Application and Evaluation of a Common Data Model to Multicenter Workers' Health Examination Data

Juho Sim

The Graduate School
Yonsei University
Department of Public Health

Application and Evaluation of a Common Data Model to Multicenter Workers' Health Examination Data

A Dissertation Thesis
Submitted to the Department of Public Health
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy of Public Health

Juho Sim

December 2022

This certifies that the dissertation thesis
of Juho Sim is approved.

_____

Thesis Supervisor: Jin-Ha Yoon

_____

Jong-Uk Won: Thesis Committee Member #1

_____

Chi Nyun Kim: Thesis Committee Member #2

_____

Jun-Hee Lee: Thesis Committee Member #3

_____

Inchul Jeong: Thesis Committee Member #4

The Graduate School
Yonsei University
December 2022

Table of Contents

List of Tables

List of Figures

Abstract

Application and Evaluation of a Common Data Model to Multicenter Workers' Health Examination Data

Introduction

Recently, as the amount of data in the pharmaceutical field in Korea has increased, research using real-world data (RWD) called real-world evidence (RWE) research is being conducted. The US is using Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) for research using medical data. In Korea, research is performed in various fields such as medical device, colorectal cancer, and drug research using the OMOP-CDM. Therefore, this study was conducted to determine if Workers' Health examination data could be combined with a CDM and be used for research.

Methods

Data from January 2015 to December 2017 were collected from three university hospitals. A data dictionary was prepared according to the electronic medical record (EMR) status of each institution, and the data were mapped. During data mapping, the values input to each institution were converted to values in the same range according to the data dictionary, and the variable names were also converted to be the same. After that, empirical research was by dividing the study into one using questionnaires and one using exposure to harmful factors. Based on the analytical method used for each subject in the empirical study,

regression analysis, survival analysis, and meta-analysis were performed for statistical analyses.

Results

A mapping table was prepared by dividing the questionnaire into three sections (general questionnaire, night questionnaire, and special examination) in the questionnaire section for each institution. The general questionnaire was divided into five areas, and the total number of variables was 85. The night questionnaire was divided into 5 areas, and the total number of variables was 64. The special examination was divided into 11 areas, and the total number of variables was 40. In the examination results section, the mapping table was produced by dividing it into physical measurements and clinical examinations. Physical measurements included 99 variables, and clinical tests included 231 variables.

In the empirical study using questionnaires, a study on the number of consecutive night work days and insomnia, a study on shift interval and insomnia, and a study on insomnia and constipation were conducted. In the empirical study using exposure to harmful factors, a study on noise exposure and fasting blood sugar, a study on noise exposure and hypertension, and a study on dust exposure and diabetes were conducted.

Discussion

Since there is no official standard code for the Workers' Health examination questionnaire, a comprehensive data dictionary was created, and the newly created code was used to include all variables for this study. In the variables of the questionnaire, Severance Hospital

was mapped to 79.4%, Ulsan University Hospital to 68.2%, and Wonju Severance Hospital to 66.1%. For harmful exposure variables, Severance hospital was mapped to 76.1%, Ulsan University Hospital to 76.1%, and Wonju Severance Hospital to 37.0%. However, Wonju Severance Hospital had a low mapping rate (%) because this study only had a small number of variables.

Although there was some dissatisfaction with the lack of information in the study using questionnaires and exposure to harmful factors, the Workers' Health examination data using a CDM was confirmed to have high validity. Since a CDM can be standardized in the same way even with additionally loaded data, if more institutions participate and improve the completeness of the survey responses, accurate results can be obtained on various topics.

Conclusion

This study has shown that a CDM could be used to study data from Workers' Health examinations. However, it was found that there are parts that need to be supplemented in the current special health examination data. If we complement this and build distributed big data, we will be able to conduct better research. If a representative institution in each region of Korea uses this CDM to perform a study, it will be possible to learn about the characteristics of all the people who undergo Workers' Health examinations in Korea.

Keywords: workers' health examinations, CDM(common data model), data standardization, application of research

I. Introduction

Recently, as the amount of data in the medical/pharmaceutical field in Korea increases, research is being conducted using real-world data (RWD) called real-world evidence (RWE) research. RWE research in the medical/pharmaceutical field processes and analyzes actual data collected through large-scale trials and retrospective or prospective observational studies and yields other results that cannot be explained in clinical trials [1]. This type of research is not the same as a clinical trial because factors such as comorbidities and age, which were controlled in clinical trials, are not the same in the real world [2]. Therefore, the importance of RWE is being emphasized, and RWE is playing an increasingly important role in research using medical data.

According to the US Food and Drug Administration (FDA), medical RWD include electronic health records (EHRs), claims and billing activities, product and disease registries, patient-generated data including data collected in home-use settings, and data gathered from other sources that can inform about health status, such as mobile devices [3]. However, because the databases that collect these data were made for different purposes and have different data structures, it is very difficult to study and combine them [4].

To solve this problem, the Observational Health Data Sciences and Informatics (OHDSI) program in the US started using the Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) to analyze already collected medical data.

The goal of the OMOP-CDM is to make it easier to move data from different observational databases into a common format with a defined vocabulary that can then be used for systematic analysis [5]. Accordingly, overseas, the feasibility of the OMOP-CDM was confirmed using various data such as comparative effectiveness data, longitudinal community registry data, active safety surveillance research, French national EHRs, Austrian claims data, and questionnaire data from a tertiary care hospital in Singapore, and based on this, as a result, the OMOP-CDM has been established as an appropriate alternative to overcome the differences in data structure[6,7,8,9,10,11].

A study using the OMOP-CDM was also conducted in Korea to analyze medical data. The following OMOP-CDM-related research has been conducted: a study of medical device information EHRs [12], a colorectal cancer study [13], a pilot study of a large volume of polysomnography (PSG) data [14],a study of anti-seizure drugs (ASDs) and adverse drug reactions (ADRs) [15], and a study of anti-seizure medication treatment pathways [16].

Korea's National Health Insurance Corporation has health examination big data; thus, research using health examination data is possible, but the content is limited to information measured in general health examinations. The Workers' Health examination has parts that do not overlap with the general health examination (e.g., nighttime questionnaire, special questionnaire), and the Health Insurance Corporation does not have data on this part. There is a need for big data research using Workers' Health examination data, but it has been difficult to conduct such research because the data from

each institution are in a different form. For this reason, the Workers' Health examination data have not yet been big data-ized. Therefore, it is necessary to process the data from each institution to big data-ize the Workers' Health examination data and to determine whether the processed data can be used for research as big data.

II. Objectives

The purpose of this study is to find out step by step whether the Common Data Model method is appropriate for research use when applying it to Workers' Health examination data.

1) First, the data are technically processed to transform the Workers' Health examination data into distributed big data.

2) Second, research applying the CDM method is conducted using distributed big data and review the results.

III. Study Background

1. Common Data Model

A CDM is used to standardize by transforming datasets of different structures and shapes of various medical institutions into datasets of the same structure and shape. However, standardized datasets are not integrated into one database, but are retained by each institution. As such, a CDM is a model that executes the same analysis code in organizations that have the same data structure and form and integrates only the results.

To date, various types of CDM methods have been developed from medical data, and these methods are known and used in research. The types of CDM are as follows:

1) Patient-Centered Outcome Research Network Common Data Model (PCORnet CDM) [17]

Created and operated by the National Patient-Centered Outcome Research Institute (PCORI), PCORnet is a nationwide resource where health data, research knowledge, and patient perspectives are accessible in order to provide quick, reliable solutions that improve health outcomes. The PCORnet CDM version 6.0, which has 23 tables showing all the data on EHRs, has recently been released.

2) Observational Medical Outcomes Partnership Common Data Model (OMOP-

CDM) [18]

Created and operated by the OHDSI (pronounced "Odyssey") program is a multi-stakeholder, interdisciplinary collaboration aimed at maximizing the value of health data via large-scale analytics, and all of our solutions are open source. Developed by the OHDSI international consortium, it is a CDM in which more than 200 institutions in 14 countries around the world participate. To date, more than 660 million medical records have been converted to this CDM. As such, many organizations in various countries use the same data structure as well as a common physical and logical model using a common medical terminology system called the OMOP code.

3) Sentinel Common Data Model (SCDM) [19]

The U.S. FDA created and operated the Sentinel Common Datal Model (SCDM). Through the Sentinel Initiative, the FDA intends to create new methods for evaluating the safety of licensed medical items, such as medications, vaccines, and medical devices. Currently, the SCDM is maintaining two models concurrently as SCDM v8.0.0 and SCDM v8.1.0. The SCDM v8.0.0 includes 16 tables.

4) Integrating Biology and the Bedside (i2b2) Common Data Model [20]

i2b2 and tranSMART were created to give clinical and translational researchers with the tools necessary to integrate medical records and clinical research data in the genomic era. This is a new CDM that was released in v1.0 in 2021 and has six tables.

According to the characteristics of each dataset, the abovementioned CDMs are used for research.

2. Data Standardization

To proceed with CDMs using medical data, standardization should be performed first. The standardization method is also used to classify CDMs [21]. This is how the standardization method classifies things.

1) Organizing CDMs

Organizing CDMs involve organizing multiple raw datasets into one standardized data structure. Standardized tables include normal inpatient records, outpatient records, treatments, laboratory tests, and drug prescriptions. Organizing CDMs provides a way to organize tables by type and order of raw data but does not interpret or transform the recorded information. Therefore, organizing CDMs is the stage of preparing a mapped CDM or adaptive rule system. These CDMs are SCDM, i2b2, and PCORnet.

2) Mapping CDMs

Mapping CDMs involve applying multiple rules to a dataset to define standardized data structures and variables that can be analyzed. Mapping generally encompasses the full range of variables that can be analyzed and, once mapped, data can be used by researchers. Therefore, researchers can always see the information in the mapped composition instead of the term or information appearing in the source data.

Mapping algorithms also vary the underlying data source according to the structure,

richness, and completeness of the source data. Therefore, new mapping rules are required whenever the basic data content is changed, and a new CDM to which it is applied is also required. A CDM of this kind is the OMOP-CDM.

3) Adaptive rule system

An adaptive rule system is created based on the experience of configuring and mapping a CDM. It is a logic query language for analyzing longitudinal data. Researchers can change the data as needed, and the rules are collected according to the shared library. This kind of CDM has an EU-ADR.

The goal of this study is to determine if it is possible to use a mapping CDM with OMOP-CDM.

3. Difference between distributed big data research and big data research

What is distributed big data research? It is data composed in such a way that only the statistical results analyzed by each institution are collected by sending the analysis program code to the institutions participating in the research rather than collecting the data in one place. The benefit of distributed big data is that it does not collect data in one place. thus, there is no security risk associated with the data collection process, and it yields the same results as if the data were collected and analyzed.

What is big data research? Statistical errors can be reduced and better predictions can be made by conducting research using many cases. On the other hand, in the process of collecting data in one place, personal information is exposed, and problems related to data design are occurring.

