteria. A total of 6,780 subjects were found to be registered as having CRC at the SNUBH between May 1st, 2003 and December 31st, 2016 (Table 1). Initially, for economy of time, we planned to select ($\alpha = 0.05$, $1-\beta =$

Table 1. Baseline characteristics of study subjects in control group and study group

Characteristic	Control group (n = 1,899)	Study group (n = 6,780)
Sex (male : female)	1,119 : 780	4,058 : 2,743
Mean age at diagnosis	45.59 ± 17.37	61.98 ± 12.07

Values are presented as number only or mean ± SD.

0.95, and effect size 0.1) 1,000 subjects randomly in the 6,780 subjects. However, to increase the confidence of our study and exclude sampling error, we changed our plan to identify all included patients of both the study and control group (Fig. 1). Analysis of all subjects of the study was completed in 10 months (between January 1st, 2017 and October 30th, 2017).

Medical charts of pathologic findings, imaging test (CT, MRI, and PET), and other hospitals' results in 6,780 patients were analyzed based on pathologic orders (Supplementary Table 3). A total of 21,305 pathologic orders were identified, 12,548 of which were found to be adenocarcinoma or squamous cell carcinoma. We deleted the overlapping results in each subject and a total of 6,657 subjects at SNUBH had pathologic findings of CRC including adenocarcinoma (n = 6,596, 99.08%) and squamous cell carcinoma (n = 61, 0.92%) (Table 2). We analyzed remaining subjects. In the remaining 123 subjects, a total of 101 patients had pathologic diagnosis of CRC from the other hospital's results including adenocarcinoma (n = 97, 96.04%) and squamous cell carcinoma (n = 4, 3.96%) (Table 2). A total of 6,758 patients were identified as actual CRC patients with accurate pathologic finding at SNUBH or other hospitals. After carefully reviewing of the medical charts by the two reviewers, 22 patients underwent pathologic diagnostic procedures but showed no evidence of pathologic finding of CRC from SNUBH or other hospitals. These 11 false positive patients were found to have another type of primary origin cancer (gastric cancer, prostate cancer, and cervical cancer) with colonic metastasis. In addition, 6 patients were referred from another hospital because of malignancy suspicious lesion with endoscopic finding. However, the pathologic finding of the specimen obtained by endoscopic mucosal resection at SNUBH was adenoma rather than cancer. In addition, 3 patients had insufficient pathologic finding to diagnose CRC (atypical cell). They were advised to receive endoscopy again to obtain more specimens but did not visit SNUBH. One patient showed possible CRC but ischemic colitis was confirmed after the diagnosis process. One patient was diagnosed with cancer of an unknown origin. To receive support by the NHIS, this patient was classified as having CRC according to the ICD 10th code of CRC.

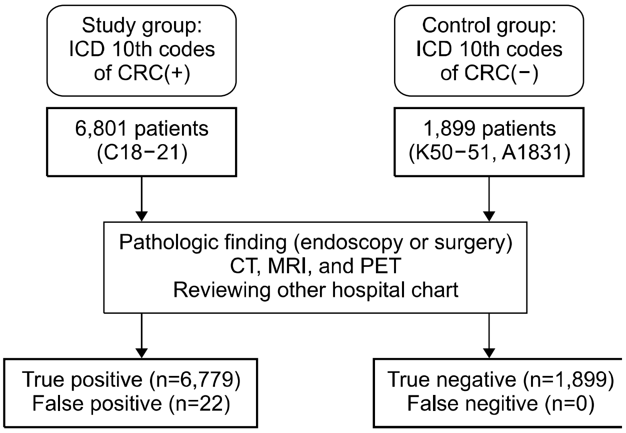

Figure 1. Proposed study algorithm for the inclusion and classification of subjects. ICD, International Classification of Disease; CRC, colorectal cancer; CT, computed tomography; MRI, magnetic resonance imaging; PET, positron emission tomography.

2. Classification of control group

We searched the control group who had not been registered with the ICD 10th code of CRC (Fig. 1). The control group subjects showed endoscopic or imaging findings of IBD or tuberculosis colitis (K50-51 and A1831) (Supplementary Table 2). In practice, they underwent pathologic diagnostic procedure to obtain colon or rectum specimens for diagnosis and to rule out CRC. Similar to the CRC group, two reviewers checked closely and identified the enroll criteria for the control group. A total of 1,899 subjects were identified as the control group (Table 1). When a total of 8,424 pathologic medical orders were checked for CRC (Supplementary Table 1), only 2 cases had pathologic findings of intramucosal adenocarcinoma, well differentiated, limited in lamina propria (pTis) in the rectum. They visit SNUBH through other hospital to remove adenomatous polyp. There is some controversy, but intramucosal adenocarcinoma was classified as dysplasia, not cancer.[26-28] So these two cases defined as not false negative cases. So in control group, there was no false negative case.

