Ontologizing Health Systems at Scale: Making Translational Discovery a Reality

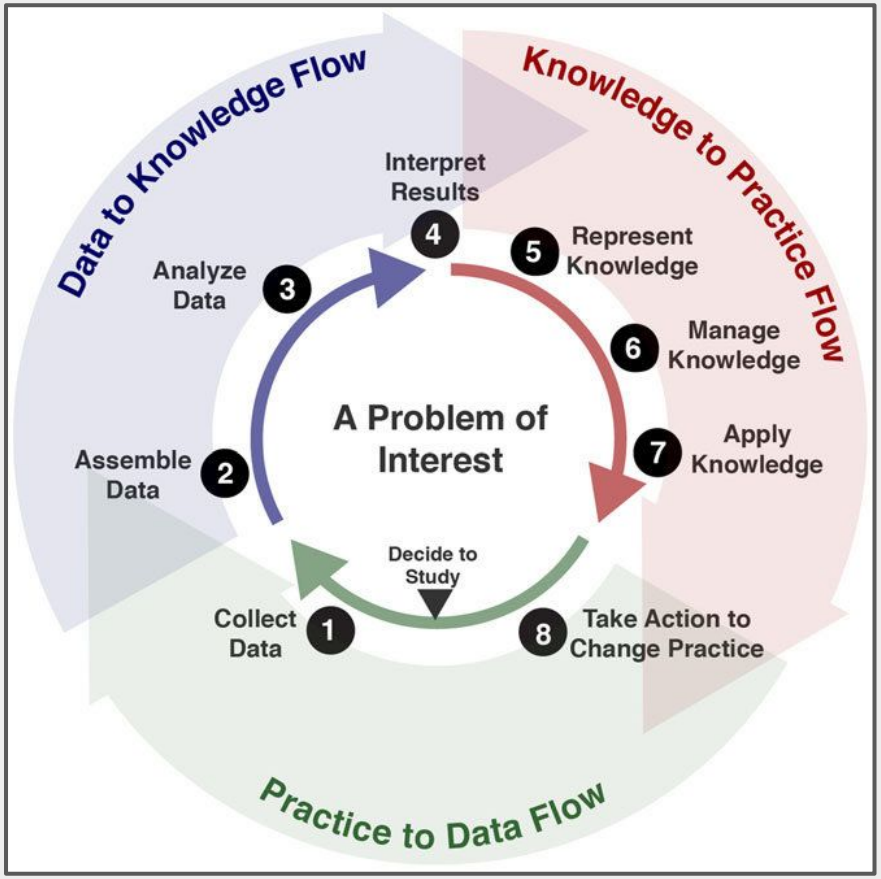
Tiffany J. Callahan MPH, PhD Candidate

OHDSI Community Call - September 2020

University of Colorado
Anschutz Medical Campus

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"[1]



[1]Norgeot et al. *Nat Med*. 2019; Flynn et al. *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

Rajkomar, Dean, Kohane *N Engl J Med* 2019

Creating Smarter Healthcare Systems

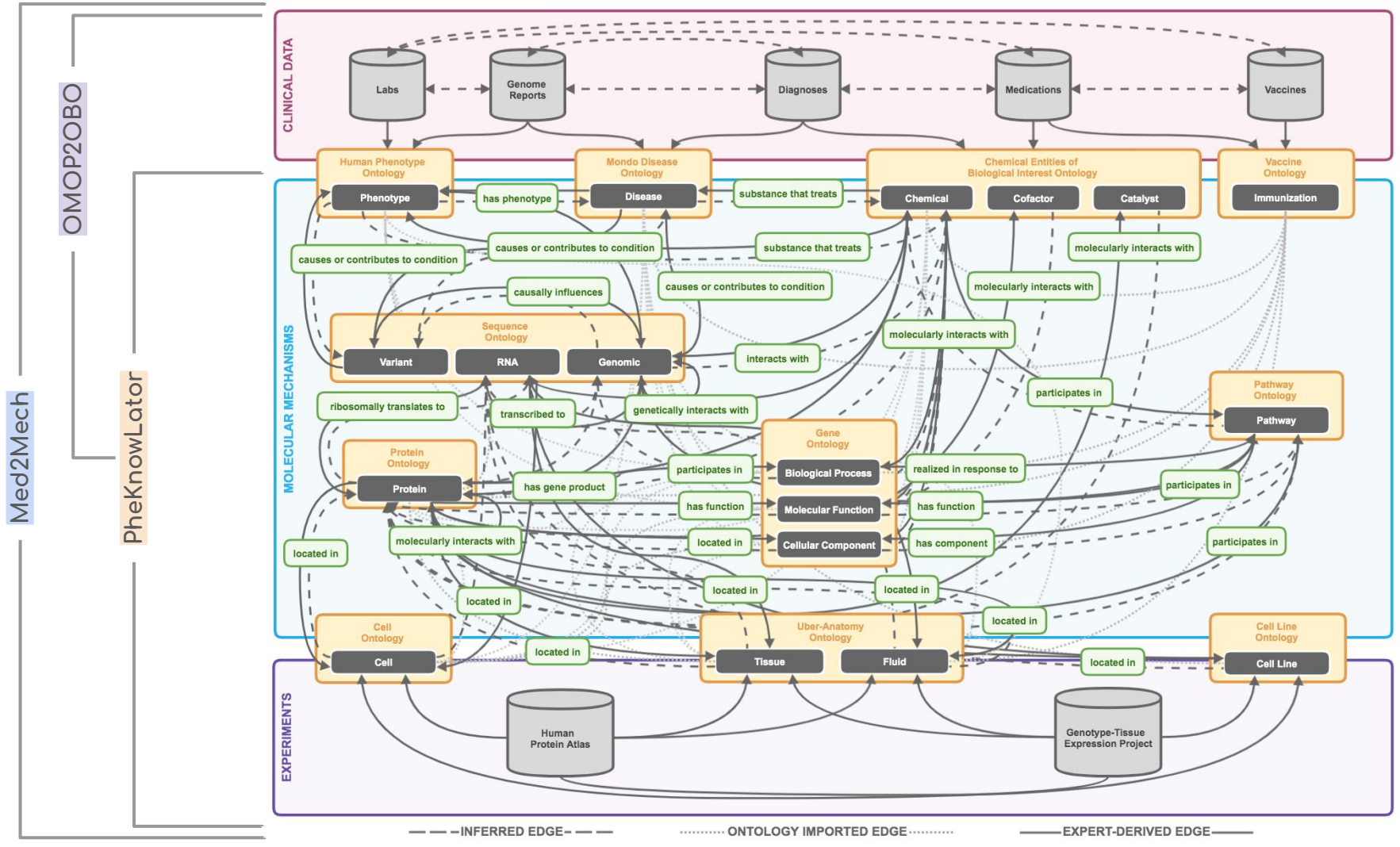
Current research use of medical record data enables

