PLOS ONE

RESEARCH ARTICLE

Development of algorithms for identifying patients with Crohn's disease in the Japanese health insurance claims database

Hiromu Morikubo[1,2,3], Taku Kobayashi[1]*, Tomohiro Fukuda[1,2], Takayoshi Nagahama[4], Tadakazu Hisamatsu[3], Toshifumi Hibi[1]

1 Center for Advanced IBD Research and Treatment, Kitasato University Kitasato Institute Hospital, Minato-ku, Tokyo, Japan, **2** Department of Gastroenterology and Hepatology, Kitasato University Kitasato Institute Hospital, Minato-ku, Tokyo, Japan, **3** Department of Gastroenterology and Hepatology, Kyorin University School of Medicine, Mitaka-shi, Tokyo, Japan, **4** Data Innovation Lab, Japan Medical Data Center Co., Ltd., Minato-ku, Tokyo, Japan

* kobataku@insti.kitasato-u.ac.jp

Citation: Morikubo H, Kobayashi T, Fukuda T, Nagahama T, Hisamatsu T, Hibi T (2021) Development of algorithms for identifying patients with Crohn's disease in the Japanese health insurance claims database. PLoS ONE 16(10): e0258537. https://doi.org/10.1371/journal.pone.0258537

Editor: Valérie Pittet, Center for Primary Care and Public Health, SWITZERLAND

Received: March 25, 2021

Accepted: September 29, 2021

Published: October 13, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0258537

Data Availability Statement: All files are available from the GitHub database (https://github.com/HiromuMorikubo/pone2021).

Abstract

Background

Real-world big data studies using health insurance claims databases require extraction algorithms to accurately identify target population and outcome. However, no algorithm for Crohn's disease (CD) has yet been validated. In this study we aim to develop an algorithm for identifying CD using the claims data of the insurance system.

Methods

A single-center retrospective study to develop a CD extraction algorithm from insurance claims data was conducted. Patients visiting the Kitasato University Kitasato Institute Hospital between January 2015–February 2019 were enrolled, and data were extracted according to inclusion criteria combining the Tenth Revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) diagnosis codes with or without prescription or surgical codes. Hundred cases that met each inclusion criterion were randomly sampled and positive predictive values (PPVs) were calculated according to the diagnosis in the medical chart. Of all cases, 20% were reviewed in duplicate, and the inter-observer agreement (Kappa) was also calculated.

Results

From the 82,898 enrolled, 255 cases were extracted by diagnosis code alone, 197 by the combination of diagnosis and prescription codes, and 197 by the combination of diagnosis codes and prescription or surgical codes. The PPV for confirmed CD cases was 83% by diagnosis codes alone, but improved to 97% by combining with prescription codes. The inter-observer agreement was 0.9903.

Funding: This study was funded by JMDC Inc. The funder provided support in the form of salaries for TN, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. There was no additional external funding received for this study. The specific roles of these authors are articulated in the author contributions section.

Competing interests: HM has received research grants from Japan Foundation for Applied Enzymology. TK has served as a speaker, a consultant or an advisory board member for Abbvie, Alfresa Pharma, Janssen Pharma, Takeda, Mitsubishi Tanabe Pharma, Pfizer, Mochida, and received research grants from Nippon Kayaku, EA Pharma, Otsuka Holdings, JIMRO, Abbie, Zeria. FT has received research grants from Mitsubishi Tanabe Pharma. TN are employees of JMDC Co. Ltd., holds shares in JMDC Co. Ltd. TaH has served as a speaker, a consultant or an advisory board member for Mitsubishi Tanabe Pharma, AbbVie GK, EA Pharma, Kyorin Pharma, JIMRO, Janssen Pharmaceutical, Mochida Pharmaceutical, Takeda Pharmaceutical, and received research grants from Alfresa Pharma, EA Pharma, Mitsubishi Tanabe Pharma, AbbVie GK, JIMRO, Zeria Pharmaceutical, Daiichi-Sankyo, Kyorin Pharmaceutical, Nippon Kayaku, Astellas Pharma, Takeda Pharmaceutical, Pfizer, Mochida Pharmaceutical. ToH has served as a speaker, a consultant or an advisory board member for Aspen Japan, Abbvie GK, Ferring, Gilead Sciences, Janssen, JIMRO, Mitsubishi Tanabe Pharma, Mochida Pharmaceutical, Nippon Kayaku, Pfizer, Takeda Pharmaceutical, Zeria, and received research grants from Abbvie, EA Pharma, JIMRO, Otsuka Holdings, Zeria, and received scholarship grants from Zeria. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Conclusions

Single ICD-code alone was insufficient to define CD; however, the algorithm that combined diagnosis codes with prescription codes indicated a sufficiently high PPV and will enable outcome-based research on CD using the Japanese claims database.

Introduction

Crohn's disease (CD) is a chronic inflammatory bowel disease (IBD) of unknown etiology [1]. Recent progress on treatment for IBD has been remarkable, and many new drugs have been launched following randomized control trials (RCTs) [2]. At the same time, multiple clinical questions have arisen to help adapt the increased treatment options to better suit patients' needs in clinical practice. Consequently, the importance of observational studies, as well as RCTs, are being reevaluated [3]. In fact, it has been demonstrated that RCTs represent only a small proportion of patients with IBD in real-world practice [4]. In this respect, large-scale observational studies are also needed.

