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

Vaclav Papez, PhD[1*]
Maxim Moinat, MSc[2]**

*v.papez@ucl.ac.uk,**maxim@thehyve.nl

BACKGROUND: CALIBER research platform links electronic health records (EHR) from Clinical Practice Research Datalink (CPRD) primary care data, Hospital Episode Statistics (HES) hospital data and Office for National Statistics (ONS) mortality and socioeconomic data. Disease phenotypes, also implemented in CALIBER, are clinically agreed and validated diagnostic/procedure/drug codes using specific terminologies (Table 1) to describe diseases in EHRs. Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) serves as a main harmonization platform between diverse data source involved in BigData@Heart project including CALIBER. This study evaluates syntactic as well as semantic transformation of all CALIBER data sources into OMOP CDM.

METHODS: We designed an Extract Transform Load (ETL) process based on existing validated mappings and consisted of syntactic mapping where data from 20 source tables were mapped onto 14 clinical data tables of CDM version 5.2 and semantic mapping translating source codes into vocabularies supported by OMOP CMD. Cohort of 502,723 patients identified with incident of heart failure (Table 3) was used in ETL process. Testing strategy consisted of direct querying into CALIBER and OMOP CDM databases and comparing retrieved numbers (Figure 1 and Table 2,3). Queries implemented clinically agreed disease phenotype algorithms.

RESULTS: We converted >1B rows of data and mapped ~110K source codes from 5 dictionaries. Mean term mapping coverage is ~98%. 356 patients were lost due to the validity of an observation period window (Table 3). All identified data losses were caused by quality of source data or by imprecise mapping (Figure 2).

A successful **structural and syntactical mapping** to the **OMOP CDM**, including validation of the mapping coverage