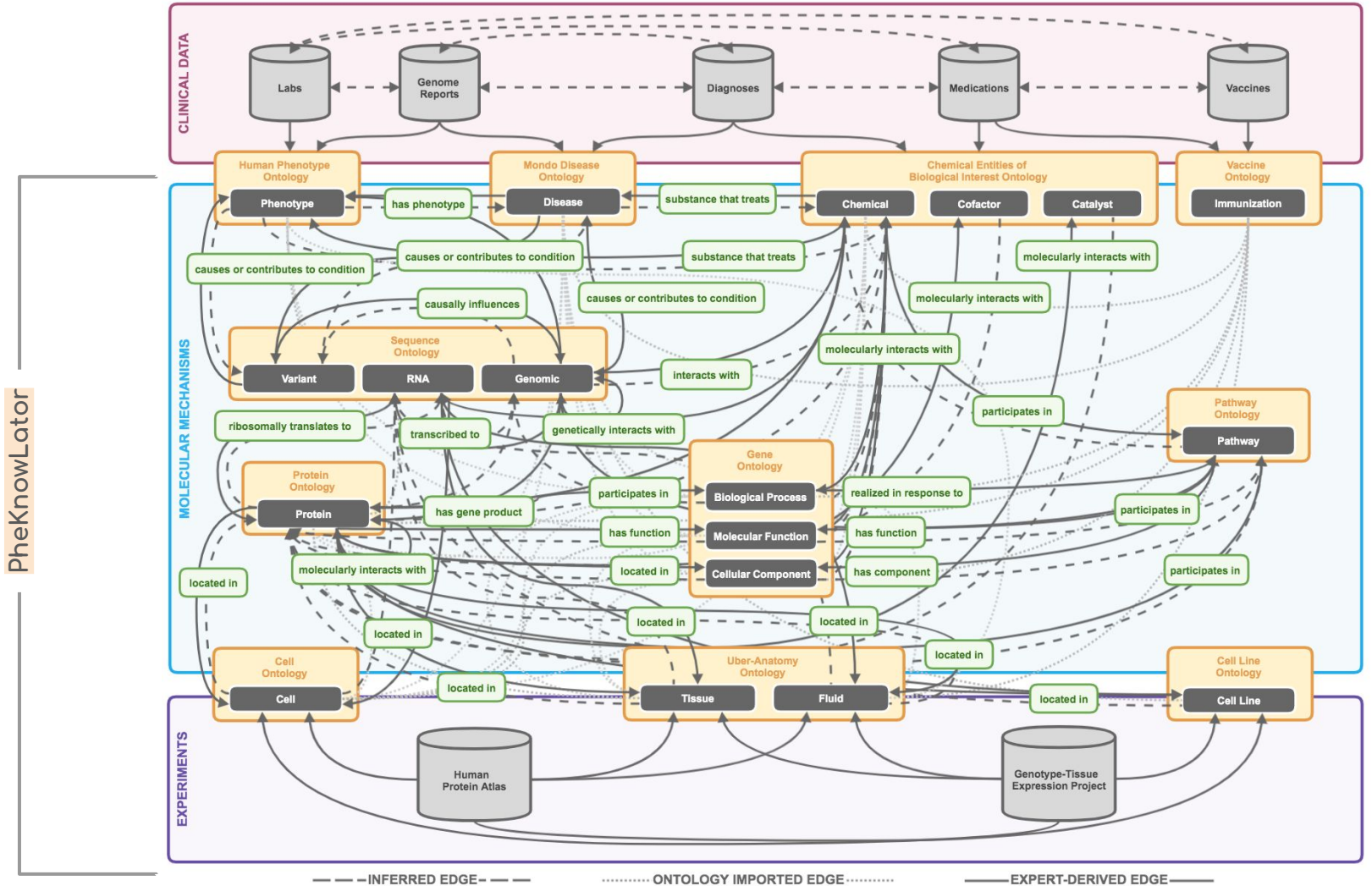
- Automatic triage of medical conditions
- Prediction of pertinent patient risk factors
- Identification of emerging or important pathogens