IV. Methods

1. Materials

In this study, Severance Hospital, Wonju Severance Hospital, Gachon University Hospital, Ajou University Hospital, and Ulsan University Hospital participated in the standardization of Workers' Health examination data from each institution. However, only three university hospitals (Severance Hospital, Wonju Severance Hospital, and Ulsan University Hospital) participated in the empirical study. Therefore, this study was conducted in three university hospitals.

For data from three medical institutions, Workers' Health examination data and general health examination data were used for workers who received Workers' Health evaluations at participating research institutes between January 2013 and December 2017. However, there was a difference in the year in which standardization was performed because each institution provided data only from the year in which data standardization and empirical research were performed, but it commonly ranged from January 2015 to December 2017.

2. Data Standardization

The three medical institutions differed in their developed industries according to their geographic location; therefore, there were some differences in the health examination data collected from the subjects. To understand this, a data dictionary was first prepared for each institution for the data collected by each institution. In the data

dictionary for each institution, the variable name of the examination item used, the meaning of the variable name, the meaning of the input value, and the input method were prepared. At this time, all variable names were made to consist only of English names and numbers.

Furthermore, since the three medical institutions used different electronic medical record (EMR) systems, the input contents of the Workers' Health examination results were different. For example, in institution A, male/female is an input as a variable of sex; in institution B, '1'/'2' is input as a variable of sex; and in institution C, M/F is input as a variable of sex. As such, the variables and values were different.

Consequently, the subject area and code name corresponding to each variable name were identified using the standardized code used in the OHDSI network. In addition, it was set according to the domain classified by ATHENA that is website of OHDSI vocabularies, and the standard code was selected by selecting the standard concept and class. At this time, the vocabulary libraries used were mainly SNOMED, LOINC, and Nebraska Lexicon, but the vocabulary libraries were limited to 'measurement' and 'observation' in domain. Therefore, there was no term that fit the variables on the health examination questionnaire, so a new concept ID was created and standardized.

The self-generated concept ID was created based on the Severance Hospital data among the three institutions and was considered a master file. The data from the other two institutions were mapped to the variable name and value input method according to the Severance Hospital concept ID.

3. Data Processing

To utilize a CDM, PCs were provided to each institution, and a work environment was established on the PCs. First, the OS can utilize the CPU core to the maximum, and Linux (CentOS) was installed to be suitable for handling large amounts of data. In addition, a server that can be easily improved for data handling can be used, and the root folder is more secure than other OSs, so Linux(CentOS) was installed.

In addition, PostgreSQL was installed to use Structured Query Language (SQL), a computer language used for data processing in the database system, and R and Rstudio, Achilles, and ATLAS were installed. Furthermore, postgresql was set so that Rstudio and ATLAS were linked through WebAPI.

However, since the PC with Linux installed was a PC acting as a server, a separate PC with Windows OS was used to connect to the Linux PC, and DBeaver, a software program acting as a viewer, was installed on the Windows PC and used as an SQL database management tool (Figure 1).

In addition, the Pentaho Spoon program was installed to preprocess and store data, and using this, data filtering was performed for each domain, the input value was converted, and the converted data were stored on the server. If an error occurred during this process, the cleaning data were saved on the server by checking the error log written in the Spoon program and repeating the error correction operation (Figure 2).

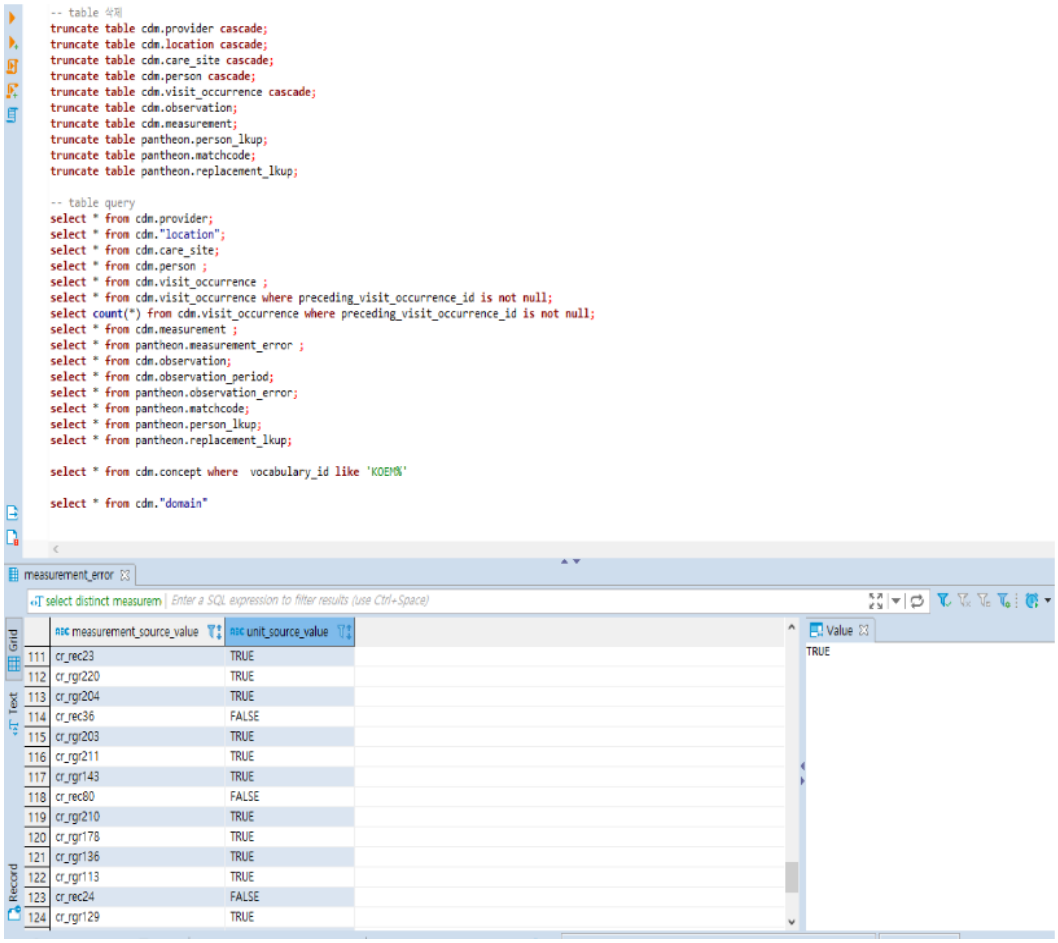

Figure 1. A screen sample of DBeaver used as a database management tool

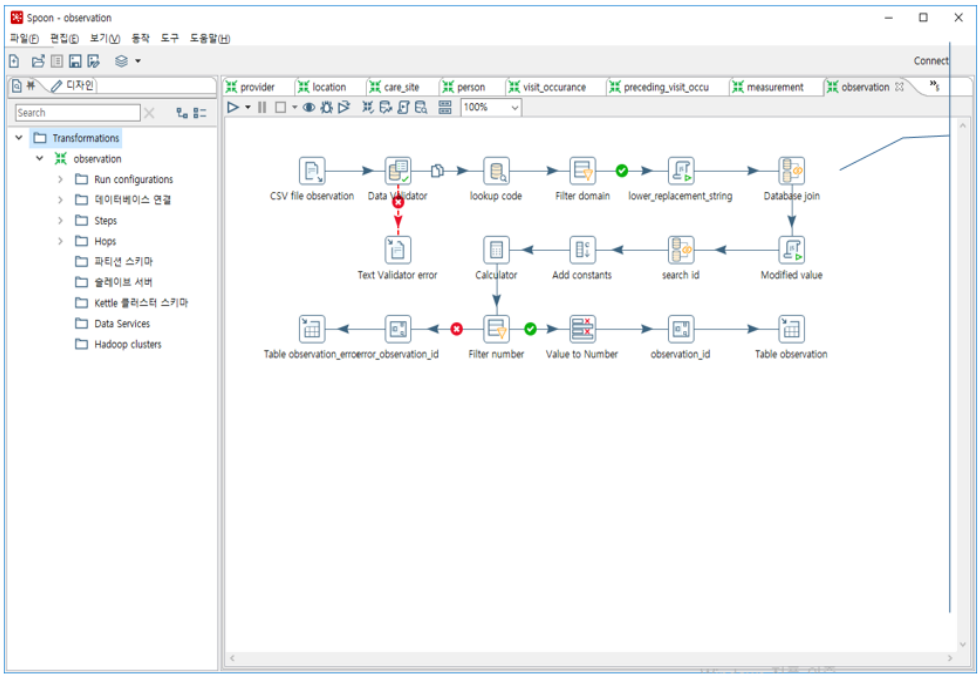

Figure 2. A screen sample of Spoon used as a server storage tool

DBeaver and Spoon, the programs used in ATALS OHDSI, allow DB management and data storage, but after extracting the final dataset, analysis must be performed using R studio. Therefore, it was judged that it is not efficient to use multiple programs; thus, R Studio and R shiny were used for all tasks from DB management to server storage, analysis, and result visualization.

The saved data are included four pieces of data: medical, general examination, special examination, and night examination information. All files that were set in wide form were converted into long form, and all were stored regardless of the year and number of variables entered in person id. made to be After that, a visit occurrence variable was created using the person's ID and the date of the examination. This way, even if the same person had more than one visit, the ID of that person indicates a visit that corresponded to that person.

However, in the case of duplicate visit dates, the source value was reviewed to make the visit occurrence unique. In addition, if there was a value such as 'negative' or 'positive' as text in the variable value where continuous numbers should be entered, the text was converted to a number, and one more variable was created and saved so that the existing text value could be checked.

The variable names were converted to English and numerals according to Severance's data dictionary, which was considered a master file so that the data dictionary and the variable names of the dataset match. At this time, categories were divided into a main category and subcategories according to the data dictionary.

The harmful factors were matched with the ‘CAS no’ used by the American Chemical Society, and if there was no ‘CAS no’, one was arbitrarily assigned (Figure 3).

A domain was created through the above process, and the data were saved according to the domain (Figure 4).

value	cascode	cascode2
1.1.2.2-테트라클로로에탄	79-34-5	79345
1.1.2-트리클로로에탄	79-00-5	79005
1.1-디클로로-1-플루오로에탄	1717-00-6	1717006
1.2-에폭시프로판	75-56-9	75569
1.3-부타디엔	106-99-0	106990
1.4-디옥산	123-91-1	123911
1-부틸알콜	71-36-3	71363
1-브로모프로판	106-94-5	106945
2.3-에폭시-1-프로판올	556-52-5	556525
2-메톡시에탄올	109-86-4	109864
2-메톡시에틸아세테이트	110-49-6	110496
2-부톡시에탄올	111-76-2	111762
2-부톡시에탄올아세테이트(에틸렌글리콜모노부틸에테르아세테이	112-07-2	112072
2-부톡시에탄올함유제제	111-76-2	111762
2-부틸알코올	78-92-2	78922
2-부틸알코올함유제제	78-92-2	78922
2-부틸알콜	78-92-2	78922
2-브로모프로판	75-26-3	75263
2-에톡시에탄올	110-80-5	110805
2-에톡시에탄올함유제제	110-80-5	110805
2-에톡시에틸아세테이트	111-15-9	111159
DMAC	127-19-5	127195
DMF	68-12-2	68122
IPA	67-63-0	67630
o-디클로로벤젠	95-50-1	95501

Figure 3. CAS no matching sample

Figure 4. data set in server

4. Statistical Analysis

In data processing, no statistical analysis was performed. However, statistical analysis was performed in empirical studies that used data.

