Doctoral Dissertation

A Methodology of Dataset Generation for Secondary Use of Health Care Big Data

Tomohide Iwao

March 2020
Graduate School of Informatics
Kyoto University
Supervisor: Prof. Tomohiro Kuroda

A Methodology of Dataset Generation for Secondary Use of Health Care Big Data

Tomohide Iwao

Abstract

Healthcare data include electronic medical record data, health insurance claims data, health check-up data etc. and is obtained by medical, health, welfare, and other specialists when diagnosing or examining on human subjects. Recently, the data is digitized and stored in respective database, and the number of medical institutions that store data that may be called big data in terms of quantity has increased. Among them, secondary use of healthcare databases has attracted attention in clinical and epidemiological study for the prevention of disease and verification of treatment methods.

However, at present, healthcare databases are not actively used. One of the reasons is that data providers such as countries have strict security policies and laws and regulations on secondary use of such data. On the other hand, even at the data analysis stage, there are challenges that arise because the healthcare database is not designed for secondary use. The first challenge is related with data contents that the attributes required for epidemiological analysis are not always stored in the healthcare database. The second challenge is related to the data structure in which information about patients is distributed in many relations because the healthcare data is generally stored in a normalized form in the database.

In epidemiological analysis, tabulation and statistical analysis are often performed according to the purpose, so it is desirable to create a dataset summarized in one tuple for each patient. When using a health care database for epidemiological analysis, it is necessary to resolve the challenges related to the data contents and data structure. As a result, considerable programming skills are required for data handling to a create dataset. For this reason, it is difficult for physician-scientists who have little skills of data handling to quickly conduct epidemiological studies using a health care database. The purpose of this research is to establish a methodology that enables even researchers with little data handling skills, such as programming, to create the dataset from health care databases on their own. As a material, the health insurance claims database known as one of health care databases is used.

This research focused on the fact that the relational model theory can be used to create a dataset summarized in one tuple for each patient. In this method, the research question is regarded as a proposition and divided into multiple propositions. This idea separates the process of generating a

dataset into a process of computing each attribute and a process of solving independent constraints between the attributes defined by the research question with the truth function.

In the process of computing attributes, this research proposes a simple and patterned relational calculus that can compute the desired attributes. Moreover, by applying the closure property, which is one of the features of the relational model theory, the simplicity of the relational calculus is maintained even in the case where there is a constraint condition dependent on any attributes between each attribute. For attributes that cannot be physically covered, a data warehouse optimized for the relational calculus is constructed. As a result of conducting five epidemiological studies using the proposed method, physician-scientists themselves were able to carry out data handling on their own for two epidemiological studies. These five epidemiological studies include cross-sectional and retrospective cohort studies that dominate epidemiological studies using helth care databases, and it can be confirmed that relational calculus cover not less than 80% of attributes in the original database. Therefore, although the scope of application is limited to epidemiological studies, it is considered that the usefulness of this method has been verified to a certain extent. Moreover, since the method is applicable to any database that includes unique ID for people such as patient, it can be applied not only to the health insurance claims databases used in this study, but also to many healthcare databases such as electronic medical records. This research simplifies the process of creating a dataset summarized in one tuple for each patient suitable for epidemiological analysis, and uses a highly versatile relational model theory as a solution. For this reason, the process of dataset generation can be implemented as software. This research is expected to initiate development of more practical research platform such as automation of dataset generation and statistical analysis.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Healthcare data include electronic medical record data, health insurance claims data, health check-up data etc. and is obtained by medical, health, welfare, and other specialists when diagnosing or examining on human subjects. Recently, the data is digitized and stored in respective database, and the number of medical institutions that store data that may be called big data in terms of quantity has increased. Healthcare databases can be used at low cost because it is automatically accumulated with daily medical care. Moreover, it can be effectively used in large number of fields if reliability of the health care data collected are qualitatively high[1]. Applications can include, for example, the diagnosis and selection of the optimum treatment by medical specialists, epidemiological study, and health management for individuals.

Among them, secondary use of healthcare databases has attracted attention in clinical and epidemiological study for the prevention of disease and verification of treatment methods. Here, secondary use refers to using data for a purpose different from the original purpose[2]. Healthcare databases are typically used by researchers as data sources for a purpose different from the research activity for which it is being used[3]. However, at present, healthcare databases are not actively used.

One of the reasons is that data providers such as countries have strict security policies and laws and regulations on secondary use of such data. These are the fundamental challenges encountered before embarking on data analysis. On the other hand, even at the data analysis stage, there are challenges that arise because the healthcare database is not designed for secondary use.

This research contributes to the promotion of research using a healthcare insurance claims database which is one of the healthcare databases by solving the challenges. In this chapter, after describing the features of epidemiological research using the health care database, the challenges in analyzing data are described, and finally the positioning of this research is explained.

1.1. Epidemiological study using health care databases

Epidemiological studies can be broadly divided into interventional studies and observational studies[4]. Interventional studies are ones that involve medical invasion of patients. On the other hand, observational research refers to one that does not invade patients. Epidemiological studies using health care databases are classified as observational studies because existing data are used.

Observational study

Observational studies generally consist of longitudinal, cross-sectional, and case-control studies [4]. Observational studies include retrospective studies and prospective studies. Longitudinal studies consist of a prospective cohort study that collects data prospectively from the start of the study and a retrospective cohort study that explores data retrospectively. A cross-sectional study is a study that

investigates a disease or exposure of people at one point in time. Case-control study is a retrospective study that seeks the cause by searching data of people who have a disease. Epidemiological studies using health care databases for secondary use are limited to retrospective cohort, cross-sectional, and case-control studies because they deal with existing data already collected.

Health insurance claims database

There are various types of health care databases, such as those based on electronic medical records[5] and collected for research purposes[6]. In general, healthcare databases often store a unique patient ID for each patient, an event such as a treatment or medication performed on the patient, and a timing such as a date or time when the event was performed. Among them, health insurance claims data have a high degree of completeness in countries with universal health care systems since comprehensive information of almost all people can be obtained. As a result, using data from such countries in large-scale epidemiological surveys and studies of rare diseases is to be expected. Moreover, because of the abovementioned reasons, the number of health insurance claims data items can be enormous and has volume and velocity, which are required for big data [7]. For that reason, this research focuses on health insurance claims data related to health care bills, which is one type of health care data.

There are several countries that accept secondary use of health insurance claims data, and their usefulness in epidemiological studies is recognized[8 9 10]. Many epidemiological studies have been conducted by researchers in Taiwan, South Korea, France, and Japan etc. where the universal health insurance system has been introduced [11]. Moreover, in the United States, although there is no universal insurance system, it manages the claims data of Medicare and Medicaid, and analysis support for researchers is progressing [12 13]. Thus, in recent years, epidemiological studies using health insurance claims database have been actively conducted worldwide[14,15].

Procedure of epidemiological study using database

In general, epidemiological studies consist of hypothesis building and hypothesis testing [4]. Hypothesis building is almost the same as building research questions, and exploratory analysis is sometimes used in epidemiological studies using databases [16 17]. On the other hand, in the hypothesis testing process, it is necessary to statistically analyze the relationship between various attributes related to patients using statistical analysis software.

1.2 Challenges of epidemiological study using database

Available health care data has been collected mainly for its original purposes (i.e., primary objectives) without assuming there would be such a secondary use in epidemiological research. As a result, when the data is subjected to secondary use in epidemiological research, there are challenges related to the reliability of the data and regulations for the protection of personal information. Especially regarding data reliability, bias in the database, which can be introduced through confounding, missing data, and misclassification, is a great threat to the validity of epidemiologic research[3]. These problems are difficult to solve immediately, but there are also technically solvable problems. These are challenges related to data content and data structure, and the secondary use of data cannot be maximized without overcoming these two main challenges.

Challenges of data contents

As for data contents of the databases, since the primary purpose of healthcare databases[18] concerns health insurance claims data, information would otherwise commonly be required for epidemiological studies may not be stored. Thus, researchers must create their own markers or indices and add them to the data using external master data, which is time consuming and inefficient.

Challenges of data structure

As for data structure of the databases, the majority of the databases of health insurance claims currently in use are relational databases (RDB); these are designed to support the relational model[19], which is created based on the set theory and logic. In the relational model, data are distributed across multiple relations, which aims to normalize redundant data and increase the efficiency of data storage and updates[20]. Normalization, which results in splitting the data up into small relations, is the usual guideline for storing relational data [21]. However, it is said that clinical data are distributed across several tables that is difficult to use for analysis [22 23].

In epidemiological analysis, tabulation and statistical analysis are often performed according to the purpose, so it is desirable to create a dataset summarized in one tuple for each patient. When using a health insurance claims database as a source of secondary use data for epidemiological analysis, it is necessary to resolve the challenges related to the data contents and data structure. For that reason, it has been reported that data handling of claims data requires complex and time-consuming [24], and the ratio of data handling in data analysis is 50% to 80% [25]. Data analysis is not easy for those who are not used to data handling of databases.

The goal of this dissertation is to develop a methodology that can allow researchers who have little skills in data handling to carry out data handling by themselves while solving above-mentioned challenges regarding data contents and data structure.

This dissertation is organized as follows. Chapter 2 describes related research including data warehousing etc. Chapter 3 describes the health insurance claims data used in this research and the proposed method. Chapter 4 describes epidemiological studies conducted using the proposed method. Chapter 5 discusses the effectiveness of the proposed method based on the results. Finally, Chapter 6 summarizes this dissertation and describes future applications.

Chapter 2 Related researches

This dissertation is based on great works by many researchers, especially on the relational model theory. First, after explaining the relational model theory, this chapter describes several researches on the data warehouse (DW) including the dimensional model which is a technology often used for exploratory analysis using databases. Finally, the purpose of this dissertation is described.

2.1 Relational model theory

The relational data model theory is a database management system theory developed in 1970 by Edgar Frank "Ted" Codd[19]. It was devised based on predicate logic and set theory. The following is a simple explanation of predicate logic and set theory.

2.1.1 Predicate logic and set theory

The concept of predicate logic is an extension of propositional logic. Propositional logic is the study dealing with the truth or falsehood of propositions.

Proposition #1:

Patient X1 is female

The “X1” part of Patient X1 in Proposition #1 can be given various names, and in predicate logic, it is called a variable[26]. The phrase “is female” in Proposition #1 expresses a quality or relationship of this variable, and this is called a predicate. More general propositions such as the one below can be given by increasing the number of predicates.

Proposition #2:

Patient X1 is a female aged 57 years’ old who was diagnosed on September 24, 2007 with stroke and died on October 15, 2007

Increasing the number of variables thus enables the handling of large number of propositions, which can be displayed in the form of tables in Table 2.1. Each line in the table corresponds to a single proposition. Codd focused on the fact that if the table headings {Patient, Sex, Age, Diagnosis Date, Disease name, Death date} were considered as a set of elements, they could be handled as the

data. As a result, each of the records in the table below is simultaneously a proposition and a set composed of individual elements.

As described above, Codd proposed a database management system that defined general data constructs based on predicate logic and set theory. This data model has two significant advantages. The first is that a data model can be expressed in the form of a table. This enables the creation of interfaces that are highly intuitive and easily visualized for users.

Table 2.1 Propositions in table format

Patient	Sex	Age	Diagnosis date	Disease name	Death date
X1	Female	57	2007-09-04	Cerebral infarction	2007-10-15
X2	Male	62	1985-06-05	Liver cancer	1988-11-30
X3	Female	89	2013-02-15	Nontuberculous mycobacteria	2016-03-15
・ ・	・ ・	・ ・	・ ・	・ ・	・ ・
XN	Male	54	1988-02-03	Cerebral hemorrhage	1988-02-05

2.1.2 Truth function, Extend and Summarize

The second advantage is that it enables efficient data manipulation by means of the use of truth functions and relational algebra, which are reliant on predicate logic and set theory, respectively, and are suited to computer processing. In addition to the truth function, operators of “Extend” and “Summarize” have been added. The Extend operator has the function to compute arbitrary attributes in one relation and add the result as a new attribute. For example, in Table 2.1, if you want to compute the number of days from diagnosis date to the date of death, you can add the result of subtracting these attributes (death date minus (-) diagnosis date) by the Extend operator as a new attribute. Summarize operator aggregates multiple tuples in one relation into one tuple.

Thanks to these two advantages, the relational data model has been broadly accepted worldwide. Today, various vendors have released Relational Database (RDB) products supporting the relational data model, and new versions are repeatedly being released. Moreover, above mentioned “Extend”, and “Summarize” operators are implemented in the SQL, which is incorporated into the various RDB products.

The SQL is a database language implemented for data manipulation in the relational data model. Strictly speaking, SQL is not a specification of the relational data model, but one form of implementation that meets its specifications. It communicates with the database management system (DBMS) to implement mainly data definitions and manipulation. The current standard is SQL99, and this is extended in individually different ways in products from different DBMS vendors[27]. Demand for analysis functions has risen in recent years, and Online Analytical Processing (OLAP) functions are now attracting attention[28]. The OLAP technology pioneered data warehouses described in the next section.

2.2 Data Warehouse

A well-known approach for supporting analyses that involve relational databases is using DW. A DW is integrated data that stores data in a time series without regard for their later deletion or updating. Concept of a DW was first proposed by William H. Inmon in 1990 as data structure that is suited to data analysis for the purpose of decision-making [29]. DWs are data aggregates that integrate time series without assuming data deletion or updating, and this structure has now been introduced by many organizations, mainly in the form of core database systems for business use.

2.2.1 Dimensional Model

A typical example of DWs is the dimensional model, which is a data model proposed in the book written by Ralph Kimball in 1996[30]. The dimensional model is based on data structure that is particularly suitable for decision-making and many studies have been conducted so far[31,32]. In recent years, it has been often used to increase the speed of database inquiries when developing software to facilitate exploratory analyses, one major example of which is business intelligence (BI) [33], which requires good response times. In recent years, database-related products that support the dimensional model have been released[34].

As shown in Fig.2.1, the dimensional model is composed of two individual tables, Fact and Dimension. The Dimensional Model includes “Fact Constellation” with multiple Facts, “Snowflake” with normalized dimensions, and “Star Shema” with only one Fact[35]. The star schema has the simplest structure of dimensional model and is considered suitable for BI tools[33]. The Fact table is generally the largest table and includes the attributes to be analyzed. The Dimension table is a table that contains analyzable attributes by virtue of its linkage to the Fact table and corresponds to the master table of a regular database. In most cases, its size is therefore smaller than that of the Fact table. In the following example, Fact is only the observation fact table, but several such tables may exist. The attributes listed in the Fact table have already been tabulated, and if any other data are needed, these can be obtained by linking to an appropriate Dimension table as required. However,

although the attributes listed in the Fact table are suitable for queries involving rapid searches and aggregations, they do not possess the means to deal with other queries with more complex conditions. As almost all the data required for decision-making in business is patterned, most cases can be dealt with by application of the dimensional model.

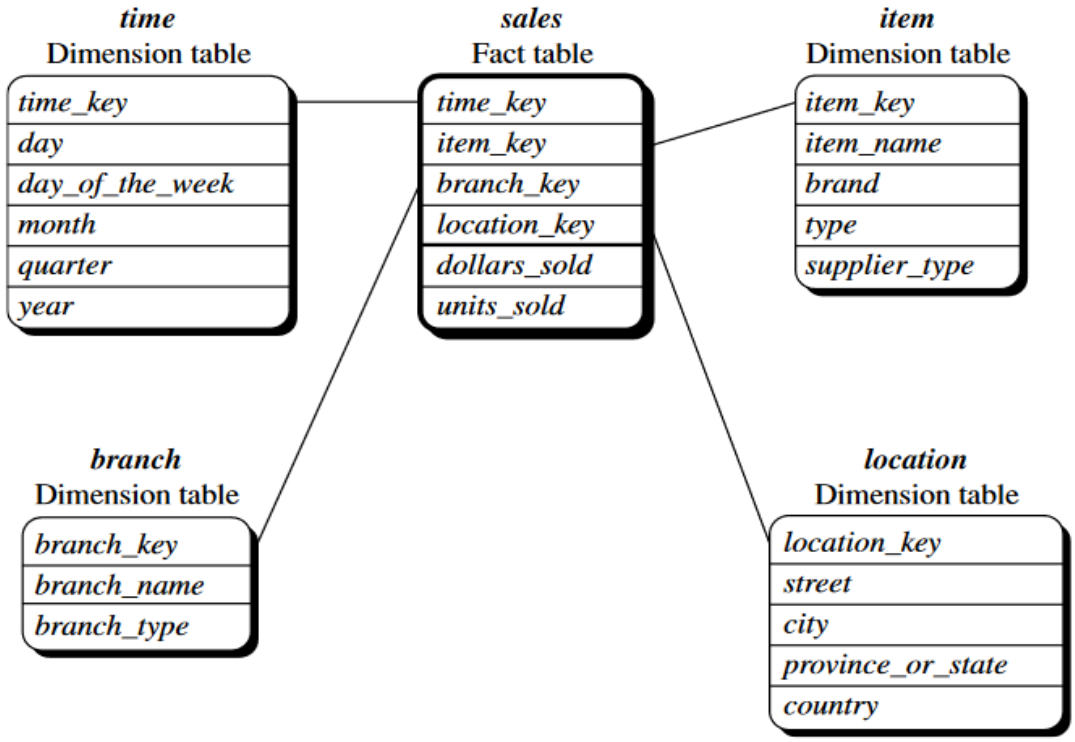


Fig.2.1 Example of the Dimensional Model (Star schema)
(Reprinted from "Data Mining Concepts and Techniques Third Edition" p.140,
by Jiawei Han et al. 2011[35])

2.2.2 Application of Data Warehouse techniques in the medical field

Approaches based on DW are also being adapted in medical field[36] [37]. In the United States, for example, the Chronic Condition Data Warehouse (CCW) [38] has been built for the purpose of supporting studies related to patients with chronic diseases by organizations that support research using health insurance claims databases, allowing for greater efficiency in specific types of epidemiological study. Another example is a decision-making tool developed for use in health insurance claims databases that are operated in Taiwan[39]. This tool allows for analyses to be performed more easily when conducting typical epidemiological studies.

Despite these advances, because the attributes required in each specific epidemiological study are different, there are many cases in which attributes that have been previously prepared in DW insufficiently cover the respective epidemiological study [40]. This is because technologies related to

DW, especially such as the dimensional model, are based on the premise that the attributes to be used are predetermined. In epidemiological studies such as public health, the required attributes are often different for each study and complex[41].

As a result, there is a country as United States where specialists provide support to those who are performing database analyses, which indicates that there is a need for the type of support that can be provided by database and computer science specialists[13]. There is a report that an epidemiological study conducted using the Japanese national claims database could not be conducted without the support of IT engineers [42]. In a country where such support is not in place, researchers want to use data that are more research friendly and processed with immediate applicability to epidemiological research[24].

2.2.3 Data structure suitable for epidemiological analysis

Generally, a structure summarized in one tuple for each patient is considered better for epidemiological analysis [43,44], because in the field of epidemiological study complicated statistical analysis is often performed for each patient. However, if one patient's data is distributed in multiple relations, experienced data handling skills are required to reconstruct these data. Thus, researchers must extract and integrate data for each patient from multiple relations [45]. In Japan, there are several studies in which researchers attempt to reconstruct the Japanese health insurance claims database into a structure summarized in one tuple for each patient, which is more suitable for statistical analysis[46 47].

Characteristics of research questions in epidemiological studies

Generally, there are some research questions which have a high level of abstraction in epidemiological studies. For example, one of the attributes with a high level of abstraction is "the monotherapy of a drug" commonly used in epidemiological study[48]. However, a normalized database often stores only "the date of prescription" and "prescription days of a drug" in one tuple. In this case, it is necessary to compute the continuous prescription period by comparing "the date of expire of a drug" with "the next date of prescription of a drug". Thus, many processes are required to compute attributes with a high degree of abstraction such as monotherapy of a drug.

In addition, it is often necessary to consider the time-series relevance of each event, since patients are often tracked based on the timing of the event that occurred for each patient.

Thus, it has been thought that it is difficult to cover in advance with software or data warehouse because the variation of attributes required for epidemiological research is unknown. Therefore, so far, there has been little research on easy creation of a dataset summarized in one tuple for each patient from already-operated healthcare databases. No software has yet been developed to easily

create a dataset for various epidemiological studies. For this reason, researchers are resigned to create a dataset for their respective epidemiological study. Thus, it is still difficult for researchers who have little skills in data handling to conduct their epidemiological study by themselves.

2.2.4 Contribution

The contribution of this dissertation is to develop a methodology that can allow researchers who have little skills in data handling to create a dataset which is summarized in one tuple for each patient from health care databases by themselves. The relational model theory is well known and based on set theory and logic, which is a field of math and the author believes it is the best tool in data handling. This research focused on the fact that the relational model theory can be used to create the dataset by treating a research question which researchers have as a proposition.

In this dissertation, the author proposed an improved relational calculus for epidemiological studies and a building method of data warehouse optimized for the relational calculus. The next chapter describes the health insurance claims databases used in this dissertation and proposed methods.

Chapter 3 Materials and Methods

In this chapter, after the materials used in the dissertation, the proposed method for a dataset generation is described. This method applies the relational model theory introduced in Chapter 2.1 and patterns and simplifies the dataset generation summarized in one tuple for each patient. First, the Japanese health insurance claims data used in this research will be explained. The material used in this research is health insurance claims data held in the National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB) managed by the Ministry of Health, Labour and Welfare (MHLW) and JMDC claims database managed by the JMDC, a private company in Japan. In this section, the NDB are explained as an example of the Japanese health insurance claims system.

3.1 National Database of Health Insurance Claims and Specific Health Checkups of Japan（NDB）

The NDB [49] was created in FY2009 and is operated by the MHLW. To date, it has accumulated over 15 billion entries of electronic data concerning health insurance claims and specific health checkups. Examples of insurance claims data used in Japan is shown in Fig.3.1. When stored in the NDB, they are grouped into similar items as shown in the lower part of Fig.3.1 and stored in tables such as the RE and IR (Appendix A1) etc.