Connecting medical record data with biomedical knowledge could enable

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

Rajkomar, Dean, Kohane *N Engl J Med* 2019

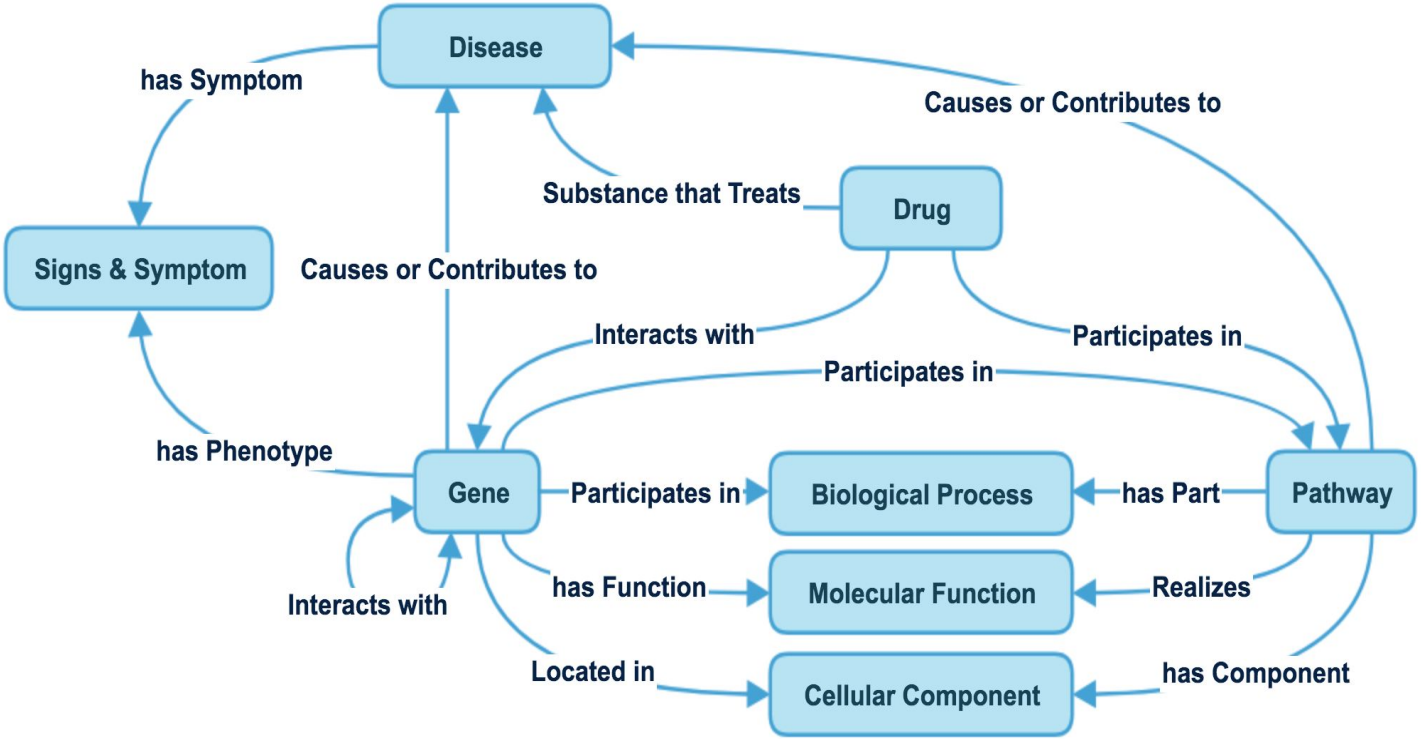

Med2Mech
OMOP2OBO
PheKnowLator
CLINICAL DATA
Labs
Genome Reports
Diagnoses
Medications
Vaccines
MOLECULAR MECHANISMS
Human Phenotype Ontology
Phenotype
has phenotype
Mondo Disease Ontology
Disease
substance that treats
Chemical Entities of Biological Interest Ontology
Chemical
Cofactor
Catalyst
Vaccine Ontology
Immunization
causes or contributes to condition
causes or contributes to condition
substance that treats
molecularly interacts with
causes or contributes to condition
causally influences
molecularly interacts with
Sequence Ontology
Variant
RNA
Genomic
interacts with
molecularly interacts with
participates in
Pathway Ontology
Pathway
ribosomally translates to
transcribed to
genetically interacts with
Gene Ontology
Protein Ontology
Protein
has gene product
participates in
Biological Process
realized in response to
participates in
has function
Molecular Function
has function
molecularly interacts with
located in
Cellular Component
has component
participates in
located in
located in
located in
located in
Cell Ontology
Cell
located in
Uber-Anatomy Ontology
Tissue
Fluid
located in
Cell Line Ontology
Cell Line
located in
EXPERIMENTS
Human Protein Atlas
Genotype-Tissue Expression Project
INFERRED EDGE
ONTOLOGY IMPORTED EDGE
EXPERT-DERIVED EDGE