The incidence and prevalence of CD are higher in Western countries [5] and are also increasing in Asian countries, including Japan [6]. Its prevalence is 1.51-322/100,000 in Western countries [5] and 55.6/100,000 in Japan [7]. When conducting real-world observational studies requiring a large number of patients in Japan, it is often difficult to obtain a sufficient sample size from a single or small number of institutions. For diseases with low prevalence, the claims database can therefore be a useful tool for conducting large-scale real-world observational studies [8, 9]. In fact, various epidemiological studies using the claims database have been successfully conducted [10–12], and the usefulness of these databases has also been proven in IBD [8, 13–15]. However, it is important to note that the diagnosis in the claims database may not always reflect the final medical diagnosis made in clinical practice, and validation studies are therefore necessary for each disease [16–19]. Furthermore, in Japan, the validity of the Tenth Revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) codes registered in the claims data for CD has not been evaluated. Thus, the reliability of claims database studies on CD using ICD-10 codes has not yet been confirmed. Therefore, the purpose of this study was to develop an algorithm for identifying CD using the claims data of the Japanese insurance system.

Materials and methods

Study design

This was a retrospective cross-sectional validation study that reviewed health insurance claims data and medical records. Patients who met the inclusion criteria and those who did not were randomly selected from the claims data of patients who visited Kitasato University Kitasato Institute Hospital (Tokyo, Japan) and filed for insurance reimbursement. The medical records of these patients were reviewed to evaluate the validity of the inclusion criteria. Case selection, random sampling, and statistical analyses were conducted in collaboration with the Japan Medical Data Center (JMDC) Corporation (Tokyo, Japan). The flow of the review process is shown in Fig 1.

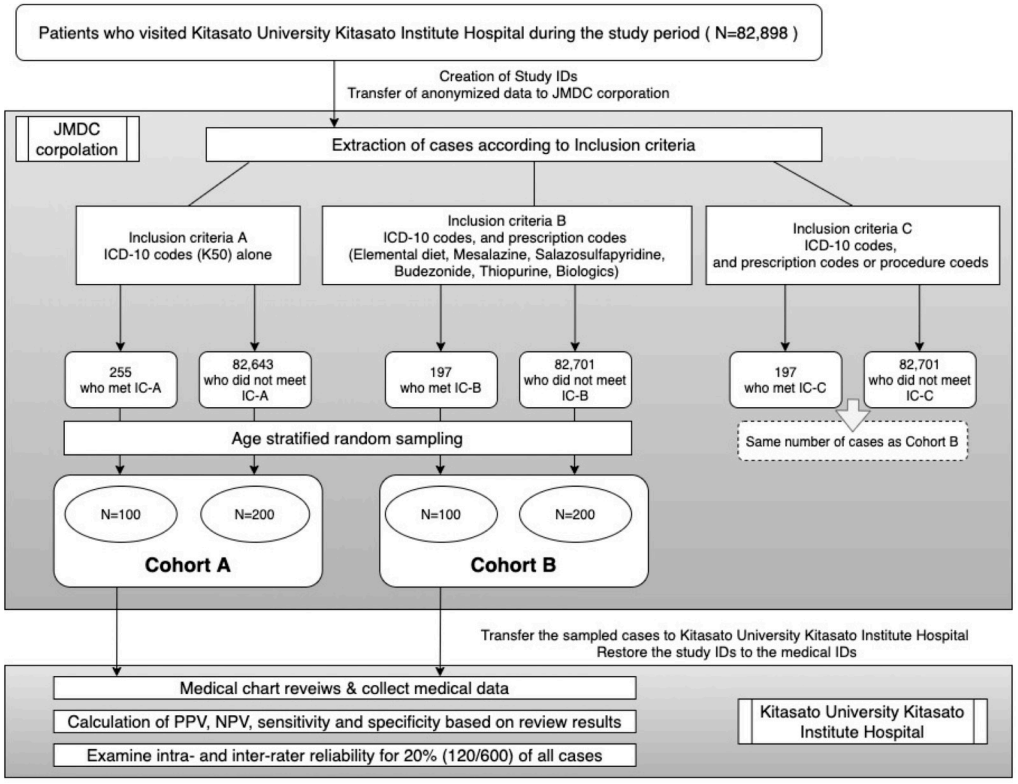

Fig 1. Study design and data flow, cohort setting. A total of 82,898 patients were enrolled during the study period. 255 and 197 patients who met the IC-A and B were extracted, respectively. Patients selected for IC-C (n = 197) were excluded from the subsequent analyses due to the same number of patients as in IC-B. JMDC; Japan Medical Data Center, ICD-10; Tenth Revision of the International Statistical Classification of Diseases and Related Health Problems, IC; Inclusion criteria, PPV; positive predictive value, NPV; negative predictive value.

https://doi.org/10.1371/journal.pone.0258537.g001

Japanese health administrative data

Japan has a universal health insurance system, which covers almost all citizens, as they are obliged to join one of the systems according to their occupation and age [20]. At the end of each month, each medical provider files a set of reimbursement invoices to the insurance payer via the review organization. For this reason, medical institutions register all processes, drugs, procedures, and devices that are subject to reimbursement according to the Ministry of Health, Labor and Welfare's standard codes, and this registration information is managed as the Japanese claims database [20, 21].

Setting

Kitasato University Kitasato Institute Hospital (Tokyo, Japan) is affiliated with Kitasato University; it has the Center for Advanced IBD Research and Treatment, which has 329 hospital beds. It received 865 outpatients and received 163 inpatients per day in FY2019.

Inclusion criteria

Patients who visited the hospital between January 2015 and December 2019 were included in the first sampling of the claims data. The observation period was set as the maximum period for which insurance claims data were available at the study site. According to the inclusion

Table 1. Inclusion criteria, and details of the confirmed diagnosis.

