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

Seongwon Lee, PhD[1]; Seng Chan You, MD[1]; Jimyung Park[2]; Jaehyeong Cho[2]; Santa Borel, MSc[3]; Khaled El Emam, PhD[3,4]; Rae Woong Park, MD PhD[1,2,5]

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

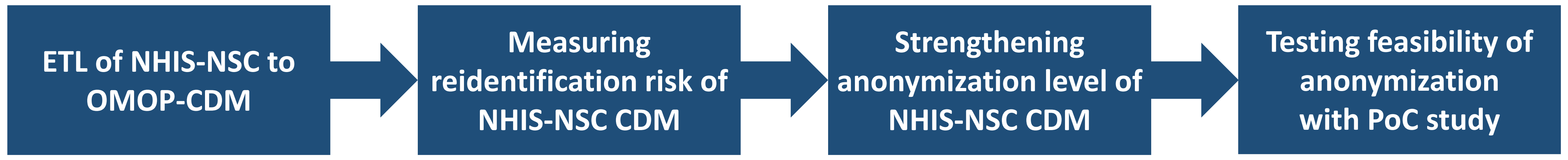

ETL of NHIS-NSC to OMOP-CDM → Measuring reidentification risk of NHIS-NSC CDM → Strengthening anonymization level of NHIS-NSC CDM → Testing feasibility of anonymization with PoC study

Figure 1. The overall process of research

ETL (Extract, Transform, Load)

- NHIS-NSC was converted OMOP-CDM v5.3.1.
- The ETL process was proceeded with the works of defining ETL rule, developing SQL scripts (for MS-SQL), executing the ACHILLES for quality check and packaging with R.
- All details are available at github: https://github.com/OHDSI/ETL---Korean-NSC

Anonymization Enhancement

- The Eclipse (version 2.11, Privacy Analytics, Canada), an automatic privacy-preserving software was used for measuring re-identification risk and strengthen the anonymization level of NHIS-NSC CDM.
- The Eclipse provides various anonymization techniques such as masking, generalization, suppression, and date shifting.
- We executed the Eclipse for the Proof-of-Concept (PoC) study's cohort, metformin or sulfonylurea-prescribed-patients.

Proof-of-Concept Study

- To validate the feasibility of anonymization enhancement, a PoC study was performed at before and after stronger anonymization.
- The PoC study is to compare hypoglycemia risk between metformin and sulfonylurea.

Results

Results of CDM Conversion

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

CDM Tables	Record count, n		Conversion rate (%)	Mapping coverage (%)
	NHIS-NSC	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
MEASUREMENT	33,440,451	33,440,451	100.00	100.00
OBSERVATION	33,218,703	33, 218,703	100.00	100.00
COST	908,678,310	609,571,436	67.08	Not applicable