PheKnowLator
CLINICAL DATA
Labs
Genome Reports
Diagnoses
Medications
Vaccines
MOLECULAR MECHANISMS
Human Phenotype Ontology
Phenotype
has phenotype
Mondo Disease Ontology
Disease
substance that treats
Chemical Entities of Biological Interest Ontology
Chemical
Cofactor
Catalyst
Vaccine Ontology
Immunization
causes or contributes to condition
causes or contributes to condition
substance that treats
molecularly interacts with
causes or contributes to condition
causally influences
molecularly interacts with
Sequence Ontology
Variant
RNA
Genomic
interacts with
molecularly interacts with
participates in
Pathway Ontology
Pathway
ribosomally translates to
transcribed to
genetically interacts with
Gene Ontology
Protein Ontology
Protein
has gene product
participates in
Biological Process
realized in response to
participates in
has function
Molecular Function
has function
molecularly interacts with
located in
Cellular Component
has component
participates in
located in
located in
located in
located in
Cell Ontology
Cell
located in
Uber-Anatomy Ontology
Tissue
Fluid
located in
Cell Line Ontology
Cell Line
located in
EXPERIMENTS
Human Protein Atlas
Genotype-Tissue Expression Project
INFERRED EDGE
ONTOLOGY IMPORTED EDGE
EXPERT-DERIVED EDGE

BARRIER 1

Medical Records are not Connected to Biomedical Knowledge

Knowledge Graphs



Knowledge Graphs Challenges

Design Challenges

- Multiple approaches to modeling biomedical knowledge[1-3]
- Decisions impact downstream learning

Implementation Challenges

- Methods have varying functional, logical, and semantic consequences[1]
- Semantic Web standard is expressive, but unwieldy and of uncertain benefit[2]

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

https://github.com/callahantiff/PheKnowLator

[1]Hunter *EPJ Data Science* 2017; [2]Callahan et al *Ann Rev Biomed Data Sci* 2020; [3]Ji et al *arXiv* 2020

Human Disease Mechanisms

Ontologies

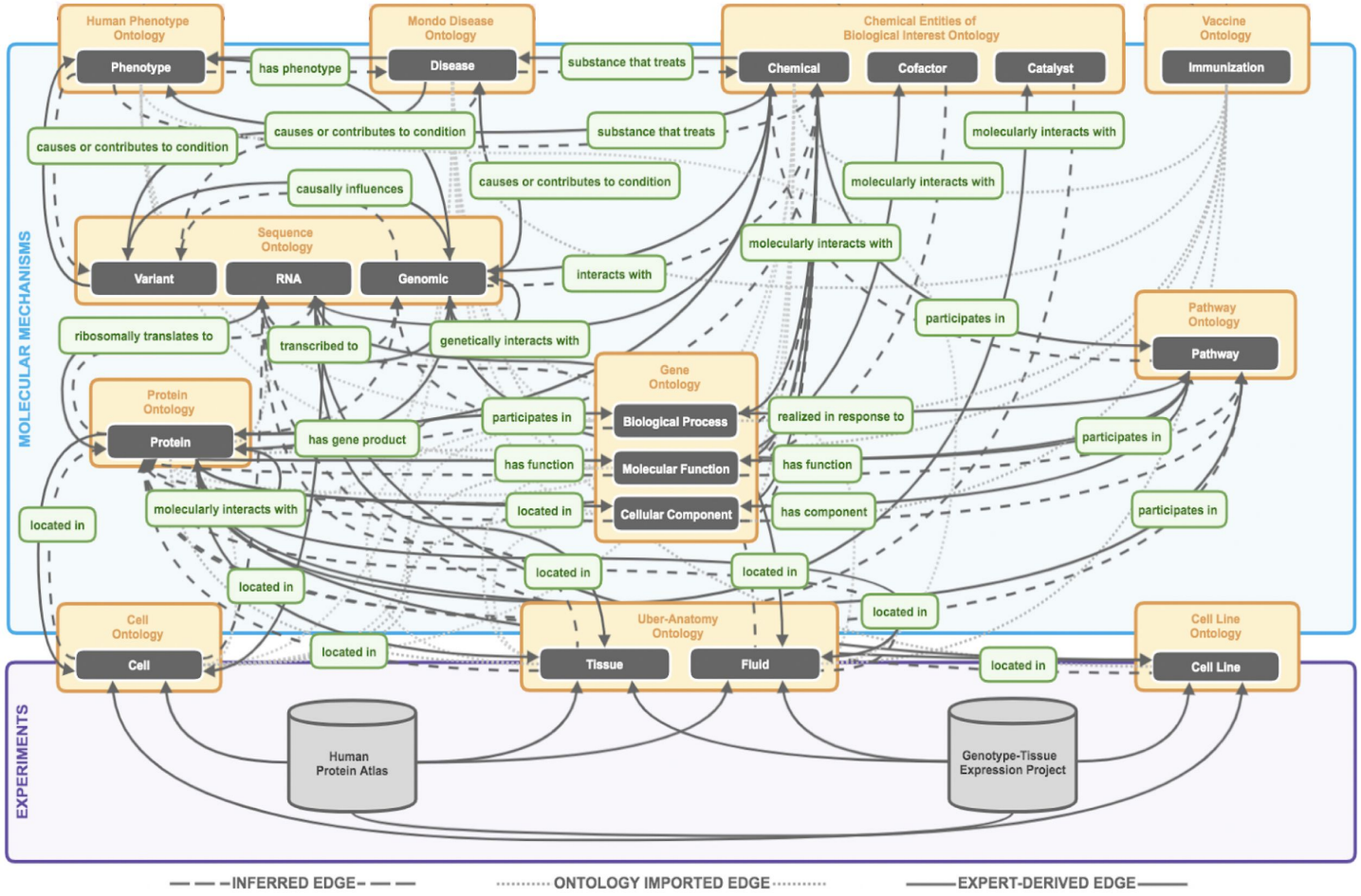
12 OBO ontologies
366,846 classes
3,923,625 axioms