3.1.1 Health insurance claims data collection pathway

The procedure for collecting health insurance claims in the NDB is as follows. Fig.3.2 shows the pathway for health insurance claims data collection. In Japan, medical institutions issue invoices for patient treatment in the form of health insurance claims, and these are passed to a review and payment agency before the insurers are billed. There are several different types of Japanese health insurance claims, comprising medical health insurance claims issued by hospitals and clinics, dental health insurance claims issued by dental hospitals and clinics, “dispensing health insurance claims” issued by dispensing pharmacies, and DPC (Diagnosis Procedure Combination) health insurance claims composed of partial DPC data. Health insurance claims are issued once a month for each patient by each medical institution that has treated that patient. Once they have been returned to the review and payment agency by the authority, they are anonymized and stored in the NDB. Although only electronic health insurance claims are collected in the NDB, recently, almost all health insurance claims have been made electronically [50].

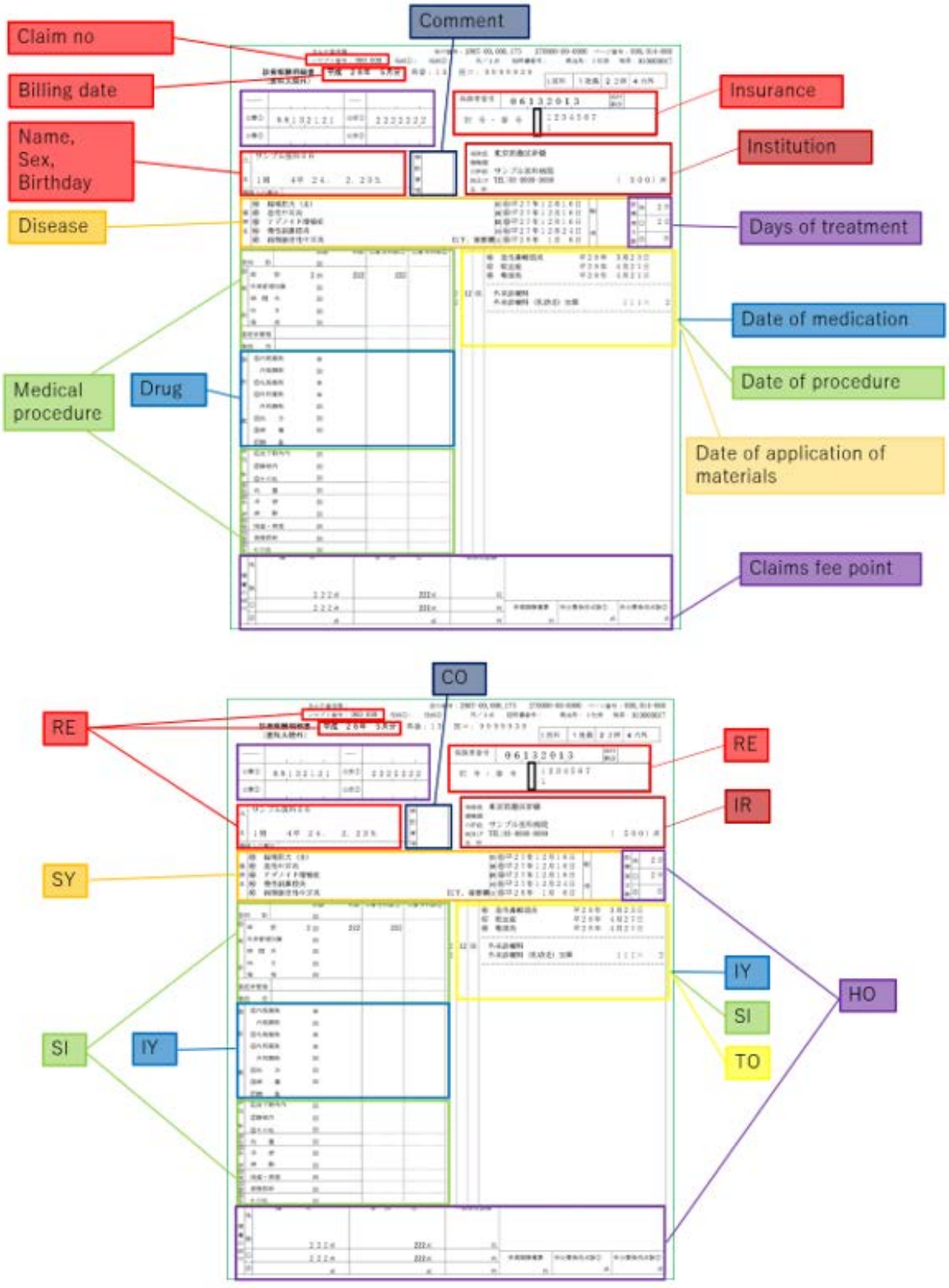


Fig.3.1 Example of health insurance claims in Japan

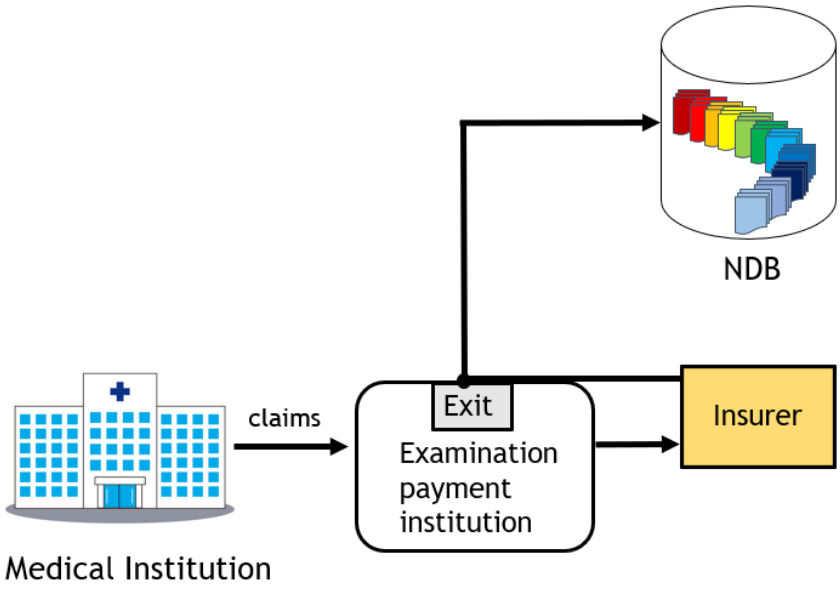


Fig.3.2 Collection pathway of health insurance claims in Japan

3.1.2 Secondary use of the NDB

As for secondary use of the NDB, if a person who wants to use the data, they can be provided with the data after undergoing screening according to certain criteria[51]. Because it contains data on almost all residents of Japan, it provides a high degree of completeness, making it a potentially effective resource for epidemiological studies. The MHLW allows secondary use of NDB mainly in the following provision formats.

(1) NDB Onsite research center

(2) NDB Special Extraction

(3) NDB Sampling Dataset

The on-site research center in (1) allows users to directly operate the NDB [52]. The on-site research center in (1) allows users to directly operate NDB. There have been some studies on the investigation and improvement of the analysis environment [43],but now they are open to the public. The provision formats in (2) and (3) provide users with data extracted from the NDB. In the NDB special extraction in (2), health insurance claims data and health checkup data specified by the user is provided, and NDB sampling dataset in (3), one month of health insurance claims data randomly sampled from the NDB is provided. So far, epidemiological studies using the NDB have been reported to some extent[53]. However, the NDB has been accumulated for medical insurance claims and is not intended for secondary use for research purposes[54,55], so it cannot be said that it is actively being used. Although there are documents that explain epidemiological analysis methods using the

NDB, high-spec machines are required [56]. Moreover, it is said that data analysis using the NDB requires the support of IT specialists [42].

3.1.3 NDB Sampling Dataset

This research uses the NDB Sampling Dataset (NDB-SD), which contains a random sample of the NDB data over a one-month period. We obtained the NDB-SD, which is comprised by the items “medical outpatient claims,” “medical inpatient claims,” “DPC claims,” and “pharmacy claims”. Fig.3.3, Fig.3.4, and Fig.3.5 show the NDB-SD relations used in this research. In the Japanese system, such claims are issued at a frequency of one per month. In Fig.3.3, the IR stores records on medical institutions while the HO relation stores records associated with health insurance information. The RE relations include patient details such as age and sex, labelling them with a unique Patient ID. Records of the, SY, SI and IY relations, which concerned drugs, medical procedures, and diseases, respectively, are correspond to the records in the RE relation with {claim no} and {absolute_no}. And, the csv files of NDB-SD were converted into the database format by using Database Management System. Similarly, Fig. 3.4 is data on DPC (Diagnosis Procedure Combination) claims. The DPC is Japanese cost calculation method that classifies diagnostic groups (1,572 classification) classified according to the patient's disease name and medical procedure and defines the hospitalization cost per day for each classification. DPC is stored in BU_DPC, disease information within 3 months after admission is stored in SB_DPC, and disease information after that is stored in SY_DPC. In addition, the CD_DPC stores the medical procedures carried out within 3 months after admission and information on medication etc. And, Fig. 3.5 is data on Pharmacy claims. Information on dispensing is stored separately in SH_PHA and CZ_PHA.

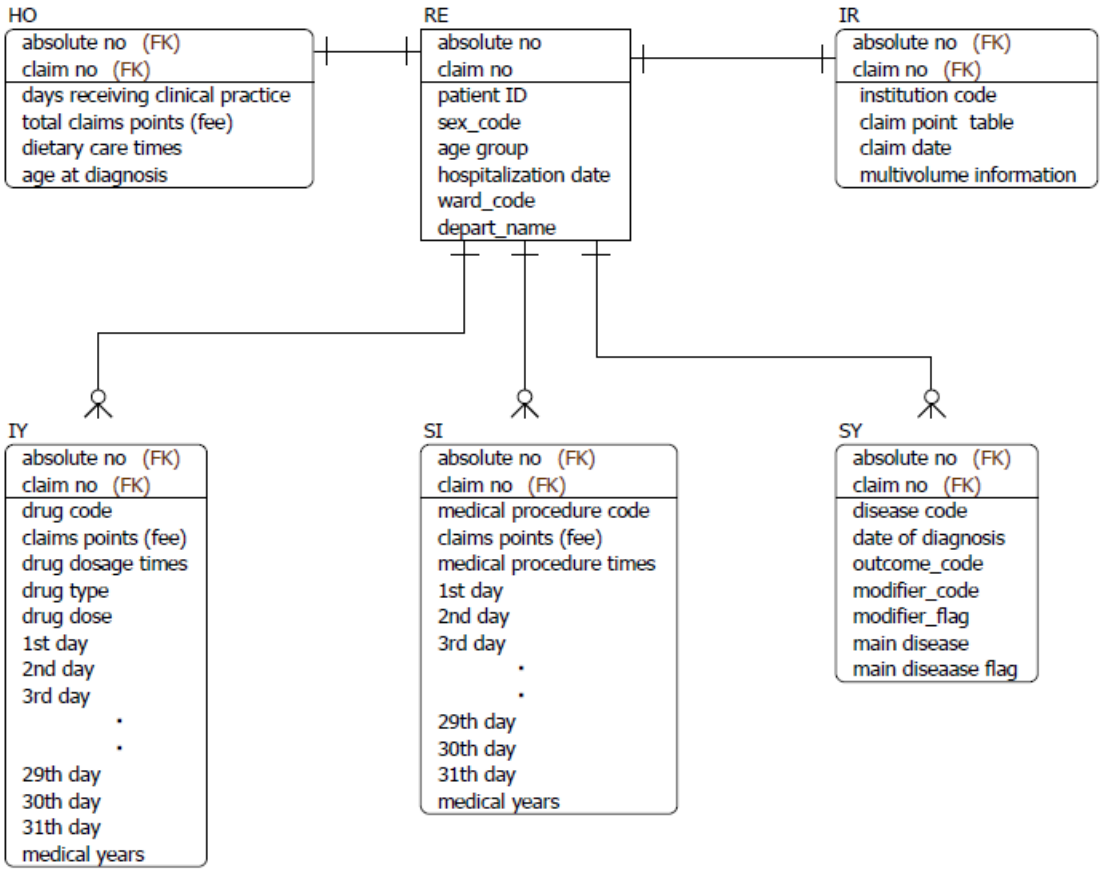


Fig. 3.3. Example of the typical NDB-SD relations and attributes (Medical claims)

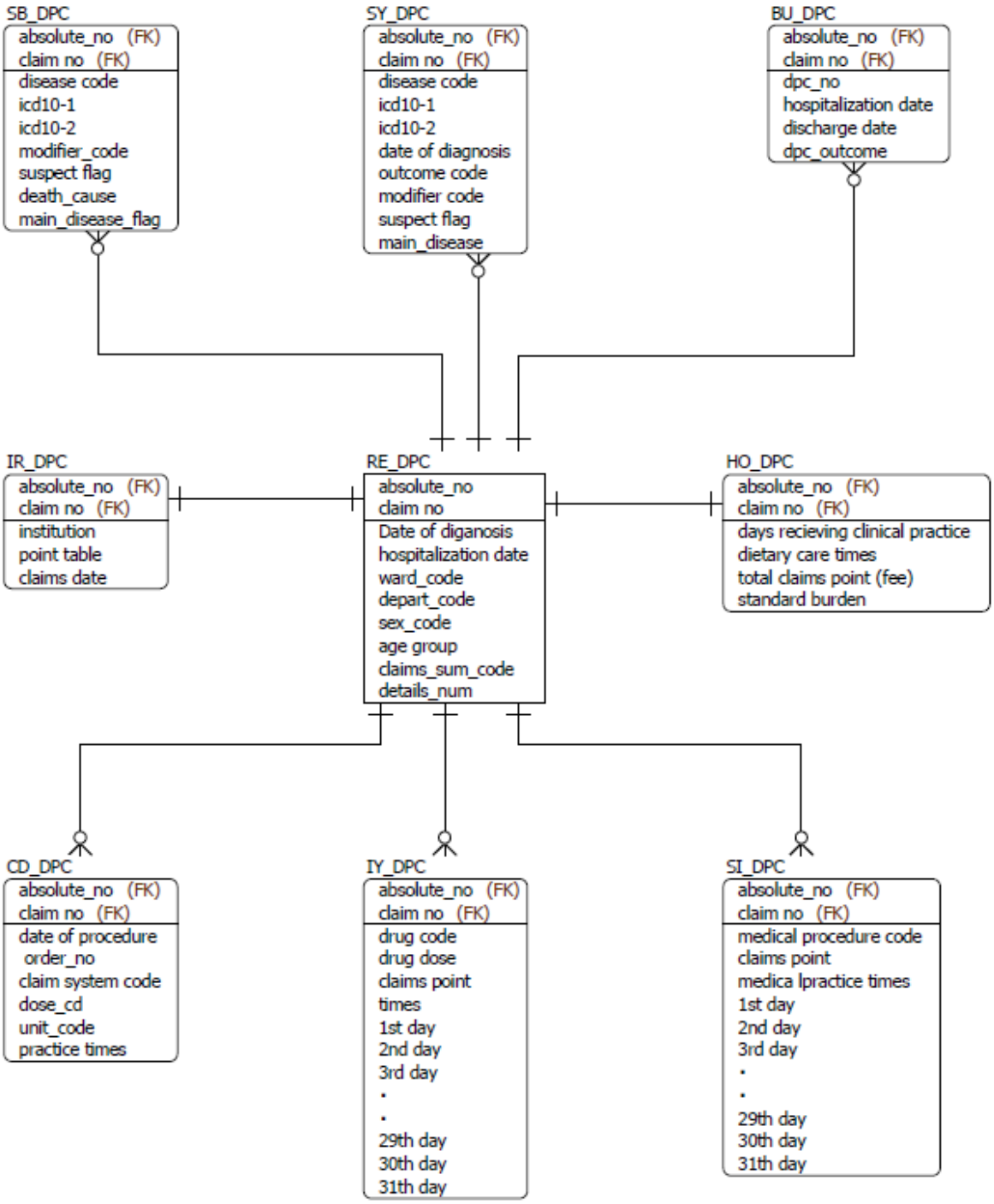


Fig. 3.4. Example of the typical NDB-SD relations and attributes (DPC claims)

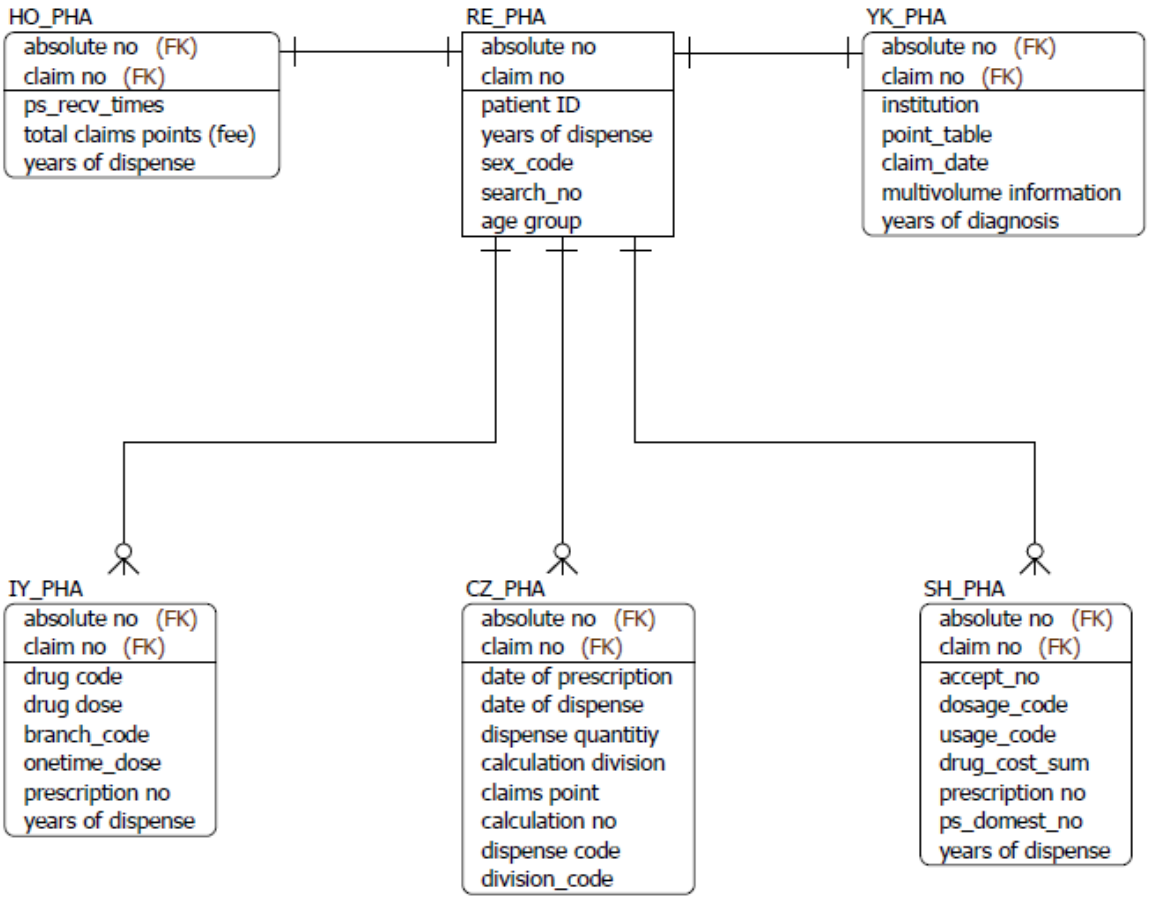


Fig. 3.5 Example of the typical NDB-SD relations and attributes (Pharmacy claims)

3.2 JMDC claims database

In addition to the NDB-SD, this research used insurance claims data from the JMDC. The JMDC compiles information from multiple insurance groups which consist of the insured belonging to each company. By June 2018, the JMDC had completely anonymized the insurance claims data of 5.4 million insured patients aged between 0 and 74 years.

The data in comma-separated values (CSV) format on patients' background were obtained, with data on patient characteristics, such as age, and sex (PATIENT Table), disease information (DISEASE Table), and drug information such as drug name, date of prescription, and drug volume etc. (DRUG Table) for 5,478,478 individuals for the period from January 2005 to December 2017.

