

Evaluating Semantic Loss in EHR Terminology Mapping in Common and Rare Diseases

Melanie Philofsky[1], Bryan J Laraway[2], Michael G Kahn[3], Melissa A Haendel[2]

[1]Epam Systems, [2]University of North Carolina, [3]University of Colorado

Background



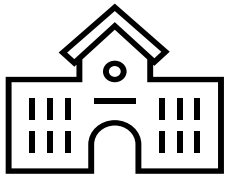
Interface terminologies

– IMO, includes a subset of Mondo



Reporting & administrative terminologies

– ICD10CM, CPT4



Clinical terminologies; comprehensive and specialized

– SNOMED; Mondo & Orphanet

Problem Statement

Previous Work: Demonstrated semantic loss in mapping IMO interface terminology to SNOMED CT (direct vs. indirect).

Current Study Scope: Evaluates generalizability across a broader spectrum of diseases.

What is Mondo?



https://mondo.monarchinitiative.org

- Rare disease ontology
- Partially incorporated into IMO
  - Available to clinicians in EHR

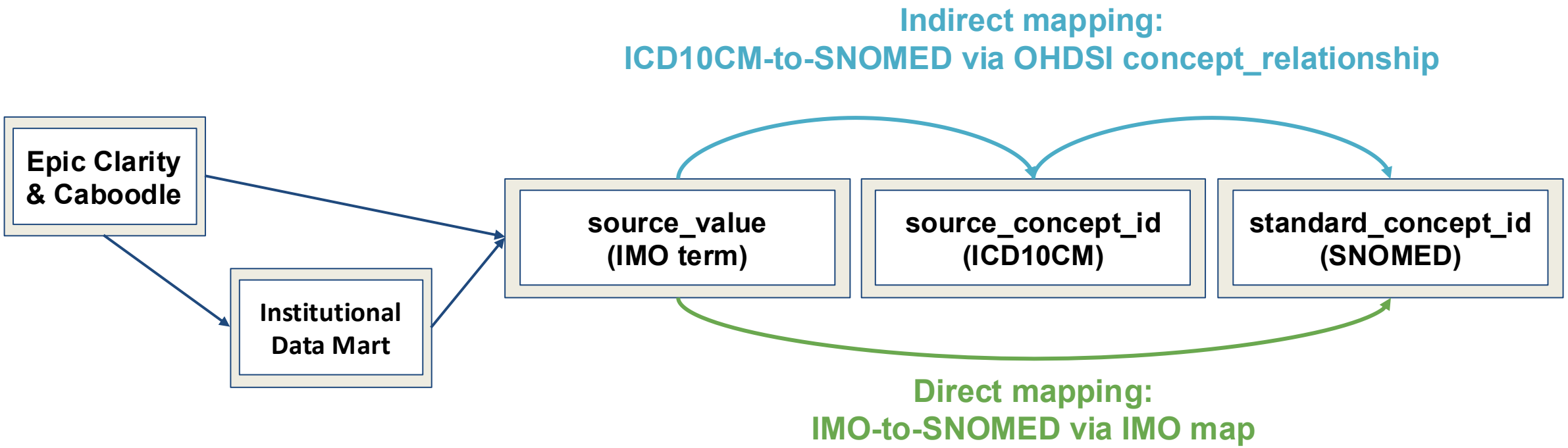
Objective

- **Objective:** Inform best practices for terminology mapping in observational research
  - **Scope Expansion:** Encompasses both common and rare conditions
  - **Multi-Site Focus:** Evaluates distinct healthcare settings, workflows, and diagnoses

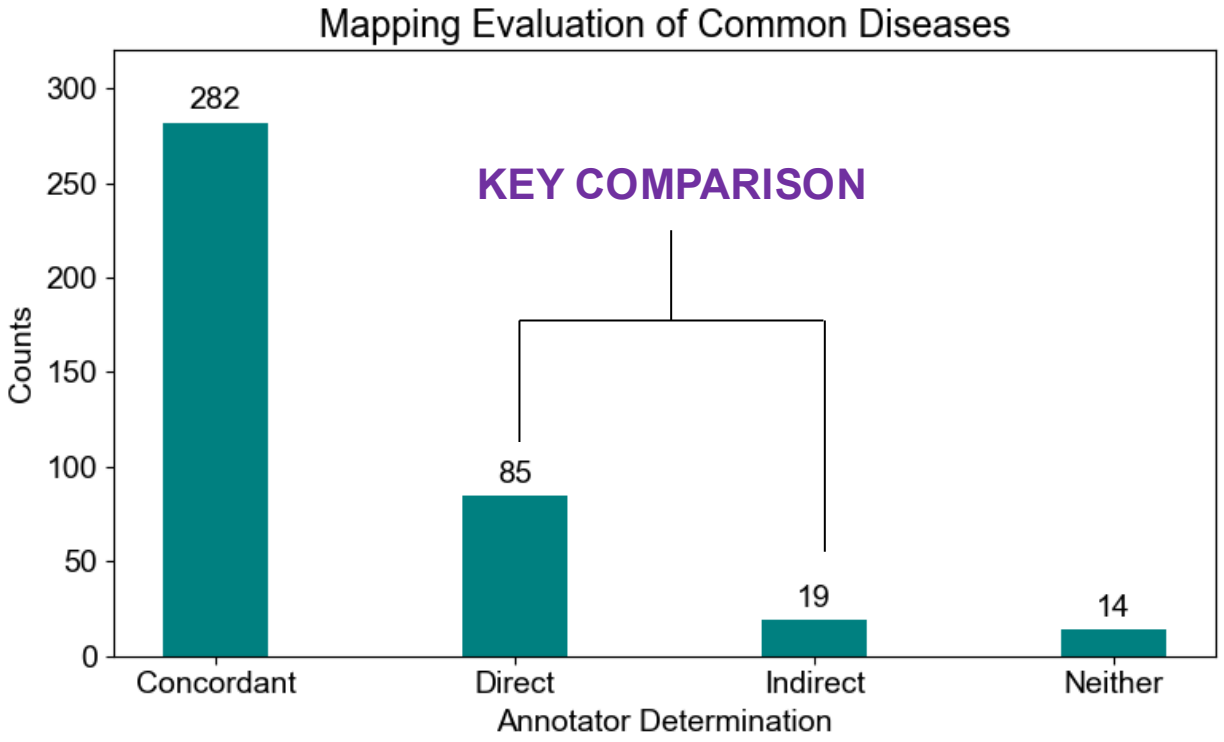
Methods

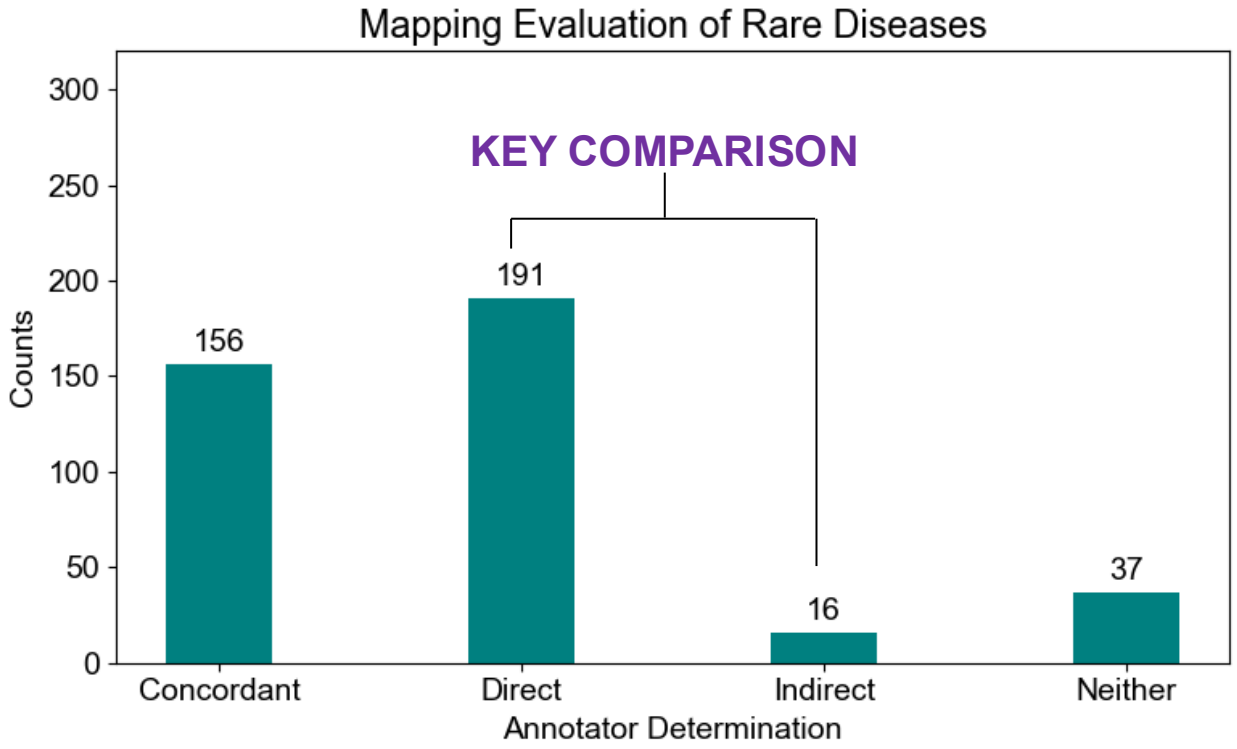
- **Identification:** Mapped IMO terms to rare conditions in the Mondo Disease Ontology
- **Cohort Selection:**
  - 200 randomly sampled rare conditions
  - 200 ‘common’ conditions

Direct vs Indirect mapping method

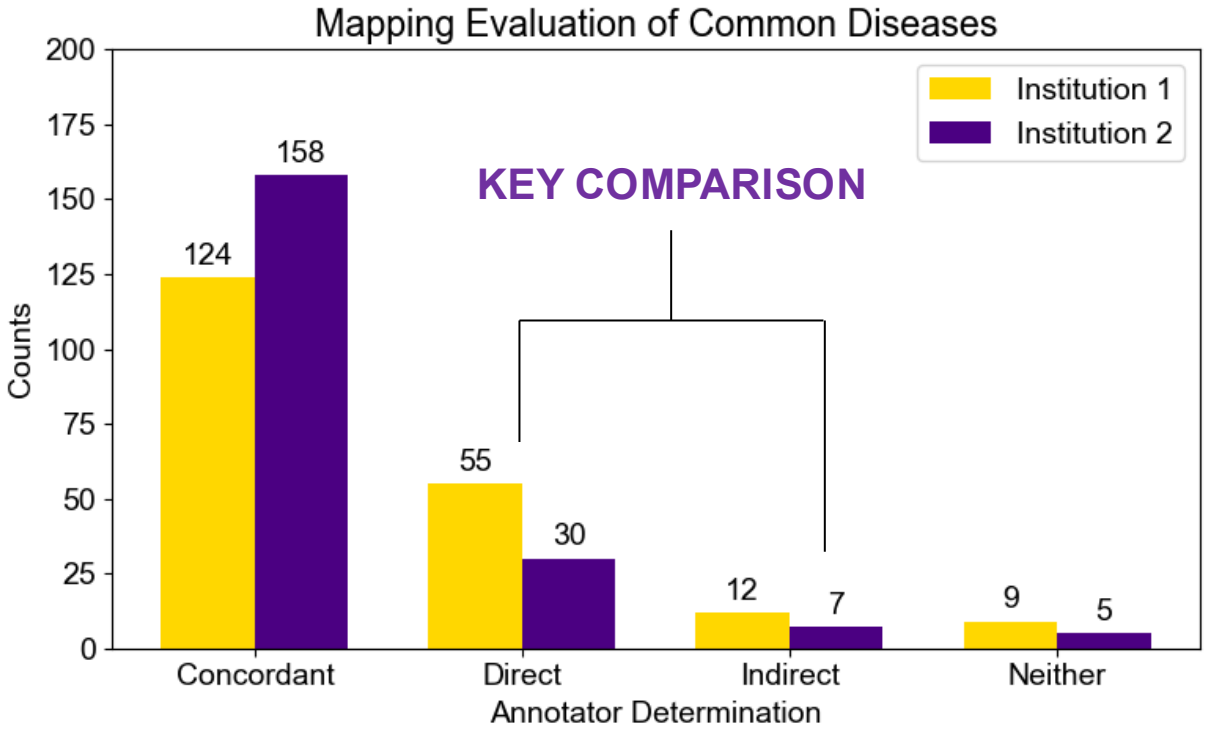


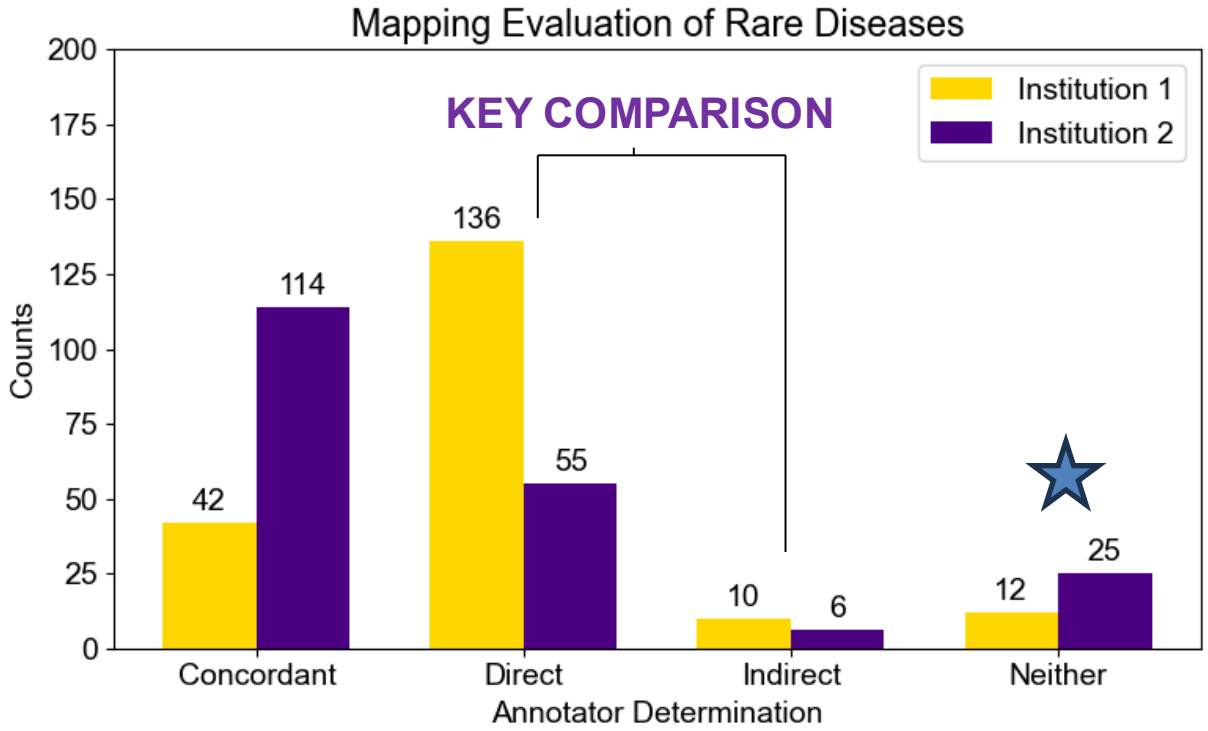
Direct Mapping Preferred for Rare and Common





Direct Mapping Preferred Across Institutions





Key Takeaways

The more direct method of mapping the terminology selected by front-line clinicians ensures that the resulting standard concept more closely captures the intended meaning of the patient-provider interaction.

