An Optimum Data Warehouse for Epidemiological Analysis using the National Database of Health Insurance Claims of Japan

Tomohide Iwao[1], Genta Kato[2]*, Shigeru Ohtsuru[3], Eiji Kondoh[4], Takeo Nakayama[5], Tomohiro Kuroda[1]

[1]Division of Medical Information Technology and Administration Planning, Kyoto University Hospital, Japan

[2]Solutions Center for Health Insurance Claims, Kyoto University Hospital, Japan

[3]Department of Primary Care and Emergency Medicine, Kyoto University Hospital, Japan

[4]Department of Gynecology and Obstetrics, Kyoto University Hospital, Japan

[5]Department of Health Informatics, Graduate School of Medicine and Public Health, Kyoto University, Japan

Abstract

Background: While administrative databases for health care are increasingly used as research tools, such databases generally contain only health insurance claims data, the contents of which are insufficient for conducting epidemiological research. Creating a dataset appropriate for specific analysis requires technical expertise and familiarity with data analysis. The aim of our research is to develop a data warehouse (DW) accessible to researchers of epidemiology without this expertise.

Methods: We began by adding commonly used attributes in the epidemiological field to the National Database of Health Insurance Claims of Japan (NDB), to construct a Research Question Oriented DB. Secondly, we developed a versatile analysis unit schema by which the Research Question Oriented DW was reconstructed as per-patient units, covering demographics including sex, age group etc. We then proposed a pattern relational calculus by which research-specific attributes can be added without expert knowledge of SQL. Finally, we applied the DW in two epidemiological studies.

Results: In both studies, the coverage of attributes constructed only by the versatile analysis unit schema was limited. The versatile analysis unit schema covered 12% (3/25) of the attributes used for the one study as well as 15% (3/20) in the other study. On the other hand, the pattern relational calculus we proposed covered all remaining attributes which researchers used for their study.

Conclusion: As the versatile analysis unit schema and the pattern relational calculus were able to cover all attributes used in the two epidemiological studies, this shows that even within a limited scope, our method allows researchers who have little knowledge of SQL to tackle respective epidemiological study.

Abbreviations and Terminologies: NDB-SD: NDB Sampling Data set; DW: Data Warehouse; Shema: 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; Tuple relational calculus: logical expression used in the relational model theory; SQL: database language based on the relational model theory

Keywords

Data warehouse; Epidemiology; Health insurance claims database; Relational model; Big data analysis

Correspondence to:

Genta Kato, MD, PhD
Solutions Center for Health Insurance Claims,
Kyoto University Hospital,
54 Shogoin Kawahara-cho, Sakyo-ku,
Kyoto 606-8507, Japan,
E-mail: qq9f8hn9@kuhp.kyoto-u.ac.jp

EJBI 2019; 15(3):31-42
Received: May 08, 2019
Accepted: September 23, 2019
Published: September 30, 2019

1. Introduction

The significance of the secondary use of health insurance claims data is beginning to be recognized in Japan. There is a flourishing new body of epidemiological studies that use these databases as source material.

However, the secondary use of these data is not maximized due to two main challenges: the data quality and the data structure of the databases.

Considering the former, since the primary purpose of these databases [1] 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, which is time consuming and inefficient.

As for data structure, the majority of the databases of health insurance claims currently in use are relational databases (RDB); these are designed to support the relational model [2], 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 [3].

A well-known approach for supporting analyses that involve relational databases is using data warehouses (DW) [4]. 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. A typical example of DW is the dimensional model [5], which is a data model proposed in a book written by Ralph Kimball in 1996. A dimensional model is a data structure that is particularly suitable for exploratory analyses. Thus, particularly 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) [6], which requires good response times.

Approaches based on DW are being adapted for epidemiological analyses using health insurance claims databases in similar ways [7]. In the United States, for example, the Chronic Condition Data Warehouse (CCW) [8] has been built for the purpose of supporting studies related to patients with chronic diseases by organizations that support research using 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 [9]. This tool allows for analyses to be performed more easily when conducting typical epidemiological researches.

Despite these advances, because the attributes (dataset) 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. As a result of this situation, 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 [10]. 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 research [11].

Generally, a per-patient data structure is considered best for epidemiological study [12], because in the field of epidemiological research, 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. In Japan, there are several studies in which researchers attempt to reconstruct the Japanese health care database into a per-patient structure, which is more suitable for statistical analysis [13,14].

However, because each epidemiological study requires a different dataset, it is difficult to prepare such data in advance in DW. For this reason, each researcher is resigned to create a dataset according to their study. As a result, it is still difficult for researchers who have little experience handling data to conduct their study by themselves.

1.1 Objectives

The first aim of our research is to solve the above-mentioned problems regarding data quality by identifying data items that are frequently used in epidemiological analysis. The second aim of our research is to reconstruct per-patient structure by integrating data which are originally dispersed in multiple relations for researchers who have little skills in data handling.