Criteria		
Inclusion Criteria	A	Patients with a confirmed ICD-10 diagnostic code of CD (K50), without a confirmed ICD-10 diagnostic code of ulcerative colitis (K51) or Behcet's disease (M35) in the same month.
	B	A + Prescription codes for CD in the same month
	C	A + Prescription or Surgical codes for CD in the same month
Details of the confirmed diagnosis	a	Confirmed diagnosis at own institution
	b	Diagnosed by an IBD specialist or gastroenterologist in another hospital
	c	Diagnosed by a primary care physician (with a description of the findings supporting the diagnosis)
	d	Diagnosed by a primary care physician (without a description of the findings supporting the diagnosis)

IBD; Inflammatory bowel disease, ICD-10; Tenth Revision of the International Statistical Classification of Diseases and Related Health Problems, CD; Crohn's disease

https://doi.org/10.1371/journal.pone.0258537.t001

criteria listed below, the cases were divided into those that met the inclusion criteria and those that did not (Table 1). Age-stratified random sampling of 100 cases each was performed in cases that met the inclusion criteria and 200 cases from those that did not. The cases extracted using each inclusion criterion (IC-A/B/C) were defined as cohorts (Cohort-A/B/C).

Inclusion criteria A (IC-A; diagnostic code alone): Patients with a confirmed ICD-10 diagnostic code of CD (K50) (S1 Table) but without a confirmed ICD-10 diagnostic code of ulcerative colitis or Behçet's disease in the same month.

Inclusion criteria B (IC-B; diagnostic and prescription codes): Cases fulfilling IC-A and with prescription codes (S2 Table) in the same month as the diagnostic codes.

Inclusion criteria C (IC-C; diagnostic, prescription, and surgical codes): Cases fulfilling IC-A with prescription codes or surgical codes (S2 Table) in the same month as the diagnostic codes.

Reviewing process

A medical chart review was independently performed by two gastroenterologists (chart reviewers with at least 5 years of clinical experience and training in IBD practice at a specialist center who are engaged in Kitasato University Kitasato Institute Hospital). The reviewers classified cases into three categories based on the gold standard according to the definition by the national guidelines [1] described in the section below as confirmed diagnosis, suspected diagnosis, and negative. If the two reviewers had different diagnoses, the final decision was made after (1) discussion between the two reviewers or (2) consultation with a third reviewer (a gastroenterologist and IBD specialist).

Gold standard and data collection of clinical information

The following data were collected for each randomly sampled case at the time when the inclusion criteria were met: age, sex, age of onset, disease type (Montreal classification), previous surgery (intestinal/anal), medications for CD, laboratory findings, examination results (upper and lower endoscopy, histopathology, small bowel radiography, magnetic resonance enterography, and intestinal ultrasound findings), discharge summary, referral letter, and registration of intractable disease application. The gold standard was based on the national guidelines of the Japanese Society of gastroenterology [1]. The details of cases with confirmed diagnoses

were categorized as follows: a) diagnosed or confirmed the diagnosis at our own institution, b) diagnosed only by an IBD specialist or gastroenterologist in another hospital; c) diagnosed only by a primary care physician (with a description of the findings supporting the diagnosis), and d) diagnosed only by a primary care physician (without a description of the findings supporting the diagnosis).

Assessment of validity

For each inclusion criterion, validity was assessed for confirmed and suspected diagnoses. A 2 × 2 contingency table was created, and the validity was mainly calculated by the positive predictive value (PPV). The sensitivity, specificity, and negative predictive value (NPV) were also calculated. A total of 20% (120/600) of the total cases were independently reviewed by two chart reviewers per case to examine inter-rater reliability and another 20% of the total cases were reviewed twice by one chart reviewer with a two-week interval, to examine the intra-rater reliability.

Statistical analysis

All statistical analyses were performed using STATA/S v. 15.1 (Stata Corporation, College Station, Texas, USA). Continuous variables were expressed as the median interquartile range (IQR) or mean standard deviation (SD). Categorical variables were expressed as integers and percentages (%). A 2 × 2 contingency table was created to calculate the sensitivity, specificity, PPV, and NPV. Inter- and intra-rater reliability was assessed using kappa, weighted kappa, and AC1.

The sample size was set at 100 for cases that met the inclusion criteria and 200 for cases that did not. If the 95% confidence interval for PPV was set to within ±0.1, the required number of cases that met the inclusion criteria was 100. Since the prevalence of CD is 55.6/100,000 in Japan [7], approximately 370,000 cases that did not meet the inclusion criteria were required to detect the exact sensitivity and specificity. However, to ensure feasibility, only 200 cases were selected.

Ethical considerations

The study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice guidelines. The Research Ethics Committee of Kitasato University Kitasato Institute Hospital approved the study protocol and all necessary documents (approval number: 19047). The study used data already recorded, and the ethics committee approved a waiver of informed consent.

Results

Case extraction and medical record review

A total of 82,898 patients who visited Kitasato University Kitasato Institute Hospital during the study period were enrolled, and 255 and 197 cases who met IC-A and B respectively, were extracted. Although 197 cases were selected for IC-C, they were excluded from later analyses because the number of cases that met IC-C was the same as IC-B (Fig 1). In Cohort-A, PPV was 83.0% for only confirmed diagnosis and 90.0% for confirmed and suspected diagnosis, and in Cohort-B, PPV was 97.0% for only confirmed diagnosis and 100.0% for confirmed and suspected diagnosis (Table 2) (The 2×2 tables are shown in S3 Table).

In Cohort-A, the positive predictive value (PPV) was 0.830 for confirmed and 0.900 for confirmed and suspected Crohn's disease (CD) cases. In Cohort-B, the PPV was 0.970 for

Table 2. Assessment of validity for each cohort.

Cohort	Diagnosis	TP	TN	FP	FN	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
A	Confirmed	83	200	17	0	1.000(0.957–1.000)	0.922(0.878–0.954)	0.830(0.742–0.898)	1.000(0.982–1.000)
	Confirmed & suspected	90	200	10	0	1.000(0.960–1.000)	0.952(0.914–0.977)	0.900(0.824–0.951)	1.000(0.982–1.000)
B	Confirmed	97	200	3	0	1.000(0.963–1.000)	0.985(0.957–0.997)	0.970(0.915–0.994)	1.000(0.982–1.000)
	Confirmed & suspected	100	200	0	0	1.000(0.964–1.000)	1.000(0.982–1.000)	1.000(0.964–1.000)	1.000(0.982–1.000)

*CD; Crohn's disease, TP; true-positive, TN; true-negative, FP; false-positive, FN; false-negative, PPV; positive predictive value, NPV; negative predictive value

https://doi.org/10.1371/journal.pone.0258537.t002

confirmed and 1.000 for confirmed and suspected CD cases. The negative predictive value (NPV) is 1.000 because there are no false-negative cases.

