MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-467

Conversion of National Health Insurance Service-National Sample Cohort (NHIS-NSC) Database into Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM)

Seng Chan You[a*], Seongwon Lee[b*], Soo-Yeon Cho[a], Hojun Park[a], Sungjae Jung[a], Jaehyeong Cho[a], Dukyong Yoon[a], Rae Woong Park[a,c]

[a] *Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea*
[b] *Center for Education Artificial Intelligence, Dankook University, Yongin, Korea*
[c] *BrainKorea21 (BK21) Plus, Korea*
[*] *The first two authors contributed equally to this work*

Abstract

It is increasingly necessary to generate medical evidence applicable to Asian people compared to those in Western countries. Observational Health Data Sciences and Informatics (OHDSI) is an international collaborative which aims to facilitate generating high-quality evidence via creating and applying open-source data analytic solutions to a large network of health databases across countries. We aimed to incorporate Korean nationwide cohort data into the OHDSI network by converting the national sample cohort into Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM). The data of 1.13 million subjects was converted to OMOP-CDM, resulting in average 99.1% conversion rate. The ACHILLES, open-source OMOP-CDM-based data profiling tool, was conducted on the converted database to visualize data-driven characterization and access the quality of data. The OMOP-CDM version of National Health Insurance Service-National Sample Cohort (NHIS-NSC) can be a valuable tool for multiple aspects of medical research by incorporation into the OHDSI research network.

Keywords:
Delivery of Health Care; Medical Informatics Application; Database

Introduction

Use of existing observation data to generate medical evidence is rapidly increased in terms of quantity of data, diversity of data sources, and the transparency [1]. Observational Health Data Sciences and Informatics (OHDSI) is an international collaborative consortium which aims to facilitate generating high-quality evidence via creating and applying open-source data analytic solutions to a large network of health databases across countries [2]. The OHDSI adopted distributed research network (DRN), which mediates observational studies to be conducted using multi-site database, while confidential personal health data remain within the original data holders [3]. The standardized same data structure, called the Observational Medical Outcomes Partnership Common Data Model (OMOP- CDM), is imperative to create network-wide results through the DRN by running the same analysis program for cooperating organizations. The OMOP-CDM was first developed in 2008 for drug surveillance study and has expanded its capacity to other research area. It supports various types of studies such as drug/procedure safety, drug/procedure comparison, and medical cost analysis.

South Korea adopts a compulsory social insurance program, covering the whole population living in the country [4]. The National Health Insurance Service (NHIS), the institution for the Korean health insurance service holds the health claim database for all Koreans. In 2015, The NHIS released the NHIS-National Sample Cohort (NSC) database, which is a population-based sample cohort as the representative of the population as a whole.

In this paper, we present the converting process of the NHIS-NSC database into the OMOP-CDM and the result, focusing the executing phases.

Methods

Data Source

Korean public health insurance system for all citizens was initiated in 1963. Universal healthcare coverage was achieved in 1989. In 2000, the NHIS was launched as a single-insurer system by integrating more than 366 medical insurance organizations, for efficient system operation in Korea. The NHIS maintained national records for healthcare utilization and prescription over 98% of Korean whole population as of 2006. The NHIS established the NHIS-NSC, which was population-based cohort to provide representative, useful health insurance and health examination data to public health researchers and policy makers in 2015. About one million subjects, 2% of the Korean whole population, were selected by stratified random sampling from 2002 Korean health insurance database. Longitudinal health records in these population were collected for 11 years from 2002 to 2013. To preserve total number of subjects in the cohort, a representative sample of newborn was added annually as expired or emigrated subjects were excluded. The NHIS-NSC database can be assessed on the website [http://nhiss.nhis.or.kr/bd/ab/bdaba021eng.do].

Code Mapping

The NHIS uses Korean national medical code system. The 6th Korean standard Classification of Diseases (KCD-6) code is used for the diagnoses, which was originated from ICD-10 code. The NHIS has its own code system, electronic document

interchange (EDI) codes for drug, procedures and measurements.