To accomplish our aim, we developed a DW using the national health insurance claims database of Japan that researchers can use without expert handling skills for epidemiological purposes aiming to solve the two aims mentioned above simultaneously.

2. Materials and Methods

2.1 Materials

The NDB (National Database of Health Insurance Claims and Specific Health Checkups of Japan) [15] was created in FY2009, and is operated by the Japanese Ministry of Health, Labor, and Welfare (MHLW). To date, it has accumulated over 10 billion entries of electronic data concerning health insurance claims and specific health checkups. Secondary use of the NDB is possible but, as mentioned in the previous section, there are several problems when using it for epidemiological study.

Our 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". Figures 1-3 show the NDB-SD relations used in our research. In the Japanese system, such claims are issued at a frequency of one per month. In Figure 1, 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, we converted csv files of NDB-SD into the database format by using Database Management System. Similarly, Figure 2 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

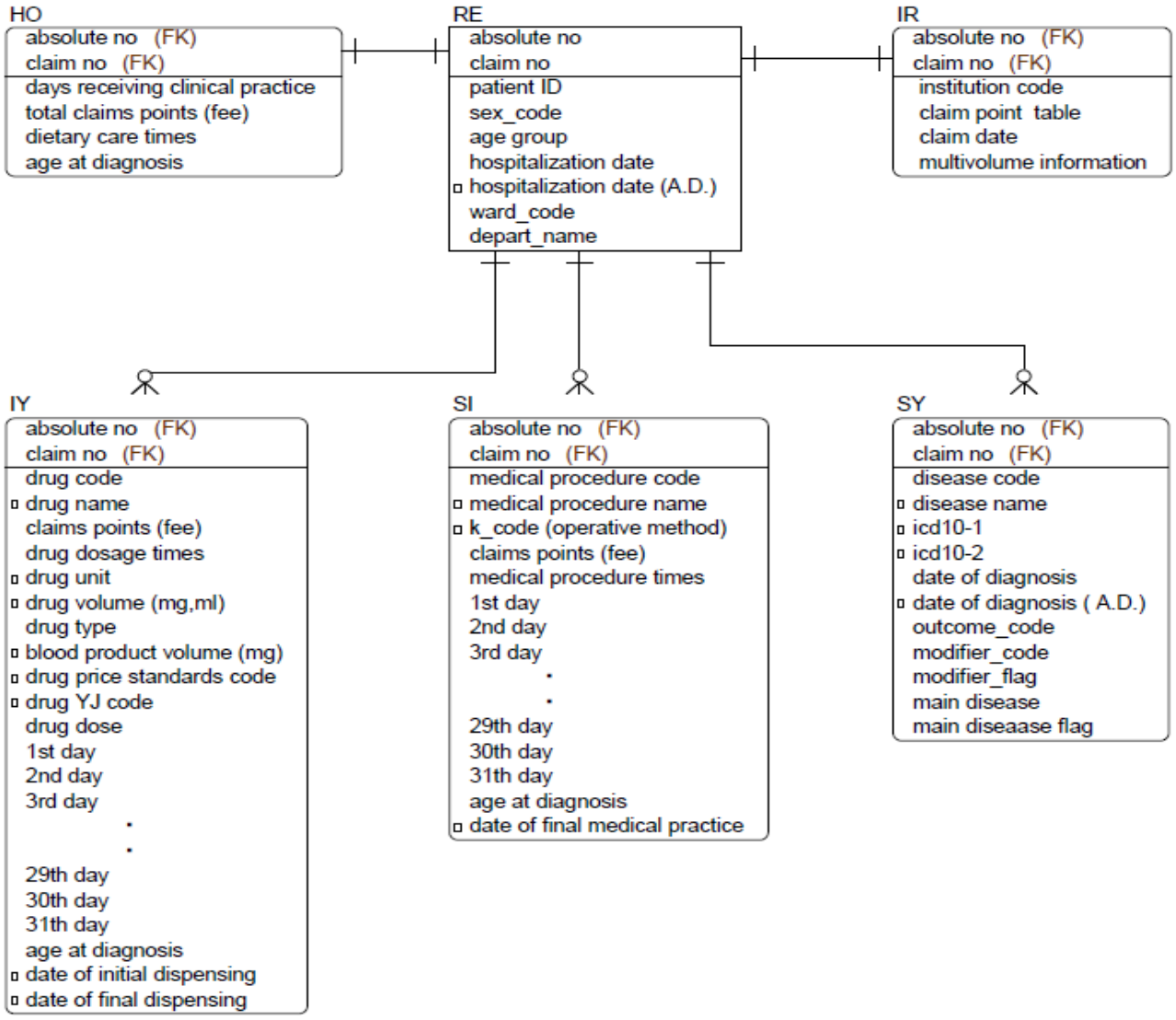


Figure 1: Example of the typical NDB-SD relations and attributes (Medical claims).

name and medical procedure and defines the hospitalization cost per day for each classification. DPC is stored in BU, disease information within 3 months after admission is stored in SB, and disease information after that is stored in SY. In addition, the CD stores the medical procedures carried out within 3 months after admission and information on medication etc. And, Figure 3 is data on Pharmacy claims. Information on dispensing is stored separately in SH and CZ. Details of relation and main attribute names are listed in Appendix A1.

