

From Chaos to Clarity: Transformative Strategies to Harmonize and Interpret Real-World Data

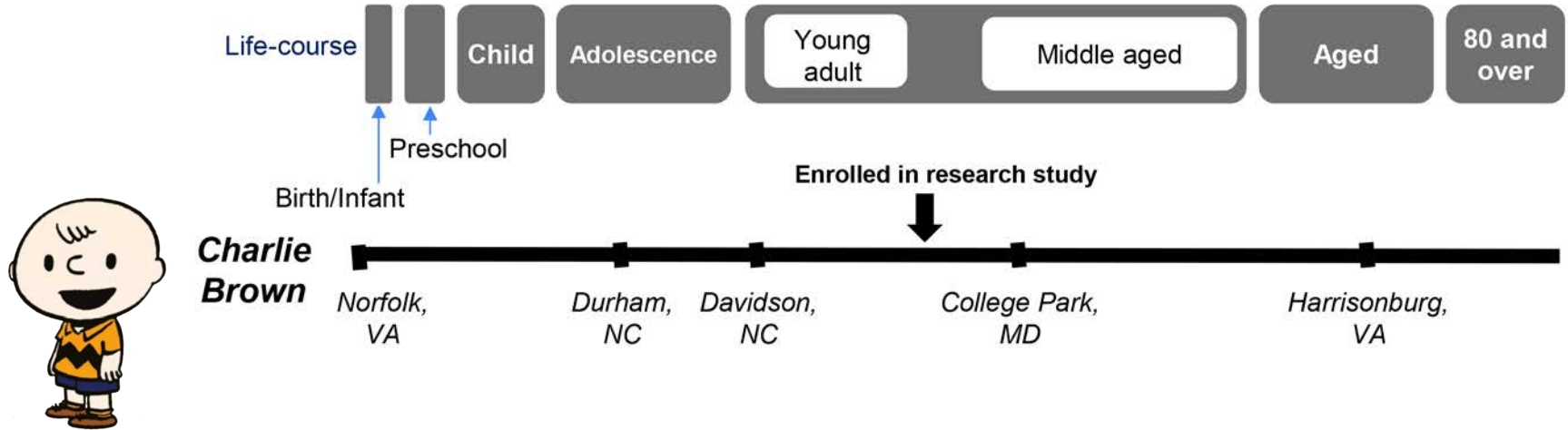
Melissa Haendel, PhD, FACMI
University of North Carolina Chapel Hill

Addressing Gaps, Challenges, and Opportunities Related to Data and Metadata Standards for
NIDDK Research Priorities



These slides: tislabs.org/niddk-2025

Putting the patient back together again

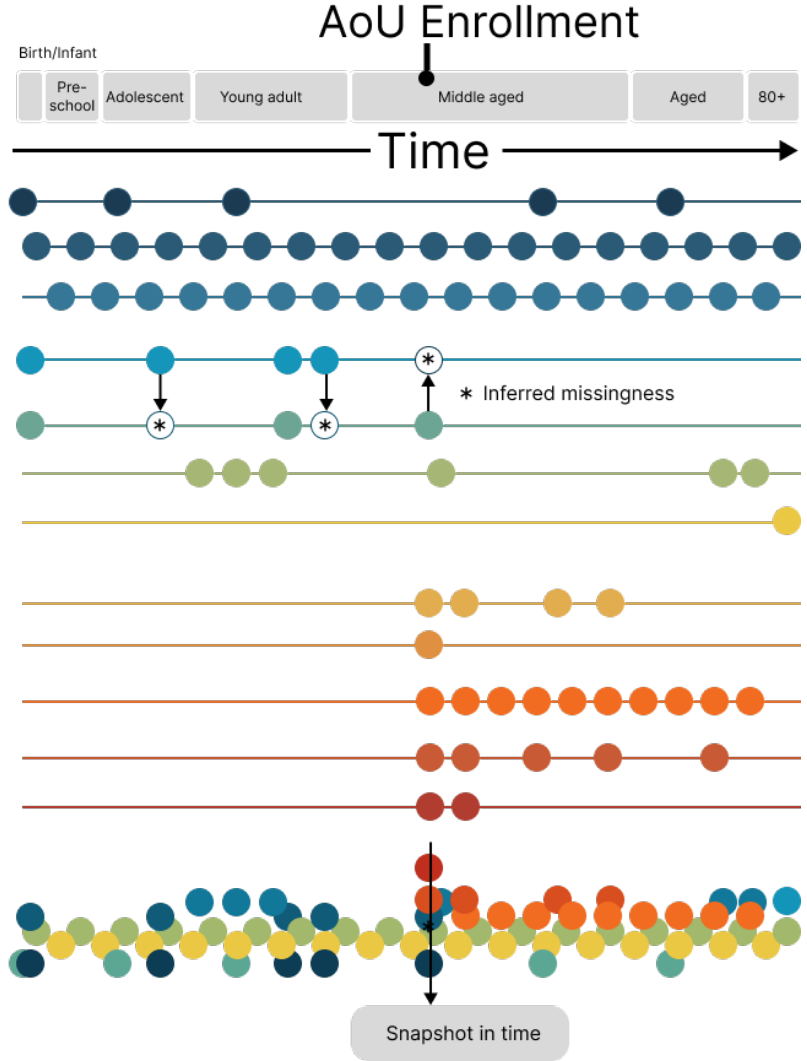


- Data about individuals is scattered across health systems and numerous sources
- A comprehensive dataset for each individual's life-course could improve health and research
- Many hurdles: different formats and standards, regulatory and legal, security and technical compatibility

