OHDSI Global Symposium Collaborator Showcase Lightning Talk Session #1

Moderator:

Harry Reyes Nieva, PhD, MAS

Co-Lead, OHDSI Early-Stage Researchers WG
Division of Infectious Diseases, Department of Medicine
Columbia University Irving Medical Center

Bridging Standards: Creating OMOP data via FHIR and Health Information Networks

Stephanie S. Hong, MS, FAMIA
Research Associate | Department of Medicine
Johns Hopkins University School of Medicine

on behalf of the
AoU CLAD HIE-HIN Team

October 8, 2025 | OHDSI Global Symposium 2025

All of Us Research Program: Over 800K participants have consented to share EHR but only about 473K have linked EHR records

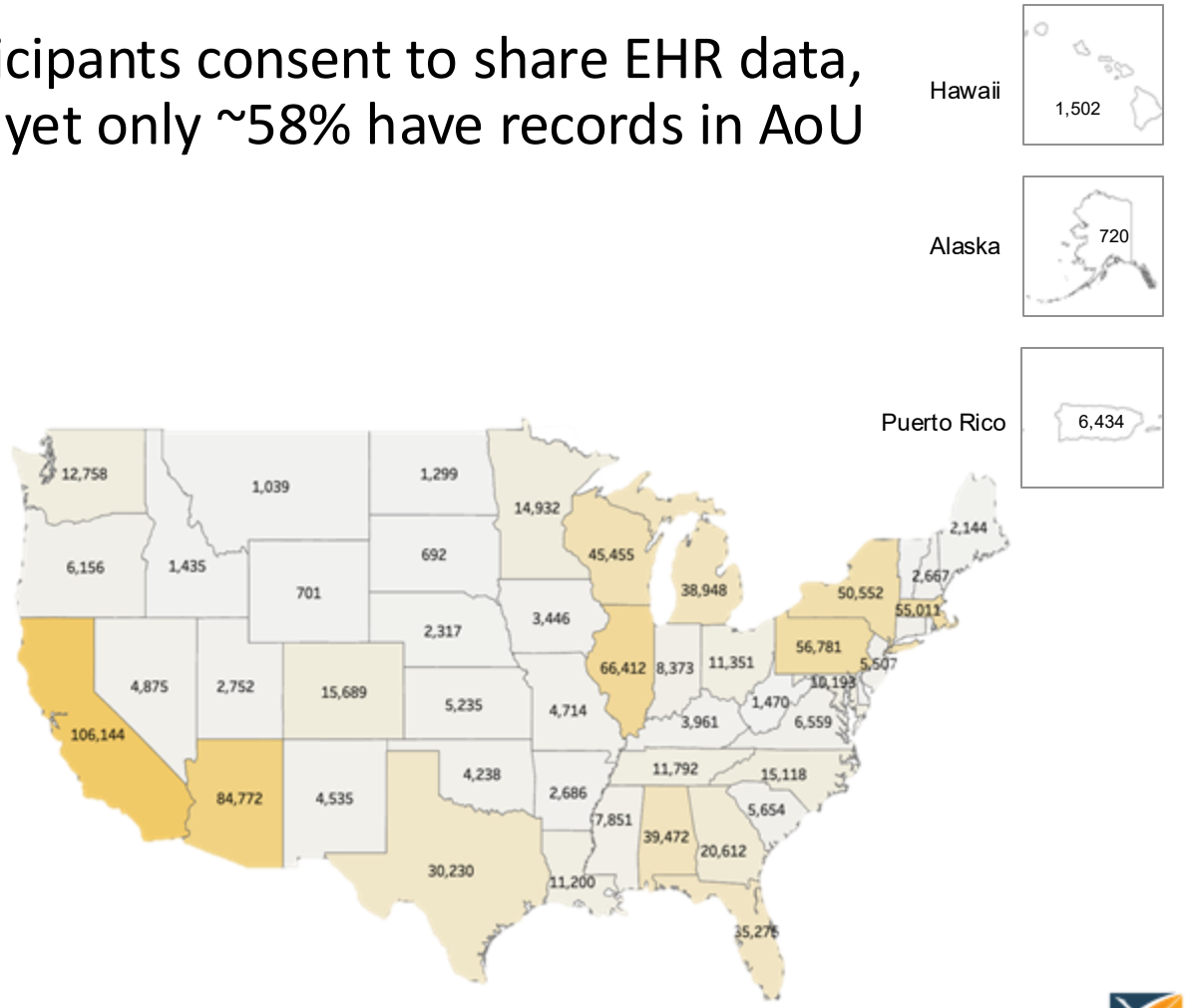
The challenge: ~94% of AoU participants consent to share EHR data, yet only ~58% have records in AoU

868,000+
Participants

473,000+
Electronic Health Records

596,000+
Participants who have completed the initial steps of the program

611,000+
Biosamples



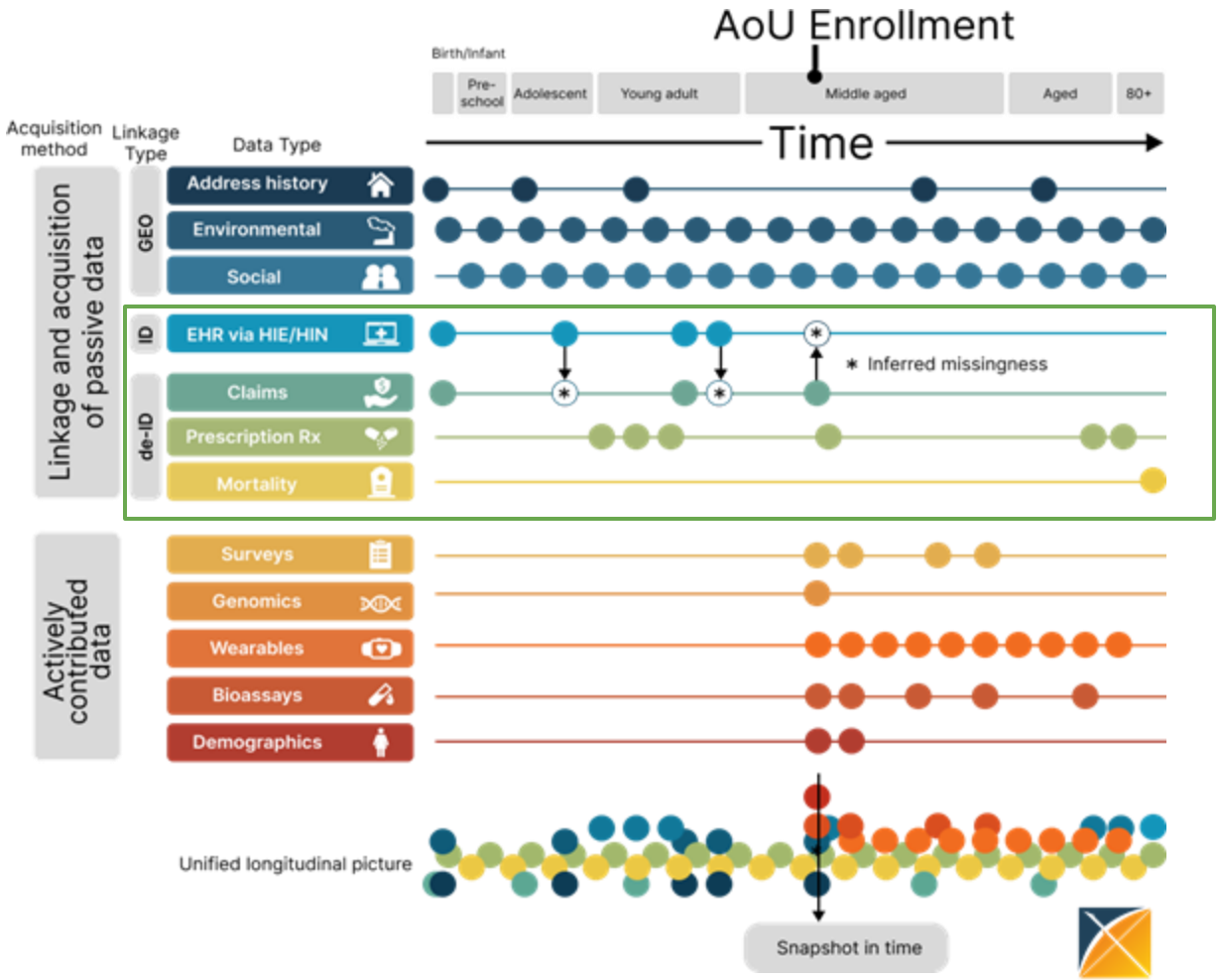
Numbers current as of October 2, 2025

OHDSI 2025

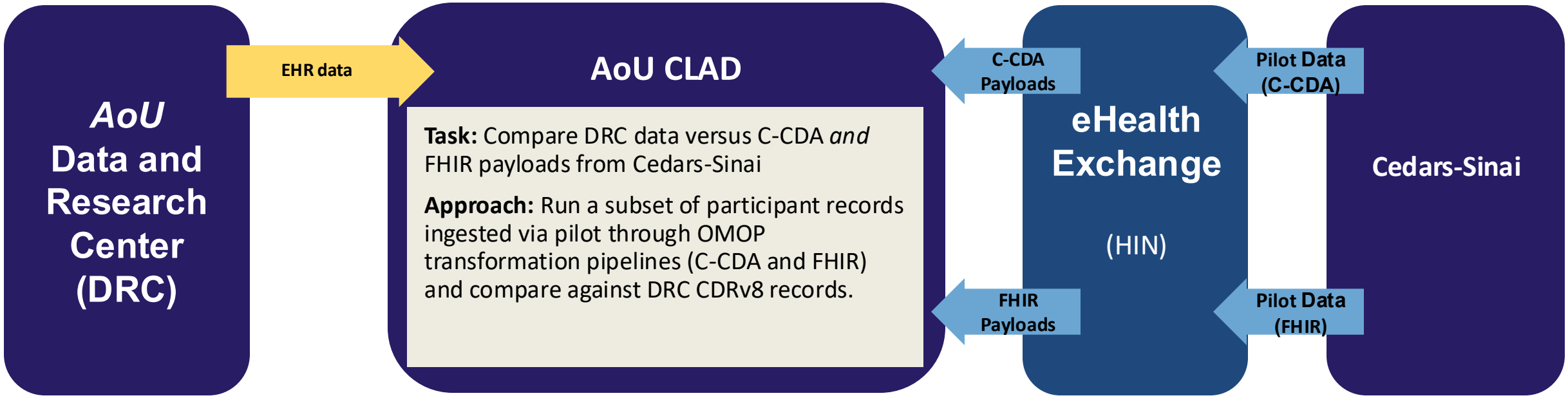
AoU Center for Linkage and Acquisition of Data (CLAD): Vision => Putting the Patient Back together Again

- New data acquisition methods are needed to capture and link each person's complete health journey

- Patient-Privacy Preserving Record Linkage (PPRL, de-identified token-based linkage) enabled the integration of patient data from disparate sources



HIN/HIE Pilot Comparison - Cedars-Sinai - DRC vs FHIR vs C-CDA



- Cedars-Sinai was chosen for high match rates, with a large population of AoU participants.
- Both FHIR **and** C-CDA payloads for Cedars-Sinai were available through **eHealth Exchange**, a nationwide nonprofit **Health Information Network (HIN)**, allowing retrieval of Cedars-Sinai data through the exchange.
- Cedars-Sinai enables data retrieval under Data Use and Reciprocal Support Agreement (DURSA), permitting **participant-authorized exchange** for Research use.

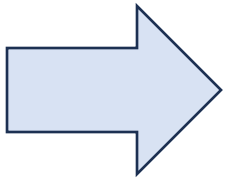
OHDSI 2025

Flattened FHIR JSON data to Prepare for OMOP Transformation

10,512,100
FHIR Resources collected:

Patient
Encounter
Observation
Condition
Procedure
Medication
MedicationRequest
Immunization
Practitioner
Device
Location
CareTeam

Preserved: CodeableConcept codings → *system, code, display*



i_patient	i_medication
i_encounter	i_medication_administration
i_condition	i_medication_request
i_observation	i_location
i_device	i_organization
i_immunization	i_practitioner

OHDSI
2025

Flattened and Collated Intermediate Datasets Enable Syntactic Translation

i_patient.race

source code	source code system
2106-3	urn:oid:2.16.840.1.113883.6.238

FHIR value-set mapping table

source concept code	source concept code system	target concept id	target concept name	target vocabulary id
2106-3	urn:oid:2.16.840.1.113883.6.238	8527	White	Race
M	http://terminology.hl7.org/CodeSystem/v3-AdministrativeGender	8507	Male	Gender

i_patient.sex

source code	source code system
M	http://terminology.hl7.org/CodeSystem/v3-AdministrativeGender

- Handles metadata field-to-concept ID mapping

OHDSI 2025

Fully Provenance-Enabled Pipeline Architecture: Processed over 10M FHIR resources through 142 transformation steps

```
Define and RegisterBaseStruct ( _struct_name, _struct_type )

Initialize intermediate processing data structures. ( i.e.
i_patient, i_encounter, i_medication, i_immunization, i_procedures,
i_device, i_observation, i_medication_request, i_organization, etc)
```

FHIR Resource
JSON Response of the AoU matching patients

fhir value-set mapping table:
FHIR codeable concepts to omop concept id lookup table