The characteristics of the patients who were diagnosed as confirmed and suspected cases in each cohort are shown in Table 3. In Cohort-A, 90 CD patients were diagnosed as confirmed and suspected cases [mean age 43.7±14.0, 62 males (68.9%)]; in Cohort-B (n = 100), the mean age was 44.3±14.7 and included 71 males (71.0%). In Cohort-A, 62% of the patients had CD confirmed by our medical records, 20% by an IBD specialist or gastroenterologist in another hospital, and 1% by primary care physicians without a description of the findings supporting the diagnosis (Fig 2). Of those, 7% were considered to have suspicious diagnoses and 10% were declared negative for CD, based on our medical records. Cases that were declared negative for CD included infectious enterocolitis (n = 4), intestinal Behçet's disease (n = 2), drug-induced enterocolitis (n = 1), intestinal tuberculosis (n = 1), unspecified intestinal stenosis (n = 1), and cirrhosis (n = 1). In Cohort-B, 74% of the patients had CD confirmed by our medical records, 23% by an IBD specialist or gastroenterologist in another hospital, and 3% were considered to have suspicious diagnoses. No cases were declared negative for CD in Cohort-B.

Inter- and intra-rater reliability

The inter- and intra-rater reliability are shown in Table 4. The weighted kappa coefficient of inter-rater reliability was 0.9903 and that of intra-rater reliability was 0.9948, suggesting that the diagnoses derived by medical record review were valid.

Discussion

In this study, we first developed algorithms to extract CD cases from the Japanese claims database by assessing the accuracy of claim codes validated by medical chart review.

For a disease with a low prevalence of CD, it is difficult to secure a sufficient number of cases from a single center. Murakami et al. reviewed the number of CD cases from various facilities, and the maximum number of cases at a single specialist center was approximately 320 [7], which is a small number of cases when compared to the 70,700 cases in Japan as a whole [22]. Another issue is that large-scale observational studies are usually conducted in specialist centers, including numerous non-specialized facilities, and may not reflect real-world practice. A large-scale study utilizing big data is therefore necessary to examine populations representing real-world practice. The insurance claims database is a useful resource and has been used in several important studies [23, 24].

Since Japan has a universal health insurance system and almost all citizens are enrolled, the Japanese claims database is a very useful resource for real-world data in database studies. In addition to the databases owned by the government (National Database), commercial databases from private companies are also available (JMDC, Medical Data Vision), which are under contract to different insurance payers, and which are used to conduct database research

Table 3. Baseline characteristics of each cohort.

	Cohort A (N = 90)	Cohort B (N = 100)
Age (mean ± SD, years)	43.65±13.99	44.33±14.66
Male, n (%)	62, 68.9%	71, 71.0%
Age at diagnosis (mean ± SD, years)	28.08±12.48	28.64±12.69
Disease duration, (mean ± SD, years)	10.52±9.13	11.46±10.80
Montreal Age at diagnosis, n (%)		
A1 (<16 years)	6 (6.7%)	4 (4.0%)
A2 (17–40 years)	68 (75.6%)	76 (76%)
A3 (>40 years)	13 (14.4%)	18(18%)
unknown	3 (3.3%)	2 (2%)
Montreal Location, n (%)		
L1 (ileal)	19 (21.1%)	21 (21.0%)
L2 (Colonic)	18 (20.0%)	15 (15.0%)
L3 (ileo-colonic)	49 (54.4%)	63 (63.0%)
+ isolated L4 (upper)	7 (7.8%)	11 (11.0%)
unknown	2 (2.2%)	1 (1.0%)
Montreal Behavior, n (%)		
B1 (Non-stricturing, non-penetrating)	46 (51.1%)	43 (43.0%)
B2 (Stricturing)	29 (32.2%)	27 (27.0%)
B3 (Penetrating)	13 (14.4%)	39 (30.0%)
+ perianal disease	29 (32.2%)	43 (43.0%)
unknown	2 (2.2%)	0 (0.0%)
Prior history of surgery, n (%)		
intestine	25 (27.7%)	36 (36.0%)
peri-anal	13 (14.4%)	21 (21.0%)
Intractable disease registration, n (%)	72 (80.0%)	92 (92.0%)
Review result, (confirmed / suspected)	83/7	97/3
Treatment		
Mesalazine, n (%)	71 (78.9%)	80 (80.0%)
Immunomodulator, n (%)	44 (48.9%)	51 (51.0%)
Elemental diet, n (%)	22 (24.4%)	26 (26.0%)
Corticosteroid, n (%)	12 (13.3%)	13 (13.0%)
Prednisolone, n (%)	6 (6.7%)	6 (6.0%)
Budesonide, n (%)	6 (6.7%)	7 (7.0%)
Biologics, n (%)	37 (41.1%)	46 (46.0%)
Infliximab, n (%)	20 (22.2%)	27 (27.0%)
Infliximab BS, n (%)	1 (1.1%)	1 (1.0%)
Adalimumab, n (%)	16 (17.8%)	18 (18.0%)
Ustekinumab, n (%)	1 (1.1%)	1 (1.0%)
Vedolizumab, n (%)	0 (0.0%)	0 (0.0%)

*SD; standard deviation.

https://doi.org/10.1371/journal.pone.0258537.t003

and to support hospital management by analyzing medical costs. The National Database is a public database that contains data supporting more than 1 billion claims, as well as data and information on specific legal health checkups and guidance [20]. The Diagnosis Procedure Combination (DPC) database holds medical information of inpatients from 1,730 DPC-registered hospitals captured in 2018. The JMDC database is a commercial database that contains