Statistical analysis for each subject was performed according to the purpose of the empirical study, which included multiple logistic regression, multiple logistic regression (fixed-effect model), multivariate time-dependent Cox proportional hazards models, landmark analysis, propensity score matching, meta-analysis.

V. Results

1. Part 1: Data Standardization

1.1. Data Profile: Questionnaires

To connect the data on the questionnaire for each institution, the variable code was mapped for each institution. There was no usable vocabulary for the variable code of the Workers' Health examination questionnaire because there was no international or domestic agreement. To accomplish this, a mapping table was made based on the internal data dictionary and split into three sections: the general questionnaire, night questionnaire, and special examination.

In the general questionnaire, a code called GQ was assigned, and the subcategories were divided into disease history (19 ea), smoking (7 ea), drinking (44 ea), exercise (5 ea), and old age (10 ea), and the total score was 85 ea. The night questionnaire was assigned a code called NQ, and the subcategories were exposure evaluation (14 ea), insomnia evaluation (7 ea), gastrointestinal evaluation (6 ea), breast cancer evaluation (6 ea), drowsy evaluation (8 ea), and sleep quality evaluation (23 ea), and the total number of variables was 64 ea. For the special examination, a code called SE was designated, and the subcategories were hearing examination (2 ea), eye examination (3 ea), urinary examination (4 ea), gastrointestinal examination (3 ea), and cardiovascular examination (5 ea), and general was divided into examinations (3 ea), mouth examinations (2 ea), neurology examinations (6 ea), musculoskeletal examinations (5 ea), nose examinations (3 ea), and skin examinations (4 ea). The total number of variables was 40 ea. In addition,

variable codes were arbitrarily assigned according to the contents of the subcategories, and a total of 189 variables corresponded to questionnaires. Variable codes according to each subcategory for each main category are shown in Table 1.

Table 1. Questionnaire variables category mapping table

Main Category	Main Category Code	Sub-category	Variables Code	EA
General Questionnaire	GQ(78)	Disease history	hist_g1~g14	14
			f_hist_g1~g5	5
		Smoking	smok_g1~g7	7
		Drinking	drink_g101~g104	4
			drink_g202~g241	40
		Exercise	exe_g4~g8	5
		Old age	old_g1~g10	10
Night Questionnaire	NQ(64)	Exposure evaluation	nw_n1~n14	14
		Insomnia evaluation	insom_n1~n7	7
		Gastrointestinal evaluation	gastro_n1~n6	6
		Breast cancer evaluation	brcn_n1~n6	6
		Drowsy evaluation	drowsy_n1~n8	8
		Sleep Quality evaluation	sqol_n1~n23	23

Table 2. Questionnaire variables category mapping table (Continuously)

Main Category	Main Category Code	Sub-category	Variables Code	EA
Special Examination	SE	Hearing examination	sd_s1~s2	2
		Eye examination	ey_s1~s3	3
		Urinary examination	ur_s1~s4	4
		Gastrointestinal examination	gs_s1~s3	3
		Cardiovascular examination	cv_s1~s5	5
		General examination	ge_s1~s3	3
		Mouth examination	mo_s1~s2	2
		Neurology examination	nr_s1~s6	6
		Musculoskeletal examination	mu_s1~s5	5
		Nose examination	ns_s1~s3	3
		Skin examination	sk_s1~s4	4

1.2. Data Profile: Exposures

Body measurements and clinical pathology results were mapped using the standardized code provided by the OHDSI program. In the OHDSI measurement domain, 197 variables were mapped with Workers' Health examination data, but the remaining variables were not mapped. Therefore, a mapping table for this study was also prepared and divided into two categories. After the general inspection results were named the "registration general result," a code called RGR was designated. The other part was the results of the health examination performed for recruitment, and it was called the "registration employment examination", for which the code REC was designated.

In addition, the subcategories were largely divided into body and pathology, and because of the nature of the Workers' Health examination, it is sometimes performed twice a year; thus, in the RGR part, it was divided into primary and secondary. In the section on general registration results, Body 1st had 54 pieces, Pathology 1st had 132 pieces, Body 2nd had 45 pieces, and Pathology 2nd had 99 pieces. In addition, there were 18 employment bodies and 96 employment pathologies in the registration employment examination sector, and a total of 189 variables corresponded to exposures. Variable codes according to each subcategory for each main category are shown in Table 2.

Table 3. exposure variables category mapping table

Main Category	Main Category Code	Sub-category	Variables Code	EA
Registration general result	RGR	Body 1st	bd_rgr1~8 bd_rgr101~146	8 46
		Pathology 1st	cr_rgr1~37 cr_rgr101~195	37 95
		Body 2nd	bd_rgr201~247	45
		Pathology 2nd	cr_rgr201~300	99
Registration employment examination	REC	employment Body	bd_rec1~18	18
		employment Pathology	cr_rec1~96	96

1.3. Data Mapping

Variable codes were assigned to all variables that could be surveyed in the Workers' Health examination, and variables that could be matched to the variable codes were counted. As a result, in Severance Hospital, there were 46 ea of GQ, 64 ea of NQ and 40 ea of SE, for a total of 150, and Ulsan University Hospital had 33 ea of GQ, 56 ea of NQ and 40 ea of SE, for a total of 129. In Wonju Severance Hospital, 22 ea of GQ, 63 ea of NQ, and 40 ea of SE were able to map, for a total of 125 (Table 3).

Variable codes were assigned to all variables of body measurements and clinical pathology results, and variables that could be mapped to the corresponding variable codes were counted for each institution. As a result, in Severance Hospital, Body 1st had 47 ea, Pathology 1st had 94 ea, Body 2nd had 47 ea, and Pathology 2nd had 63 ea, for a total of 251. In Ulsan University Hospital, Body 1st had 47 ea, Pathology 1st had 94 ea, Body 2nd had 47 ea, and Pathology 2nd had 63 ea. The was the same as in Severance Hospital, with a total of 251. In Wonju Severance Hospital, a total of 122 could be mapped with 32 ea of Body 1st, 76 ea of Pathology 1st, 3 ea of Body 2nd, and 11 ea of Pathology 2nd (Table 4).

Table 4. Number of variables mapped for Questionnaire in each institution

	Severance Hospital	Ulsan university Hospital	Wonju Severance Hospital
General Questionnaire	46	33	22
Night Questionnaire	64	56	63
Special Examination	40	40	40

Table 5. Number of variables mapped for Exposure in each institution

Sub-category	Severance Hospital	Ulsan university Hospital	Wonju Severance Hospital
Body 1st	47	47	32
Pathology 1st	94	94	76
Body 2nd	47	47	3
Pathology 2nd	63	63	11

Variables that must additionally be written in text or variables that require numbers to be written but with an answer written in text, a number that can be replaced was randomly assigned from 1 to a continuous natural number. Therefore, it was possible to give each of the 17 variables a random replacement value of natural numbers, and a total of 1,184 values could be data-ized (Table 5).

Table 6. Examples of values of subjective variables and the number of converted values

Main Category Code	Variables	Input value example	Count
REC	cr_rec37	congruence, lack of inspection, result unconfirmed etc.	6
REC	cr_rec58	None, Yes, nothing special etc.	7
REC	cr_rec59	None, nothing special, yes etc.	21
REC	cr_rec60	None, No specific symptoms, nothing special etc.	23
REC	cr_rec61	None, nothing special, No special findings etc	11
RGR	cr_rgr123	normal, periodontal disease, caries etc	115
RGR	cr_rgr125	normal, nothing special, nothing strange etc.	64
RGR	cr_rgr132	normal, No special findings, limited etc	68
RGR	cr_rgr147	normal, No special findings, arrhythmia (Hunting of the Mac. Heart rhythm abnormality) etc.	179
RGR	cr_rgr150	No special findings, normal, degenerative changes etc	66

Table 7. Examples of values of subjective variables and the number of converted values (continuously)

Main Category Code	Variables	Input value example	Count
RGR	cr_rgr154	class1, negative, sample defect etc	26
RGR	cr_rgr18	normal(a), inactivity (c), non-tuberculous disease (F) etc	273
RGR	cr_rgr193	normal, anterior capsule formation, bleeding etc	36
RGR	cr_rgr247	Not good quality of sleep, Moderate daytime sleepiness and poor sleep quality, poor quality of sleep etc	13
RGR	cr_rgr249	Lt. B 4K (-), Rt. B 4K (-),Lt.B 3~4K (-) etc	220
RGR	cr_rgr287	normal, degenerative changes, minor degenerative changes etc	49
RGR	cr_rgr295	normal, tartar formation, destruction of enamel etc	7

2. Part 2: Empirical

2.1. Questionnaires

2.1.1 Consecutive Night Shift and Insomnia Study

It was possible to conduct a study applying a CDM using the night work and insomnia questionnaires, and the research results are as follows.

Objectives and Methods

The purpose of this study was to evaluate the number of consecutive nights of work that aggravate insomnia in night shift workers [22].

The subjects of this study were cases in which both the values of the independent variable 'number of consecutive nights of workdays' response and the dependent variable 'Insomnia Severity Index (ISI)' were present among the data from January 2015 to December 2017. The subjects included 13,311, 6,429, and 13,929 patients from Institutions No. 1, 2, and 3, respectively, for a total of 33,669 patients.

The independent variable was the number of consecutive night shifts, which included the responses 1) no continuous night shifts, 2) two consecutive nights shift, 3) three consecutive nights shift, 4) four consecutive nights shift, and 5) five or more consecutive nights shifts. The dependent variable was the same questionnaire as the ISI, an insomnia measurement tool, and the total score was 0 to 28 points. According to the score, 0–7 points were categorized as non-insomnia,

and 8 points or more were categorized as insomnia.

Statistical analysis was performed to calculate the odds ratios (ORs) of insomnia by multiple logistic regression. Then, using the CDM analytical method, the results of each institution were combined and evaluated as a whole to determine the pooled ORs.

Results

The baseline characteristics of the three institutions were as follows. In Institution No. 1, statistical analysis using the chi-square test and t-test showed significant differences in all variables except working hours. When all variables were corrected, the consecutive night shifts was the highest at three nights (Tables 6).