2.2 Methods

2.2.1 Research question splicing: One of the reasons why it is difficult to create a data set is that the attributes required for each research question vary from study to study. Therefore, we considered simplifying the problem to be solved by dividing the research question. In this research, we named this idea "research question splicing" and the concept is shown in the Figure 4. 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 Figure 4, 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 shown in Figure 4 can be made simple and patterned, anyone can easily create a data set for each analysis unit. The following "the pattern tuple relational calculus" is a method for making the process of „calculation" simple and patterned.

2.2.2 Pattern tuple relational calculus: Figure 5 shows the principle of the pattern tuple relational calculus we proposed. 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 manuscript. A PR stores the analysis unit and is a mandatory relation when adding all attributes. 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.).

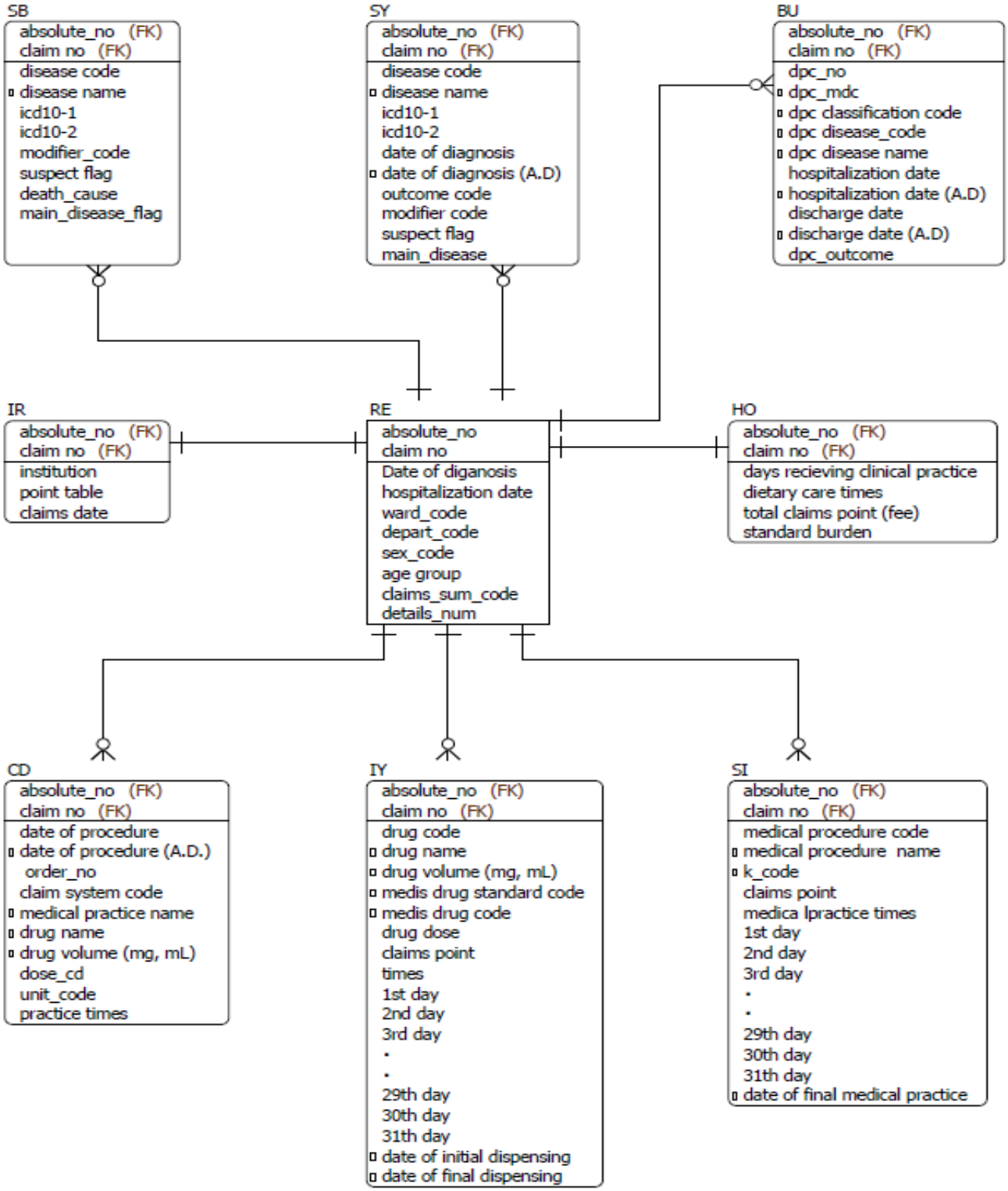


Figure 2: Example of the typical NDB-SD relations and attributes (DPC claims).