When neither mapping is acceptable, the EHR source term is the most granular term for rare diseases.

Limitations

- The study used OHDSI's February 2026 vocabulary version
- Differing ETL processes
- Utilized term matching to identify rare diseases

Future Directions

- Bring the topic to Themis
- Recommend all sites with IMO clinical interface terminology utilize the direct approach when mapping to SNOMED
- When neither mapping is acceptable, what should we do?
  - Collaborate!
- Cohort comparison study – does this affect RWE generation?

Acknowledgements

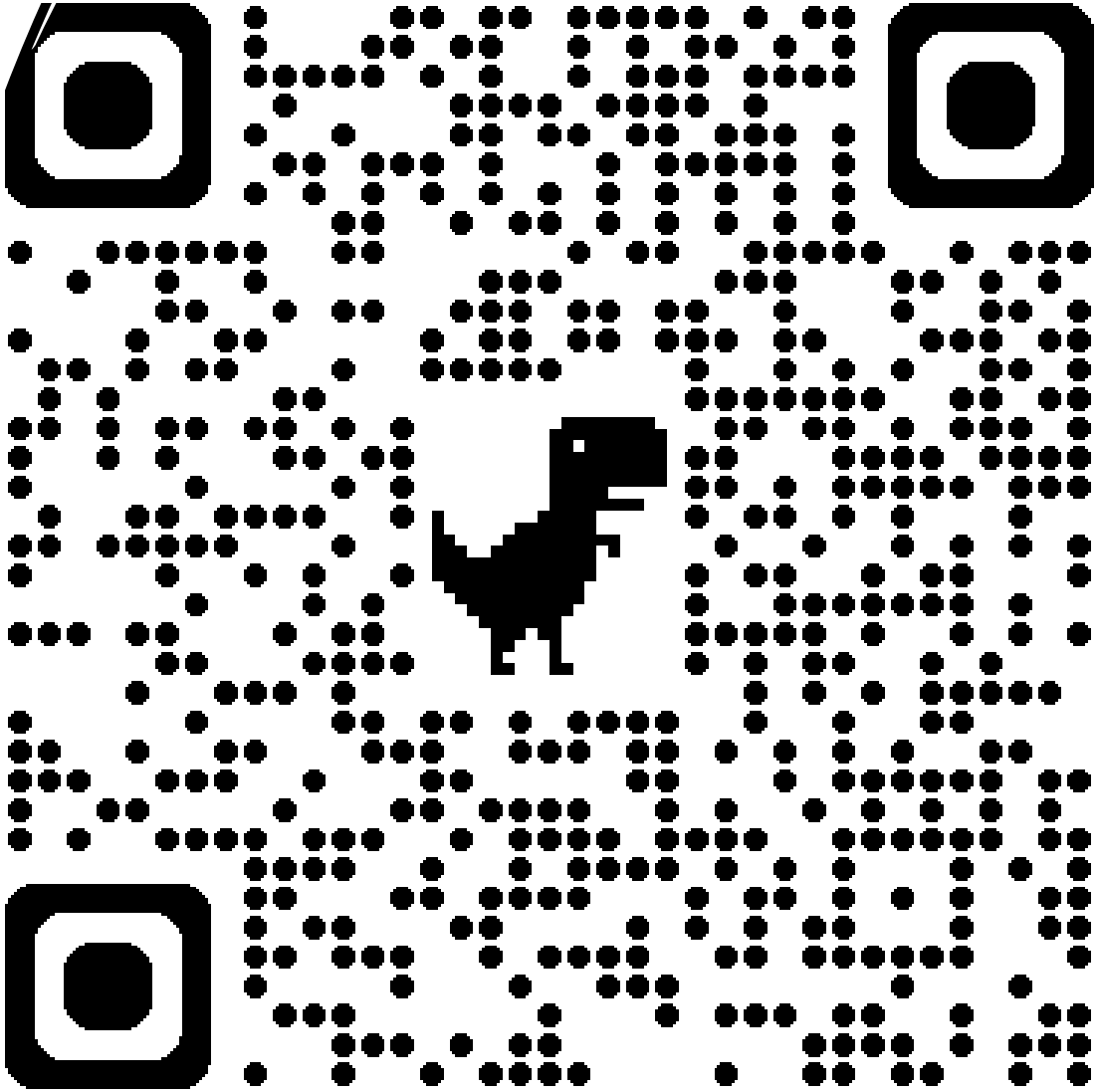
- **Co-Authors:** For your dedication, insightful discussions, and partnership.
- **Annotators:** For your meticulous evaluations and rigorous data annotation.
- *This work would not have been possible without your expertise and collaborative effort.*

Thank you!

Questions?

Come to poster

154



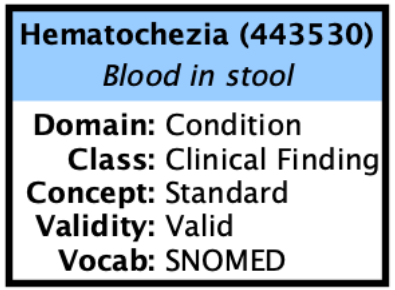


Using OWL Reasoning to Surface Hidden Tensions in the OHDSI Standardized Vocabularies

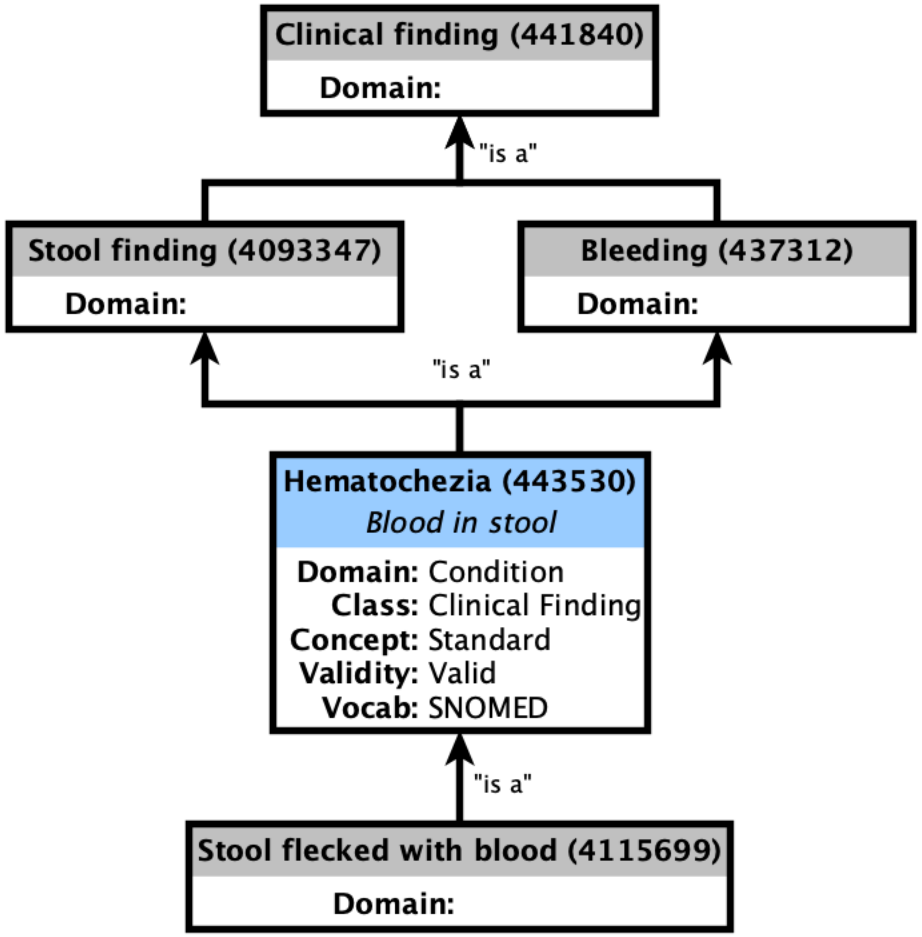
William Baumgartner Jr and Lisa Schilling