In Institution No. 2, statistical analysis using the chi-square test and t-test showed significant differences in all variables. When all variables were corrected, the consecutive night shift was the highest at three nights (Tables 7).

In Institution No. 3, all variables were significantly different in statistical analysis using the chi-square test and t-test. When all variables were adjusted, the consecutive night shifts were similar, but three nights was the highest, with a slight difference (Tables 8).

Table 8. Baseline characteristics and Multivariable Logistic Regression Models of Each Institution No. 1

Institution No. 1	Insomnia (N=5817)	Non Insomnia (N=7494)	P-value	Model 1	Model 2
(Intercept)				0.531 (0.448-0.629)	0.463 (0.383-0.561)
Sex					
Male	2707 (46.5%)	4022 (53.7%)	<0.001	(reference)	(reference)
Female	3110 (53.5%)	3472 (46.3%)		**1.089 (1.007-1.179)**	**1.120 (1.033-1.214)**
Age				**0.991 (0.988-0.995)**	**0.992 (0.989-0.996)**
Mean±SD	36.8±10.0	38.1±10.2	<0.001		
Median [Min, Max]	34.0 [19.0, 69.0]	36.0 [19.0, 69.0]			
Working hours					
under 52h	4088 (70.3%)	5224 (69.7%)	0.490		(reference)
over 52h	1729 (29.7%)	2270 (30.3%)			0.998 (0.915-1.088)
Rest time between shifts					
Slow return (more 11hr)	4259 (73.2%)	5767 (77.0%)	<0.001		(reference)
Quick return (< 11hr)	1558 (26.8%)	1727 (23.0%)			**1.295 (1.188-1.412)**
Consecutive night shifts					
None	635 (10.9%)	1563 (20.9%)	<0.001	(reference)	(reference)
two nights	655 (11.3%)	715 (9.5%)		**2.176 (1.187-2.508)**	**2.182 (1.891-2.516)**
three nights	1235 (21.2%)	867 (11.6%)		**3.301 (2.898-3.760)**	**3.392 (2.973-3.871)**
four nights	575 (9.9%)	880 (11.7%)		**1.734 (1.502-2.001)**	**1.852 (1.596-2.148)**
five or more nights	2717 (46.7%)	3469 (46.3%)		**2.006 (1.804-2.231)**	**2.056 (1.847-2.289)**

Table 9. Baseline characteristics and Multivariable Logistic Regression Models of Each Institution No. 2

Institution No. 2	Insomnia (N=2158)	Non Insomnia (N=4271)	P-value	Model 1	Model 2
(Intercept)				0.716 (0.570-0.898)	0.566 (0.445-0.720)
Sex					
Male	1311 (60.8%)	2782 (65.1%)	<0.001	(reference)	(reference)
Female	847 (39.2%)	1489 (34.9%)		1.034 (0.921-1.160)	1.115 (0.990-1.257)
Age				**0.982 (0.978-0.987)**	**0.983 (0.979-0.988)**
Mean±SD	41.4±12.0	44.8±12.5	<0.001		
Median [Min, Max]	40.0 [20.0, 77.0]	45.0 [20.0, 85.0]			
Working hours					
Under 52hrs	1182 (54.8%)	2588 (60.6%)	<0.001		(reference)
Over 52hrs	976 (45.2%)	1683 (39.4%)			**1.368 (1.221-1.533)**
Rest time between shifts					
Slow return (more 11hr)	1563 (72.4%)	3338 (78.2%)	<0.001		(reference)
Quick return (< 11hr)	595 (27.6%)	933 (21.8%)			**1.181 (1.044-1.335)**
Consecutive night shifts					
None	487 (22.6%)	1539 (36.0%)	<0.001	(reference)	(reference)
two nights	394 (18.3%)	766 (17.9%)		**1.549 (1.315-1.826)**	**1.578 (1.336-1.862)**
three nights	399 (18.5%)	506 (11.8%)		**2.053 (1.711-2.464)**	**2.114 (1.759-2.539)**
four nights	112 (5.2%)	172 (4.0%)		**1.915 (1.475-2.486)**	**1.966 (1.511-2.557)**
five or more nights	766 (35.5%)	1288 (30.2%)		**1.761 (1.534-2.023)**	**1.668 (1.449-1.919)**

Table 10. Baseline characteristics and Multivariable Logistic Regression Models of Each Institution No. 3

Institution No. 3	Insomnia (N=5050)	Non Insomnia (N=8879)	P-value	Model 1	Model 2
(Intercept)				0.415 (0.350-0.492)	0.395 (0.332-0.469)
Sex					
Male	4373 (86.6%)	7945 (89.5%)	<0.001	(reference)	(reference)
Female	677 (13.4%)	934 (10.5%)		**1.284 (1.118-1.475)**	**1.309 (1.139-1.504)**
Age				0.997 (0.994-1.001)	0.997 (0.994-1.000)
Mean±SD	43.0±11.3	43.8±11.5	<0.001		
Median [Min, Max]	44.0 [19.0, 76.0]	45.0 [18.0, 73.0]			
Working hours					
Under 52hrs	4119 (81.6%)	7394 (83.3%)	0.0111		(reference)
Over 52hrs	931 (18.4%)	1485 (16.7%)			**1.155 (1.051-1.269)**
Rest time between shifts					
Slow return (more 11hr)	3995 (79.1%)	7172 (80.8%)	0.0188		(reference)
Quick return (< 11hr)	1055 (20.9%)	1707 (19.2%)			**1.151 (1.054-1.258)**
Consecutive night shifts					
None	833 (16.5%)	2181 (24.6%)	<0.001	(reference)	(reference)
two nights	282 (5.6%)	462 (5.2%)		**1.401 (1.170-1.676)**	**1.396 (1.166-1.670)**
three nights	461 (9.1%)	573 (6.5%)		**1.796 (1.524-2.115)**	**1.790 (1.519-2.109)**
four nights	1296 (25.7%)	2031 (22.9%)		**1.712 (1.538-1.905)**	**1.785 (1.601-1.990)**
five or more nights	2178 (43.1%)	3632 (40.9%)		**1.610 (1.461-1.773)**	**1.599 (1.451-1.762)**

When the results of the multivariate logistic regression with sex, age, work hours, and rest time between shifts as covariates were pooled, they yielded an OR of 2.65 (95% confidence interval (CI) 1.97–3.56) for insomnia was found compared with 'no continuous night shifts' at three consecutive nights (Table 9).

Furthermore, although there was a difference in the second and third highest values in each institution, it was confirmed that the ORs showed the highest value for 'three consecutive night shifts' in all three institutions (Table 10).

However, it was difficult to accurately assess sleep quality because lifestyle habits such as drinking and smoking were not confirmed. In addition, it was hard to elucidate the effects of continuous night work because there was not enough information about each person's night work profile.

Table 11. Pooled Odds Ratios of Insomnia in Multivariable Logistic Regression Models

	Model 1	Model 2
Consecutive night shifts		
None	1.00 (reference)	1.00 (reference)
two nights	1.69 (1.29–2.21)	1.81 (1.45–2.26)
three nights	**2.32 (1.59–3.39)**	**2.65 (1.97–3.56)**
four nights	1.78 (1.64–1.93)	1.68 (1.55–1.82)
five or more nights	1.78 (1.54–2.06)	1.78 (1.56–2.03)

Adjusted for Model 1: sex, age.

Adjusted for Model 2: sex, age, working hours, Rest time between shifts

Table 12. Odds ratios for prevalence of insomnia by consecutive night shift in each institution

Consecutive night shifts	Model 2
two nights	
U	1.60(1.35-1.89)
W	1.63(1.39-1.90)
S	2.25(1.96-2.60)
three nights	
U	2.11(1.82-2.44)
W	2.49(2.11-2.94)
S	3.51(3.09-3.98)
four nights	
U	1.67(1.50-1.86)
W	2.06(1.59-2.67)
S	1.61(1.40-1.85)
five or more nights	
U	1.57(1.43-1.73)
W	1.88(1.64-2.15)
S	1.93(1.74-2.14)

2.1.2 Rest Time Between Shifts (Quick Return) and Insomnia Study

It was possible to conduct a study applying a CDM using night work and insomnia questionnaires, and the research results are as follows.

Objectives and Methods

This study confirmed the relationship between rest time between shifts and insomnia in shift workers [23].

The subjects of this study were cases in which both the values of the independent variable 'rest time between shifts' response and the dependent variable 'ISI' were present among the data from January 2015 to December 2017. The subjects included 13,311, 6,429, and 13,929 patients from Institutions No. 1, 2, and 3, respectively, for a total of 33,669 patients.

As for rest time between shifts, the independent variables were 1) slow return (over 11 hours) and 2) quick return (less than 11 hours). Dependent variables were classified as non-insomnia due to the absence of insomnia, subthreshold insomnia, moderate insomnia, and severe insomnia according to the categories of the ISI.

Statistical analysis yielded ORs of insomnia using multiple logistic regression.

Results

The baseline characteristics of the three institutions were as follows (Table

11). In Institutions No. 1 and 3, quick returns were more frequent among men than slow returns, but in Institution No. 2, quick returns were more frequent among women than slow returns. In Institutions No. 1 and 2, the age of quick return was lower than that of slow return, but in Institution No. 3, the age of quick return was higher than that of slow. In the quick return of all institution, there were more responses of 'over 52hours' than slow return.

Institution No. 1 had many quick returns during cases of no and two consecutive nights, and Institution No. 2 had many quick returns during cases of two, four, and five or more consecutive nights. Institution No. 3 had many quick returns, except for during cases of four consecutive nights. With regard to insomnia, all three institutions had a high number of quick returns, which was statistically significant.