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 Figure 5. Since tuple relational calculus can be expressed in SQL, as a result, they can be realized with simple and patterned code that can be implemented even by those who have little knowledge of SQL. For example, if M1 is an attribute to be added, it can be added using sub attributes from S1 to SN as conditions. Note that there may be more than one sub attribute or none at all, contingent on research questions.

Example of SQL based on pattern tuple relational calculus

(1) Relational complete

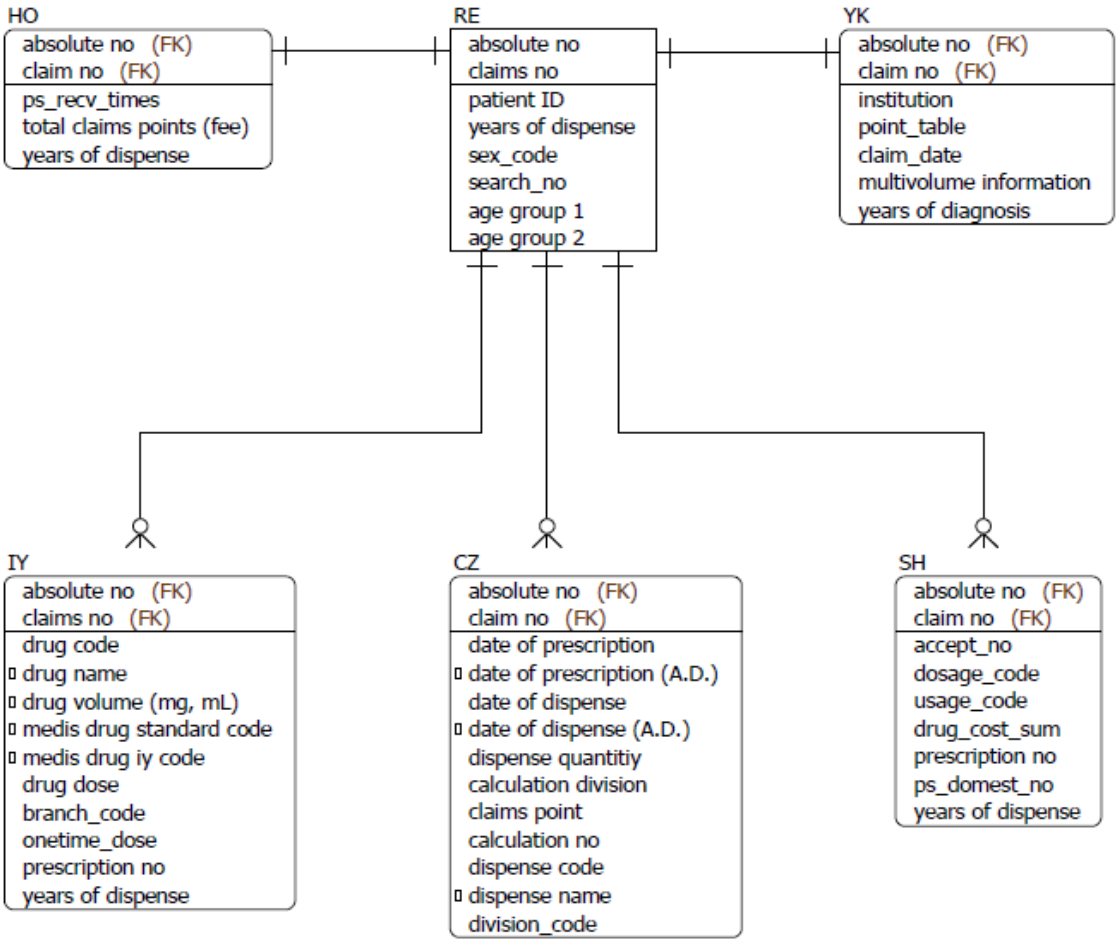
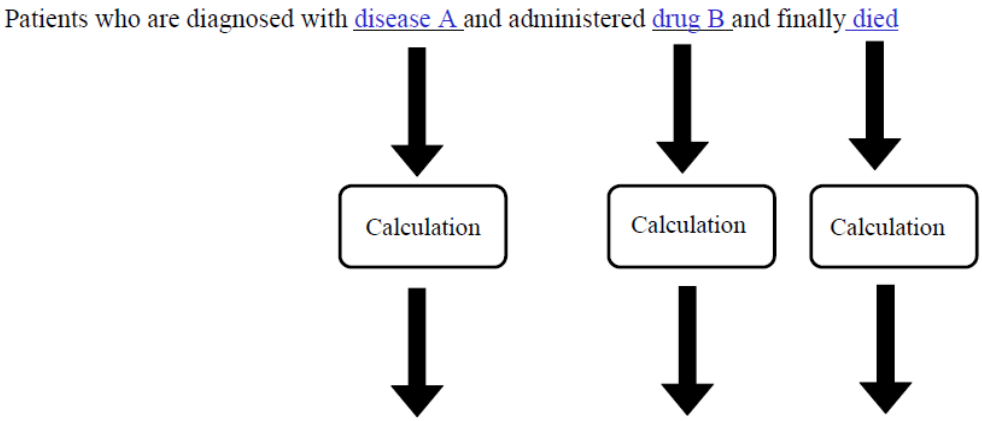


Figure 3: Example of the typical NDB-SD relations and attributes (Pharmacy claims).



