Retrospective Cohort Study

Accuracy of an administrative database for pancreatic cancer by international classification of disease 10th codes: A retrospective large-cohort study

Young-Jae Hwang, Seon Mee Park, Soomin Ahn, Jong-Chan Lee, Young Soo Park, Nayoung Kim

ORCID number: Young-Jae Hwang (0000-0003-3979-711X); Seon Mee Park (0000-0002-5835-2741); Soomin Ahn (0000-0002-2464-0665); Jong-Chan Lee (0000-0001-7862-3257); Young Soo Park (0000-0001-8945-0507); Nayoung Kim (0000-0002-9397-0406).

Author contributions: Hwang YJ collected data, analyzed data and drafted the article; Park SM advised design of the protocol and edited the manuscript; Ahn S performed the pathologic diagnosis; Lee JC advised the protocol and revised the manuscript; Park YS and Kim N designed the protocol, and edited the manuscript; All authors have read and approved the final draft of this paper.

Supported by the National Research Foundation of Korea, No. 2011-0030001.

Institutional review board statement: This study was approved by the institutional review board of the Ethics Committee of Seoul National University Bundang Hospital.

Informed consent statement: Patients were not required to give informed consent to the study because our study was done retrospectively. Data for study were obtained after each patient agreed to treatment.

Conflict-of-interest statement: The authors have no conflicts of interest to disclose.

Abstract

BACKGROUND

Korean National Health Insurance (NHI) claims database provides large-cohort. However, studies regarding accuracy of administrative database for pancreatic cancer (PC) have not been reported. We aimed to identify accuracy of NHI database regarding PC classified by international classification of disease (ICD)-10 codes.

AIM

To identify the accuracy and usefulness of administrative database in PC and the accurate ICD codes for PC with location.

METHODS

Study and control groups were collected from 2003 to 2016 at Seoul National University Bundang Hospital. Cases of PC were identified in NHI database by international classification of diseases, 10th revision edition (ICD-10 codes) supported with V codes. V code is issued by medical doctors for covering 95% of medical cost by Korean government. According to pathologic reports, definite or possible diagnoses were defined using medical records, images, and pathology.
Hwang YJ et al. Diagnostic accuracy of pancreatic cancer database

INTRODUCTION

Pancreatic cancer (PC) has a very poor prognosis because most are diagnosed at advanced stages, are inoperable state due to invasion of adjacent arteries, or are intractable to chemotherapy[1-3]. Accurate diagnosis of PC remains challenging despite the widespread use of endoscopic ultrasound-guided fine needle aspiration (EUS-FNA) and biopsy. Therefore, pathological diagnosis of PC is not always possible, and most are diagnosed based on clinical features including image findings, clinical course, and laboratory data.

Location of primary PC is important for prognosis[4]. Patients with PC in head showed a 5% increased survival benefit as compared with PC in body or tail[5]. It may be associated with early symptom of PC in the head by obstruction of bile duct or pancreatic duct. Further research is needed about epidemiology and risk factor of PC in body or tail for screening and early diagnosis. If the primary location of PC is well described in database, it might be easier to do research for PC.

Recently, an administrative database has been widely used for medical research[5-8]. The administrative database includes personal medical information of a large number of the population with long-term follow-up. In addition, administrative database can provide easy access for study of PC location such as identification of information regarding this PC location. For proper interpretation of the results derived from this database, the reliability on the database is critical. Nevertheless, their accuracy in identifying cancer patients for the claims databases collect data for the purposes of reimbursement remains in doubt[9]. Furthermore, there have been limited studies regarding accuracy and usefulness of the administrative database[9,10].

The Korea National Health Insurance System (NHIS) contains a complete set of health information pertaining to 50 million members[9,10]. The source of the NHIS is the Health Insurance Review and Assessment (HIRA) database, including all insurance claims information of approximately 97% of the Korean population. In this database, the name of the disease is usually coded according to the international classification of diseases, 10th revision edition (ICD-10 codes) of pancreatic cancer in an administrative database are acceptable for use for population-based large-cohort studies. To enhance the diagnostic accuracy, we recommend patient identification by the ICD-10 code with tumor location information.

RESULTS

A total of 1846 cases with PC and controls were collected. Among PC, only 410 (22.2%) cases were identified as specific cancer sites including head in 234 (12.7%) cases, tail in 104 (5.6%) cases and body in 72 (3.9%) cases. Among PC, 910 (49.3%) cases were diagnosed by definite criteria. Most of these were adenocarcinoma (98.0%). The rates of definite diagnosis of PC were highest in head (70.1%) followed by body (47.2%) and tail (43.3%). False-positive cases were pancreatic cystic neoplasm and metastasis to the pancreas. In terms of the overall diagnosis of PC, sensitivity, specificity, positive predictive value, and negative predictive value were 99.95%, 98.72%, 98.70%, and 99.95%, respectively. Diagnostic accuracy was similar both in terms of diagnostic criteria and tumor locations.

