Name:	Vaclav Papez
Affiliation:	Institute of Health Informatics, UCL, London
Email:	vpapez@kiv.zcu.cz
Presentation type (select one):	Poster

Evaluating the transformation of UK national linked electronic health records to the OMOP CDM

Vaclav Papez, PhD[1], Maxim Moinat, MSc[2], Richard Dobson, Prof[1], Folkert Asselbergs, Prof[1], Spiros Denaxas, Prof[1]
[1]Institute of Health Informatics, University College London, London, United Kingdom;
[2]The Hyve, Utrecht, Netherlands

Abstract

Given by an increasing trend in adopting Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) in Europe for observational research, OMOP CDM has become a main harmonization platform for diverse data sources provided by countries involved in running project Innovative Medicine Initiative (IMI) BigData@Heart. CALIBER research platform containing structured linked electronic health records from three national sources (primary care, hospital care and mortal registry) is one of the participating data resources. Main challenge was to preserve CALIBER's ability to implement disease phenotypes defined across all presented data sources as these differ in their data structures as well as terminologies used. The aim of this study is to evaluate the quality and consistency of a transformation process from CALIBER to OMOP CDM from both syntactic and semantic perspective.

Introduction

CALIBER is a platform[1] consisting of linked electronic health records (EHR) from three diverse national data sources: Clinical Practice Research Datalink (CPRD) primary care data, Hospital Episode Statistics (HES) hospital data and Office for National Statistics (ONS) mortality and socioeconomic data. CALIBER also implements disease phenotypes, clinically agreed and validated codes using specific terminologies to describe diseases in EHRs. Native encodings for diagnostic/procedure/drug codes used for these phenotype definitions are Read codes, CPRD product codes and CPRD entity types for CPRD data, International Classification of Diseases 10th revision (ICD-10) and OPCS Classification of Interventions and Procedures version 4 (OPCS-4) codes for HES data and ICD-9 and ICD-10 codes for ONS data. In comparison with other studies focusing to a single data sources[2, 3] our study evaluates a transformation of all three data sources at once. For a transformation we used a subset of CALIBER data containing patients diagnosed with heart failure (HF).

Methods

We designed an Extract Transform Load (ETL) process based on existing and validated mappings consisted of syntactic mapping where data from 20 source tables were mapped onto 14 clinical data tables of CDM version 5.2[4] and semantic mapping translating source codes into vocabularies supported by OMOP CMD (Table 1). Internal OMOP CDM representation of all used codes (source and target) is in the form of unique identifier across all terminologies used, so called concepts. ETL process was executed over data extracted from all 20 source tables for a cohort of 502,723 patients identified with incident of HF. Testing strategy consists of direct querying into CALIBER and OMOP CDM databases and comparing retrieved numbers.

Table 1. Mapping of source (CALIBER) to target (OMOP CDM) vocabularies.

Source vocabulary	Intermediate mapping	Target vocabulary
Read / ICD10 / ICD9 / OPCS4	native	SNOMED-CT
CPRD Product	gemscript,DM+D	RxNorm
CPRD Entity Type	JNJ_CPRD_ET_LOINC[5]	LOINC
CPRD Units	native	UCUM

This study was approved by the Medicines and Healthcare Products Regulatory Agency Independent Scientific Advisory Committee (protocol reference: 17_015R).

Results

We converted 1,099,195,384 rows of data in total. 356 patients were lost due to the validity of an observation period window. All data identified data losses were caused by quality of source data or by incomplete mapping (Table 2 – mapping coverage).

Table 2. Mapping coverage for disease and drug clinical terminologies used (ET – Entity Type)

	Used unique terms	Used mapped terms (%)	Total unique events	Total excluded events (%)	Total mapped events (%)
Read	67 886	97.58	320328788	0.22	97.42
ICD-9	495	100	13130	0.92	100
ICD-10	10158	88.53	31905144	0.01	99.09
OPCS-4	8474	99.45	8453813	0	99.88
Drugs	40647	62.53	264589509	1	92.67
Units	22	72.72	27036	1.55	99.95
ET – Lab. results	245	54.28	125581411	0.59	54.06
ET - Test	324	97.22	151645201	12.24	98.16

Conclusion

Structural as well as syntactic mapping was successfully evaluated from the perspective of mapping coverage. Evaluation of data consistency for disease phenotypes application is in progress.

References

1. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc. 2019;26: 1545–1559.
2. Matcho A, Ryan P, Fife D, Reich C. Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. Drug Saf. 2014;37: 945–959.
3. Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug Saf. 2013;36: 119–134.
4. https://github.com/OHDSI/CommonDataModel/releases/tag/v5.2.0
5. https://github.com/OHDSI/ETL-CDMBuild