COMPUTABLE BIOMEDICAL KNOWLEDGE AND THE LANGUAGE OF DISCOVERY: A CROSS-REPOSITORY METADATA EVALUATION

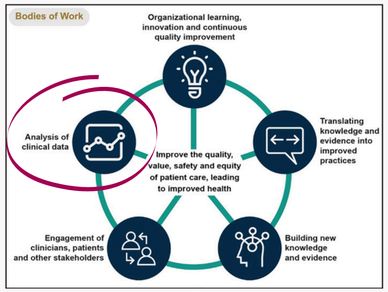
Annabel Pinkney, MSLIS; Hanieh Razzaghi, PhD, MPH
PEDSnet Learning Health System
Applied Clinical Research Center, Children's Hospital of Philadelphia

This research was developed through the training and support provided by the MLA Research Training Institute (RTI).

What barriers and opportunities are present for relating computable biomedical knowledge objects across repositories through metadata; how might standard vocabularies and unified domain language support interoperability?

BACKGROUND

- Patient cohorts in real world data analysis (RWD) are defined by computable, machine-executable definitions, or computable phenotypes.
- Access to well-described, reusable computable phenotypes, can reduce redundancy and improve reproducibility in RWD research. Therefore, **systems to store and share phenotypes in accordance with FAIR principles is crucial.**
- Despite broad consensus on the importance of sharing cohort-defining CBK, a key barrier to interoperability remains: the absence of a shared descriptive language.
- Phenotype and codeset repositories often operate in isolation and apply heterogeneous terminology and metadata, even for similar artifacts. **A lack of standardized vocabularies limits discovery, comparison, and reuse across systems.**



Learning Health Systems Consolidated Framework (LHS-CF) Easterling et al. Learn Health Syst. 2021 Jul 22;6(2):e10287. doi: 10.1002/lrh2.10287

This figure illustrates a Learning Health System, in which RWD research informs clinical decision making. This investigation aims to improve the "Analysis of Clinical Data" portion of the cycle.

METHODS

- 7 repositories were identified for analysis based on the inclusion criteria:
  1. Repository contains cohort-defining CBK artifacts (specifically phenotypes).
  2. Repository is non-commercial and open access.
  3. Repository is accessible in English.
  4. Repository supports algorithmic cohort definition using electronic health record (EHR) data. Repositories lacking executable or reproducible cohort definitions were excluded. (E.g., dbGaP excluded).
- Metadata elements were extracted and controlled fields were identified.
- Elements were grouped into "knowledge classes" based on an adapted version of Alpers et al.'s metadata categories for CBK. Inter-rater reliability testing (Cohen's κ and Jaccard similarity) was used to validate classification consistency.
- Within each knowledge class, term lists were normalized and analyzed for lexical similarity.
- Then, semantic similarity of the terms in each knowledge class was assessed through mappings of synonymous and related terms.
- Findings will inform development of a prototype taxonomy.

SAMPLE OF DATA, Knowledge Class: Ontological Components

REPOSITORY	FIELD	ORIGINAL TERM	NORMALIZED TERM	SEMANTIC MAPPING
CIPHER	Algorithm Components	Healthcare Common Procedure Coding System (HCPCS)	Healthcare Common Procedure Coding System (HCPCS)	HCPCS
CIPHER	Algorithm Components	ICD-9 Diagnostic Codes	ICD-9 Diagnostic Codes	ICD-9
CIPHER	Algorithm Components	Observational Medical Outcomes Partnership (OMOP) Concept	Observational Medical Outcomes Partnership (OMOP) Concept	OMOP Value Set
PEDSpace	Algorithm Components	HCPCS	HCPCS	HCPCS
PEDSpace	Algorithm Components	ICD-10	ICD-10	ICD-10
PEDSpace	Algorithm Components	ICD-9	ICD-9	ICD-9
PEDSpace	Algorithm Components	OMOP Value Set	OMOP Value Set	OMOP Value Set
HDR UK	coding_system	ICD9 codes	ICD-9 Codes	ICD-9
HDR UK	coding_system	OMOP codes	OMOP Codes	OMOP Value Set
HDR UK	coding_system	ICD10 codes	ICD-10 Codes	ICD-10
PheKB	Phenotype Attributes	ICD 9 Codes	ICD-9 Codes	ICD-9
PheKB	Phenotype Attributes	ICD 10 Codes	ICD-10 Codes	ICD-10
PheKB	Structured Data	ICD10	ICD-10	ICD-10
PheKB	Structured Data	ICD9	ICD-9	ICD-9
4 Repositories	**5 Fields**	**15 Terms; No identical terms across repositories.**	**4 lexical pairs -> 11 unique terms**	**4 semantic groupings -> 4 unique terms**

