Repeatable ETL process to transform Estonian health data to OMOP CDM

Conversion of Estonian health data into the OMOP CDM: *insurance claims, prescription data and electronic health records*

PRESENTER: **Marek Oja**

INTRO:

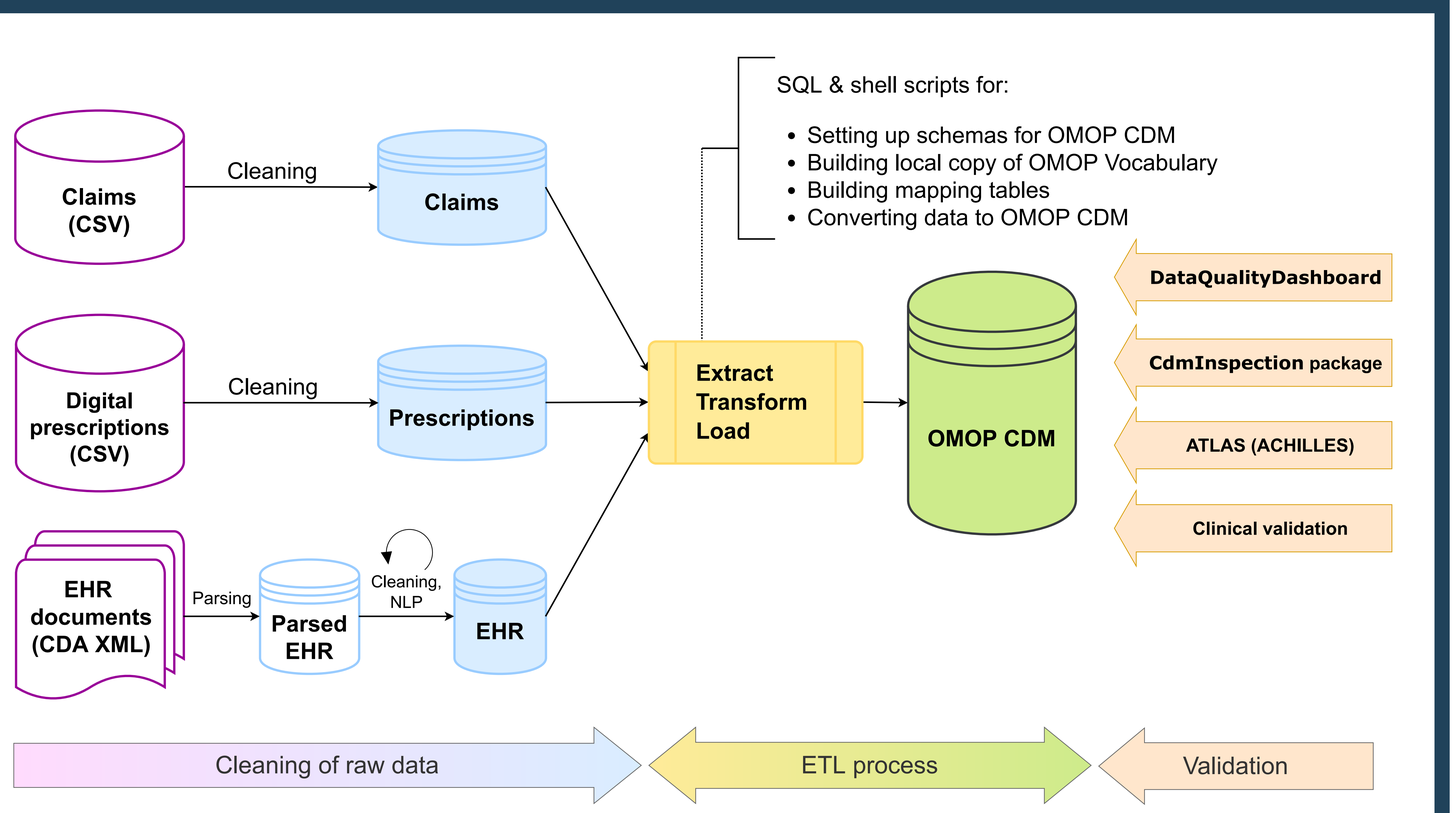
- Estonia needs a research database where all health data is standardized and ready to use for observational research.