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

Figure 4: Research question splicing.

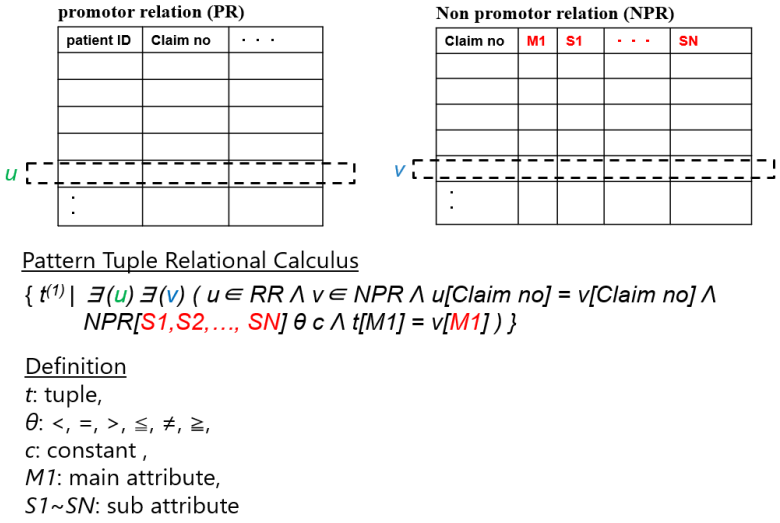


Figure 5: The principle of the pattern tuple relational calculus.

In the relational model theory, the SQL is guaranteed to be relational complete. For that reason, a tuple relational calculus theoretically can be implemented in SQL. In this manuscript, we explain examples of the SQL based on the pattern tuple relational calculus for readability.

(2) Example of SQL for complementing an attribute

Here we assume that there is a schema consisting only of patient ID. Let us define this "per-patient unit schema".

Figure 6 shows the example of SQL to complement an attribute to the per-patient unit schema. In our method, attributes other than the attributes in the per-patient unit schema, such as A1 and A2 in Figure 6 (e.g., disease name, drug name), are complemented and extended in addition to the per-patient unit schema area. Figure 6 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 Figure 6), and in the case of patients who do not have the disease name, nothing is stored (shown as "False" in Figure 6). Basically, researchers can complement new attributes only by editing the underlined section from ① to ⑤ in Figure 6.

Underlined section ① and ② in Figure 6 showed an attribute name to be complemented. Here, the SQL function underlined ③ ("row_number ()" function in Figure 6) 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 complemented. In Table 1, "row_number ()" function is mainly used in complementing 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, "count ()" function is mainly used for "number of administrations of medical procedure", "sum ()" function for "volume of drug" respectively. 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 [16], but only SQL functions shown in Table 1 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, the 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 complement (e.g., SY, SI, IY, etc.).

As an example, if researchers want to complement 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 complement attributes such as drug name or date of prescription for each drug per analysis unit.

2.2.3 Construction of Data Warehouse (DW): Here, we introduce a DW which we developed to make the pattern tuple relational calculus applicable, as well as to complement essential attributes for epidemiological studies. The procedure of developing the Data Warehouse is shown in Figure 7. This DW is developed from the NDB-SD. The construction procedure of the DW can be divided into two Phases. During Phase 1, we constructed a Research Question Oriented Data Warehouse (Research Question Oriented DW) to complement attributes that are commonly used for epidemiological studies. In Phase 2, we designed a versatile analysis unit schema which transforms the Research Question Oriented DW into a more easily tractable format, the structure summarized to one record for each analysis unit. Finally, by extending the schema using "SQL based on pattern tuple relational calculus" as proposed, researchers can create the desired dataset for the respective epidemiological study. The pattern tuple relational calculus is our method used to make a "per-analysis unit" dataset for the individual study by extending, one by one, attributes required in individual studies to a "versatile analysis unit schema."

```
left outer join
(select dsb.patient ID, ① dsb.disease name
 from
 (select re.patient ID, ② sy.disease name, ③ row_number() over (partition by re.patient ID) as rank
  from
 ④ RE re inner join SY sy
  on re.claim no = sy.claim no
 ⑤ where sy.disease code in ( 8834172 )
 ) dsb
 where dsb.rank = 1
) A3
on analysis unit schema.patient ID = A3.patient ID
```

Extend

patient ID	A1	A2	. . .	. . .
1	True	False		
2	False	False		
3	False	True		
4	True	False		
5	False	True		
. .				

Figure 6: An example of SQL based on the pattern tuple relational calculus.

Table 1: SQL functions for summarizing into analysis unit.