3. Statistical analysis result

The results of the administrative database with the ICD 10th code for CRC are shown in Table 3. Calculated statistic values are summarized in Table 4. The sensitivity and specificity of ICD 10th code of CRC in the administrative medical database of SNUBH

Table 2. Study results of administrative database with ICD 10th code of colorectal cancer with evidence of pathologic finding. We classified these subjects by pathologic findings

Variable	Adenocarcinoma	Squamous cell carcinoma	Total
Total	6,693 (99.04)	65 (0.96)	6,758
Diagnosis in SNUBH	6,596 (99.08)	61 (0.92)	6,657
Diagnosis in other hospital	97 (96.04)	4 (3.96)	101
Sex (male : female)	3,999 : 2,594	39 : 26	
Mean age at diagnosis	61.96 ± 12.06	35.14 ± 8.03	

Values are presented as number (%) or mean ± SD. ICD, International Classification of Disease; SNUBH, Seoul National University Bundang Hospital.

Table 3. Study results of administrative database with ICD 10th code of colorectal cancer

Variable		Status of colorectal cancer		Total
		Positive	Negative	
ICD codes of colorectal cancer	Outcome positive	True positive 6,758	False positive 22	6,780
	Outcome negative	False negative 0	True negative 1,899	1,899
	Total	6,779	1,921	8,700

ICD, International Classification of Disease.

Table 4. Statistical analysis results of administrative database with ICD 10th code of colorectal cancer

Variable	Point estimate (%)	95% CI
Sensitivity	100.00	100.00-100.00
Specificity	98.86	98.38-99.33
Positive predictive value	99.68	99.54-99.81
Negative predictive value	100.00	100.00-100.00

ICD, International Classification of Disease.

were 100.00% (95% CI: 100.00-100.00) and 98.86% (95% CI: 98.38-99.33), respectively. The PPV and NPV were 99.68% (95% CI: 99.54-99.81) and 100.00% (95% CI: 100.00-100.00), respectively.

DISCUSSION

Recently, several reports have been presented in which a big-cohort database was used, especially in Taiwan[29,30] and South Korea,[3,4] where the NHIS is well established. When validating the ICD 10th code for CRC in the administrative big-cohort database for the diagnosis of CRC, the sensitivity, specificity, PPV, and NPV were all close to 100%. To the best of our knowledge, this is the first report regarding this type of validation for such a long period (13.5 years) with 6,780 patients of the CRC group and 1,899 patients of the control group.

The administrative medical database is a very useful resource for researchers, especially when a randomized control study cannot be performed for economic or practical reasons. Actually, the randomized control study has several limitations because the protocol cannot be changed and other factors such as pharmacologic values cannot be added. In contrast, various other values can be easily added during an ongoing study in the administrative database. Furthermore, an administrative big-cohort database study can verify the tendency of risk factors, cancer location, diagnostic rate, and prevalence of diseases by using the ICD 10th code or V code in South Korea. From such analysis, researchers can easily obtain information about treatment trends and prognoses. However, a significant limitation of the administrative database is the degree of imprecision of the diagnostic information. As these data are stored by doctors using the ICD 10th code, for any study, the validity of the database should be checked before using the administrative database of the ICD 10th code. If the validation study reveals high accuracy of CRC diagnosis in the ICD 10th code which is the condition for the research, the reliability of the study on the trend of case information, CRC development, treatment, and prognosis will be verified. In addition, the validation can provide cancer related outcomes, geographic and private variations, and cancer related costs.