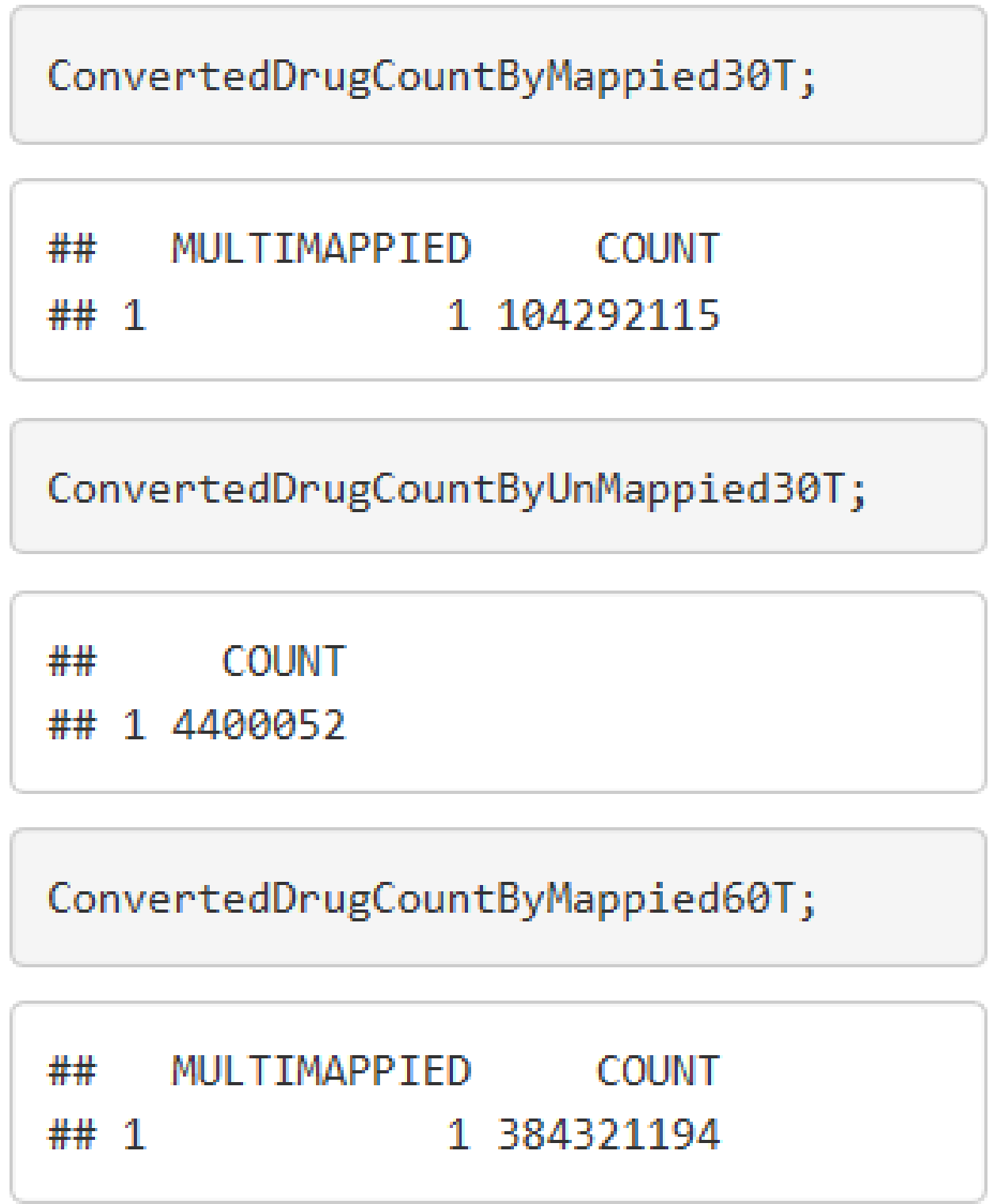
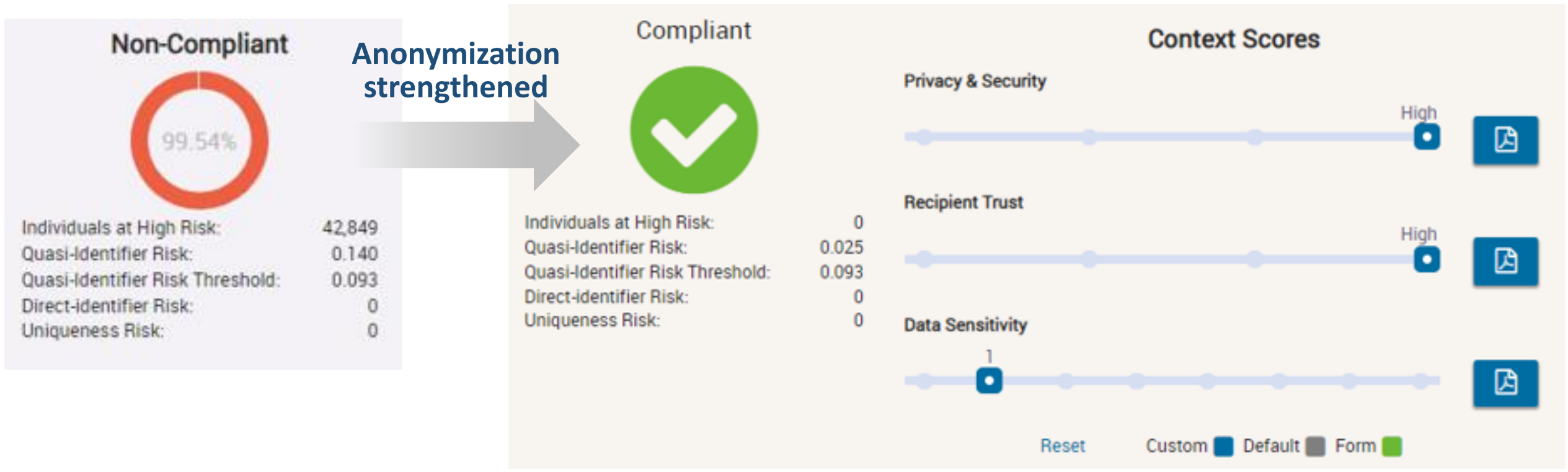

```
ConvertedDrugCountByMappied30T;

##    MULTIMAPPIED     COUNT
## 1             1 104292115

ConvertedDrugCountByUnMappied30T;

##      COUNT
## 1 4400052

ConvertedDrugCountByMappied60T;

##    MULTIMAPPIED     COUNT
## 1             1 384321194
```

Figure 2. Sample of ETL results report, developed by R markdown

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.



