Ontologizing Health Systems at Scale: Making Translational Discovery a Reality

Tiffany J. Callahan MPH, PhD Candidate
98% of Hospitals Use Electronic Health Records

“...now is the time to create smarter healthcare systems in which the best treatment decisions are computationally learned from electronic health record data by deep-learning methodologies”\textsuperscript{1}

\textsuperscript{1}Norgeot et al. \textit{Nat Med}. 2019; Flynn et al. \textit{Knowledge Grid}. 2018
Creating Smarter Healthcare Systems

Current research use of medical record data enables

- Automatic triage of medical conditions
- Prediction of pertinent patient risk factors
- Identification of emerging or important pathogens
Creating Smarter Healthcare Systems

Current research use of medical record data enables

- Automatic triage of medical conditions
- Prediction of pertinent patient risk factors
- Identification of emerging or important pathogens

Connecting medical record data with biomedical knowledge could enable

- Integration of patient -omics data
- Identification of causal biomarkers
- Mechanistic explanations for specific patient characteristics
BARRIER 1

Medical Records are not Connected to Biomedical Knowledge
Knowledge Graphs

Diagram showing relationships between concepts such as Disease, Signs & Symptoms, Gene, Drug, Biological Process, Molecular Function, Cellular Component, Pathway, Substance that Treats, Causes or Contributes to, Interacts with, Participates in, has Part, Realizes, has Component, Located in, has Function.
Knowledge Graphs Challenges

Design Challenges
- Multiple approaches to modeling biomedical knowledge\textsuperscript{1-3}
- Decisions impact downstream learning

Implementation Challenges
- Methods have varying functional, logical, and semantic consequences\textsuperscript{1}
- Semantic Web standard is expressive, but unwieldy and of uncertain benefit\textsuperscript{2}

PheKnowLator: A Python framework to build large-scale heterogeneous biomedical knowledge graphs

https://github.com/callahantiff/PheKnowLator

\textsuperscript{1}Hunter EPJ Data Science 2017; \textsuperscript{2}Callahan et al Ann Rev Biomed Data Sci 2020; \textsuperscript{3}Ji et al arXiv 2020
Human Disease Mechanisms

Ontologies
12 OBO ontologies
366,846 classes
3,923,625 axioms

Edge Data
22 Linked Open Data
2 Experimental
34 edge-types

Validation
PhD Molecular Biologist

OBO = Open Biomedical Ontologies
Summary

- First fully custom biomedical knowledge graph construction framework
- Evaluation in progress
- Promising applications
  - Biomedical hypergraphs → Joslyn, Aksoy, Callahan et al., arXiv. 2020
  - Toxicogenomic mechanistic inference → Tripodi, Callahan et al. Tox InVitro. 2020
- Full testing and continuous integration; Dockerized

https://github.com/callahantiff/PheKnowLator
Medical Records were not Built to Facilitate Translational Research
Clinical Research Challenges

Problem
- Hospital databases/EHRs incomplete and not standardized\(^1\)-\(^3\)
- Diagnosis codes used for billing, not research\(^2\),\(^4\)

Solution
- CDMs help different data sources speak the same language
- Ontologies contain meaningful representations of molecular knowledge

EHRs = Electronic Health Records; CDMs = Common Data Models

Ontologies

Ontology → Graph of entities and relationships\textsuperscript{1,2}
- Domain-specific
- Community consensus
- Hierarchical

Open Biomedical Ontologies
- Phenotypes (Human Phenotype Ontology)
- Diseases (Disease Ontology)
- Chemicals/Metabolites/Hormones (ChEBI)
- Anatomical Entities (UBERON)
- Cell Types (Cell Line Ontology)

\textsuperscript{1}Hunter, Data Science, 2017; \textsuperscript{2}Callahan et al., Annu. Rev. Biomed. Data Sci, 2020
Ontologies Translate Clinical Concepts