Table 13. Basic characteristics of each institution with respect to rest time between shifts

	Institution No. 1			Institution No. 2			Institution No. 3		
	Quick return (N=3285)	Slow return (N=10028)	P-value	Quick return (N=1528)	Slow return (N=4901)	P-value	Quick return (N=2762)	Slow return (N=11167)	P-value
Sex									
Female	1458 (44.4%)	5124 (51.1%)	< 0.001	618 (40.4%)	1718 (35.1%)	< 0.001	306 (11.1%)	1305 (11.7%)	0.390
Male	1827 (55.6%)	4904 (48.9%)		910 (59.6%)	3183 (64.9%)		2456 (88.9%)	9862 (88.3%)	
Age									
Mean ± SD	36.1 ± 9.53	38.0 ± 10.3	< 0.001	41.5 ± 11.8	44.3 ± 12.6	< 0.001	45.1 ± 11.6	43.1 ± 11.4	< 0.001
Median [min, max]	33.0 [19.0, 69.0]	36.0 [19.0, 69.0]		40.0 [20.0, 79.0]	44.0 [20.0, 85.0]		46.0 [18.0, 76.0]	44.0 [19.0, 73.0]	
Working hours									
Under 52 h	1376 (41.9%)	7938 (79.2%)	< 0.001	824 (53.9%)	2946 (60.1%)	< 0.001	2021 (73.2%)	9492 (85.0%)	< 0.001
Over 52 h	1909 (58.1%)	2090 (20.8%)		704 (46.1%)	1955 (39.9%)		741 (26.8%)	1675 (15.0%)	

Table 14. Basic characteristics of each institution with respect to rest time between shifts (continuously)

	Institution No. 1			Institution No. 2			Institution No. 3		
	Quick return (N=3285)	**Slow return (N=10028)**	**P-value**	**Quick return (N=1528)**	**Slow return (N=4901)**	**P-value**	**Quick return (N=2762)**	**Slow return (N=11167)**	**P-value**
Consecutive night shifts									
None	735 (22.4%)	1463 (14.6%)	< 0.001	283 (18.5%)	1743 (35.6%)	< 0.001	728 (26.4%)	2286 (20.5%)	< 0.001
two nights	427 (13.0%)	943 (9.4%)		318 (20.8%)	842 (17.2%)		189 (6.8%)	555 (5.0%)	
three nights	455 (13.9%)	1648 (16.4%)		210 (13.7%)	695 (14.2%)		234 (8.5%)	800 (7.2%)	
four nights	156 (4.7%)	1300 (13.0%)		81 (5.3%)	203 (4.1%)		301 (10.9%)	3026 (27.1%)	
five or more nights	1512 (46.0%)	4674 (46.6%)		636 (41.6%)	1418 (28.9%)		1310 (47.4%)	4500 (40.3%)	
ISI									
Insomnia	1558 (47.4%)	4260 (42.5%)	< 0.001	595 (38.9%)	1563 (31.9%)	< 0.001	1055 (38.2%)	3995 (35.8%)	0.019
Non-insomnia	1727 (52.6%)	5768 (57.5%)		933 (61.1%)	3338 (68.1%)		1707 (61.8%)	7172 (64.2%)	

When the covariates were adjusted, there was an OR of 1.21 (95% CI 1.12–1.31) for insomnia in subjects with quick returns compared with those with slow returns (Table 12).

However, a causal relationship between a quick return and insomnia could not be confirmed. Furthermore, a quick return was determined based on a questionnaire asking about past experiences, and the frequency of quick returns was not measured. Interactions could happen if people did not know about diseases or ways of living in the past that could affect insomnia.

Table 15. Insomnia prevalence pooled odds ratio by rest time between shifts

	Model 1	Model 2
Rest time between shifts		
Slow return	1.00 (reference)	1.00 (reference)
Quick return	1.20 (1.11–1.29)	1.21 (1.12–1.31)

Model 1: gender, age.

Model 2: gender, age, working hours, consecutive night shifts.

2.1.3 Insomnia and Constipation Study

It was possible to conduct a study applying a CDM using the special examination and insomnia questionnaires, and the research results are as follows.

Objectives and Methods

This study examined the link between poor sleep in shift workers and frequency of constipation [24].

The subjects of this study were shift workers aged 30 years or older with data from January 2015 to December 2017 and cases with an independent variable of the 'ISI' and a dependent variable of 'constipation/. A total of 17,529 people participated in the study, including 12,879 people from Institution No. 1 and 4,650 people from Institution No. 2.

According to the classification criteria of the ISI, the independent variable categorizes 0–7 points as absence of insomnia, 8–14 points as subthreshold insomnia, 15–21 points as moderate insomnia, and 22–28 points as severe insomnia. Constipation was the dependent variable, and there were three levels: 1) none, 2) mild, and 3) severe. Mild and severe constipation were considered.

Statistical analysis calculated ORs of constipation were determined with multiple logistic regression (fixed-effect model). When all p-values of the heterogeneity tests were less than 0.05, a fixed-effect model was used. Sensitivity analysis was performed.

Results

The baseline characteristics of subjects from the two institutions participating in this study were as follows (Tables 13–14). In Institution No. 1, women were younger when they had constipation than men. The more severe the symptoms of insomnia were, the more constipation appeared. Non-smokers also had more constipation. Individuals with a BMI was <18.5 had the most constipation. When exercise was not performed, more constipation was found. Finally, when the number of working years was less than 15 years, there was more constipation.

The age of constipation was lower in the Institution No. 2, and there were more women than men. The more severe the symptoms of insomnia were, the more constipation appeared. When the working hours were less than 52 hours, non-smokers experienced more constipation. When the BMI ranged from 18.5 to 22.9, constipation was the most prevalent, and when exercise was not performed, there was more constipation. When the working interval was short and when the number of working years was less than 15 years, more constipation was present.

Table 16. Basic characteristics of Institution No. 1

	normal	constipation	P-value
Age(≧30years)			
Mean ± SD	43.79 ± 8.03	42.29 ± 8.00	< 0.001
Sex			
Female	4110 (69.78%)	1780 (30.22%)	< 0.001
Male	6146 (87.94%)	843 (12.06%)	
Insomnia			
none	6126 (84.25%)	1145 (15.75%)	< 0.001
sub-threshold	3216 (75.28%)	1056 (24.72%)	
moderate	801 (70.39%)	337 (29.61%)	
severe	113 (57.07%)	85 (42.93%)	
Working hours			
Under 52 h	9087 (79.45%)	2350 (20.55%)	0.161
Over 52 h	1169 (81.07%)	273 (18.93%)	
Smoking history			
non-smoker	5424 (75.43%)	1767 (24.57%)	<0.001
ex-smoker	2118 (85.13%)	370 (14.87%)	
current-smoker	2714 (84.81%)	486 (15.19%)	
BMI			
18.5-22.9	3798 (75.34%)	1243 (24.66%)	<0.001
<18.5	437 (73.45%)	158 (26.55%)	
23-24.9	2553 (82.01%)	560 (17.99%)	
≧25	3468 (83.97%)	662 (16.03%)	

Table 17. Basic characteristics of Institution No. 1 (continuously)

	normal	constipation	P-value
exercise			
yes	4192 (84.18%)	788 (15.82%)	<0.001
no	6064 (76.77%)	1835 (23.23%)	
working interval			
Long	8577 (79.82%)	2168 (20.18%)	0.242
Short	1679 (78.68%)	455 (21.32%)	
Shift type			
three shifts	9270 (79.57%)	2380 (20.43%)	0.612
others	986 (80.23%)	243 (19.77%)	
Working year			
<15	4703 (76.88%)	1414 (23.12%)	<0.001
≧15	5553 (82.12%)	1209 (17.88%)	

Table 18. Basic characteristics of Institution No. 2

	normal	constipation	P-value
Age(≧30years)			
Mean ± SD	47.05 ± 10.17	45.61 ± 9.71	< 0.001
Sex			
Female	1157 (69.53%)	507 (30.47%)	< 0.001
Male	2499 (83.69%)	487 (16.31%)	
Insomnia			
none	2499 (83.38%)	498 (16.62%)	< 0.001
sub-threshold	940 (71.48%)	375 (28.52%)	
moderate	183 (64.89%)	99 (35.11%)	
severe	34 (60.71%)	22 (39.29%)	
Working hours			
Under 52 h	2172 (77.46%)	632 (22.54%)	0.019
Over 52 h	1484 (80.39%)	362 (19.61%)	
Smoking history			
non-smoker	1679 (74.62%)	571 (25.38%)	<0.001
ex-smoker	860 (83.82%)	166 (16.18%)	
current-smoker	1116 (81.28%)	257 (18.72%)	
BMI			
18.5-22.9	1221 (75.32%)	400 (24.68%)	<0.001
<18.5	110 (77.46%)	32 (22.54%)	
23-24.9	879 (79.19%)	231 (20.81%)	
≧25	1446 (81.37%)	331 (18.63%)	

Table 19. Basic characteristics of Institution No. 2 (continuously)

	normal	constipation	P-value
exercise			
yes	1782 (81.56%)	403 (18.44%)	<0.001
no	1874 (76.02%)	591 (23.98%)	
working interval			
Long	2872 (79.38%)	746 (20.62%)	0.021
Short	784 (75.97%)	248 (24.03%)	
Shift type			
three shifts	1630 (76.35%)	505 (23.65%)	<0.001
others	2026 (80.56%)	489 (19.44%)	
Working year			
<15	2623 (78.23%)	730 (21.77%)	0.309
≧15	1033 (79.65%)	264 (20.35%)	

When all covariates were corrected, when insomnia was severe, the OR was 4.15 (95% CI 3.18–5.41) for constipation, which was high. Furthermore, the crude model showed that the pooled ORs for constipation tended to increase as the severity of insomnia increased, even when covariates were considered (Table 15).

Furthermore, the pooled ORs also increased as the severity of insomnia for both men and women stratified by sex increased (Table 16).

Table 20. Pooled ORs and 95% CIs for constipation with insomnia

	Model 0	Model 1	Model 2	Model 3
Insomnia				
None	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Sub-threshold	1.82 (1.68-1.97)	1.77 (1.63-1.92)	1.76 (1.62-1.91)	1.76 (1.62-1.91)
Moderate	2.35 (2.07-2.66)	2.31 (2.03-2.63)	2.29 (2.02-2.61)	2.28 (2.01-2.60)
Severe	3.84 (2.98-4.95)	4.18 (3.21-5.44)	4.17 (3.20-5.44)	4.15 (3.18-5.41)
P for trend	<0.001			

Model 0: crude model.

Model 1: adjusting for age and sex.

Model 2: Model 1 + adjusting for BMI, exercise, and smoking history

Model 3: Model 1 + adjusting for working year, shift type, working interval, and working hours

Table 21. Pooled ORs and 95% CIs for constipation with insomnia stratified by sex

	Model 0	Model 1	Model 2	Model 3
Insomnia(male)				
None	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Sub-threshold	2.20 (1.93–2.50)	2.18 (1.91–2.48)	2.15 (1.89–2.45)	2.13 (1.87–2.43)
Moderate	3.34 (2.76–4.04)	3.29 (2.72–3.98)	3.21 (2.65–3.88)	3.12 (2.57–3.79)
Severe	5.52 (3.89–7.81)	5.43 (3.83–7.69)	5.40 (3.81–7.67)	5.22 (3.66–7.42)
Insomnia(female)				
None	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Sub-threshold	1.55 (1.39–1.72)	1.51 (1.36–1.68)	1.52 (1.36–1.69)	1.52 (1.36–1.69)
Moderate	1.80 (1.51–2.13)	1.72 (1.44–2.04)	1.74 (1.46–2.07)	1.74 (1.46–2.07)
Severe	3.29 (2.22–4.89)	3.18 (2.14–4.72)	3.23 (2.17–4.81)	3.23 (2.17–4.81)

Model 0: crude model.

Model 1: adjusting for age and sex.