CLAD Vision: putting the patient back together again

New data acquisition methods are needed to acquire and link data to each person:

- Patient-Privacy Preserving Record Linkage (PPRL, de-identified token-based linkage)
- Geo-spatial & temporal linkage
- HIE/HIN identified linkage to acquire EHR data for research

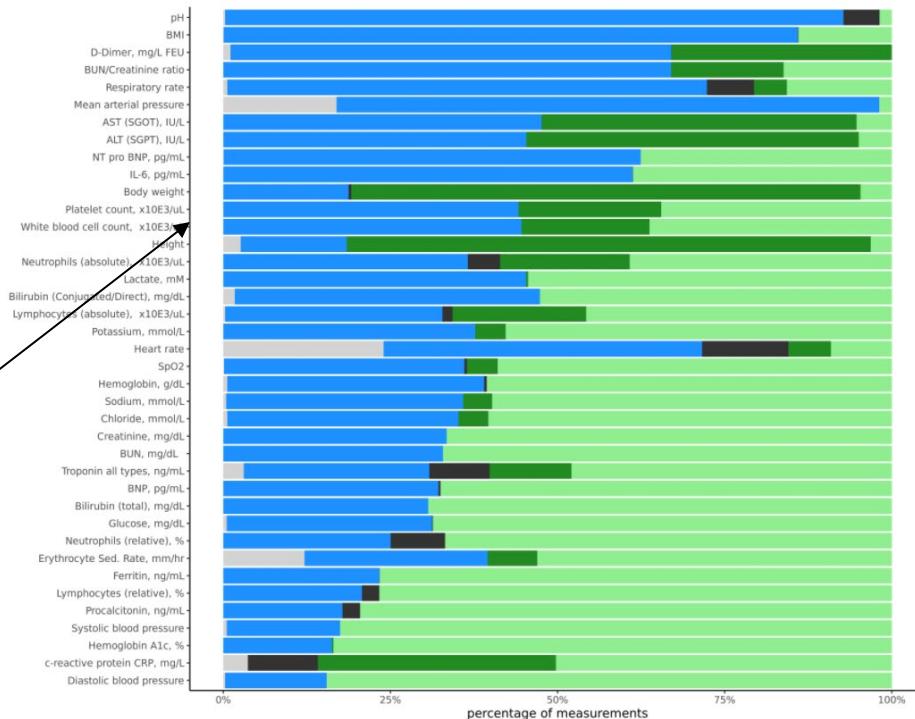


Algorithmic data repair: made more feasible with many sites' EHR data - and helps improve the sources

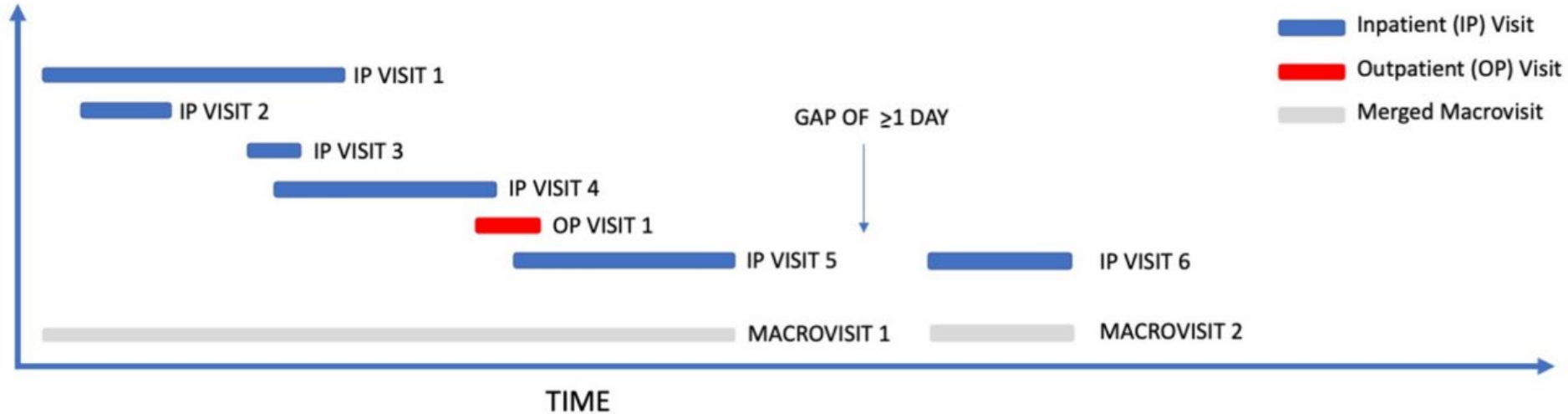
Humans measured in **grams** do not look the same as humans measured in **kilograms**!