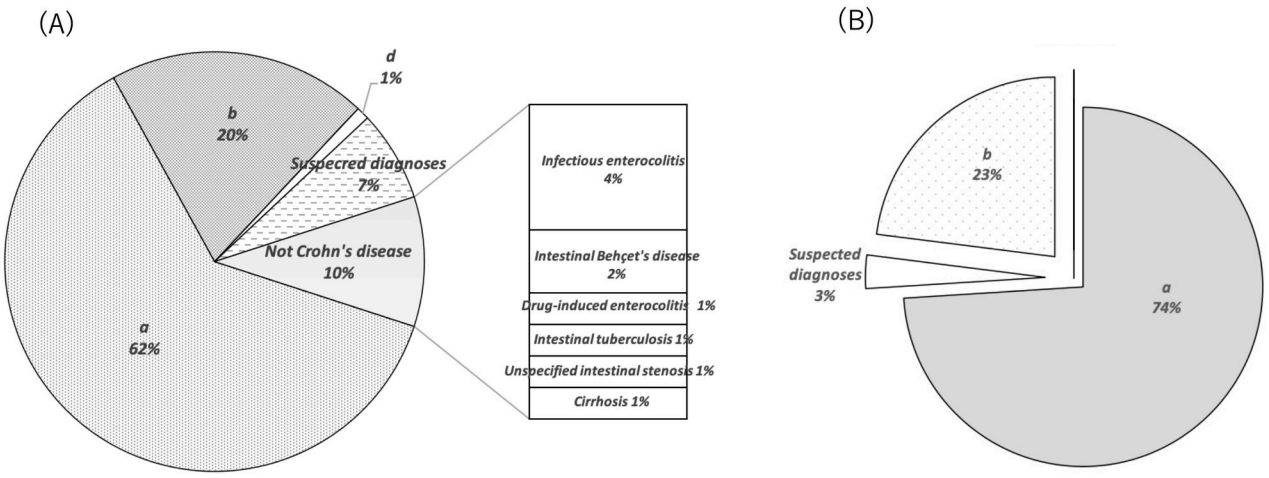

Fig 2. Details of medical chart review. (A) Details of medical chart review for Cohort-A. 83% of cases were confirmed for Crohn's disease (CD) (confirmed diagnosis at own institution, or another hospital). 7% were considered suspected diagnosis. Cases denied for CD (10%) included infectious enterocolitis (n = 4), intestinal Behçet's disease (n = 2), drug-induced enterocolitis (n = 1), intestinal tuberculosis (n = 1), unspecified intestinal stenosis (n = 1), and cirrhosis (n = 1). (B) Details of medical chart review for Cohort-B. 97% of cases were confirmed CD. 3% were considered suspected diagnoses. None of the cases were denied CD. *a; Confirmed diagnosis at own institution, b; Diagnosed by an IBD specialist or gastroenterologist in another hospital, c; Diagnosed by a primary care physician (with a description of the findings supporting the diagnosis), d; Diagnosed by a primary care physician (without a description of the findings supporting the diagnosis), CD; Crohn's disease.

https://doi.org/10.1371/journal.pone.0258537.g002

claims data for up to 7.3 million insured individuals, which represents approximately 6.1% of the Japanese population between 2005 and April 2020 and includes some salaried employees and their families. A previous study extracted 150 CD cases treated with biologic agents from this database [15]. The Medical Data Vision database is a commercial database that contains data on about 29.8 million patients who received treatment from approximately 400 DPC hospitals in Japan between April 2008 and October 2019. According to a previous database study, about 75,000 CD and ulcerative colitis cases were registered [25]. The validation of our present study is based on claims filed from the medical provider independently of payers; therefore, it is expected to be applicable in any of the claims databases.

There have been many studies on other diseases using the various databases mentioned above. Some utilized prior validation studies [26, 27], but others did not [11, 28]. However, it is possible that a lack of validated algorithms may significantly reduce the reliability of each database study. It is, therefore, extremely important to develop a validated algorithm to extract target diseases from the relevant databases [29].

In fact, in a previous study, the PPV was often remarkably low (60%) for extractions with only a single disease code, while an acceptable PPV (82–91%) was achieved by using repeated detection of the disease code as the extraction protocol [30]. In this study, we found that extraction by diagnostic codes alone (Cohort-A) resulted in the inclusion of other diseases, such as infectious enteritis and Behçet's disease, which suggests that extraction from claims data by ICD-10 code alone is not sufficient. The PPV of confirmed CD cases Cohort-A of this study was 83%. In general, other studies have set the target PPV as 85% or higher [31]. These

Table 4. Inter- and intra-rater reliability of the medical chart review.

	Kappa (95% CI)	Weighted Kappa (95% CI)	Gwet's AC1 (95% CI)
Inter-rater reliability	0.9634(0.9136–1.0000)	0.9903(0.9768–1.0000)	0.9784(0.9481–1.0000)
Intra-rater reliability	0.9816(0.9457–1.0000)	0.9948(0.9845–1.0000)	0.9892(0.9678–1.0000)

https://doi.org/10.1371/journal.pone.0258537.t004

PPV values in Cohort-A did not reach this level. However, Cohort-B, in which prescription codes were added, resulted in a remarkably improved PPV of 97.0%. This is comparable to the PPVs for other diseases in Japan [19, 32, 33] and is therefore considered to be acceptable for general extraction algorithms. The number of cases extracted by IC-B and IC-C, which had additional surgical codes was the same (n = 197). In other words, most cases that underwent a CD-related procedure or surgery were likely to receive the prescription code for CD at the same time, showing that there was little significance in adding the procedure or surgery code.