University of Colorado Anschutz

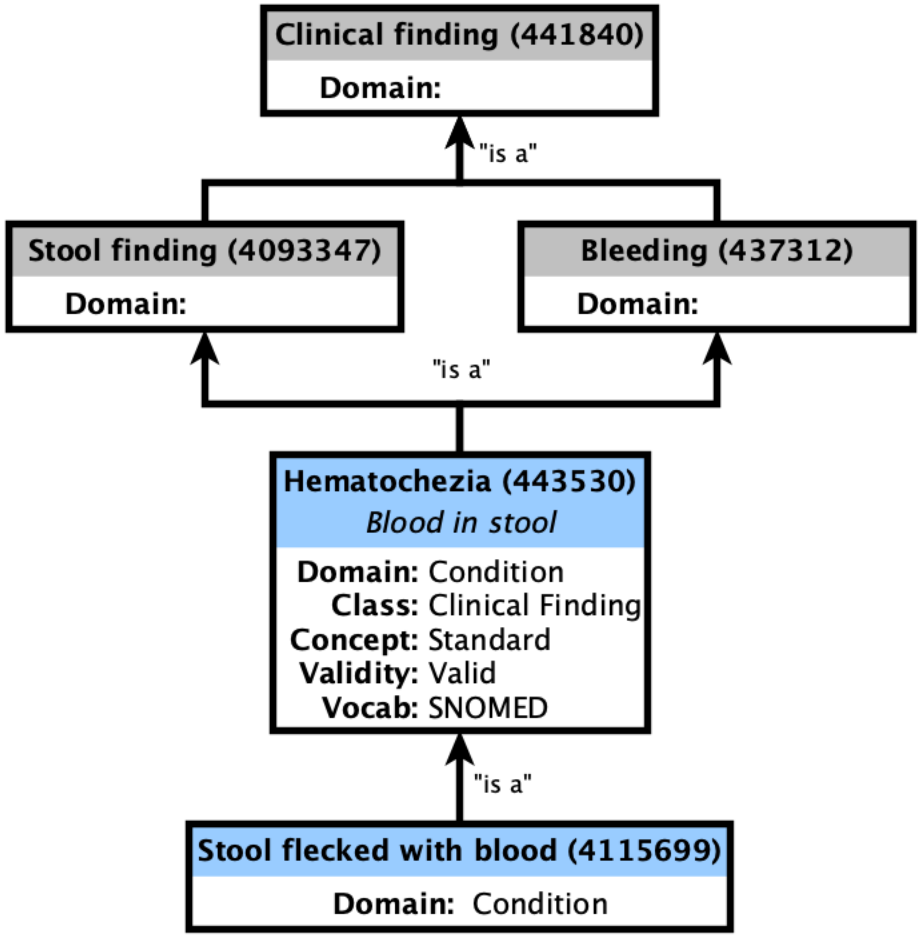
Domain tension example: Hematochezia



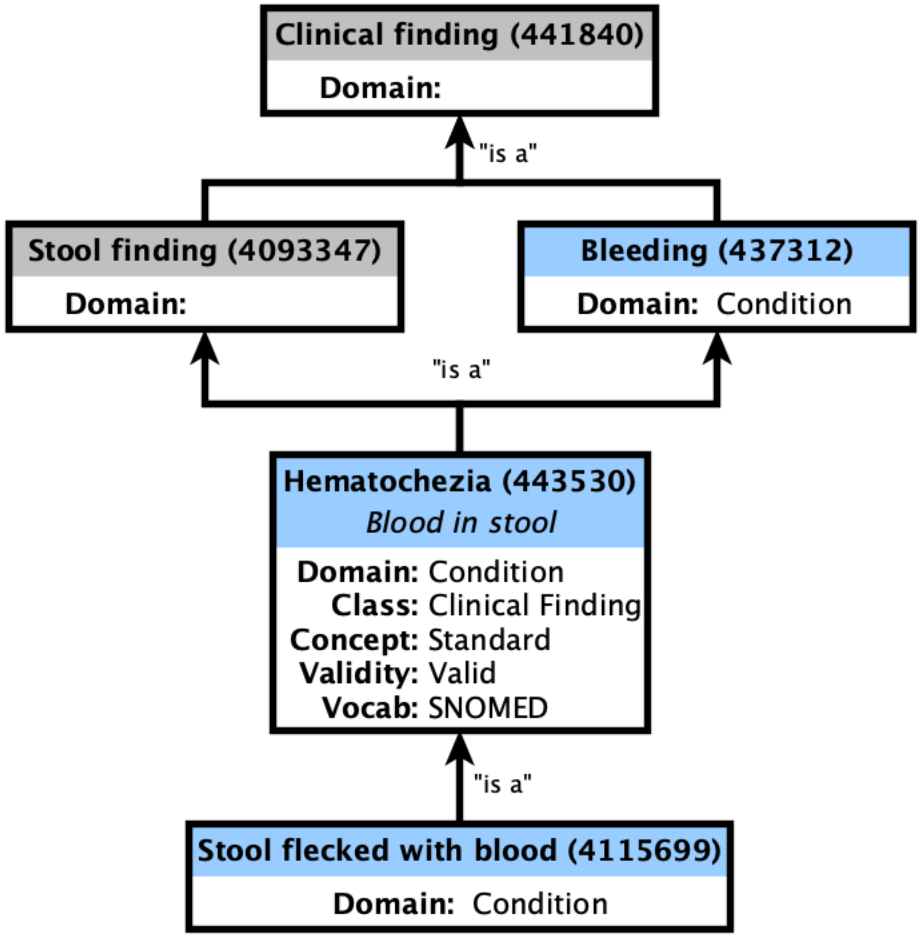
Domain tension example: Hematochezia



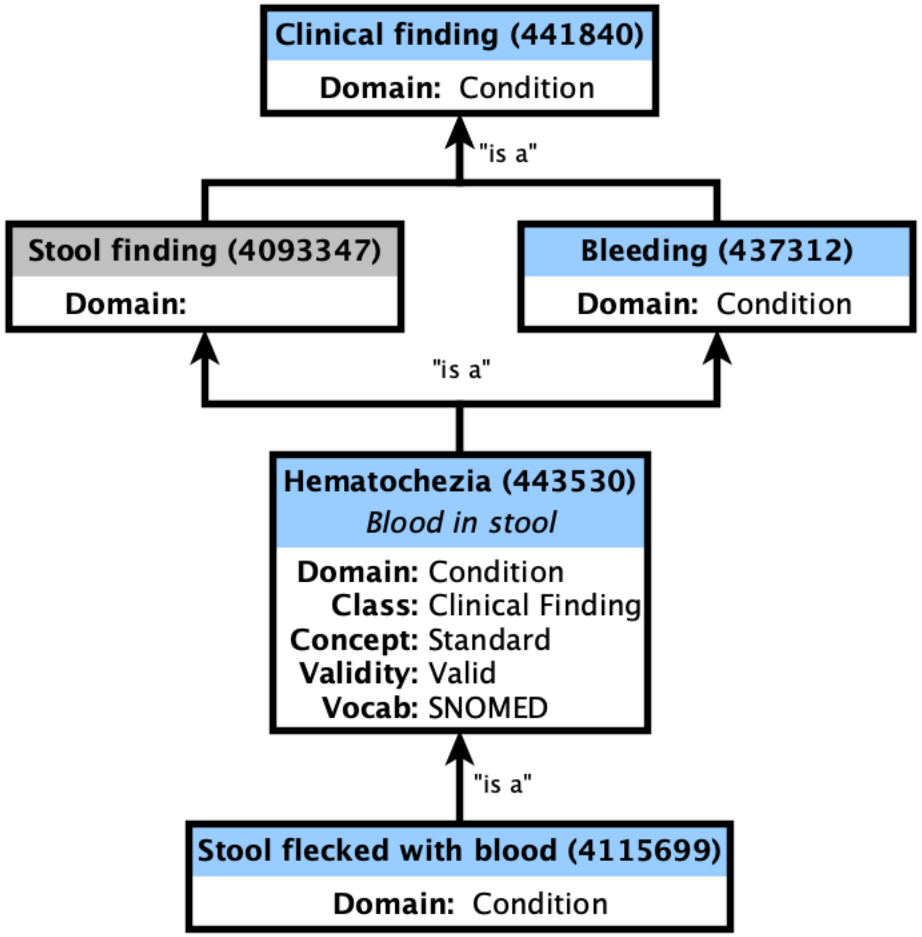
Domain tension example: Hematochezia



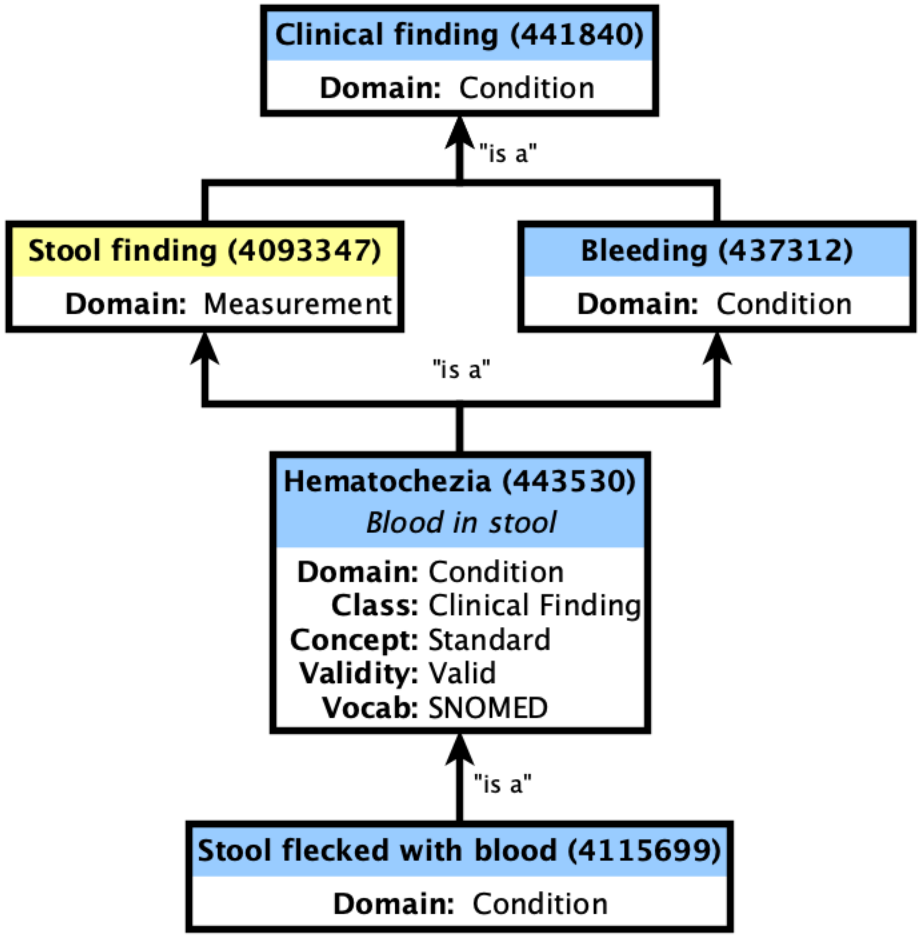
Domain tension example: Hematochezia



Domain tension example: Hematochezia



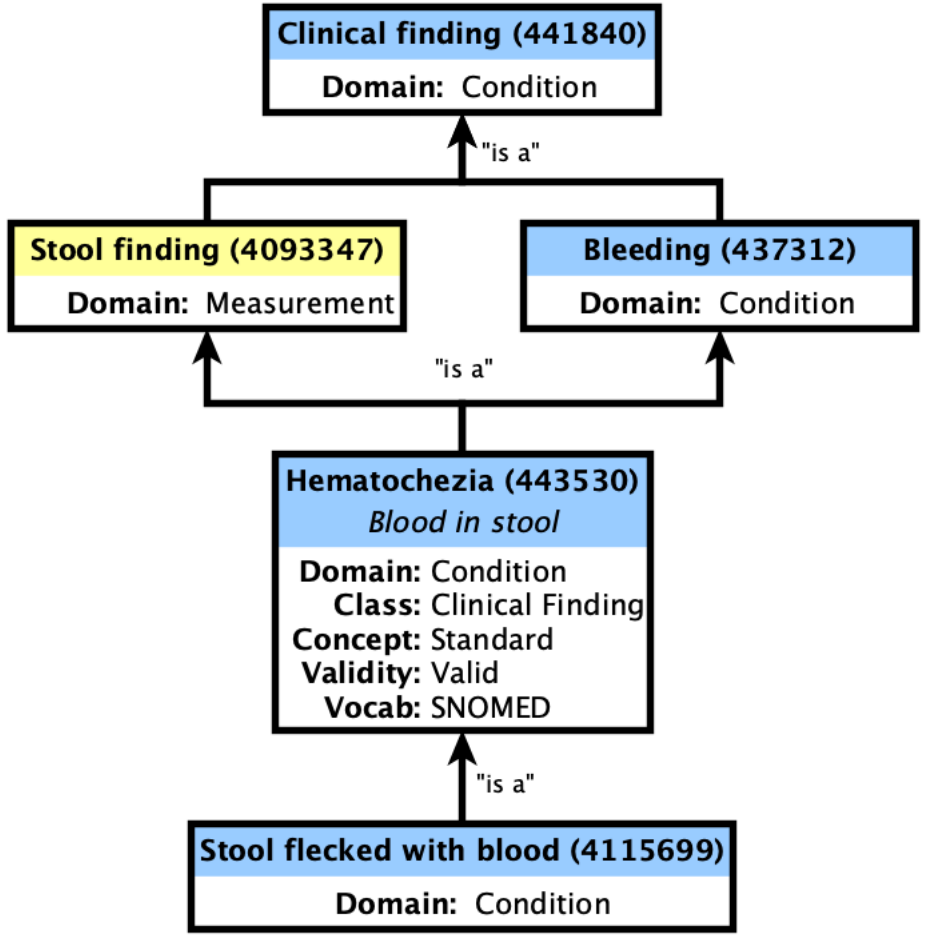
Domain tension example: Hematochezia



What is a vocabulary tension?

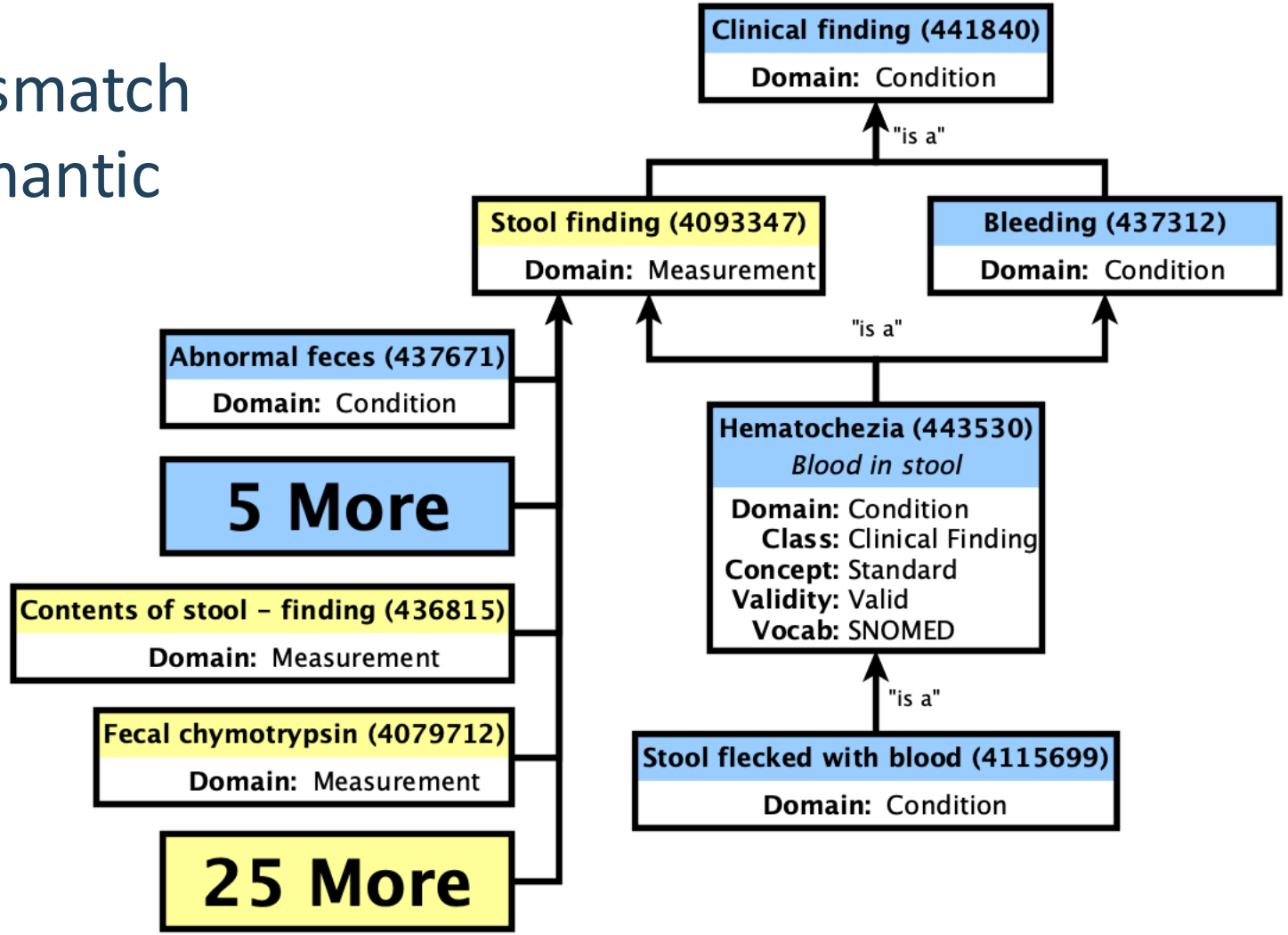
- **Vocabulary tension** = a mismatch between structure and semantic expectations

- **Structure:** A *Measurement*-domain concept occurs within a *Condition* inheritance hierarchy.
- **Expectation:** Concepts in a hierarchical neighborhood should generally convey a consistent semantic story.
- **Result:** Vocabulary tension