Canonical unit
Uses a known conversion
Unit not plausible
Missing unit inferred
Unit still missing



Macrovisits: harmonizing visits across healthcare systems

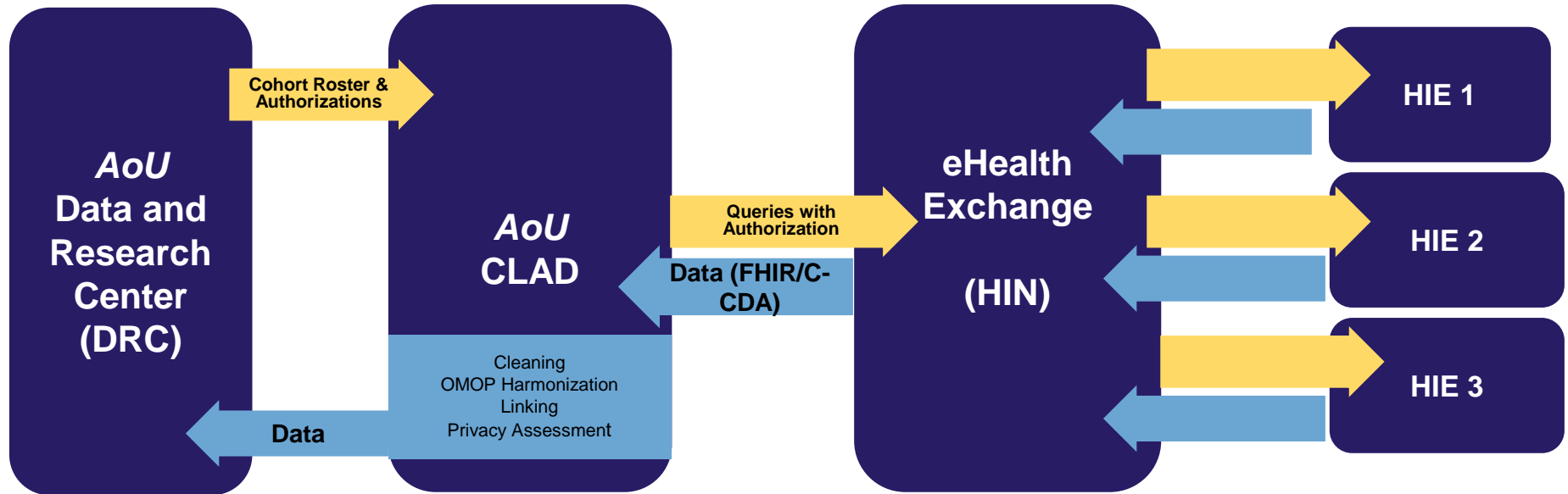


Microvisit to Macrovisit Map is an augmented version of the OMOP visit_occurrence table

[10.1093/jamia/ocad057](https://doi.org/10.1093/jamia/ocad057)

Using HIE/HINs to acquire EHR data for research: a Pilot

- >90% of *AoU* participants consent to share EHR data, yet only ~69% have records in *AoU*
- Current methods of collecting EHR data are **expensive** and **fragmented**
- Acquire national EHR data from Health Information networks (HINs) and Health Information Exchanges





Included checks are:

- Total persons
- Persons with birth dates prior to 1900
- Persons with birth dates after the current year
- Breakdown of reported sex
- Month of birth distribution
- Breakdown of race
- Breakdown of ethnicity

Screenshot of the Person tab in the FHIR to OMOP Data Quality Portal

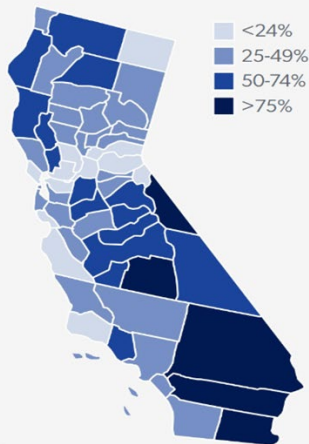
Match rates with Manifest MedEx – a California statewide HIE



Data to Serve Every County in California

The MX network had an active health record for more than 18 million Californians in 2023 alone, demonstrating the opportunity for a digital health data safety net that bridges large urban centers and rural communities alike.

Active Records on MX Network in 2023 as a Percent of County Population*



Growth in Top 10 Largest California Counties

Here are a few highlights of our growth in the 10 largest counties in California, totaling more than 22 million records in 2023.