Several studies have been conducted around the world on the validation of administrative databases. In 2008 in Turin, Italy, Ileana Baldi et al.[13] performed validation of the ICD 9th code including lung cancer, breast cancer, and CRC. They corrected the study group cases by searching the ICD 9th code and compared the data with that from the Piedmont Cancer Registry of Turin.[13] Through an algorithm, they calculated the sensitivity (lung cancer: 80.8%; breast cancer: 76.7%; CRC: 72.4%, respectively) and PPV (lung cancer: 78.7%; breast cancer: 87.9%; CRC: 92.6%, respectively). Penberthy et al.[31] also examined the diagnosis codes from inpatient data of a Virginia hospital (based on standard universal billing forms) to validate central cancer register reporting. They searched cases from the Virginia Cancer Registry and compared these with cases collected from the Virginia statewide hospital discharge file.[31] Through an algorithm, they reported PPV (breast cancer: 98%; cervical cancer: 86%; CRC: 95%; lung cancer: 96%; prostate cancer: 94%, respectively). Furthermore, Ganry et al.[14] validated the Programme de médicalisation des systèmes d'information (PMSI) as an independent source to identify incident cancer cases by using the French hospital database adapted from the Diagnosis Related Group classification. Through their algorithm, they reported that the PMSI database had a sensitivity of 85%, a specificity of 99.9%, and a PPV of 97% for breast cancer. In Denmark, Rostgaard et al.[15]

validated data in a clinical database and showed that 78% of breast cancer patients were registered on both the traditional national cancer register and the national clinical database. In these studies, administrative data were validated using various algorithms,[13-15,31] where by most of the studies validated the clinical database compared with the cancer register organization database. These studies investigated hospital database or insurance database compared to registries to ascertain the selected diseases by calculating the sensitivity, specificity, PPV, and NPV depending on the definitions used and the disease under evaluation.

Our study compensated the defect of earlier mentioned studies to enhance reliability of study. Different from other studies that used only one diagnostic code, we used the CDW system of a tertiary hospital (SNUBH) that had both the ICD 10th code of CRC and the V code of CRC in the NHIS for accurate study algorithm. Second, to validate the ICD 10th code of CRC, both the outpatient records and hospital medical records including admission notes, progress notes, colonoscopy findings, imaging readings, and pathologic records were verified for all 6,780 patients of the CRC group and 1,899 patients of the control group (Fig. 1). This analysis was time-consuming and needed considerable effort in the examination of all the medical records of each patient. Furthermore, we investigated other hospital medical records that had been formally uploaded on the SNUBH's EMR. Third, for an accurate algorithm, we established strictly enroll criteria. In fact, such enroll criteria were not found in other studies except in the ICD codes.[6] To obtain reliable data we examined subjects had adequate medical records about the evaluation of CRC. The study algorithm in the present study could provide an example for similar studies in which other administrative databases are used.[13-15,31] Furthermore, we set up a control group of patients with IBD or tuberculosis of colon to calculate the statistical values (sensitivity, specificity, and NPV) which could not be calculated without control group.

After this preparation procedure, the sensitivity and specificity of the ICD 10th codes of CRC in the administrative medical data of SNUBH were found to be 100.00% and 98.86%, respectively. Furthermore, the PPV and NPV were 99.68% and 100.00%, respectively (Table 4). The sensitivity and specificity of our study were higher than those of other studies.[13,14,31] This suggests the high validity of administrative data in the ICD 10th code of CRC, proving that administrative data on CRC patients are useful for obtaining epidemiologic or medical services information. If some researches about CRC using administrative database is done, this result could be the basis for validity of these researches. The number of enrolled CRC patients (n = 6,780) in the present study is compatible with that in a previous study emphasizing the importance of a large study sample size to enhance the sensitivity and specificity of validation.[32] A large study sample size is needed to achieve high sensitivity and specificity. And V code, specific diagnostic code in South Korea, maybe contributed accuracy of ICD 10th codes of CRC.

Despite these strengths, our study also has some limitations. First, the number of false positive patients was found to be 22, indicating a lower specificity than sensitivity. These results can be explained by the following factors. Before the South Korea government started to subsidize 90% of cancer patients' medical fees in 2008, the system of V code was established. This V code is used in all hospitals. Even if the application of V code is incorrect, the claim for the doctor who inserted this code would not be significant. Thus, several false positive cases of CRC could have occurred when the doctor registered the V code at the NHIS with the ICD 10th code based on highly probable colonoscopy finding of CRC prior to the formal histologic diagnosis of CRC at SNUBH. Similarly, CRC patients could visit another tertiary hospital where they could be diagnosed with CRC without being required to consider the SNUBH database (follow-up loss bias). However, our study was a retrospective study for the duration of 13.5 years. We could examine other hospitals' medical results when they had been uploaded into the SNUBH's data base in any format. And it might be possible that patients who were treated for CRC in other hospitals were included in the false positive patients. These patients were registered with the ICD 10th code of CRC without evidence of CRC at SNUBH. We could not determine the patients' history of CRC if they did not provide medical reports from other hospitals. The specificity would be increase if this weakness was overcome. And patients with cancer of an unknown origin such as adenocarcinoma could have been included in the false positive subjects. If doctors suspect CRC to be the most likely diagnosis of cancer, they will sometimes register the case with the ICD 10th code of CRC without pathologic evidence of CRC. However, these cases are very rare and did not have a high impact on the present study because of the large number of patients enrolled 6,780.