Some algorithms have been used to extract IBD from other claims databases, such as the algorithm for the Korean National database, which achieved a PPV of approximately 98% by combining the ICD-10 codes, treatment with the incurable disease application code, and the number of hospital visits for IBD [34]. CD is one of the diseases included in the Intractable Disease Registry by the Ministry of Health, Labor and Welfare. However, a certain proportion of patients (20.0% of cohort A and 8.0% of cohort B) were not registered in the registry. This means that Intractable Disease Registry may not necessarily reflect the real world.

Other algorithms that combine Ninth Revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-9) codes with the number of visits and hospitalizations have also reported good PPVs: 81–91% from the Veterans Affairs Health Care System and 94–98% from the Canadian claims database. Ananthakrishnan et al. [35] also reported a PPV of 98% by combining the ICD-9 codes, medical record information, and the complications of IBD for claims data from two tertiary referral hospitals. The results of this study are also comparable to those of other such studies.

We confirmed the extracted patient population in two additional ways. Inter- and intra-rater agreements of the chart review results were confirmed to ensure the reliability of the validation. In addition, the validated cohort in our study was similar to the characteristics of patients in terms of the sex ratio, Montreal classification, prior history of surgery, and previously reported treatment from other specialist centers [35–38].

This study has several limitations. First, although the algorithm developed in this study successfully demonstrated excellent PPV, it is important to note that the study was conducted at a single specialist center, where the prior probability of CD patients among all patients is likely to be much higher than that in the non-specialist centers. Therefore, it is possible that PPV was overestimated compared to real-world clinical practice. In addition, it is also unclear whether cases extracted from the claims database using our algorithm would represent real-world practice in the entire patient population. Further studies to validate our algorithm are warranted from various types of facilities, including non-specialist general hospitals and private clinics.

Second, it is possible that IC-B may inappropriately exclude the true CD patients who have stopped medications and are no longer prescribed. Using the PPV in this study, the numbers of patients who met the IC–A and B before random sampling were estimated to be 211 (PPV 83%) and 191 (PPV 97%), respectively. In other words, IC-B might have excluded approximately 20 patients who had no prescription for several years. If the aim of the study requires to extract such patients together, IC-A should be used with caution to its low PPV. However, considering the disease behavior of CD, it is very rare that a whole set of treatment is discontinued for several years once it has been prescribed.

Third, although the sensitivity (100%), specificity (92–100%), and NPV (100%) shown in our study were excellent, the sample size considered sufficient to accurately calculate these parameters was calculated as 37,000, considering the actual prevalence of CD (55.6/100,000), and thus our sample size (200 cases) is too small. Therefore, the accuracy of these parameters cannot be assured, and 2 × 2 tables with adjusted weights are also assumptive (S3 Table). However, PPV is generally considered to be the most important to develop the extraction

algorithms. Moreover, the sample size required for the calculation of PPV is reported to be much smaller [39, 40]. Therefore, our algorithm is still likely to help appropriately define CD cases from the large-scale claims database.

In conclusion, this study established an algorithm to extract CD from the Japanese claims database and will be of importance in future large-scale real-world studies using the claims database.

Supporting information

S1 Table. ICD-10 diagnostic code.
(XLSX)

S2 Table. Prescription codes and surgical codes for this study.
(XLSX)

S3 Table. A 2 × 2 contingency tables for inclusion criteria and validation criteria. The number listed is the actual number of validated cases, and the number in parentheses is the assumed number of cases in the entire Kitasato Research Institute Hospital, calculated based on the prevalence calculated from all cases extracted in this study (82,898).
(XLSX)

S1 File.
(DOCX)

Acknowledgments

The authors are grateful to Hiroki Kiyohara, Yuki Watanabe (Center for Advanced IBD Research and Treatment, Kitasato University Kitasato Institute Hospital), Takashi Tanaka, Katsuhiko Nagai (Japan Medical Data Center Co., Ltd.) for their assistance in this study.

Author Contributions

Conceptualization: Hiromu Morikubo, Taku Kobayashi, Takayoshi Nagahama.

Data curation: Hiromu Morikubo, Tomohiro Fukuda, Takayoshi Nagahama.

Formal analysis: Hiromu Morikubo, Taku Kobayashi, Tomohiro Fukuda, Takayoshi Nagahama.

Funding acquisition: Hiromu Morikubo, Taku Kobayashi.

Investigation: Hiromu Morikubo, Taku Kobayashi, Tomohiro Fukuda.

Methodology: Hiromu Morikubo, Taku Kobayashi, Takayoshi Nagahama.

Project administration: Hiromu Morikubo, Taku Kobayashi, Takayoshi Nagahama.

Supervision: Taku Kobayashi, Tadakazu Hisamatsu, Toshifumi Hibi.

Validation: Hiromu Morikubo, Taku Kobayashi.

Writing – original draft: Hiromu Morikubo, Taku Kobayashi.

Writing – review & editing: Hiromu Morikubo, Taku Kobayashi, Tomohiro Fukuda, Takayoshi Nagahama, Tadakazu Hisamatsu, Toshifumi Hibi.