Haendel et al., Annu. Rev. Biomed. Data Sci., 2018
Existing Mapping Work

- Manually curated 2,923 LOINC lab tests to HPO
- Validated with 15,681 patients with respiratory complaints → known asthma biomarkers
  - Abnormal metabolism/vitamin metabolism
  - ↑ Red blood cell count and VLDL cholesterol concentration
Mapping Strategies

“Chronic deep venous thrombosis of right calf”

1:1 Mapping Approach:

“Chronic deep venous thrombosis of right calf” → “Abnormality of the calf”

1:Many Mapping Approach:

“Chronic deep venous thrombosis of right calf” → “Chronic”
“Deep venous thrombosis”
“Abnormality of the calf”
“Right”
Objective

- Extend and expand existing mapping work → more than 1:1 mappings
- Map ontologies to OMOP
- Hospital scale, disease-agnostic

**OMOP2OBO**: the first health system-wide integration and alignment between OMOP standardized clinical terminologies and eight OBO biomedical ontologies

[https://github.com/callahantiff/OMOP2OBO](https://github.com/callahantiff/OMOP2OBO)
<table>
<thead>
<tr>
<th>Clinical Domain</th>
<th>OMOP Table</th>
<th>Concept Class</th>
<th>Concept Vocabularies</th>
<th>Ontologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditions</td>
<td>Condition Occurrence</td>
<td>Conditions</td>
<td>SNOMED CT Source Codes: “Mapped to”, “Mapped From”, “Concept poss_eq from” “Concept same_as from”</td>
<td>Human Phenotype Ontology Mondo Disease Ontology</td>
</tr>
<tr>
<td>Medications</td>
<td>Drug Exposure</td>
<td>Drugs</td>
<td>RxNorm Standard Non-Standard</td>
<td>ChEBI Protein Ontology NCBiTaxon Vaccine Ontology</td>
</tr>
<tr>
<td>Measurements</td>
<td>Measurements</td>
<td>Measurements</td>
<td>LOINC Standard Non-Standard</td>
<td>Human Phenotype Ontology ChEBI Uber Anatomy Ontology</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Protein Ontology NCBiTaxon Cell Ontology</td>
</tr>
</tbody>
</table>
# Available Data

## Clinical Data - OMOP CDM

<table>
<thead>
<tr>
<th>Concept ID*</th>
<th>138994</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Code</td>
<td>snomed:109995007, icd10cm:D46.9, mesh:D009190, icd9cm:238.75</td>
</tr>
<tr>
<td>Concept Name</td>
<td>Myelodysplastic syndrome</td>
</tr>
<tr>
<td>Concept Synonym</td>
<td>Myelodysplastic syndrome (disorder)</td>
</tr>
</tbody>
</table>

*OMOP Data also includes all ancestors

## Ontology Data - Open Biomedical Ontologies

<table>
<thead>
<tr>
<th>Class ID</th>
<th>HP_0002863</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Label</td>
<td>Myelodysplasia</td>
</tr>
<tr>
<td>Class Synonym</td>
<td>Hypoplastic myelodysplasia</td>
</tr>
<tr>
<td>Class Definition</td>
<td>Clonal hematopoietic stem cell disorders characterized by dysplasia (ineffective production) in one or more hematopoietic cell lineages, leading to anemia and cytopenia</td>
</tr>
<tr>
<td>Class DbXRef</td>
<td>UMLS:C1851971, MSH:D009190, SNOMEDCT_US:109995007</td>
</tr>
</tbody>
</table>
Mapping Approach