METHODS

- 10% random sample of the Estonian population (n=149K patients) from 2012 to 2019
- Dataset included three national data sources:
  - insurance claims (n=6.2M)
  - digital prescriptions (n=9.6M)
  - electronic health records (EHR) (n=4.97M)
- OMOP CDM v5.3
- Technologies used:
  - Git for version control
  - PostgreSQL for database
  - Python (luigi), SQL, bash for ETL pipelines
  - EstNLTK for NLP
- Translation of local vocabularies to standard vocabularies

RESULTS

- All three different data sources were combined successfully into one OMOP CDM. With this, we have a full view of patient data over the observation period.
- Process is repeatable and used for different datasets and projects in Estonia:
  - Asthma specific dataset
  - COVID specific dataset
  - Estonian Biobank health data
- Participation in network studies:
  - Prostate cancer study - PIONEER



Statistics on the ETL procedure to convert Estonian health data to OMOP CDM

Table	Total number of records	Mapped records	Mapping rate	Number of source codes	Number of codes mapped	Mapping rate
location	1					
care_site	1,820					
person	149,351					
death	8,277					
observation_period	149,351					
visit_occurrence	18,194,512	18,194,512	100.0%	10	10	100.0%
visit_details	48,002	48,002	100.0%	8	8	100.0%
condition_occurrence	20,238,707	20,220,877	99.9%	9,240	9,215	99.7%
procedure_occurrence	6,950,829	5,386,453	77.5%	5,999	1,506	25.1%
drug_exposure	8,596,491	8,151,931	94.8%	1,153	915	79.4%
device_exposure	77,842	47,296	60.8%	159	4	2.5%
observation	15,159,794	14,721,482	97.1%	755	548	72.6%
measurement	30,440,242	27,457,547	90.2%	3,524	2,904	82.4%
measurement-value	1,544,690	1,544,690	100.0%	49	49	100.0%
measurement-unit	11,193,028	11,193,028	100.0%	79	79	100.0%
drug_era	4,313,993	4,313,993	100.0%	774	774	100.0%
condition_era	9,347,695	9,347,695	100.0%	6,401	6,401	100.0%

Source vocabulary mappings to standardized vocabularies

Source vocabulary	Target vocabulary	Count	Percent
ATC	RxNorm	5,111	87.7%
ATC	RxNorm Extension	368	6.3%
ATC (mostly combination drugs)	0	352	6.0%
Cancer related findings (TNM codes, cancer stages, etc.)	Cancer Modifier	9	8.0%
Cancer related findings (TNM codes, cancer stages, etc.)	NCIt	36	32.1%
Cancer related findings (TNM codes, cancer stages, etc.)	SNOMED	67	59.8%
Pathology findings, body measurements	SNOMED	18	100.0%
Drug administration routes	SNOMED	87	100.0%
Local codes from claims (procedures, drugs, measurements, etc)	LOINC	32	0.9%
Local codes from claims (procedures, drugs, measurements, etc)	OMOP Extension	3	0.1%
Local codes from claims (procedures, drugs, measurements, etc)	RxNorm	85	2.5%
Local codes from claims (procedures, drugs, measurements, etc)	RxNorm Extension	14	0.4%
Local codes from claims (procedures, drugs, measurements, etc)	SNOMED	1,188	34.6%
Local codes from claims (procedures, drugs, measurements, etc)	0	2,110	61.5%
ICD10	Cancer Modifier	21	0.1%
ICD10	OMOP Extension	3	0.0%
ICD10	SNOMED	20,016	99.9%
ICD10	0	1	0.0%
LOINC	LOINC	81,452	97.9%
LOINC	SNOMED	731	0.9%
LOINC (mostly local and temporary LOINC codes)	0	1,019	1.2%
NOMESCO Classification of Surgical Procedures (NCSP)	RxNorm	2	0.0%
NOMESCO Classification of Surgical Procedures (NCSP)	SNOMED	728	9.8%
NOMESCO Classification of Surgical Procedures (NCSP)	0	6,671	90.1%
Measurement units	UCUM	922	100.0%

Marek Oja[1], Sirli Tamm[1], Sulev Reisberg[1,2,3], Raivo Kolde[1], Sven Laur[1], Hendrik Šuvalov[1], Harry-Anton Talvik[1,2,3], Jaak Vilo[1,2]

[1] Institute of Computer Science (ICS), University of Tartu, Tartu, Estonia
[2] STACC, Tartu, Estonia
[3] Quretec, Tartu, Estonia

Contact: marek.oja@ut.ee

Acknowledgements This work was supported by the Estonian Research Council grants number PRG1095, RITA1/02-96; the European Union through the European Regional Development Fund grant number EU48684; and the European Social Fund via IT Academy programme. The whole conversion was carried out in the High Performance Computing Center of the University of Tartu.