Model 2: Model 1 + adjusting for BMI, exercise, and smoking history

Model 3: Model 1 + adjusting for working year, shift type, working interval, and working hours

A strong correlation between insomnia and constipation could be confirmed. However, since it was a cross-sectional study, no causal relationship could be verified. Furthermore, since the survey assessing constipation is not a medical diagnosis but a question asking about an individual's thoughts, the accuracy may be decreased. In addition, even though there were enough covariates, lifestyle variables that could have affected the dependent variable were not considered.

2.2. Exposure Assessment

This study, which applied a CDM, could be conducted using the exposure assessment, and the research results are as follows.

2.2.1 Noise and High Fasting Blood Glucose Study

Purpose and Methods

The purpose of this study was to confirm the association between occupational noise and the occurrence of fasting blood glucose (FBG) [25].

The subjects of this study were those who underwent a health examination in 2013 or 2014, were followed up three years later, and then underwent a health examination in 2016 or 2017. There were 11,932 patients from Institution No. 1 and 31,926 patients from Institution No. 3 among the subjects. A total of 43,858 people participated in the study.

As an independent variable, participants exposed to occupational noise exceeding 85 dB for six hours a day during the follow-up period were classified into 1) a noise exposure group, and the remaining participants were classified into 2) a non-exposure group. For noise exposure, the Korea Occupational Safety and Health Agency used a database to measure noise levels by monitoring the workplace.

The dependent variable was a high FBG greater than 100 mg/dL or a response of ‘yes’ to the questions 'Have you been diagnosed with diabetes by a

physician?' or 'Are you taking any drugs for diabetes?' The rest of the subjects were considered to have a non-high FBG.

For statistical analysis, the hazard ratios (HRs) of each institution were calculated using multivariate time-dependent Cox proportional hazard models, and pooled HRs values were calculated. Furthermore, the survival rate of subjects with a high FBG was confirmed using Kaplan–Meier plots.

Results

The baseline characteristics of the subjects from the two institutions participating in this study were as follows (Table 17). The subject age was older in the noise exposure group at Institution No. 1 and No. 3, and there were more males than females in the noise exposure group. In Institution No. 1, non-smokers had the highest noise exposure, but in Institution No. 3, current smokers had the highest noise exposure. For BMI, the noise exposure group had the highest number of subjects with a normal BMI in both institutions. alcohol had a higher Institution No. 1 of 'no', but Institution No. 3 had more 'yes'.

Questions about hypertension had more 'no' responses at both institutions, and questions about had more 'yes' responses for physical exercise at both institutions, but only that at Institution No. 1 was statistically significant. In terms of cardiovascular-related exposure, there were more replies of 'no' at Institution No. 1 and more replies of 'yes' at Institution No. 3.

Table 22. Demographic Characteristics of Study Population in each hospital

	Institution No. 1 (n=11,932)			Institution No. 3 (n=31,962)		
	Unexposed group	**Noise exposed group**	**P-value**	**Unexposed group**	**Noise exposed group**	**P-value**
Age			<0.001			<0.001
Mean ± SD	35.97 ± 9.3	40.28 ± 10.91		39.84 ± 9.84	44.24 ± 10.26	
Sex			<0.001			<0.001
Male	3858(37.86%)	1555(89.32%)		10532(81.24%)	18328(96.66%)	
Female	6333(62.14%)	186(10.68%)		2432(18.76%)	634(3.34%)	
Smoking history			<0.001			<0.001
non-smoker	7573(74.31%)	731(41.99%)		6233(48.08%)	4923(25.96%)	
ex-smoker	894(8.77%)	331(19.01%)		3189(24.60%)	5312(28.02%)	
current-smoker	1724(16.92%)	679(39.00%)		3542(27.32%)	8727(46.02%)	

Table 23. Demographic Characteristics of Study Population in each hospital (continuously)

	Institution No. 1 (n=11,932)			Institution No. 3 (n=31,962)		
	Unexposed group	**Noise exposed group**	**P-value**	**Unexposed group**	**Noise exposed group**	**P-value**
BMI			<0.001			<0.001
underweight	1121(11.00%)	38(2.18%)		353(2.72%)	188(0.99%)	
normal	5780(56.72%)	767(44.06%)		5150(39.73%)	7656(40.38%)	
overweight	1729(16.97%)	438(25.16%)		3569(27.53%)	5811(30.65%)	
obese	1561(15.31%)	498(28.60%)		3892(30.02%)	5307(27.98%)	
Alcohol consumption			<0.001			<0.001
Yes	2473(24.27%)	828(47.56%)		6205(47.86%)	10257(54.09%)	
No	7718(75.73%)	913(52.44%)		6759(52.14%)	8705(45.91%)	

Table 24. Demographic Characteristics of Study Population in each hospital (continuously)

	Institution No. 1 (n=11,932)			Institution No. 3 (n=31,962)		
	Unexposed group	**Noise exposed group**	**P-value**	**Unexposed group**	**Noise exposed group**	**P-value**
Hypertension			<0.001			<0.001
Yes	742(7.28%)	274(15.74%)		1136(8.76%)	1877(9.90%)	
No	9449(92.72%)	1467(84.26%)		11828(91.24%)	17085(90.10%)	
Physical exercise			<0.001			0.625
Yes	4303(42.22%)	1011(58.07%)		11365(87.67%)	16659(87.85%)	
No	5888(57.78%)	730(41.93%)		1599(12.33%)	2303(12.15%)	
Cardiovascular related exposure			0.028			<0.001
Yes	1835(18.01%)	275(15.80%)		1327(10.24%)	12261(64.66%)	
No	8356(81.99%)	1466(84.20%)		11637(89.76%)	6701(35.34%)	

The high HR of FBG at Institution No. 1 was 1.35 (95% CI 1.24–1.48), the high HR of FBG at Institution No. 3 was 1.22 (95% CI 1.17–1.28), and the pooled high HR of FBG was 1.28 (95% CI 1.16–1.41) in the multivariate time-dependent Cox proportional hazard analysis (Table 18).

When the gender and age groups were stratified, the risk of high occurrence of FBG was higher in men and those over 40 years of age due to exposure to noise (Table 19).

However, since the health examination is performed once a year, the exact timing of hyperglycemia could not be confirmed. In addition, it was not possible to correct for non-professional noise such as noise generated by residential areas. Furthermore, since retirees were not included, there was a possibility that healthy workers were working.

Table 25. Hazard Ratio from multivariate time-dependent Cox analysis

	Institution No. 1	Institution No. 3	Pooled
Noise exposure			
Non noise exposure	1.00 (reference)	1.00 (reference)	1.00 (reference)
Noise Exposure	1.35 (1.24–1.48)	1.22 (1.17–1.28)	1.28 (1.16-1.41)

Table 26. hazard ratio of high FBG from stratification analysis

	Institution No. 1	Institution No. 3	Pooled
Sex			
Male	1.36 (1.24–1.49)	1.23 (1.17–1.28)	1.28 (1.16-1.41)
Female	1.10 (0.84–1.44)	0.97 (0.74–1.27)	1.03 (0.85-1.25)
Age			
≥40	1.44 (1.29–1.61)	1.26 (1.20–1.33)	1.33 (1.17-1.52)
<40	1.27 (1.11–1.46)	1.24 (1.14–1.34)	1.25 (1.17-1.33)

2.2.2 Noise and Hypertension

Purpose and Methods

The purpose of this study was to investigate exposure to noise and the incidence of hypertension [26].

The subjects of this study were those who underwent annual health examinations from 2014 to 2017, and individuals without hypertension in 2014 were included. Of the 19,113 people who were studied, only 12,141 were finalists. Those who did not work in a company or who were not exposed to loud noise were not included in the study.

The independent variable of severe noise exposure was classified as 1) 'yes' when subjects were exposed to noise over 85 dB for six hours a day and 2) 'no' when they were not. All screening institutions in Korea have a database that measures the exposure of each worker, including severe exposure to noise; therefore, that database was used.

The dependent variable was the incidence of hypertension. Subjects were asked ‘Have you ever been diagnosed with hypertension?’ and ‘Are you taking antihypertensive medications?’ Blood pressure was evaluated for those who answered ‘yes’. People with a systolic blood pressure (SBP) of 140 mmHg or diastolic blood pressure (DBP) of 90 mmHg were assumed to have hypertension.

For statistical analysis, the appearance of hypertension was confirmed using a multivariate Cox proportional hazard model, and multivariate time-dependent

Cox proportional hazard models and landmark analysis were also performed.

Results

In the crude model, severe noise exposure yielded an HR of 2.37 (95% CI 2.08–2.69), which was higher than that for no severe noise exposure, and in the model to which all covariates were applied, severe noise exposure yielded an HR of1.28 (95% CI 1.11–1.47), indicating a higher incidence of hypertension (Table 20).

Similar results were obtained when various statistical methods were applied and compared. Based on the final model, the time-dependent Cox regression had the highest incidence of hypertension with an HR of 1.60 (95% CI 1.38–1.85), and the time-fixed Cox regression [HR 1.28 (95% CI 1.11–1.47)] and time-fixed Cox regression with landmark [HR 1.33 (95% CI 1.13–1.57)] showed similar results (Table 21).

It was possible to confirm the results of the comparison using various statistical analysis methods and to control for factors that were presumed to be confounding factors. Although only one institution (Severance Hospital) participated in this study, enough subjects were followed for four consecutive years for confirmation.

The lack of information on the presence or absence of hearing protection devices and previous work history could have caused bias. Furthermore, since the

information came from questionnaires, there was no date for the diagnosis of hypertension, which is inaccurate.

Table 27. Hazard ratios of hypertension in time-fixed Cox proportional hazard models

	Crude	Model 1	Model 2	Final Model
Noise exposure				
No	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Yes	2.37 (2.08-2.69)	1.21 (1.06-1.39)	1.21 (1.06-1.39)	1.28 (1.11-1.47)

Model 1: sex, age.

Model 2: sex, age, diabetes, smoking history.

Final Model: sex, age, diabetes, smoking history, Waist circumference, Exercise history, Drinking history, Family history of hypertension, Number of exposures related to cardiovascular risk.

Table 28. Hazard ratios of hypertension incidence by occupational noise exposure of each model

Statistical methods	Crude	Model 1	Model 2	Final Model
Time-fixed Cox regression	2.37 (2.08-2.69)	1.21 (1.06-1.39)	1.21 (1.06-1.39)	1.28 (1.11-1.47)
Time-dependent Cox regression	2.94 (2.56-3.37)	1.59 (1.38-1.84)	1.58 (1.37-1.82)	1.60 (1.38-1.85)
Time-fixed Cox regression with Landmark	2.34 (2.02-2.71)	1.22 (1.05-1.43)	1.22 (1.05-1.43)	1.33 (1.13-1.57)

2.2.3 Dust and Diabetes Study

Objectives and Methods

This study tried to confirm that male workers who are exposed to dust at work are more likely to develop diabetes [27].