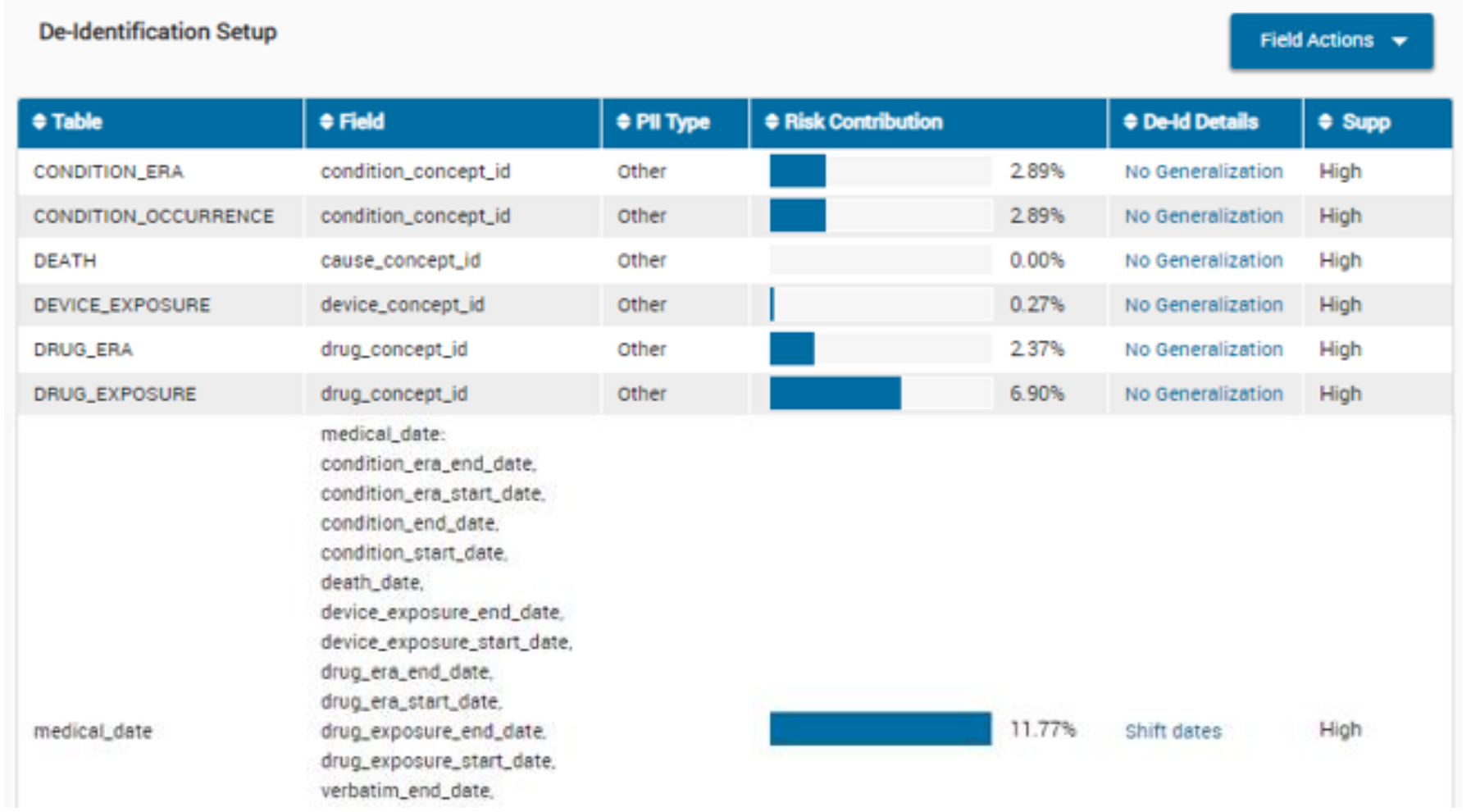
Non-Compliant

Individuals at High Risk:	42,849
Quasi-Identifier Risk:	0.140
Quasi-Identifier Risk Threshold:	0.093
Direct-identifier Risk:	0
Uniqueness Risk:	0

Compliant

Individuals at High Risk:	0
Quasi-Identifier Risk:	0.025
Quasi-Identifier Risk Threshold:	0.093
Direct-identifier Risk:	0
Uniqueness Risk:	0

Figure 3. Risk score before and after anonymization



De-Identification Setup

Table	Field	PII Type	Risk Contribution	De-Id Details	Supp
CONDITION_ERA	condition_concept_id	Other	2.89%	No Generalization	High
CONDITION_OCCURRENCE	condition_concept_id	Other	2.89%	No Generalization	High
DEATH	cause_concept_id	Other	0.00%	No Generalization	High
DEVICE_EXPOSURE	device_concept_id	Other	0.27%	No Generalization	High
DRUG_ERA	drug_concept_id	Other	2.37%	No Generalization	High
DRUG_EXPOSURE	drug_concept_id	Other	6.90%	No Generalization	High
medical_date	medical_date: condition_era_end_date, condition_era_start_date, condition_end_date, condition_start_date, death_date, device_exposure_end_date, device_exposure_start_date, drug_era_end_date, drug_era_start_date, drug_exposure_end_date, drug_exposure_start_date, verbatim_end_date,		11.77%	Shift dates	High

Figure 4. Example of anonymization setup in the Eclipse

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**.

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	(0.35 to 0.69)	0.00

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)