HL7 terminology id to omop **vocabulary id lookup table**

OMOP Transformer

code_map: Standard terminology code to concept id mapper dataset

Deploy()
OMOP CDM-compliant datasets

- ✓ Schema & Integrity Checks
- ✓ Data Validation Checks

Entire pipeline runs < 30 mins

Novel FHIR mapping pipeline

OHDSI 2025

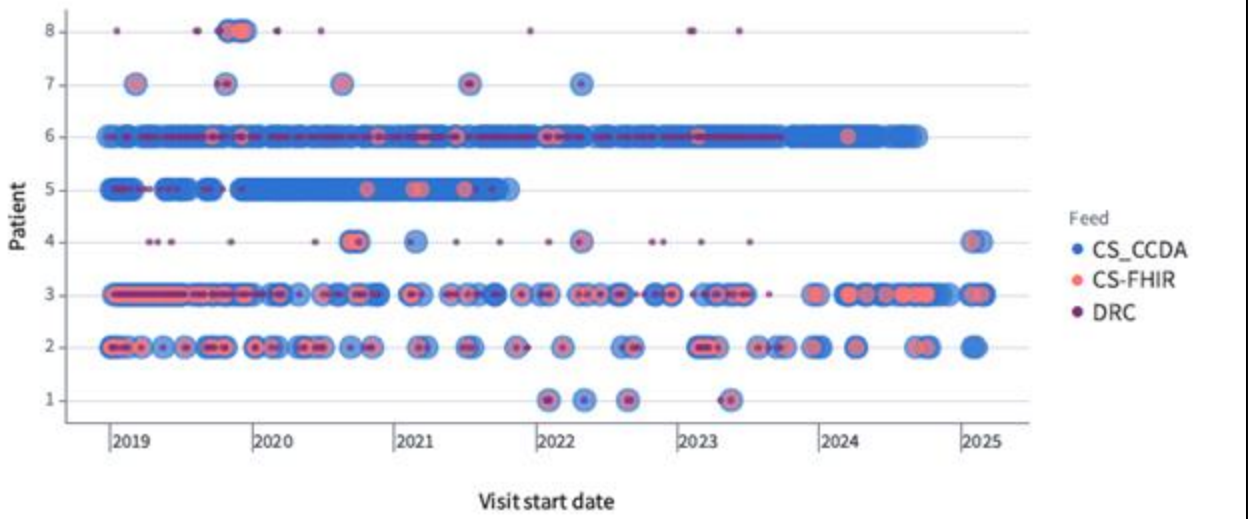
Non-USCDI v2 (US Core 3.1.1) ⇒ Presents OMOP Mapping Challenges

Domain	Mapped	Mapping gap	Mapping Gap Explained
Condition	2,194,240	0%	
Device	369	93%	missing code/code system and local code used
Drug	703,433	4%	missing code/code system
Measurement	4,057,184	39%	missing code/code system and missing unit of measure string
Observation	3,670,804	38%	missing code/code system and local codes used
Person	7,047	1%	Duplicate MSPI
Procedure	37,715	51%	use of local code/code system
Provider	53,186	51%	missing information
Visit	591,054	46%	local code used

OHDSI 2025

FHIR and C-CDA significantly improved Visit and Vaccination record coverage despite FHIR coding gaps

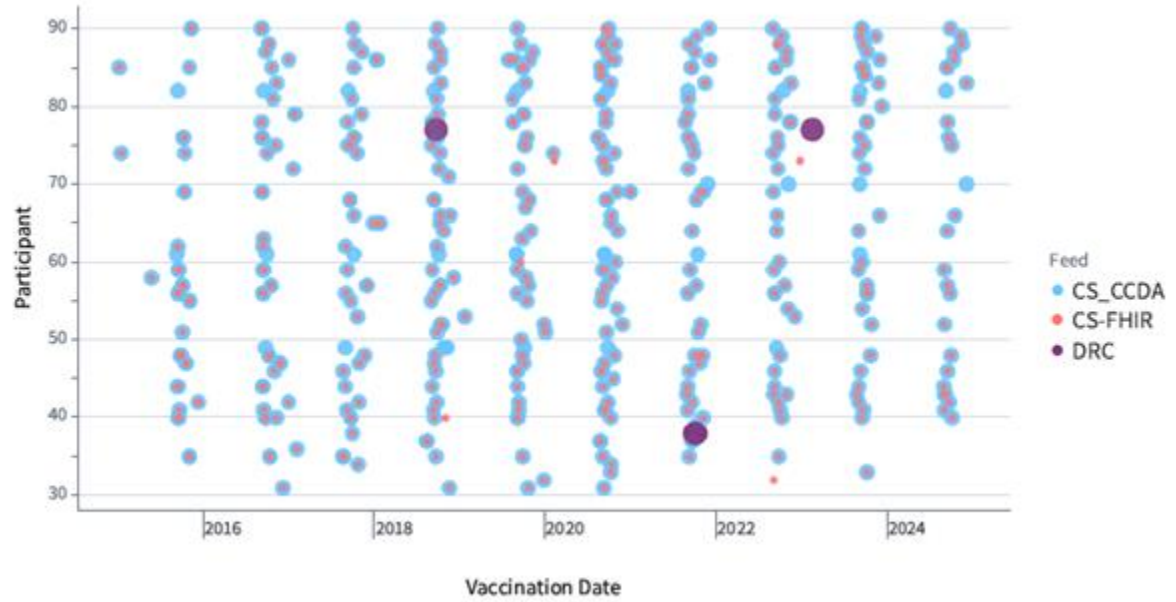
Cedars Sinai FHIR vs Cedars Sinai CCDA vs DRC: Visit timeline for 8 random patients



HIN/HIE data filled in missing visits

HIN/HIE data greatly improved vaccination records

Cedars Sinai FHIR vs Cedars Sinai CCDA vs DRC: Influenza vaccine records for 60 random patients



OHDSI 2025

Lessons Learned

- **USCDI/US Core gaps:** The returned payload was not fully USCDI v2–compliant per FHIR US Core 3.1.1; local EPIC code system used
- **HIE coverage gaps:** Certain data elements and care transitions may be absent from HIE networks; for instance, hospital discharges to skilled nursing facilities may be missing.
- **Transform loss:** The **FHIR** → **OMOP** conversion can be **lossy** (reduced granularity/context) FHIR resources are richer in content. Additional OMOP Domain to support multimodal data
- **What we found in pilot implementation:** Even with mapping gaps, Cedars-Sinai sample analyses demonstrated research value and filled important data gaps.
- As an interoperable standard, **FHIR** enables research-quality EHR capture and ***reduces key data gaps.***

Thank You

CLAD HIE-HIN FHIR Pilot Team	
Stephanie Hong MS, FAMIA Thanaphop Na Nakhonphanom, MD Bryan Laraway, MS Tanner Zhang MD, MS Yvette Chen, MS Richard Moffitt, PhD Rob Schuff, MS	Tursynay Issabekova, MBA Christopher Chute, MD, DrPH Josh Lemieux, BA Melissa Haendel, PhD William Hogan MD, MS Emily Pfaff, MS, PhD Shahim Essaid, MD

For collaboration inquiries, contact Melissa Haendel at info@cladteam.io.

This work was supported by NIH All of Us Research Program (Award OT2 OD036113-01)

OHDSI 2025

OMOP Waveform Extension: A Schema for Integrating Physiological Signals and Derived Features into the OMOP CDM

Jared Houghtaling[1], Polina Talapova[1,4], Brian Gow[2], Manlik Kwong[1], Andrew J King[3], Benjamin Moody[2], Mike Kriley[3], Tom Pollard[2], Andrew E Williams[1]

[1] *Tufts University School of Medicine, Boston, MA, USA*
[2] *Massachusetts Institute of Technology, Cambridge, MA, USA*
[3] *University of Pittsburgh, Pittsburgh, PA, USA*
[4] *SciForce, Ukraine*

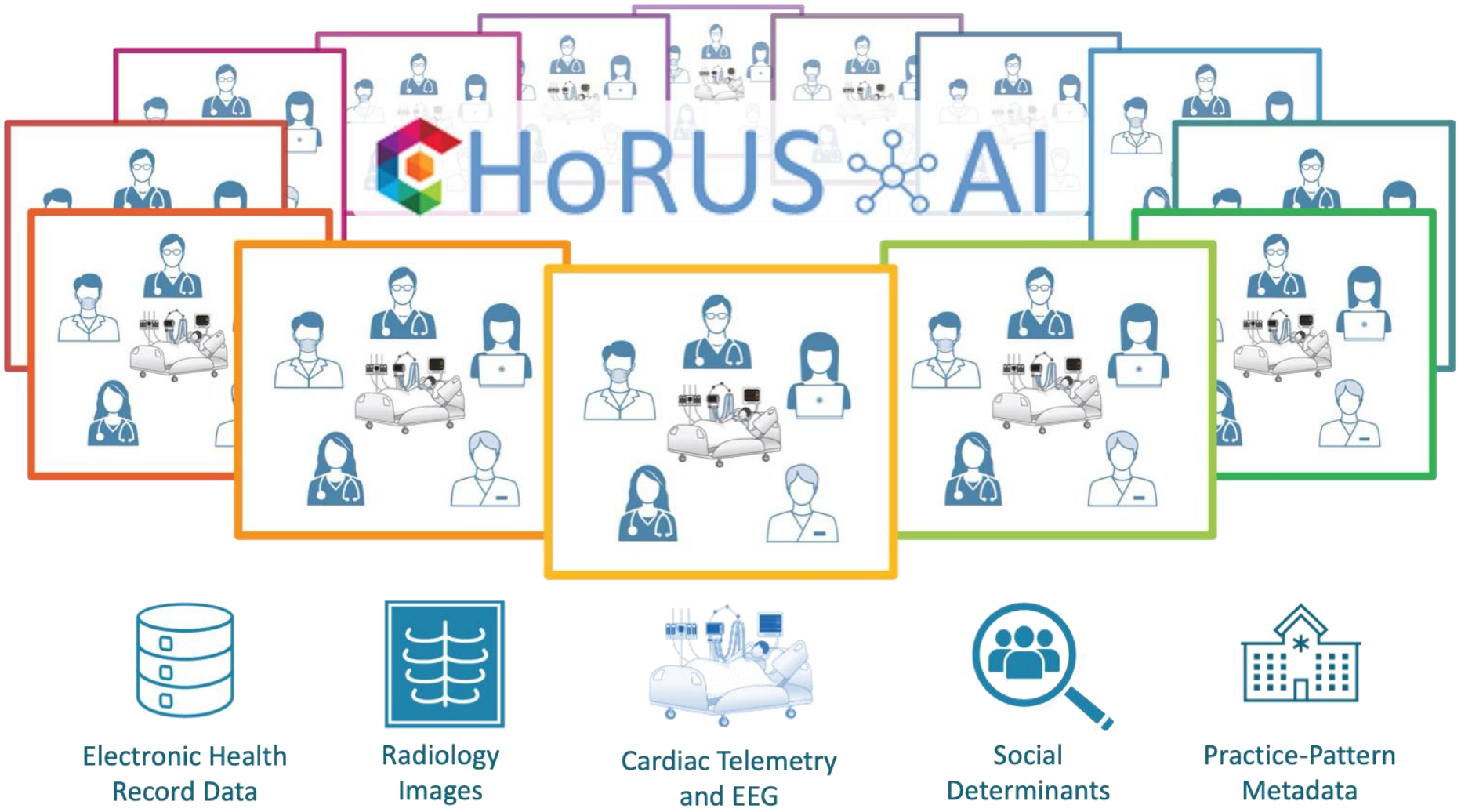
2025 OHDSI Global Symposium – Lightning Talk

Jared Houghtaling
Asst. Professor – Clinical Informatics

Content

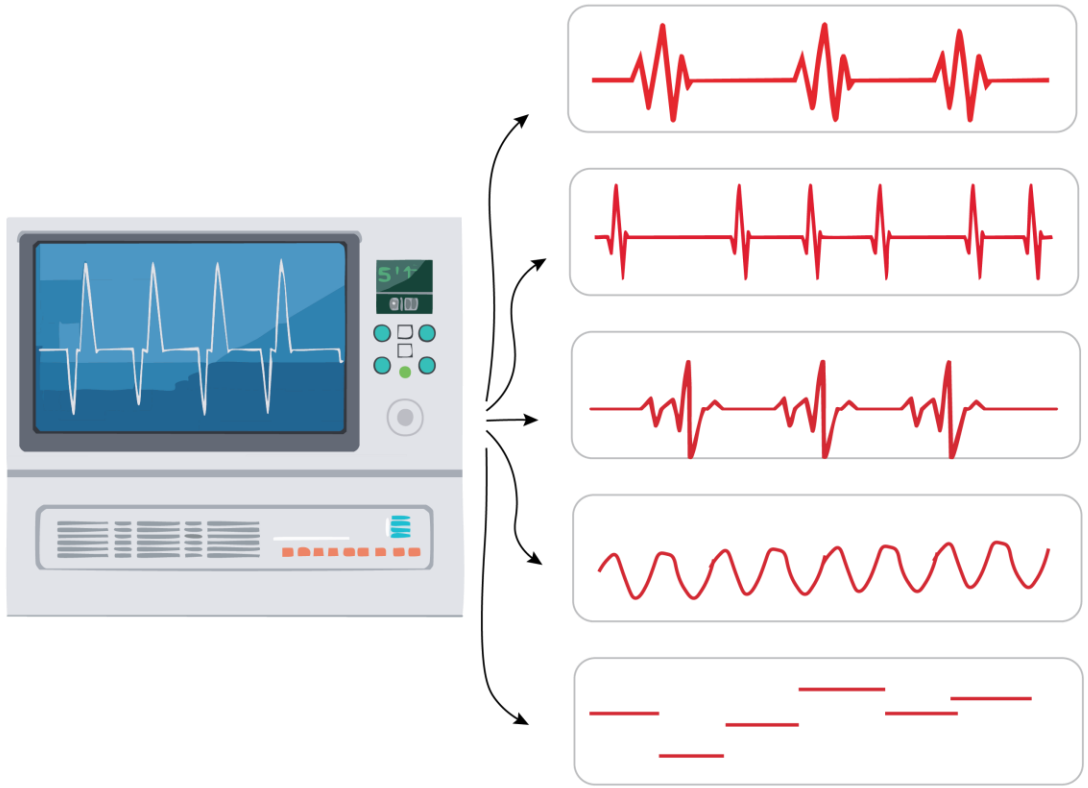
- Background: Bridge2AI – CHoRUS
- Motivations for Extension
- Waveform Data Collection and Organization
- Key Table Elements
- Next Steps
- Acknowledgements

Background: Bridge2AI – CHoRUS



Another OMOP Extension?