County by size (largest to smallest)	Unique records in 2022	Unique records in 2023	Percent increase in records
1. Los Angeles	5,651,248	7,368,789	30.4%
2. San Diego	1,991,270	2,719,015	36.5%
3. Orange	1,774,294	2,244,857	26.5%
4. Riverside	2,599,110	3,242,770	24.8%
5. San Bernardino	2,280,161	2,820,915	23.7%
6. Santa Clara	820,726	957,520	16.7%
7. Alameda	705,078	816,936	15.9%
8. Sacramento	670,685	803,474	19.8%
9. Contra Costa	382,817	433,444	13.2%
10. Fresno	924,599	1,223,981	32.4%

Increase in Records According to Last Known Addresses on MX Network by County From 2022 to 2023

Participant Match Query/ Response to Date

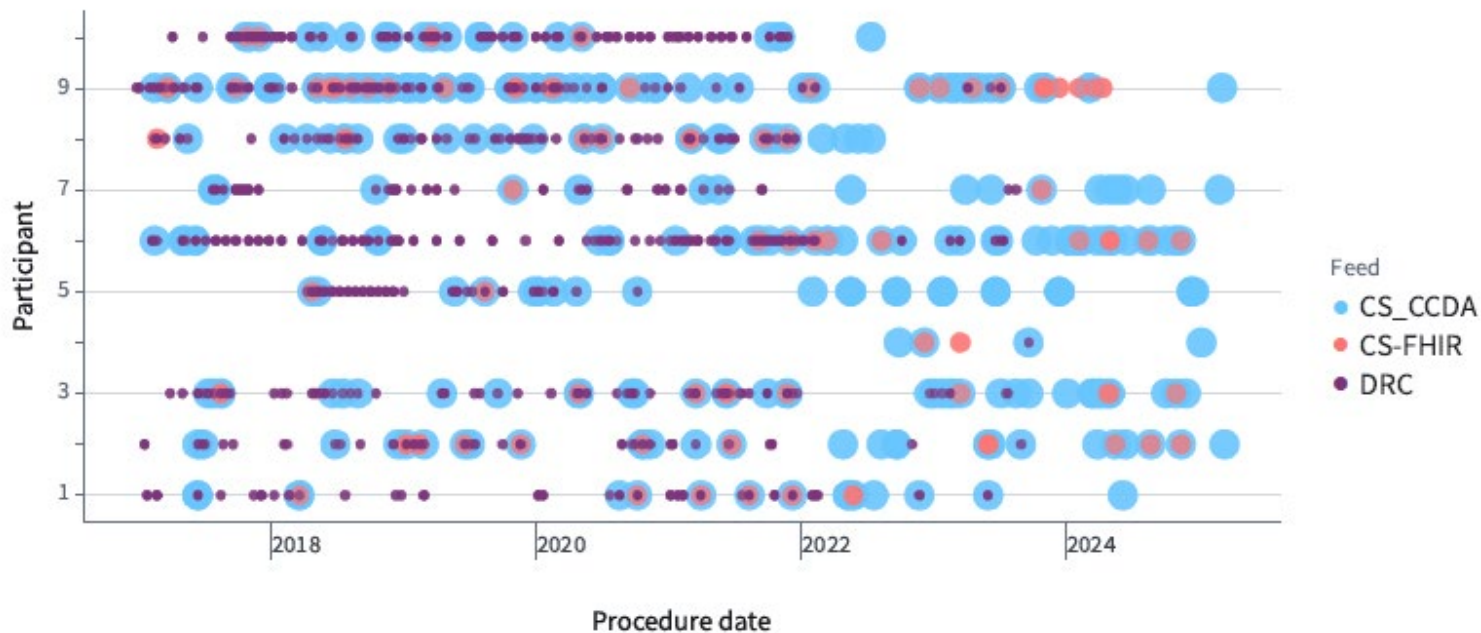
Queries	Matches	%
77,995	50,685	65%

Query for Records / Response to Date

# of Queries for Documents	# of AoU participants with C-CDAs* retrieved from Manifest MedEx	%
50,685	49,101	97%

*Manifest MedEx consolidates data from all its contributing sources into a consolidated C-CDA per participant