CHALLENGES

- Many controlled fields found in the content analysis were actually attributes, or metadata describing the value of a field. These were excluded from analysis.
- Some fields were controlled to Yes/No responses. I equated these to controlled terms as in Yes = positive <field name>, no = negative <field name>.
- Semantic mapping was difficult to systematize.
- Instances of variation in classification granularity within the same knowledge class.
- Several repositories provide an option for user-supplied terms, effectively making the fields uncontrolled.
- Some identical terms came from the same repository.

RESULTS

Trends across repositories vary depending on the knowledge class. Some classes are highly standardized across the board, possessing identical terms, e.g., **Ontological Components (13% lexically identical)**, and **Administrative Status (21% lexically identical)**.

High semantic or conceptual overlap occur in **Validation (71% semantic overlap)** and **Development Population (41% semantic overlap)** knowledge classes.

Some classes are structurally inconsistent across repositories, such as Domain and Development Population terms, which varied in granularity.

28%
Average proportion of controlled fields in each metadata profile

48% of fields are mandatory

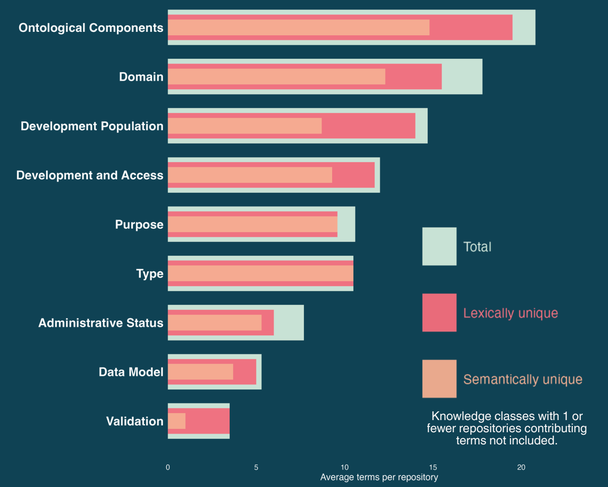
31% of fields are repeatable

50% of fields are faceted

Count of Repositories with a Controlled Field per Knowledge Class
Circle area proportional to the number of repositories controlling this information type



Term Agreement by Knowledge Class
Layered bars show reduction from total terms → lexically unique → semantically unique. Taxonomy, and repo-specific terms not included.



Repositories are aligned on the descriptive information necessary for computable phenotype discovery, but there is terminology discordance in the communication of this information.

These findings highlight the need for community standards for existing vocabularies and the development of new taxonomies for FAIR phenotype repositories.

NEXT STEPS

- Revisit some of the listed challenges to refine data analysis methods.
- Survey repository owners to assess involvement of MLIS credentialed professionals in metadata profile development.
- Build out a taxonomy for cohort-defining knowledge object descriptors based on the needs identified from this research. Long term goal: publish the vocabulary in the National Center for Biomedical Ontology's BioPortal.

REFERENCES

Alper BS, Flynn A, Bray BE, Conte ML, Eldredge C, Gold S, Greenes RA, Haug P, Jacoby K, Koru G, McClay J, Sainvil ML, Sottara D, Tuttle M, Visweswaran S, Yurk RA. Categorizing metadata to help mobilize computable biomedical knowledge. Learn Health Syst. 2021 May 9;6(1):e10271. doi: 10.1002/lrh2.10271. PMID: 35036552; PMCID: PMC8753304.