SQL function name	**Function**
row_number ()	Assign a sequential integer to each record
count ()	Calculate the number of records
sum ()	Calculate the total value of an attribute

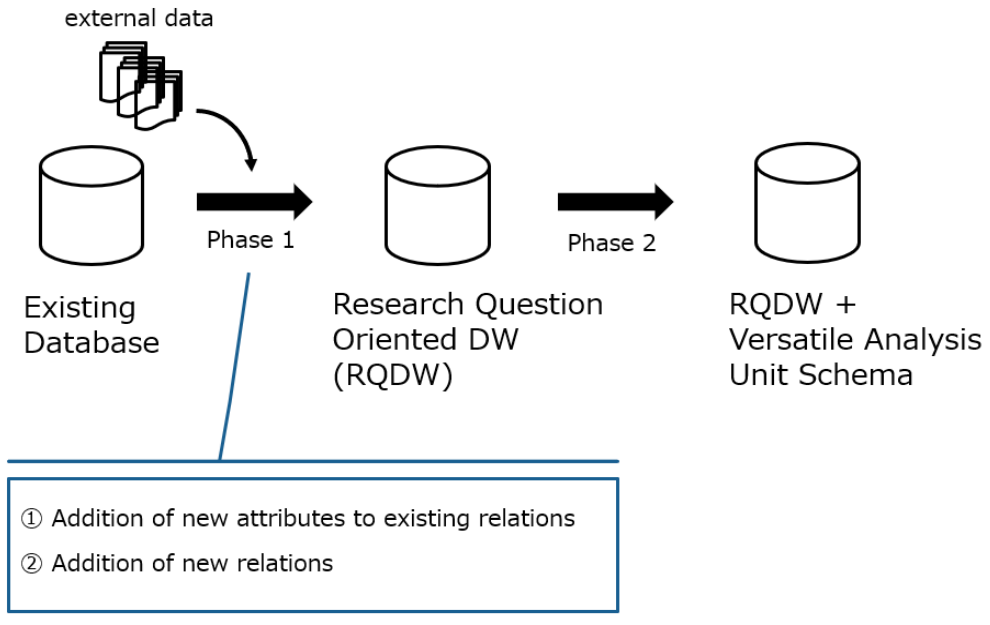


Figure 7: Procedure of developing the data warehouse.

Table 2: Attributes obtained from the NDB-Literature.

Groups	Attribute (variable)	Analysis Unit Schema (patient ID)	Analysis Unit Schema (Claim no)	Research Question Oriented Data Warehouse	Frequency (category)
G1	Age group	○	○	-	17
	Sex	○	○	-	14
G2	Discharge date (A.D. format)	×	○	-	6
	Hospitalization date (A.D. format)	×	○	Hospitalization date (AD 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 (AD 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
G3	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 (AD 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 (AD format)	6
	Date of commencement of diagnosis of a disease specific to the research	×	×	Diagnosis commencement date (AD 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

Table 3: How comprehensively the attributes required for an epidemiological research of PPH, which was previously published in the Journal of Maternal-Fetal & Neonatal Medicine (DOI: 10.1080/14767058.2018.1465921), were covered by the Ext.

Data item used in the paper	Attribute (variable)	Groups	Covered
Sex	Sex	G1	○
Age group	Age group		
Death	Death	G2	○
Uterine atony	Disease name(s) specific to the research	G3	×
Placental abruption	Disease name(s) specific to the research		
Placenta previa	Disease name(s) specific to the research		
Retained placenta	Disease name(s) specific to the research		
Placenta accrete	Disease name(s) specific to the research		
Cervical laceration	Disease name(s) specific to the research		
Vaginal hematoma	Disease name(s) specific to the research		
Multiple pregnancy	Disease name(s) specific to the research		
Low-lying placenta	Disease name(s) specific to the research		
Uterine rupture	Disease name(s) specific to the research		
Placenta previa accrete	Disease name(s) specific to the research		
Uterine inversion	Disease name(s) specific to the research		
Amniotic fluid embolism	Disease name(s) specific to the research		
Volume of RCC (ml)	Volume of drug specific to the research		
Volume of FFP (ml)	Volume of drug specific to the research		
Volume of PC (ml)	Volume of drug specific to the research		
Cesarean delivery	Name of medical procedure(s) specific to the research		
Operative vaginal delivery	Name of medical procedure(s) specific to the research		
Surgical intervention	Name of medical procedure(s) specific to the research		
Intrauterine balloon tamponade	Name of medical procedure(s) specific to the research		
Arterial embolization	Name of medical procedure(s) specific to the research		
Hysterectomy	Name of medical procedure(s) specific to the research		
Note: ○ = covered; × = not covered.			

Table 4: How comprehensively the attributes required for epidemiological research of patients after cardiopulmonary resuscitation were covered by the Research Question Oriented DW.