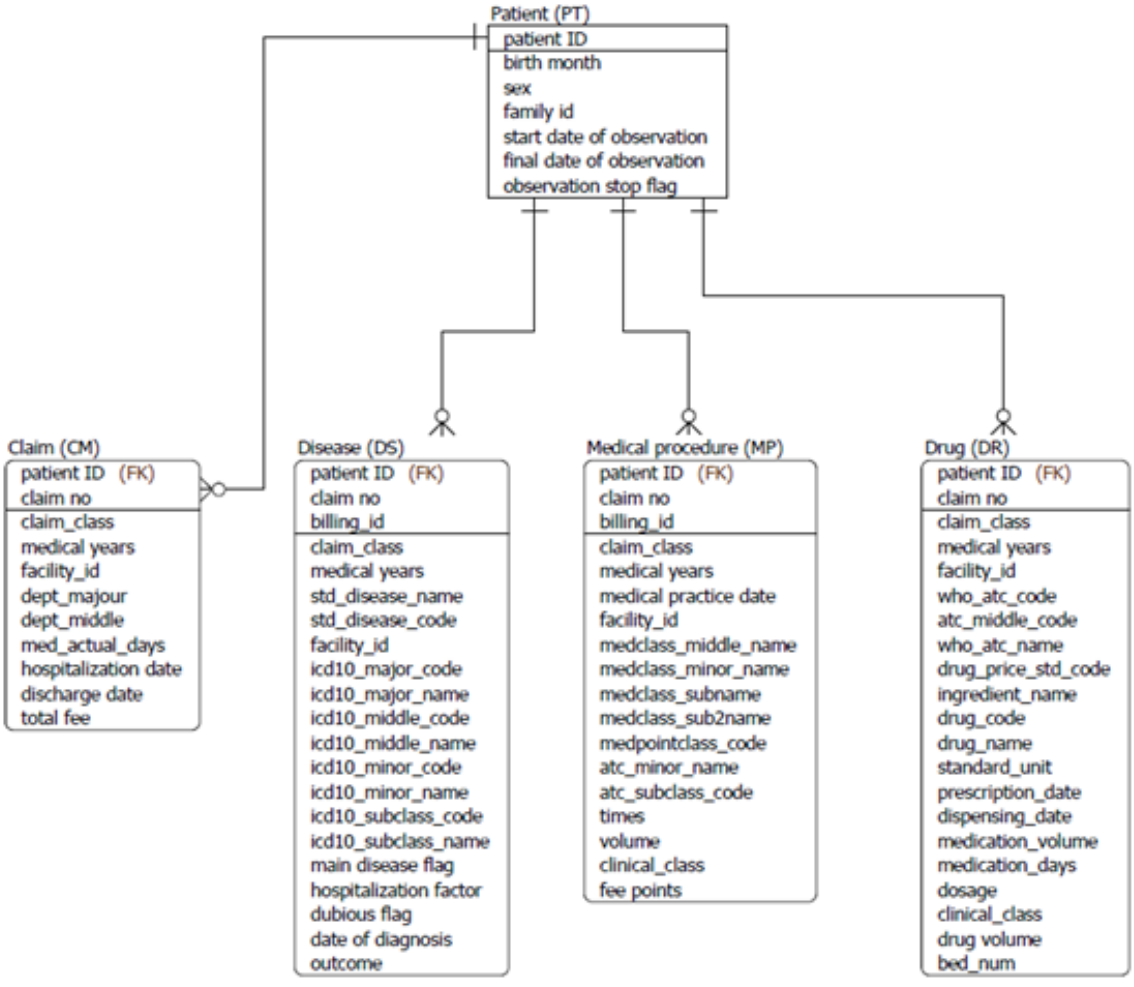
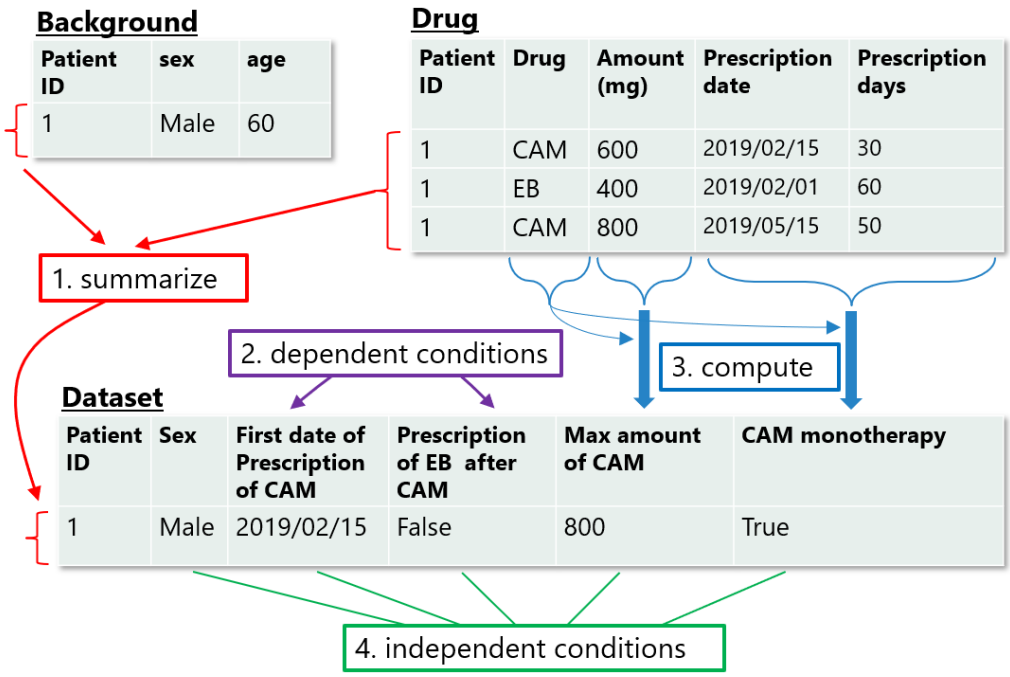


Fig. 3.6 Example of the typical JMDC claims relations and attributes

3.3 Pattern tuple relational calculus

For example, taking a retail store database as an example, there are multiple analysis units [57] such as employee, stores and sales etc. in the DW. These analysis units are all important for decision making and exploratory analysis. On the other hand, when creating datasets in epidemiological studies, the unit of analysis is almost “human” in many cases. Therefore, in this research, the analysis unit is limited to human.



CAM: Clarithromycin, EB: Ethambutol

Fig. 3.7. Challenges in generating a dataset

3.3.1 Methodology for generating a dataset summarized in one tuple for each patient

In order to create a dataset for each patient from multiple relations, it is necessary to solve some problems. An example of creating a dataset is shown in Fig. 3.7. To create the Dataset from the relation Background and Drug in the figure, it is necessary to tackle for challenges simultaneously. The first problem is “1. Summarize” in Fig.3.7. As shown in the figure, there are multiple relations with the same patient ID in the relation Drug, so researchers must choose one. The second problem is “2. dependent conditions” in Fig.3.7. In this example, to compute the attribute {Prescription of EB after CAM}, researchers need to refer to the attribute {First date of Prescription of CAM}, and these two attributes have dependent conditions. Next, the third problem is “3. compute” in Fig.3.7. In this example, to compute the attributes "Max amount of CAM" and "CAM monotherapy", researchers must refer to all the attributes of relation Drug. In the actual health care databases, there are other

relations than Drugs, and the required attributes are different for each epidemiological study, so this problem is most troublesome. The last problem is “4. independent conditions” in Fig.3.7. Generally, in research questions, for each attribute, it is necessary to distinguish between those who have that attribute and those who do not. Thus, it is the work to narrow down the patients along the research question.

Some people may think that the four tasks in Fig.3.7 can be easily achieved by computer programming. However, even those who are skilled in programming skills need to implement many functions. This is because there are many attribute variations required by each epidemiological study. Therefore, it is difficult for researchers who have research questions and are not familiar with computer programming to create a dataset summarized in one tuple for each patient.

Methodology for generating a dataset

In this research, the author focuses on the fact that the relational model theory can be applied to dataset generation by treating a research question as a proposition. As a result, powerful relational calculus and operators defined in the relational model theory can be used to generate a dataset. The procedure of this method is as follows.

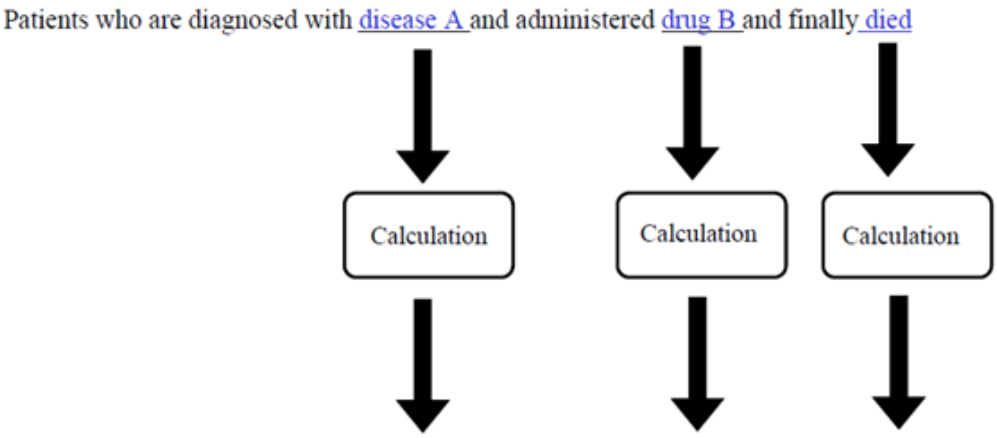
Step 1. Research question splicing. Step 2. Pattern relational calculus. Step 3. Truth function, Extend.

First, Step 1 is the process of dividing a research question for the purpose of simplifying each calculation process called research question splicing. Next, the pattern relational calculus is the process of calculating each attribute. Step 3 is for narrowing down patients based on the conditions specified in a research question. In this step, the truth function and Extend operator are used. The process of Step 2 and Step 3 can be realized by using the SQL. The Step 3 can be easily performed by users who understand the basic grammar of SQL such as “Select”, “From”, “Where”. On the other hand, Step 2 requires complex SQL when implemented straightforwardly.

3.3.2 Research question splicing

In general, a research question consists of multiple propositions, so it is complicated to handle as it is. René Descartes said that divide each “difficulty” into as many parts as is feasible and necessary to resolve it[58]. In epidemiological studies, the phrase “difficulty” could be replaced with “a research question”. Therefore, we considered simplifying the problem to be solved by splicing the research question. In this research, we named this idea “research question splicing” and the concept is shown in the Fig.3.8. For example, let us assume that a researcher wants to analyze data for patients with disease A who were administered drug B and finally died. As shown in Fig.3.8, in this case, the three conditions corresponding to attribute contained in the above sentence (patients with disease A who were administered drug B and finally died) are divided into units corresponding to

attributes. If each "calculation" process in Fig.3.8 can be made simple and patterned, anyone can create a dataset for each analysis unit. The following “the pattern tuple relational calculus” is a method for making the process of "calculation" simple and patterned. Moreover, as shown in Fig. 3.8, once each attribute is tabulated, the truth function and Extend operator in the relational model theory can be used, so that patients can be narrowed down as necessary.



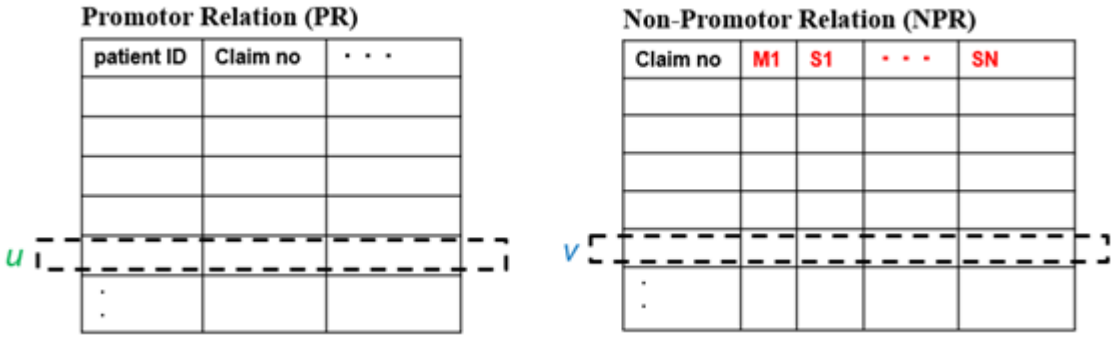
Patient ID	Disease A	Drug B	Died
1	True	True	False
2	True	False	
. . .			
N	False	False	False

Fig. 3.8. Research question splicing

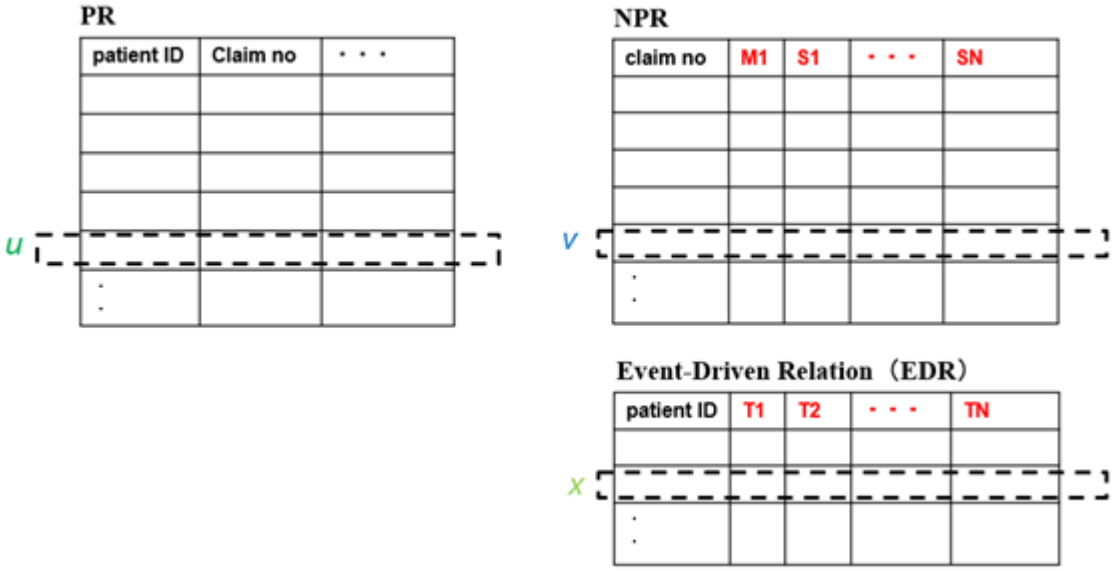
3.3.3 Pattern tuple relational calculus for cross sectional study

In the “calculation” process described above, this research uses the relational model theory. This is for the following two reasons. First, this is because the relational model theory has a manipulative feature of the *closure property*. The closure property is the specification of the relational model theory and also is what lets us write nested relational expressions[20]. For this reason, it is possible to express complex operations with simple relational calculus.

Second, database technology is most suitable for big data operations. Since the SQL is relational complete, it can express all relational calculus without resorting programming loop or any other form of branched execution[59] , and is supported by many DBMS products such as RDB, and other NoSQL database software. Neither general programming languages nor SQL-like techniques devised for manipulating tables [60] meet the above criteria and are not suitable for this task.



(a) Promotor Relation (PR) and Non-Promotor Relation (NPR)



(b) Combination of PR, NPR and Event-Driven Relation (EDR)

Fig. 3.9. The Relations for the pattern tuple relational calculus

Fig.3.9 shows the principle of the pattern tuple relational calculus (TRC) we proposed[61]. This method takes measures to store attributes expected to be necessary for only two relations according to research questions. One is a promotor relation (PR) we called in this dissertation. A PR stores the analysis unit and is a mandatory relation when computing an attribute. In the case of NDB-SD, the RE with the patient ID is the promoter relation. Usually, in the case of a health care database built based on the relational model, PR exists in most cases. The other is non promotor relation (NPR). In the case of NDB-SD, all relations other than PR (e.g. SY, SI, IY, etc.).

As a result of applying equation (3.1), several tuples are obtained for the patient ID. Therefore, researchers need to select only one tuple for each patient ID. Let *R* be the relation obtained from

equation (3.1). Next, by applying equation (3.2), it is aggregated for each patient ID. Functions for aggregation are implementation-dependent, and functions such as rank(), row number(), count(), max(), min() are generally implemented in the SQL.

If a PR and NPR have attributes required for respective study, each attribute can be obtained with a simple tuple relational calculus as shown in equation (3.1).

Since tuple relational calculus can be expressed in the SQL, they can be realized with simple and patterned code. For example, if M1# is an attribute to be computed, it can be computed using sub attributes from *S1* to *SN* as conditions. Note that there may be more than one sub attribute or none, contingent on research questions.

Pattern TRC

$$\{ t^{(1)} \mid \exists(u) \exists(v) (u \in PR \wedge v \in NPR \wedge u(\text{Claim no}) = v(\text{Claim no}) \wedge NPR(S1, S2, \ldots, SN] \ \theta \ c \wedge t(M1) = v(M1)) \} \quad \text{---(3.1)}$$

Summarize *R* PER (*R* {patient ID}) ADD (*summary* as *M1#*) ---(3.2)

Where $t^{(1)}$ is tuple with one attribute; θ is operator (e.g. <, =, >, $\leqq$, $\neq$, $\geqq$); *c* is constant; *M1* is a main attribute; *S1~SN* are sub attributes; *N* is natural number greater than or equal to 1;
ADD is the Extend operator in the relational model theory;
Summarize [62] is the summarize operator in the relational model theory;
Summary is the aggregate function (e.g. count, sum, min, max, etc.);

3.3.4 Pattern tuple relational calculus for cohort studies

In cohort studies, the date on which an event occurred is often important. For example, let us consider patients who have been administered pyrazinamide within 30 days from the date of diagnosis for tuberculosis. Since the date of diagnosis of tuberculosis is different for each patient, it is difficult to prepare an attribute called “the date of diagnosis of tuberculosis” in advance in the NPR.

Therefore, as shown in Fig.4.9, we introduce the Event-Driven Relation (EDR). The EDR is created by the pattern TRC and consists of only event-driven attributes such as date-related attribute. Using attributes stored in the EDR and applying the following the pattern TRC, a dataset that considers the context of events can be created.

Pattern TRC for an event-driven attribute

$$\{ t^{(1)} \mid \exists(u)\, \exists(v)\, (u \in PR \wedge v \in NPR \wedge u(\text{claim no}) = v(\text{claim no}) \wedge NPR(S1,S2,\ldots,SN)\, \theta$$
$$\exists(u)\, \exists(x)\, (u \in PR \wedge v \in EDR \wedge u(\text{patient ID}) = x(\text{patient ID})) \wedge t(M1) = v(M1)) \} \quad \text{--(3.3)}$$

Summarize *R* PER (*R* {patient ID}) ADD (*summary* as *M1#*) --(3.4)

Where $t^{(1)}$ is tuple with one attribute; θ is operator (e.g. <, =, >, ≦, ≠, ≧); *M1* is a main attribute; *S1~SN* are sub attributes; *N* is natural number greater than or equal to 1;
ADD is the Extend operator in the relational model theory;
Summarize [62] is the summarize operator in the relational model theory;
Summary is the aggregate function (e.g. count, sum, min, max, etc.);

3.3.5 Example of SQL based on the pattern TRC

(1) Relational complete

In the relational model theory, the SQL is guaranteed to be *relational complete*[59]. For that reason, a TRC in equation (3.1) and (3.3) theoretically can be implemented in the SQL. In this dissertation, we explain examples of the SQL based on the pattern TRC. Moreover, "*Summary*" in equation (3.3) is supported by SQL's window functions[63].

(2) Example of SQL for computing an attribute

Here we assume that there is a schema consisting only of patient ID. In this dissertation, it is called the "analysis unit schema".

Fig.3.10 shows the example of SQL to compute an attribute. In our method, attributes other than the attributes in the analysis unit schema, such as A1 and A2 in Fig.3.10 (e.g., disease name, drug name), are computed and the analysis unit schema is extended with it. Fig.3.10 shows example of SQL for evaluating the attributes related to a disease name.

For patients who have the disease name, one representative disease name is stored in the schema (shown as "True" in Fig.3.10), and in the case of patients who do not have the disease name, nothing is stored (shown as "False" in Fig.3.10). Basically, researchers can compute new attributes only by editing the underlined section from ① to ⑤ in Fig.3.10.

Underlined section ① and ② in Fig.3.10 showed an attribute name to be computed. Here, the SQL function underlined ③ ("row_number ()" function in Fig.3.10) is used to consolidate multiple records of a single patient into one tuple. This function has three patterns according to the type of attribute to be computed. The "row_number ()" function is mainly used in computing a disease name (code), drug name (code), medical procedure (code), date-related attribute (e.g., date of diagnosis,

date of prescription) etc. and is most frequently used in our research. Similarly, the “sum ()” function is mainly used for “volume of drug”. These functions are responsible for selecting just one tuple from the many tuples that result from the join of two relations (RE and SY).

There are many SQL functions in SQL’s specification [63], but only above-mentioned two SQL functions were able to cover all the attributes obtained from the NDB-literatures. In the underlined section ⑤, extraction conditions such as disease code, disease name, drug code, medical procedure code, date of diagnosis, etc. are listed. In the underlined section ④, researchers specify the name of two relations.

Basically, two relations consist of a relation (e.g., RE relation in the NDB-SD) that contains analysis unit (e.g., patient ID, claim no) and a relation that contains an attribute researcher would like to compute (e.g., SY, SI, IY, etc.).

As an example, if researchers want to compute drug related attributes, by specifying relations such as combination of "RE (Medical claims) and IY," "RE (Pharmacy claims) and CZ," "RE (Pharmacy claims) and IY," and "RE (DPC claims) and CD," researchers can compute attributes such as drug name or date of prescription for each drug per analysis unit. In summary, the skill sets required by users are as follows.

(1) Basic grammar of SQL (e.g. SELECT, FROM, WHERE)
(2) Basic grammar of SQL window functions (e.g. ROW_NUMBER (), SUM ())
(3) Basic knowledge of the target database

As mentioned in (1) and (2), in my proposed method, basic knowledge about SQL is required. However, relatively complicated syntax of SQL such as "JOIN" or "subquery" is not required. In addition, knowledge about the database structure such as what attributes are stored in which relations is necessary. In the NDB-SD example, the relation IY contains information about drugs, and the relation SY contains name of diseases.

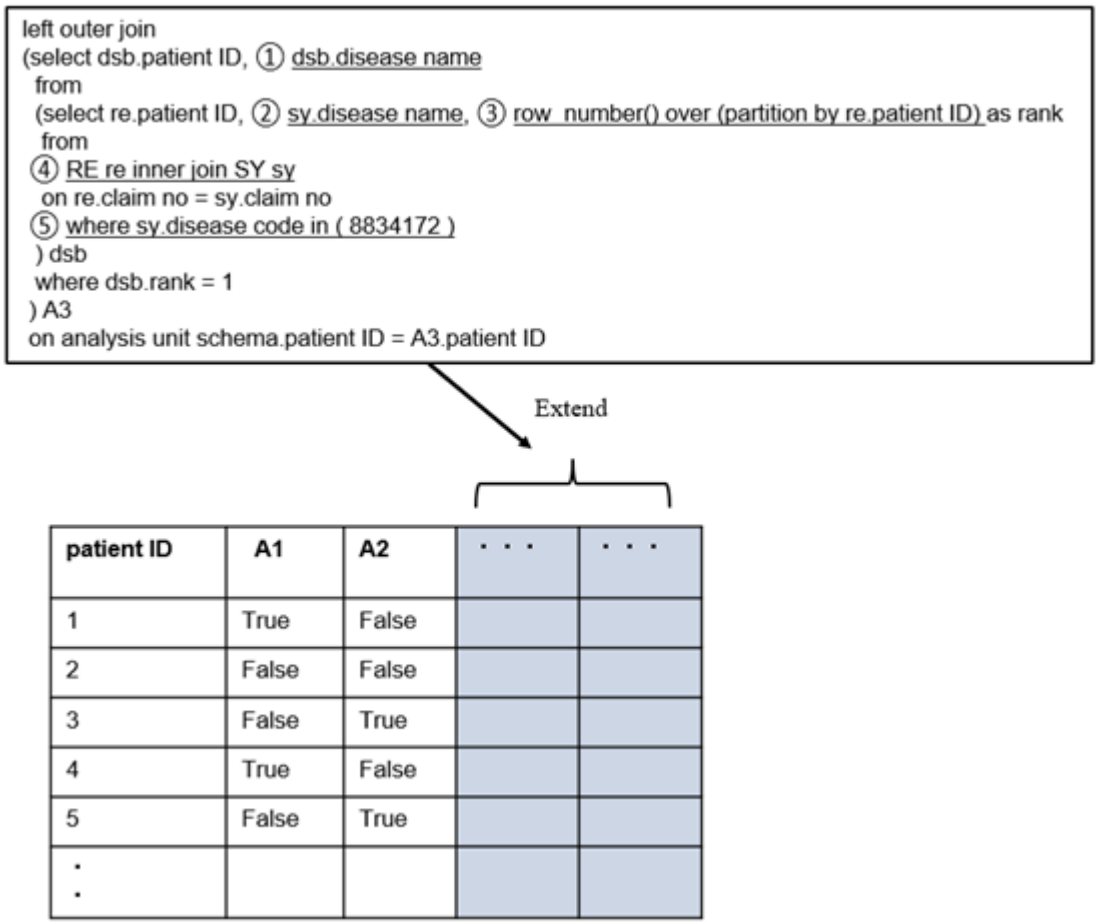


Fig. 3.10. An example of SQL based on the pattern tuple relational calculus

3.4 Research Question oriented Data Warehouse

Here, we introduce a DW which we developed to make the pattern TRC applicable, as well as to compute essential attributes for epidemiological studies. The procedure of developing the Data Warehouse is shown in Fig.3.11. This DW is developed from the NDB-SD. The building procedure of the DW can be divided into two Phases. During Phase 1, we constructed a Research Question oriented Data Warehouse (RQDW) optimized for inputting aggregate functions used in the pattern TRC to compute attributes that are commonly used for epidemiological studies. In Phase 2, by applying the pattern TRC, we created the versatile analysis unit schema which is the structure summarized to one record for each analysis unit from the RQDW. Attributes implemented in the versatile analysis unit schema depend on the health care database used.