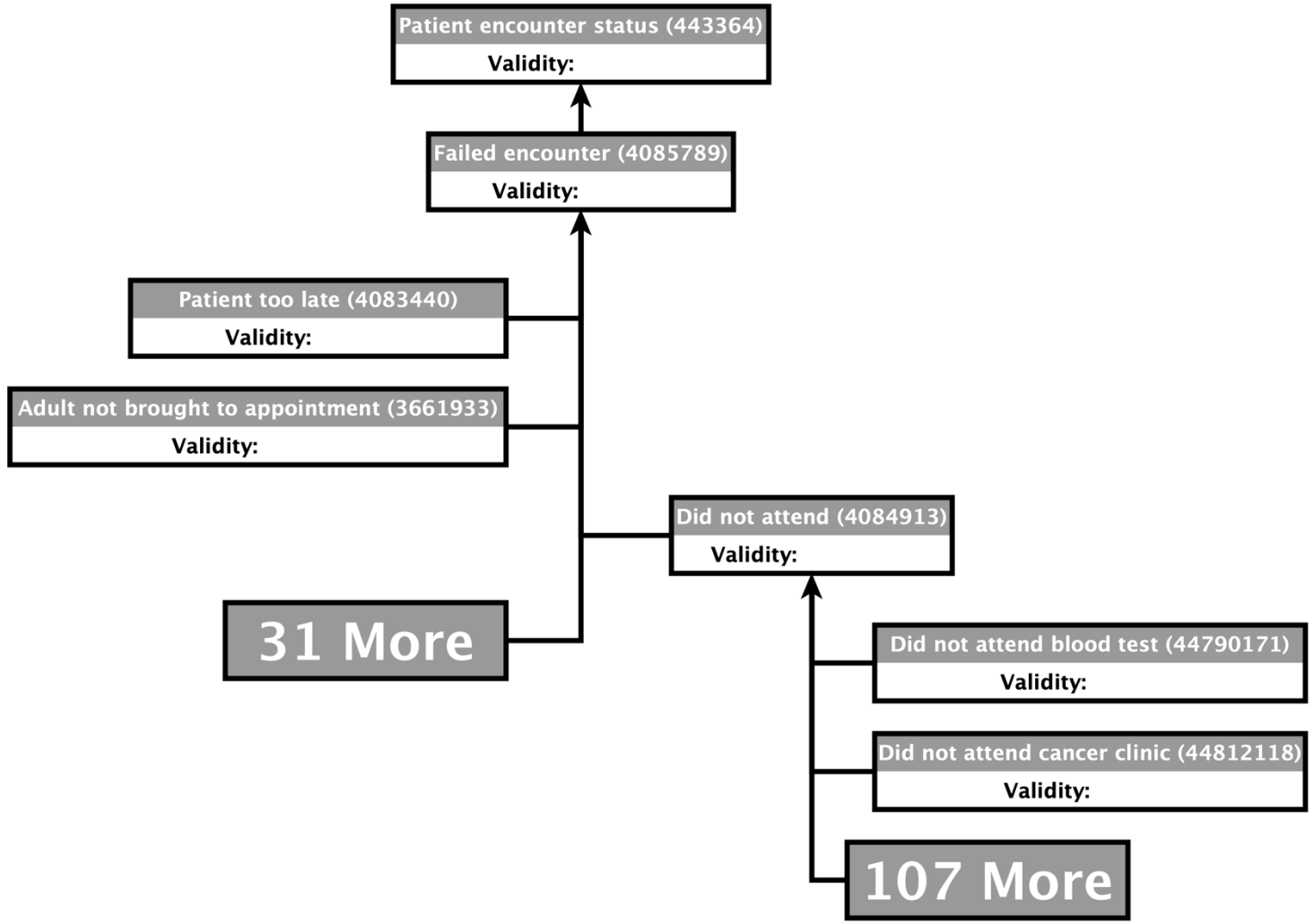


What is a vocabulary tension?

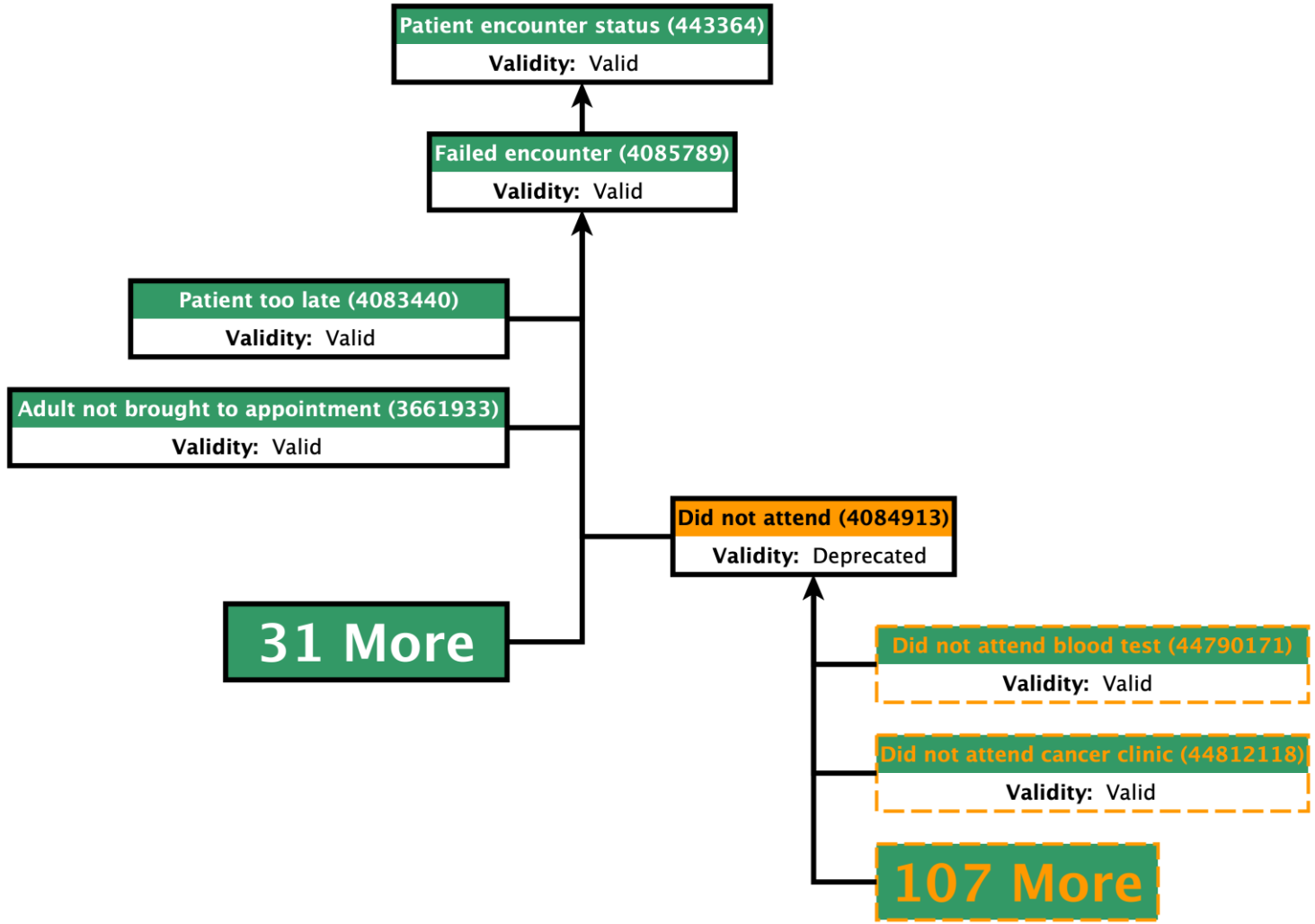
- **Vocabulary tension** = a mismatch between structure and semantic expectations



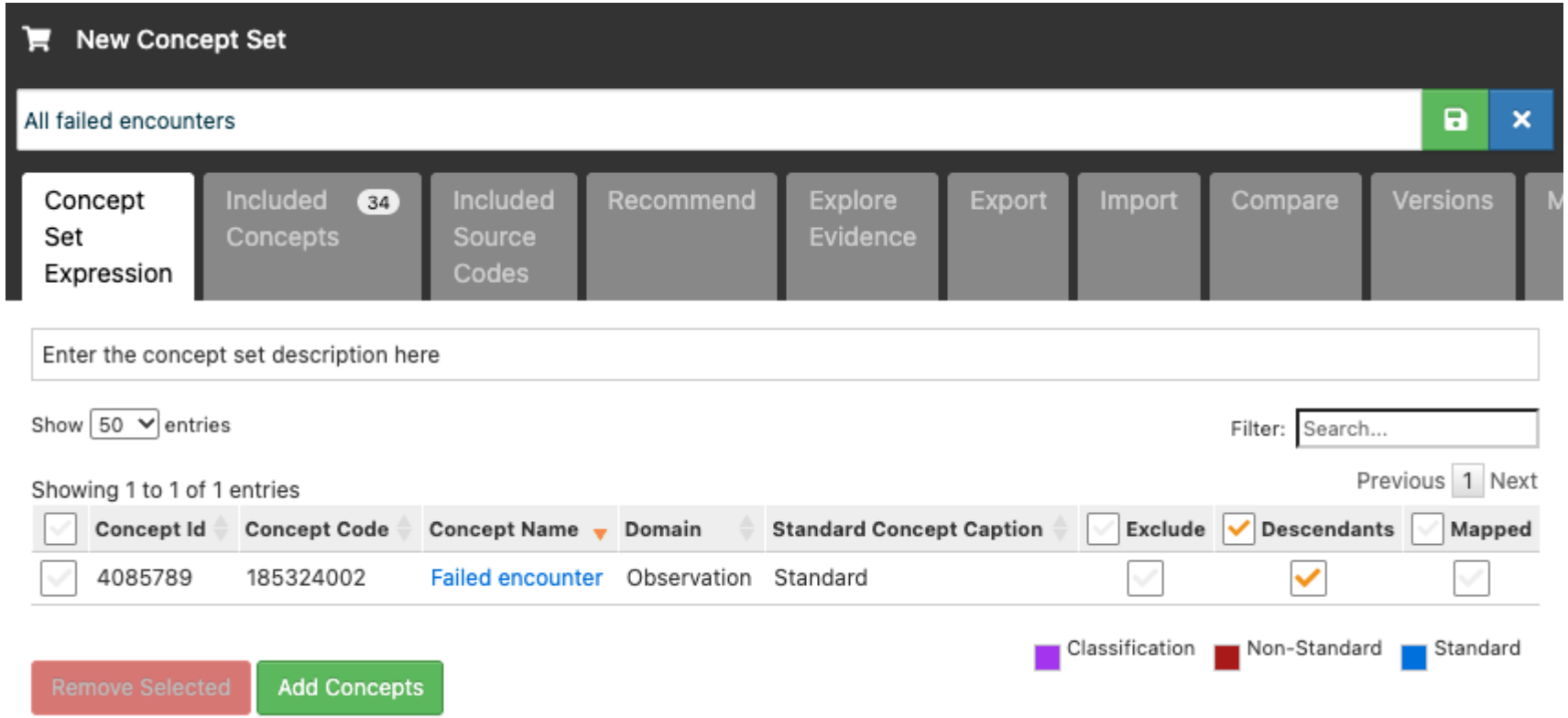
When valid concepts become disconnected



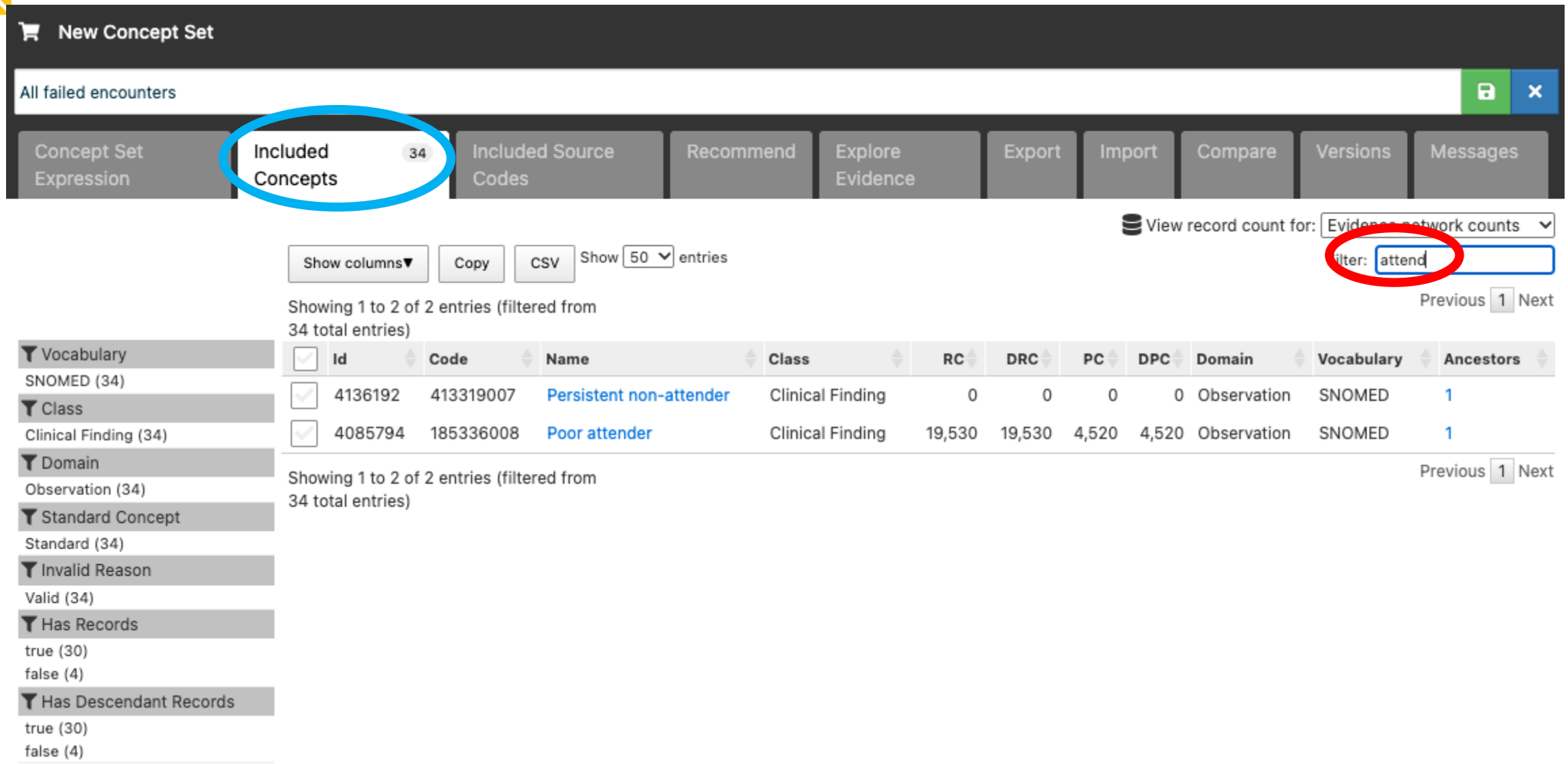
When valid concepts become disconnected



When valid concepts become disconnected



When valid concepts become disconnected





Can we systematically detect vocabulary tensions?