Second limitation is that our study was done from only single tertiary referral center due to the Personal Information Protection Actin Korea. To reduce inter-observer variability like pathologic finding and imaging reading, we planned our study that attended restricted researchers in SNUBH. We could not get other hospital database of CRC, study subjects were limited who visit SNUBH. But SNUBH is tertiary hospital that many patients suspected or diagnosed with CRC visit SNUBH for definite diagnosis and treatment. In 6,780 subjects who got colon specimen in SNUBH (Table 2), 1,861 subjects had first visited other multiple primary clinic or secondary hospital and referred to SNUBH because CRC was suspected. And subjects from various areas visit SNUBH to receive cancer screening examination including colonoscopy. So, our study might have some amount of generalization. However, research involving various institutes should be needed to enhance generalization with validation of ICD 10th codes based on V-coed in Korea. Our study could be guideline for coming multicenter study. And ICD 10th codes of CRC registered by doctors (medical specialist and resident) of SNUBH were highly reliable. So it was helpful to analyze medical big-data if other hospitals make an effort to manage quality of diagnostic codes in the information technology era.

In conclusion, an administrative database validation study was performed using data from the ICD 10th code of CRC. Our study suggests that the big-cohort administrative database using ICD 10th code for CRC appears to be accurate, supporting the CRC studies thus far.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Research Foundation of Korea (NRF) (No. 2011-0030001) for the Global Core Research Center (GCRC) funded by the Korean government (MSIP).

CONFLICTS OF INTEREST

No potential conflicts of interest were disclosed.

Supplementary Materials

Supplementary materials can be found via https://doi.org/10.15430/JCP.2018.23.4.183.

REFERENCES

1. Barnabe C, Joseph L, Belisle P, Labrecque J, Edworthy S, Barr SG, et al. Prevalence of systemic lupus erythematosus and systemic sclerosis in the First Nations population of Alberta, Canada. Arthritis Care Res (Hoboken) 2012;64:138-43.
2. Bernatsky S, Lix L, Hanly JG, Hudson M, Badley E, Peschken C, et al. Surveillance of systemic autoimmune rheumatic diseases using administrative data. Rheumatol Int 2011;31:549-54.
3. Shin CM, Han K, Lee DH, Choi YJ, Kim N, Park YS, et al. Association among obesity, metabolic health, and the risk for colorectal cancer in the general population in Korea using the national health insurance service-national sample cohort. Dis Colon Rectum 2017;60:1192-200.
4. Choi YJ, Lee DH, Han KD, Kim HS, Yoon H, Shin CM, et al. The relationship between drinking alcohol and esophageal, gastric or colorectal cancer: a nationwide population-based cohort study of South Korea. PLoS One 2017;12:e0185778.
5. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol 2004;57:131-41.
6. Bernatsky S, Joseph L, Pineau CA, Tamblyn R, Feldman DE, Clarke AE. A population-based assessment of systemic lupus erythematosus incidence and prevalence--results and implications of using administrative data for epidemiological studies. Rheumatology (Oxford) 2007;46:1814-8.
7. Barber C, Lacaille D, Fortin PR. Systematic review of validation studies of the use of administrative data to identify serious infections. Arthritis Care Res (Hoboken) 2013;65:1343-57.
8. Bae S, Kim N, Kang JM, Kim DS, Kim KM, Cho YK, et al. Incidence and 30-day mortality of peptic ulcer bleeding in Korea. Eur J Gastroenterol Hepatol 2012;24:675-82.
9. Bae S, Shim KN, Kim N, Kang JM, Kim DS, Kim KM, et al. Incidence and short-term mortality from perforated peptic ulcer in Korea: a population-based study. J Epidemiol 2012;22:508-16.
10. Office of the Secretary, HHS. Administrative simplification: adoption of a standard for a unique health plan identifier; addition to the National Provider Identifier requirements; and a change to the compliance date for the International Classification of Diseases, 10th edition (ICD-10-CM and ICD-10-PCS) medical data code sets. Final rule. Fed Regist 2012;77:54663-720.
11. World Health Organization. ICD-10: international statistical classification of diseases and related health problems: tenth revision. Gevena, World Health Organization, 2004.
12. Abraham NS, Cohen DC, Rivers B, Richardson P. Validation of administrative data used for the diagnosis of upper gastrointestinal events following nonsteroidal anti-inflammatory drug prescription. Aliment Pharmacol Ther 2006;24:299-306.
13. Baldi I, Vicari P, Di Cuonzo D, Zanetti R, Pagano E, Rosato R, et al. A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. J Clin Epidemiol 2008;61:373-9.
14. Ganry O, Taleb A, Peng J, Raverdy N, Dubreuil A. Evaluation of an algorithm to identify incident breast cancer cases using DRGs data. Eur J Cancer Prev 2003;12:295-9.
15. Rostgaard K, Holst H, Mouridsen HT, Lynge E. Do clinical databases render population-based cancer registers obsolete? The example of breast cancer in Denmark. Cancer Causes Control 2000;11:669-74.
16. International Agency for Research on Cancer. World cancer report. Lyon, International Agency for Research on Cancer, 2008.
17. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin 2005;55:74-108.
18. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:E359-86.
19. Jung KW, Won YJ, Oh CM, Kong HJ, Lee DH, Lee KH. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2014. Cancer Res Treat 2017;49:292-305.
20. Kim SE, Paik HY, Yoon H, Lee JE, Kim N, Sung MK. Sex- and gender-specific disparities in colorectal cancer risk. World J

