Challenges in harmonising data across multiple biobanks

Karyn Mégy[1], Rebecca Akhanemhe[1], Ben Hollis[1], Ali Abbasi[1], Amanda O'Neill[1], Shikta Das[1], Stewart MacArthur[1], Sean O'Dell[1], Sebastian Wasilewski[1], Quanli Wang[2], Slavé Petrovski[1], Jen Harrow[1].

1. Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK. 2. Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Waltham, MA, USA

1. Background

Early-stage incorporation of human genomic data into the assessment of drug targets has been shown to significantly increase drug pipeline success rates.

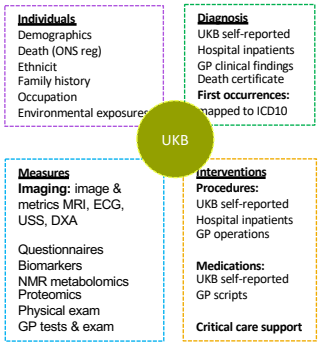
Large biobanks such as UK Biobank, combining genetic and clinical data on 0.5 million individuals, offer an unprecedented opportunity to evaluate effects of genetic variants on a broad collection of traits.

Statistical power, however, comes from the size as well as the ethnic diversity of those biobanks.

AstraZeneca's Centre for Genomics Research is establishing one of the world most comprehensive and diverse genetic resource, combining genetic and phenotypic data for multiple biobanks. **This work describes the challenges faced when harmonizing such datasets,** enabling their cross analysis.

Different data sets, several data types, multiple standards

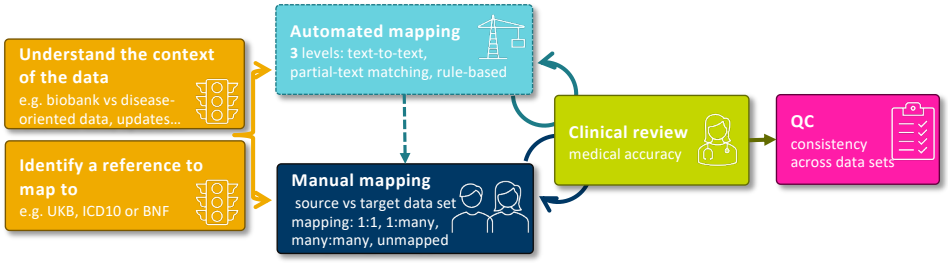
UK Biobank (UKB) is one of the golden standard, in terms of data diversity, sources but also coding systems. However, it is very reflective of the UK population and health care system. Biobanks from different countries will be using different coding system, different units (e.g. *Hba1c: mmol/mol vs. %*), medication names (e.g. *metformin vs. metformina*), and the local language, making comparison of those data sets challenging.

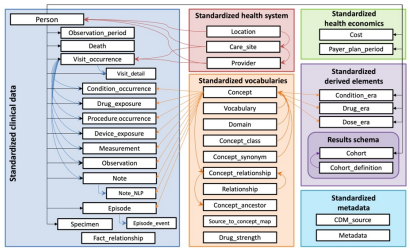


	Source of health data available in our cohorts					And also….
	Hospital data	Primary care	Cancer data	Questions	Free text	
UK Biobank	WHO ICD9 & 10	Read2 & 3	yes	formatted	-	Lab, proc.
US cohort #1	CM ICD9 & 10	CM ICD9 & 10	CM ICD9 & 10	-	-	Lab, proc.
UK cohort #1	WHO ICD9 & 10	Read2 & 3	yes	-	-	Lab, proc.
FinnGen	WHO ICD8, 9 & 10					
MCPS	-	-	WHO ICD10	WHO ICD10	yes	Lab.

ICD: International Classification of Diseases, versions 8, 9 and 10, in the WHO or US (CM) system
Snomed: Systematized Nomenclature of Medicine Clinical Terms ; **Lab** & **proc.**: laboratory results and procedures

2. Harmonisation process





Understand the data | Mapping strategies | Review | QC | *If relevant,* mapped data integrated into OMOP CDM

3. Results

E.g.1: mapping a small disease-oriented data set to a large biobank

We mapped one of our data sets, a small disease-oriented resource, to UK Biobank. An initial mapping was done manually, following the process described above, high-quality but timely. As a test, we then performed NLP (Natural Language Processing) on that same data set, reducing the mapping time from months to a week, however <50% of terms could be mapped.

=> NLP followed by manual mapping would be most efficient strategy in term of time and accuracy.

The final mapping will be transformed into the **OMOP common data model**

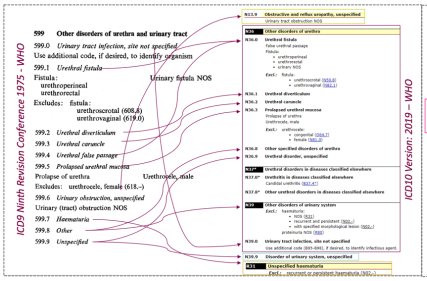
Approach	Description	Example
Text-to-text match	Exact Text match	'Age' = 'Age at recruitment'
Partial text match	**Step 1:** Sub-word matching, part of substring matches Field ID **Step 2:** Sub-word matching and matching all words present in column	'Supplement' = 'Supplements' 'weight loss' = 'loss in weight'
Rule-based match	Identify key terms and match to UKB Field ID **Layer 1:** Age or family history-based fields can be identified using 'age', 'fh', '<brother/mother>' terms **Layer 2:** Diseases and symptoms can be identified using ontologies and classified as cancer vs non-cancer **Layer 3:** Category and Question information to identify terms like self-reported, subsection etc	20002 (self-reported non-cancer)

E.g.2: mapping across ontologies, from UKB ICD9 to ICD10

UK Biobank diagnoses are encoded both in the ICD9 and in ICD10 WHO classifications. Following our harmonisation process, we have mapped the ICD9 terms present in the UKB data set to ICD10 and, according to the FAIR principles, are returning the results to UK Biobank so that they can be **shared with the community** (*manuscript in preparation*).

⇒ In total, we mapped 751 ICD9 codes to 573 unique ICD10 codes, with 85% having a 1:1 mapping.

Challenge #1
"not elsewhere classified"



Challenge #2
category restructuration

E.g. accidents:

- ICD9: group by type of injuries

Sprains and strains of joints and adjacent muscles (840-848)

- ICD10: group by body site affected

-> across S03; S13; S23; S33; S43; S53; S63; S73; S83; S93

4. Take home messages

- Harmonising data sets require to understand the data and adapt the strategy when needed.
- NLP followed by manual mapping is be most efficient strategy for mapping across cohorts
- Importance of return of data for use by the community (FAIR principle)
- Gap in OMOP: mapping of images & medications

Acknowledgements - We would like to thank the **participants and investigators in the UK Biobank** study who made this work possible.
We also acknowledge contribution from members of the **AstraZeneca Genomics Initiative** and the AZ IT Knowledge EngineeringTeam.

AstraZeneca