Phase 1: Building of the Research Question oriented DW

The purpose of the RQDW is to add attributes or relations to an original database so that the pattern TRC can be applied. There are following two ways to build the RQDW by adding the original database.

Method 1 for adding attributes: Addition of new attribute to existing relations,

Method 2 for adding attributes: Addition of new relations

In above-mentioned Method 1, essential attributes are often added by external master data or simple data shaping of such as numerical-type or date-type attributes. Regarding a master data, in Japan, medicinal master data managed by the MEDIS is famous [64]. As for above-mentioned Method 2, it may be necessary for epidemiological studies with relatively complex research designs. In this case, it is often required to analyze attributes existing in multiple relations and perform data shaping. In addition, it seems that it is better to build the RQDW in database operation languages such as SQL, but general programming languages may be used.

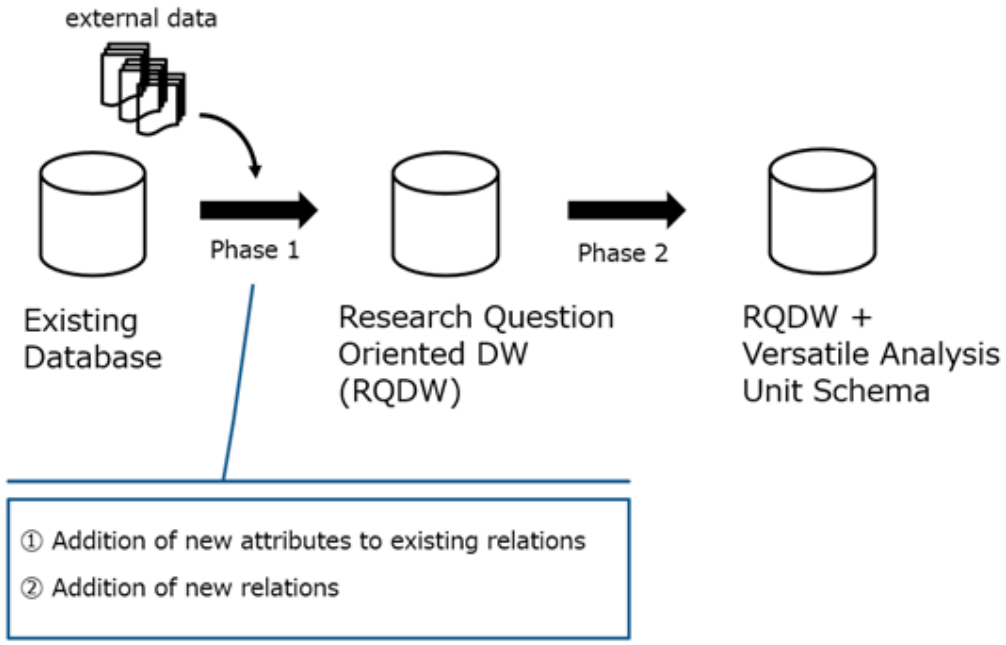


Fig. 3.11. Building procedure of the Research Question oriented Data Warehouse

Phase 2: Implementation of the versatile analysis unit schema

The purpose of creating the versatile analysis unit schema is to prepare the attributes that are often used in epidemiological studies in the schema that is summarized in one tuple for each analysis unit. In the relational model design, unique attribute in each relation is called “primary key.” In many epidemiological studies researchers focus on human populations, so in that case they place the analysis unit as each “human.”

Taking the NDB as an example, the closest attributes to “human” in the Research Question Oriented DW are {patient ID} and {claim no} respectively; The reason why {claim no} is close to “human” is that the NDB-SD contains only one month sampled data, so with some exceptions in such cases that one patient visits multiple medical institutions and accidentally extracted simultaneously, one claim data corresponds to one patient. Thus, in our research, these attributes were treated as the primary key respectively.

Next, by applying the pattern TEC we created the versatile analysis unit schema from the RQDW, the primary keys of which are {patient ID} and {claim no}. To ensure the versatility of the versatile analysis unit schema, we selected attributes that can be summarized in one tuple for each analysis unit such as “Death,” “Age group,” etc. in the NDB-Literatures and implemented them to the schema. To build the analysis unit schema, we used SQL [27], a database language and MS-DOS commands. And, we developed a software that make the versatile analysis unit schema from NDB-SD’s CSV files provided by the NDB-SD automatically with 20,000 steps source codes. We used a

laptop computer to build the DW in this research; furthermore, we used PostgreSQL (ver. 9.5) , which is well equipped with online analytical processing functions[28], as a database management system.

In this dissertation, five epidemiological studies were conducted using the proposed method and RQDW. The next chapter describes epidemiological studies using the NDB-SD and JMDC claims database.

Chapter 4 Results

In this chapter, first, it is described how much of the attributes required for respective study were covered by the proposed method. Next, it is described the Research Question Oriented DW of the NDB-SD and JMDC claims DB implemented to compensate for missing attributes. Finally, we introduce epidemiological studies conducted by the author of this dissertation.

4.1 Results of epidemiological studies

Table 4.1 shows the epidemiological studies conducted using the proposed method. The coverage rate in Table 4.1 indicates the ratio that the pattern TRC was able to cover the attributes required for each study and practitioner indicates the person who performed data handling in the respective study. As a result, not less than 80% of the attributes were covered without building the RQDW in the five epidemiological studies.

Table 4.1 Epidemiological studies using the proposed method

Case	Study name	Practitioner	Study type	Database	Coverage (Original)	Coverage (RQDW)
1	PPH	Author	Cross sectional	NDB-SD	88% (22/25)	100% (25/25)
2	CPR	Physician-scientist A	Cross sectional	NDB-SD	100% (20/20)	100% (20/20)
3	NTM	Author	Cross sectional	JMDC	85% (11/13)	100% (13/13)
4	MAC Lung Disease	Author	Retrospective cohort	JMDC	80% (20/25)	92% (23/25)
5	Influenza	Physician-scientist B	Retrospective cohort	JMDC	97% (34/35)	100% (35/35)

Coverage rate: Pattern TRC's coverage rate of attributes using the original database
PPH: Postpartum Hemorrhage, NTM: Nontuberculous mycobacteriosis,
CPR: Cardiopulmonary Resuscitation, MAC: *Mycobacterium Avium-intracellulare* Complex

Case 1: Postpartum Hemorrhage (PPH)

The aim of the PPH study was to investigate epidemiological and clinical aspects of severe postpartum hemorrhage (PPH) in Japan. The research question advocate is an obstetrician at the Kyoto University Hospital. The results of this epidemiological study have already been published in the journal [65]. Table 4.2 shows the attributes used in the PPH study. Window function refers to the SQL functions that implement the *summary* function in equation (3.2). As shown in the table, the pattern TRC was able to cover 88% (22/25) of the used attributes in the original NDB-SD.

Case 2: Cardiopulmonary Resuscitation (CPR)

The aim of the CPR study was to investigate epidemiological and clinical aspects of Cardiopulmonary Resuscitation (CPR) in Japan. The research question advocate is an emergency physician (Physician-scientist A in Table 4.1) at the Kyoto University Hospital. The results of this epidemiological study have already been published at academic conference [66]. Table 4.3 shows the attributes used in the CPR study. Window function refers to the SQL functions that implement the *summary* function in equation (3.2). As shown in the table, the pattern TRC was able to cover 100% (20/20) of the used attributes in the original NDB-SD.

Case 3: Nontuberculous mycobacteriosis (NTM)

The aim of the NTM study was to investigate epidemiological and clinical aspects of Nontuberculous mycobacteriosis (NTM) in Japan. The research question advocate is first author of this manuscript. The results of this epidemiological study have already been published in the journal [67]. Table 4.4 shows the attributes used in the NTM study. Window function refers to the SQL functions that implement the *summary* function in equation (3.2). As shown in the table, the pattern TRC was able to cover 85% (11/13) of the used attributes in the original JMDC claims database.

Case 4: *Mycobacterium Avium-intracellulare* Complex Lung Disease (MAC Lung Disease)

The aim of the MAC Lung Disease study was to investigate prescription pattern for patients with MAC Lung Disease in Japan. The research question advocate is first author of this manuscript. The results of this epidemiological study have already been published in the journal [68]. Table 4.5 shows the attributes used in the MAC Lung Disease study. Window function refers to the SQL functions that implement the *summary* function in equation (3.2). As shown in the table, the pattern TRC was able to cover 80% (20/25) of the used attributes in the original JMDC claims database.

Case 5: Influenza

The aim of the Influenza study was to investigate epidemiological and clinical aspects of patients with Influenza in Japan. The research question advocate is an obstetrician (Physician-scientist B in Table 4.1) at the Kyoto University Hospital. Table 4.6 shows the attributes used in this study. Window function refers to the SQL functions that implement the *summary* function in equation (3.2). As shown in the table, the pattern TRC was able to cover 97% (34/35) of the used attributes in the original JMDC claims database.

Table 4.2. The attributes required for the epidemiological study of PPH

Attribute	**Covered**	**PR × NPR**	**Window function**
Sex	○	RE	Row_number ()
Age group	○	RE	Row_number ()
Death	○	RE × SY	Row_number ()
Uterine atony	○	RE × SY	Row_number ()
Placental abruption	○	RE × SY	Row_number ()
Placenta previa	○	RE × SY	Row_number ()
Retained placenta	○	RE × SY	Row_number ()
Placenta accrete	○	RE × SY	Row_number ()
Cervical laceration	○	RE × SY	Row_number ()
Vaginal hematoma	○	RE × SY	Row_number ()
Multiple pregnancy	○	RE × SY	Row_number ()
Low-lying placenta	○	RE × SY	Row_number ()
Uterine rupture	○	RE × SY	Row_number ()
Placenta previa accrete	○	RE × SY	Row_number ()
Uterine inversion	○	RE × SY	Row_number ()
Amniotic fluid embolism	○	RE × SY	Row_number ()
Volume of RCC (ml)	×	RE × IY	Sum ()
Volume of FFP (ml)	×	RE × IY	Sum ()
Volume of PC (ml)	×	RE × IY	Sum ()
Cesarean delivery	○	RE × SI	Row_number ()
Operative vaginal delivery	○	RE × SI	Row_number ()
Surgical intervention	○	RE × SI	Row_number ()
Intrauterine balloon tamponade	○	RE × SI	Row_number ()
Arterial embolization	○	RE × SI	Row_number ()
Hysterectomy	○	RE × SI	Row_number ()

Note: ○ = covered in the original NDB-SD; × = Not covered in the original NDB-SD,

PR: Promotor Relation, NPR: Non-Promotor Relation

RCC: Red Cell Concentrate, FFP: Fresh Frozen Plasma, PC: Platelet Concentrated

Table 4.3. The attributes required for the epidemiological study of CPR

Attribute	**Covered**	**PR × NPR**	**Window function**
Sex	○	RE	Row_number ()
Age group	○	RE	Row_number ()
Death	○	RE × SY	Row_number ()
Closed chest cardiac massage	○	RE × SY	Row_number ()
Adrenalin	○	RE × SI	Row_number ()
Tracheal intubation	○	RE × SI	Row_number ()
Post resuscitation encephalopathy	○	RE × SI	Row_number ()
Artificial respiration	○	RE × SI	Row_number ()
Defibrillation	○	RE × SI	Row_number ()
EEG	○	RE × SI	Row_number ()
CT	○	RE × SI	Row_number ()
Induced Hypothermia	○	RE × SI	Row_number ()
CAG	○	RE × SI	Row_number ()
PCI	○	RE × SI	Row_number ()
Hemodialysis	○	RE × SI	Row_number ()
PCPS	○	RE × SI	Row_number ()
Medical expenses for first aid	○	RE × SI	Row_number ()
Administering of blood products for transfusion	○	RE × SI	Row_number ()
Medical expenses for intensive care	○	RE × SI	Row_number ()
Post resuscitation encephalopathy	○	RE × SI	Row_number ()

Note: ○ = covered in the original NDB-SD; × = Not covered in the original NDB-SD,

CPR: CardioPulmonary Resuscitation,

PR: Promotor Relation, NPR: Non-Promotor Relation, EEG: Electroencephalogram,

CT: Computed Tomography, CAG: Coronary angiography, PCI: Percutaneous coronary intervention,

PCPS: Percutaneous cardiopulmonary support

Table 4.4. The attributes required for the epidemiological study of NTM

Attribute	**Covered**	**PR × NPR**	**Window function**
Sex	○	PT	Row_number ()
Age	×	PT	Row_number ()
Diagnosis of NTM	○	PT × DS	Row_number ()
Date of diagnosis of NTM	○	PT × DS	Row_number ()
Date of prescription of Pyrazinamide	○	PT × DS	Row_number ()
Number of antibacterial agents	○	PT × DS	Row_number ()
First date of prescription of CAM	○	PT × DS	Row_number ()
Final date of prescription of CAM	○	PT × DS	Row_number ()
First date of prescription of RFP	○	PT × DR	Row_number ()
Final date of prescription of RFP	○	PT × DR	Row_number ()
First date of prescription of EB	○	PT × DR	Row_number ()
Final date of prescription of EB	○	PT × DR	Row_number ()
Maximum amount of EB	×	PT × DS	Sum ()

Note: ○ = covered in the original JMDC database; × = Not covered in the original JMDC database,

PR: Promotor Relation, NPR: Non-Promotor Relation,

NTM: Nontuberculous mycobacteriosis, CAM: clarithromycin, EB: ethambutol,

RFP: rifampicin, SM: streptomycin, KM: kanamycin

Table 4.5. The attributes required for the epidemiological study of MAC Lung Disease

Attribute	**Covered**	**PR × NPR**	**PR × EDR**	**Window function**
Sex	○	PT	-	Row_number ()
Age	×	PT	-	Row_number ()
Diagnosis of NTM	○	PT × DS	-	Row_number ()
Date of diagnosis of NTM	○	PT × DS	-	Row_number ()
Date of diagnosis of chronic bronchitis	○	PT × DS	-	Row_number ()
Date of diagnosis of tuberculosis	○	PT × DS	-	Row_number ()
Date of diagnosis of bronchiectasis	○	PT × DS	-	Row_number ()
Date of diagnosis of diffuse panbronchiolitis	○	PT × DS	-	Row_number ()
Date of diagnosis of ethambutol optic neuropathy	○	PT × DS	-	Row_number ()
Date of prescription of antibacterial agent after NTM diagnosis	○	PT × DS	PT × EDR	Row_number ()
Number of prescriptions of CAM after NTM diagnosis	○	PT × DS	PT × EDR	Row_number ()
Number of prescriptions of azithromycin after NTM diagnosis	○	PT × DS	PT × EDR	Row_number ()
Date of prescription of Imipenem after NTM diagnosis	○	PT × DR	PT × EDR	Row_number ()
Date of prescription of Isoniazid after NTM diagnosis	○	PT × DR	PT × EDR	Row_number ()
Date of prescription of azithromycin after NTM diagnosis	○	PT × DR	PT × EDR	Row_number ()
Date of prescription of pyrazinamide after NTM diagnosis	○	PT × DR	PT × EDR	Row_number ()
Date of prescription of CAM after NTM diagnosis	○	PT × DR	PT × EDR	Row_number ()
Date of prescription of RFP after NTM diagnosis	○	PT × DR	PT × EDR	Row_number ()
Date of prescription of EB after NTM diagnosis	○	PT × DR	PT × EDR	Row_number ()
Date of prescription of KM after NTM diagnosis	○	PT × DR	PT × EDR	Row_number ()
Date of prescription of SM after NTM diagnosis	○	PT × DR	PT × EDR	Row_number ()
Maximum amount of EB	×	PT × DR	PT × EDR	Sum ()
Amount of CAM	×	PT × DR	PT × EDR	Sum ()
Period of EB prescription	×	PT × PP	PT × EDR	Row_number ()
Period of CAM monotherapy	×	PT × PP	PT × EDR	Row_number ()

Note: ○ = covered in the original JMDC database; × = Not covered in the original JMDC database,

PR: Promotor Relation, NPR: Non-Promotor Relation, EDR: Event-driven Relation,

NTM: Nontuberculous mycobacteriosis, CAM: clarithromycin, EB: ethambutol,

RFP: rifampicin, SM: streptomycin, KM: kanamycin, CR: clarithromycin + rifampicin

Table 4.6. The attributes required for the epidemiological study of Influenza

Attribute	**Covered**	**PR × NPR**	**PR × EDR**	**Window function**
Sex	○	PT	-	Row_number ()
Age	×	PT	-	Row_number ()
Diagnosis of Influenza	○	PT × DS	-	Row_number ()
Date of diagnosis of Influenza	○	PT × DS	-	Row_number ()
Date of prescription of anti-influenza drug after diagnosis of Influenza	○	PT × DR	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 1 day after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 2 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 3 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 4 days after diagnosis of influenza	○	PT × DR	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 5 days after diagnosis of influenza	○	PT × DR	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 6 days after diagnosis of influenza	○	PT × DR	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 7 days after diagnosis of influenza	○	PT × DR	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 8 days after diagnosis of influenza	○	PT × DR	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 9 days after diagnosis of influenza	○	PT × DR	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 10 days after diagnosis of influenza	○	PT × DR	PT × EDR	Row_number ()
continued the following page				

Attribute	**covered**	**PR × NPR**	**PR × EDR**	**Window function**
Date of diagnosis of pneumonia within 11 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 12 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 13 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 14 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 15 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 16 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 17 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 18 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 19 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Date of diagnosis of pneumonia within 20 days after diagnosis of influenza	○	PT × DS	PT × EDR	Row_number ()
Diagnosis of bronchial asthma 7 days before and after the day when influenza is diagnosed	○	PT × DS	PT × EDR	Row_number ()
Diagnosis of COPD 7 days before and after the day when influenza is diagnosed	○	PT × DS	PT × EDR	Row_number ()
Diagnosis of cryptogenic organizing 7 days before and after the day when influenza is diagnosed	○	PT × DS	PT × EDR	Row_number ()
Continued the following page				

Attribute	**covered**	**PR × NPR**	**PR × EDR**	**Window function**
Diagnosis of chronic heart failure 7 days before and after the day when influenza is diagnosed	○	PT × DS	PT × EDR	Row_number ()
Diagnosis of end-stage renal failure 7 days before and after the day when influenza is diagnosed	○	PT × DS	PT × EDR	Row_number ()
Diagnosis of metastatic lung tumor 7 days before and after the day when influenza is diagnosed	○	PT × DS	PT × EDR	Row_number ()
Diagnosis of type 2 diabetes mellitus 7 days before and after the day when influenza is diagnosed	○	PT × DS	PT × EDR	Row_number ()
Diagnosis of rheumatoid arthritis 7 days before and after the day when influenza is diagnosed	○	PT × DS	PT × EDR	Row_number ()
Date of prescription of steroid 180 days before the date of diagnosis of influenza	○	PT × DR	PT × EDR	Row_number ()
Date of prescription of immunosuppressive drug 180 days before the date of diagnosis of influenza	○	PT × DR	PT × EDR	Row_number ()

Note: ○ = covered in the original JMDC database; × = Not covered in the original JMDC database,

PR: Promotor Relation, NPR: Non-Promotor Relation, EDR: Event-driven Relation,

COPD: Chronic Obstructive Pulmonary Disease

4.2 Result of the Research Question oriented DW

In this section, the RQDW built in this research is explained. First, we describe the RQDW built for NDB-SD, and then describe the RQDW built for the JMDC claims database.

4.2.1 RQDW of the NDB-SD

The results of RQDW for the NDB-SD are described. The master data of drugs managed by the MEDIS-DC [64] compensated for missing drug doses (Volume of RCC, FFP, PC) in the PPH study in Table 4.2.

In this research, for further improvement of the RQDW of the NDB-SD, additional attributes frequently used in epidemiological studies were added to RQDW.

First, we investigated epidemiological studies that used the NDB[53] as source material. The MHLW introduced 105 published articles as NDB-related publications via its website. These publications include conference presentations, public surveys, academic articles and so on. However, the number of publications in which NDB data analysis was directly performed were only 20. We named these publications as “NDB-Literatures” (Appendix A) in this dissertation. We carefully reviewed tables, figures, and sentences in the NDB-literatures and picked up attributes which were used one or more times. The attributes obtained from the NDB-literatures is shown in Table 4.7. The “Frequency” refers to the times of appearance (maximum value: 20) in the NDB-literatures. In Table 4.7, the column titled “Research Question Oriented DW” shows attributes that were added to the original NDB-SD when building the RQDW.

Second, we added the attributes obtained from the NDB-literatures to the original NDB-SD, to build the RQDW. In Fig.4.1, Fig.4.2 and Fig.4.3, attributes marked “o” at the prefix in each table are newly added to the NDB-SD when building the Research Question Oriented DW. When constructing the DW, the most important thing is conversion of data-related attributes. This is because the attributes that can be covered by the pattern TRC are greatly increased by simply converting the date-related attributes into appropriate format. Some date-related attributes, such as dates of admission, diagnosis, etc. are stored in era name format in many relations in the NDB-SD. So, we converted these attributes into an A.D. format suitable for calculation by using functions of statistical analysis software R and SQL.