One option:

- Find a tension
- Write SQL
- Review results
- Repeat

Problems:

- Requires knowing what to look for
- One query per pattern
- Difficult to generalize
- New tension types require new code

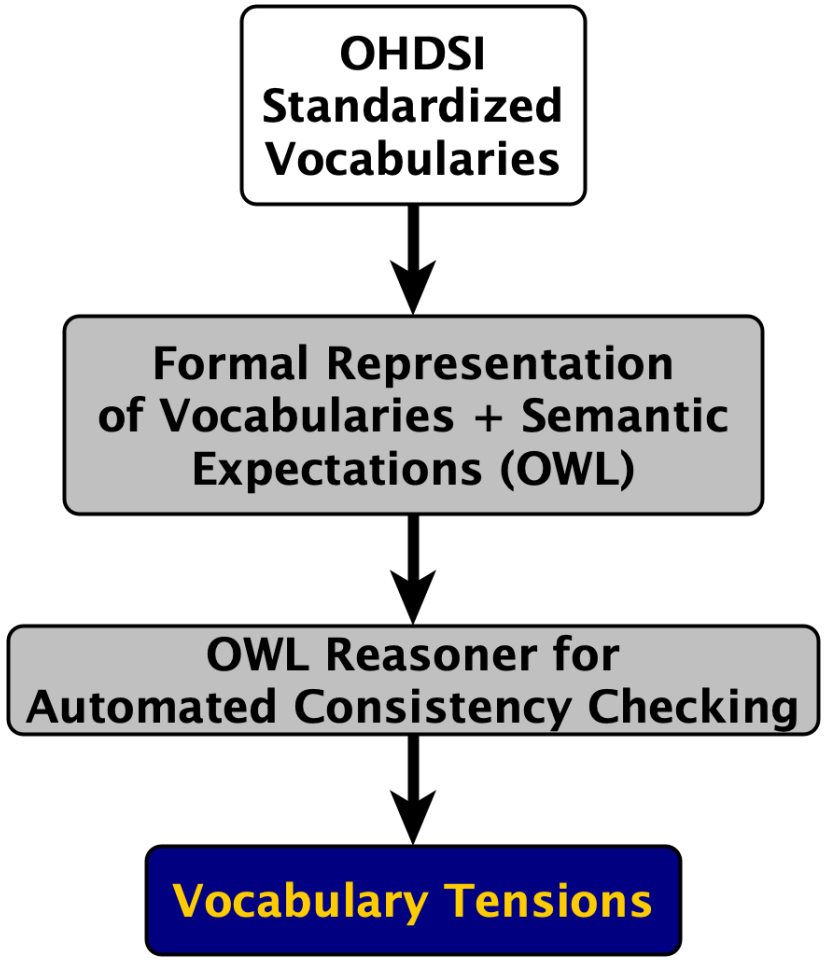
Alternative:

- Define semantic expectations once and evaluate automatically

Example expectations:

- Concepts should not belong to incompatible domains
- Valid concepts should not inherit through deprecated concepts
- Relationships should agree with concept classification

Testing semantic expectations at scale



Vocabulary Structure

- Concepts
- Domains
- Validity
- Relationships
- Hierarchies

Semantic Expectations

- Incompatible domains
  - (e.g., Condition ≠ Measurement)
- Valid should not inherit from invalid
- Relationships should respect domain/range constraints

Why this matters to OHDSI

- Makes hidden semantic assumptions more visible
- Identifies concepts that may warrant expert review
- Enables monitoring across vocabulary releases
- Provides a deterministic, scalable complement to existing QA

A vocabulary tension is not necessarily an error. It is a place where vocabulary structure and semantic expectations deserve closer examination.



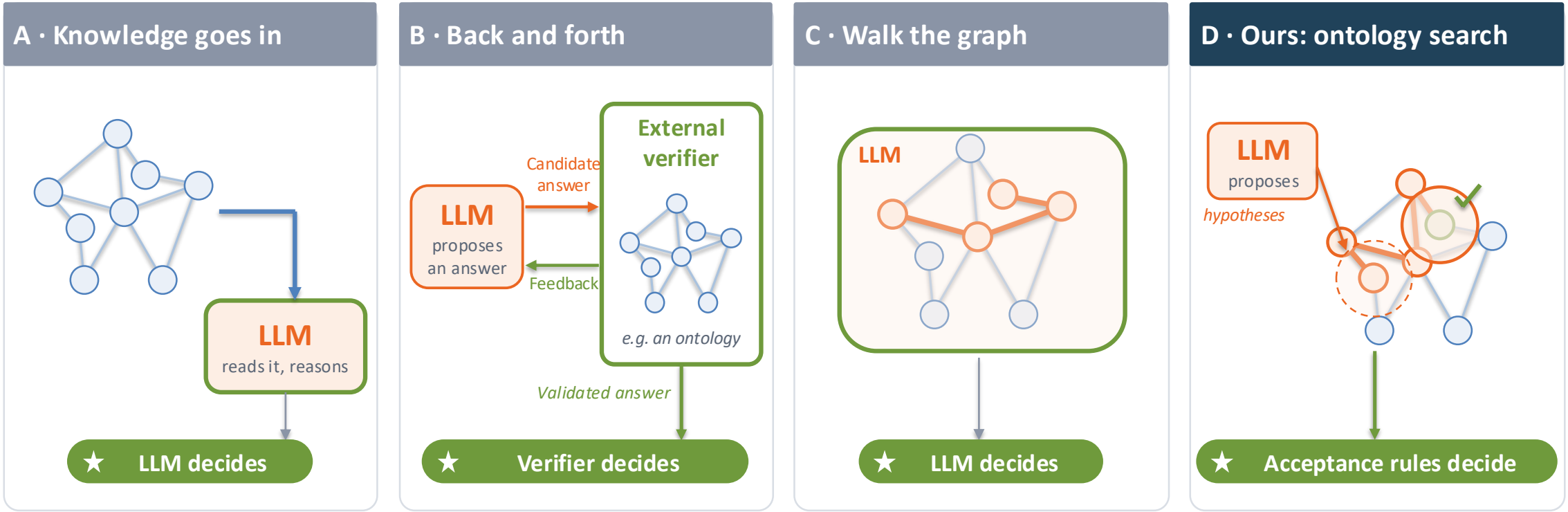
Ontology-Governed LLM Search for Safe Drug Concept Mapping in Air-Gapped Healthcare Environments

Seung Ju Yun, Eun Kyung Ko, Myung ji Yoo

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2025-02373039).

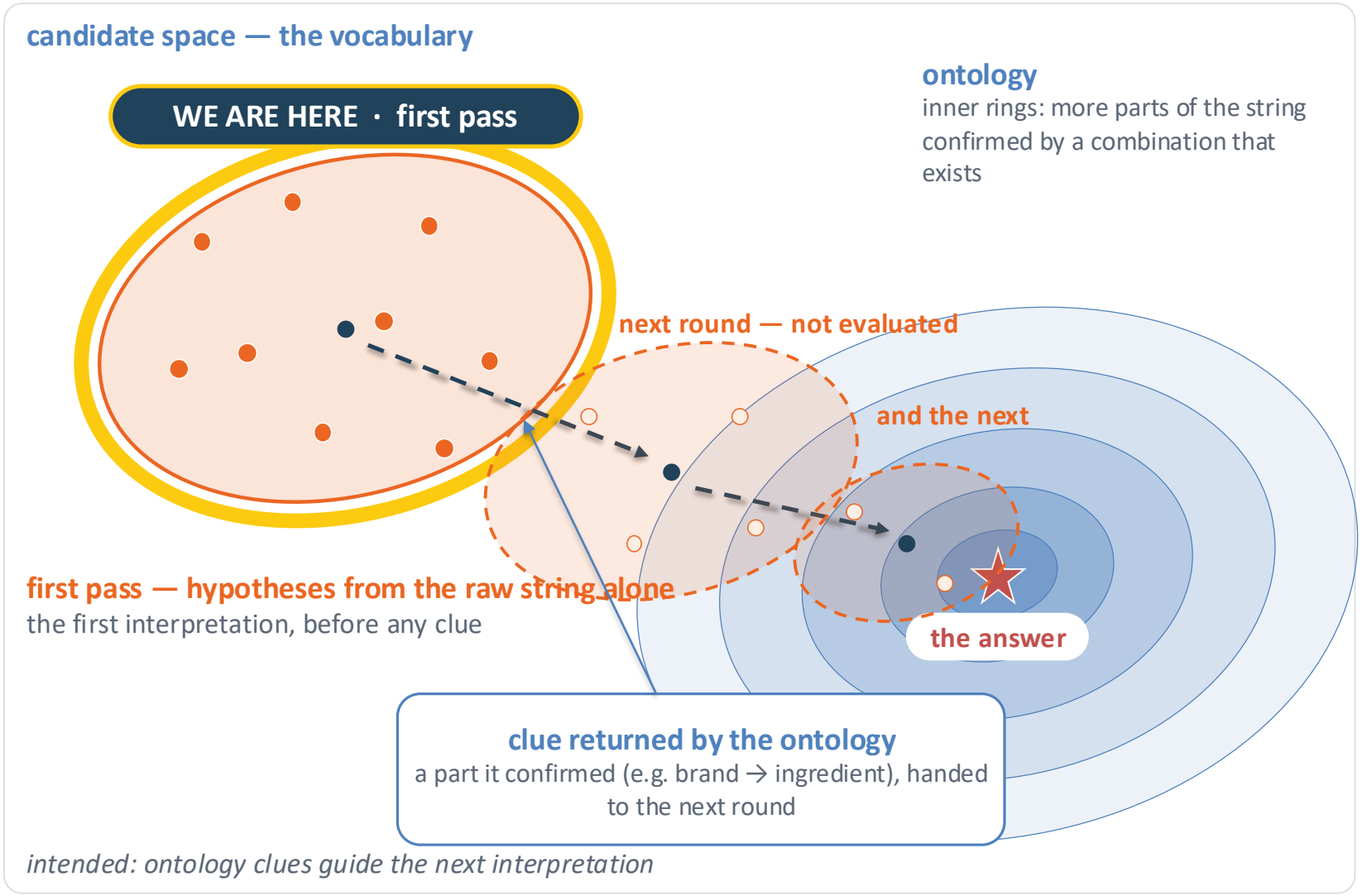
How an LLM can work with an ontology

Same two ingredients — a language model and a knowledge graph. In each picture: where is the answer decided?



Ours: part hypotheses · ontology search · rule-based acceptance — the answer is an existing concept, or an abstention

From design to first-pass evaluation



1 LLM proposes
reads each part of the string, offers several hypotheses

2 Ontology checks
grounds each hypothesis; records what is confirmed, in conflict, or unconfirmed

3 Clues guide next round
not evaluated in this first pass
what it confirmed returns as a clue

- LLM hypothesis space (per round)
- ontology: more parts confirmed
- candidate the LLM measured
- not evaluated in this first pass

From hypotheses to acceptance

attribute evidence is checked against the vocabulary · the whole is measured by meaning · explicit rules decide

① propose

raw local string
"A-gel 2% (oph) 10g"

LLM proposes pieces

ingredient	brand	dose form	strength	route
A	G	gel	2%	oph
A'		oint	20mg/g	
A''				

② search

the whole string
"A-gel 2% (oph) 10g"

LLM as a ruler
measures ***d*** per candidate

- ***d* = 0.31** — A · G · gel · 2% · oph (full product)
- ***d* = 0.36** — A · gel (clinical drug form)
- ***d* = 0.42** — A (ingredient)
- A' · A'' · oint · 20mg/g × — **No existing combination** → not a candidate

***d* = semantic distance**

Liu et al., ICLR 2024

raw string R ↔ concept name C

$d(R, C) = E_t \mid \log M_R(t) - \log M_C(t) \mid$

smaller = read alike → visited first

③ decide

- A · G · gel · 2% · oph ***d* = 0.31**
- A · gel ***d* = 0.36**
- A ***d* = 0.42**

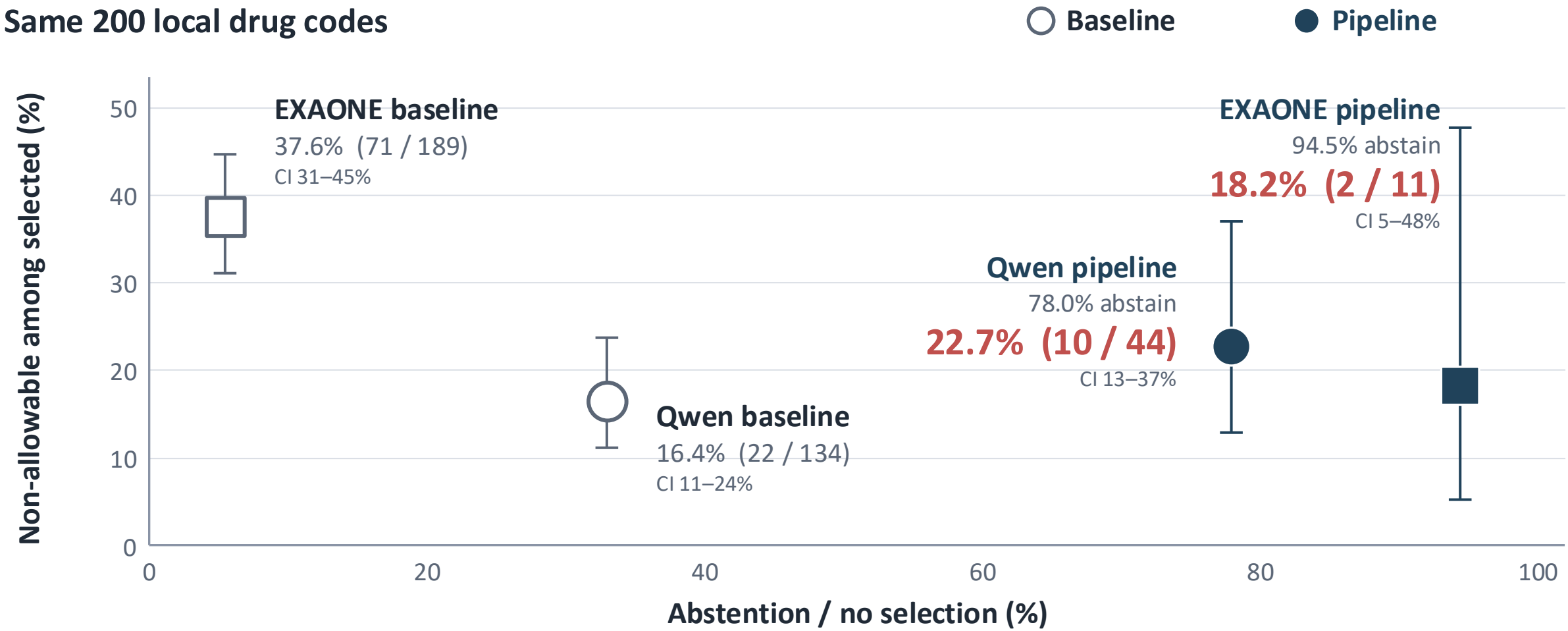
in visiting order

GATE
Acceptance rules
Detailed levels first

- **ACCEPT** A · G · gel · 2% · oph
- **ABSTAIN** reason on record
- **NEXT ROUND** not evaluated in this first pass

Illustrative example

High abstention did not eliminate accepted errors



200 codes drawn at random from 4,747 single-ingredient local codes with an EDI→OMOP mapping; allowability of each output judged by the CRC team against that reference. Error bars: 95% Wilson intervals. Qwen baseline with thinking (not shown): 28 / 146 = 19.2% (CI 14–26).