References

1. Matsuoka K., Kobayashi T., Ueno F., Matsui T., Hirai F., Inoue N., et al., Evidence-based clinical practice guidelines for inflammatory bowel disease. J Gastroenterol, 2018. 53(3): p. 305–353.
2. Feagan B.G., Sandborn W.J., Gasink C., Jacobstein D., Lang Y., Friedman J.R., et al., Ustekinumab as Induction and Maintenance Therapy for Crohn's Disease. N Engl J Med, 2016. 375(20): p. 1946–1960. https://doi.org/10.1056/NEJMoa1602773 PMID: 27959607
3. Issa J., D. and David M.K., Can the learning health care system be educated with observational data? JAMA, 2014. 312: p. 129–130. https://doi.org/10.1001/jama.2014.4364 PMID: 25005647
4. Ha C., Ullman T.A., Siegel C.A., and Kornbluth A., Patients enrolled in randomized controlled trials do not represent the inflammatory bowel disease patient population. Clin Gastroenterol Hepatol, 2012. 10 (9): p. 1002–7; quiz e78. https://doi.org/10.1016/j.cgh.2012.02.004 PMID: 22343692
5. Ng S.C., Shi H.Y., Hamidi N., Underwood F.E., Tang W., Benchimol E.I., et al., Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. The Lancet, 2017. 390(10114): p. 2769–2778. https://doi.org/10.1016/S0140-6736(17)32448-0 PMID: 29050646
6. Vegh Z., Kurti Z., and Lakatos P.L., Epidemiology of inflammatory bowel diseases from west to east. J Dig Dis, 2017. 18(2): p. 92–98. https://doi.org/10.1111/1751-2980.12449 PMID: 28102560
7. Murakami Y., Nishiwaki Y., Oba M.S., Asakura K., Ohfuji S., Fukushima W., et al., Estimated prevalence of ulcerative colitis and Crohn's disease in Japan in 2014: an analysis of a nationwide survey. J Gastroenterol, 2019. 54(12): p. 1070–1077. https://doi.org/10.1007/s00535-019-01603-8 PMID: 31309327
8. Kosa F., Kunovszki P., Borsi A., Ilias A., Palatka K., Szamosi T., et al., Anti-TNF dose escalation and drug sustainability in Crohn's disease: Data from the nationwide administrative database in Hungary. Dig Liver Dis, 2020. 52(3): p. 274–280. https://doi.org/10.1016/j.dld.2019.09.020 PMID: 31669077
9. Bergmann M.M., Hernandez V., Bernigau W., Boeing H., Chan S.S., Luben R., et al., No association of alcohol use and the risk of ulcerative colitis or Crohn's disease: data from a European Prospective cohort study (EPIC). Eur J Clin Nutr, 2017. 71(4): p. 512–518. https://doi.org/10.1038/ejcn.2016.271 PMID: 28120853
10. Sruamsiri R., Iwasaki K., Tang W., and Mahlich J., Persistence rates and medical costs of biological therapies for psoriasis treatment in Japan: a real-world data study using a claims database. BMC Dermatol, 2018. 18(1): p. 5. https://doi.org/10.1186/s12895-018-0074-0 PMID: 29996929
11. Uno S., Goto R., Suzuki K., Iwasaki K., Takeshima T., and Ohtsu T., Current treatment patterns and medical costs for multiple myeloma in Japan: a cross-sectional analysis of a health insurance claims database. J Med Econ, 2020. 23(2): p. 166–173. https://doi.org/10.1080/13696998.2019.1686870 PMID: 31682533
12. Sato H., Yokomichi H., Takahashi K., Tominaga K., Mizusawa T., Kimura N., et al., Epidemiological analysis of achalasia in Japan using a large-scale claims database. J Gastroenterol, 2019. 54(7): p. 621–627. https://doi.org/10.1007/s00535-018-01544-8 PMID: 30607612
13. Schwartz D.A., Tagarro I., Carmen Diez M., and Sandborn W.J., Prevalence of Fistulizing Crohn's Disease in the United States: Estimate From a Systematic Literature Review Attempt and Population-Based Database Analysis. Inflamm Bowel Dis, 2019. 25(11): p. 1773–1779. https://doi.org/10.1093/ibd/izz056 PMID: 31216573
14. Kobayashi T., Udagawa E., Uda A., Hibi T., and Hisamatsu T., Impact of immunomodulator use on treatment persistence in patients with ulcerative colitis: A claims database analysis. J Gastroenterol Hepatol, 2020. 35(2): p. 225–232. https://doi.org/10.1111/jgh.14825 PMID: 31397010
15. Yokoyama K., Yamazaki K., Katafuchi M., and Ferchichi S., A Retrospective Claims Database Study on Drug Utilization in Japanese Patients with Crohn's Disease Treated with Adalimumab or Infliximab. Adv Ther, 2016. 33(11): p. 1947–1963. https://doi.org/10.1007/s12325-016-0406-6 PMID: 27664107
16. Eindhoven D.C., van Staveren L.N., van Erkelens J.A., Ikkersheim D.E., Cannegieter S.C., Umans V., et al., Nationwide claims data validated for quality assessments in acute myocardial infarction in the Netherlands. Neth Heart J, 2018. 26(1): p. 13–20. https://doi.org/10.1007/s12471-017-1055-3 PMID: 29119544
17. Langner I., Ohlmeier C., Haug U., Hense H.W., Czwikla J., and Zeeb H., Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data. BMJ Open, 2019. 9(7): p. e026834. https://doi.org/10.1136/bmjopen-2018-026834 PMID: 31350240
18. Nakayama T., Imanaka Y., Okuno Y., Kato G., Kuroda T., Goto R., et al., Analysis of the evidence-practice gap to facilitate proper medical care for the elderly: investigation, using databases, of utilization measures for National Database of Health Insurance Claims and Specific Health Checkups of Japan

(NDB). Environ Health Prev Med, 2017. 22(1): p. 51. https://doi.org/10.1186/s12199-017-0644-5 PMID: 29165139