CONCLUSION

Korean NIH claims database collected according to ICD-10 code with V code for PC showed good accuracy.

Key words: Korean national health insurance; Accuracy; Pancreatic cancer; International classification of disease; Sensitivity; Specificity

©The Author(s) 2019. Published by Baishideng Publishing Group Inc. All rights reserved.
diseases, 10th revision edition (ICD-10 code) published by the World Health Organization\textsuperscript{[13,14]}. Direct validation for the accuracy between the administrative dataset and NHIS data is impossible because of the Personal Information Protection Act in Korea. Therefore, validation for accuracy and usefulness of diagnostic codes could only be performed at individual hospitals where the diagnosis of each disease was performed and reported to HIRA for insurance claims. Furthermore, even though ICD-10 code includes the information for location of PC sometimes it is difficult to define the location of PC. In this situation ICD-10 code without location of PC is used by medical doctors. If current situation is analyzed, it might be good information for approaching the patients with PC.

From this background we aimed to evaluate the accuracy and usefulness of administrative database in PC. To certify the accuracy of diagnosis, we calculated the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of PC by ICD-10 codes compared to controls\textsuperscript{[15]}. In addition, we aimed to identify the location of PC in detail using ICD-10 codes and electronic medical records (EMR) to define how much the doctors insert the accurate ICD codes for PC with location.

**MATERIALS AND METHODS**

**Data source**

From May 2003 to December 2016, cases of PC were retrospectively collected using the Seoul National University Bundang Hospital (SNUBH) Clinical Data Warehouse (CDW)\textsuperscript{[16]}, which was its own database analysis program. The EMR system contains information on the visiting hospital departments, the principal diagnoses and surgical and diagnostic procedures for each patient\textsuperscript{[17]}. In addition, it includes pathologic results of specimens and imaging modalities, including computerized tomography (CT), endoscopic retrograde cholangiopancreatography (ERCP), EUS, magnetic resonance imaging (MRI), and positron emission tomography (PET)\textsuperscript{[3,18-21]}.

**Study population**

Information regarding patients, including hospital visit dates, subject characteristics, diagnostic procedures, pathologic results, and surgeries was collected. These information was easily obtained from administrative database. Other hospital medical data were identified though the uploaded database in SNUBH EMR. After approval of the study protocol by the Ethics Committee at SNUBH (IRB number B-1701/378-105), a list of patients with PC according to the ICD-10 code as primary diagnosis was acquired: (1) C25.0-25.3 (malignant neoplasm of pancreas at head, body, tail, and duct, respectively); (2) C25.4 (Malignant neoplasm of endocrine pancreas); and (3) C25.7-9 (malignant neoplasm of pancreas at other parts, overlapping lesion, and unspecified, respectively)\textsuperscript{[22]}. Then, searched cases were checked as being registered as V codes in the NHIS to confirm diagnostic codes\textsuperscript{[23]}. The V code is a special code for patients with any ICD-10 cancer codes in South Korea, established by the Korean Ministry of Health and Welfare in 2008. Cancer patients who are registered in the NHIS have issued a V code and are reimbursed at 95% of the medical cost by the Korean government for 5 years. Control cases are defined as individuals without ICD-10 codes for PCs (C25.0-25.9) during the study periods, who experience work-up pathways similar to those of PC, including images (CT, MRI, ERCP, or EUS) and surgery.

**Analyzing accuracy of PC diagnosis from administrative database**

Medical records of the study and control groups were analyzed to identify definite and possible diagnostic criteria. Definite diagnoses were made according to pathologic reports compatible with PC\textsuperscript{[24-26]}. Possible diagnoses were made according to image findings, clinical courses, or increased CA 19-9 > 100 U/m compatible with PC\textsuperscript{[3,21,27-29]}. Typical image finding of PC was defined as focal hypo-attenuated lesions, pancreatic ductal dilation, distal pancreatic parenchymal atrophy, and involvement of the surrounding vascular structures or other organs on radiologic examinations (Table 1 and Figure 1)\textsuperscript{[30-32]}.

In the definite diagnosis group, cancer cell types (adenocarcinoma, adenosquamous carcinoma, or neuroendocrine tumor) and methods of pathologic diagnosis through surgery, endoscopic biopsy or FNA were analyzed (Tables 1 and 2)\textsuperscript{[25,33]}. In the possible diagnosis group, we examined reports of images (CT, MRI, ERCP, EUS, endoscopy, and PET) by a radiologist or medical records of a physician’s reading of the images. We used serum levels of CA 19-9 to differentiate PC from other cancers\textsuperscript{[21,29]}. To enhance the study reliability, three reviewers carefully examined medical records and compared the final decisions for each case. For discordant cases, they discussed the
### Table 1  Diagnostic criteria of pancreatic cancer