In order to compensate for attributes not stored in the NDB-SD, we made use of the master data officially published by the MHLW[69], for “disease name,” “medical procedure,” and “drug name,” and the one of MEDIS-DC[64] for the ICD10-1 and ICD10-2 code and the drug HOT code etc. Additionally, we referenced a study that is proposing new CCI scoring method based on the ICD10[70] for calculating the Charlson Comorbidity Index (CCI) score of each patient ID/claim no.

Table 4.7 The attributes obtained from the NDB-Literature

Attribute	Research Question Oriented Data Warehouse	Frequency
Age group	-	17
Sex	-	14
Discharge date (A.D. format)	-	6
Hospitalization date (A.D. format)	Hospitalization date (A.D. format)	5
Charlson comorbidity index	ICD10 code	4
Days receiving clinical procedure	-	4
Presence or absence of medical treatment	-	2
Administering of blood products for transfusion	Type of blood products used	2
Date of initial dispensing of drugs (A.D. format)	Date of dispensing of drugs (A.D. format)	2
Application of internal medicine	-	2
Death	-	1
Application of drug	-	1
Administering of injection	-	1
Application of prescription	-	1
Administering of inspection	-	1
Application of external medicine	-	1
Date of final dispensing of drugs (A.D. format)	Date of final dispensing of drugs (A.D. format)	1
Date of final application of medical procedure (A.D. format)	Date of final application of medical procedure (A.D format)	1
Drug name(s) specific to the research	Drug name	14
Name of medical procedure(s) specific to the research	Medical procedure name	12
ICD10 code(s) specific to the research	ICD10 code	10

Prescription date specific to the research (AD format)	Prescription date (A.D. format)	8
Disease name(s) specific to the research	Disease name	7
Date of administering medical procedure(s) specific to the research	Medical procedure date (A.D format)	6
Date of commencement of diagnosis of a disease specific to the research	Diagnosis commencement date (A.D. format)	4
Volume of drug specific to the research	Volume of drug (unit: mg. ml)	4
Classification by efficacy of drugs specific to the research	Classification drugs by efficacy	4
Number of administrations of medical procedure(s) specific to the research	Medical procedure name	2
Health insurance claims points	-	1
Drug’s YJ code (drug’s product code in Japan)	Drug’s YJ code	1

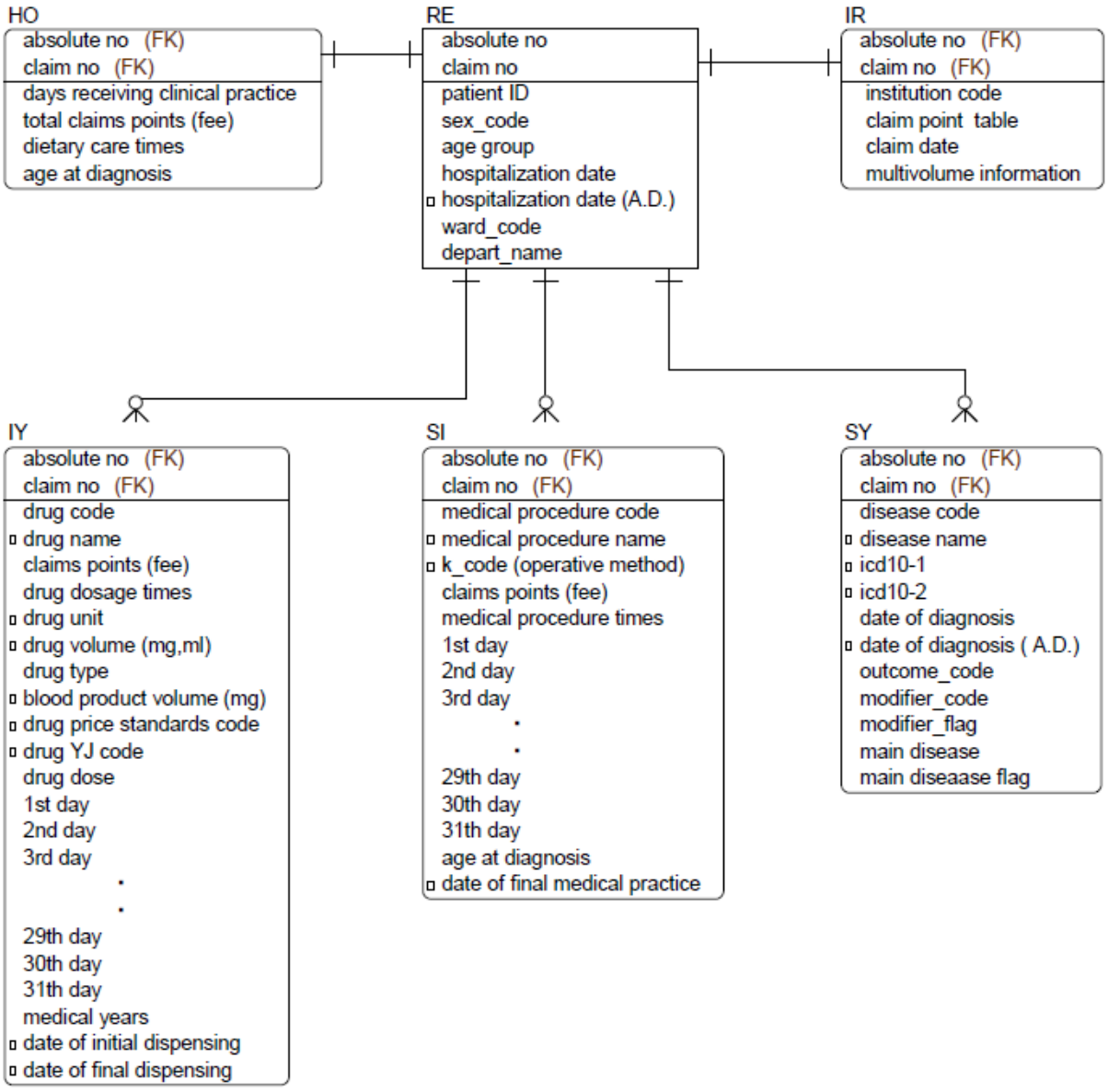


Fig.4.1 RQDW of the NDB-SD (Medical claims)

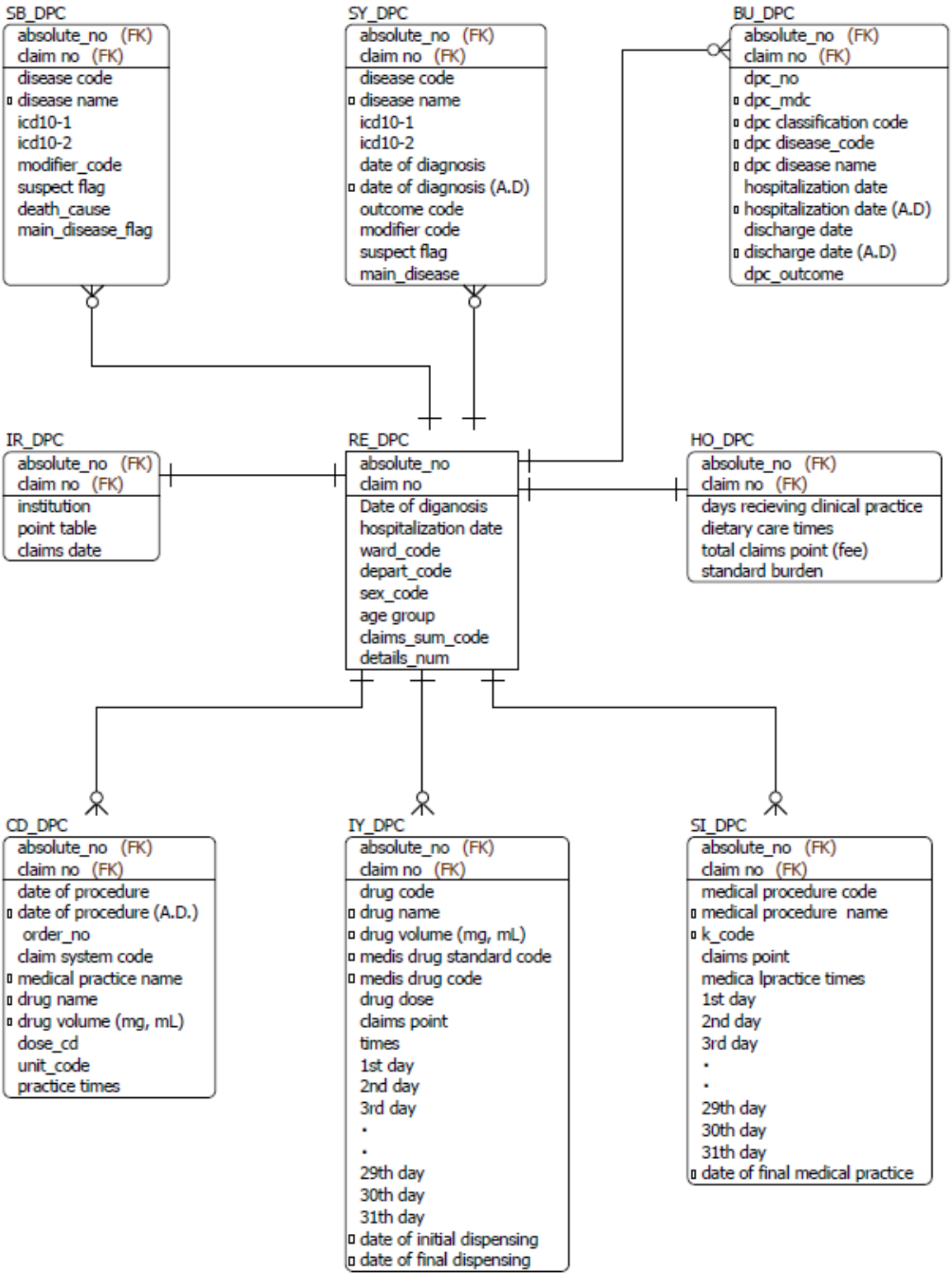


Fig.4.2 RQDW of the NDB-SD (DPC claims)

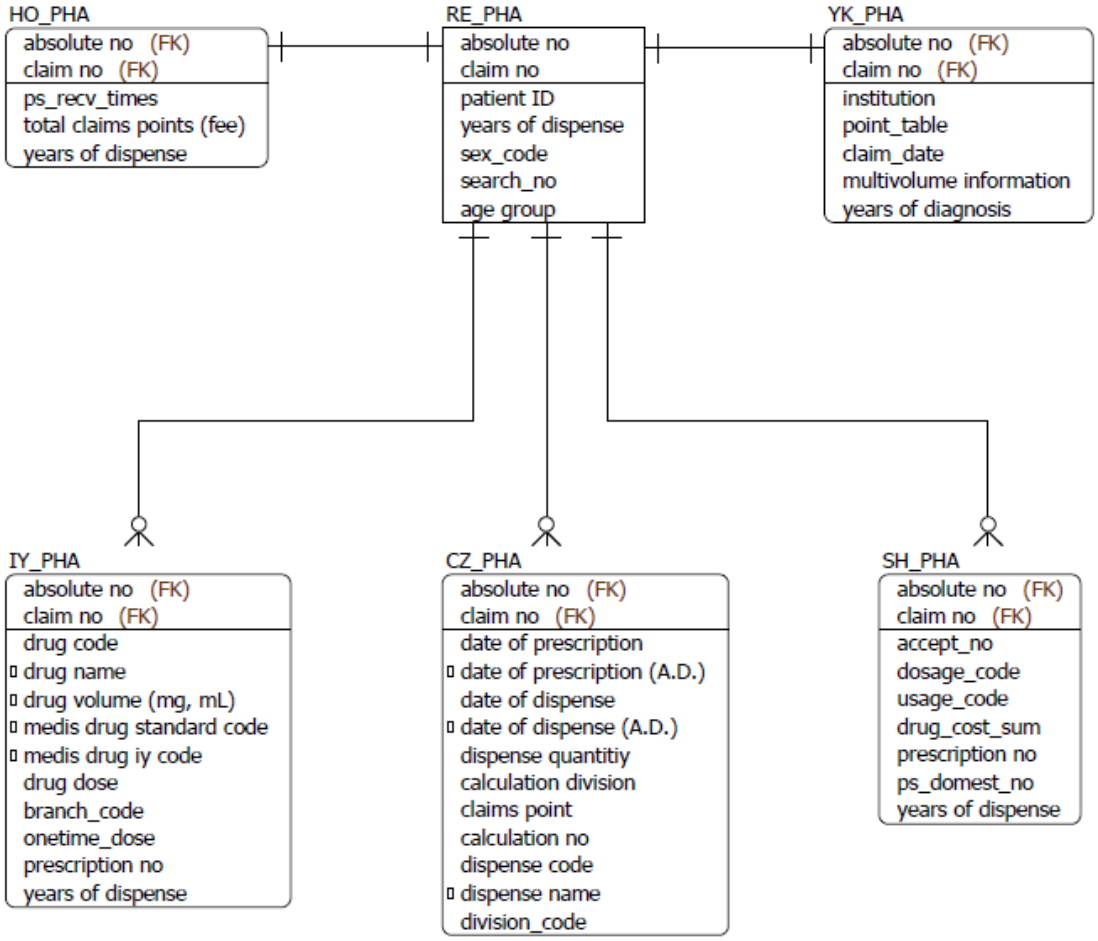


Fig.4.3 RQDW of the NDB-SD (Pharmacy claims)

4.2.2 RQDW of the JMDC claims database

Fig.4.4 shows the RQDW for JMDC claims database. Regarding the JMDC claims database, the RQDW was created in accordance with the research question of the epidemiological studies in Table 4.1. Regarding the study of NTM (Case 3) and Influenza (Case 5) in Table 4.1, all the attributes could be covered with the pattern TRC by using the master data of the MEDIS-DC and data conversion of date-related attributes. On the other hand, the study of MAC Lung Disease (Case 4) is a relatively complex pharmacoepidemiology study. Pharmacoepidemiology is the epidemiological study of the use, and effects, of drugs and other medical devices in large numbers of people [71]. In the case 4 study, it was necessary to compute the prescription period and the single agent prescription period for specific drugs. Therefore, the Prescription relation (PP) was created by analyzing the Drug relation (DR) in Fig.4.4. In the Prescription relation, information about whether a single drug is

administered is added as an attribute for each drug. In addition, in order to compute the continuous prescription period, the days when the drug runs out is also computed and added as attribute.

In the next section, we describe the epidemiological findings obtained from the study of MAC Lung Disease (Case 4).

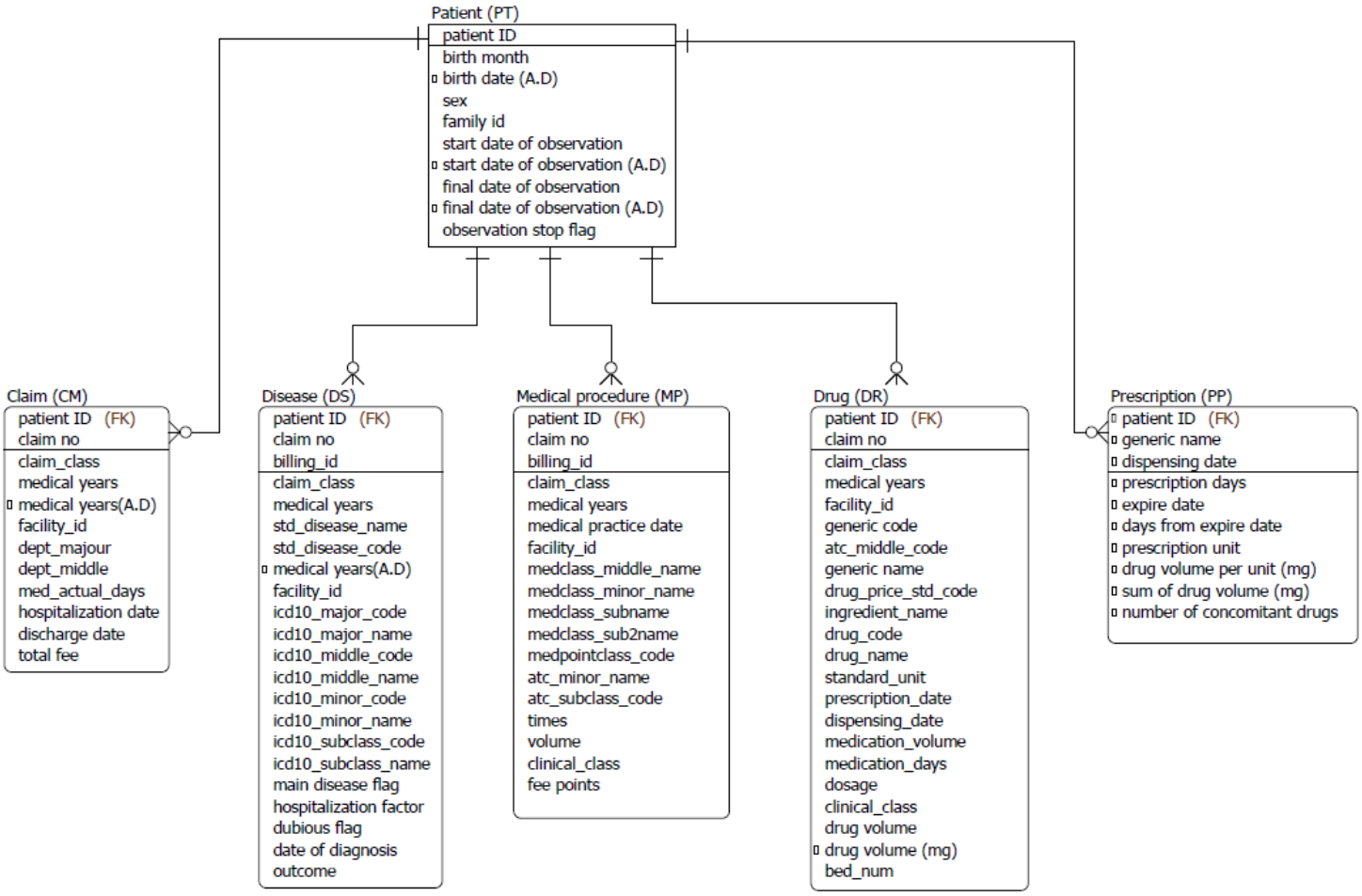


Fig.4.4 RQDW of the JMDC claims database

4.3 Epidemiological study of the MAC Lung Disease

Pharmacoepidemiology is one of the research fields where database utilization is expected[72,73]. Pharmacoepidemiologic studies often require long-term follow-up of pharmacy claims, and complex data handling is often required. The study of MAC Lung Disease (Case 4 in Table 4.1) conducted in this research is one of them. In this section, we describe the epidemiological findings of the MAC Lung Disease.

Background of NTM and MAC Lung Disease

Nontuberculous mycobacteriosis (NTM) is increasingly becoming a public health issue, and is associated with a growing financial burden, particularly for elderly patients [74]. A national survey conducted in Japan in 2014 found that 14.7 of every 100,000 people are infected with NTM, approximately 2.6 times higher than the prevalence found in a similar national survey conducted in 2007.Of lung disease caused by NTM, 88.8% are caused by *Mycobacterium avium-intracellulare* complex (MAC), 4.3% are caused by *M. kansasii*, and 3.3% is caused by *M. abscessus* [75] . This study focused on MAC lung disease, which is a major contributor to the rapid increase in the prevalence of NTM.

Among patients with MAC lung disease, chemotherapy with antibiotics is the most common form of treatment. Although clear treatment guidelines have not been developed in all countries[76 77], in Japan, triple-agent therapy with clarithromycin (CAM), ethambutol (EB), and rifampicin (RFP) is recommended by specialists as the standard treatment regimen. In addition to these three drugs, the latest Japanese treatment guideline specifies streptomycin (SM) and kanamycin (KM) may also be used to treat patients with severe disease [78]. However, the recommended regimens are associated with numerous side effects such as diarrhea, nausea, hepatopathy, hematopenia, and visual disturbance, due to the use of multiple drugs and the long duration of treatment [79]. Regarding the duration of treatment, the American Thoracic Society (ATS) guidelines state that providing an antibiotic regimen for a period of ≥12 months from the time of conversion to a bacilli negative state is necessary because patients with MAC lung disease are unusually difficult to treat due to the moderate effectiveness of antibiotics [80]. Conversion to a bacilli-negative state generally takes six to twelve months from the start of antibiotic administration[81] . The United Kingdom guideline recommends that chemotherapy for MAC lung disease be maintained for a minimum of 12 months from the time of culture conversion [82].

In Japan, a clinical cohort study demonstrated that treatment periods longer than those suggested by ATS lead to better outcomes [83]. However, it is the proportion of individuals with MAC lung disease who are treated in accordance with the guidelines has not been investigated in Japan.

Moreover, there is also risk of ethambutol optic neuropathy due to the long-term administration of high doses Ethambutol (EB). Visual impairment associated with EB was the most common adverse drug reaction. According to Japanese treatment guideline for MAC lung disease, the maximum dose of EB is 750 mg regardless of patient's body weight. Therefore, in this manuscript, we define EB doses not less than 1000 mg as high doses.

In recent years, insurance claims data have been used in epidemiological research [84] . As insurance claims data are based on real clinical settings, they can be used to determine the pattern of use of chemotherapy in medical facilities. In this study, we used insurance claims data to clarify the issue.

Materials and Patient Eligibility

This study used insurance claims data from the Japan Medical Data Center (JMDC). We obtained data in comma-separated values (CSV) format (data size: 5 gigabyte) on patients' background, with data on patient characteristics, such as age, and sex (PATIENT Table), disease information (DISEASE Table), and drug information such as drug name, date of prescription, and drug volume etc. (DRUG Table) for 5,151,648 individuals for the period from January 2007 to December 2017. Moreover, in this epidemiological study, we used the RQDW of JMDC in Fig.4.4.Table 4.8 shows eligibility and exclusion criteria. First, we selected patients assigned the pulmonary mycobacterial infection (ICD-10 code A310) because this study involved patients with pulmonary diseases (Eligibility Criterion 1). Inclusion criteria comprised patients with MAC lung disease who had been prescribed clarithromycin (Eligibility Criterion 2) at least twice, to account for the possibility that patients who had been prescribed CAM only once could may have had it prescribed for another medical condition, such as a cold. As for EB prescription, in addition to CAM, azithromycin (AZM), which is not described in Japanese guidelines but is commonly used for patients with MAC lung disease worldwide, was added.