19. Ando T., Ooba N., Mochizuki M., Koide D., Kimura K., Lee S.L., et al., Positive predictive value of ICD-10 codes for acute myocardial infarction in Japan: a validation study at a single center. BMC Health Serv Res, 2018. 18(1): p. 895. https://doi.org/10.1186/s12913-018-3727-0 PMID: 30477501
20. *Ministry of Health, Labour and Welfare*. https://www.mhlw.go.jp/english/index.html.
21. Ikegami N., Yoo B.-K., Hashimoto H., Matsumoto M., Ogata H., Babazono A., et al., Japanese universal health coverage: evolution, achievements, and challenges. The Lancet, 2011. 378(9796): p. 1106–1115.
22. *Japan Intractable Disease Information Center.*; http://www.nanbyou.or.jp.
23. Clayton J.L., Jones S.G., Dunn J.R., Schaffner W., and Jones T.F., Enhancing Lyme Disease Surveillance by Using Administrative Claims Data, Tennessee, USA. Emerg Infect Dis, 2015. 21(9): p. 1632–4. https://doi.org/10.3201/eid2109.150344 PMID: 26291336
24. Paller A.S., Siegfried E.C., Vekeman F., Gadkari A., Kaur M., Mallya U.G., et al., Treatment patterns of pediatric patients with atopic dermatitis: A claims data analysis. J Am Acad Dermatol, 2020. 82(3): p. 651–660. https://doi.org/10.1016/j.jaad.2019.07.105 PMID: 31400453
25. Kobayashi T., Uda A., Udagawa E., and Hibi T., Lack of Increased Risk of Lymphoma by Thiopurines or Biologics in Japanese Patients with Inflammatory Bowel Disease: A Large-Scale Administrative Database Analysis. J Crohns Colitis, 2020. 14(5): p. 617–623. https://doi.org/10.1093/ecco-jcc/jjz204 PMID: 31867632
26. Yamana H., Moriwaki M., Horiguchi H., Kodan M., Fushimi K., and Yasunaga H., Validity of diagnoses, procedures, and laboratory data in Japanese administrative data. J Epidemiol, 2017. 27(10): p. 476–482. https://doi.org/10.1016/j.je.2016.09.009 PMID: 28142051
27. Hara K., Tomio J., Svensson T., Ohkuma R., Svensson A.K., and Yamazaki T., Association measures of claims-based algorithms for common chronic conditions were assessed using regularly collected data in Japan. J Clin Epidemiol, 2018. 99: p. 84–95. https://doi.org/10.1016/j.jclinepi.2018.03.004 PMID: 29548842
28. Izumi K., Morimoto K., Hasegawa N., Uchimura K., Kawatsu L., Ato M., et al., Epidemiology of Adults and Children Treated for Nontuberculous Mycobacterial Pulmonary Disease in Japan. Ann Am Thorac Soc, 2019. 16(3): p. 341–347. https://doi.org/10.1513/AnnalsATS.201806-366OC PMID: 30339468
29. Benchimol E.I., Manuel D.G., To T., Griffiths A.M., Rabeneck L., and Guttmann A., Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. J Clin Epidemiol, 2011. 64(8): p. 821–9. https://doi.org/10.1016/j.jclinepi.2010.10.006 PMID: 21194889
30. Hou J.K., Tan M., Stidham R.W., Colozzi J., Adams D., El-Serag H., et al., Accuracy of diagnostic codes for identifying patients with ulcerative colitis and Crohn's disease in the Veterans Affairs Health Care System. Dig Dis Sci, 2014. 59(10): p. 2406–10. https://doi.org/10.1007/s10620-014-3174-7 PMID: 24817338
31. Lacasse Y., Daigle J., Martin S., and Maltais F., Validity of chronic obstructive pulmonary disease diagnoses in a large administrative database. Can Respir J, 2012. 19: p. e5–9. https://doi.org/10.1155/2012/260374 PMID: 22536584
32. Sato I., Yagata H., and Ohashi Y., The Accuracy of Japanese Claims Data in Identifying Breast Cancer Cases. Biol Pharm Bull, 2015. 38(1): p. 53–57. https://doi.org/10.1248/bpb.b14-00543 PMID: 25744458
33. Ooba N., Setoguchi S., Ando T., Sato T., Yamaguchi T., Mochizuki M., et al., Claims-based definition of death in Japanese claims database: validity and implications. PLoS One, 2013. 8(5): p. e66116. https://doi.org/10.1371/journal.pone.0066116 PMID: 23741526
34. Lee C.K., Ha H.J., Oh S.J., Kim J.W., Lee J.K., Kim H.S., et al., Nationwide validation study of diagnostic algorithms for inflammatory bowel disease in Korean National Health Insurance Service database. J Gastroenterol Hepatol, 2020. 35(5): p. 760–768. https://doi.org/10.1111/jgh.14855 PMID: 31498502
35. Ananthakrishnan A.N., Cai T., Savova G., Cheng S.C., Chen P., Perez R.G., et al., Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflamm Bowel Dis, 2013. 19(7): p. 1411–20. https://doi.org/10.1097/MIB.0b013e31828133fd PMID: 23567779
36. Yasukawa S., Matsui T., Yano Y., Sato Y., Takada Y., Kishi M., et al., Crohn's disease-specific mortality: a 30-year cohort study at a tertiary referral center in Japan. J Gastroenterol, 2019. 54(1): p. 42–52. https://doi.org/10.1007/s00535-018-1482-y PMID: 29948302
37. Huang S., Li L., Ben-Horin S., Mao R., Lin S., Qiu Y., et al., Mucosal Healing Is Associated With the Reduced Disabling Disease in Crohn's Disease. Clin Transl Gastroenterol, 2019. 10(3): p. e00015. https://doi.org/10.14309/ctg.0000000000000015 PMID: 30839440

38. Peyrin-Biroulet L., Loftus E.V. Jr., Colombel J.F., and Sandborn W.J., The natural history of adult Crohn's disease in population-based cohorts. Am J Gastroenterol, 2010. 105(2): p. 289–97. https://doi.org/10.1038/ajg.2009.579 PMID: 19861953

39. Cutrona S.L., Toh S., Iyer A., Foy S., Cavagnaro E., Forrow S., et al., Design for validation of acute myocardial infarction cases in Mini-Sentinel. Pharmacoepidemiol Drug Saf, 2012. 21 Suppl 1: p. 274–81. https://doi.org/10.1002/pds.2314 PMID: 22262617

40. Semins M.J., Trock B.J., and Matlaga B.R., Validity of administrative coding in identifying patients with upper urinary tract calculi. J Urol, 2010. 184(1): p. 190–2. https://doi.org/10.1016/j.juro.2010.03.011 PMID: 20478584