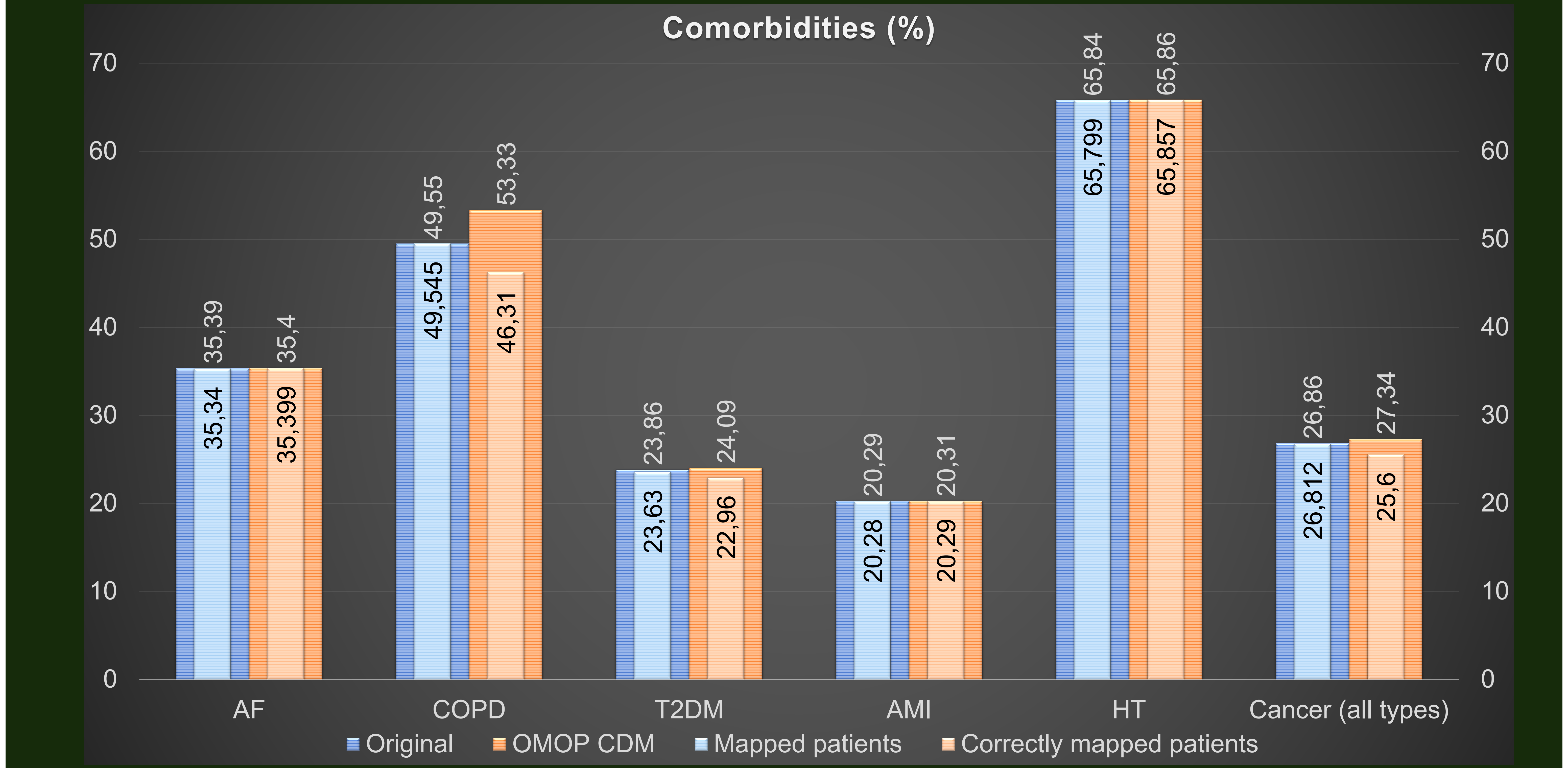

Figure 1. Comorbidity metrics comparison of original CALIBER dataset and OMOP CDM conversion. The figure shows small differences, but overall in agreement.

Take a picture to **download** the **abstract**

Table 1. Validated mappings between source (CALIBER) and target (OMOP CDM) vocabularies.

Source vocabulary	Intermediate mapping	Target vocabulary
Read / ICD-10 / ICD-9 / OPCS-4	native	SNOMED-CT
CPRD Product	gemscript, DM+D	RxNorm
CPRD Entity Type	JNJ_CPRD_ET_LOINC	LOIN
CPRD Units	native	UCUM

Table 2. Mapping coverage for disease and drug clinical terminologies used in CPRD, HES and ONS and converted to the OMOP CDM Standard dictionary (entire cohort).