We also set an exclusion criterion to eliminate patients with diseases that were difficult to distinguish from MAC lung disease at the time of diagnosis. First, in order to exclude patients with a lung disease other than MAC lung disease, we excluded patients who had been prescribed CAM <400 mg and patients with chronic bronchitis, bronchiectasis and diffuse panbronchiolitis. Second, we excluded patients who had been prescribed pyrazinamide, isoniazid, or imipenem, which are first-line drugs for pulmonary tuberculosis, *M. kansasii* disease, and *M. abscessus* disease, and are not prescribed for MAC lung disease.

In this study, we defined CAM monotherapy as not having had a prescription for any generalized antibacterial drug (ATC code: J01) or antimycobacterial drug (ATC code: J04) other

than CAM, after being diagnosed with pulmonary mycobacterial infection (ICD-10 code A310).

Table 4.8 Patient eligibility criteria

Observation Period
January 2007 to December 2017
Eligibility criterion
1. Patients with the ICD10 code: A310 2. Patients with ≥2 previous CAM prescriptions
Exclusion criterion
1. Patients who had been prescribed <400 mg CAM after diagnosis 2. Patients who had been diagnosed with chronic bronchitis, bronchiectasis or diffuse panbronchiolitis 3. Patients with a who had been prescribed pyrazinamide 4. Patients with a who had been prescribed isoniazid 5. Patients with a who had been prescribed imipenem
Definition of continuous administration Repeat prescription of the same drug within 30 days, upon completion of the first prescription

4.3.1 Dataset generation with the pattern TRC

As one problem here, unlike data sets built for physician-scientist-led clinical trials, information about the patient is distributed and stored in different tables, therefore, the data were not suitable for epidemiological analysis.

To address this issue, we created an analysis unit schema which summarized data in one tuple for each patient using the pattern TRC described in chapter 3 [61].

So, we developed a method to simplify the SQL code to enable us to do complex data analysis such as computing the drug medication period and drug choice pattern for a long period for each patient.

To simplify the SQL code, we divided the research questions into smaller units. For example, in order to analyze data for patients with MAC lung disease who had been administered CAM and who had died, three conditions had to be fulfilled, namely:, (i) patients with MAC lung disease; (ii) and who had been administered CAM; and (iii) who had died could be divided into three attributes (variables). As for the JMDC claims database, the PATIENT relation corresponds to PR and the analysis unit schema (Time Series) to the EDR.

When merging one of these relations with the PATIENT relation that contained unique patient's identification (ID), attributes such as A1 and A2 (e.g., the name of any disease or drug) could be computed and extend it to the analysis unit schema in Fig.4.5 respectively. Through this process, the PATIENT relation, which contains the analysis units (patient ID), and a DISEASE or DRUG relation that contains clinical information can be merged to compute a single attribute. In this way, the method allows users to compute an attribute in a routine pattern.

Fig.4.5 also shows the SQL source codes for computing a single attribute. In this method, attributes other than date-related data, such as A1 and A2, were assigned to the analysis unit schema, whereas date-related data such as T1 and T2 were assigned to the analysis unit schema (Time Series).

The underlined section No.1 in Fig.4.5 is the attribute name which users wish to compute. Similarly, the underlined section No.2 and No.3 were obtained through 'JOIN' on the PATIENT and No.4 DISEASE or DRUG relation. The underlined section No.3 is a SQL function which can aggregate patient information and compute parameters such as the total, the number of cases, and the maximum value. The underlined section No.5 lists the extraction conditions (e.g. name of the disease, drug code, prescription date). Through listing the attributes that were contained in the analysis unit schema (Time Series) in a conditional clause, these attributes could also be used as conditions for extracting time series data (e.g., 30 days after the date of diagnosis of a specific disease). By using above-mentioned technique, we created a dataset summarizing the data for each patient with MAC lung disease in one tuple. Thus, we determined the number of patients who were prescribed CAM and other drugs and the number of patients who were administered not less than 1000 mg of EB per day continuously for 90 days or longer, considering that administration of not less than 1000 mg of EB per day for the two months directly after initial administration is permitted for treatment of pulmonary tuberculosis [85].

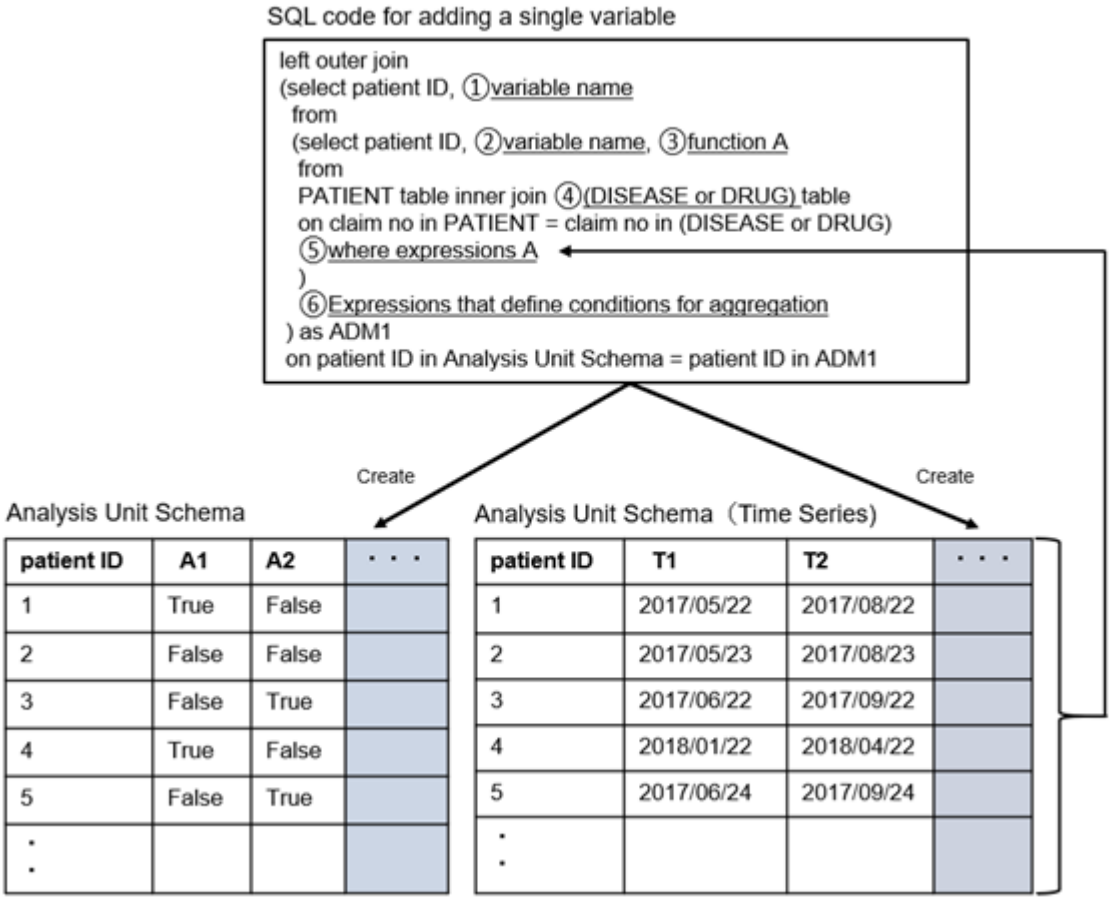


Fig.4.5 The SQL based on the pattern TRC and schema structure

4.3.2 Results of prescription pattern

Fig.4.6 shows a flow chart for patient selection. The patient data extracted using the eligibility criteria established for this study resulted in the identification of 1037 patients with NTM who met the eligibility criterion, of whom 659 were excluded according to the exclusion criterion, leaving 378 patients with MAC lung disease for inclusion in the analysis. Table 4.9 shows the number of patients and the incidence rate according to age and sex, relative to the total population. Among both males and females, the incidence rate of MAC lung disease increased with age. As for the survey of EB prescription, the patient data extracted using the eligibility criteria established for this study resulted in identification of 1149 insured individuals who met the eligibility criteria and 14 individuals who were excluded according to the exclusion criteria, leaving 1135 individuals as our analysis subjects.

Table 4.10 shows the drug regimens prescribed for patients with MAC lung disease. As a criterion to determine how long chemotherapy had been administered, we defined the CAM prescription end date as the end of chemotherapy. Moreover, the criterion for determining the CAM prescription end date was when there was no repeat prescription during a period of more than 30 days after a patient

after the end of the prescription period. The majority of patients (63%) were prescribed triple-agent therapy with CAM, RFP, and EB (CR+EB). This included patients who had been prescribed the drug for more than six months (73.1%), and for more one year (48.3%). In addition, some patients (2.9%) had been administered CAM monotherapy.

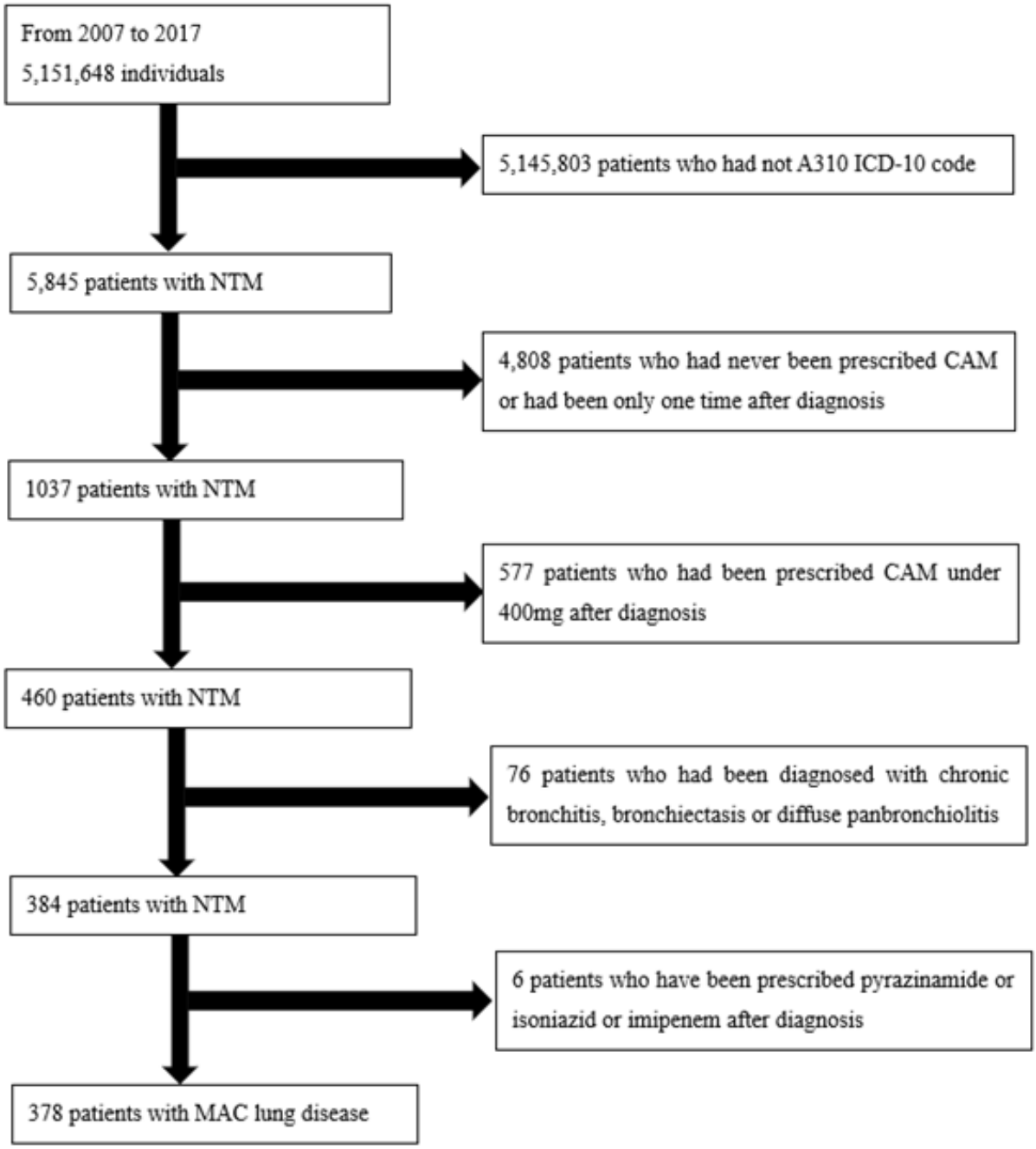


Fig.4.6 Patient selection flow chart

Table 4.9 Age distribution of patients with *Mycobacterium avium-intracellulare* complex lung disease and the age-specific incidence rates

		Population (N = 5,151,648)		Patients in the study (N = 378)		Morbidity (Patient/Population)	
		Male	Female	Male	Female	Male (%)	Female (%)
Age (years)	0–20	783,797	696,405	0	0	0.000	0.000
	21–30	544,887	402,657	3	3	0.001	0.001
	31–40	465,330	430,007	5	9	0.001	0.002
	41–50	437220	407059	14	47	0.003	0.012
	51–60	317043	294033	25	126	0.008	0.043
	61–70	198429	142082	45	79	0.023	0.056
	71–75	13238	19461	10	12	0.076	0.062
Total		2,759,944	2,391,704	102	276	0.004	0.012

CAM: clarithromycin

Table 4.10 Drug prescription pattern for the patient with MAC Lung Disease

	Total (N = 378)		1-89 (days)		90-180 (days)		181-270 (days)		271-365 (days)		≥365 (days)	
	N	%	N	%	N	%	N	%	N	%	N	%
1. CAM	11	2.9	6	1.6	2	0.5	1	0.3	1	0.3	1	0.3
2. CAM+RFP	17	4.5	0	0.0	2	0.5	1	0.3	3	0.8	11	2.9
3. CAM+EB	15	4.0	1	0.3	2	0.5	2	0.5	2	0.5	8	2.1
4. CAM+SM	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
5. CAM+KM	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
6. CR+SM	2	0.5	0	0.0	0	0.0	0	0.0	2	0.5	0	0.0
7. CR+KM	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
8. CAM+EB+SM	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
9. CAM+EB+KM	1	0.3	0	0.0	0	0.0	0	0.0	0	0.0	1	0.3
10. CR+EB	238	63.0	24	6.3	40	10.6	27	7.1	32	8.5	115	30.4
11. CR+EB+SM	16	4.2	2	0.5	3	0.8	3	0.8	1	0.3	7	1.9
12. CR+EB+KM	10	2.6	0	0.0	1	0.3	0	0.0	2	0.5	7	1.9
13. CAM+other drugs	68	6.9	-	-	-	-	-	-	-	-	-	-

CAM: Clarithromycin, EB: ethambutol, RFP: Rifampicin, SM: Streptomycin, KM: kanamycin

CR: Clarithromycin and Rifampicin

4.3.3 Results of prescription state of Ethambutol

Table 4.11 shows drug prescription conditions of EB for patients with MAC lung disease. Out of 507 cases who were prescribed EB, 22 cases (4.3%, 22/507) were prescribed not less than 1000 mg of EB per day. Among these 22 cases, 10 were prescribed 1000 mg per day, 1 was prescribed 1250 mg per day, and 11 were prescribed 1500 mg per day. 18 patients were prescribed high doses of EB (not less than 1000mg/day) continuously for 90 days or longer. As for ethambutol optic neuropathy, none of 22 patients who were prescribed high doses of EB had ethambutol optic neuropathy. In all patients receiving ethambutol, 21 patients (4.1%, 21/507) had ethambutol optic neuropathy. Regarding standard regimen with CAM, RFP, and EB, 474 patients were prescribed EB in Table 4.11. Aa result, 16 cases (3.6%, 16/474) were prescribed not less than 1000 mg of EB for 90 days or longer.

Table 4.11 Ethambutol prescription conditions for patients with *Mycobacterium avium-intracellulare* complex lung disease

	EB	EB not less than 1000 mg (1-89days)		EB not less than 1000 mg (For 90 days or longer)		EB with CR	EB with CR not less than 1000 mg (1-89days)		EB with CR not less than 1000 mg (For 90 days or longer)	
	N	N	%	N	%	N	N	%	N	%
Male	133	1	0.8	7	5.3	126	0	0.0	6	4.8
Female	374	3	0.8	11	2.9	348	1	0.3	10	2.9
Total	507	4	7.9	18	3.6	474	1	0.2	16	3.4

CAM: Clarithromycin, EB: Ethambutol, CR: Clarithromycin and Rifampicin

4.3.4 Discussion and Conclusion of MAC lung disease studies

In this study, we determined the actual drug prescribing pattern for patients with MAC lung disease using claims data. Approximately 48.3% of the patients received basic triple-agent therapy for more than one year in this regimen. In addition, there were some cases of CAM monotherapy prohibited by the Japanese treatment guideline.

First, with regard to the prescribing pattern in Table 4.10, it was noteworthy that CAM monotherapy, which is strongly contraindicated in patients with MAC lung disease, had been prescribed for 11 (2.9%) patients, and had been continued for at least three months in 5 (1.3%) patients. One study found that CAM monotherapy may have been prescribed for patients who experienced adverse drug reactions with multiple agents' chemotherapy and who strongly dislike long-term treatment with multiple agents [86] . Another study found that MAC bacteria acquire drug resistance to CAM quite easily in vitro environments [87] . As MAC lung disease is usually a chronic infection, development of CAM-resistant bacterial strains is problematic. Another study found that some MAC bacteria acquired genes from external bacterial species and performed chromosomal recombination with other species, a phenomenon that does not occur with M.tuberculosis [88] . Generally, long-term CAM monotherapy induced CAM-resistant bacterial strains at an early stage (between the 2nd and 5th months) [78] . In cases where surgical resection and long-term injection of aminoglycoside drugs in combination could not be used after diagnosis of resistant bacteria, 45% of patients whose condition had deteriorated due to resistant bacteria died in 2 years [89] . Therefore, it is important to avoid acquiring drug-resistance resulting from CAM monotherapy. However, our results indicate the possibility that evidence-based drug treatment for patients with MAC lung disease is sometimes inadequately applied in current clinical practice in Japan. Second, in Table 4.10, although triple-agent therapy with CAM, RFP and EB (CR+EB) was administered for the largest number of patients, combination therapy with CAM, RFP, EB and SM (CR+EB+SM) and combination therapy with CAM, RFP, EB and KM (CR+EB+KM) prescribing patterns accounted for 6.8% (26 patients), which is extremely low. Aminoglycosides such as SM are highly active against extracellular organisms and decrease the time to bacilli negative conversion[81]. In addition, a clinical study has confirmed the effectiveness of SM combined with CAM, RFP, and EB [90] . The result was adopted in the latest Japanese treatment guidelines. The reason why there have been few cases of combined treatment with aminoglycosides is mainly because of the toxicity risks associated with the long-term use of aminoglycosides, especially nephrotoxicity and ototoxicity, and because of the need for frequent therapeutic drug monitoring. It appears that 46 (12.2%) patients in the prescription patterns No.1 to No.9 in Table 4.10 could not be treated with the basic triple-agent presumably due to side effects. In addition, no more than 48.3% of the patients with basic triple-agent therapy in

No.10 in Table 4.10 had been prescribed continuously for more than one year. From these facts, it is suggested that it would be practically difficult to continue treatment with basic triple-agent for a long time because of problems such as side effects and poor adherence to guideline and patient's complaint etc. There are several limitations in this study.

First, although we applied criteria to exclude patients who did not have MAC lung disease, clinical diagnostic criteria (a positive sputum culture on at least two occasions or a single positive culture for bronchoscopic samples) could not be confirmed due to the limitations of the claims data. Therefore, in this study, the sensitivity and specificity of the patients with MAC lung disease has not been discussed. Second, the data used in this study does not include the elderly, who are considered to occupy the largest number of patients with MAC lung disease in Japan [91] . Since elderly patients would be more vulnerable to side effects of drugs, the standard triple-agent therapy might be less often tolerable in this population. Moreover, due to the limitations of the claims data, it was difficult to infer why the standard treatment had been discontinued. Finaly, the JMDC claims data include those who withdrew from the insurer for some reason during the observation period. In this analysis, such people were included in patients with MAC lung disease, and the drug prescription period was determined as the time until they became no longer insurers.

Next, Table 4.11 shows cases where not less than 1000 mg of EB per day was prescribed continuously for 90 days. Although administration of not less than 1000 mg per day is permitted in Japan, depending on patient weight, for up to two months starting from initial administration as treatment for pulmonary tuberculosis, it is not recommended to continue beyond that point regardless of patient weight. Incidence rate is low at standard doses, but optic nerve phase is the most likely side effect of ethambutol [92]. Other studies have reported loss of vision in response to long-term high doses of EB [93].

This study of MAC lung disease clarified the status of chemotherapy currently being applied in medical institutions for patients with MAC lung disease. Chemotherapy for patients with MAC lung disease was often not continued for the standard treatment period. Therefore, widely publicizing the need for application of standard treatment for at least two years is necessary, as is the need to develop methods to overcome side effects or provide an alternative treatment regimen for MAC lung disease. We hope that the results obtained in this study will lead to further clinical trials and other related research in the future.

Chapter 5 Discussions

In this dissertation, we focused on the characteristics of research questions in epidemiological studies and closure property of the relational model theory and proposed the pattern TRC that can easily compute desired attributes. Next, by building the RQDW optimized for the pattern TRC, all the attributes required for each study in Table 4.1 were supplemented.

5.1 Simplicity of the pattern TRC

The reason why physician-scientists were able to carry out data handling on their own is thought to be the simplicity of the pattern TRC. To use the pattern TRC in equation (3.1) and (3.3), users only learn how to use the function equivalent to "*summary*" in equation (3.2) and (3.4) and some basic SQL knowledge of “Select”, “From”, and “Where”. The “*summary*” may be implemented by the window function of SQL. As shown in Table 4.2 to 4.6, only two window functions, "row_number ()" and "sum ()", are necessary for all epidemiological studies. Therefore, the physician-scientists were able to perform data handling by learning only these two functions. At the same time, the pattern TRC is also useful for almost software engineers. This is because the pattern TRC can quickly complete the programming tasks for dataset generation that they think are not creative, bothersome and time-consuming.