| Diagnostic criteria | Methods                        | Positive finding                                    |
|---------------------|--------------------------------|-----------------------------------------------------|
| **Definite criteria** |                                |                                                     |
| Pathologic analysis | Endoscopy, ERCP, EUS-FNA, Surgery | Adenocarcinoma Mucinous carcinoma Other malignant neoplasm |
|                     |                                |                                                     |
| **Possible criteria** |                                |                                                     |
| Imaging finding     | CT, MRI, ERCP, MRCP, EUS, endoscopy, PET | Focal hypo-attenuated lesion Pancreatic duct dilation Distal pancreatic parenchymal atrophy Involvement of the surrounding vascular structures or other organs |
| Clinical features   | Medical record                 | Clinical courses compatible with PC                 |
| Tumor marker        | CA19-9                         | To exclude other malignancies including hepatocellular carcinoma |

CT: Positron emission tomography; MRI: Magnetic resonance imaging; ERCP: Endoscopic retrograde cholangiopancreatography; MRCP: Magnetic resonance cholangiopancreatography; EUS: Endoscopic ultrasound; PET: Positron emission tomography; FNA: Fine needle aspiration; CA19-9: Carbohydrate antigen 19-9; αFP: α-fetoprotein; PC: Pancreatic cancer.

cases and reached consensus. After reviewing medical records and classifying each case, the sensitivity, specificity, PPV, and NPV with 95% confidence intervals (CI) were calculated. We also compared diagnostic power according to cancer sites at the head (C25.0), body (C25.1) and tail (C25.2). In addition, we analyzed patients with ICD 10-code of PC with primary location (C25.0, C25.1, and C25.2).

### RESULTS

#### Characteristics of cases diagnosed as PC by the international classification of diseases, 10th revision edition

A total of 1846 subjects were identified as registered with ICD-10 codes for PC at the SNUBH during the study period (Table 3). Among PC, 1428 (77.4%) cases were registered as unspecified PC, and only 410 (22.2%) cases could be identified with specific cancer sites. PC in the head [234, (12.7%)] was the most common, followed by the tail [104, (5.6%)] and body [72, (3.9%)]. Proportions of PC cases in the pancreatic duct, neuroendocrine tumor, or overlapping were very rare, with only 0.3%, 0.1%, or 0.1%, respectively. Primary cancer location couldn’t be identified in patients with C25.9 [1428, (77.4%)]. In these cases with C25.9, we carefully examined all medical records one by one to identify primary cancer location.

Among PC, 910 (49.3%) cases had pancreatic pathologic results associated with the definite diagnostic criteria and classified as definite diagnosis group. Other 936 (50.7%) cases were classified as possible diagnosis group. Pathologic diagnosis was accomplished by surgery in 717 (78.8%) cases, by ERCP or endoscopy in 163 (17.9%) cases, and by EUS-FNA or percutaneous biopsy in 30 (3.5%) cases. Among 1198 cases with serum levels of CA19-9, 684 (57.1%) cases had elevated levels (> 100 UL).

#### Diagnostic accuracy of PC by the international classification of diseases, 10th revision edition in the administrative database

We analyzed accuracy of ICD-10 codes of PC by definite or possible diagnostic criteria (Table 4). Among 910 cases with pathologic diagnosis, 904 cases satisfied definite diagnostic criteria of PC. Pathologic diagnoses were adenocarcinoma in 886 (98.0%) cases, adenosquamous carcinoma in 3 (0.3%), and neuroendocrine tumor in 15 (1.7%) cases. Six cases who were identified as false-positives, were pancreatic cystic neoplasms, including serous cystic neoplasms, mucinous cystic neoplasms and intraductal pancreatic mucinous neoplasms (Table 3). Among 938 cases with possible diagnoses, 924 subjects satisfied possible diagnostic criteria for PC. Fourteen cases identified as false-positive were pancreatic metastasis from other primary cancers in 6 cases, pancreatic cystic neoplasms in 5 cases, pancreatitis in 2 cases, and accessory spleen in 1 case.

Among 1846 cases of control, only one case of PC was identified (Table 4). This patient underwent distal pancreatectomy because of a pancreatic tail mass and pancreatic ductal dilatation on CT scan. Pathologic diagnosis was invasive carcinoma originating from an intraductal papillary mucinous neoplasm of the pancreas. This
ICD: International classification of diseases; PC: Pancreatic cancer; EUS: Endoscopic ultrasound; ERCP: Endoscopic retrograde cholangiopancreatography; CT: Computerized tomography; MRI: Magnetic resonance imaging; PET: Positron emission tomography.

The diagnostic accuracy of PC differed according to tumor sites (Table 5). The rate of definite diagnosis in the pancreas head was 70.0%, while those in pancreas tail and body were 46.2% and 43.1%, respectively. Incorrect diagnoses including false-positives and false-negatives were 1.4% for pancreatic body cancer, 1.0% for pancreatic tail cancer, and 0% for pancreatic head cancer.