Terminology	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

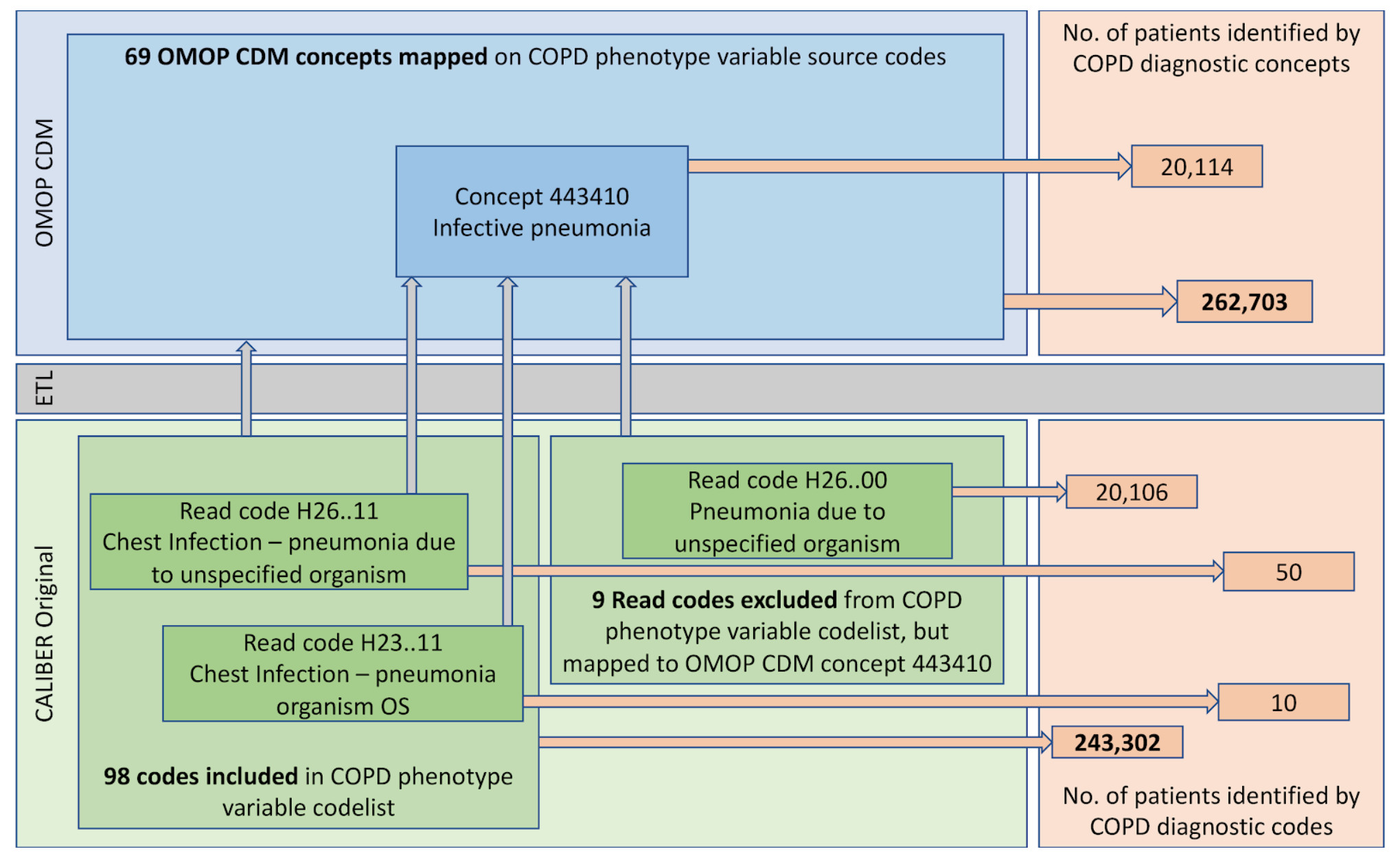
OMOP CDM
69 OMOP CDM concepts mapped on COPD phenotype variable source codes
Concept 443410
Infective pneumonia
No. of patients identified by COPD diagnostic concepts
20,114
262,703
ETL
CALIBER Original
Read code H26..11
Chest Infection – pneumonia due to unspecified organism
Read code H23..11
Chest Infection – pneumonia organism OS
98 codes included in COPD phenotype variable codelist
Read code H26..00
Pneumonia due to unspecified organism
9 Read codes excluded from COPD phenotype variable codelist, but mapped to OMOP CDM concept 443410
20,106
50
10
243,302
No. of patients identified by COPD diagnostic codes

Figure 2. Example of origin of inconsistency between source and target data demonstrated on COPD phenotype variable. Multiple source codes (Read codes in green boxes) are mapped onto the same OMOP CDM target concept (blue box), however not all of these source codes are part of the examined codelist (COPD). Thus, patient counts (orange boxes) based on the codelists retrieved from original CALIBER (243,302) does not match counts based on mapped concept lists retrieved from OMOP CDM CALIBER transformation (262,703). Main result difference is caused by the Read code H26..00 found in more that 20k patients, which is excluded from COPD phenotype variable, but mapped to the same concept of Infective pneumonia as other codes from the source codelist.

Table 3. Cohort summary and comparison of main metrics (subpopulations of used heart failure cohort serves as validation metrics) between the raw data and data transformed to the OMOP CDM. 356 persons are lost in the transformation due to an invalid observation period. The other metrics are comparable.

Metric	CALIBER (raw)	CALIBER (OMOP CDM)
Number of persons	502,723	502,367
Median follow up (IQR)	9.56 (10.39)	9.56 (10.39)
Demographics		
Female (%)	52.39	52.4
Caucasian (%)	90.81	90.46
Most deprived fifth (%)	15.18	15.18

Vaclav Papez, PhD[1*]
Maxim Moinat, MSc[2]**
Stefan Payrable, MSc[2]
Richard Dobson, Prof[1]
Folkert Asselbergs, Prof[1]
Spiros Denaxas, Prof[1]

[1] UCL
[2] the hyve