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

BARRIER 2

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

[1]Sinnott et al. *J Amer Med Inform Assoc.* 2018; [2]Haendel et al. *N Engl J Med.* 2018; [3]Kahn, **Callahan** et al. *eGems. 2016*; [4]Zhang, Yates, Vasilevsky, Gourdine, **Callahan et al.** *Nat Digit Med. 2019*

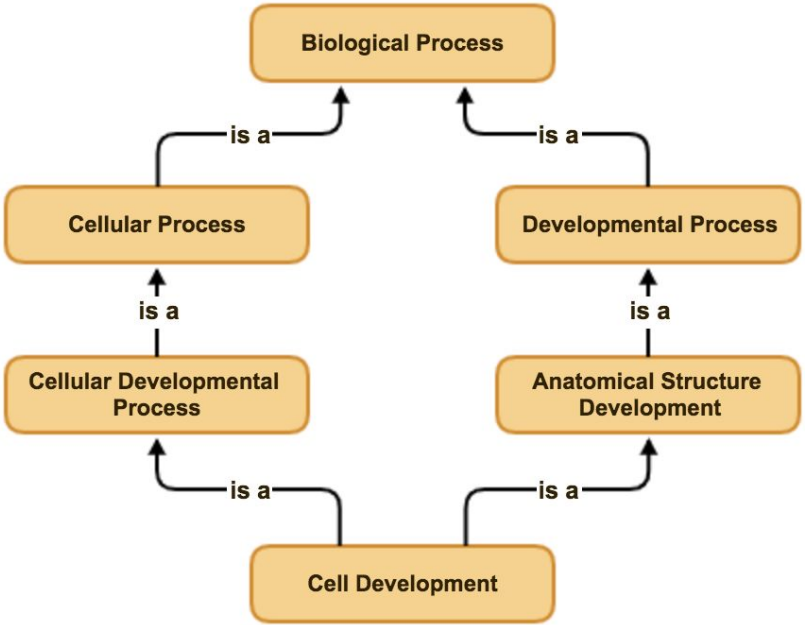
Ontologies

Ontology → Graph of entities and relationships[1,2]