Mapping Strategies

- Database Cross-References

- Exact String Match
  - Labels
  - Synonyms

- Bag-of-Words + TF-IDF Weighting
  - Labels
  - Synonyms
  - Definitions (ontologies only)
### Mapping Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Automatic Exact - Concept</strong></td>
<td>Exact label or synonym, dbXRef, or expert validated mapping @ concept-level; 1:1</td>
</tr>
<tr>
<td><strong>Automatic Exact - Ancestor</strong></td>
<td>Exact label or synonym, dbXRef, or expert validated mapping @ concept ancestor-level; 1:1</td>
</tr>
<tr>
<td><strong>Manual Exact - Concept Similarity</strong></td>
<td>Concept similarity score suggested mapping -- manually verified; 1:1</td>
</tr>
<tr>
<td><strong>Automatic Constructor - Concept</strong></td>
<td>Exact label or synonym, dbXRef, cosine similarity, or expert validated mapping @ concept-level; 1:Many</td>
</tr>
<tr>
<td><strong>Automatic Constructor - Ancestor</strong></td>
<td>Exact label or synonym, dbXRef, cosine similarity, or expert validated mapping @ concept-level; 1:Many</td>
</tr>
<tr>
<td><strong>Manual Constructor</strong></td>
<td>Hand mapping created using expert suggested resources; 1:Many</td>
</tr>
<tr>
<td><strong>Manual</strong></td>
<td>Hand mapping created using expert suggested resources; 1:1</td>
</tr>
<tr>
<td><strong>UnMapped</strong></td>
<td>No suitable mapping or not mapped type</td>
</tr>
</tbody>
</table>
## Mapping Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic Exact - Concept</td>
<td>Exact label or synonym, dbXRef, or expert validated mapping @ concept-level; 1:1</td>
</tr>
<tr>
<td>Automatic Exact - Ancestor</td>
<td>Exact label or synonym, dbXRef, or expert validated mapping @ concept ancestor-level; 1:1</td>
</tr>
<tr>
<td>Manual Exact - Concept Similarity</td>
<td>Concept similarity score suggested mapping -- manually verified; 1:1</td>
</tr>
<tr>
<td>Automatic Constructor - Concept</td>
<td>Exact label or synonym, dbXRef, cosine similarity, or expert validated mapping @ concept-level; 1:Many</td>
</tr>
<tr>
<td>Automatic Constructor - Ancestor</td>
<td>Exact label or synonym, dbXRef, cosine similarity, or expert validated mapping @ concept-level; 1:Many</td>
</tr>
<tr>
<td>Manual Constructor</td>
<td>Hand mapping created using expert suggested resources; 1:Many</td>
</tr>
<tr>
<td>Manual</td>
<td>Hand mapping created using expert suggested resources; 1:1</td>
</tr>
<tr>
<td>UnMapped</td>
<td>No suitable mapping or not mapped type</td>
</tr>
</tbody>
</table>

### Example: Apraxia (OMOP_132342)

#### Mapping Type: Automatic Exact - Concept

#### Ontology Concept: Apraxia (HP_0002186)

#### Mapping Evidence:

- CONCEPT_DBXREF:snomed_68345001
- OBO_LABEL-OMOP_CONCEPT_LABEL:apraxia
- OBO_LABEL-OMOP_CONCEPT_SYNONYM:apraxia
- CONCEPT_SIMILARITY:HP_0002186_1.0
- ANCESTOR_DBXREF:snomed_68345001
- OBO_LABEL-OMOP_ANCESTOR_LABEL:apraxia
## Conditions Occurrence Mappings

<table>
<thead>
<tr>
<th>Mapping Type</th>
<th>Condition Concept ID</th>
<th>Phenotype (Human Phenotype Ontology)</th>
<th>Disease (Mondo Disease Ontology)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Automatic Mapping</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concept</td>
<td>Macular Hole (OMOP_4338894)</td>
<td>Macular Hole (HP_0011508)</td>
<td>Macular Hole (DOID_7633)</td>
</tr>
<tr>
<td>Ancestor</td>
<td>Major histocompatibility complex class I deficiency (OMOP_4100979)</td>
<td>Severe Combined Immunodeficiency (HP_0004430)</td>
<td>Severe Combined Immunodeficiency (DOID_627)</td>
</tr>
<tr>
<td><strong>Manual Mapping</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similarity</td>
<td>Malakoplakia of colon (OMOP_4024250)</td>
<td>Abnormality of the large intestine morphology (HP_0005210)</td>
<td>Colonic Disease (DOID_5353)</td>
</tr>
<tr>
<td>Constructor</td>
<td>Macular edema and retinopathy due to type 2 diabetes mellitus (OMOP_45770830)</td>
<td>AND</td>
<td>Type II Diabetes Mellitus (HP_0005978)</td>
</tr>
</tbody>
</table>
Conditions Occurrence Mappings