Accuracy of the international classification of diseases, 10th revision edition of PC in the administrative database

Calculated statistical values are summarized in Table 6. For overall diagnostic criteria of PC, the sensitivity and specificity of ICD-10 codes for PC were 99.95% (95%CI: 99.94-99.95) and 98.72% (95%CI: 98.70-98.73), respectively. The PPV and NPV were 98.70% (95%CI: 98.68-98.72) and 99.95% (95%CI: 99.94-99.95), respectively. For definite diagnostic criteria of PC, the sensitivity and specificity of ICD-10 codes for PC were 99.89% (95%CI: 99.88-99.90) and 99.68% (95%CI: 99.67-99.68), respectively. The PPV and NPV were 99.34% (95%CI: 99.32-99.36) and 99.95% (95%CI: 99.94-99.95), respectively. For possible diagnostic criteria for PC, the sensitivity and specificity were 99.89% (95%CI, 99.88-99.90) and 99.03% (99.02-99.05), respectively. The PPV and NPV were 98.08% (98.05-98.11) and 99.95% (99.94-99.95), respectively.

DISCUSSION

This study demonstrated that ICD-10 codes for PC in the administrative database are valid for use in population-based large-cohort studies. Although half of the cases were diagnosed by clinical and radiological features, they showed high diagnostic accuracy. Our results suggest the reliability of previous large-cohort studies using the administrative database in South Korea.

Administrative large databases from various disease registries have been used for population-based studies. However, the quality of a database may be suggested by the quotation of previous studies or by demonstrating similar trends in national estimates instead of validation of their database. Jon et al. studied cancer trends in liver, gallbladder, bile duct, and pancreas in an elderly population in Denmark. They identified cases by ICD-10 codes using the NORDCAN database, widely used in a previous study, without validation.

Previous studies for accuracy of ICD-9 codes revealed that interpretation of administrative databases relying only on ICD-9 codes requires caution. Arous et al. identified a total of 1107 PC patients by ICD-9 codes from institutional health care information system (HIS)-linked data sets and surgical databases. They reviewed all patients manually to validate the diagnoses. Analysis regarding pancreatic pathology revealed that 80.3% of patients had true pancreatic neoplasms and 19.7% had other

Figure 1 Proposed study algorithm for the inclusion and classification of subjects. ICD: International classification of diseases; PC: Pancreatic cancer; EUS: Endoscopic ultrasound; ERCP: Endoscopic retrograde cholangiopancreatography; CT: Computerized tomography; MRI: Magnetic resonance imaging; PET: Positron emission tomography.
pancreatic pathologies. When they used only the HIS-linked dataset, only 36.3% of patients were consistent with pancreatic neoplasms. Friedlin et al.\textsuperscript{[9]} compared the diagnostic accuracy of ICD-9 codes and natural language processing (NLP) technology to identify PC in a cohort of pancreatic cysts. They reported that ICD-9 codes achieved lower specificity than did the NLP method (46% and 94%, respectively) in spite of the high sensitivity for identifying PC by both ICD-9 codes and NLP (95% and 84%, respectively).

Our study identified a study group of PC by ICD-10 codes by adding V code using two disease registries, the SNUBH database and the NHIS. Previous population-based large-cohort studies identified cancer populations by both V code and ICD-10 codes\textsuperscript{[41,42]}. They reported the usefulness of the NHIS database collected by V code in South Korea\textsuperscript{[41,42]}. Seo et al.\textsuperscript{[42]} compared the cancer incidence rates found in the NHIS against in the National Cancer Registry of Korea. The results showed similar overall cancer incidences as well as age-, sex-, and disease-specific rates in both databases. The reason why we tried to identify the accuracy of ICD-10 code for PC registered in the NHIS in the present study was because the disease entity of PC is difficult to diagnose. We used two disease registries, the SNUBH database and the NHIS, to identify PC cases and controls. We analyzed the diagnostic accuracy according to definite diagnostic criteria in the presence of pathologic reports. Although the rates of pathologic diagnosis were only 49.3%, they achieved a high sensitivity of 99.89%, specificity of 99.68%, PPV of 99.34%, and NPV of 99.95%. These results provide scientific evidence of the results of previous studies using the administrative database.

The rates of definite diagnosis and identification of specific cancer sites were higher for pancreatic head cancer ($n = 163$) than for pancreatic body ($n = 33$) or tail ($n = 45$) cancers. These results suggest that pancreatic head cancer is detected earlier and specimens are obtained more easily than for other sites\textsuperscript{[28]}. In addition, we suggest that it is rather difficult to diagnose pancreatic body or tail cancer, respectively, based on pathologic finding.