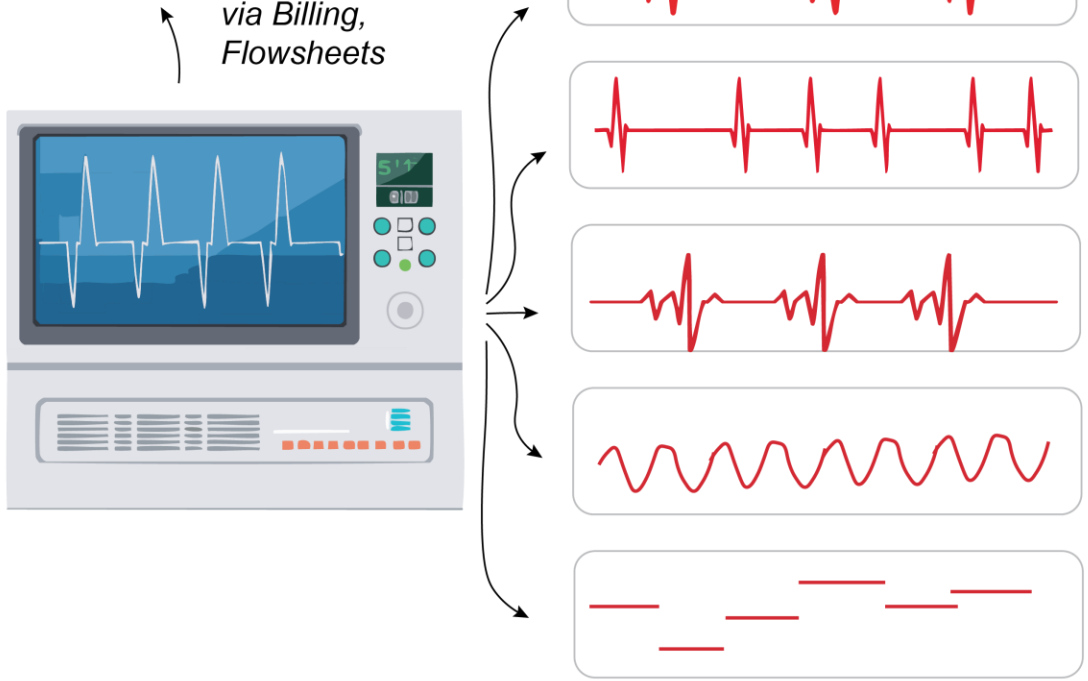
1. CTS data are DIVERSE – Interoperability is challenging
   - Multiple relevant metadata elements
   - Broad range of recording durations (seconds to weeks or months)
   - Many recording paradigms and formats (hdf5, wfdb, atriumDB, parquet, etc.)
   - Many file structures, extensions, segments
   - Many signals relevant to health (e.g. acoustic, electrical, physical movement)



DEVICE EXPOSURE

PROCEDURE OCCURRENCE

via Billing, Flowsheets



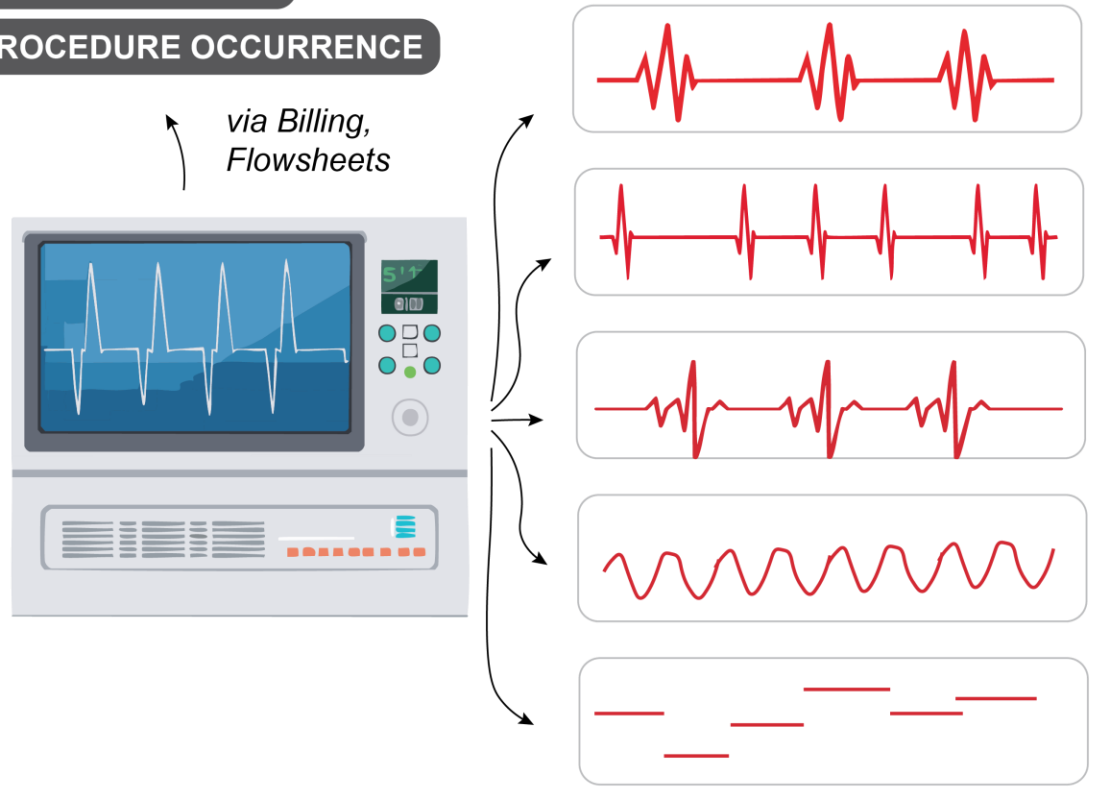
MEASUREMENT

via EHR-Device Integrations

DEVICE EXPOSURE

PROCEDURE OCCURRENCE

via Billing, Flowsheets



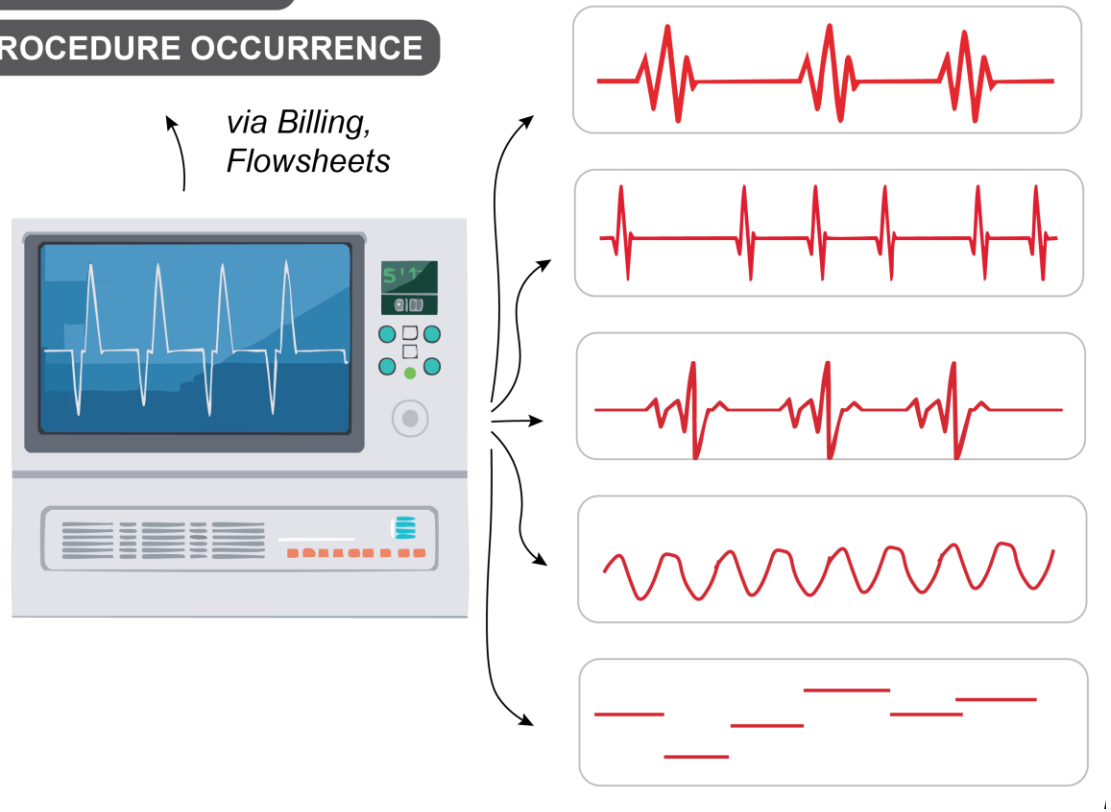
MEASUREMENT

via EHR-Device Integrations

DEVICE EXPOSURE

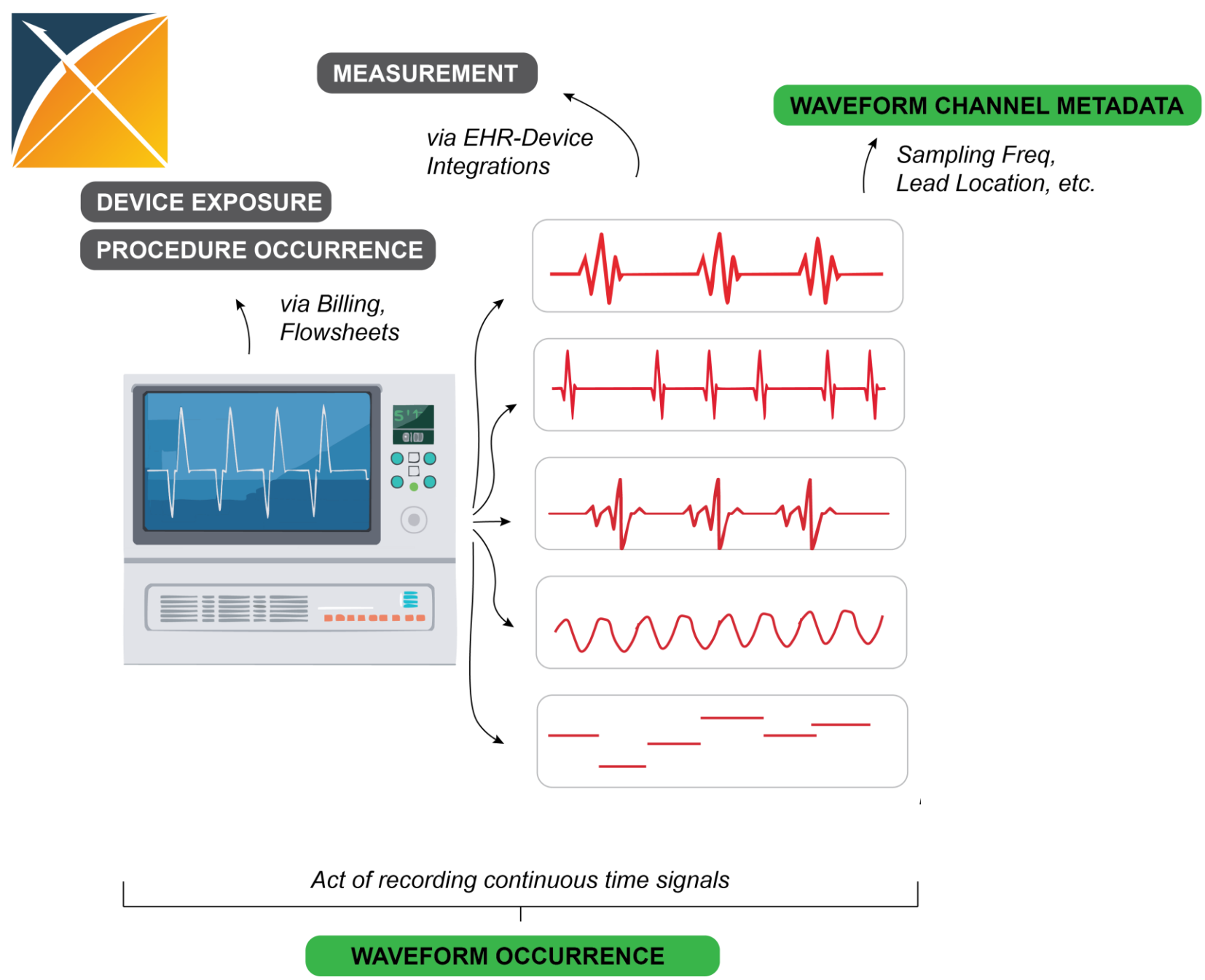
PROCEDURE OCCURRENCE

via Billing, Flowsheets



Act of recording continuous time signals

WAVEFORM OCCURRENCE


MEASUREMENT
via EHR-Device Integrations
WAVEFORM CHANNEL METADATA
Sampling Freq, Lead Location, etc.
DEVICE EXPOSURE
PROCEDURE OCCURRENCE
via Billing, Flowsheets
Act of recording continuous time signals
WAVEFORM OCCURRENCE

MEASUREMENT

via EHR-Device Integrations

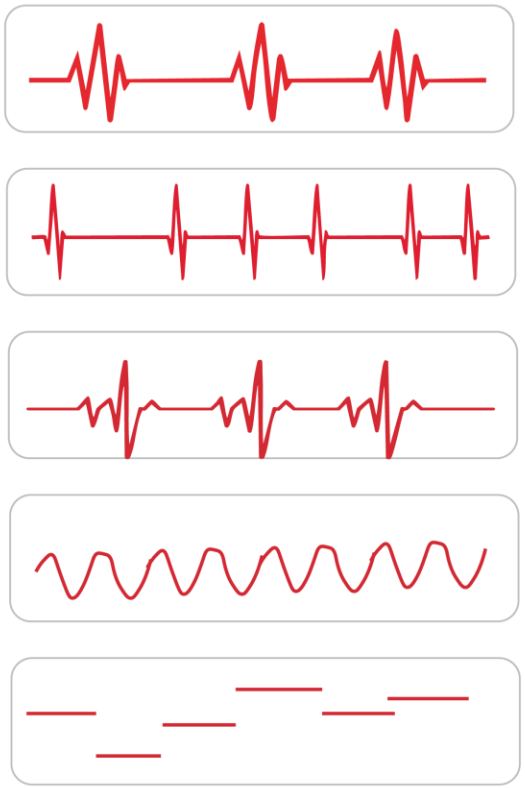
WAVEFORM CHANNEL METADATA

Sampling Freq, Lead Location, etc.

