Four Complexities when mapping NCRAS to the OMOP CDM

Laura Kerr, Abigail Carter
Genomics England

INTRODUCTION

Genomics England (GEL) is a global leader in enabling genomic medicine and research, focused on creating a world where everyone benefits from genomic healthcare.

- It enriches its primary clinical data (participant information) with secondary data (supporting healthcare records) which includes the National Cancer Registration and Analysis (NCRAS) data.
- The data is made available to researchers in an isolated Research Environment.
- There is a desire from researchers to run federated queries, however this is hampered by the disparate sources of data and their differing data models.
- GEL have therefore mapped NCRAS to the OMOP CDM to empower researchers and have made the mappings publicly available.

MAPPING APPROACH

1. Retrieve source attributes and enumerations from NCRAS data dictionary.
2. Identify OMOP domain best suited to source attribute.
3. Manually map source attributes to concepts by matching on descriptions.
4. Standard attributes are used wherever possible.

COMPLEXITIES

1. MAPPING GRANULARITY

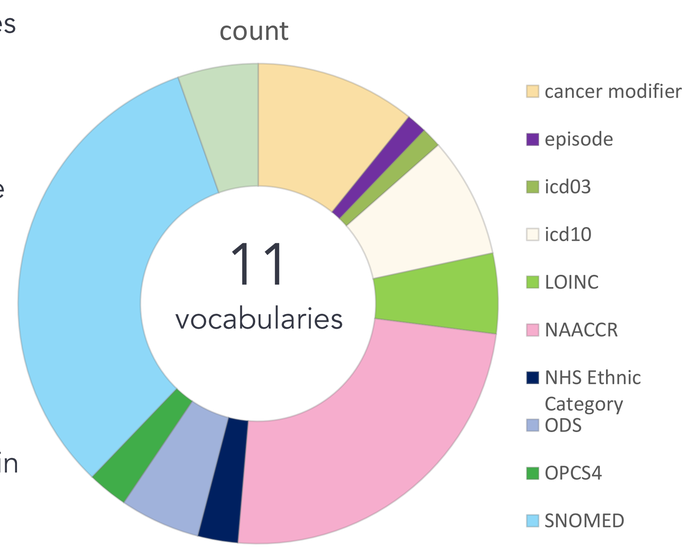
- Opted to always map to the most granular concept that wouldn't lead to researchers inferring information that wasn't present in the source data.
- Utilised the CDM to provide context where desired levels of granularity have not been possible.
- For example, the most granular match found for the source enumeration shown below is 'Index of Multiple Deprivation (England)'. The value_as_string' field, has then been used to provide more detail.

Source Attribute	Source Enumeration	OMOP Domain Fields	Field Value
quintile2004	Deprivation Quintile 2004. 1 least deprived	observation_concept_id	35812882
		value_as_string	Deprivation Quintile 2004. 1 – least deprived

- GEL is interested in defining a new vocabulary to support these niche cases.

2. VOCABULARIES

- The NCRAS dataset describes a broad scope of events.
- The use of a broad range of vocabularies to represent the data reflects this.
- GEL are interested in understanding how researchers will interact with the diversity of vocabularies in the future.



3. CLINICAL CODES

- The NCRAS data is a collection of secondary data sources.
- Clinical coded source attributes in cases therefore vary in format.
- Deviations in format need reformatting to maximise mapping quality. A simple example is shown below.

Source Attribute	Source Value	Target Value
site_icd10_02	C049	C04.9

- A benefit of mapping the data to OMOP CDM is the identification of coded entries that require such reformatting, and their target format.

4. ONCOLOGY EXTENSION

- Used the Oncology extension on GEL mental health datasets as the structure is ideal for grouping many different episodes of care.
- More data is required to use the NCRAS dataset to populate episodes of care and patient pathway domains.
- GEL do not wish to infer information that we do not have so have not used the oncology extension here.

CONCLUSION

The OMOP mapping gave a high success rate, with almost all clinical information being mapping to the OMOP CDM. GEL expects the mapped NCRAS data to vastly improve the user experience in our research environment.

It is recognised that there is scope for improvement in the mappings and feedback on the mappings is very welcome as GEL hope to iteratively improve their quality and depth.

SCAN ME

https://gitlab.com/genomicsengland/genomics_england_publications/public-omop-mappings

Laura Kerr
Data Manager
laura.kerr@genomicsengland.co.uk

Abigail Carter
Data Wrangler
abigail.carter@genomicsengland.co.uk