Half of the cases registered as PC by ICD-10 codes were validated by possible diagnostic criteria. Because obtaining pancreatic specimens by non-surgical methods is difficult and most would not be candidates for surgery. Only 15-20 percent of patients could be candidates for surgery\textsuperscript{[38]}. In our study, 717 (38.8%) patients got pancreatectomy. In patients who were not candidates of surgery or procedure because of advanced stages, PC was diagnosed only by clinical, radiologic or serologic features. For the diagnostic accuracy of PC we did not absolutely depend on the level of CA19-9. Instead we used tumor markers of CA19-9 and αFP to differentiate them from other cancer such as hepatocellular carcinoma when image findings and clinical symptoms were insufficient to diagnose PC. Cases registered as PC by ICD-10 codes without pathologic confirmation achieved a high sensitivity of 99.89%, specificity of 99.03%, PPV of 98.08%, and NPV of 99.95%.

We analyzed false-positive and false-negative cases. Cases with incorrect diagnostic pathologic codes were pancreatic cystic neoplasm. Malignant transformation can occur in premalignant pancreatic cystic neoplasm. The differential diagnosis between them is very difficult\textsuperscript{[38]}. Among cases with possible diagnoses, the wrong diagnosis was caused by pancreatic metastases, pancreatic cystic neoplasm, pancreatitis, or ectopic adjacent organs. PC was difficult to differentiate from invasion, metastasis from adjacent organs or benign cystic lesion.

We found that diagnosis according to cancer sites was not accurate in spite of the high overall diagnostic accuracy for PC. Unspecified PC (C25.9) comprised 77.4% of all PC, and most of the false-positive cases (23 out of 24) were recorded as C25.9.

### Table 2: Cancer cell type of pancreatic cancer

| Adenocarcinoma | Adeno-squamous carcinoma | Neuroendocrine tumor | Total |
|----------------|--------------------------|----------------------|-------|
| C250 163       | 0                        | 1                    | 164   |
| C251 33        | 0                        | 0                    | 33    |
| C252 45        | 0                        | 0                    | 45    |
| C253 2         | 0                        | 0                    | 2     |
| C254 2         | 0                        | 0                    | 2     |
| C257 1         | 0                        | 0                    | 1     |
| C259 640       | 3                        | 14                   | 657   |
| **886 (98.0)**| **3 (0.3)**              | **15 (1.7)**         | **904**|

\(^1\)Number of patients (%).


| Parameters                                      | N (%)   | C250       | C251       | C252       | C253       | C254       | C257       | C259       |
|------------------------------------------------|---------|------------|------------|------------|------------|------------|------------|------------|
| Number of patients                             | 1846    | 234 (12.7) | 72 (3.9)   | 104 (5.6)  | 5 (0.3)    | 2 (0.1)    | 1 (0.1)    | 1428 (77.4) |
| Age at diagnosis                               | 65.22 ± 11.97 | 64.72 ± 10.14 | 63.39 ± 10.48 | 65.43 ± 11.32 | 65.20 ± 11.08 | 70.50 ± 4.95 | 45.00 ± 0.00 | 65.39 ± 12.35 |
| Gender (male: Female)                          | 1116/730 | 150/84     | 45/27      | 52/52      | 1/4        | 2/0        | 1/0        | 865/563    |
| Diagnostic criteria                            |         |            |            |            |            |            |            |            |
| Definite                                       | 910 (49.3) | 164 (70.1) | 34 (47.2)  | 45 (43.3)  | 2 (40.0)   | 2 (100.0)  | 1 (100.0)  | 662 (46.4) |
| Possible                                       | 936 (50.7) | 70 (29.9)  | 38 (52.8)  | 59 (56.7)  | 3 (60.0)   | 0 (0.0)    | 0 (0.0)    | 766 (53.6) |
| Methods of pathologic diagnosis (n = 910)      |         |            |            |            |            |            |            |            |
| ERCP or endoscopy                              | 163 (17.9) | 34 (20.7)  | 2 (5.9)    | 8 (17.8)   | 0 (0.0)    | 0 (0.0)    | 0 (0.0)    | 119 (18.0) |
| EUS-FNA or percutaneous biopsy                 | 30 (3.3)  | 4 (2.4)    | 0 (0.0)    | 2 (4.4)    | 0 (0.0)    | 0 (0.0)    | 0 (0.0)    | 26 (3.9)   |
| Surgery                                        | 717 (78.8) | 126 (76.8) | 32 (94.1)  | 35 (77.8)  | 2 (100.0)  | 2 (100.0)  | 1 (100.0)  | 517 (78.1) |
| CA19-9 > 100 U/L                               |         |            |            |            |            |            |            |            |
| Yes                                            | 684 (37.1) | 95 (40.6)  | 29 (40.3)  | 44 (42.3)  | 1 (20.0)   | 0 (0.0)    | 0 (0.0)    | 515 (36.1) |
| No                                             | 514 (27.8) | 63 (26.9)  | 24 (33.3)  | 23 (22.1)  | 1 (20.0)   | 1 (50.0)   | 0 (0.0)    | 402 (28.2) |
| Missing                                        | 648 (35.1) | 76 (32.5)  | 19 (26.4)  | 37 (35.6)  | 3 (60.0)   | 1 (50.0)   | 1 (100.0)  | 511 (35.8) |

1Median ± standard deviation.
2Number (%).