Gastroenterol 2015;21:5167-75.

21. Lee SM, Kim N, Son HJ, Park JH, Nam RH, Ham MH, et al. The effect of sex on the azoxymethane/dextran sulfate sodium-treated mice model of colon cancer. J Cancer Prev 2016;21:271-8.
22. Yoo S, Lee KH, Lee HJ, Ha K, Lim C, Chin HJ, et al. Seoul National University Bundang hospital's electronic system for total care. Healthc Inform Res 2012;18:145-52.
23. de Mul M, Alons P, van der Velde P, Konings I, Bakker J, Hazelzet J. Development of a clinical data warehouse from an intensive care clinical information system. Comput Methods Programs Biomed 2012;105:22-30.
24. Song SO, Jung CH, Song YD, Park CY, Kwon HS, Cha BS, et al. Background and data configuration process of a nationwide population-based study using the Korean national health insurance system. Diabetes Metab J 2014;38:395-403.
25. Kim HS, Choi YJ, Shin DW, Han KD, Yoon H, Shin CM, et al. Secondary primary prostate cancer after colorectal cancer: a nationwide population-based cohort study in Korea. J Cancer Prev 2017;22:241-7.
26. Green FL, Rage DL, Fleming ID, Fritz A, Balch CM, Haller DG, et al. AJCC cancer staging handbook: from the AJCC cancer staging manual. 6th ed. New York, Springer-Verlag, 2002.
27. Schlemper RJ, Riddell RH, Kato Y, Borchard F, Cooper HS, Dawsey SM, et al. The Vienna classification of gastrointestinal epithelial neoplasia. Gut 2000;47:251-5.
28. Schlemper RJ, Kato Y, Stolte M. Diagnostic criteria for gastrointestinal carcinomas in Japan and Western countries: proposal for a new classification system of gastrointestinal epithelial neoplasia. J Gastroenterol Hepatol 2000;15 Suppl:G49-57.
29. Leung WK, Ho HJ, Lin JT, Wu MS, Wu CY. Prior gastroscopy and mortality in patients with gastric cancer: a matched retrospective cohort study. Gastrointest Endosc 2018;87:119-27.e3.
30. Wu CY, Chang YT, Juan CK, Shieh JJ, Lin YP, Liu HN, et al. Risk of inflammatory bowel disease in patients with rosacea: results from a nationwide cohort study in Taiwan. J Am Acad Dermatol 2017;76:911-7.
31. Penberthy L, McClish D, Pugh A, Smith W, Manning C, Retchin S. Using hospital discharge files to enhance cancer surveillance. Am J Epidemiol 2003;158:27-34.
32. Bujang MA, Adnan TH. Requirements for minimum sample size for sensitivity and specificity analysis. J Clin Diagn Res 2016; 10:YE01-06.