Data item used in the paper	Attribute (variable)	Groups	Covered
Sex	Sex	G1	○
Age group	Age group		
Death	Death	G2	○
Closed chest cardiac massage	Name of medical procedure(s) specific to the research	G3	×
Adrenalin	Name of medical procedure(s) specific to the research		
Tracheal intubation	Name of medical procedure(s) specific to the research		
Postresuscitation encephalopathy	Name of medical procedure(s) specific to the research		
Artificial respiration	Name of medical procedure(s) specific to the research		
Defibrillation	Name of medical procedure(s) specific to the research		
EEG	Name of medical procedure(s) specific to the research		
CT	Name of medical procedure(s) specific to the research		
Induced Hypothermia	Name of medical procedure(s) specific to the research		
CAG	Name of medical procedure(s) specific to the research		
PCI	Name of medical procedure(s) specific to the research		
Hemodialysis	Name of medical procedure(s) specific to the research		
PCPS	Name of medical procedure(s) specific to the research		
Medical expenses for first aid	Name of medical procedure(s) specific to the research		
Administering of blood products for transfusion	Name of medical procedure(s) specific to the research		
Medical expenses for intensive care	Name of medical procedure(s) specific to the research		
Postresuscitation encephalopathy	Name of medical procedure(s) specific to the research		
Note: ○ = covered; × = not covered.			

2.2.4 Phase 1: Construction of the research question oriented DW: The purpose of construction of the Research Question Oriented DW is to add attributes or relations to an existing database so that the pattern tuple relational calculus can be applied. In this research, we use the published articles regarding NDB instead of research questions.

First, we investigated epidemiological studies that used the NDB [17] as source material. The MHLW introduced 105 published articles as NDB-related publications *via* its website. These publications include conference presentations, public surveys, and academic articles and so on. However, the number of publications in which NDB data analysis was directly performed was only 20. We named these publications as "NDB-Literatures" (Appendix A2) in this manuscript. We carefully reviewed tables, figures, and sentences in the NDB-literatures and picked up attributes which were used one or more times.

Second, we added the attributes obtained from the NDB-literatures to the converted NDB-SD, to construct the research question-oriented Database. In Figures 1-3, attributes marked "o" at the prefix in each table are newly added to the NDB-SD when constructing the Research Question Oriented DW. The process of conversion is shown in Figure 7. At this conversion process, we applied minor changes for existing attributes to make the data handling easier. One example of a minor change is the conversion of date-related attributes. Some date-related attributes, such as dates of admission, diagnosis, etc. are stored in era name format in many relations in the NDB-SD. When constructing the Research Question Oriented DW, 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 [18], for "disease name," "medical procedure," and "drug name," and the one of MEDIS-DC [19] 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 [20] for calculating the Charlson Comorbidity Index (CCI) score of each patient ID/claim no.

3.2.5 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." 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, we transformed the Research Question Oriented DW into the versatile analysis unit schema, the primary key of which are {patient ID} and {claim no} using above-mentioned SQL based on tuple relational calculus. To ensure the versatility of the versatile analysis unit schema, we selected attributes that can be summarized in one record for each analysis unit such as "Death," "Age group," etc. in the NDB-Literatures and implemented them to the schema. To construct, we used SQL [21], 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 construct the DW in this research; furthermore, we used PostgreSQL (ver. 9.5), which is well equipped with online analytical processing functions [22], as a database management system.

The versatile analysis unit schema embodies easily tractable format, but researchers can't finish their own specific studies by using this schema, since it contains only the commonly used attributes.

Finally, two epidemiological studies were conducted using the DW. They used SQL based on the pattern relational calculus. One is a study on postpartum hemorrhage (PPH) in the field of obstetrics and gynecology, in which we used "claim no" as an analysis unit. The other is a study to investigate the treatment situation after cardiopulmonary resuscitation, in which we used "claim no" and "patient ID" as analysis unit

3. Results

The attributes obtained from the NDB-literatures are displayed in Table 2. The "groups" of attributes in Table 2 are classified as follows: "G1" are attributes that are used in most epidemiological studies; "G3" are those which indicate the contents of specific diseases and treatments such as specific drug names; "G2" are those not categorized into G1 or G3. Although some attributes in G2 such as the history of injection, history of drug etc., can be assessed as specific, we did not categorize them into G3 because we want to expand the function of versatile analysis unit schema as possible. "Frequency" refers to the times of appearance (maximum value: 20) in the NDB-literatures. The attributes included in the versatile analysis unit schema are shown with a "o," while those that were not included are indicated with an "x." Furthermore, the column titled "Research Question Oriented DW" in Table 2 shows attributes that were added to the NDB-SD when constructing the Research Question Oriented DW. Finally, we implemented G1 and G2 attributes in the schema.

We applied the versatile analysis unit schema for two epidemiological studies and evaluated the cover ratio of the attributes in the schema. The versatile analysis unit schema covered 12% (3/25) of the attributes for one study (Table 3) and 15% (3/20) for the other study (Table 4). However, after the application of the pattern relational calculus, in both studies the cover ratio improved to 100%. In the former study regarding postpartum hemorrhage (PPH), we published an academic article [23] using this DW. And in the latter study regarding

cardiopulmonary resuscitation shown in Table 4, it was possible for the physician-scientists with little knowledge of SQL to work out their study by themselves.