Therefore PC by ICD-10 code adding a V-code in the NHIS data was not sufficient to study cancer sites. For the accurate study regarding primary PC location, we excluded PC patients of C25.9 or examined these patients one by one. If patients with C25.9 are excluded, the advantages of large administrative database disappear. If patients with C25.9 need to be checked primary cancer location individually, the advantage of easy access for medical information is eliminated. Both methods reduce the usefulness of administrative database. So we should try to fill in the ICD-10 codes with primary location of PC. Another weak point of PC coded by ICD-10 in the NHIS data was that it was not adequate for evaluation of neuroendocrine tumors. All neuroendocrine tumors were coded as C25.0 or C25.9, whereas they should be coded as C25.4. Furthermore, two adenocarcinoma cases were coded as C24.4 and should have been coded as C25.4. For the study for accuracy of diagnostic codes in the administrative database, institutions require two conditions: A high burden of cancer patients and a well-established CDW system. SNUBH might be an adequate hospital to perform this study because of its comprehensive EMR system. SNUBH developed an in-house comprehensive EMR in 2003. The warehouse system provides easy access to diagnostic information for research. In addition, SNUBH is a tertiary referral hospital to which regional hospitals would refer patients; therefore, sufficient numbers of PC cases would be enrolled in this study to enhance the power of the study results. To satisfy statistical requirements (α = 0.05, 1-β = 0.95, and effect size 0.1), more than one thousand cases are needed. The size of our study group was sufficient to fulfill the statistical criteria. We provided a new study model for evaluating the accuracy and usefulness of large administrative databases. Many studies using large administrative databases of PC have been done, and our study could support the reliability of these studies. To enhance the reliability of studies with large administrative databases, our study could be cited as a reference.

Our study has several limitations. One-half of cases were diagnosed by possible diagnostic criteria without pathologic confirmation. Pathologic diagnosis of PC is sometimes impossible because of poor patient conditions and technical difficulty. Therefore, if we adopted only definite diagnostic criteria of PC for accuracy of diagnosis, selection bias could occur. Another limitation was that the study was done only in a single hospital, SNUBH. The diagnostic accuracy might be increased in a tertiary referral hospital rather than a multicenter study. Because most PC cases are treated in referral hospitals in South Korea, we believe that our data may represent the entire PC data of the NHIS in South Korea. In spite of this limitation, our study demonstrated the excellent diagnostic accuracy of the PC data of the NHIS.

In conclusion, ICD-10 codes of PC in an administrative database are acceptable for use for population-based large-cohort studies. To prove reliability of administrative database, we examined subjects dividing two groups, definite and possible diagnosis. In addition, we analyzed both disease registries, SNUBH and NHIS. This study also compared with control group for calculating sensitivity, specificity, PPV and NPV.
### Table 4 Diagnostic accuracy of pancreatic cancer diagnosed by the international classification of diseases, 10th revision edition in the administrative database

| Status of PC | Positive | Negative | Total |
|--------------|----------|----------|-------|
| **Definite diagnostic criteria** |          |          |       |
| ICD codes of PC |          |          |       |
| Outcome positive | True positive | False positive | 910 |
| | 904 | 6 |
| Outcome negative | False negative | True negative | 1846 |
| | 1 | 1845 |
| Total | 905 | 1851 | 2756 |
| **Possible diagnostic criteria** |          |          |       |
| ICD codes of PC |          |          |       |
| Outcome positive | True positive | False positive | 936 |
| | 918 | 18 |
| Outcome negative | False negative | True negative | 1846 |
| | 1 | 1845 |
| Total | 919 | 1863 | 2782 |
| **Overall diagnostic criteria** |          |          |       |
| ICD codes of PC |          |          |       |
| Outcome positive | True positive | False positive | 1846 |
| | 1822 | 24 |
| Outcome negative | False negative | True negative | 1846 |
| | 1 | 1845 |
| Total | 1823 | 1869 | 3692 |

ICD: International classification of disease. PC: Pancreatic cancer.

To identify usefulness of database, we examined cancer location. If researchers could get information of PC site through only ICD-10 code, they can perform the study more easily.

