OMOP-CDM Conversion and Anonymization of National Health Insurance Service-National Sample Cohort (NHIS-NSC)

Seongwon Lee, PhD1; Seng Chan You, MD1; Jimyung Park2; Jaehyeong Cho2; Santa Borel, MSc3; Khaled El Emam, PhD4,5; Rae Woong Park, MD PhD1,2,5

1Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea; 2Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea; 3Privacy Analytics, Toronto, Canada; 4Department of Paediatrics, University of Ottawa, Ottawa, Ontario, Canada; 5FEEDER-NET (Federated E-health Big Data for Evidence Rovation Network)

Background

• The National Health Insurance Service (NHIS), the institution for the Korean health insurance service holds the health claim database for all Koreans and provides the National Sample Cohort (NSC) database for research and policy purposes.
• NHIS-NSC are fully pseudonymized data without any direct identifiers. Still, re-identification risk was claimed under the certain situation such as researchers’ breach of confidentiality.
• The objectives of this study are:
  1) to present the ETL process of NHIS-NSC into CDM
  2) to establish stronger anonymization techniques for NHIS-NSC CDM

Methods

Data source

• NHIS-NSC (National Health Insurance Service-National Sample Cohort)
• 1 million persons’ 12 years (from 2002 to 2013) of claim data (about 2% sample of Korea population)
• NHIS-NSC includes:
  - Participants’ insurance eligibility
  - Medical records (including diagnosis, prescription, device, procedure)
  - Annual general health examination data
  - Medical cost

Results

Results of CDM Conversion

• The 1.13 million subjects was converted to OMOP-CDM, resulting in average 95.4% conversion rate.

Results of Measuring/Enhancing Anonymization

• A risk score for NHIS-NSC, measured by the Eclipse, was 0.140 and it was higher than the threshold (0.093).
• We strengthen the anonymization by using anonymization techniques of the Eclipse:
  - High suppression for all columns of quasi-identifier risk
  - Date-shift for the medical date group (with maintaining date interval for all medical date)
• Anonymization enhancement decreased to 0.025 below the threshold.

Results of PoC Study

• The PoC study which compares hypoglycemia risk between metformin and sulfonylurea before and after anonymization.
• We found that statistical attributes were retained after anonymization enhancement.

Conclusions

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.

• 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 : Santa Borel, and Khaled El Emam are employees at Privacy Analytics.
• Contact : Rae Woong Park (veritas@ajou.ac.kr)