- Domain-specific
- Community consensus
- Hierarchical

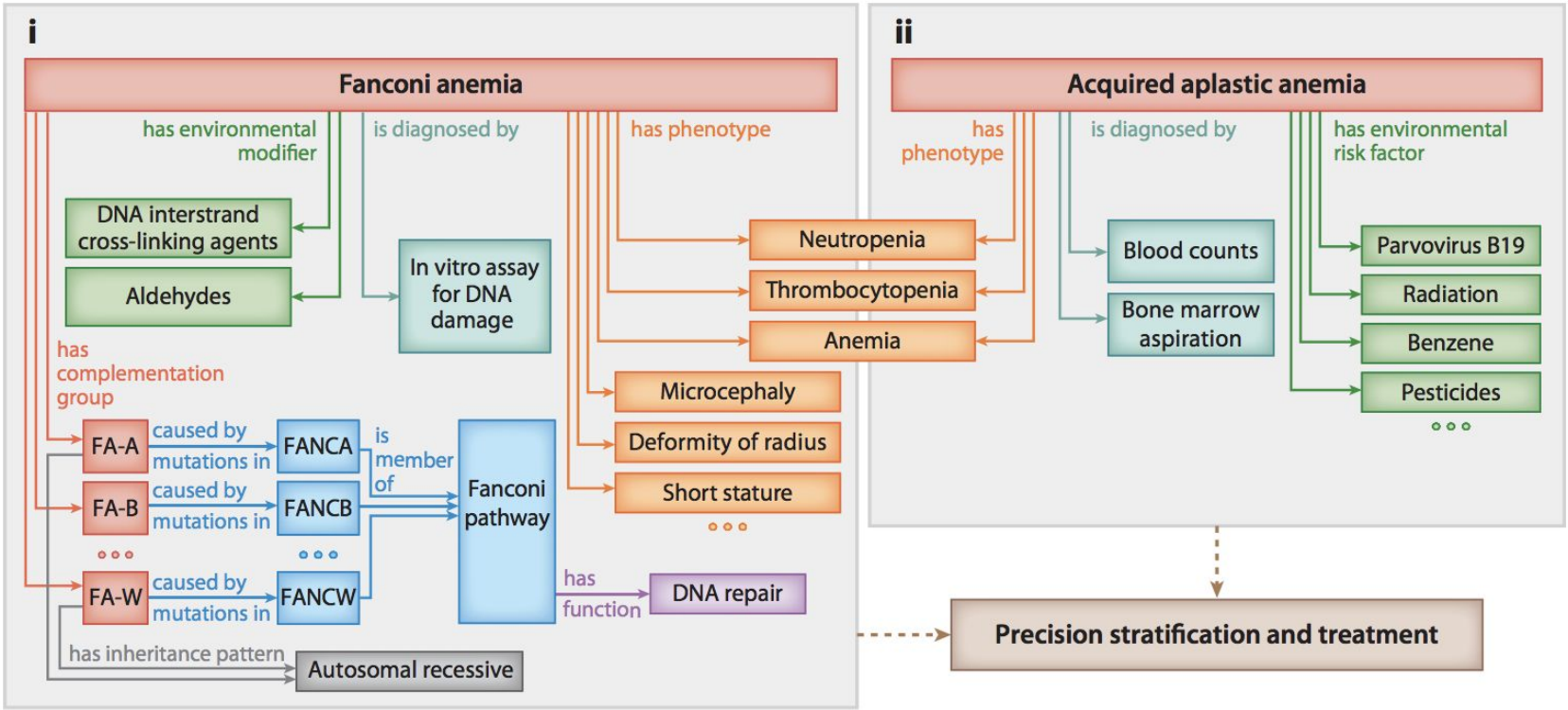
Open Biomedical Ontologies

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



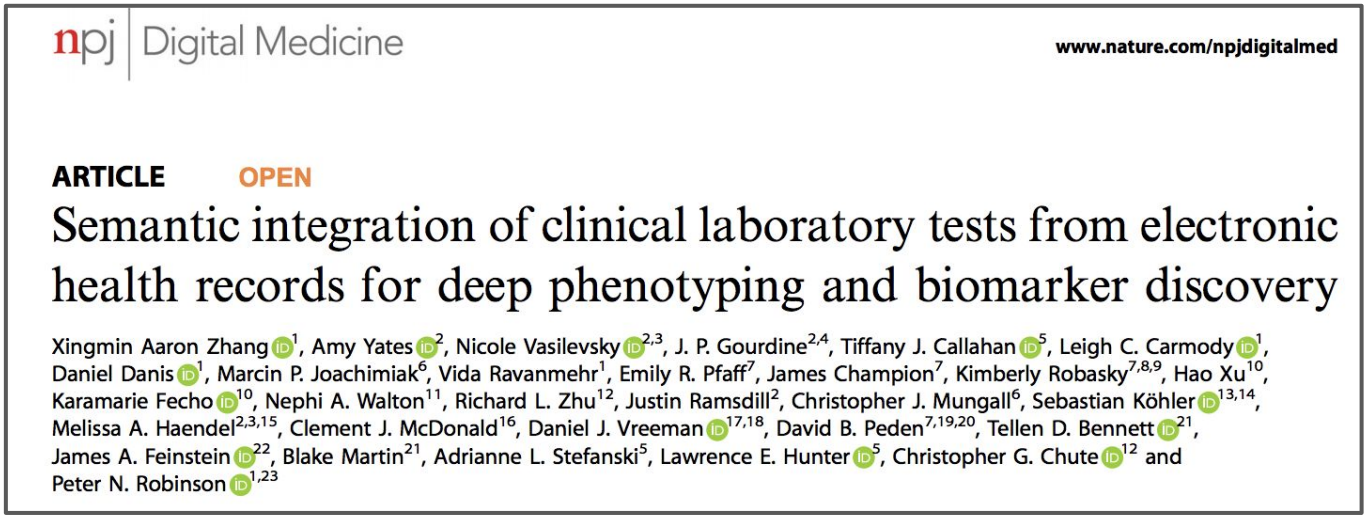
[1]Hunter. *Data Science*. 2017; [2]**Callahan** et al. *Annu. Rev. Biomed. Data Sci.* 2020

Ontologies Translate Clinical Concepts



Haendel et al., *Annu. Rev. Biomed. Data Sci.* 2018

Existing Mapping Work



npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery

Xingmin Aaron Zhang[1], Amy Yates[2], Nicole Vasilevsky[2,3], J. P. Gourdine[2,4], Tiffany J. Callahan[5], Leigh C. Carmody[1], Daniel Danis[1], Marcin P. Joachimiak[6], Vida Ravanmehr[1], Emily R. Pfaff[7], James Champion[7], Kimberly Robasky[7,8,9], Hao Xu[10], Karamarie Fecho[10], Nephi A. Walton[11], Richard L. Zhu[12], Justin Ramsdill[2], Christopher J. Mungall[6], Sebastian Köhler[13,14], Melissa A. Haendel[2,3,15], Clement J. McDonald[16], Daniel J. Vreeman[17,18], David B. Peden[7,19,20], Tellen D. Bennett[21], James A. Feinstein[22], Blake Martin[21], Adrianne L. Stefanski[5], Lawrence E. Hunter[5], Christopher G. Chute[12] and Peter N. Robinson[1,23]

- 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:

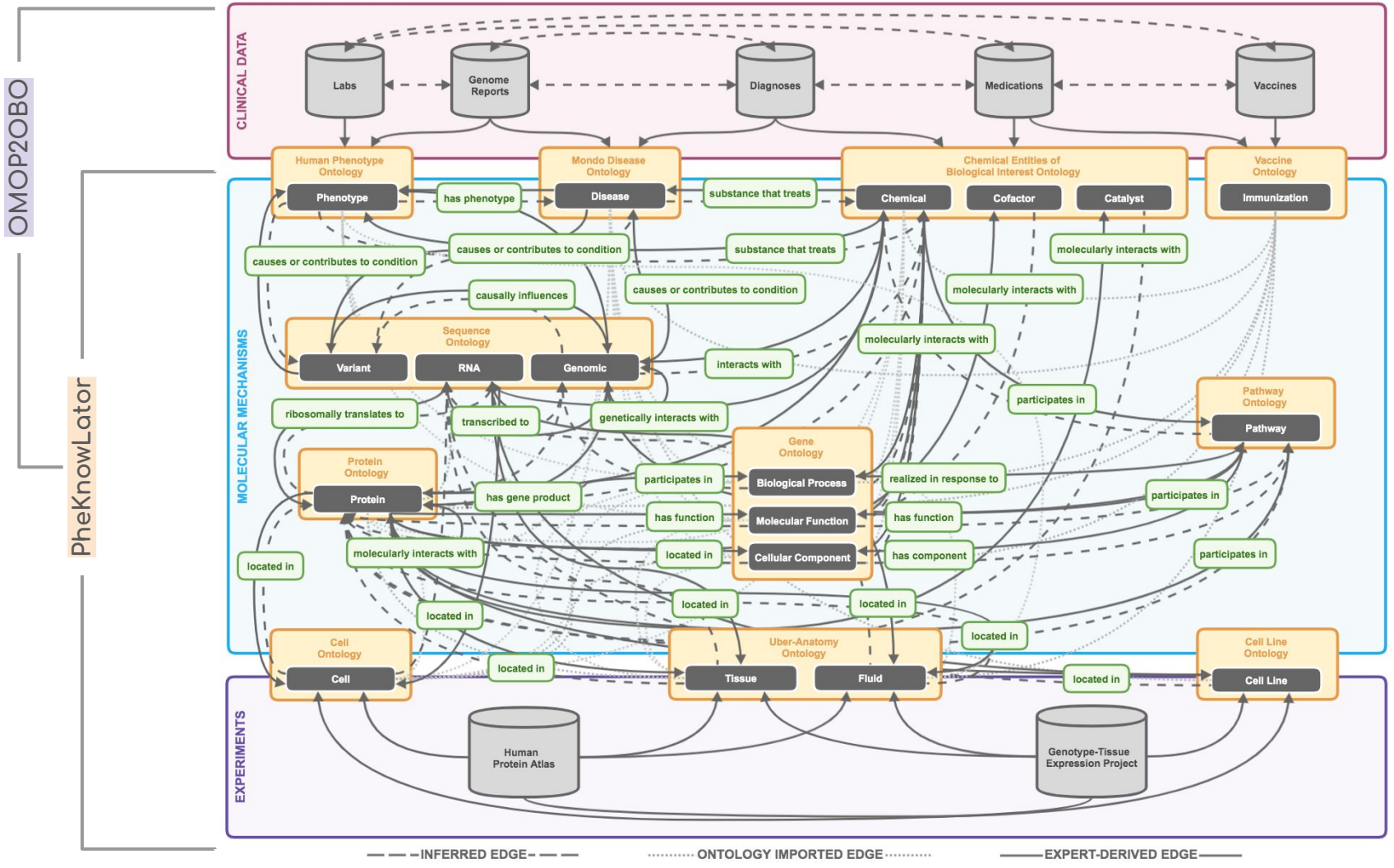
"Chronic deep venous thrombosis of right calf" → "Chronic"
"Deep venous thrombosis"
"Abnormality of the calf"
"Right"

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


OMOP2OBO
PheKnowLator
CLINICAL DATA
Labs
Genome Reports
Diagnoses
Medications
Vaccines
MOLECULAR MECHANISMS
Human Phenotype Ontology
Phenotype
Mondo Disease Ontology
Disease
Chemical Entities of Biological Interest Ontology
Chemical
Cofactor
Catalyst
Vaccine Ontology
Immunization
has phenotype
substance that treats
causes or contributes to condition
causally influences
molecularly interacts with
Sequence Ontology
Variant
RNA
Genomic
interacts with
ribosomally translates to
transcribed to
genetically interacts with
participates in
Pathway Ontology
Pathway
Protein Ontology
Protein
has gene product
Gene Ontology
Biological Process
Molecular Function
Cellular Component
realized in response to
has function
has component
located in
Cell Ontology
Cell
Uber-Anatomy Ontology
Tissue
Fluid
Cell Line Ontology
Cell Line
EXPERIMENTS
Human Protein Atlas
Genotype-Tissue Expression Project
INFERRED EDGE
ONTOLOGY IMPORTED EDGE
EXPERT-DERIVED EDGE

Available Data

Clinical Domain	OMOP Table	Concept Class	Concept Vocabularies	Ontologies
Conditions	Condition Occurrence	Conditions	SNOMED CT Source Codes: *“Mapped to”, “Mapped From”,* *"Concept poss_eq from"* *"Concept same_as from"*	Human Phenotype Ontology Mondo Disease Ontology
Medications	Drug Exposure Immunizations	Drugs Ingredients	RxNorm Standard Non-Standard	ChEBI Protein Ontology NCBITaxon Vaccine Ontology
Measurements	Measurements	Measurements	LOINC Standard Non-Standard	Human Phenotype Ontology ChEBI Uber Anatomy Ontology Protein Ontology NCBITaxon Cell Ontology