5.2 Coverage rate of the pattern TRC

In this section, we consider the coverage of the pattern TRC. As shown in Table 4.1, the pattern TRC covered not less than 80% of the required attributes in the five epidemiological studies. Since the pattern TRC can cover any attribute stored in the database, it is speculated that the attributes originally prepared in the database are often used in epidemiological studies. The reason for the high pattern TRC coverage is that it also supports date-related (event-driven) attributes frequently used in cohort studies.

In the cross-sectional studies in Tables from 4.2 to 4.4, only PT and PT × NPR combinations were used to compute attributes. In general, because cross-sectional studies do not need to track patients in time series, there is often no association between attributes.

In the cohort studies in Tables from 4.5 to 4.6, the combination of PT × EDR is frequently used. In the cohort studies, because it is necessary to resolve the relationship such as the occurrence order or date-related relationship that exist between attributes, the pattern TRC in equation 3.3 is effective. If the pattern TRC for an event-driven attribute and the EDR did not exist, the attribute coverage in

case 4 and case 5 studies is 32% (8/25) and 9% (3/35) respectively, and the coverage would be greatly reduced.

Thus, although the attribute coverage by the pattern TRC is relatively high, there are some attributes that must be supplemented to the original database. Examples of such attributes are those does not physically exist in the original database and those required certain data transformations and computations. In the next section, the RQDW to supplement these attributes is discussed.

5.3 Role of the RQDW

As shown in Table 4.7, age and sex were used as attributes in most epidemiological studies investigated. In addition, a lot of date-type attributes such as date of prescription or diagnosis have appeared. The date-type attributes are very important in our method. Because date-type attributes are the target of the pattern TRC and Extend operator, it is important that these attributes are stored in the A.D. format (e.g. 2007/10/15). For this reason, as shown in Fig. 4.4, the RQDW supplements the attribute of birth date by interpolating the day in the birth month. Thanks to this process, the attribute of age can be computed with the *summary* function and Extend operator.

Next, let us consider the attributes that can be supplemented from external master data. Since the volume of RCC, FFP, and PC in Table 4.2 do not exist in the original NDB-SD, the RQDW in Fig.4.1 supplements the drug volume (mg, ml) by processing using the MEDIS master data. Similarly, for the RQDW of JMDC claims data, the attribute drug volume (mg) was added to the Drug relation in Fig. 4.4 to supplement the maximum amount of EB in Table 4.4 and 4.5 and the amount of CAM in Table 4.5. In the five epidemiological studies in Table 4.1, most attributes can be handled by the pattern TRC by converting date-related attributes into an appropriate format and supplementing the missing attributes with external master data.

However, in the study of MAC Lung disease in case 4, there are attributes that cannot be handled by the pattern TRC only by the processing described above. In order to handle the period of EB prescription and period of CAM monotherapy in Table 4.5 with the pattern TRC, it is necessary to create a new relation PP. One reason for this is that the drug prescription period can only be computed by searching all prescription histories for each patient. Since only the prescription date and the number of prescription days are described in the Drug relation, it is necessary to compute expire date. In addition, it is necessary to compute how many days have passed from the expire date until the next time the same drug is prescribed.

Such a complex computation cannot be realized with just the *summary* function, so it is necessary to compute it in advance when building the RQDW. In this way, in our method, the

summary function of the pattern TRC in equation (3.2) and the RQDW are sharing work to cover the attributes required for epidemiological studies.

5.4 Extensibility of the RQDW

As mentioned in the previous section, it is necessary to build the RQDW for complex research questions. In principle, the RQDW becomes more sophisticated as new epidemiological studies are conducted. For example, it may be possible to deal with more abstract research questions by referring to the techniques that supplements time-series information for chronic diseases[94].

This is because the more epidemiological studies are conducted, the more attributes required for research can be added to the RQDW. Thus, the RQDW is made step by step, each new relations and attributes that are added forming a basis for further advances. Moreover, the OODA loop that are said to be effective in achieving goal will rotate[95]. The goal here refers to having all attributes used in all epidemiological studies in the RQDW. If the RQDW is refined by conducting more epidemiological studies in the future, it is possible to obtain more generality allowing the pattern TRC alone to cover most epidemiological studies.

On the other hand, when building the RQDW, it is necessary to consider the nature of the database to be analyzed. As for the NDB, there is a problem that the reliability of the patient ID is low. In order to solve above-mentioned problem, researches to create a highly reliable patient ID are also being conducted[96,97]. In addition, there are many things that need to be taken care of when analyzing data in health insurance claims database[98]. A research is also underway to prepare tables suitable for analysis of information such as hospitalization and discharge dates that are considered difficult to interpret in health insurance claims data[99]. It will be further refined by developing the RQDW with reference to such researches. Thus, when building the RQDW, it is necessary to carefully consider the characteristics of the databases to be analyzed.

5.5 Versatility of the pattern TRC and RQDW

Let us consider the versatility of the patterns TRC and RQDW proposed in this dissertation. Since this method does not depend on the data structure and contents, so it can be applied not only to health insurance claims databases in the world but also to databases such as electronic medical records. It is also suitable for distributed database management systems [100]. This is because the combinations of relation PR and NPR in equation (3.1) can be computed in parallel by storing them in separate computers. Researchers can obtain the target dataset by calculating the attributes separately on each computer and combining all the calculation results later.

Chapter 6 Conclusions

The purpose of this research was to establish a methodology for creating a dataset summarized in one tuple for each patient from a healthcare database on their own, even for users without data handling skills. The author focused on the fact that the relational model theory can be used to generate the dataset by regarding research questions as a proposition. As mentioned in this manuscript, this research proposed a method that can create a dataset for each patient in a simple procedure by cooperating the pattern relational calculus and a data warehouse optimized for it.

With proposed method, two physician-scientists who have little data handling skills were able to create the dataset for respective epidemiological study on their own on. These two epidemiological studies include cross-sectional and retrospective cohort studies, which account for most of the epidemiological study using health care databases. Therefore, although the scope of application is limited to epidemiological studies, it is considered that the usefulness of this method has been verified to a certain extent. Moreover, since the method is applicable to any database that includes unique ID for people such as patient, it can be applied not only to the health insurance claims databases used in this study, but also to many healthcare databases such as electronic medical records.

The author thinks engineers, epidemiologists, medical doctors, and other occupations, all play a big part in epidemiological studies using the health care databases. The author supposes that there are some people who have a disease that they want to investigate, not just medical doctors.

For many people to participate in epidemiological studies using health care databases in the future, it is necessary to lower the threshold as much as possible. Such efforts to make data handling simpler and more patterned than it is at present might be necessary to broaden the reach of epidemiological studies to be conducted by many types of people. This dissertation lowered the threshold from the viewpoint of data handling.

Although proposed method cannot theoretically cover all epidemiological studies, the author believe that it may cover most of the studies that researchers may come up with. The answer will be clear in the immediate future.

6.1 Future Application

As mentioned above, this research will be able to create datasets with simple work in most epidemiological studies by refining the data warehouse.

Once the data warehouse is complete, the author believes that entire process from dataset generation to statistical analysis can be automated in the future. The datasets generation process could be automated by analyzing research questions using such as natural language processing

technology. It is also possible to automate statistical analysis by collecting research questions and applying them to techniques such as machine learning. This dissertation is the basis for achieving these goals from a technical perspective.

Abbreviations and Terminologies

NDB: National Database of Health Insurance Claims and Specific Health Checkups of Japan
NDB-SD: NDB Sampling Dataset
DPC: Diagnosis Procedure Combination
DW: Data Warehouse
Schema: design of attributes in relations in the relational model theory
Relation: table with no duplicate tuple
Attribute: column name or variable name in relations
Primary key: one or more attributes that uniquely identify each tuple in a relation
Tuple: combination of attributes in a relation, almost the same meaning as row in general
Tuple Relational Calculus (TRC): logical expression used in the relational model theory
SQL: database language based on the relational model theory
RQDW: Research Question oriented Data Warehouse

Ethical approval

The epidemiological studies using the NDB-SD in this dissertation has passed the ethical application of Kyoto University Hospital. On the other hand, as the studies regarding the JMDC claims data was anonymized data without intervention in a completely anonymized form, approval from an ethics committee was not necessary according to the rules of the affiliated institution, Nara Institute of Science and Technology.

Acknowledgement

I am convinced that this doctoral dissertation was able to be accomplished thanks to the cooperation of various people and my good luck.

Firstly, I would like to express my sincere gratitude to Professor Tomohiro Kuroda who gave me the opportunity to enter this laboratory. Without his generous support, this dissertation would not have been completed. I would like to offer sincere thanks to Associate Professor Genta Kato who took me as his student and inspired me with many discussions during my Ph.D. course. Without his teaching me how to write a journal paper, I wouldn't have been able to publish several journal papers. I am very grateful to Professor Masatoshi Yoshikawa who gave me much informative advice from a professional perspective on database studies. I am grateful to Professor Takeo Nakayama who gave me much informative support from a professional perspective on epidemiology. Also, I am thankful to Professor Hiroshi Tamura, Associate Professor Kazuya Okamoto, Assistant Professor Shusuke Hiragi, and Shosuke Ohtera for my research support.

Moreover, I am also thankful to Program-Specific Professor Masayuki Nambu, Program-Specific Senior Lecturer Hiroshi Sasaki, Goshiro Yamamoto and Osamu Sugiyama who gave me some advice in the research meeting.

I am very thankful to Associate professor Shigeru Otsuru, Eiji Kondo and researcher Yuka Nakatani and Mai Sato who gave the opportunity to do the epidemiological studies. I would like to express my sincere thanks to Senior Lecturer in the Department of Respiratory Medicine Isao Ito for giving me valuable guidance on how to write journal papers in the medical field and researcher Masahiro Shirata who used my methodology in conducting the epidemiological study. My experience in the doctor course gave me a better glimpse than I had had before of the real nature of human beings.

In addition, I am very grateful to the teachers who are not affiliated with Kyoto University.

I am very thankful to Michi Sakai who gave me much informative advice and great achievement. I am also very grateful to Associate Professor of NAIST Eiji Aramaki who gave me the opportunity to work as a researcher and gave me chance of tackling several epidemiological studies. I am also grateful to Shigeko Noguchi who gave me epidemiological research questions. I would like to thank secretaries Yuko Furusawa, Yoko Hara, Kaori Shiomi and Kuniko Hiramatsu who assisted me with daily work, always gave me great support. Finally, I am very grateful to all the Ph.D. course students and Master course students in the Medical Informatics laboratory.

Appendix

Appendix A1: Relation names and main attribute names used in the NDB-SD in this dissertation

(The name of the relation was the one given by the Japanese Ministry of Health, Labor and Welfare.)

Relation (Table)	**Explanation**	
RE	Information of patient's background such as sex, age group	
HO	Information of medical fee	
IR	Information of medical institution	
IY	Information of drug such as drug name etc.	
SI	Information of medical procedure	
SY	Information of disease such as disease code, disease name, etc.	
SB	Information of disease in DPC claims	
BU	Information of classified diagnostic groups	
CD	Information of drug, medical procedure in DPC claims	
YK	Information of pharmacy data	
CZ	Information of prescription such as date of prescription etc.	
SH	Information of prescription management such as order number etc.	
Attribute (Variable)	**Explanation**	**Relation**
Patient ID	Unique number per patient (insured)	RE
Claim no	Bill number issued monthly by the medical institution	All relation
Absolute no	Row number in one insurance claim (bill)	All relation
Disease code	Code assigned to the disease name used for insurance claims	SY, SB
Disease name	Disease name corresponding to disease code	SY, SB
Drug code	Code assigned to the drug name used for insurance claims	IY, CD
Drug name	Drug name corresponding to drug code	IY, CD

Appendix A2: List of references used for in creating the Research Question oriented DW of NDB. (These lists are called the NDB literatures in the dissertation.)

Author	Title	Source
Mayumi Seki et al.	Drug Utilization Study of Potential Drug-drug Interactions using the Sampling Dataset of the National Database of Health Insurance Claim Information (In Japanese)	An Official Journal of the Japan Association for Medical Informatics. Vol.34(6), pp.293-304, 2015
Naomi Iihara et al.	Survey of Usage of Medication with Driving with Prohibition or Caution by the National Health Insurance Claims Database in Japan (In Japanese)	Japanese journal of Pharmaceutical Health Care and Sciences Vol.40(2), pp.67 -77, 2014
Ryosuke Arakawa et al.	Examination of Prescription Condition of Anxiolytics Hypnotics in Outpatient Clinical Treatment Using National Database (In Japanese)	Japanese journal of clinical psychiatry. Vol. 44(7), pp1003 -1010, 2015
Yuko Sato et al.	A Preliminary Survey to Measure the Quality Indicators of End-of-life Cancer Care Using the Japanese National Database (In Japanese)	Palliative Care Research, Vol.11(2), pp.156-165, 2016
Yasuyuki Okumura et al.	Prescription patterns of antipsychotic drugs for schizophrenia patients in Japan: Utilization of national database (In Japanese)	Japanese journal of clinical psychiatry. Vol16(8), pp.1201-1215, 2013
Yasuyuki Okumura et al.	Epidemiology of overdose episodes from the period prior to hospitalization for drug poisoning until discharge in Japan: an exploratory descriptive study using a nationwide claims database	Journal of Epidemiology, Vol.27(8), pp.373-380, 2017
Yasuyuki Okumura et al.	Risk of recurrent overdose associated with prescribing patterns of psychotropic medications after nonfatal overdose	Neuropsychiatric Disease and Treatment, Vol.13, pp.653-665, 2018
Pharmaceuticals and Medical Devices Agency: Analysis Division	Characteristic survey and prescription survey report using sampling data set of National Insurance claims data (In Japanese)	Survey report by PMDA, http://www.pmda.go.jp/files/000148196.pdf, 2014.
Michio Kimura et al.	Same Examinations in Different Healthcare Providers in the Same Month of Referral, Analysis by Reimbursement Claims Database (In Japanese)	An Official Journal of the Japan Association for Medical Informatics, Vol.35(5), pp.213-217, 2015

Continued the following page

Kiyoshi Kubota et al.	Epidemiology of psoriasis and palmoplantar pustulosis: a nationwide study using the Japanese national claims database	BMJ OPEN. Vol. 5(1), http://dx.doi.org/10.1136/bmjopen-2014-006450, 2015
Yuuki Nakamura et al.	Evaluation of Estimated Number of Influenza Patients from National Sentinel Surveillance Using the National Database of Electronic Medical Claims (In Japanese)	Japanese Journal of Infectious Diseases. Vol. 68 (1), pp. 27-29, 2015
Chieko Ishiguro et al.	A case example of using the sampling dataset of the National insurance claims database: An example survey of formulation of biguanide type diabetes drug (In Japanese)	Japanese Society for Pharmacoepidemiology. Vol.20, pp.80-81, 2014
Kouichi Hosomi et al.	Association of Antipsychotic Use with Extrapyramidal Symptoms: Data Mining of the Japanese National Insurance Claims Database (In Japanese)	Japanese journal of Pharmaceutical Health Care and Sciences, Vol.42(2), pp.87-97, 2016
Mitsutaka Takada et al.	Study on Risk of Gastrointestinal Complications in Low-dose Aspirin Therapy Using the National Receipt Database (In Japanese)	Japanese journal of Pharmaceutical Health Care and Sciences Vol.39(8), pp.471-481, 2013
Toyokawa S et al.	Estimation of the number of children with cerebral palsy using nationwide health insurance claims data in Japan	Developmental Medicine & Child Neurology, Vol.59(3), pp.317-321, 2017
Kitazawa T et al.	Cost Analysis of Transplantation in Japan, Performed with the Use of the National Database	Transplant Proc. Vol.49(1), pp.4-9, 2016
Naomi Iihara et al.	Polypharmacy of medications and fall-related fractures in older people in Japan: a comparison between driving-prohibited and driving-cautioned medications	Journal of Clinical Pharmacy and Therapeutics. Vol.41(3), pp.273-278, 2016
Hagiwara H et al.	The effectiveness of risk communication regarding drug safety information: a nationwide survey by the Japanese public health insurance claims data	Journal of Clinical Pharmacy and Therapeutics. Jun, Vol.40(3), pp.273-278, 2015
Mai Sato et al.	Consideration on actual condition of postpartum hemorrhage cases using a national insurances claims data in Japan (In Japanese).	Journal of Japan Society of Perinatal and Neonatal Medicine, Vol.52(2), p724, 2016
Tetsuya Otsubo et al.	Regional Variations in In-hospital Mortality, Care Processes, and Spending in Acute Ischemic Stroke Patients in Japan	Journal of Stroke and Cerebrovascular Diseases. Vol.24(1), pp.239-251, 2015

References

1. Lusignan D, Crawford L, Munro N. Creating and using real-world evidence to answer questions about clinical effectiveness. *Journal of innovation in health informatics.* 2015;22(3):368-373.
2. Kenneth R. *Modern Epidemiology.* LWW; 2012.
3. Yoshitaka M. Secondary Data Analysis of Epidemiology in Asia. *J Epidemiol.* 2014;24(5):345-346.
4. Hiroshi Y, Syusaku I, Hideaki W. *Textbook of Biostatistical and Clinical Research Design.* Tokyo: Medical Publications; 2015.
5. PMDA. Medical Information Database Network. 2018; https://www.pmda.go.jp/safety/mid-net/0001.html. Accessed 2019/1/1.
6. National Clinical Database. 2019; http://www.ncd.or.jp/. Accessed 10/21, 2019.
7. Laney D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *META Group.* 2001(File:949).
8. Obermeyer. Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med.* 2016;375(13):1216-1219.
9. Ha J, Cho S, Shin Y. Utilization of health insurance data in an environmental epidemiology. *Environ Health Toxicol.* 2015;30:e2015012.
10. Beck M, Velten M, Rybarczyk-Vigouret MC, Covassin J, Sordet C, Michel B. Analysis and Breakdown of Overall 1-Year Costs Relative to Inpatient and Outpatient Care Among Rheumatoid Arthritis Patients Treated with Biotherapies Using Health Insurance Claims Database in Alsace. *Drugs Real World Outcomes.* 2015;2(3):205-215.
11. WIP. Survey of the current state of databases in the fields of health, medical care and nursing care in other countries. 2019; https://www.mhlw.go.jp/content/12400000/000548451.pdf. Accessed 9/29, 2019.
12. Leonard CE, Brensinger CM, Nam YH, et al. The quality of Medicaid and Medicare data obtained from CMS and its contractors: implications for pharmacoepidemiology. *BMC Health Serv Res.* 2017;17(1):304.
13. ResearchDataAssistanceCenter. Find, Request and Use CMS Data. 2018; https://www.resdac.org/. Accessed 1/1, 2019.
14. Lakhani C, Tierney B, Manrai A, Yang J, Visscher P, Patel C. Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nat Genet.* 2019;51(2):327-334.

15. Zhu. W, Chernew. ME, Sherry. TB, Maestas N. Initial Opioid Prescriptions among U.S. Commercially Insured Patients, 2012-2017. *N Engl J Med.* 2019;380(11):1043-1052.
16. John T. *Exporative Data Analysis.* New York: ADDISON WESLEY PUB CO INC; 1977.
17. Sergeant E, Perkins N. *Epidemiology for Field Veterinarians: An Introduction.* London: Springer; 2013.
18. Kim L, Kim JA, Kim S. A guide for the utilization of Health Insurance Review and Assessment Service National Patient Samples. *Epidemiol Health.* 2014;36:e2014008.
19. Codd EF. A relational model for large shared bank. *Communications of the ACM - Special 25th Anniversary Issue.* 1970;26(1):64-69.
20. Date CJ. *Database in Depth: Relational Model for Practitioners.* New York: Oreilly & Associates Inc; 2005.
21. BARC. Relational databases http://barc-research.com/relational-databases/. Accessed 11/25, 2019.
22. Ferdynus C, Huiart L. Technical improvement of cohort constitution in administrative health databases: Providing a tool for integration and standardization of data applicable in the French National Health Insurance Database (SNIIRAM). *Rev Epidemiol Sante Publique.* 2016;64(4):263-269.
23. Johnson SB, Chatziantoniou D. Extended SQL for Manipulating Clinical Warehouse Data. Paper presented at: Proceedings of the AMIA Symposium1999; USA.
24. Kim JA, Yoon S, Kim LY, Kim DS. Towards Actualizing the Value Potential of Korea Health Insurance Review and Assessment (HIRA) Data as a Resource for Health Research: Strengths, Limitations, Applications, and Strategies for Optimal Use of HIRA Data. *J Korean Med Sci.* 2017;32(5):718-728.
25. Lohr S. For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. 2014; https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html. Accessed 08/30, 2019.
26. Shigeki Y. *Logic.* University of Tokyo press; 1994.
27. Standardization IOf. ISO/IEC 9075-1:2008 Information technology -- Database languages -- SQL -- Part 1: Framework (SQL/Framework). 2008; https://www.iso.org/standard/45498.html. Accessed 1/1, 2019.
28. Codd EF. *Providing OLAP to User Analysts: An IT Mandate.* New York: Codd & Associates; 1993.
29. Inmon. WH. *Building the Data Warehouse.* New York: Wiley; 2005.