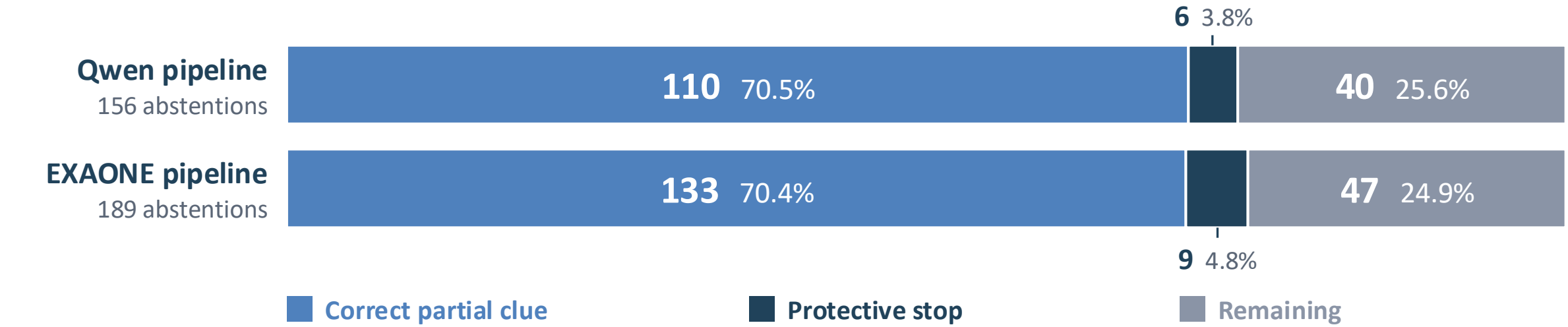
What useful abstention can look like

Clues for the next interpretation

Search candidates → Compare interpretations → Record conflicts / unconfirmed parts

Reference-based audit

unit: first-pass abstentions — Qwen 156 · EXAONE 189 of the same 200 codes



About 70% contained a reference-matched *partial clue*

Recovery not yet evaluated

e.g. Meqsel: ingredient conflict → abstained
kept: **trametinib** ✓ matches reference

Why the system abstained

LLM · interpretation | Pipeline · check & record | LLM · next round

No support → no acceptance

Amphogel

Read **'gel' in the name** as the dose form

→ **Gel candidates** found — support **W = 0** for every combination
No supported combination → **ABSTAIN**

Abstain — clue left for the next interpretation

Feedback not yet evaluated

Pakinol PD

Brand: Requip XL
Form: **plain Tablet**

→ Plain tablet ↔ **Extended-Release Tablet** candidate
Dose form mismatch → **ABSTAIN**

Ultravist

Proposed **iohexol**
Unit of '300' unresolved

→ Brand → **iopromide** · '300' ↔ **623 mg/mL** candidate
Strength unconfirmed → **ABSTAIN**

Meqsel

Initial ingredient: **mequinol**

→ **mequinol** ↔ **trametinib 2 mg [MEQSEL]**
Ingredient mismatch → **ABSTAIN**

→ **Next interpretation**

input:
Candidate + decision record

What false acceptances reveal about the architecture

Selected examples

Release cues	Invega **ER** → Invega Sandostatin **LAR** → Sandostatin
Product variants	Ketotop-**L** → Ketotop Isentress **HD** → Isentress
Device detail	Skyrizi **Pen → Injectable Solution**
Device mismatch	Humalog **Cartridge → Pen Injector**

① Source cues not preserved

LAR proposed → accepted without LAR

② Form / device distinctions missed

Raw string + semantic distance: insufficient here

Next: add product context for finer distinctions

Future work

Lessons from the first pass

LLM proposes interpretations; ontology constrains candidates.

1 Preserve useful interpretations

LAR proposed → accepted without LAR.

2 Keep interpretations revisable

Retained conflicts provide material for re-examination.

3 Match evidence to the distinction

Form/device mismatches still passed acceptance checks.

Our first-pass evaluation motivates three design requirements for ontology-governed LLM mapping: preserve useful interpretations, enable revision, and match evidence to the distinction.

Future Work

1 Refine policies and strengthen evidence

Preserve cues; test product-label context for fine distinctions.

2 Test recovery from feedback

Measure correct recovery and new false acceptances.

3 Track multi-round behavior

Observe correct resolution, repetition, and persistent errors.

EPISODE/EPISODE_EVENT-based Representation of Microbiology Results & Antimicrobial Susceptibility Testing in OMOP CDM v5.4

Don’t flatten microbiology

Preserve the chain of evidence:

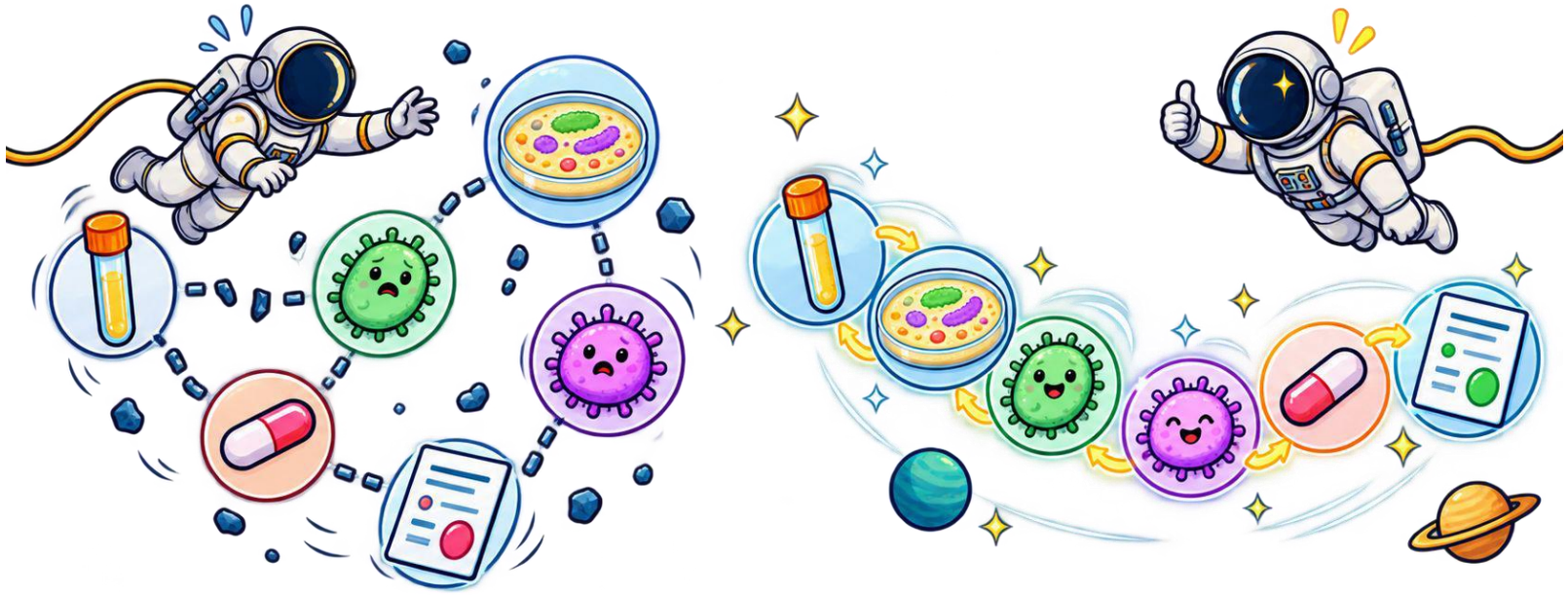
Specimen → Analysis → Organism → Antimicrobial → Result

Presenter: **Polina Talapova, MD, PhD**
on behalf of **Jared Houghtaling**, **Gilles Clermont**,
and **Andrew E. Williams†**

Every value can survive, and the meaning can still be lost

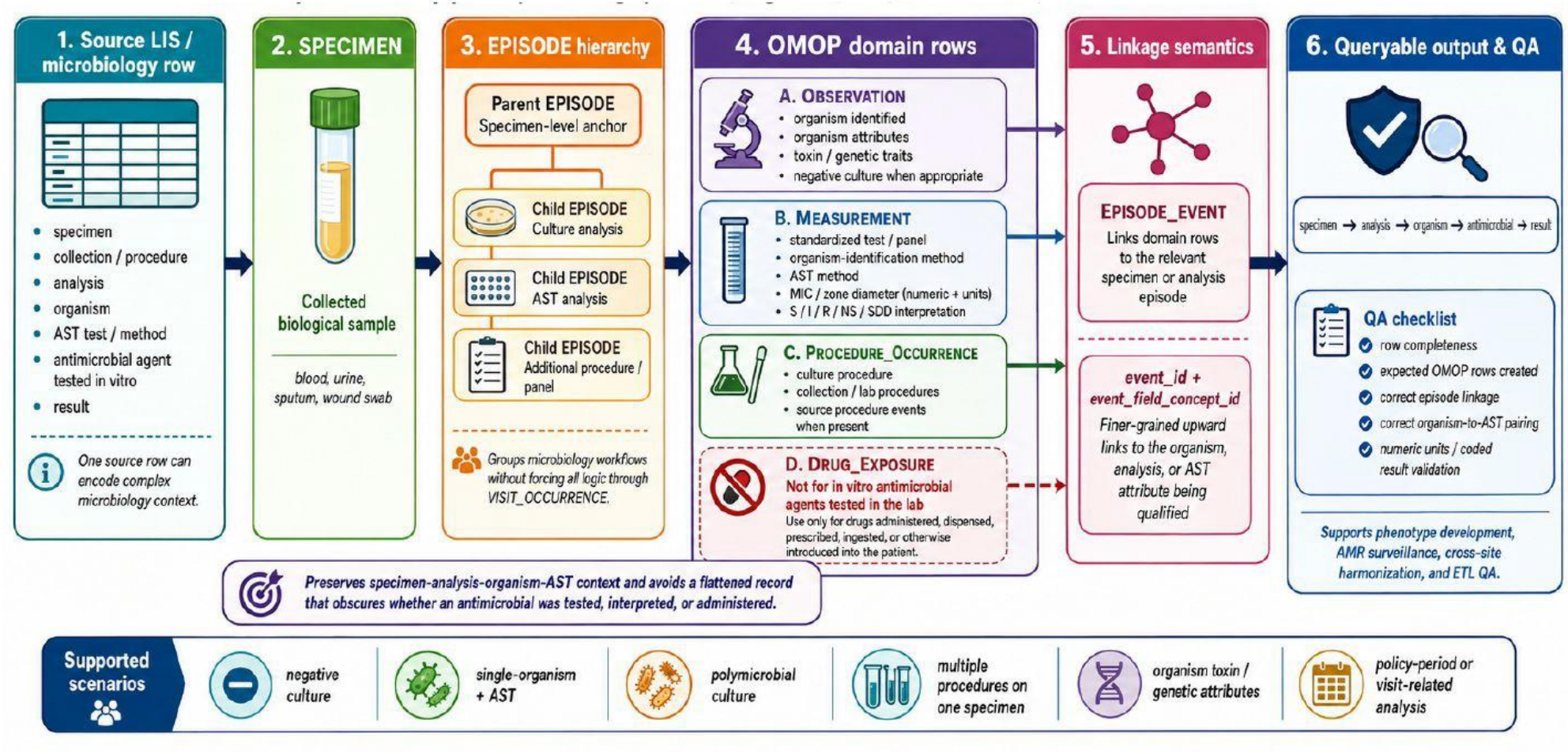
Disconnected rows

Recoverable evidence



Syntactically valid. Semantically unrecoverable.

