OMOP-CDM Conversion and Anonymization of National Health Insurance Service-National Sample Cohort

Seongwon Lee[1], Seng Chan You[1], Jimyung Park[1], Jaehyeong Cho[1], Santa Borel[2], Khaled El Emam[2,3],Rae Woong Park[1,4]

[1]Department of Biomedical Informatics, Ajou University School of Medicine; [2]Privacy Analytics, Toronto, Canada; [3]Department of Paediatrics, University of Ottawa, Ottawa, Ontario, Canada; [4]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea;

Is this the first time you have submitted your work to be displayed at any OHDSI Symposium?

Yes _____ No___O___

Abstract

Common Data Model (CDM) is a primary method for distributed research network. We converted the National Health Insurance Service-National Sample Cohort (NHIS-NSC) database into Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) and enhanced its anonymization level with the privacy-preserving software. We expect that the anonymized NHIS-NSC data can be a more valuable resource for OHDSI network.

Introduction

Along with the tremendous progress in adoption of OMOP-CDM in Korea, it is increasingly necessary to share the ETL conventions of Korean healthcare. South Korea adopts a compulsory social insurance program, covering the virtually all citizens. In 2015, The NHIS released the NHIS-National Sample Cohort (NSC) database, which is a population-based sample cohort as the representative of Korean population. While providing access to the NHIS-NSC database has contributed to generate numerous medical evidences for Korean population, confusion and fear surrounding the privacy issues of health-care big data prevented further release of recent databases.

The purpose of this study is to promote the unification of ETL process across Korean hospitals by sharing detailed process of ETL and ease the public fear about medical big data by establishing measuring the re-identification risk of CDM data and stronger anonymization.

Method

Data Source

NHIS-NSC contains comprehensive, rich, and pseudonymized information for health care utilization and health examination of 1.13 million subjects. Longitudinal health records in these population were collected for 12 years from 2002 to 2013. This database contains information of participants' insurance eligibility, medical treatment history, and results from general health examination[1].

ETL (Extract, Transform, and Load)

Total of fourteen OMOP-CDM tables of OMOP-CDM v 5.3.1 were converted from NHIS-NSC. The details are described in the github at https://github.com/OHDSI/ETL---Korean-NSC

Anonymization and Validation

NHIS-NSC and CDM themselves are fully pseudonymized data without any direct identifiers. Still, re-identification risk was measured under the certain situation such as researchers' breach of confidentiality[2]. If the risk was above the pre-specified threshold, further anonymization was conducted by using an automatic privacy-preserving software, the Eclipse (version 2.11, Privacy Analytics, Canada). The identical proof-of-concept comparative research was conducted before and after anonymization to test feasibility.

Results

The data of 1.13 million subjects was converted to OMOP-CDM, resulting in average 95.4% conversion rate (Table 1).

Table 1. Result of CDM Conversion

CDM Tables	Record count, n		Conversion rate (%)	Mapping coverage (%)
	Sample Cohort	CDM		
PERSON	1,125,691	1,125,691	100.00	Not applicable
DEATH	55,940	55,940	100.00	96.03
VISIT	121,572,555	121,570,475	100.00	Not applicable
CONDITION	296,252,657	299,419,634	101.07	98.65
DRUG	504,951,817	422,492,469	83.67	80.34
PROCEDURE	445,492,445	452,449,166	101.56	53.41
DEVICE	11,316,127	11,381,608	100.58	69.70
MEASURE MENT	33,440,451	33,440,451	100.00	100
OBSERVATION	33,218,703	33,218,703	100.00	100
COST	908,678,310	609,571,436	67.08	Not applicable

The probability of risk re-identifying an individual was .140 in converted CDM database which was higher than the prespecified threshold (.093). By suppressing data of medical concept IDs and date, it was decreased to .025 below the threshold.

Table 2 describes the result of proof-of-concept study which compares hypoglycemia risk between metformin and sulfonylurea before and after anonymization. We found that the statistical attributes were retained even after anonymization.

Table 2. Results of PoC Study

Data Source	Metformin		Sulfonylurea		RR	95% CI	*p* value
	Total	Events	Total	Events			
Converted CDM	20,349	72	22,051	140	0.49	(0.35 to 0.69)	0.00
Anonymized CDM	20,346	72	22,051	140	0.49	(0.34 to 0.69)	0.00

Conclusion

The whole process from conversion to strong anonymization of National Health Insurance Service-National Sample Cohort (NHIS-NSC) can be valuable for medical research by incorporation into the OHDSI research network.

Acknowledgement

This work was supported by the Bio Industrial Strategic Technology Development Program (20001234) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from 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].

Conflict of Interest

Employee: Santa Borel, and Khaled El Emam are employees at Privacy Analytics.

References

1. Lee J, Lee JS, Park S, Shin SA, Kim K. Cohort profile: the National Health Insurance Service-National Sample Cohort (NHIS-NSC), South Korea. Int J Epidemiol. 2015;46;e15.
2. Emam KE, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Rose S, Howard J, Gluck J. De-identification methods for open health data: the case of the heritage health prize claims dataset. J Med Internet Res. 2012;14;1-16.