The subjects of this study were those who underwent medical examinations for four consecutive years between January 2013 or January 2014 and December 2017 and those who did not have diabetes at the time of the first medical examination. Exposure was defined in the case of direct occupational exposure to dust. Due to the nature of the job, very few women are directly exposed to occupational dust; thus, only men were targeted.

Of the total of 13,835 people, 5,141 people participated in the study, except those who smoked, took hypoglycemic agents, or had an FBG of 7 mmol/L or higher.

The independent variable of dust exposure was classified into 1) ‘yes’ for exposure to mineral dust, wood dust, glass fiber dust, cotton dust or grain dust and 2) ‘no’ for the rest. In Korea, the level of exposure to inorganic and organic dust, welding fumes, and wood dust was measured; thus, this information was used.

Subjects were asked ‘Have you ever been diagnosed with type 2 diabetes?’ and ‘Are you taking blood sugar-lowering drugs?’. FBG levels were determined for those who answered ‘yes’. A person with an FBG of 7 mmol/L or greater was defined as having diabetes.

Statistical analysis confirmed the incidence of diabetes using multivariate time-dependent Cox proportional hazard models. In addition, as a sensitivity analysis, time-dependent Cox proportional hazard analysis was performed with HRs for a propensity score-matched dataset, and landmark analysis was also performed.

Results

In Model 1, the HR was1.61 (95% CI 1.05–2.12) times higher for dust exposure than for no dust exposure. And in the model controlling all covariates, the HR was 1.66 (95 % CI 1.26–2.20) times higher than for no dust exposure (Table 22).

There was no statistically significant interaction between occupational dust exposure and lifestyle factors. As a sensitivity analysis, the incidence of diabetes was 1.67 (95% CI 1.24–2.24) times higher in time-dependent Cox proportional hazard models using the propensity score-matched dataset (Table 23).

It was possible to confirm the incidence rate by comparing the results of the analysis using various statistical analysis methods and to confirm the association of lifestyle factors, which are estimated as confounding factors.

However, although sufficient cases were used in this study, only one institution (Institution No. 1) participated and the exact date of diabetes diagnosis could not be determined due to the nature of the questionnaire data. Since female workers were not included in the study, the effects on women could not be

determined. The level of exposure is also believed to be different in each workplace of the same company, but this could not be shown because the data were not complete enough.

Table 29. Hazards ratios from various multivariable Cox models of incident diabetes for diabetes risk factors

	Model 1	Model 2	Final Model
Dust exposure			
No	1.00 (reference)	1.00 (reference)	1.00 (reference)
Yes	1.61 (1.05-2.12)	1.70 (1.29-2.23)	1.66 (1.26-2.20)

Model 1: age.

Model 2: age, hypertension, Family history of diabetes.

Final Model: age, hypertension, Family history of diabetes, High LDL-cholesterol, High triglyceride, Smoking history, Waist circumference, Drinking history, No physical activity, Exposure to chemical/physical cardiovascular risk factors.

Table 30. Hazard ratios of diabetes incidence by interaction variables between dust exposure and lifestyle factors

Interaction variable	Non-exposure and healthy lifestyle	Dust exposure and healthy lifestyle	Non-exposure and unhealthy lifestyle	Dust exposure and unhealthy lifestyle	p for interaction
Abnormal waist circumference	1.00 (reference)	1.67 (1.24-2.24)	1.86 (1.34-2.58)	2.91 (1.28-6.59)	0.885
Smoking history	1.00 (reference)	1.85 (1.08-3.16)	1.21 (0.93-1.59)	1.99 (1.37-2.89)	0.697
Drinking history	1.00 (reference)	1.37 (0.88-2.14)	1.19 (0.94-1.52)	2.25 (1.58-3.22)	0.258
No physical activity	1.00 (reference)	1.40 (0.92-2.13)	1.05 (0.83-1.34)	2.01 (1.41-2.87)	0.269

VI. Discussion

This study tried to determine if it is possible to apply a CDM to big data when Workers' Health examination data from university hospitals in Korea can be used for research.

1. Part 1: Data Standardization

Technical processing was performed to standardize the Workers' Health examination data. Data standardization was performed through data mapping, which is the method of the OMOP-CDM, and the mapping range is all the variables that can be created in a Workers' Health examination. This was done to cover the entire scope of the mapping because there were different items from the three participating university hospitals. The reason for the different items was that the types of hazardous substances subjects were exposed to were different due to the different occupational industries of the subjects undergoing the Workers' Health examinations conducted by each institution and because the items tested were different. As a result, there was a difference in the source data for each institution, which also caused the differences in the mapping data.

As a result, in the mapping percentage of the questionnaire variables, Severance Hospital achieved 79.4%, Ulsan University hospital achieved 68.2%, and Wonju Severance Hospital achieved 66.1%. In the mapping percentage of exposure variables, Severance Hospital achieved 76.1%, Ulsan University Hospital achieved 76.1%, and Wonju Severance Hospital achieved 37.0%. In Wonju Severance Hospital, the mapping

percentage was low because few variables were provided for this study. If the variables are expanded in the future, it is expected that the mapping percentage of the relevant institution will increase.

Since the GQ smoking question was changed in 2018 and the drinking question changed in 2019, a data dictionary was created with all the questions from 2013 to 2019. Thus, approximately 40 variables were changed and were not mapped. If these variables are removed from the data dictionary, all three institutions achieved a mapping percentage of 80% or more. Although the country and fields such as claim data and EMRs were different, the mapping percentage that applied the OMOP-CDM was approximately 90%, and when the diagnosis name was mapped among the claim data in Korea, it was 87.8% based on the KCD7 standard [5,28]. Considering that the mapping percentage for diagnosis alone was considered, the overall variable mapping percentage in this study was considered to be relatively well done.

When the OMOP-CDM is applied, the standard code can be mapped through ATHENA operated by the OHDSI program, but the standard code did not exist in ATHENA for the questionnaire in this study. To this end, a data dictionary was created based on comprehensive variables, and a newly created code was used. Loss of information was identified as a source of potential damage in a prior study to confirm the feasibility of the OMOP-CDM [5], and it was determined that a new standard code was needed as a way to reduce this. As it has not been created in ATHENA or in previous studies, although this code is not an official standard code, it has become a newly defined

standard for this study. Through this, it is possible to completely prevent the loss of source data that may occur during the data conversion process.

DBeaver and Spoon, the programs used in the ATALS OHDSI program, allow DB management and data storage, but after extracting the final dataset, analysis must be performed using R Studio. Therefore, it was judged that it is inefficient to use several programs; thus, R Studio and R Shiny were used for all the work from DB management to server storage, analysis, and result visualization. As in 2019, data from EHRs were made into an R package so that the OMOP-CDM could be utilized; thus, in this study, the process was simplified using R Studio [29]. As a result, there was no difference between DB management and data storage using DBeaver and Spoon and DB management and data storage using R Studio.

The standardization of Workers' Health examination data through this study has strengths. First, it is the first study to apply the OMOP-CDM using Workers' Health examination data. To date, CDM studies have been conducted to evaluate medical devices, ADRs, and specific diseases [12,15,16]. On the other hand, as in this study, CDM research on Workers' Health examination data has not been performed; therefore, this study can serve as a very important cornerstone.

Second, results can be expected for all participants in Workers' Health examinations in Korea. Workers' Health examinations are conducted every year and are conducted by many domestic medical institutions, but data analysis was performed for each medical institution; thus, it could not reflect the overall situation in Korea. However, since this

study confirmed the possibility of applying a CDM to Workers' Health examination data, if a CDM is applied to representative institutions in each region, it will be possible to elucidate the characteristics of all participants in Workers' Health examinations in Korea.

2. Part 2: Empirical

After completion of data processing, questionnaires and exposures were divided, and an empirical study was conducted to evaluate the possibility of data utilization by applying a CDM. This is because the questionnaires created a new standard for the standardization of all variables, and the exposures were standardized based on Atlas' ATHENA, but some items were newly created. Therefore, it was divided into two sections.

In a study using a questionnaire, the relationship between insomnia and consecutive night shifts was investigated for the first time. Insomnia was evaluated using the ISI, and the validation of a Korean version of the ISI has been shown to obtain relatively accurate results based on many previous articles[30-32]. Consecutive nights shifts were inquired about by asking about the average of consecutive nights with a conditional question regarding the past year. This is not a question from which accurate results can be obtained; however, because the subject's thoughts may be included, information bias may occur. However, since almost identical results were found in all medical institutions that participated in this study, the bias could be said to have been eliminated.

In a study investigating the relationship between quick return to work and insomnia,

a quick return indicated that rest time between shifts was insufficient, as the time from one work shift to the next work shift was less than 11 hours. In this question, the frequency of overtime was not accurately indicated; therefore, there was a possibility of information bias or selective perception. Nevertheless, quick returns and insomnia were found to be related.

In addition, in a study on the relationship between insomnia and constipation, the symptoms of constipation were assessed with a question regarding instances ‘within the last 6 months'. Because the response was also divided into three stages of 'severe', 'mild' and 'none', the accuracy was low. However, the correlation between the two variables was found to be strong. Although the accuracy of the information in the questionnaire was low, similar results were obtained in several hospitals, and the same results were obtained when pooled. This is thought to have overcome the possible bias.

A regrettable fact in studies using questionnaires was that missing data on lifestyle habits such as drinking, smoking, and exercising could not be reflected. Furthermore, it was not possible to secure a larger sample size because there were many missing responses to the number of hours worked and years of service. In addition, only workers on the night shift responded to the NQ questionnaire, and there were no subjects who could be used as a control group. As such, it is believed that more accurate results would have been obtained if the responses had been faithfully induced to reduce missing values and to obtain information from night shift workers.

In the study on exposure, we investigated whether harmful substances such as noise

and dust affect hyperglycemia, hypertension, and diabetes. Since health examinations are generally performed once a year, the exact time of disease occurrence could not be confirmed, but it was possible to transform it into longitudinal data according to the characteristics of CDM data. Through this, it was possible to determine the year of occurrence the disease, and part of the results could be supplemented with statistical analysis methods.

As such, although there is some concern with the lack of information in studies using questionnaires and exposure, the Workers' Health examination data using a CDM were confirmed to have high validity. Since a CDM can be standardized in the same way even with more data loaded, if more institutions take part and the quality of survey responses improves, accurate results can be obtained for many themes. Our results demonstrated the application of CDM analysis and indicated the potential for future CDM distributed network research in the area of Workers' Health examinations in Korea.

VII. Conclusion

1. Part 1: Data Standardization

To apply the OMOP-CDM to the Workers' Health examination data, the data from each institution were standardized through mapping. The mapping percentage of the Workers' Health examination data of the three university hospitals participating in this study differed by classification, but on average it was approximately 70%. For your reference, for the standardization of values entered by each institution, checking and converting all values had to be performed more than once.