The OMOP-CDM uses OMOP Standard Vocabulary(hereafter OMOP code), which requires transformation process from the local medial code system. OMOP code is based on RxNorm for medical drug and SNOMED-CT for medical diagnosis. Figure 1 shows how a drug code from Korean local drug code system is transformed to standard OMOP code. To consolidate the meaning of multi-site data, code transformation system is essential.

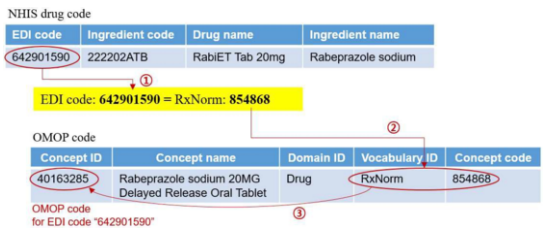

Figure 1– The example of transformation process from Korean local drug code to OMOP code.

We adopted the Korean Code Mapping Dictionary (KCMD) as a means of transoformation from the NHIS codes to the OMOP codes. The KCMD is the code mapping dictionary between Korean local codes and the OMOP codes, which is developed by the authors. It shows 99% coverage for the diagnoses (KCD-6), 95% for drugs (EDI codes), and 99% for procedure (EDI codes) by the number of patients treated. For measurement, we additionally developed a mapping dictionary.

Extraction, Transformation, Loading (ETL) process

The OMOP-CDM embraces a variety of medical data such as patients, encounters, diagnoses, drug exposures, procedures, devices, results from laboratory tests, anthropometric measurements or questionnaire, and medical costs.

We converted the original data necessary for drug safety monitoring into OMOP-CDM with priority. Therefore, the seven tables in OMOP-CDM were generated: person, death, visit_occurrence, condition_occurrence, drug_exposure, procedure_occurrence, and location. With these converted tables, the various research including drug effect analysis, pharmacovigilance analysis can be performed.

The logical ETL mapping rules between the NHIS-NSC and OMOP-CDM were first defined. There were problems in developing ETL mapping documents. First, the NHIS-NSC provides subjects' age as a form of five-year-interval age group. Therefore, the approximate birth year of each subject had to be presumed. Second, all treatments such as drugs, procedures, and medical devices were stored in a single table in the NHIS-NSC and there was no reference data to accurately distinguish the type of treatments. We had to execute ETL by applying the KCMD for each treatment to source data.

We developed the physical ETL scripts as a form of Standard Query Language (SQL) and performed them as actual ETL process.

Data Characterization and Quality Management

ACHILLES (Automated Characterization of Health Information at Large-scale Longitudinal Evidence System) is a open-source OMOP-CDM-based data profiling tool, developed by the OHDSI community. The ACHILLES creates a data-driven characterization and visualizes this result, by generating high-level aggregating statitics in a database based on the OMOP-CDM. The quality of the data can be assessed by the ACHILLES Heel, which is a subcomponent of the ACHILLES.

It is possible that mistakes in ETL processes occur [5]. To verify that the transformed data is equivalent to the original cohort data, we used statistical results from the ACHILLES to compare the population-level statitics with the previous reported cohort profile [6]. The ACHILLES Heel was also conducted to evaluation the data quality. The ACHILLES Heel generates two types of notifications: errors and warnings. "Errors" indicate more serious data quality issues, while "warnings" represent faults anticipated to have smaller impact.

Results

CDM Conversion Performance

Table 1 is the result of ETL, which shows counts of converted database records and conversion rates for OMOP-CDM tables. All data in the NHIS-NSC equivalent to subject, death, hospital visit and location were totally converted, resulting in 100% conversion rate. For the condition table, 89.4% diagnosis codes were mapped to SNOMED-CT and 99.9% of condition data were converted to OMOP-CDM with the mapping diagnosis codes. This means that 89.4% of codes showed the 99.9% prevalence in condition data. For drug and procedure, the total about 975 million data were converted to OMOP-CDM. Among them, drug and procedure comprised 53.6% (522,575,793 records) and 46.4% (452,147,182 records). Its whole conversion rate was 94.5%.