DEVICE EXPOSURE

PROCEDURE OCCURRENCE

via Billing, Flowsheets

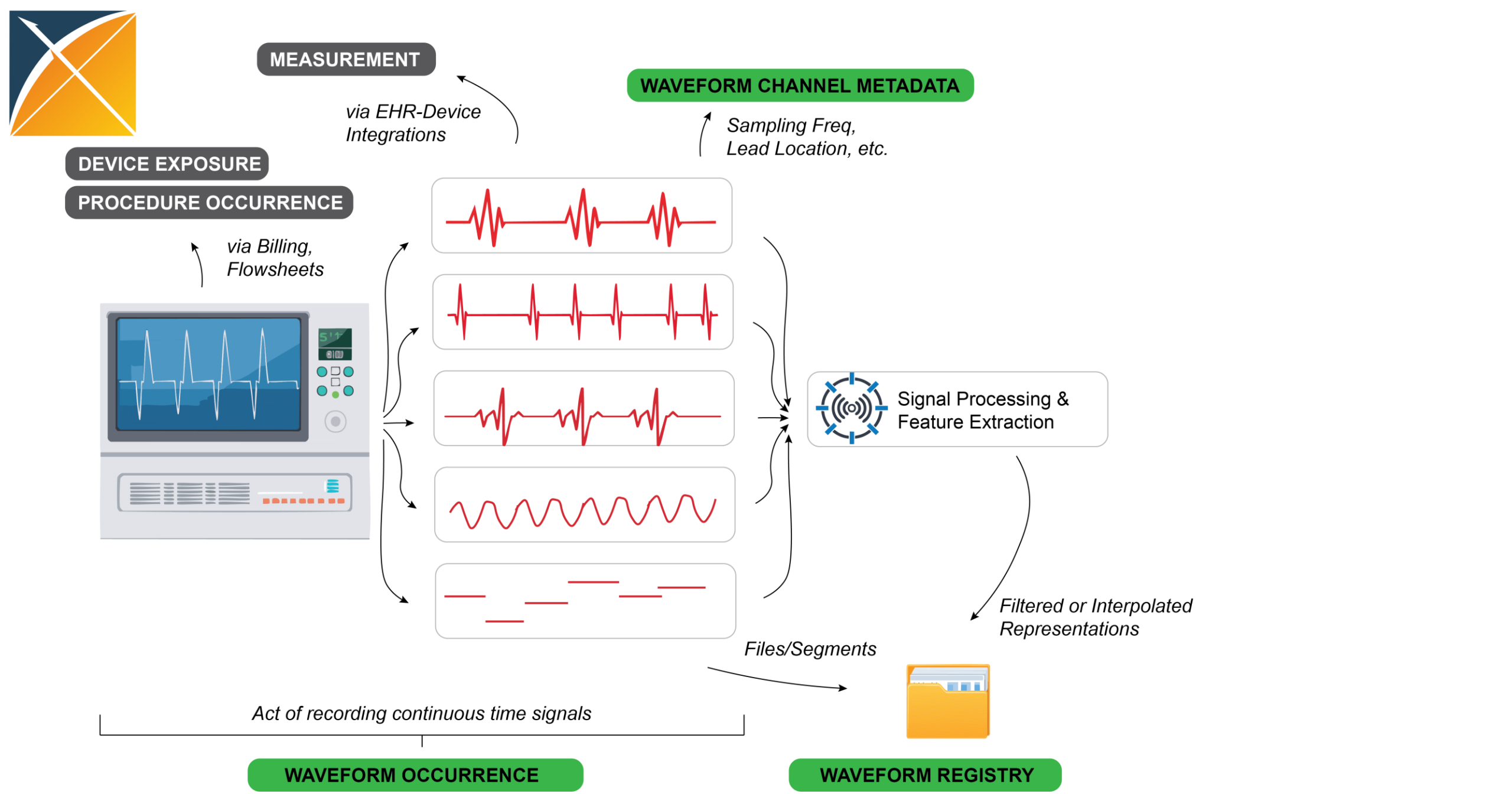


Files/Segments

Act of recording continuous time signals

WAVEFORM OCCURRENCE

WAVEFORM REGISTRY


MEASUREMENT
via EHR-Device Integrations
WAVEFORM CHANNEL METADATA
Sampling Freq, Lead Location, etc.
DEVICE EXPOSURE
PROCEDURE OCCURRENCE
via Billing, Flowsheets
Signal Processing & Feature Extraction
Filtered or Interpolated Representations
Files/Segments
Act of recording continuous time signals
WAVEFORM OCCURRENCE
WAVEFORM REGISTRY

MEASUREMENT

via EHR-Device Integrations

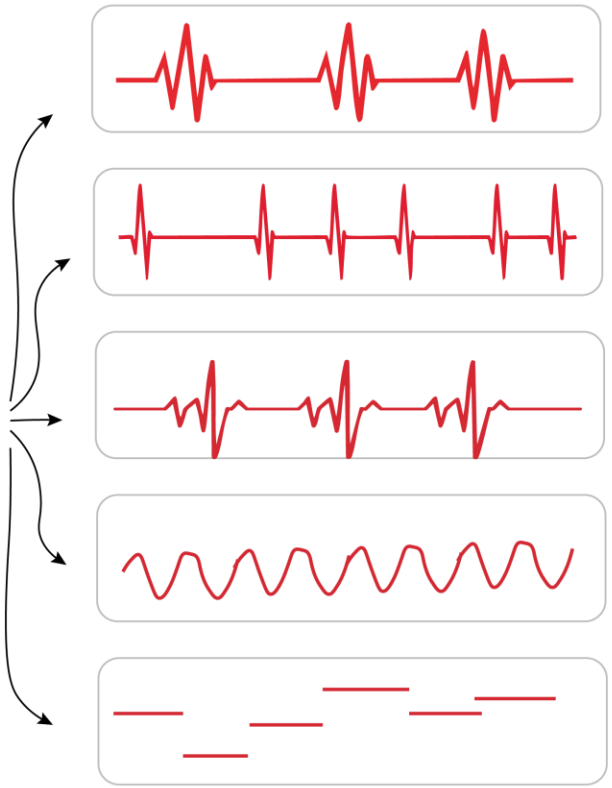
WAVEFORM CHANNEL METADATA

Sampling Freq, Lead Location, etc.

DEVICE EXPOSURE

PROCEDURE OCCURRENCE

via Billing, Flowsheets



Signal Processing & Feature Extraction

WAVEFORM FEATURE

Feature Name	Feature Value	Wave Occ Id	
P-Wave Dur.	112 ms	73	
QT Duration	394 ms	73	
RR Interval	837 ms	73	
Avg. Blood Ox.	94 %	73	
Avg. Resp. Rt.	14 br/m	73	
Avg. Heart. Rt.	83 bpm	73	

Filtered or Interpolated Representations

Files/Segments

Act of recording continuous time signals

WAVEFORM OCCURRENCE

WAVEFORM REGISTRY

MEASUREMENT

via EHR-Device Integrations

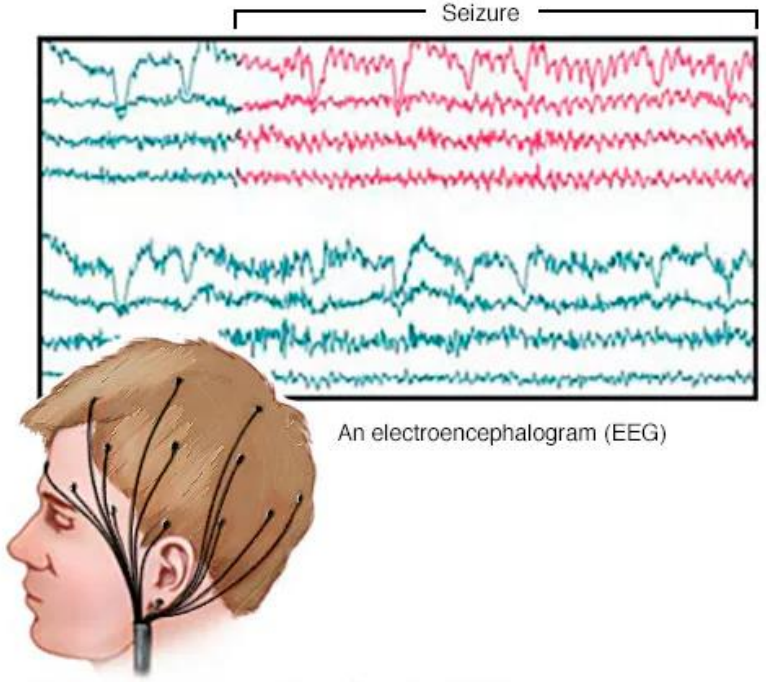
WAVEFORM CHANNEL METADATA

Sampling Freq, Lead Location, etc.

DEVICE EXPOSURE

PROCEDURE OCCURRENCE

via Billing, Flowsheets



An electroencephalogram (EEG)

Signal Processing & Feature Extraction

WAVEFORM FEATURE

Feature Name	Feature Value	Wave Occ Id	
P-Wave Dur.	112 ms	73	
QT Duration	394 ms	73	
RR Interval	837 ms	73	
Avg. Blood Ox.	94 %	73	
Avg. Resp. Rt.	14 br/m	73	
Avg. Heart. Rt.	83 bpm	73	

Filtered or Interpolated Representations

Files/Segments

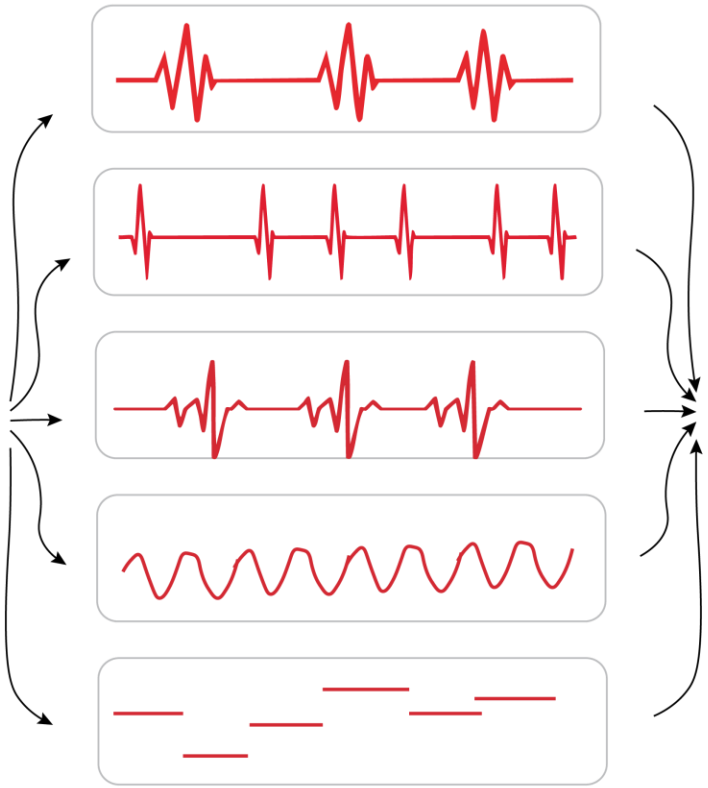
Act of recording continuous time signals

WAVEFORM OCCURRENCE

WAVEFORM REGISTRY

WAVEFORM CHANNEL METADATA

Sampling Freq,
Lead Location, etc.



Signal Processing & Feature Extraction

WAVEFORM FEATURE

Feature Name	Feature Value	Wave Occ Id	
P-Wave Dur.	112 ms	73	
QT Duration	394 ms	73	
RR Interval	837 ms	73	
Avg. Blood Ox.	94 %	73	
Avg. Resp. Rt.	14 br/m	73	
Avg. Heart. Rt.	83 bpm	73	

Filtered or Interpolated Representations

Files/Segments

Act of recording continuous time signals

WAVEFORM OCCURRENCE

WAVEFORM REGISTRY

Key Table Elements and Motivations

1. WAVEFORM_OCCURRENCE
    1. Which type of recording process was performed (*waveform_occurrence_concept_id*)
    2. When did the acquisition start (*waveform_occurrence_start_datetime*)
    3. How many files were acquired (*num_of_files*)

Less Granular

More Granular

Next Steps

- Finalize semantic representations of waveform-specific terminology (see CVB)
- Evaluate model across 14-site consortium in Bridge2AI – CHoRUS
- Launch an OHDSI workgroup (!) to evaluate new use cases and expand upon the extension
  - Come talk to me if you're interested!