2. Part 2: Empirical Research

It is considered a valid research method to conduct a study using the CDM method for Workers' Health examination data. However, even with standardized data, it is difficult to analyze all variables because there are many missing values because not all tests are performed. Therefore, additional data cleansing must be performed according to the study design. In addition, health effects were able to be predicted using questionnaire and exposure information from the Workers' Health examinations. Consequently, when the CDM approach is used to data from Workers' Health examinations, it is possible to conclude that the data is eligible for study.

VIII. References

1. Kim H-S, Lee S, Kim JH. Real-world Evidence versus Randomized Controlled Trial: Clinical Research Based on Electronic Medical Records. *jkms.* 2018;33(34):0-0.
2. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence — What Is It and What Can It Tell Us? *New England Journal of Medicine.* 2016;375(23):2293-2297.
3. Administration USFD. Real-World Evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Published 2022. Updated 09/08/2022. Accessed 22/09, 2022.
4. Khosla S, White R, Medina J, et al. Real world evidence (RWE) - a disruptive innovation or the quiet evolution of medical evidence generation? (2046-1402 (Electronic)).
5. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association.* 2015;22(3):553-564.
6. FitzHenry F, Resnic FS, Robbins SL, et al. Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. *Appl Clin Inform.* 2015;06(03):536-547.
7. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. *Journal of Biomedical Informatics.* 2016;64:333-341.
8. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association.* 2011;19(1):54-60.
9. Lamer A, Depas N, Doutreligne M, et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. *Appl Clin Inform.* 2020;11(01):013-022.
10. Haberson A, Rinner C, Gall W. Standardizing Austrians Claims Data Using the OMOP Common Data Model: A Feasibility Study. Paper presented at: EFMI-STC2019.
11. Sathappan SMK, Jeon YS, Dang TK, et al. Transformation of Electronic Health Records and Questionnaire Data to OMOP CDM: A Feasibility Study Using SG_T2DM Dataset. *Appl Clin Inform.* 2021;12(04):757-767.
12. Choi S, Choi SJ, Kim JK, et al. Preliminary feasibility assessment of CDM-based active surveillance using current status of medical device data in medical records and OMOP-CDM. *Scientific Reports.* 2021;11(1):24070.
13. Ryu B, Yoon E, Kim S, et al. Transformation of Pathology Reports Into the Common Data Model With Oncology Module: Use Case for Colon Cancer. *J Med Internet Res.* 2020;22(12):e18526.

14. Kim J-W, Kim S, Ryu B, Song W, Lee H-Y, Yoo S. Transforming electronic health record polysomnographic data into the Observational Medical Outcome Partnership's Common Data Model: a pilot feasibility study. *Scientific Reports.* 2021;11(1):7013.
15. Choi SA, Kim H, Kim S, et al. Analysis of antiseizure drug-related adverse reactions from the electronic health record using the common data model. *Epilepsia.* 2020;61(4):610-616.
16. Kim H, Yoo S, Jeon Y, et al. Characterization of anti-seizure medication treatment pathways in pediatric epilepsy using the electronic health record-based common data model. *Frontiers in neurology.* 2020;11:409.
17. Institute P-COR. PCORnet. https://pcornet.org/. Published 2013. Updated Aug/2022. Accessed 23/09, 2022.
18. Informatics TOHDSa. OHDSI. https://www.ohdsi.org/. Published 2008. Accessed 23/09, 2022.
19. Administration TUSFaD. Sentinel Initiative. https://www.sentinelinitiative.org/. Published 2007. Accessed 23/09, 2022.
20. Community ib. i2b2 Common Data Model. https://community.i2b2.org/wiki/. Published 2020. Accessed 2022, 23/09.
21. Schneeweiss S, Brown JS, Bate A, Trifirò G, Bartels DB. Choosing Among Common Data Models for Real-World Data Analyses Fit for Making Decisions About the Effectiveness of Medical Products. *Clinical Pharmacology & Therapeutics.* 2020;107(4):827-833.
22. Sim J, Yun BY, Lee J, et al. The Association Between the Number of Consecutive Night Shifts and Insomnia Among Shift Workers: A Multi-Center Study. (2296-2565 (Electronic)).
23. Sim J, Yun B, Yoon J-H, et al. Relationship between insomnia and rest time between shifts among shift workers: A multicenter cross-sectional study. *Journal of Occupational Health.* 2022;64(1).
24. Yun B-Y, Sim J, Yoon J-H, Kim S-K. Association Between Insomnia and Constipation: A Multicenter Three-year Cross-sectional Study Using Shift Workers' Health Check-up Data. *Safety and Health at Work.* 2022;13(2):240-247.
25. Kim S, Yun B, Lee S, et al. Occupational Noise Exposure and Incidence of High Fasting Blood Glucose: A 3-Year, Multicenter, Retrospective Study. *International Journal of Environmental Research and Public Health.* 2021;18(17):9388.
26. Yun B, Sim J, Jeong I, et al. Does severe subacute noise exposure increase risk of new onset hypertension beyond conventional risk factors? A 30 000 person-years cohort study. *Journal of Hypertension.* 2022;40(3):588-595.
27. Yun B, Sim J, Lee S, et al. The relationship between occupational dust exposure and incidence of diabetes in male workers: A retrospective cohort study. *Diabetic Medicine.* 2022;39(6):e14837.
28. Ko SJ, Park SJ, Chang D-J. Experience of Converting Clinical Data Warehouse to Common Data Model and Additional Data Loading. *HIRA.* 2021;1(2):179-195.

29. Glicksberg BS, Oskotsky B, Giangreco N, et al. ROMOP: a light-weight R package for interfacing with OMOP-formatted electronic health record data. *JAMIA Open.* 2019;2(1):10-14.
30. Cho YW, Song ML, Morin CM. Validation of a Korean Version of the Insomnia Severity Index. *jcn.* 2014;10(3):210-215.
31. La YK, Choi YH, Chu MK, Nam JM, Choi Y-C, Kim W-J. Gender differences influence over insomnia in Korean population: A cross-sectional study. *PLOS ONE.* 2020;15(1):e0227190.
32. Jang T-W, Jeong KS, Ahn Y-S, Choi K-S. The relationship between the pattern of shift work and sleep disturbances in Korean firefighters. *International Archives of Occupational and Environmental Health.* 2020;93(3):391-398.

국문초록

다기관 특수건강검진 데이터의 공통데이터 모델(Common Data Model) 적용 및 평가

서론

최근 우리나라 의약학 분야에서 데이터의 양이 많아지면서 Real world evidence (RWE) 라는 이름으로 real world data (RWD) 를 활용하는 연구가 유행처럼 시행되고 있다. 미국은 의료데이터를 활용하여 연구하기 위해 OMOP-CDM을 활용하고 있다. 우리나라에서도 OMOP-CDM을 활용하여 Medical device, 대장암, 약물연구 등 다양한 분야에서 연구를 진행하고 있다. 특수건강검진 데이터를 활용한 빅데이터 연구가 필요하지만, 각 기관의 데이터가 다른 형태를 지니고 있어서 그 동안은 연구가 진행되기에 어려움이 있었다. 따라서 본 연구는 특수건강검진 데이터를 CDM에 적용하여 연구에 활용할 수 있는지 알아보고자 하였다.

방법

3개의 대학병원에서 2015년 1월부터 2017년 12월까지의 데이터를 사용하였다. 각 기관의 EMR 현황에 맞게 data dictionary를 제작하고, data를 매핑하였다.

data 매핑은 각 기관에 입력된 값을 data dictionary에 따라 동일한 범위의 값으로 변환하고, 변수명 또한 동일하도록 변환하였다. 이후 실증연구를 하였으며, 설문지를 활용한 연구와 유해인자 노출을 활용한 연구로 구분하여 진행하였다. 통계분석은 시행된 실증연구의 각 주제별 분석방법에 맞게 회귀분석, 생존분석, 메타분석 등을 시행하였다.

결과

각 기관별 설문지 부문에서 General Questionnaire, Night Questionnaire, Special Examination으로 나누어 매핑테이블을 제작하였다. General Questionnaire는 5개 영역으로 나누었고, 전체 변수는 85개였다. Night Questionnaire는 5개 영역으로 나누었고, 전체 변수는 64개였다. Special Examination는 11개 영역으로 나누었고, 전체 변수는 40개였다. 검사결과 부문에서는 신체계측, 임상검사로 나누어 매핑테이블을 제작하였다. 신체계측은 99개의 변수 였고, 임상검사는 231개의 변수였다.

설문을 활용한 실증연구에서는 연속야간근무일수와 불면증에 관한 연구와 교대 간격과 불면증에 관한 연구, 그리고, 불면증과 변비에 관한 연구를 시행하였다. 유해인자 노출을 활용한 실증연구에서는 소음 노출과 공복혈당에 관한 연구, 소음 노출과 고혈압에 관한 연구, 그리고, 먼지 노출과 당뇨에 관한 연구를 시행하였으며, 6개의 실증연구 관련 논문이 저널에 게재되었다.

고찰

특수건강검진 설문지는 공식적인 표준 코드가 없기 때문에 본 연구를 위해 모든 변수를 포함할 수 있도록 포괄적으로 데이터 사전을 만들고 새롭게 생성한 코드를 사용하였다. 설문지의 변수에서 세브란스 병원은 79.4%, 울산대학교 병원은 68.2%, 원주세브란스 병원은 66.1%를 매핑하였으며, 유해인자 노출 변수에서는 세브란스 병원은 76.1%, 울산대학교병원은 76.1%, 원주세브란스 병원은 37.0%를 매핑하였다. 다만, 원주세브란스 병원은 본 연구를 위해 제공된 변수가 적었기 때문에 매핑률(%)이 낮았으나, 추후 변수를 확장한다면, 해당 기관의 매핑률(%)은 높아질 것으로 기대된다.

설문지와 유해인자 노출을 활용한 연구에서 일부 부족한 정보에 대한 아쉬움은 있지만, CDM을 적용한 특수건강검진 data는 연구에 적용이 가능한 것으로 확인되었다. CDM 은 추가로 적재되는 데이터에서도 같은 방식으로 표준화가 이루어질 수 있기 때문에 더 많은 기관이 참여하고, 설문 응답의 완성도를 높인다면 다양한 주제로 정확한 결과를 얻을 수 있을 것이다.

결론

본 연구를 통해 특수건강검진 데이터는 CDM을 활용한 연구 가능성을 확인하였다. 다만, 현재의 특수건강검진 데이터에서 보완이 필요한 부분이 있음을 알

수 있었다. 이를 보완하여 분산형 빅데이터를 구축한다면 더 나은 연구를 할 수 있을 것이다. 또한, 우리나라 각 지역별 대표 의료기관과 함께 CDM을 적용한 연구를 시행한다면 우리나라 전체 특수건강검진 대상자의 특성에 대해 파악할 수 있을 것이다.

핵심어: 특수건강검진, 공통데이터모델, 데이터 표준화, 연구 적용