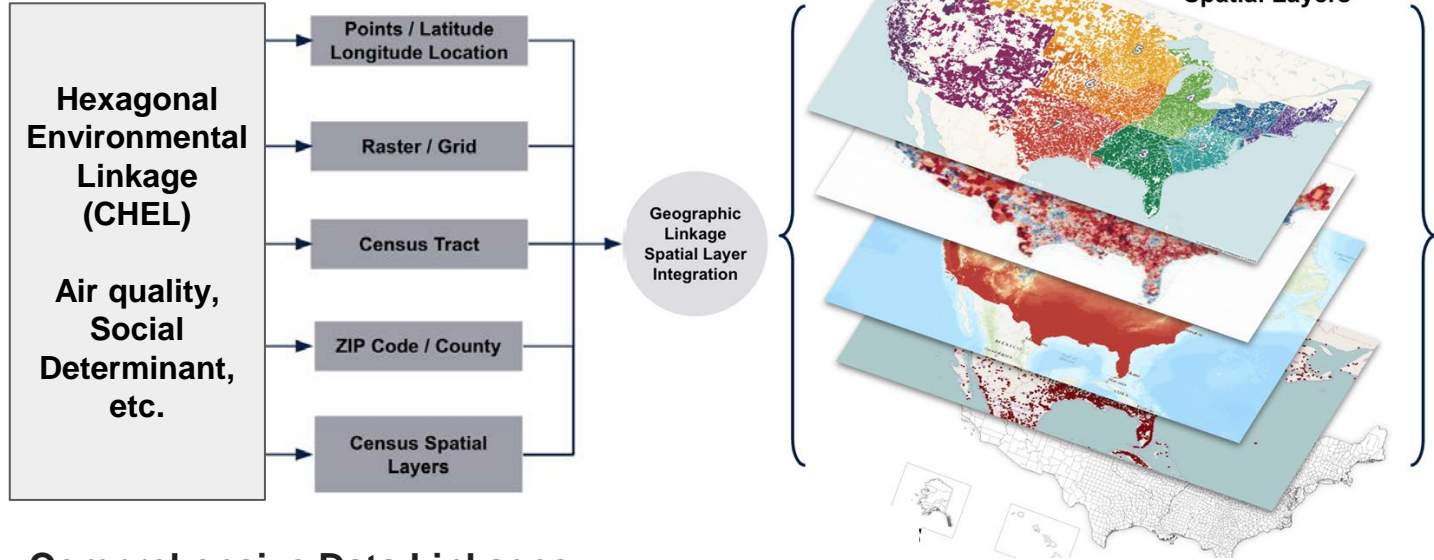
Scalable FHIR & CCDA feeds complement EHR data sent directly



10 random patients' procedures shown from one HPO

C-CDA and FHIR payloads eclipse the DRC's procedure coverage, with more recent data

Linking Place- Based Environment and Health Data

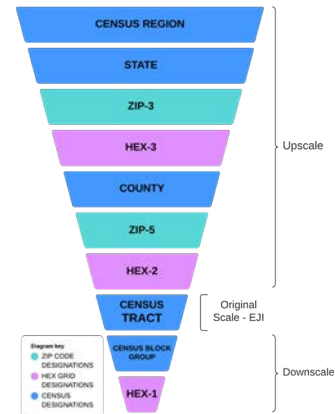


Comprehensive Data Linkages

- Hierarchical designations—Integrates census & non-census data
- Layered geocoding—More reporting opportunities
- Multiple spatial layers—Overcomes irregular/sparse census tracts or ZIP Codes

