Name:	Maxim Moinat
Affiliation:	The Hyve
Email:	maxim@thehyve.nl
Presentation type (select one):	Lightning Talk

Transforming Electronic Health Records from Swedish registers to the OMOP CDM

Maxim Moinat, MSc, Marinel Cavelaars, MSc PhD, Kees van Bochove, MSc Bed
The Hyve, Utrecht, Netherlands; Bayer Pharmaceuticals

Abstract

The Nordic countries have one of the best centralized Electronic Health Records (EHR) in Europe. At The Hyve we created an Extract Transform Load procedure to transform Swedish inpatient, outpatient, drug, death and population registers into the OMOP[1] Common Data Model. This ETL procedure was executed remotely for an Atrial Fibrillation study with about 450k patients. Multiple Nordic healthcare coding systems, including a Swedish version of ICD10, Swedish drug product codes and the NOMESCO[2] coding system, have been translated partly to the OMOP standard vocabulary.

Data structure

The source data consisted of four tables; a table with patient care (inpatient and outpatient), a drug registry, death registry and a social-economic registry. The full Swedish social-economic registry, called LISA, contains many more variables of which a subset was selected for this study.

The source data structure had three major differences compared to the OMOP CDM: a registry with unique patients was missing, each row contained up to 21 diagnoses and 30 procedures and multiple causes of death with additional circumstances were present.

Table 1. Tables and fields in the Swedish EHR source data.

Patient Care	Drug	Death	Socio-economic
70 fields (3 tables)	*48 fields*	*76 fields*	*10 fields*
Patient details (id, gender, age, civil status, emigration)	Patient details (id, gender, age)	Patient details (id, gender, age)	Patient details (id, civil status)
Diagnosis codes (up to 21 per row)	Drug details (name, ATC, dose form, strength)	Cause of death (up to 48 per row)	Income
Procedure codes (up to 30 per row)	Packsize, amount, instructions	Circumstances (alcohol, work related, narcotics, …)	Education
Date of care (in/out)	Date of prescription/dispense	Date of death	Employment
Place of care	Place of prescription	Place of death	
	Provider specialty		

A table of unique patients was created by selecting all unique patient ids form the patient, drug and death registries. Conflicting genders and ages were solved by choosing the most frequent value.

Multiple causes of death were stored in an addendum table. This table also contained additional circumstances, like work or alcohol related deaths.

The multiple diagnoses and procedures per row were split to load them separately into the OMOP CDM. To conserve the position of the diagnosis, we used 'Primary Condition', 'Secondary Condition' and 'First Position Condition'. The latter was used for all the third and higher diagnoses, which resulted in a slight loss of information.

Concept mapping

Four source vocabularies were mapped to the OMOP standard vocabulary, see Table 2. Firstly, the ICD10-SE and drugs could be mapped automatically. The ~33k unique ICD10-SE codes had a large overlap with the international ICD10 codes. Either the first five, four or three code characters were mapped to a ICD10 equivalent. The drug concepts were mapped using ATC code, drug strength and drug form. Using this information, most concepts could be mapped to RxNorm clinical drug concepts. For both the ICD10-SE and RxNorm codes, an assessment of the quality of the mapping still needs to be performed.

Secondly, a subset of NOMESCO codes was mapped manually to the standard vocabulary. The full NOMESCO vocabulary contained 6012 chirurgical and non-chirurgical procedures and this vocabulary did not exist in the OMOP vocabulary. At the time of writing, 104 NOMESCO codes were mapped. Lastly, the 163 provider specialties were all manually mapped.

Besides as a database export (source to concept map), the Usagi data file is also provided in the repository. The Usagi file allows for easy viewing and updating the mappings.

To assess the mapping quality, we developed an extension of Achilles. This extension provides the mapping coverage per concept class for the loaded dataset and the most frequent mapped and unmapped concepts.

Table 1. Vocabularies mapped to the OMOP standard vocabulary

Concept Type	Source Vocabulary	Target Vocabulary	Mapping type
Condition, Death cause	ICD10-SE	SNOMED	automatic
Procedure	NOMESCO	SNOMED	manual
Drug	ATC	RxNorm	automatic
Provider Specialty	*Local codes*	Specialty	manual

Conclusion

We successfully developed an ETL procedure to transform Swedish EHR data to the OMOP CDM. The source code is available on Github (https://github.com/thehyve/ohdsi-etl-sweden). In the future, the concept mapping still needs to be manually curated, especially the NOMESCO mapping.

References

1. Observational Medical Outcomes Partnership [Internet]. OMOP. 2016 [cited 22 June 2016]. Available from: http://omop.org/
2. Nordic Medico-Statistical Committee (NOMESCO) 2011; NOMESCO Classification of Surgical Procedure; http://www.socialstyrelsen.se/klassificeringochkoder/atgardskoderkva