To enhance the diagnostic accuracy, we recommend patient identification by the ICD-10 code with tumor location information and V-code system. From this, we preserved huge administrative database without exclusion. More researches with multiple institutions and various diseases should be needed to practice researches with administrative database.
Table 5 Diagnostic accuracy of pancreatic cancer according to tumor sites by the international classification of diseases, 10th revision edition

| ICD codes | Overall diagnosis | Definite diagnosis | Possible diagnosis |
|-----------|-------------------|--------------------|-------------------|
|           | True (+) | False (+) | Total | True (+) | False (+) | Total | True (+) | False (+) | Total |
| Cell types | 234      | 0        | 234  | 164   | 0        | 164  | 70      | 0        | 70    |
| C251      | 71       | 1        | 72   | 30    | 1        | 31   | 41      | 0        | 41    |
| C252      | 104      | 0        | 104  | 48    | 0        | 48   | 56      | 0        | 56    |
| C253      | 5        | 0        | 5    | 2     | 0        | 2    | 3       | 0        | 3     |
| C254      | 2        | 0        | 2    | 2     | 0        | 2    | 0       | 0        | 0     |
| C257      | 1        | 0        | 1    | 1     | 0        | 1    | 0       | 0        | 0     |
| C259      | 1405     | 23       | 1428 | 657   | 5        | 662  | 748     | 18       | 766   |
|           | 1822     | 24       | 1846 | 904   | 6        | 910  | 918     | 18       | 936   |

1Real cancer patients with ICD 10th codes of cancer.
2No cancer patients with ICD 10th codes of cancer. ICD: International classification of disease.

Table 6 Diagnostic power of international classification of diseases, 10th revision edition for pancreatic cancer

| Overall diagnosis | Definite diagnosis | Possible diagnosis |
|-------------------|-------------------|-------------------|
|                  | Point estimate (%) | 95%CI (%)         | Point estimate (%) | 95%CI (%)         | Point estimate (%) | 95%CI (%)         |
| Sensitivity       | 99.95              | 99.94-99.95       | 99.89              | 99.88-99.90       | 99.89              | 99.88-99.90       |
| Specificity       | 98.72              | 98.70-98.73       | 99.68              | 99.67-99.68       | 99.03              | 99.02-99.05       |
| Positive predictive value | 98.70 | 98.68-98.72       | 99.34              | 99.32-99.36       | 98.08              | 98.05-98.11       |
| Negative predictive value | 99.95 | 99.94-99.95       | 99.95              | 99.94-99.95       | 99.95              | 99.94-99.95       |

CI: Confidence interval.

ARTICLE HIGHLIGHTS

Research background
Pancreatic cancer (PC) is usually diagnosed at advanced stages, resulting in the poor prognosis. Large-cohort studies should be performed to evaluate epidemiology and prognosis of PC. However, there are not enough researches about the accuracy of administrative database to avoid coding discrepancies. This study identified accuracy of the administrative large-cohort database of PC. This study is important to support the validation of other large cohort study for PC.

Research motivation
Administrative database was useful for research because of easy access and much information. So, administrative database has been widely used for medical research. However accuracy of the administrative database may be problem. In addition, it was difficult to perform study to identify this. We tried to examine each case and prove accuracy of database of PC. Future study using administrative database of PC should be supported by this study.

Research objectives
We evaluated the accuracy and usefulness of administrative database in PC. In addition, we identified much the doctors insert the accurate ICD codes for PC with location.

Research methods
We evaluated the diagnostic accuracy of PC according to tumor sites from total of 1846 cases with PC and controls. To enhance the study reliability, three reviewers carefully examined medical records and compared the final decisions for each case. After reviewing, we calculated the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). In addition, we analyzed patients with ICD 10-code of PC with primary location.

Research results
Among PC, 1428 (77.4%) cases were registered as unspecified PC, and only 410 (22.2%) cases could be identified with specific cancer sites. For overall diagnostic criteria of PC, the sensitivity and specificity of ICD-10 codes for PC were 99.95% and 98.72%, respectively. The PPV and NPV were 98.70% and 99.95%, respectively.
Research conclusions
We showed accuracy of administrative database of PC in seoul national university Bundang hospital. In addition, we identified the location of PC to usefulness of database. Administrative database is useful and important for research. However, validation of database is necessary. From this result, study based on administrative database might be reliable. Future study with administrative database of PC could receive credibility from this result. In addition, this study presented a research method how to identify validation of administrative database.

Research perspectives
We thought that future study involved multiple institute should be planned. In addition, it is important to gather data in a unified way. We think there is a need for researches for accuracy of administrative database on other disease. These researches should be necessary for studies base on administrative database.