From LIS data to reconstructable evidence



A specimen-centered OMOP CDM v5.4 approach that preserves the full microbiology chain

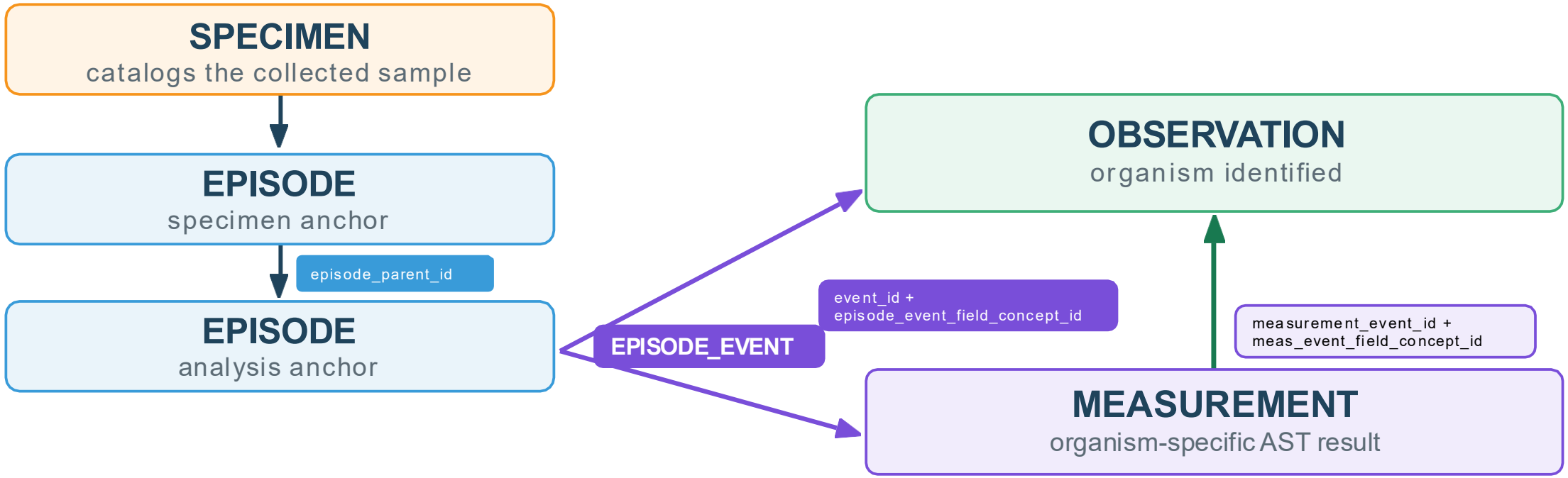
Two linkage layers. Two different questions.

Workflow membership

What belongs to this microbiology process?

Semantic qualification

What exactly does this result qualify?



One culture. Two organisms. Four organism-specific AST results.

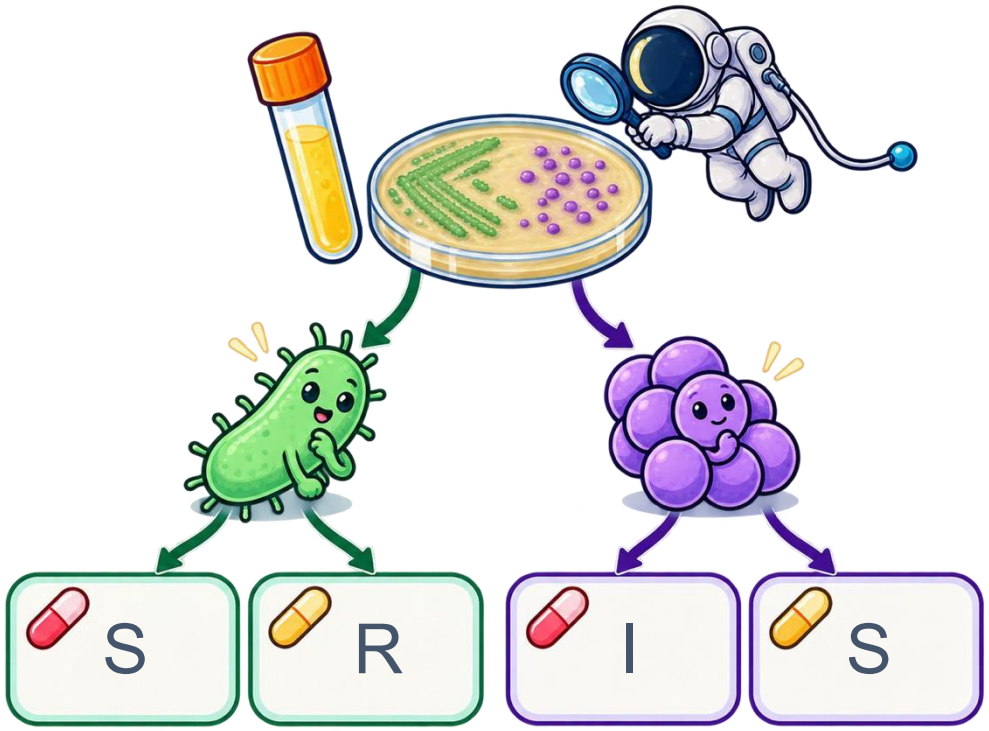
Urine culture

***E. coli* isolated**

Nitrofurantoin MIC 32 µg/mL → S
Ampicillin MIC 64 µg/mL → R

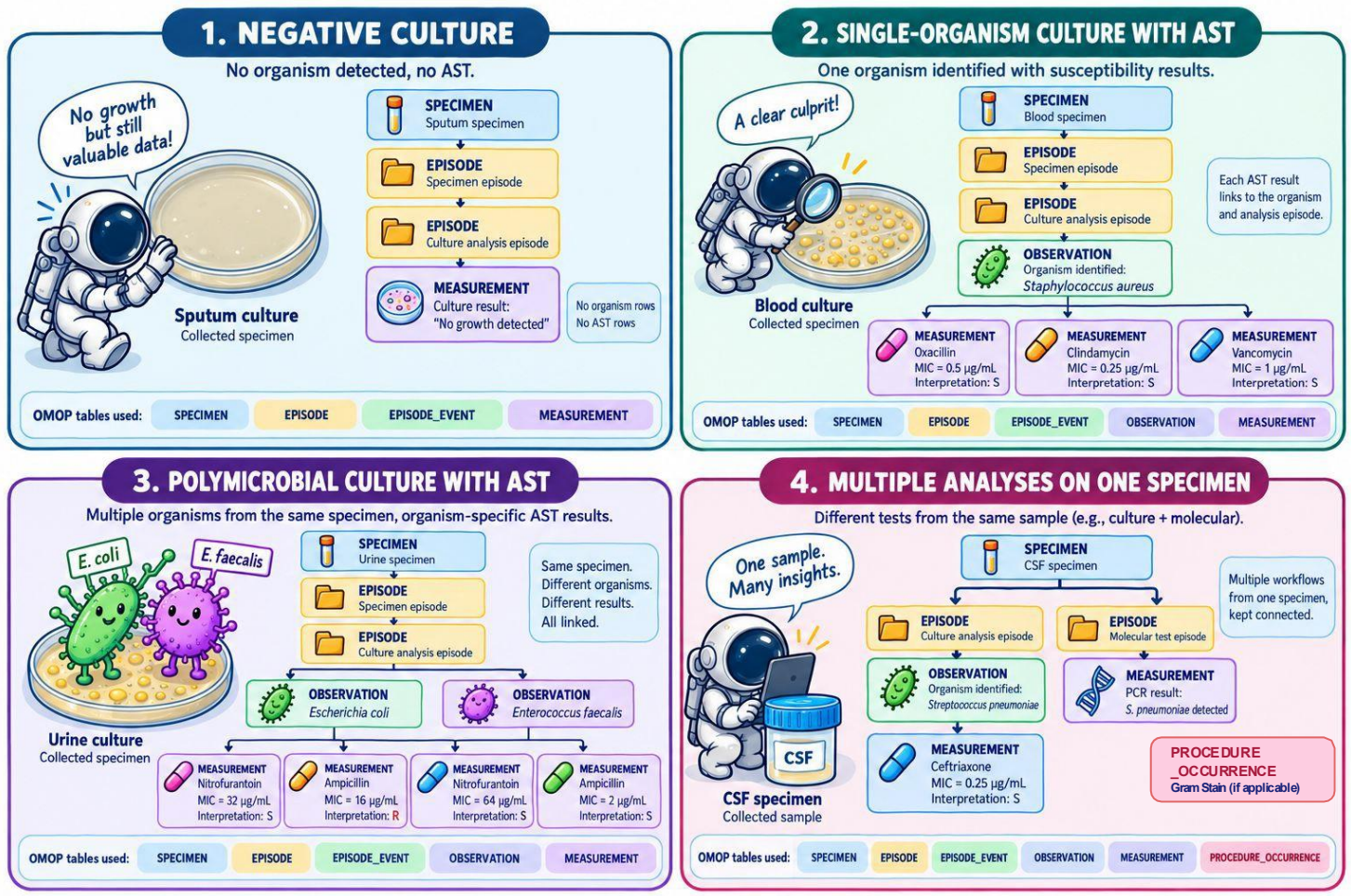
***E. faecalis* isolated**

Nitrofurantoin MIC 64 µg/mL → I
Ampicillin MIC 2 µg/mL → S



Each AST result remains linked to the organism it qualifies, while all rows remain in the same culture workflow.

The approach survives the hard cases

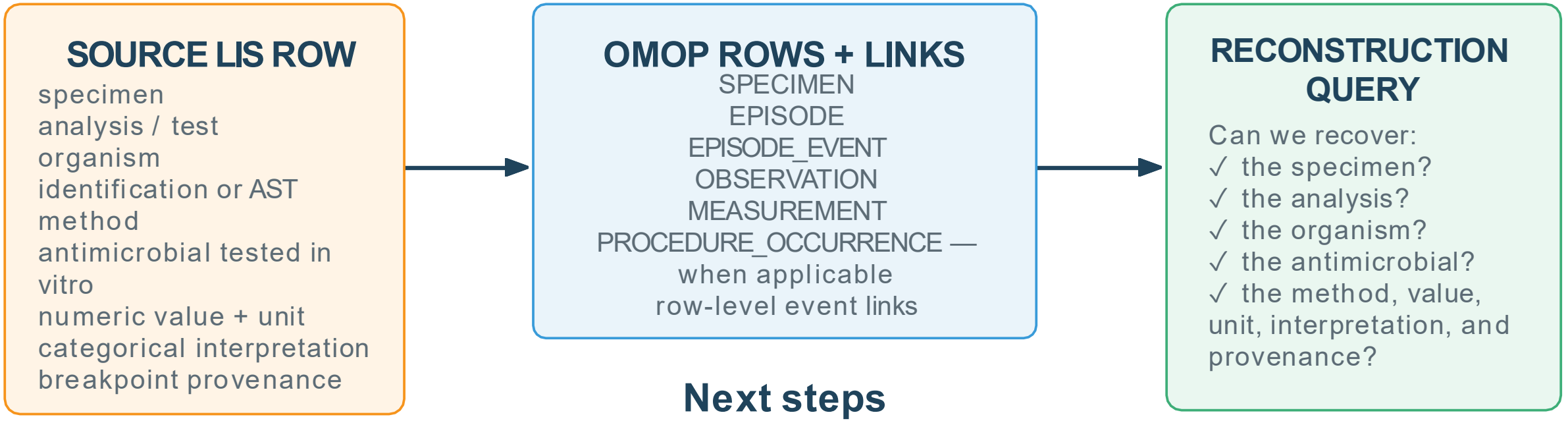


Also supports organism attributes, policy-period grouping, and visit-level organism profiles.

Proposed EPISODE concepts

Concept Code	Concept Name	Destination	Comment
MICRO_EPISODE	Microbiology episode	episode_concept_id	Hierarchical umbrella concept
MICRO_SPECIMEN_EPISODE	Microbiology specimen episode		Groups all microbiology activity attributable to one physical specimen
MICRO_ANALYSIS_EPISODE	Microbiology analysis episode		Represents one distinct culture, assay, panel, identification, stain, toxin test, or AST analysis
MICRO_ACCESSION_EPISODE	Microbiology accession episode		Parent for a laboratory order/accession spanning multiple specimens or analyses

Recoverability is the validation target



Next steps

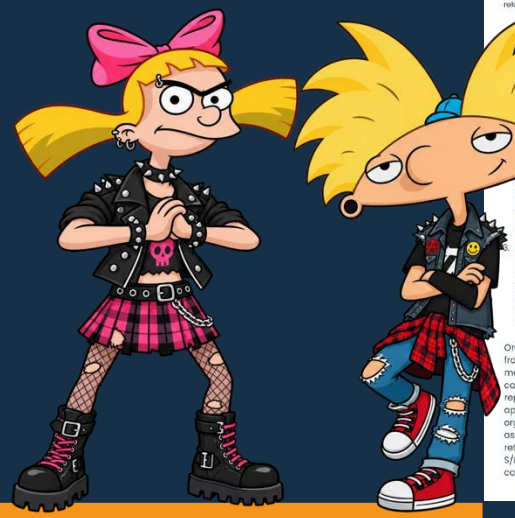
Submit microbiology EPISODE concepts as an OHDSI community contribution

Preserve how S / I / R was derived: CLSI / EUCAST / local rule set + version

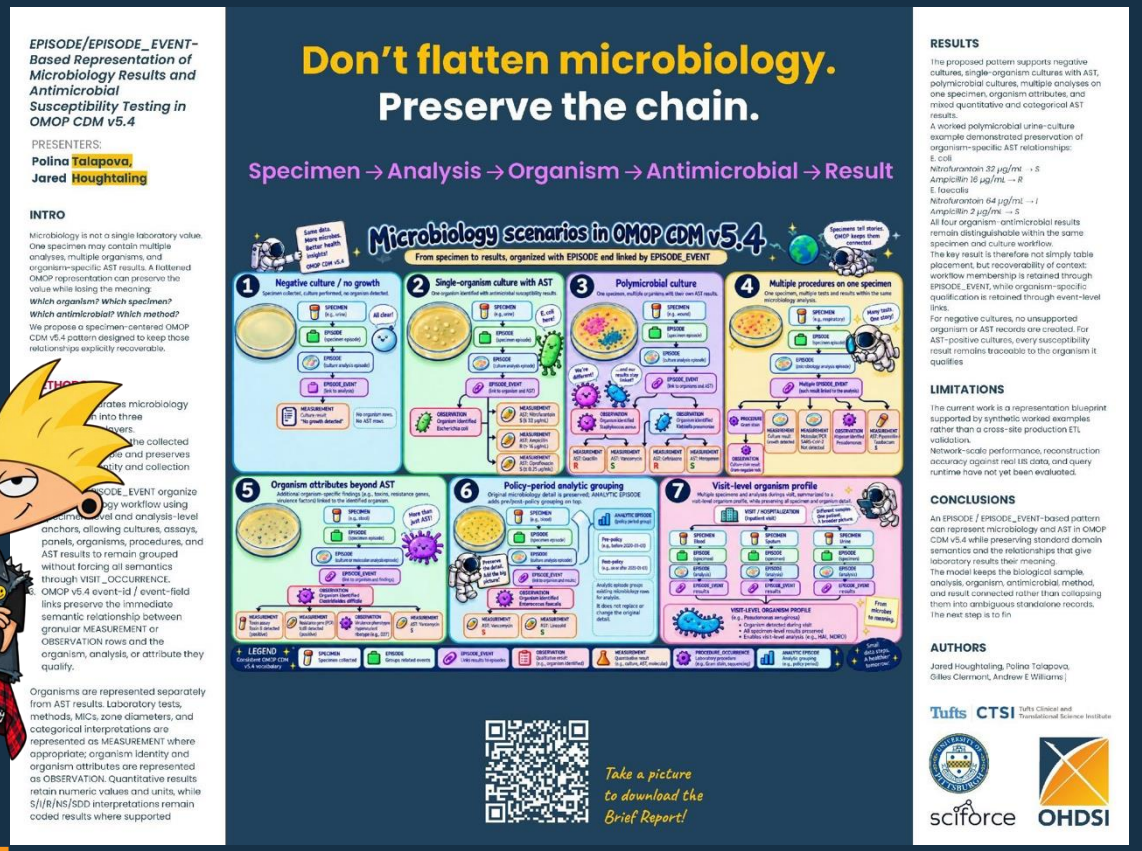
Validate reconstruction across real LIS implementations

Bacteria can be resistant.