Table 1– Count of data converted and conversion rate from the NHIS-NSC to the OMOP-CDM

Table	Data count in data source	Data count converted	Conversion rate
Subject	1,125,691	1,125,691	100.0%
Death	55,921	55,921	100.0%
Hospital visit	119,362,188	119,362,188	100.0%
Condition	299,379,695	299,053,439	99.9%
Drug/Procedure	1,031,760,325	974,722,975	94.5%
Location	317	317	100%

Data Characteristics and Visualization

The ACHILLES estimated that total number of subject in the cohort was 1.13 million. Total of 560,640 (49.8%) subjects were female (Figure 2). The reported number of all subjects and annual included infants in the cohort are compared with the result from the ACHILLES (Table 2).

The number estimated by the ACHILLES was equal to the reported number in 2002, 2010, 2011, 2012 and 2013. The ACHILLES overestimated the number of subjects in the other years. The numbers of annual subjects estimated by the ACHILLES were compared with the previous reported numbers. The numbers of infants match each other except 2009. The difference in the number of infants was only four in 2009.

The ACHILLES calculates and visualizes the statistics for the population-level data from Korean nationwide cohort. Since the

ACHILLES visualizes statistics for diagnosis, drug, procedures and hospital visits by calculating prevalence and frequency per person by using the size and color of the boxes in the tree maps. Users easily assess the statistics for population-level data from Korean nationwide cohort.

Graphs from the ACHILLES about visiting emergency room (ER) and use of cimetidine were depicted in figure 3 and figure 4, respectively. The total number of people visiting ER during whole cohort period was 188,954 with a rising trend. One of the most commonly used drug was cimetidine. Number of subjects using cimetidine was increased from 2002 to 2007, and then decreased. The most commonly used population was women in their 60s and 70s.

Table 2– Comparison between previously reported numbers of all subjects and infants in the cohort and the results from ACHILLES

Year	Number of subjects in cohort		Number of infants aged 0 in the cohort	
	Reported	ACHILLES	Reported	ACHILLES
2002	1.025M	1.025M	9565	9565
2003	1.017M	1.032M	9437	9437
2004	1.016M	1.035M	9320	9320
2005	1.016M	1.036M	8557	8557
2006	1.002M	1.037M	7872	7872
2007	1.020M	1.039M	9766	9766
2008	1.000M	1.027M	9393	9393
2009	0.998M	1.024M	8616	8604
2010	1.002M	1.002M	9032	9032
2011	1.006M	1.006M	9694	9694
2012	1.011M	1.011M	9851	9851
2013	1.014M	1.014M	8825	8825

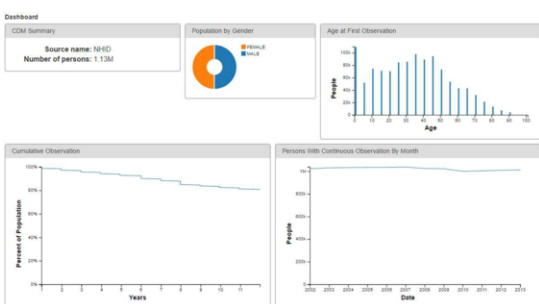

Figure 2– Dashboard of ACHILLES describes the total number of subjects in cohorts according to their gender, race, and year of birth.

Results from ACHILLES Heel

The ACHILLES Heel issued 13 errors and 15 warnings from converted the NHIS-NSC OMOP-CDM data. Most of errors and warnings were resulted from discordance between range of observation period and real date of medical claims

Discussion

We demonstrate successful conversion process of the NHIS-NSC, Korean nationwide cohort database, to the OMOP-CDM version 5.0 in this paper. To date, this has been the first attempt to convert nationwide cohort database to universal standardized OMOP-CDM format in Asian countries. Although the randomized clinical trial (RCT) undoubtedly remains as a gold standard for developing medical evidence, applicability of evidenced produced by RCTs can be restricted because of gaps between environments in RCTs and real world routine practice [7]. Geographic and racial variations in the risks for disease or the results of treatment also exist. However, these differences have been often neglected in multinational RCTs [8]. Incorporation of Korean longitudinal nationwide large cohort data into the OHDSI network will be the the first milestone to generate global high-quality medical evidence, which is applicable to Asian population and real-world practice.