Acknowledgements

- Jen Park and Paul Nagy

- Marty Alvarez

- OHDSI members and developers

- Eric Rosenthal & CHoRUS Research Consortium

- All amazing co-authors: *Polina Talapova, Brian Gow, Manlik Kwong, Andrew J King, Benjamin Moody, Mike Kriley, Tom Pollard,* ***Andrew E Williams***

Useful Links

github.com/TuftsCTSI/CVB

github.com/chorus-ai

Improving VSAC to OMOP Mapping Using LLM Assisted Curation

Robert B Barrett[a], Star Liu[a]; Kyle Zollo-Venecek[b]; Benjamin Riesser, MPA[c] ; Benjamin Martin, PhD[a]

[a]Biomedical Informatics and Data Science, Johns Hopkins University, Baltimore, MD, USA

[b]CTSI, Tufts University School of Medicine, Boston, MA

[c]Improving Health Outcomes, American Medical Association, Greenville, SC

Objectives

- Understanding the limitations in Value Set usage
- Understanding current challenges with mapping Value Sets to OMOP Concepts
- Understanding potentials and pitfalls of LLMs to validate these mappings

What is Type 2 Diabetes?

- SNOMED CT => 44054006 (Type 2 diabetes mellitus)
- ICD10 => E11.x (Type 2 diabetes mellitus)
- RxCUI =>
  - 253182 (insulin, regular, human)
  - 253181 (Insulin isophane)
  - 2380230 (Insulin lispro)
  - ……
- **Computable phenotype**

Value Sets and VSAC

- **The Value Set Authority Center (VSAC)** attempts to standardize these clinical concepts as shareable concepts
- They have **over 15,214 public value sets** (July 2024)
  - VSAC **does not** review these
  - Stewards including the Joint Commission, CMS, Mathematica, etc. **do**

☐	Type 2 Diabetes Diagnoses	ICD10CM	Extensional	AMA	2.16.840.1.113762.1.4.1160.27	Active	2025-07-30	96
☐	Type 2 Diabetes Diagnoses	SNOMEDCT	Extensional	AMA	2.16.840.1.113762.1.4.1160.28	Active	2025-04-19	20
☐	Type 2 Diabetes Diagnoses	ICD10CM SNOMEDCT	Grouping	AMA	2.16.840.1.113762.1.4.1160.29	Active	2025-07-30	116
☐	Type 2 Diabetes Mellitus, Type Two Diabetes Mellitus	ICD9CM	Extensional	VU eMERGE	2.16.840.1.113762.1.4.1053.2	Not Maintained	2015-10-21	18
☐	Type II Diabetes	SNOMEDCT	Extensional	NCQA	2.16.840.1.113883.3.464.1003.103.11.1	Not Maintained	2025-04-19	45
☐	Type II Diabetes	ICD10CM	Extensional	NCQA	2.16.840.1.113883.3.464.1003.103.11.1	Not Maintained	2025-07-30	92
☐	Type II Diabetes	ICD10CM SNOMEDCT	Grouping	NCQA	2.16.840.1.113883.3.464.1003.103.12.1	Not Maintained	2025-07-30	137

Application Matters

- The application defines phenotype
  - Clinical Decision Support (uncontrolled type 2 diabetes)
  - Research (all patients with any kind of diabetes)
  - **Quality Measures (???)**

How does concept variance affect quality measures?

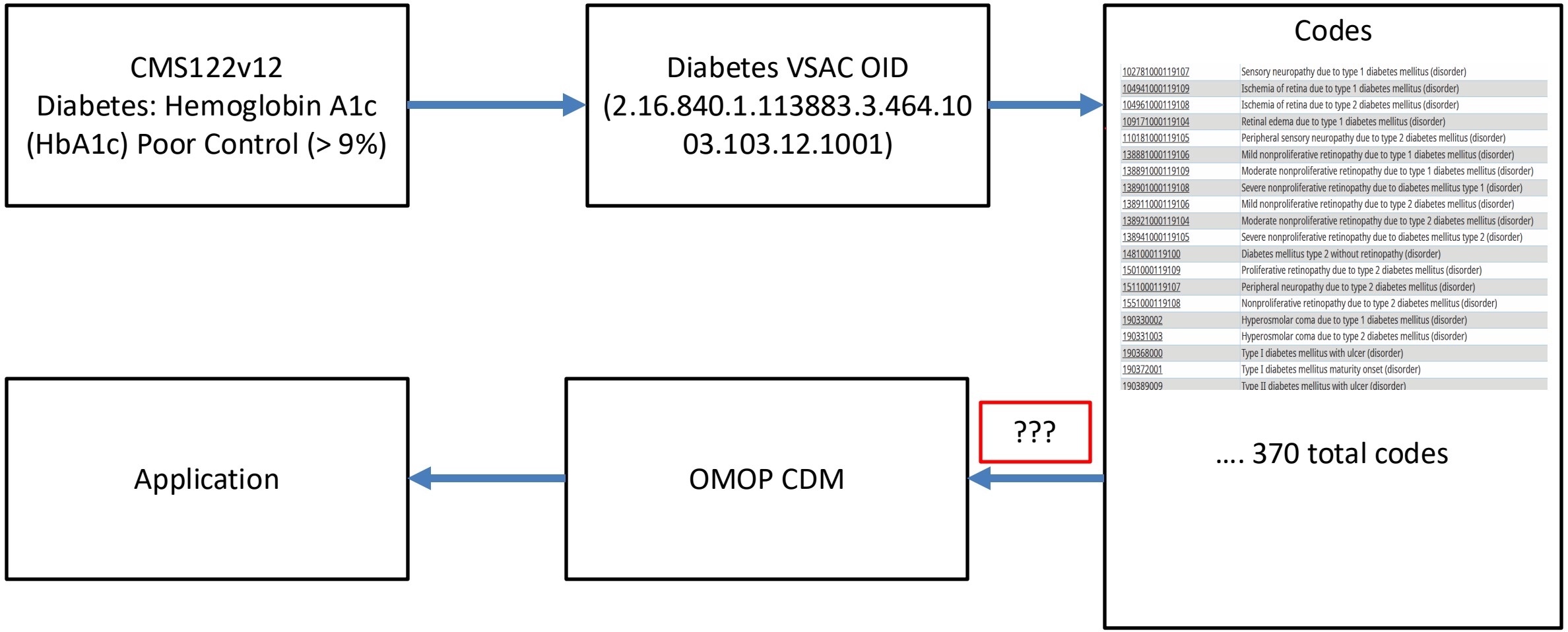
- Value set variation cause big measure change

Figure 2. Heat map showing the percentage of patients having MI and taking a beta blocker using various combinations of value sets.

	ACO ERX BETA BLOCKER ORDERED	ACO ERX PRESCRIBED BETA BLOCKER	AMB ERX GENERAL BETA-BLOCKERS NON-COMBO,NON-SOTALOL ORAL	ERX CV ICD BETA BLOCKER (ANY)	ERX GENERAL BETA BLOCKERS	ERX GENERAL BETA BLOCKERS PQRS MEASURE 7,8	ERX GENERAL HEDIS CARDIO BETA BLOCKERS	ERX GENERAL JCCM BETA BLOCKERS	ERX PERIOPERATIVE BETA BLOCKERS	ERX SYSTEMIC BETA-BLOCKERS
EDG CONCEPT CV ACC MYOCARDIAL INFARCTION (n= 20808)	58.3	58.3	72.6	72.6	74.9	75.1	74.6	74.8	57.7	74.8
EDG CONCEPT HX MYOCARDIAL INFARCTION (n=20808)	58.3	58.3	72.6	72.6	74.9	75.1	74.6	74.8	57.7	74.8
EDG ICD 2018 ACO AMI (IVD-2) (n=20179)	58.4	58.4	72.7	72.7	75	75.2	74.7	74.8	57.8	74.8
EDG ICD CHARLSON COMORBIDITY MYOCARDIAL INFARCTION (n=22594)	57.2	57.2	71.8	71.8	74.1	74.3	73.8	73.9	57.1	73.9
EDG ICD CHARLSON SCORE MYOCARDIAL INFARCTION (n=22512)	57.4	57.4	72.1	72.1	74.4	74.5	74	74.2	57.3	74.2
EDG ICD CMS CCM ACUTE MYOCARDIAL INFARCTION (n=13974)	60.8	60.8	75.2	75.2	77.2	77.3	76.9	77	60.1	77
EDG ICD CMS READMISSION RATE AMI DIAGNOSES (n=20874)	57.9	57.9	72.1	72.1	74.4	74.6	74.1	74.2	57.2	74.2
EDG ICD CMS-HCC 86: ACUTE MYOCARDIAL INFARCTION (n=20196)	58.4	58.4	72.7	72.7	75	75.2	74.7	74.8	57.8	74.8

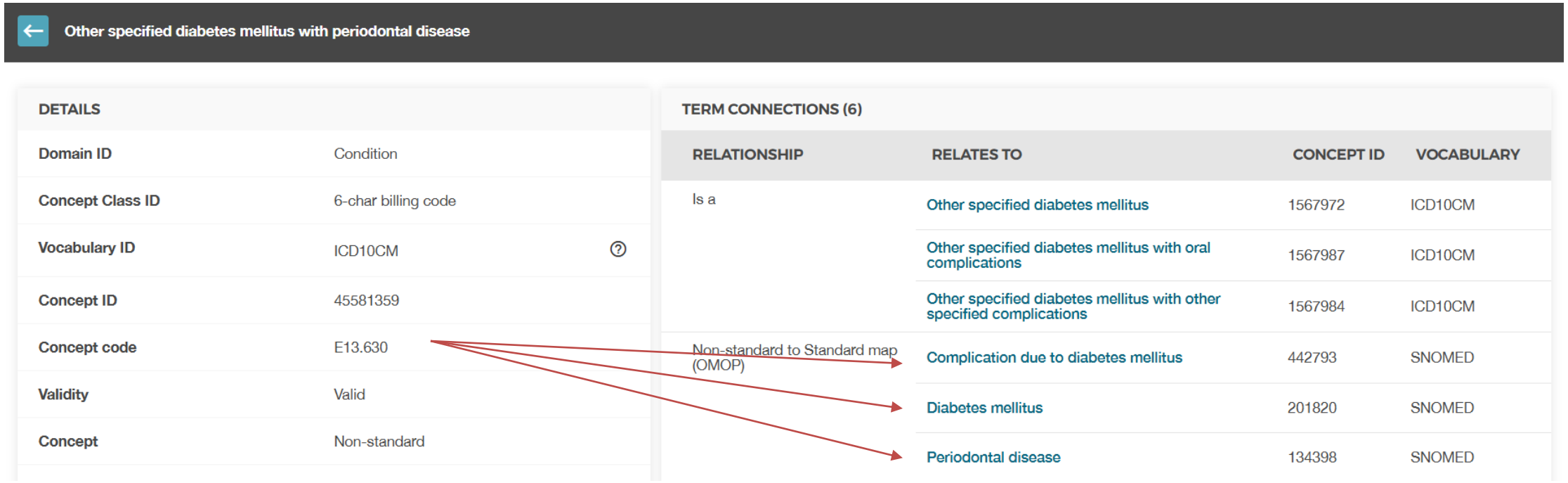
1. Zahn LA, Ahmad H, Sittig DF, et al. The Fault in Our Sets: A Mixed Methods Analysis of Clinical Value Set Errors. *medRxiv*. Preprint posted 2025-02-27; doi:10.1101/2025.02.27.25323054.

From Measure to OMOP



Mapping Problems

- VSAC Value Sets may contain non-standard concepts
- Potential loss of semantic fidelity on mapping
- Value Set concepts are not equally appropriate for all applications



Goal

- **Evaluate** the use of LLMs for **mapping VSAC concepts to OMOP** and **relevance to value set intention**.

Methods

Diabetes VSAC OID (2.16.840.1.113883.3.464.1003.103.12.1001)

VSAC APIs

Value Set Concepts (Source)

102781000119107	Sensory neuropathy due to type 1 diabetes mellitus (disorder)
104941000119109	Ischemia of retina due to type 1 diabetes mellitus (disorder)
104961000119108	Ischemia of retina due to type 2 diabetes mellitus (disorder)
109171000119104	Retinal edema due to type 1 diabetes mellitus (disorder)
110181000119105	Peripheral sensory neuropathy due to type 2 diabetes mellitus (disorder)
138881000119106	Mild nonproliferative retinopathy due to type 1 diabetes mellitus (disorder)
138891000119109	Moderate nonproliferative retinopathy due to type 1 diabetes mellitus (disorder)
138901000119108	Severe nonproliferative retinopathy due to diabetes mellitus type 1 (disorder)
138911000119106	Mild nonproliferative retinopathy due to type 2 diabetes mellitus (disorder)
138921000119104	Moderate nonproliferative retinopathy due to type 2 diabetes mellitus (disorder)
138941000119105	Severe nonproliferative retinopathy due to diabetes mellitus type 2 (disorder)
1481000119100	Diabetes mellitus type 2 without retinopathy (disorder)
1501000119109	Proliferative retinopathy due to type 2 diabetes mellitus (disorder)

'Maps to'

OMOP Standard Concepts

omop_concept_id	omop_source_concept_id	concept_code
134398	45581359	E13.630
134398	45557111	E10.630
192279	45552388	E13.29
192279	45542741	E13.21
192279	45561955	E13.22
192279	192279	127013003
194700	194700	76751001

Value Set Metadata

Purpose:

Clinical Focus:

The purpose of this value set is to represent concepts for a diagnosis of diabetes.

Inclusion Criteria:

Includes concepts that represent a diagnosis of type I or type II diabetes.

GPT 4o

Concepts Equivalent?
Broader? Narrower?