Available Data

Clinical Data - OMOP CDM	
Concept ID*	138994
Source Code	snomed:109995007 icd10cm:D46.9 mesh:D009190 icd9cm:238.75
Concept Name	Myelodysplastic syndrome
Concept Synonym	Myelodysplastic syndrome (disorder) | Myelodysplastic syndrome (clinical)

*OMOP Data also includes all ancestors

Ontology Data - Open Biomedical Ontologies	
Class ID	HP_0002863
Class Label	Myelodysplasia
Class Synonym - Broad - Exact - Narrow - Related	Hypoplastic myelodysplasia
Class Definition	Clonal hematopoietic stem cell disorders characterized by dysplasia (ineffective production) in one or more hematopoietic cell lineages, leading to anemia and cytopenia
Class DbXRef	UMLS:C1851971 MSH:D009190 SNOMEDCT_US:109995007

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

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

Mapping Categories

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

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

Mapping Type		Condition Concept ID	Phenotype (Human Phenotype Ontology)	Disease (Mondo Disease Ontology)
Automatic Mapping	Concept	Macular Hole (OMOP_4338894)	Macular Hole (HP_0011508)	Macular Hole (DOID_7633)
	Ancestor	Major histocompatibility complex class I deficiency (OMOP_4100979)	Severe Combined Immunodeficiency (HP_0004430)	Severe Combined Immunodeficiency (DOID_627)
Manual Mapping	Similarity	Malakoplakia of colon (OMOP_4024250)	Abnormality of the large intestine morphology (HP_0005210)	Colonic Disease (DOID_5353)
	Constructor	Macular edema and retinopathy due to type 2 diabetes mellitus (OMOP_45770830)	AND | Type II Diabetes Mellitus (HP_0005978) | Macular Edema (HP_0040049) | Retinopathy (HP_0000488)	AND | Type 2 Diabetes (DOID_9352) | Macular Retinal Edema (DOID_4449)

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

Ontology	Automatic Exact - Concept	Automatic Exact - Ancestor	Manual Exact - Concept Similarity	Automatic Constructor - Concept	Automatic Constructor - Ancestor	Manual Constructor	Manual
HPO	3465	2851	1055	174	1825	9477	5438
MONDO	4965	6103	484	723	2301	3109	1979

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

Ontology	Automatic Exact - Concept	Automatic Exact - Ancestor	Manual Exact - Concept Similarity	Automatic Constructor - Concept	Automatic Constructor - Ancestor	Manual Constructor	Manual	Unmapped
CHEBI	1176	3	64	9	1	58	306	80
PRO	2	0	10	0	0	8	119	1558
NCBITaxon	20	0	0	0	0	10	287	1380
VO	92	0	8	1	0	1	25	1570

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

Ontology	Automatic Exact - Concept	Automatic Exact - Ancestor	Manual Exact - Concept Similarity	Automatic Constructor - Concept	Automatic Constructor - Ancestor	Manual Constructor	Manual
HPO	6891	22	171	4	0	49	3751
UBERON	0	1612	0	0	0	3307	5957
CL	0	227	207	0	87	30	524
CHEBI	5	3535	1098	151	241	922	3758
NCBITaxon	0	693	0	0	0	459	7117
PRO	0	44	259	0	0	175	4364

OMOP2OBO

- Determine coverage of mappings on two independent samples

Data Source	Domain	Unique Concepts	Coverage
MIMIC-III	Conditions	5608	92.13%
	Medications	4426	96.36%
UCHealth	Conditions	15056	79.78%
	Medications	23972	91.35%

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)

National COVID Cohort Collaborative

Next Steps:

- Coverage study in subset of Concept Prevalence data
- Comparing to Juan Banda's Mappings

https://github.com/callahantiff/OMOP2OBO

ACKNOWLEDGEMENTS

University of Colorado Anschutz Medical Campus

Trivisory

Dr. Michael G. Kahn
Professor of Clinical Informatics
Co-Director of Colorado CTSI
Director, Health Data Compass
Assoc Director, Colorado CPM

Dr. Lawrence E. Hunter
Professor of Pharmacology, Computer Science (CU Boulder), Director of the Computational Bioscience Program

Dr. Tellen D. Bennett
Assoc Professor of Pediatric Critical Care
Co-Director, Analytics Core, CU Data Science to Patient Value
Assoc Director, Informatics Core, CCTSI

Translational Research Experts

Adrianne 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

Clinical Experts

James Feinstein, MD
Associate Professor
General Pediatrics
CU Anschutz

Blake Martin, MD
Pediatric Critical Care
Medicine Fellow
CU Anschutz

Jordan Wyrwa, DO
Physical Medicine and Rehabilitation Resident
CU Anschutz

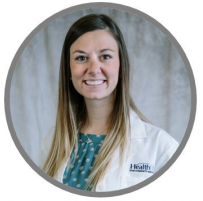
Emily Swenson
CU Anschutz
Medical Student

Kelsey Andrews
CU Anschutz
Medical Student

Pharmacology Experts

Katy Trinkley, PharmD
Associate Professor
CU Anschutz

Catherine Derington, PharmD
Post Doctoral Fellow
CU Anschutz



Jessica Sinclair, PharmD
Pharmacy Resident
Purdue

Organizations

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Children's Hospital Colorado

HEALTH DATA Compass

PEDSnet
A Pediatric Learning Health System

NIH U.S. National Library of Medicine

Work funded by the NIH, No. T15LM009451