- UMLS CUIs (*MRCONSO 2020AB*) and Semantic Types (*MRSTY 2020AB*)
- 28,129 unique condition occurrence codes

**Evaluation:** 20% of manually mapped concepts verified by clinicians; several iterations
  - 24,285 OMOP concepts → 4,661 Phenotypes
  - 19,664 OMOP concepts → 3,614 Diseases

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HPO</td>
<td>3465</td>
<td>2851</td>
<td>1055</td>
<td>174</td>
<td>1825</td>
<td>9477</td>
<td>5438</td>
</tr>
<tr>
<td>MONDO</td>
<td>4965</td>
<td>6103</td>
<td>484</td>
<td>723</td>
<td>2301</td>
<td>3109</td>
<td>1979</td>
</tr>
</tbody>
</table>
Drug Exposure Mappings

**Mapping:** Ingredients and mechanism of action (DrugBank and CHEMBL)

**Example:** balsalazide (OMOP_934262)

**Ingredient mapping:**
- Automatic Exact - Concept
- balsalazide (CHEBI_267413)

**Mechanism of Action:**
- Automatic Constructor - Concept
- **agonist:** peroxisome proliferator-activated receptor gamma (PR_P37231)
- **Inhibitor:** prostaglandin G/H synthase 2 (PR_P35354), prostaglandin G/H synthase 1 (PR_P23219), arachidonate 5-lipoxygenase-activating protein (PR_P20292)
- **organism:** homo sapien (NCBITaxon_9606)
Drug Exposure Mappings

- 11,937 drug-ingredient combinations → 1,697 unique ingredients

Evaluation: 20% of manually mapped concepts verified by clinical pharmacist; 3 iterations
- 1,618 OMOP concepts → 1,422 Chemicals, hormones, or metabolites
- 139 OMOP concepts → 91 Proteins
- 317 OMOP concepts → 39 Organisms
- 127 OMOP concepts → 54 Vaccines/Immunizations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CHEBI</td>
<td>1176</td>
<td>3</td>
<td>64</td>
<td>9</td>
<td>1</td>
<td>58</td>
<td>306</td>
<td>80</td>
</tr>
<tr>
<td>PRO</td>
<td>2</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>119</td>
<td>1558</td>
</tr>
<tr>
<td>NCBI Taxon</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>287</td>
<td>1380</td>
</tr>
<tr>
<td>VO</td>
<td>92</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>25</td>
<td>1570</td>
</tr>
</tbody>
</table>
Measurement Mappings

- Leverage LOINC scale and results type
- Data-drive confirmation

Somatotropin [Mass/volume] in Serum or Plasma (OMOP_3023709)
  - OR blood serum(UBERON_0001977) | blood plasma (UBERON_0001969)
  - Somatotropin (PR_000007968)
    - Normal: NOT Abnormality of circulating hormone level (HP_0003117)
    - Low: Growth Hormone Deficiency (HP_0000824)
    - High: Growth Hormone Excess (HP_0000845)

Transitional cells [Presence] in Urine sediment by Light microscopy (OMOP_3028475)
  - Urine (UBERON_0001088), Transitional epithelial cell (CL_0000244)
    - Negative: NOT Increased urinary transitional epithelial cell count (HP_0032214)
    - Positive: Increased urinary transitional epithelial cell count (HP_0032214)
### Measurement Mappings