Relevant to Value Set?
In Scope/Out of Scope

Methods

Diabetes VSAC OID (2.16.840.1.113883.3.464.1003.103.12.1001)

VSAC APIs

Value Set Concepts (Source)

102781000119107	Sensory neuropathy due to type 1 diabetes mellitus (disorder)
104941000119109	Ischemia of retina due to type 1 diabetes mellitus (disorder)
104961000119108	Ischemia of retina due to type 2 diabetes mellitus (disorder)
109171000119104	Retinal edema due to type 1 diabetes mellitus (disorder)
110181000119105	Peripheral sensory neuropathy due to type 2 diabetes mellitus (disorder)
138881000119106	Mild nonproliferative retinopathy due to type 1 diabetes mellitus (disorder)
138891000119109	Moderate nonproliferative retinopathy due to type 1 diabetes mellitus (disorder)
138901000119108	Severe nonproliferative retinopathy due to diabetes mellitus type 1 (disorder)
138911000119106	Mild nonproliferative retinopathy due to type 2 diabetes mellitus (disorder)
138921000119104	Moderate nonproliferative retinopathy due to type 2 diabetes mellitus (disorder)
138941000119105	Severe nonproliferative retinopathy due to diabetes mellitus type 2 (disorder)
1481000119100	Diabetes mellitus type 2 without retinopathy (disorder)
1501000119109	Proliferative retinopathy due to type 2 diabetes mellitus (disorder)

'Maps to'

OMOP Standard Concepts

omop_concept_id	omop_source_concept_id	concept_code
134398	45581359	E13.630
134398	45557111	E10.630
192279	45552388	E13.29
192279	45542741	E13.21
192279	45561955	E13.22
192279	192279	127013003
194700	194700	76751001

Value Set Metadata

Purpose:

Clinical Focus:

The purpose of this value set is to represent concepts for a diagnosis of diabetes.

Inclusion Criteria:

Includes concepts that represent a diagnosis of type I or type II diabetes.

Human Reviewers

Concepts Equivalent?
Broader? Narrower?

Relevant to Value Set?
In Scope/Out of Scope

Results

Concept Set	H vs LLM Concept Rel.	H vs LLM Value Set Rel.	Inter-Human Concept Rel.	Inter-Human Value Set Rel.
Diabetes	91.1%	95.7%	93.1%	96.6%
Beta Blocker	100.0%	100.0%	100.0%	100.0%
Dialysis Services	100.0%	98.0%	100.0%	100.0%
Hypertension	98.5%	100.0%	97.1%	100.0%
Kidney Transplant	95.0%	100.0%	96.7%	100.0%
Office Visit	85.7%	100.0%	71.4%	100.0%
Overall (mean)	**95.2%**	**99.0%**	**93.0%**	**99.4%**

- **Humans agreed with the LLM** more than each other for **concept equivalence**
- **Humans agreed with each other** more than LLM for **value set relevance**

Results

LLM identified nuance related to metadata that reviewers did not

- Hemoperfusion is not exclusively a dialysis service
- Diabetes mellitus in mother complicating pregnancy may meet exclusion criteria: *Excludes concepts that represent a diagnosis of gestational diabetes or steroid-induced diabetes.*

Value Set	VSAC Concept	OMOP Concept	LLM Perspective
Dialysis Services	**Hemoperfusion (eg, with activated charcoal or resin)**	**Hemoperfusion**	**Exclude (out-of-scope): "not a dialysis service"**
Diabetes	Type 1 diabetes mellitus with periodontal disease	Periodontal disease	Narrower relationship; exclude: missing diabetes qualifier
Diabetes	Other specified diabetes mellitus with periodontal disease	Periodontal disease	Narrower relationship; exclude: missing diabetes qualifier
Diabetes	**Diabetes mellitus in mother complicating pregnancy**	**Diabetes mellitus in mother complicating pregnancy**	**Exclude despite equivalence (85% confidence)**
Diabetes	Type 1 diabetes mellitus with diabetic nephropathy	Renal disorder due to type 1 diabetes mellitus	Equivalent (LLM) vs Broader (Humans)
Diabetes	Uncontrolled type 1 diabetes mellitus	Type 1 diabetes mellitus	Broader (LLM) vs Equivalent (Humans)

Future Directions

- Automated generation of quality measures on OMOP CDM
- Tuning quality measures
  - Evaluating quality measure change when adjusted for LLM-curated value sets
- Human-in-the-loop interventions
  - Flagging potentially inappropriate concepts for review and refinement

Conclusion

- **LLMs serve as an effective screening process**, accounting for **application intent, relevance, and equivalence**
- **This process is efficient and scalable**
  - For CMS165 (Controlling High Blood Pressure), **over 5000 OMOP Concepts are mapped from VSAC Value Sets in a few minutes!**
  - Manual review of these concepts at scale is impractical
  - **Human-in-the-loop intervention** may improve reliability and governance of Value Set use in conjunction with LLM screening processes

Evaluating the effectiveness of using Large Language Models for the development of concept sets.

Joel Swerdel, PhD MS MPH[1,2], Dmytro Dymshyts MD[1,2], Anna Ostropolets MD PhD[1,2], Azza Shoaibi, PhD[1,2], Patrick Ryan, PhD[1,2], and Martijn Schuemie PhD[1,2]

[1]Johnson & Johnson, Titusville, NJ USA; [2]Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA

Presented at OHDSI Symposium; October 8, 2025; New Brunswick, NJ USA

Disclosures

Joel Swerdel, Dmytro Dymshyts, Anna Ostropolets, Azza Shoaibi, Patrick Ryan, and Martijn Schuemie are employees and shareholders of Johnson & Johnson.

Background

- When developing phenotype definitions for health conditions, one of the first steps is to develop the list of codes used to determine the phenotype.
- In OHDSI, health condition standard codes are usually SNOMED concepts mapped from diagnosis source codes (such as ICD-10-CM) used in the native data.
- Selected concepts are then grouped into concept sets
- Our current method for creating concept sets for phenotype development is inefficient, subjective, and inconsistent.

Objective

The objective of this study was to evaluate the use of Large Language Models (LLMs) for the selection of appropriate condition codes for concept sets used in phenotype algorithms.

Methods

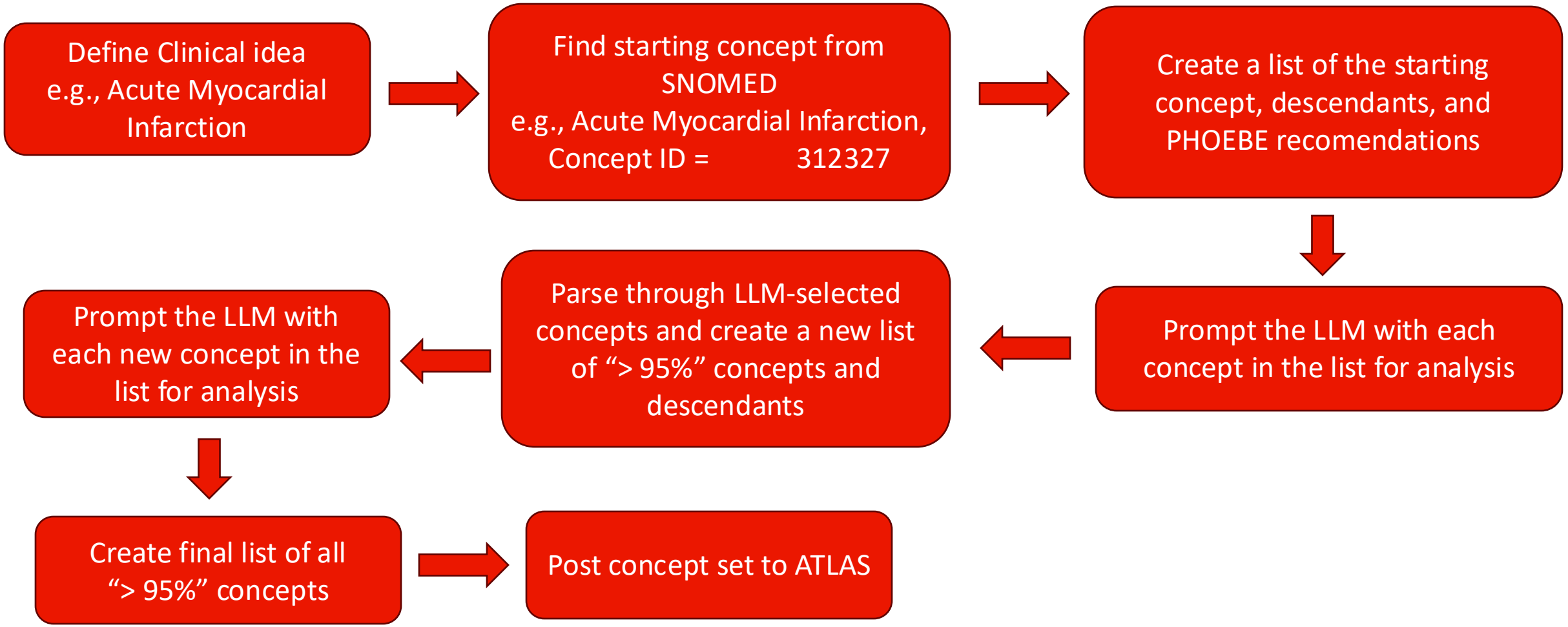
Using the LLM to adjudicate concepts for a concept set

- We used the LLM to adjudicate whether a concept belongs in a concept set for a clinical idea.

- Example: does “Coronary occlusion” (suggested concept) belong in a concept set for “Acute Myocardial Infarction” (clinical idea)

- The LLM was prompted to tell us whether “greater than 95% of patients with Coronary occlusion have Acute Myocardial Infarction”

Process Steps



Methods

- We tested the LLM process on 15 health conditions
  - Acute health conditions such as acute myocardial infarction
  - Chronic health conditions such as plaque psoriasis

- For each condition, we created a human-adjudicated concept set developed through a collaboration of at least two researchers knowledgeable in concept set development.

- These concept sets were the “reference standard” to be used for comparison with the LLM-adjudicated concept sets.

Methods (cont.)

- For this study we used the licensed Johnson & Johnson version of the OpenAI LLM, (OpenAI Model GPT-4o, trained through October 2023).
- Procedural calls to the application programming interface (API) for the LLM were made using the R platform.

Evaluation

Concept Evaluation:

1) The number of **common concepts** in the LLM and human generated concept sets
2) The number of **concepts only in the LLM** generated concept set
3) The number of **concepts only in the human** generated concept set

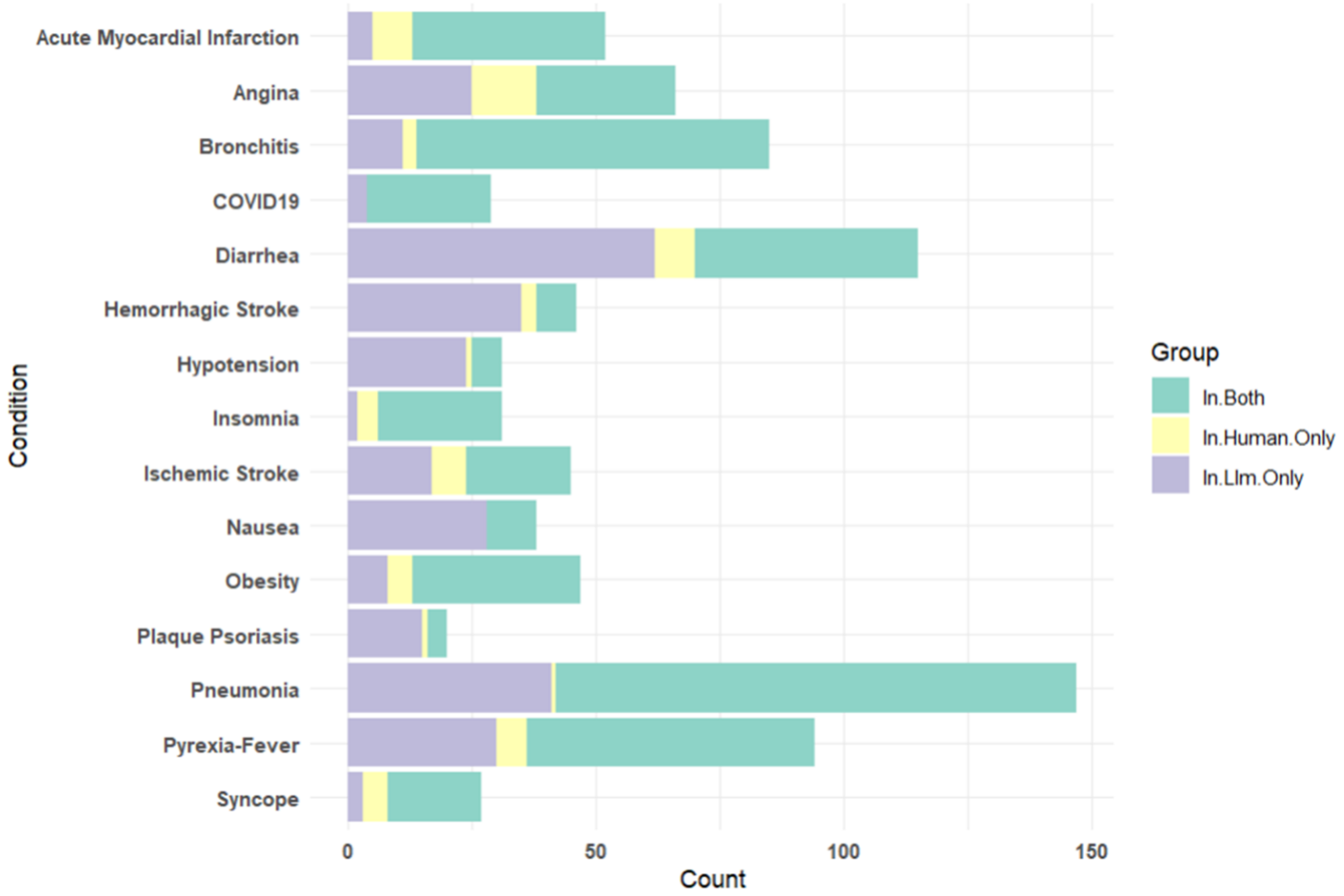
Subject Evaluation:

- Cohort comparison using cohorts based on the LLM and human-generated concept sets.
- Used the Merative Commercial Claims and Encounters (CCAE) database.