4. Discussion

This research examined the usefulness of a novel data warehouse for epidemiological analysis, with the combination of the versatile analysis unit schema and the pattern relational calculus, by using the Research Question Oriented DW. The conventional application of SQL for raw data is underutilized by most of the users who are physician-scientists, likely due to their unfamiliarity to SQL.

The versatile analysis unit schema which users could handle without SQL programing skills could only cover a few attributes as shown in our research: While G1 group attributes were able to apply these studies, the G2 group attributes were hardly utilized. This outcome shows that commonly used attributes in the existing related publications are not necessarily applicable to other new studies.

Our novel DW can cover all attributes used in two studies with the combination of pattern relational calculus, which contains only three functions with "row_number," "count," and "sum" is much easier for the users, to the versatile analysis unit. By showing our novel DW can cover all attributes, we can say that the pattern relational calculus still has a generality for analyzing claims data.

Next, let us consider the generality of our method. First, the pattern relational calculus can create all attributes theoretically if you build the Research Question Oriented DW (PR and other one relation) as shown in Figure 4. On the other hand, there are cases where the Research Question Oriented DW needs to complement attributes required for each research question. In the two epidemiological studies conducted, it was not necessary to complement new attributes to the Research Question Oriented DW. This means that the NDB-Literatures covered all the attributes needed for the pattern relational calculus. Therefore, if the Research Question Oriented DW is refined by conducting more epidemiological studies in the future, it is possible to obtain more generality allowing the pattern relational calculus alone to cover most epidemiological studies.

There are limitations to our research. Since the NDB-SD contains claims data for a short period of time (one month), we have determined it is only suitable for cross-sectional studies. Future studies are needed to examine if this method could be applied also for cohort study as well, as we plan to obtain NDB data for a longer period.

5. Conclusion

This research developed the DW including the versatile analysis unit schema having attributes that are frequently used in epidemiological analysis and proposed the pattern relational calculus that allows researchers to complement attributes which they want to use in their epidemiological study. As the versatile analysis unit schema and the pattern relational calculus were able to cover all attributes used in the two epidemiological studies, it shows that even within a limited scope, our method allows researchers who have little knowledge of SQL to tackle an epidemiological study.

6. Disclosure

The authors declare no conflicts of interest with respect to this research and paper.

7. Acknowledgements

We are deeply grateful to Dr. Taro Minami for giving us valuable advices about our research.

8. Funding

This article was performed with the assistance of a research grant from the Ministry of Education, Culture, Sports, Science, and Technology (grant number: 17K1781600).

References

1. 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.
2. Codd EF. A relational model for large shared bank. Communications of the ACM - Special 25th Anniversary Issue; 1970; 26(1): 64-69.
3. Date CJ. Database in Depth: Relational Model for Practitioners. New York: Oreilly & Associates Inc; 2005.
4. Inmon WH. Building the Data Warehouse. New York: Wiley; 2005.
5. Ralph K, Margy R. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling New York: Wiley; 2013.
6. Chuck B, Daniel F, Amit G, Carlos M, Stanislav V. Dimensional Modeling: In a Business Intelligence Environment: IBM International Technical Support Organization; 2006.
7. Tebourski W, Kara BA, Wahiba B, Ghezela H. A survey on medical datawarehouse. International Conference on Control, Decision and Information Technologies (CoDIT); 2013.
8. 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.
9. Chan CL, Van PD, Yang NP. Building a Decision Support Tool for Taiwan's National Health Insurance Data - An Application to Fractures Study. Intelligent Decision Technologies. Smart Innovation, Systems and Technologies; 2012. p. 407-17.

10. https://www.resdac.org/

11. 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.

12. Minjoe S. CDISC ADaM Application: Does All One-Record-per-Subject Data Belong in ADSL?. PharmaSUG 2012 Conference Proceedings; San Francisco; 2012.

13. Matsui H, Sato D, Ohe K. Current status and future prospects of on-site research centers for the Japanese National Insurance Claims Database. 37th Japan journal of Medical Informatics; Osaka; 2017.

14. Okamoto K, Mori Y, Kato G, Kuroda T, Muto M. Reconstruction of Japanese Receipt Data to Investigate Actual Treatments for Stomach Cancer. 35th Japan journal of Medical Informatics; Okinawa; 2015.

15. Shinya m, Kenji F. The Claims Database in Japan. Asian Pacific Journal of Disease Management. 2014; 6: 55-59.

16. Arie J, Ryan S, Ronald P, Robert G, Alex K. SQL Functions Programmer‘s Reference. New York: John Wiley & Sons; 2005.

17. http://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000155475.pdf

18. http://www.iryohoken.go.jp/shinryohoshu/downloadMenu/

19. https://www.medis.or.jp/4_hyojyun/medis-master/

20. 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.

21. https://www.iso.org/standard/45498.html

22. Codd EF. Providing OLAP to User Analysts: An IT Mandate. New York: Codd & Associates; 1993.

23. Sato M, Kondoh E, Iwao T, Hiragi S, Okamoto K, 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.