- 4,382 lab tests, 11,072 lab test results

#### Evaluation:
- 270 results verified by 3 MDs, 1 epidemiologist
- 15% verified by a biocurator, 3 iterations
  - 10,888 OMOP concepts → 920 Phenotypes | 10,876 OMOP concepts → 25 Anatomical entities
  - 1,075 OMOP concepts → 27 Cell Types | 9,710 OMOP concepts → 338 Chemicals, Hormones
  - 8,269 OMOP concepts → 194 Organisms | 4,842 OMOP concepts → Proteins

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HPO</td>
<td>6891</td>
<td>22</td>
<td>171</td>
<td>4</td>
<td>0</td>
<td>49</td>
<td>3751</td>
</tr>
<tr>
<td>UBERON</td>
<td>0</td>
<td>1612</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3307</td>
<td>5957</td>
</tr>
<tr>
<td>CL</td>
<td>0</td>
<td>227</td>
<td>207</td>
<td>0</td>
<td>87</td>
<td>30</td>
<td>524</td>
</tr>
<tr>
<td>CHEBI</td>
<td>5</td>
<td>3535</td>
<td>1098</td>
<td>151</td>
<td>241</td>
<td>922</td>
<td>3758</td>
</tr>
<tr>
<td>NCBI Taxon</td>
<td>0</td>
<td>693</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>459</td>
<td>7117</td>
</tr>
<tr>
<td>PRO</td>
<td>0</td>
<td>44</td>
<td>259</td>
<td>0</td>
<td>0</td>
<td>175</td>
<td>4364</td>
</tr>
</tbody>
</table>
OMOP2OBO

- Determine coverage of mappings on two independent samples

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Domain</th>
<th>Unique Concepts</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMIC-III</td>
<td>Conditions</td>
<td>5608</td>
<td>92.13%</td>
</tr>
<tr>
<td></td>
<td>Medications</td>
<td>4426</td>
<td>96.36%</td>
</tr>
<tr>
<td>UCHHealth</td>
<td>Conditions</td>
<td>15056</td>
<td>79.78%</td>
</tr>
<tr>
<td></td>
<td>Medications</td>
<td>23972</td>
<td>91.35%</td>
</tr>
</tbody>
</table>
Summary

- First hospital scale mapping between OMOP and the OBOs
- Preliminary coverage examined on adult and ICU populations
- Added over 200 new concepts to HPO
- Adopted by National COVID-19 Cohort (N3C)

Next Steps:
- Coverage study in subset of Concept Prevalence data
- Comparing to Juan Banda’s Mappings

https://github.com/callahantiff/OMOP2OBO
ACKNOWLEDGEMENTS

Translational Research Experts

Adrienne L. Stefanski, PhD
CU Anschutz

Peter Robinson, MD
Professor
Jackson Labs

Nicole Vasilevsky, PhD
Research Assistant Professor
OHSU

Xingmin Aaron Zhang, PhD
Post Doctoral Fellow
Jackson Labs

Pharmacology Experts

Katelyn Trinkley, PharmD
Associate Professor
CU Anschutz

Catherine Derington, PharmD
Post Doctoral Fellow
CU Anschutz

Jessica Sinclair, PharmD
Pharmacy Resident
Purdue

Clinical Experts

James Feinstein, MD
Associate Professor
General Pediatrics
CU Anschutz

Blake Martin, MD
Pediatric Critical Care Medicine Fellow
CU Anschutz

Jordan Wyrra, DO
Physical Medicine and Rehabilitation Resident
CU Anschutz

Emily Swenson
CU Anschutz
Medical Student

Kelsey Andrews
CU Anschutz
Medical Student

Organizations

OHDSI
International Health Data Sciences and Informatics

Children's Hospital Colorado

HEALTH_DATA
Compass
A Pediatric Learning Health System

PEDSnet
U.S. National Library of Medicine

Work funded by the NIH, No. T15LM009451