The data model doesn’t have to be.



Curious about culture, resistance, and a little scientific rebellion? Come find us!



EPISODE/EPISODE_EVENT-Based Representation of Microbiology Results and Antimicrobial Susceptibility Testing in OMOP CDM v5.4

PRESENTERS:
Polina Talapova,
Jared Houghtaling

INTRO

Microbiology is not a single laboratory value. One specimen may contain multiple analyses, multiple organisms, and organism–specific AST results. A flattened OMOP representation can preserve the value while losing the meaning:
Which organism? Which specimen? Which antimicrobial? Which method?
We propose a specimen-centered OMOP CDM v5.4 pattern designed to keep those relationships explicitly recoverable.

...rates microbiology ...in into three ...layers.
...the collected ...and preserves ...ility and collection ...

...SODE_EVENT organize ...gy workflow using ...level and analysis-level anchors, allowing cultures, assays, panels, organisms, procedures, and AST results to remain grouped without forcing all semantics through VISIT_OCCURRENCE. OMOP v5.4 event-id / event-field links preserve the immediate semantic relationship between granular MEASUREMENT or OBSERVATION rows and the organism, analysis, or attribute they qualify.

Organisms are represented separately from AST results. Laboratory tests, methods, MICs, zone diameters, and categorical interpretations are represented as MEASUREMENT where appropriate; organism identity and organism attributes are represented as OBSERVATION. Quantitative results retain numeric values and units, while S/I/R/NS/SDD interpretations remain coded results where supported

Don’t flatten microbiology.
Preserve the chain.

Specimen → Analysis → Organism → Antimicrobial → Result

Microbiology scenarios in OMOP CDM v5.4

From specimen to results, organized with EPISODE and linked by EPISODE_EVENT

1 Negative culture / no growth
2 Single-organism culture with AST
3 Polymicrobial culture
4 Multiple procedures on one specimen
5 Organism attributes beyond AST
6 Policy-period analytic grouping
7 Visit-level organism profile

Take a picture to download the Brief Report!

RESULTS

The proposed pattern supports negative cultures, single-organism cultures with AST, polymicrobial cultures, multiple analyses on one specimen, organism attributes, and mixed quantitative and categorical AST results.
A worked polymicrobial urine-culture example demonstrated preservation of organism–specific AST relationships:
E. coli
Nitrofurantoin 32 µg/mL → S
Ampicillin 16 µg/mL → R
E. faecalis
Nitrofurantoin 64 µg/mL → I
Ampicillin 2 µg/mL → S
All four organism–antimicrobial results remain distinguishable within the same specimen and culture workflow.
The key result is therefore not simply table placement, but recoverability of context: workflow membership is retained through EPISODE_EVENT, while organism–specific qualification is retained through event-level links.
For negative cultures, no unsupported organism or AST records are created. For AST–positive cultures, every susceptibility result remains traceable to the organism it qualifies

LIMITATIONS

The current work is a representation blueprint supported by synthetic worked examples rather than a cross-site production ETL validation.
Network-scale performance, reconstruction accuracy against real EIS data, and query runtime have not yet been evaluated.

CONCLUSIONS

An EPISODE / EPISODE_EVENT–based pattern can represent microbiology and AST in OMOP CDM v5.4 while preserving standard domain semantics and the relationships that give laboratory results their meaning.
The model keeps the biological sample, analysis, organism, antimicrobial, method, and result connected rather than collapsing them into ambiguous standalone records.
The next step is to fin

AUTHORS

Jared Houghtaling, Polina Talapova, Gilles Clermont, Andrew E Williams

Tufts CTSI
sciforce OHDSI

Poster 157

Making the Most of What We Have

Why do OHDSI databases fail network-study diagnostics?

Evelyn Goh | National University of Singapore , Saw Swee Hock School of Public Health

21 October 2026 | OHDSI Global Symposium 2026

In 2024 we ran a global OHDSI study

9 of 15

cohorts passed the diagnostics

GLP-1 RA and acute liver injury

Almost all the passes were North American

8/9

N. America

●●●●●●●●○

cohorts passed

1/4

APAC-ANZ

●○○○

cohorts passed

0/2

Europe

○○

cohorts passed



Are diagnostics unable to handle APAC/EU/regional data?

Why are some databases failing? ☹



Objectives

Understand

Why are our databases failing?



Objectives

Understand

Why are our databases failing?

Characterise

Are there specific patterns resulting in these failures?



Objectives

Understand

Why are our databases failing?

Characterise

Are there specific patterns resulting in these failures?

Resolve

Can we do something about it?

Why did some GLP1 databases fail diagnostics?

AKA: more evidence, please

Three diagnostics decide who counts

ASDM

covariate balance after matching

Empirical equipoise

preference-score overlap

EASE

systematic error

In most non-U.S. databases

The GLP-1 study DB diagnostic results:

ASDM

covariate balance after matching

Often failed

Empirical equipoise

preference-score overlap

Often failed

EASE

systematic error

Mostly passed

In most non-U.S. databases

Pilot

Evidence (?) Pt. 1

Pilot: two completed OHDSI studies

PILOT

15

databases

GLP-1 RA and acute liver injury

9.1M people

N. Am 9 · ANZ 1 · APAC 3 · EU 2

9

databases

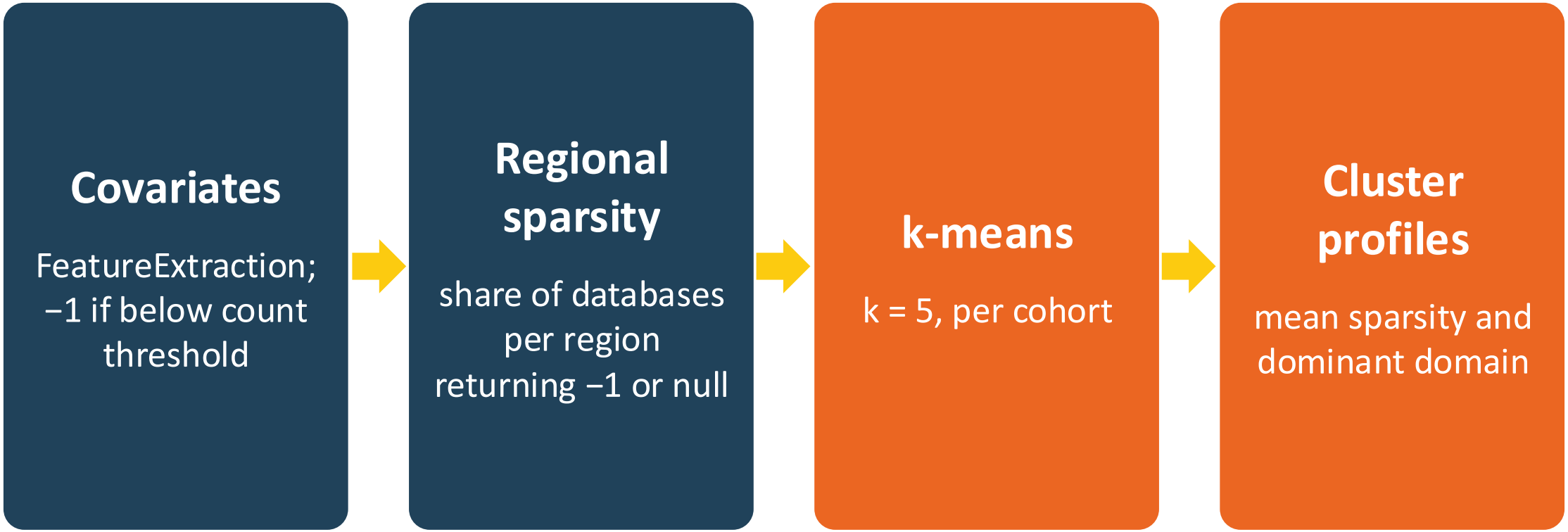
Fluoroquinolones and aortic aneurysm

1.7M people

N. Am 8 · APAC 1

Method: where is each covariate missing?

PILOT



Three patterns.
Every cohort. Both studies.

~20,000 covariates per cohort, five clusters each

Pattern 1: universal sparsity

PILOT

30–40%

of covariates are sparse everywhere (largest cluster)

N. Am	ANZ	APAC	EU

Dominant domains: observation period · rare conditions

■ available ▢ sparse (−1 / missing)

Pattern 2: North America preferential

PILOT

5–20%

of covariates available in N. Am, sparse elsewhere

N. Am	ANZ	APAC	EU

Dominant domains: drug era · condition era

■ available ☐ sparse (−1 / missing)

Pattern 3: APAC enriched

PILOT

10–15%

of covariates enriched in APAC (JMDC)

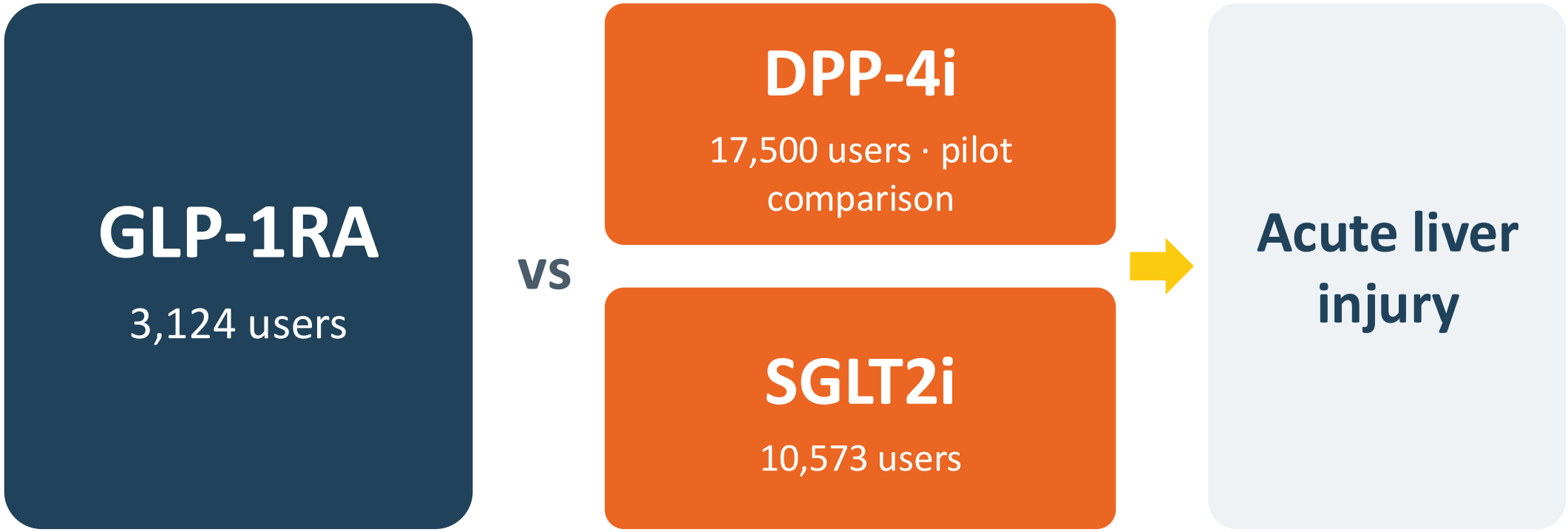
N. Am	ANZ	**APAC**	EU

Dominant domains: measurements · procedures

■ available ☐ sparse (−1 / missing)

Deep diving into one database (GLP-1-RA)

PILOT



On-treatment · 1:1 PS matching (caliper 0.2) · 130 negative controls

Why the DB failed

PILOT

Data capture

46% of users lost at "≥ 1 day at risk"

Small sample

~550 per arm
MDRR fails · EASE not computable

Chance failures

~27,000 covariates tested
42-45% of failures are rare

Plus one real failure: sulfonylureas flipped after matching

Does low density explain it? Could be – need to run a proper study!



Phase 1

OK, so let’s do it properly now

HYPOTHESIS (H1)

Lower structural density, more balance failure

H2: imbalance tracks systematic error less in low-SDI databases

The Structural Density Index (SDI)

PHASE 1

SDI = ½ Volume + ½ Breadth

0 = sparse → 1 = dense

Volume

how many people have each covariate

mean subject-count tier, scaled 0-1

Breadth

how many OMOP domains are populated

≥1% of covariates non-null, of 6 domains

Primary SDI: whole-database person counts per concept

Tiers (dense · moderate · sparse) are descriptive only

Phase 1 outcomes

PHASE 1

PRIMARY (H1)

Balance failure

max ASMD > 0.1 after adjustment

Imbalance burden

share of covariates with ASMD > 0.1

SECONDARY

Equipoise

fails below 0.2

EASE

fails above 0.25, with μ and σ

MDRR

fails above 10

Availability

by OMOP domain

Thresholds: CohortMethod defaults, recomputed from continuous values

Tested the LEGEND-T2DM study on Synthea

PHASE 1

100,000
synthetic persons

3 h 5 min
full run

23 min
LEGEND-T2DM study

Synthea · 6 runs · 7–8 Oct 2026

Next

Partner sites

TMU and LEGEND running now

More studies

extend the characterisation

EASE and negative controls

formal analysis of availability

Thank you very much!

Evelyn Goh | e0983111@u.nus.edu