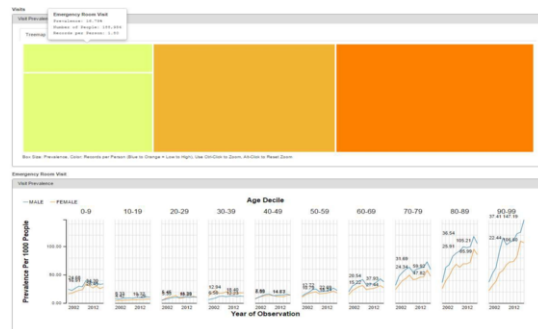

Figure 3– Visit tab of ACHILLES depicts the trend in emergency room visits of subjects in the cohort according to their gender and age.

Figure 4– Drug exposure tab of ACHILLES shows the trend in usage of cimetidine in the cohort according to their gender and age.

Some limitations still exist in the process and the result of our data conversion. Despite regular process in code mapping as previously described [9], information loss was inevitable in the process of code mapping and ETLs owing to unapproved drugs or procedures in US and Korean traditional medical diagnosis and drug. Some tables in the OMOP-CDM, especially regarding to medical cost, were left for the conversion because we planned to construct data set for drug safety monitoring first. Errors issued from the ACHILLES Heel have not fully

investigated. However, the number of error was lower than the median number of 19, which revealed by multi-site evaluation of data quality by using the ACHILLES Heel [5].

Conclusions

We report the successful transformation of the Korean nationwide cohort database, the National Health Insurance Service -National Sample Cohort (NHIS-NSC), into the OMOP-CDM model with acceptable loss of information. Converting additional information including cost data and further verification of OMOP-CDM by replicating previous research are now under way. The OMOP-CDM version of the NHIS-NSC can be a valuable resource for multiple aspects of medical research by incorporation into the OHDSI research network.

Acknowledgements

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI16C0992) and supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI14C3201).

References

[1] Murdoch TB, Detsky AS. The inevitable application of big data to health care. The Journal of American Medical Association 2013;309(13):1351-1352.
[2] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Studies in Health Technology and Informatics. 2015;216:574-578.
[3] Platt R, Wilson M, Chan, KA, Benner JS, Marchibroda J, McClellan M, The new sentinel network — improving the evidence of medical-product safety. The New England journal of medicine. 2009 Aug;361(7):645-647.
[4] Kwon S, Thirty years of national health insurance in South Korea: lessons for achieving universal health care coverage. Health Policy Plan. 2009 Jan;24(1):63–71.
[5] Huser V, DeFalco FJ, Schuemie M, Ryan PB. Multisite evaluation of a data quality tool for patient-level clinical datasets. eGEMs. 2016 Nov;4(1).
[6] Lee J, Lee JS, Park SH, Shin SA, Kim KW. Cohort profile: the National Health Insurance Service–National Sample Cohort (NHIS-NSC), South Korea. International journal of epidemiology. 2016 Jan 28;pii:dyv319.
[7] Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - What is it and what can it tell us?," The New England Journal of Medicine. 2016 Dec 8;375(23):2293–2297.
[8] Yusuf S, Wittes J. Interpreting geographic variations in results of randomized, controlled trials. The New England Journal of Medicine. 2016 Dec 8;375(23):2263–2271.
[9] Yoon D, Ahn EK, Park MY, Cho SY, Ryan PR, Schuemie MJ, et al. Conversion and data quality assessment of electronic health record data at a Korean tertiary teaching hospital to a Common Data Model for Distributed Network Research. Healthcare Informatics Research. 2016 Jan 31;22(1):54–58.

Address for correspondence

Rae Woong Park, MD, PhD

Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea

164, World cup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, Korea
Tel: +82-31-219-4470

Fax: +82-219-4472

E-mail: veritas@ajou.ac.kr