1) The number of **common subjects** in the LLM and human generated concept set cohorts
2) The number of **subjects only in the LLM** generated concept set cohorts
3) The number of **subjects only in the human** generated concept set cohorts

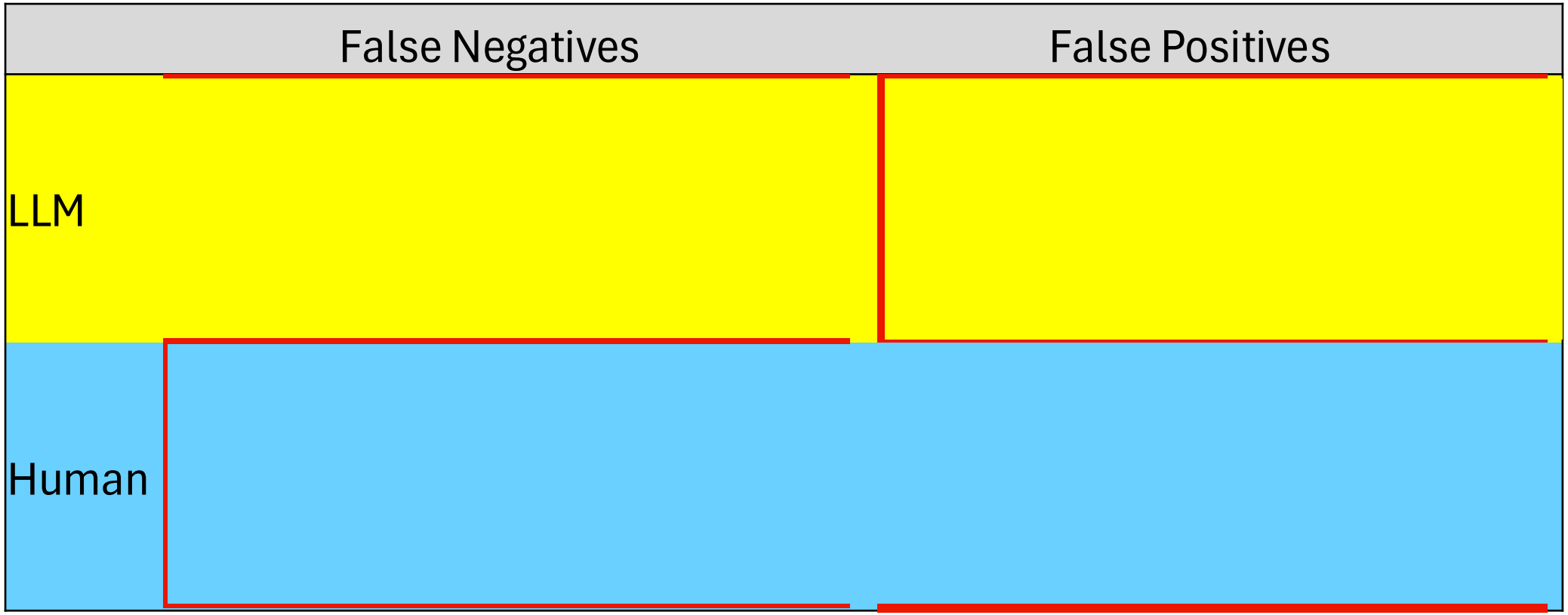
Results

Concept Comparison

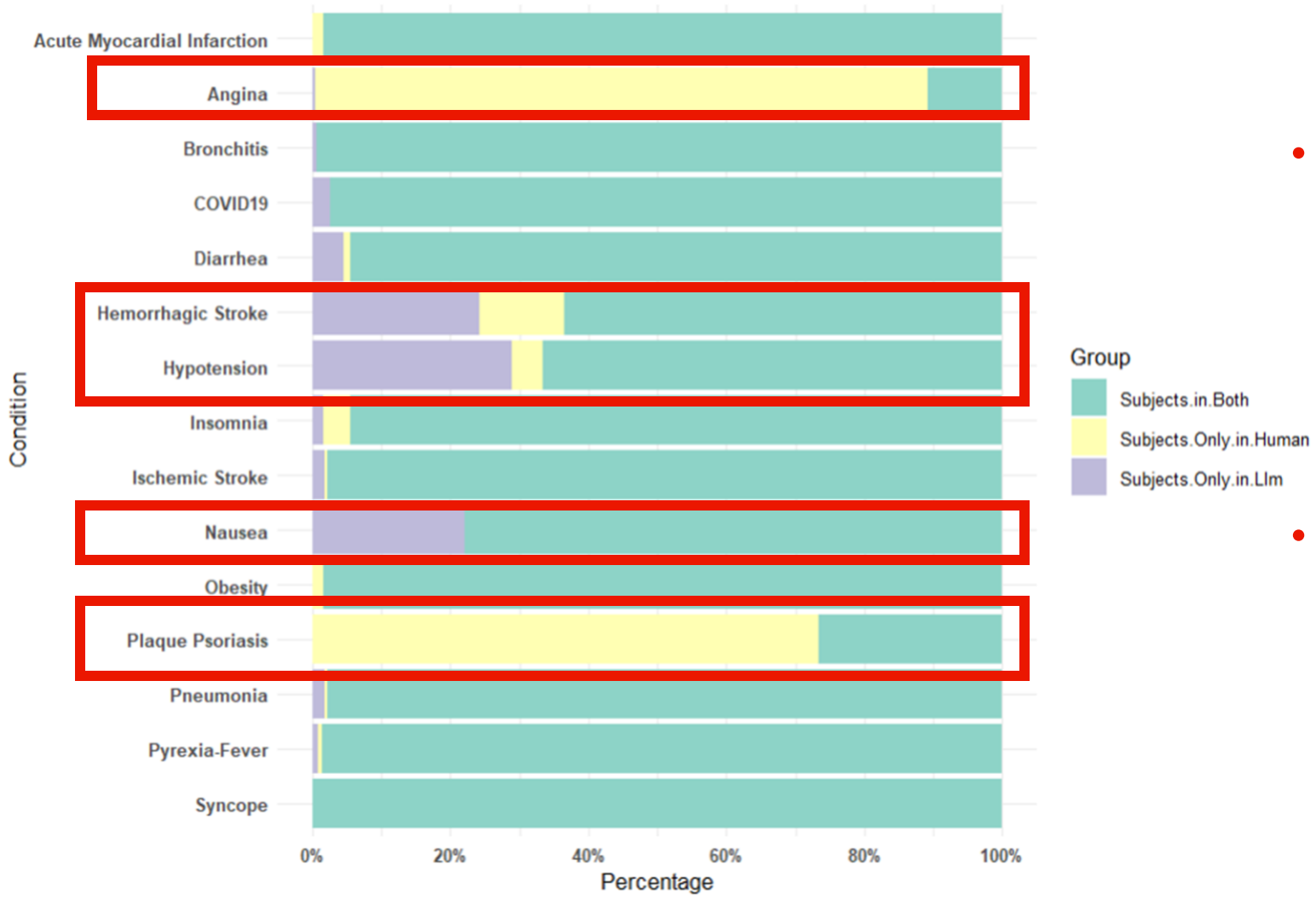


- The majority of concepts were shared by the human and the LLM adjudicated concept sets. (10 conditions/15 total conditions)
- In the other five conditions, the LLM determined more concepts than were overall shared.

Example Errors – LLM vs. Human for Obesity



Subject Comparison



- In 10 conditions, the LLM and human created similar subject counts
- In 2 conditions, the human concept set had more subjects. Example: Angina - human inclusion of the code for concept of “chest pain”. The LLM argued that most people with chest pain do not have angina.
- In 3 conditions the LLM had more subjects than the human concept set. Example: Nausea: LLM included vomiting where it argued that 95% or greater of the subjects with vomiting also have nausea.

Conclusions

- In our use of Large Language models to adjudicate whether concepts belong in a concept set, we found many differences in those generated by the LLM compared to those generated by humans.
- We found that it is valuable to use the LLM to adjudicate concepts for concept sets as a starting point to help reduce the recommended concepts by PHOEBE and speed development.
- Further evaluation is required to determine whether the use of LLM constitutes a significant improvement in quality in the overall phenotype development and evaluation process.

Thank you!

Questions?

Come to Poster

604

Want some real excitement?!?

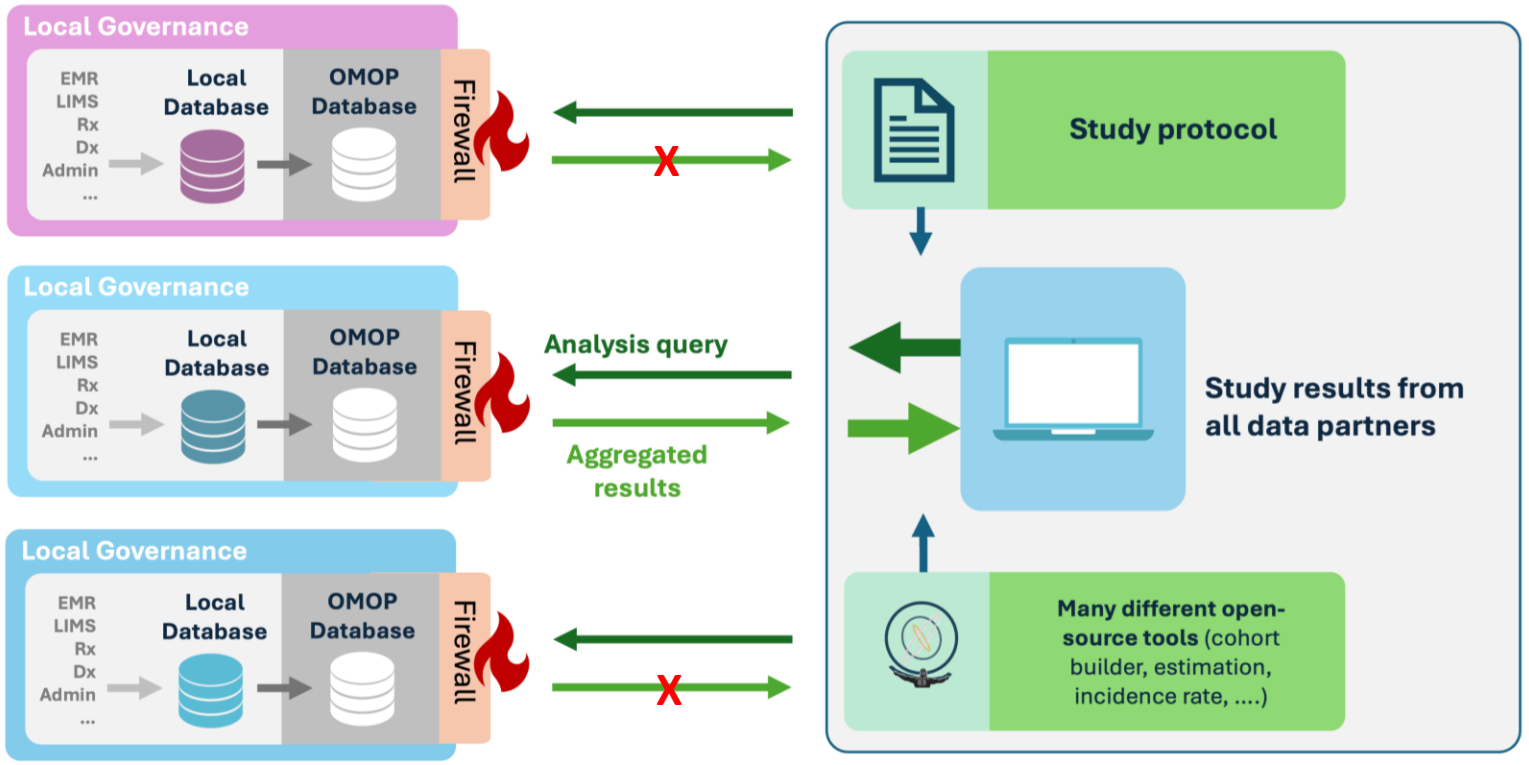
Come to the Phenotype Development Workshop Thursday 8am!

(actual participants)

Validating a Scalable Approach to Data Fitness-for-Purpose: Database Diagnostics Applied to LEGEND-T2DM

Clair Blacketer, Patrick B. Ryan, George Hripcsak, Marc A. Suchard, Fan Bu, Can Yin, Martijn J. Schuemie, Peter R. Rijnbeek