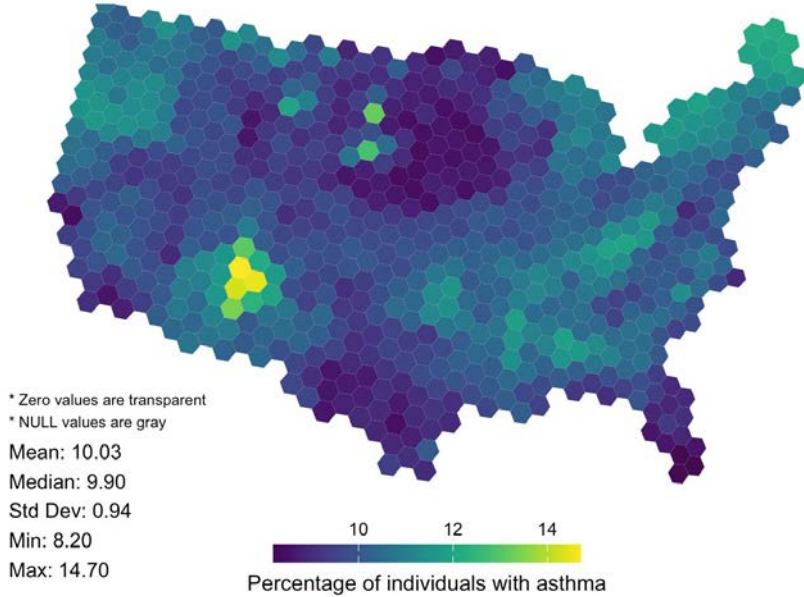
Risk of Re-identification Assessment

- Analytical risk determination over all linked data to determine risk

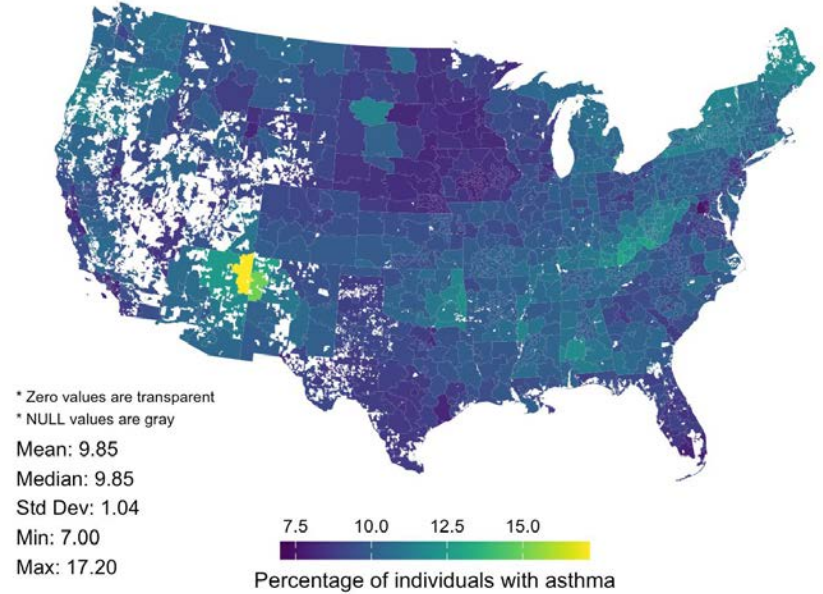


Benefits of Hex 3 in Rural Areas | Avoid dead zones

Hex 3 Aggregation



ZIP 3 Aggregation



What about derived variables/CDEs?

Distributed Data



Aggregated Data



Integrated Data



>1800 DATABASES

>1500
STANDARDS

>900 ONTOLOGIES

>23K CDES

- Many overlapping CDE repositories (NLM, caDSR, PhenX, Heal, and more)
- Context and metadata are implied - including mappings, provenance, etc

Search Term	CDEs in NLM	Ontologies in BioPortal
Date of Birth	28	49
Blood Pressure	104	33
Smoking Status	22	48

Example: Many CDEs for Blood Pressure

Blood Pressure measurement

Blood pressure measurement with systolic measurement over diastolic measurement

[Qualified](#)

Steward: NINDS
Used By: NINDS
Source: NINDS

Blood pressure systolic measurement

Measurement of **pressure** of the participant's/subject's **blood** against the artery walls during systole (the contraction phase) in millimeters of mercury

[Qualified](#)

Steward: NINDS
Used By: NHLBI, NINDS
Source: NINDS

Blood pressure diastolic measurement

Measurement of **pressure** of the participant's/subject's **blood** against the artery walls during diastole (the relaxation phase) in millimeters of mercury

[Qualified](#)

Steward: NINDS
Used By: NHLBI, NINDS
Source: NINDS

Blood pressure mean measurement

Mean measurement of the participant's/subject's **blood pressure**

[Qualified](#)

Steward: NINDS
Used By: NINDS

Label	Code	ConceptID
< 120/70		
120 - 140/70 - 90		
< 140/> 90		
> 140/< 90		

Blood Pressure measurement

Question Text

Submitter did not provide a Question Text

Definition

Blood pressure measurement with systolic measurement over diastolic measurement

Data Type: Number

Steward: NINDS

Origin:

Vital Signs Type

A textual description of a person's vital signs measurement category.

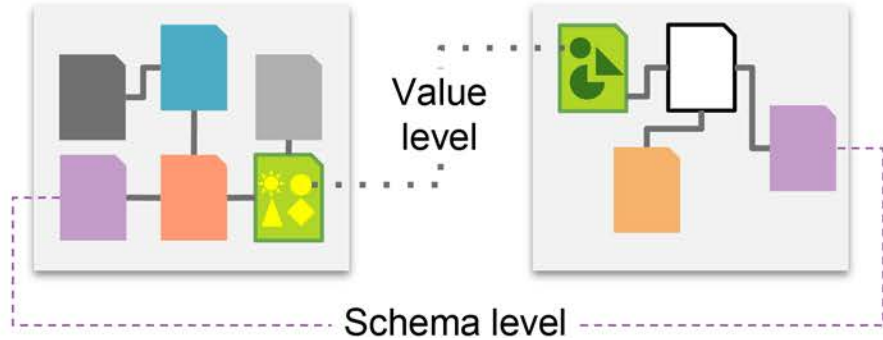
[Qualified](#)

Steward: Project 5 (COVID-19)
Used By: Project 5 (COVID-19)

Label	Code	ConceptID
Systolic blood pre...		C25298
Diastolic blood pr...		C25299
Heart rate		C49677
Respiratory rate		C49678

(8 total) See full table in [Detail View](#)

Mapping CDEs (And More) Is Necessary For Interoperability



Data model alignment: Each source models things differently

For example, no direct link from Sample-to-Diagnosis in one model

Would need to “remodel” Sample-to-Case, and Diagnosis-to-Case to align with Sample-to-Diagnosis

Value Set alignment:

Each source uses different values

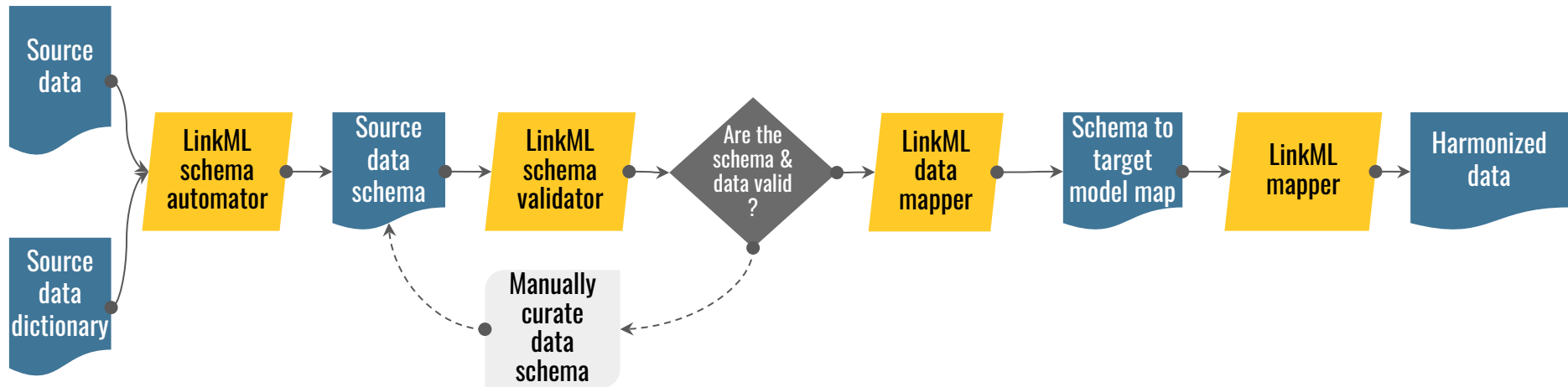
For example, one node encodes race like this:

- not reported
- white
- american indian or alaska native
- black or african american

While another does it like this:

- not allowed to collect
- unknown
- white
- native hawaiian or other pacific islander
- american indian or alaska native
- asian
- other
- black or african american

Designing an ingest & harmonization pipeline for clinical research

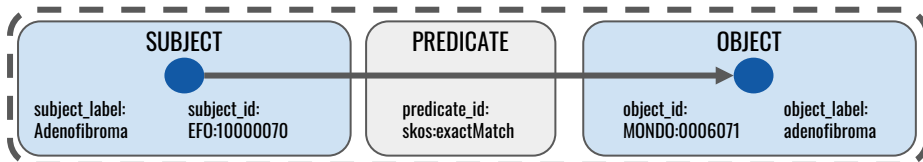


<https://github.com/linkml/dm-bip>

Semantizing CDEs: LinkML modeling + SSSOM mapping



sssom
SIMPLE STANDARD FOR SHARING
ONTOLOGY MAPPINGS



JUSTIFICATION

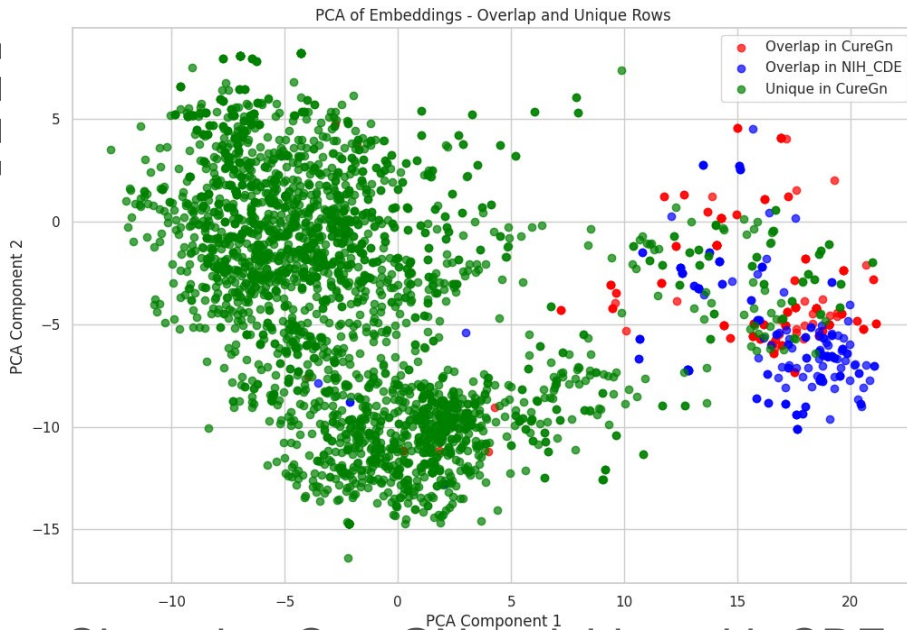
mapping_justification:
semapv:LexicalMatching

subject_match_field: rdfs:label
object_match_field: oio:hasExactSynonym
match_string: adenofibroma
mapping_date: 2022-12-13
reviewer_id: orcid:0000-0002-7356-1779
mapping_tool: wikidata:Q64360017
confidence: 0.8