30. Ralph K, Margy R. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* New York: Wiley; 2013.
31. Yusof SM, Sidi F, Ibrahim H, Affendey LS. A study of multidimensional modeling approaches for data warehouse. AIP Conference Proceedings; 2016.
32. Oscar R, Alberto A. A survey of multidimentional modeing. *International Journal of Data Warehousing and Mining.* 2009;5(2):1-23.
33. Chuck B, Daniel F, Amit G, Carlos M, Stanislav V. Dimensional Modeling: In a Business Intelligence Environment. 2006; https://www.redbooks.ibm.com/redbooks/pdfs/sg247138.pdf. Accessed 1/1, 2019.
34. Gavin P. *Oracle Data Warehouse Tuning for 10g.* New York: Digital Press; 2005.
35. Han J, Pei J, Kamber M. *Data Mining, third Edition.* Elsevier; 2011.
36. Tebourski W, Kara. BA, Wahiba., Ben Ghezela H. A survey on medical datawarehouse. International Conference on Control, Decision and Information Technologies (CoDIT); 2013.
37. Wisniewski MF, Kieszkowski P, Zagorski BM, Trick WE, Sommers M, Weinstein RA. Development of a clinical data warehouse for hospital infection control. *J Am Med Inform Assoc.* 2003;10(5):454-462.
38. Gorina Y, Kramarow EA. Identifying chronic conditions in Medicare claims data: evaluating the Chronic Condition Data Warehouse algorithm. *Health Serv Res.* 2011;46(5):1610-1627.
39. Chan C-L, Van PD, Yang N-P. Building a Decision Support Tool for Taiwan's National Health Insurance Data – An Application to Fractures Study. In: *Intelligent Decision Technologies.*2012:407-417.
40. TAKEDA. T, MANABE. S, MATSUMURA Y. The Current Situation and Issues of the Secondary Use of Electronic Medical Record Data. *Japanese Society for Medical and Biological Engineering.* 2017;55(4):151-158.
41. Alain V, Anita B, Catherine Q. *Medical Informatics, e-Health: Fundamentals and Applications* Springer; 2013.
42. Institute J. Japan Medical Association Research Institute working paper No.421 2018; http://www.jmari.med.or.jp/download/WP421.pdf. Accessed 10/05, 2019.
43. Shosuke O, Michi S, Miho A, Genta K, Tomohiro K. Report on Onsite Research Center (Kyoto) operation and activities. 2018; https://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000211812.pdf. Accessed 2019/1/1, 2018.
44. Minjoe. S. CDISC ADaM Application: Does All One-Record-per-Subject Data Belong in ADSL? . PharmaSUG 2012 Conference Proceedings; 2012; San Francisco.

45. Yu e, Lewin Group SF, CA. Summarizing to the one-record-per-person using PROC SUMMARY: when to use CLASSand when to use BY. 2001; https://support.sas.com/resources/papers/proceedings/proceedings/sugi26/p083-26.pdf. Accessed 07/27, 2019.
46. Matsui H, Sato D, Ohe K. Current status and future prospects of on-site research centers for the Japanese National Insurance Claims Database. 37th(suppl.)Japan journal of Medical Informatics; 2017; Osaka.
47. Okamoto K, Mori Y, Kato G, Kuroda T, Muto M. Reconstruction of Japanese Receipt Data to Investigate Actual Treatments for Stomach Cancer. 35th(suppl.) Japan journal of Medical Informatics; 2015; Okinawa.
48. Tomohide I, Genta K, Isao I, Eiji A, Tomohiro K. A survey of clarithromycin monotherapy and side effect of ethambutol for patients with MAC Lung Disease in Japan: a retrospective cohort study using the database of health insurance claims. *Drug Safety and Pharmacoepidemiology.* 2019.
49. Shinya M, Kenji F. The Claims Database in Japan. *Asian Pacific Journal of Disease Management.* 2014;6:55-59.
50. HealthInsuranceClaimsReviewAndReimbursementServices. Billing status for each claim form（The 1st year of Reiwa era）. https://www.ssk.or.jp/tokeijoho/tokeijoho_rezept/tokeijoho_rezept_r01.html. Accessed 10/29, 2019.
51. Website on the provision of healthcare claims data and data from specific health examinations. http://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryouhoken/reseputo/. Accessed 07/01, 2019.
52. MHLW. Outline of new database analysis system of Health Insurance Claims and Specific Health Checkups of Japan (new NDB system). 2017; https://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000072551.pdf. Accessed 2019/1/1, 2018.
53. Summary of deliverables provided by third parties. 2019; http://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000155475.pdf. Accessed 07/01, 2019.
54. Genta K. History of the Secondary Use of National Database of Health Insurance Claims andSpecific Health Checkups of Japan(NDB). *Transactions of Japanese Society for Medical and Biological Engineering.* 2017;55(4):143-150.
55. Mitsutaka T. An Example of Utilization of the National Database from an Academic Standpoint. *Jpn J Pharmacoepidemiol.* 2013;17(2):155-162.

56. Yasuyuki O, Hiroto I. Experience with sampling data sets. 2013; https://www.mhlw.go.jp/stf/shingi/2r9852000002ss9z-att/2r9852000002sse8.pdf. Accessed 1/1, 2019.
57. Sedgwick P. Unit of Analysis. *Br Med J.* 2013;346.
58. Descartes R. *Discours de la méthode.* Tokyo: IWANAMI SHOTEN; 1997.
59. Codd EF. *Relational Completeness of Data Base Sublanguages.* IBM Corporation; 1972.
60. PyDataDevelopmentTeam. pandas. 2019; https://pandas.pydata.org/. Accessed 1/1, 2019.
61. Tomohide I, Genta K, Shigeru O, Eiji K, Takeo N, Tomohiro K. An Optimum Data Warehouse for Epidemiological Analysis using the National Database of Health Insurance Claims of Japan. *European Journal of Biomedical Informatics.* 2019;15(3):31-42.
62. Date C, Darwen H. Tutorial D. 2005; https://www.dcs.warwick.ac.uk/~hugh/TTM/CHAP05.pdf. Accessed 10/08, 2019.
63. Arie J, Ryan S, Ronald P, Robert G, Alex K. *SQL Functions Programmer's Reference.* New York: John Wiley & Sons; 2005.
64. MEDIS standard master. 2019; https://www.medis.or.jp/4_hyojyun/medis-master/. Accessed 07/01, 2019.
65. Sato M, Kondoh E, Iwao T, et al. Nationwide survey of severe postpartum hemorrhage in Japan: an exploratory study using the national database of health insurance claims. *J Matern Fetal Neonatal Med.* 2018:1-6.
66. Nakatani. Y, cho. K, Otsuru. S, Iwao. T, shosuke. O, Kato. G. Survey of actual conditions of medical treatment for patients after cardiopulmonary resuscitation using the NDB. Journal of Japanese Association for Acute Medicine; 2017.
67. Tomohide I, Ken Y, Tomohiro K. Survey of Drug Prescription Patterns Among Patients with Nontuberculous Mycobacterial Disease Using the Database of Health Insurance Claims. *Jpn J Pharmacoepidemiol.* 2018;23(2):89-94.
68. Iwao. T, Kato. G, Ito. I, Kuroda T. Treatment of Mycobacterium avium-intracellulare complex lung disease in the real world: a retrospective big data analysis. *Drugs & Therapy Perspectives,.* 2019.
69. Various information of medical fee. 2019; http://www.iryohoken.go.jp/shinryohoshu/. Accessed 10/24, 2019.
70. Hude Q, Vijaya S, Patricia H. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care.* 2005;43:1130-1139.
71. Strom. B, Kimmel. S, Sean H. *Textbook of Pharmacoepidemiology.* New York: Wiley-Blackwell; 2013.

72. Shiro T, Seto. K, Kawakami K. Pharmacoepidemiology in Japan: medical databases and research achievements. *Journal of Pharmaceutical Health Care and Sciences.* 2015;1:1-16.
73. Tomomi K, Daisuke K, Takao O. Large, Automated Administrative and Clinical Databases Available for Pharmacoepidemiology Studies in Japan. *Japanese Society for Pharmacoepidemiology.* 2013;17(2):135-144.
74. Yon Ju R, Won-Jung K, Daley C. Diagnosis and Treatment of Nontuberculous Mycobacterial Lung Disease: Clinicians' Perspectives. *Tuberculosis and Respiratory Diseases.* 2016;79(2):74-78.
75. Namkoong H, Kurashima A, Morimoto K. Epidemiology of Pulmonary Nontuberculous Mycobacterial Disease, Japan. *Emerg Infect Dis.* 2016;22(6):1116-1117.
76. Adjemian J, Prevots D, Gallagher J. Lack of adherence to evidence-based treatment guidelines for nontuberculous mycobacterial lung disease. *ATS Journal.* 2014;11(1):9-16.
77. Jakko van I, Dirk W, Jack G. Poor adherence to management guidelines in nontuberculous mycobacterial pulmonary diseases. *Eur Respir J.* 2017;49:e1601855.
78. Katsuhiro S, Junichi K, Kazunori T. Guidelines for Chemotherapy of Pulmonary Nontuberculous Mycobacterial Diseas 2012 Revised Version. *Kekkaku.* 2013;88(1):29-32.
79. Kwon Y, Koh W. Diagnosis and Treatment of Nontuberculous Mycobacterial Lung Disease. *J Korean Med Sci.* 2016;31(5):649-659.
80. Griffith D, Aksamit T, Barbara A. An Official ATS/IDSA Statement: Diagnosis, Treatment and Prevention of Nontuberculous Mycobacterial Diseases. *Am J Respir Crit Care Med.* 2007;175(4):367-416.
81. David S. *Tuberculosis and Nontuberculous Mycobacterial Infections.* New York: ASM Press; 2017.
82. BritishThoracicSocietyNTMGuidelineDevelopmentGroup. British Thoracic Society Guidelines for the Managemen of NONTUBERCUROUS MYCOBACTERIAL PULMONARY DISEASE (NTM-PD) 2017. *Thorax.* 2017;72.
83. Kobashi Y, Matsushima T. The microbiological and clinical effects of combined therapy according to guidelines on the treatment of pulmonary Mycobacterium avium complex disease in Japan-including a follow-up study. *Respir Physiol.* 2007;74(4):394-400.
84. Owusu-Edusei K, Marks S, Miramontes R. Tuberculosis hospitalization expenditures per patient from private health insurance claims data. *Int J Tuberc Lung Dis.* 2017;21(4):398-404.
85. Committee J. A review of the standards of Tuberculosis medical care—2014 Revised. *Kekkaku.* 2014;89(7):683-690.

86. Katsuhiro K, Toshiaki T. Clinical features and treatment history of clarithromycin resistance in M. avium-intracellulare complex pulmonary disease patients. *Annals of the Japanese Respiratory Society.* 2007;45(8):587-592.
87. Meier A, Kirschner P, Springer B. Identification of mutations in 23S rRNA gene of Clarithromycin-resistant Mycobacterium intracellulare. *Antimicrob Agents Chemother.* 1994;38(2):381-384.
88. Hirokazu Y, Tomotada I, Yukiko N. Population structure and local adaptation of MAC lung disease agent Mycobacterium avium subsp. Hominissuis. *Genome Biol Evol.* 2017;9(9):2403-2417.
89. Griffith D, Brown-Elliott B, Langsjoen B, et al. Clinical and Molecular Analysis of Macrolide Resistance in Mycobacterium avium Complex Lung Disease. *Am J Respir Crit Care Med.* 2006;174:928-934.
90. Yoshihiro K, Toshiharu M, Mikio O. A double-blind randomized study of aminoglycoside infusion with combined therapy for pulmonary Mycobacterium avium complex disease. *Respir Med.* 2007;101:130-138.
91. Kozo M, Kazuro I, Kazuhiro U. A Steady Increase in Nontuberculous Mycobacteriosis Mortality and Estimated Prevalence in Japan. *ATS Journal.* 2012;11(1):487-490.
92. Koul PA. Ocular toxicity with ethambutol therapy: Timely recaution. *Lung India.* 2015;32(1):1-3.
93. Takanori C, Ruriko S, Yusuke K. A case of Leber's hereditary optic neuropathy with blindness triggered by ethambutol for treatment of tuberculosis. *The Journal of the Japanese Respiratory Society.* 2012;1(1):42-45.
94. Khotimah PH, Sugiyama Y, Yoshikawa M, et al. Medication Episode Construction Framework for Retrospective Database Analyses of Patients With Chronic Diseases. *IEEE J Biomed Health Inform.* 2018;22(6):1949-1959.
95. Enck RE. The OODA Loop. *Home Health Care Management & Practice.* 2012;24(3):123-124.
96. Shinichiro. K, Tatsuya. N, Tomoya. M, Tsuneyuki H. The need and key points for patient matching in clinical studies using the National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB). *Japanese Journal of Health & Research.* 2017;38:11-19.
97. Tatsuya N, Shinichiro K, Tomoya M, et al. Improvement and verification of patient matching method in clinical studies using the National Database of Health Insurance Claims

and Specific Health Checkups of Japan (NDB). *Journal of health and welfare statistics.* 2017;64(12):7-13.

98. Yasuyuki O, Nobuo S, Sayuri S, Hiroki M. Toward the academic use of the national database: the pitfalls of the insurance claims data. *Monthly IHEP.* 2017;268:16-25.

99. Fukuda H, Sato D, Shiroiwa T, Fukuda T. The Development of Dataset Tables for NDB Analyses. *Journal of the National Institute of Public Health.* 2017;68(2):158-167.

100. Özsu. MT, Valduriez P. *Principles Of Distributed Database Systems third Edition.* Springer; 2011.

List of Achievements

Achievements related to the contents of doctoral dissertation

Peer-reviewed Articles

・Tomohide Iwao, Genta Kato, Shigeru Otsuru, Eiji Kondo, Takeo Nakayama, Tomohiro Kuroda: An Optimum Data Warehouse for Epidemiological Analysis using the National Database of Health Insurance Claims of Japan, European Journal of Biomedical Informatics, 15(3), pp.31-42, 2019. (Chapter 1, 2, 3, 4, 5)

・Tomohide Iwao, Genta Kato, Isao Ito, Eiji Aramaki, Tomohiro Kuroda: A survey of clarithromycin monotherapy and side effect of ethambutol for patients with MAC Lung Disease in Japan: a retrospective cohort study using the database of health insurance claims, Drug Safety and Pharmacoepidemiology, https://doi.org/10.1002/pds.4951, 2019. (Chapter 4)

・Tomohide Iwao, Genta Kato, Isao Ito, Tomohiro Kuroda: Treatment of Mycobacterium avium-intracellulare complex lung disease in the real world: a retrospective big data analysis", Drugs & Therapy Perspectives, 36, pp.75-82, 2019. (Chapter 4)

・岩尾友秀，矢野憲，黒田知宏．レセプトデータに基づく非結核性抗酸菌症患者への薬剤処方事例に関する調査，薬剤疫学. 23(2). pp.89-94, 2018. (Chapter 4)

・Sato Mai, Kondoh Eiji, Iwao Tomohide, Hiragi Shusuke, Okamoto Kazuya, Tamura Hiroshi, Mogami, Haruta, Chigusa Yoshitsugu, Kuroda Tomohiro, Mandai, Masaki, Konishi Ikuo, Kato Genta：Nationwide survey of severe postpartum haemorrhage in Japan: an exploratory study using the national database of health insurance claims. The Journal of Maternal-Fetal & Neonatal Medicine, DOI: 10.1080/14767058.2018.1465921. 2018. (Chapter 4)

Domestic Conferences

・中谷友香, 趙晃済, 大鶴繁, 岩尾友秀, 加藤源太: NDB データを用いた心肺蘇生後患者に対する診療の実態調査, 第 45 回日本救急医学会総会・学術集会,大阪, Oct.24.2017（査読あり）(Chapter 4)

・Tomohide Iwao, Shosuke Ohtera, Michi Sakai, Shusuke Hiragi, Shigeru Ohtsuru, Eiji Kondo, Hiroshi Tamura, Genta Kato, Tomohiro Kuroda: Research on the reconstruction method of health insurance claims database suitable for secondary use for epidemiological analysis, Proceedings of Symposium on Medical and Biological Engineering 2017, 23-23, Sep.15.2017（査読あり）(Chapter 3)

・岩尾友秀, 平木秀輔, 大寺祥佑, 酒井未知, 加藤源太, 田村寛, 黒田知宏: レセプト情報・特定健診等情報データベース（NDB）を対象とした疫学研究に適した分析用データベースの構築, IT ヘルスケア学会第 11 回学術大会,名古屋, May.28.2017（査読あり）(Chapter 3)

Grant

・平成 29 年度 文部科研費若手研究(B), 疫学分析に適したデータモデル構築に関する研究 研究代表者：岩尾友秀

Achievements not directly related to the contents of doctoral dissertation

Peer-reviewed Articles

・Michi Sakai, Shosuke Ohtera, Tomohide Iwao, Yukiko Neff, Genta Kato, Yoshimitsu Takahashi, Takeo Nakayama. Validation of Claims Data to Identify Death among Aged Persons Utilizing Enrollment Data from Health Insurance Unions. Environmental Health and Preventive Medicine. 24(1), Article No:63. https://doi.org/10.1186/s12199-019-0819-3. 2019.

・荒牧英治，若宮翔子，岩尾友秀，川上庶子，中江睦美，松本妙子，友廣公子，栗山猛. 小児頻用医薬品に関する医薬品添付文書における記載状況の調査，医療情報学, 38(6). pp.337-348. 2019.

・荒牧英治, 岩尾友秀, 若宮翔子, 伊藤薫, 矢野憲，大江和彦. 症例検索システムの試行運用に基づいた利用状況に関する基礎的検討，医療情報学. 38(4). pp.245-248. 2018.

・Misa Usui, Eiji Aramaki, Tomohide Iwao, Shoko Wakamiya, Tohru Sakamoto, Mayumi Mochizuki. Extracting and Standardizing Patient Complaints from Electronic Medication Histories in Japanese. JMIR Medical Informatics. 6(3) e11021.DOI: 10.2196/11021. 2018.

・西田栄美，神宮司希和子，古島大資，山本尚子，岩尾友秀，山田知美. EDC システムに依存しない動的データチェックが可能なソフトウェアの開発. 医療情報学. 38(2). pp.103-114, 2018.

International Conferences

・Tomohide Iwao: A novel software for efficient construction of datasets using health insurance claims database. ISPE’s 12th Asian Conference on Pharmacoepidemiology. Oct 12, 2019. Kyoto, Japan. (Poster, 査読あり)

・Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, Eiji Aramaki: J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage. 11th International Conference on Language Resources and Evaluation. May 8, 2018. Miyazaki, Japan. (Full paper, 査読あり)

・Tomohiro Kuroda, Hiroki Shiomi, Eri Minamino-Muta, Yugo Yamashita, Tomohide Iwao, Hiroshi Tamura, Kazuo Ueshima, Takeshi Kimura: Evaluation of NISHIJIN e-Textile for 12-lead ECG measurement through Automatic ECG Analyzer. Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp.1234-1237 (2017/07/12) Jeju/Korea(Full paper, 査読あり).

・T Sengoku, T Ishizaki, T Iwao, S Ohtera, M Sakai, G Kato, T Nakayama, Y Takahashi. The lifestyle-related diseases among Japanese public assistance recipient. European Journal of Public Health, Volume 28, Issue suppl_4, 1 November 2018, cky214.050, https://doi.org/10.1093/eurpub/cky214.050 (Poster, 査読あり)

Domestic Conferences

・黒田知宏, 塩見紘樹, 上島一夫, 岩尾友秀, 田村寛, 木村剛. 西陣織 12 誘導心電布テクノセンサーER の評価 (第一報)　ICU と CCU, 43(4):222.2018.

・矢野憲, 岩尾友秀, 荒牧英治: MedInput: 病名の自動予測補完による医療テキスト入力支援ツールの構築. 言語処理学会 第 24 回年次大会, 岡山.2018.

・酒井未知, 大寺祥佑, 岩尾友秀, 岡本和也, 加藤源太, 中山健夫, 黒田知宏:レセプト情報等オンサイトリサーチセンター（京都）の試行的利用に基づく今後の活用可能性に関する検証, 第 36 回医療情報学連合大会論文集, pp.142-145, Nov.23.2016.

・佐藤麻衣, 近藤英治, 岩尾友秀, 平木秀輔, 岡本和也, 川崎薫, 黒田知宏, 小西郁生, 加藤源太: レセプト情報・特定健診等情報データベースを用いた弛緩出血についての実態調査, 日本周産期・新生児医学会雑誌 52(2): p724, July. 17. 2016 （査読あり）

・今中健, 岡本和也, 疋田智子, 岩尾友秀, 浦西友樹, 田村寛, 齋藤永, 加藤源太, 黒田知宏:勤務表から抽出した制約条件を用いたナース・スケジューリングシステム, 第 60 回システム制御情報学会研究発表講演会, May. 27. 2016 （査読あり）

・福士雄太, 岡本和也, 岩尾友秀, 浦西友樹, 田村寛, 齊藤永, 加藤源太, 黒田知宏: 外来病棟における位置情報とオーダ情報を用いた患者待ち時間の分析, システム制御情報学会研究発表講演会予稿集, May. 25. 2016 （査読あり）

・江指未紗, 中野友裕, 岩尾友秀, 浦西友樹, 岡本和也, 加藤源太, 齊藤永,田村寛, 野間春生, 黒田知宏: 医療機器と病院情報システムを接続する試み, 生体医工学, Vol.54, ,p.140, Suppl.1,201 第 55 回日本生体医工学会大会 プログラム・抄録集 （査読あり）

・神原誠之, 岩尾友秀, 横矢直和: 拡張現実感のための動的シャドウマップを用いたシャドウレンダリング手法, 画像の認識・理解シンポジウム(MIRU2005)講演論文集, pp. 297-304, July 2005.

・岩尾友秀, 神原誠之, 横矢直和: 視点とオブジェクトの位置関係を考慮したシャドウマップの動的生成法, 電子情報通信学会 技術研究報告, ITS2004-80, pp. 83-88, Feb. 2005.

・岩尾友秀, 神原誠之, 横矢直和: 視点と光源の位置関係を考慮したシャドウマップの動的生成手法, 情報処理学会関西支部大会講演論文集, pp. 45-48, Oct. 2004.

・目加田慶人, 田村徹也, 保田和隆, 森田友幸, 柿崎貴也, 越智健治, 岩尾友秀: 2003 年 PRMU アルゴリズムコンテスト「そこにいるのは何人?」実施報告とその入賞アルゴリズムの紹介.電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 103(516), pp.49-58.

Award

・賞名:平成 16 年度情報処理学会関西支部学生奨励賞
受賞対象論文 岩尾友秀, 神原誠之, 横矢直和
"視点と光源の位置関係を考慮したシャドウマップの動的生成手法",
情報処理学会関西支部大会講演論文集, 2004 年 10 月

・賞名: 電子情報通信学会 PRMU 研究専門委員会主催第 7 回アルゴリズムコンテスト入賞. 2003 年 9 月

Patent

・録画装置：池之上博美，渡部裕之，木村信浩，上田俊史，足立康伸，岩尾友秀，
特許 5546835（2014 年）

・デジタルレコーダー：池之上博美，渡部裕之，木村信浩，上田俊史，岩尾友秀，
特許 5542491 (2014 年）

・多チャンネル画像転送システム：太田知英，池之上博美，岩尾友秀，
特許 4679444 (2011 年）

・複数チャンネル画像転送装置：太田知英，池之上博美，岩尾友秀，
特許 4828354 (2011 年）