REFERENCES
1. Molina V, Visa L, Conill C, Navarro S, Escudero JM, Auge JM, Filella X, Lopez-Boado MA, Ferrer J, Fernandez-Cruz L, Molina R. CA 19-9 in pancreatic cancer: retrospective evaluation of patients with suspicion of pancreatic cancer. *Tumour Biol* 2012; 33: 799-807 [PMID: 22203495 DOI: 10.1007/s13277-011-0297-8]
2. Wang W, Shlipan A, Krishna SG, Ross WA, Bhatiuri MS, Tamam EP, Raju GS, Xiao L, Wolf R, Fleming JB, Lee JH. Use of EUS-FNA in diagnosing pancreatic neoplasm without a definitive mass on CT. *Gastrointest Endosc* 2013; 78: 73-80 [PMID: 23523302 DOI: 10.1016/j.gie.2013.01.040]
3. Kamisawa T, Wood LD, Ito T, Takaori K. Pancreatic cancer. *Lancer* 2016; 388: 73-85 [PMID: 26830752 DOI: 10.1016/S0140-6736(16)00141-4]
4. Tomassello G, Ghidini M, Costanzo A, Ghidini A, Russo A, Barni S, Passalacqua R, Petrelli F. Outcome of head compared to body and tail pancreatic cancer: a systematic review and meta-analysis of 93 studies. *J Gastrointest Oncol* 2019; 10: 259-269 [PMID: 31072953 DOI: 10.21037/jgo.2018.12.08]
5. Franchi C, Giussani G, Messina P, Montesano M, Romi S, Nobili A, Fortino I, Bortolotti A, Merlino L, Begli E; EPIRES Group. Validation of healthcare administrative data for the diagnosis of epilepsy. *J Epidemiol Community Health* 2013; 67: 1019-1024 [PMID: 24022813 DOI: 10.1136/jech-2013-202528]
6. Abraham NS, Cohen DC, Rivers B, Richardson P. Validation of administrative data used for the diagnosis of upper gastrointestinal-eyg-disease following nonsteroidal anti-inflammatory drug prescription. *Aliment Pharmacol Ther* 2006; 24: 299-306 [PMID: 16842456 DOI: 10.1111/j.1365-2036.2006.02985.x]
7. Baldi I, Vicari P, Di Cuonzo D, Zanetti R, Pagano E, Rosato R, Sacerdoti C, Segnan N, Merletti F, Ciccone G. A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. *J Clin Epidemiol* 2008; 61: 373-379 [PMID: 18315562 DOI: 10.1016/j.jclinepi.2007.05.017]
8. Aroue EJ, McDrade TP, Smith JK, Ng SC, Sullivan ME, Zottola RJ, Ranauro PJ, Shah SA, Whalen GF, Tseng JF. Electronic medical record: research tool for pancreatic cancer? *J Surg Res* 2014; 187: 466-470 [PMID: 24226179 DOI: 10.1016/j.jss.2013.10.030]
9. Friedla J, Overhage M, Al-Haddad MA, Waters JA, Aguilar-Saavedra JJ, Kesterson J, Schmidt M. Comparing methods for identifying pancreatic cancer patients using electronic data sources. *AMIA Annu Symp Proc* 2010; 2013: 237-241 [PMID: 21346976]
10. Beg MS, Dwivedi AK, Ahmad SA, Ali S, Oloowokure O. Impact of diabetes mellitus on the outcome of pancreatic cancer. *PLoS One* 2014; 9: e98511 [PMID: 24879130 DOI: 10.1371/journal.pone.0098511]
11. Niewenhuis LS, van den Brandt PA. Tree nut, peanut, and peanut butter consumption and the risk of gastric and esophageal cancer subtypes: the Netherlands Cohort Study. *Gastric Cancer* 2018; 21: 900-912 [PMID: 29594821 DOI: 10.1016/j.gcen.2018.08.021]
12. Kim HJ, Kang TU, Swan H, Kang MJ, Kim N, Ahn HS, Park SM. Incidence and Prognosis of Subsequent Cholangiocarcinoma in Patients with Hepatic Resection for Bile Duct Stones. *Dig Dis Sci* 2018; 63: 3465-3473 [PMID: 30517140 DOI: 10.1007/s10620-018-5262-6]
13. 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: 54663-54720 [PMID: 22950146]
14. WHO. ICD-10. International classification of diseases and related health problems: tenth revision. 2004.
15. Hwang YJ, Kim N, Yun CY, Yoon H, Shin CM, Park YS, Son IT, Oh HK, Kim DW, Kang SB, Lee HS, Park SM, Lee DH. Validation of Administrative Big Database for Colorectal Cancer Searched by ICD-10: International statistical classification of diseases and related health problems: tenth revision. 2004.
16. Hwang YJ, Kim N, Yun CY, Yoon H, Shin CM, Park YS, Son IT, Oh HK, Kim DW, Kang SB, Lee HS, Park SM, Lee DH. Validation of Administrative Big Database for Colorectal Cancer Searched by ICD-10: International statistical classification of diseases and related health problems: tenth revision. 2004.
17. WHO. ICD-10. International classification of diseases and related health problems: tenth revision. 2004.
18. Hwang YJ, Kim N, Yun CY, Yoon H, Shin CM, Park YS, Son IT, Oh HK, Kim DW, Kang SB, Lee HS, Park SM, Lee DH. Validation of Administrative Big Database for Colorectal Cancer Searched by ICD-10: International statistical classification of diseases and related health problems: tenth revision. 2004.
19. WHO. ICD-10. International classification of diseases and related health problems: tenth revision. 2004.
Hwang YJ et al. Diagnostic accuracy of pancreatic cancer database