Documents provenance & rules for terminological and value set mappings

<https://mapping-commons.github.io/sssom/tutorial/>

<https://doi.org/10.1093/database/baac035>



Clustering CureGN variables with CDEs

<https://linkml.io/>

Harmonizing data across cohorts: final step is ontology annotation for search and classification



Cohort A

LabID	Race	Atrial septal defect
HTP0888A	>1 race	TRUE

Raw Data



Cohort B

patient_id	race_white	race_asian	Which of the following congenital heart defects have been diagnosed? (Select all that apply.) - Atrial septal defect (ASD)
3162	1	1	1

Raw Data



Cohort C

Participant ID	Race	Phenotypes Text
KFDS1343676948	White; Asian	DEVELOPMENTAL DELAY; GASTROESOPHAGEAL REFLUX; HYPOTHYROIDISM; REACTIVE AIRWAY DISEASE; SECUNDUM ATRIAL SEPTAL DEFEC ; SLEEP APNEA; SMOKER IN FAMILY; TRISOMY 21

Raw Data

Harmonica  Ontology Annotation

INCLUDE Data Hub

Participant Global ID	Participant External ID	Race	Condition (Source Text)	Diagnosis (MONDO)	Phenotype (HPO)
pt-dgcr2bykx8	HTP0888A	More than one race	Atrial septal defect	MONDO:0006664	HP:0001631
pt-fihjpm84r	3162	More than one race	Atrial septal defect (ASD)	MONDO:0006664	HP:0001631
pt-fdagpi2fwb	KFDS1343676948	More than one race	Secundum atrial septal defect	MONDO:0006664	HP:0001631

Transformed
Data

Harmonizing equivalent data elements across Down syndrome cohorts.

Our goal is to enable within- and cross-cohort querying and multi-modal analytics

Challenges in standardizing RWD and CDEs



- Real data is heterogeneous despite using Common Data Models (e.g. OMOP, PCORnet) and Exchange formats (e.g. FHIR, CCDA)
- Algorithmic repair and data quality improvement benefits from sites working together
- Piggybacking research on HIN/HIEs designed for care can efficiently scale acquisition (TEFCA!)
- FHIR is easier to transform to OMOP, but is less available than CCDA from healthcare systems
- Geospatial data can be linked to RWD with Hex3 grids and maintain privacy
- CDEs can be “semanticized” for analytic use and *a priori* improved standardization using LinkML tools and improved mappings

Acknowledgements



National
Clinical
Cohort
Collaborative



LEADS

Chris Mungall

Anne Thessen
Bryan Laraway
Harold Solbrig
Corey Cox
Eric Hurwitz
Harry Caufield
Justin Reese
Kevin Schaper
Matt Brush
Madan Krishnamurthy
Nico Matentzoglou
Patrick Golden
Sarah Gehrke
Sierra Moxon
Trish Whetzel
Tursynay Issabekova



**Chris Siege
Anne Thessen**

Bryan Furner
Corey Cox
David Beaumont
Oswaldo Lozoya
Patrick Golden
Sigfried Gold
Sabrina McCutchan



INCLUDE

Pierrette Lo
Madan Krishnamurthy
Matt Brush
Trish Whetzel
Tursynay Issabekova

**Chris Chute
Emily Pfaff
Richard Moffitt**

Adit Anand
Andrew Girvin
Anita Walden
Bryan Laraway
Chris Roeder
Dave Eichmann
Kate Bradwell
Kristin Kostka
Jason Yoo
Julie McMurry
Matthew Owens
Peter Leese
Sofia Dard
Sruthi Magesh
Stephanie Hong
Tursynay Issabekova

**Chris Chute
Emily Pfaff
Richard Moffitt
Charisse Madlock-
Brown**

Adam Lee
Andrew Laitman
Andrew Poley
Anita Walden
Anne Bailey
Bren Becker
Brian Cass
Bryan Laraway
Chris Roeder
Dan Angelelli
Daniel Barth-Jones
Gabi Duhon
James Cavallon
Jaron Lee
Jimmy Phuong
Josh Lemieux
Julie McMurry
Karen Albright
Kristen Hansen
Lakshmi Anandan
Lisa Eskenasi
Margaret Hall
Matthew Lehnert
Matthew Owens
Matthew Pagel
Monique Bangudi
Patrick Baier
Paul Matthews
Rae Crist
Richard Zhu
Rishi Limaye
Rob Schuff
Shahim Essaid
Sofia Dard
Sruthi Magesh
Stephanie Hong
Tanner Zhang
Taz Khan
Tricia Francis
Thanaphop Na
Nakhonphanom
Tursynay Issabekova
William Hogan
Yvette Chen