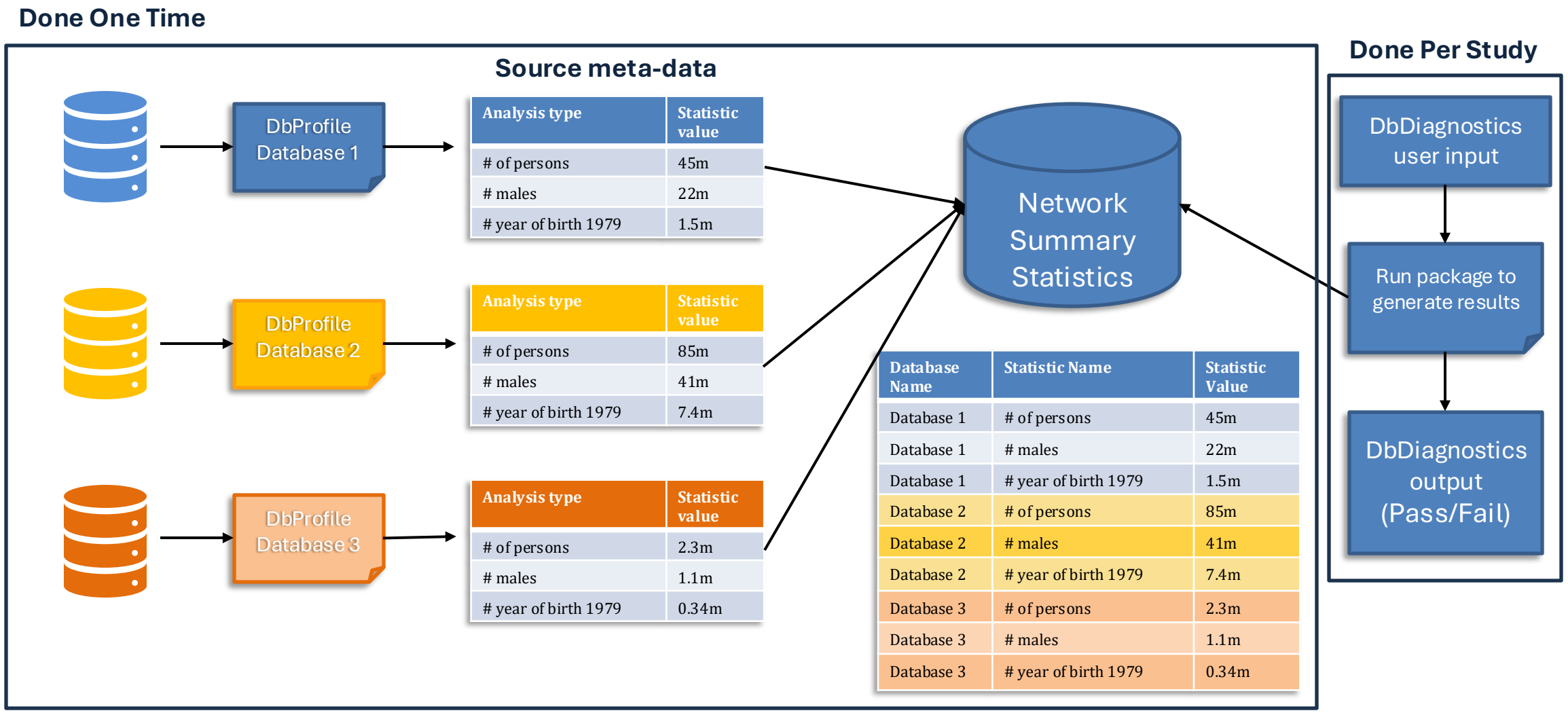
Federated Data Networks



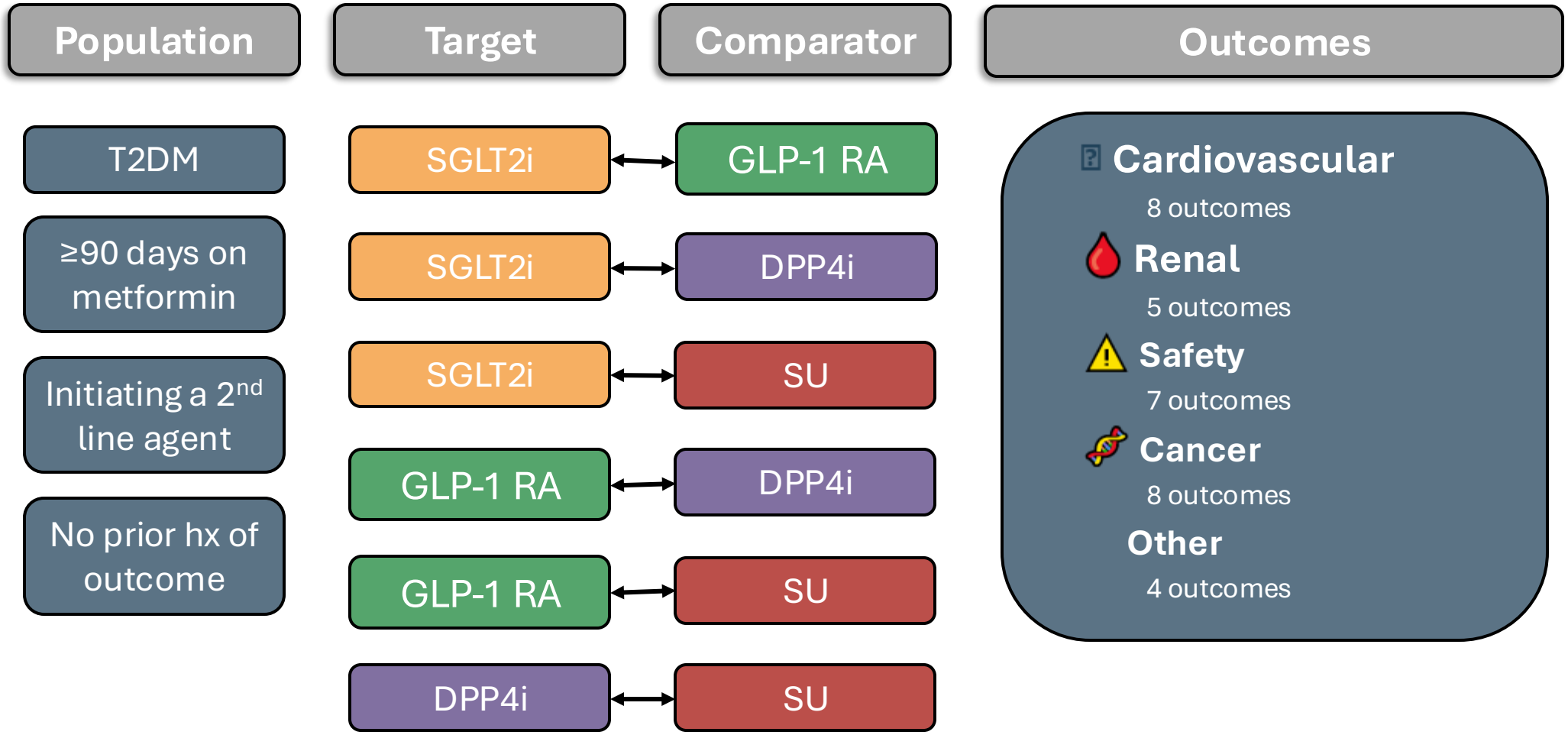
Federated data network processes including query and results sharing

- Federated networks are key to generating evidence at scale that engenders trust in the results
- Identifying the databases potentially fit to generate evidence given the study question remains a challenge

Database Diagnostics Process



Validation: LEGEND-T2DM Class v. Class Study Design



LEGEND-T2DM Study Diagnostics to Ensure Power

Eligible Target-Comparator Comparisons Across Data Sources

≥1000 patients per arm	Sufficient Sample Size
PS stratification to achieve SMD <0.15	Covariate Balance
Minimum detectable risk ratio <4	Proxy for Statistical Power
0.3< Preference score <0.7 in 25%	Empirical Equipoise
Negative control calibration	Residual Confounding
Kaplan-Meier plots	HR proportionality assumptions

LEGEND-T2DM x Database Diagnostics

Objective 1

Targets and Comparators

Can Database Diagnostics accurately identify databases with ≥1000 persons exposed?

Objective 2

Outcomes

Can Database Diagnostics accurately identify databases with representation of the outcomes of interest?

LEGEND-T2DM x Database Diagnostics Validation Methods

1

Translate Study Design into Database Diagnostics Settings objects

6 T➡C comparisons x 32 outcomes = 192 settings objects

2

Execute Database Diagnostics

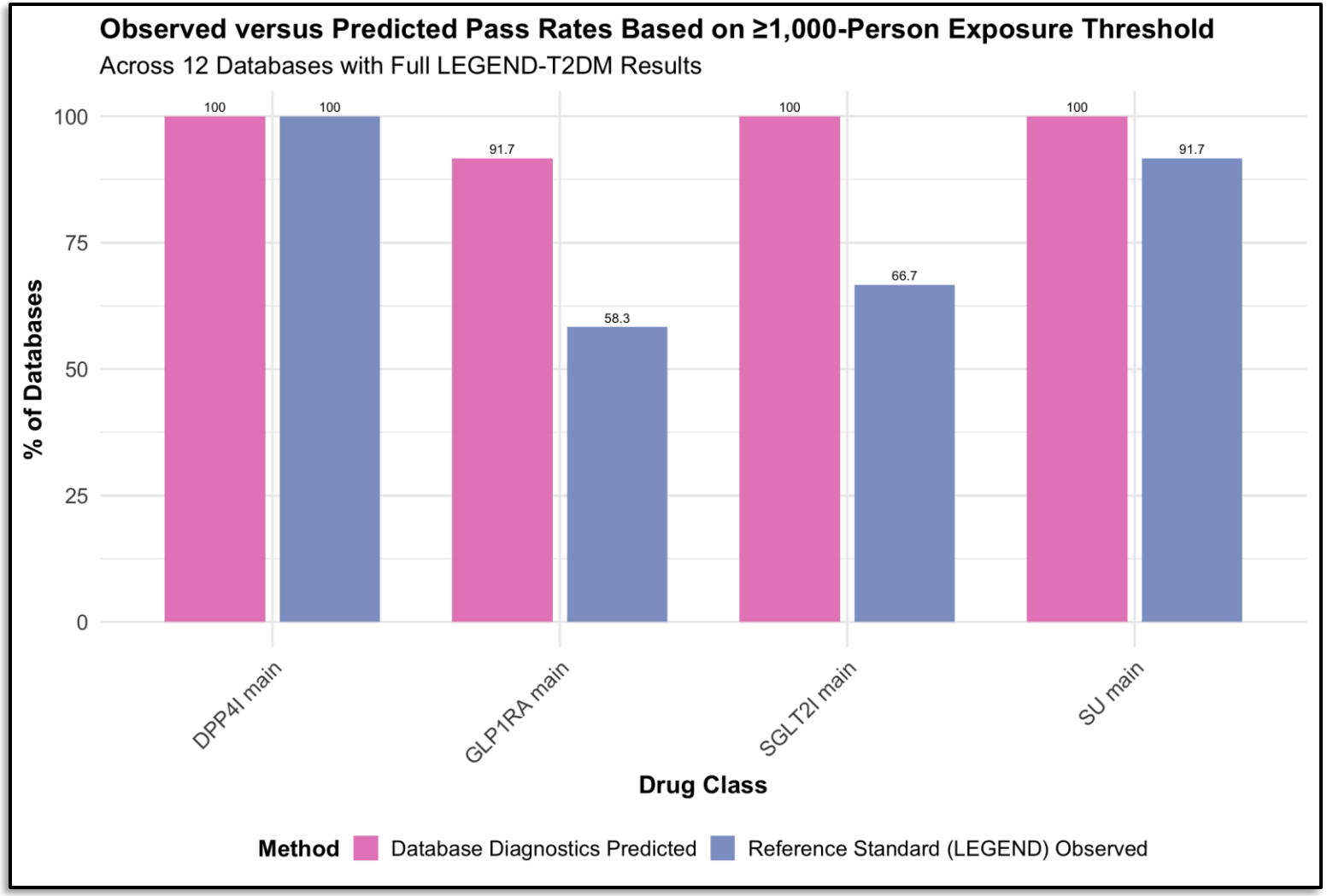
192 Settings x 12 Databases* = 2,304 evaluations

3

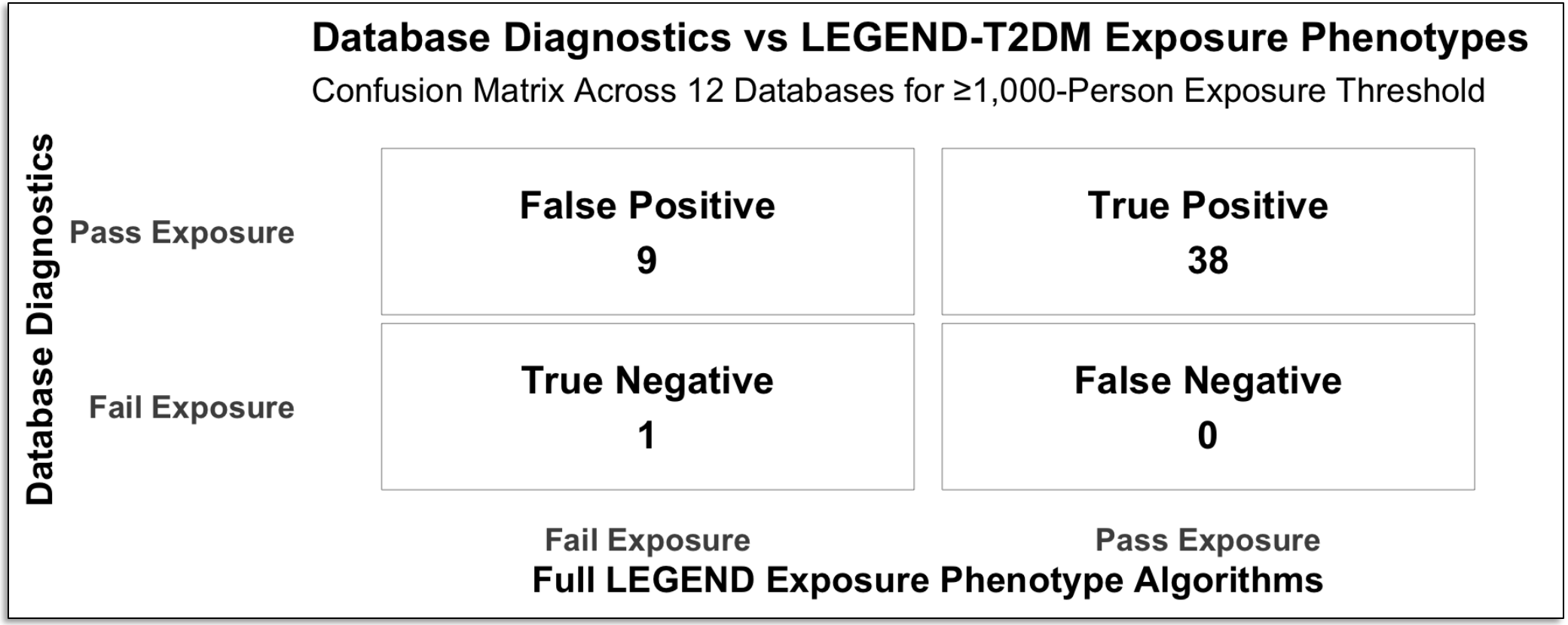
Compare Database Diagnostics results to LEGEND-T2DM results

**with both database profiles and LEGEND-T2DM results available*

Validation Results



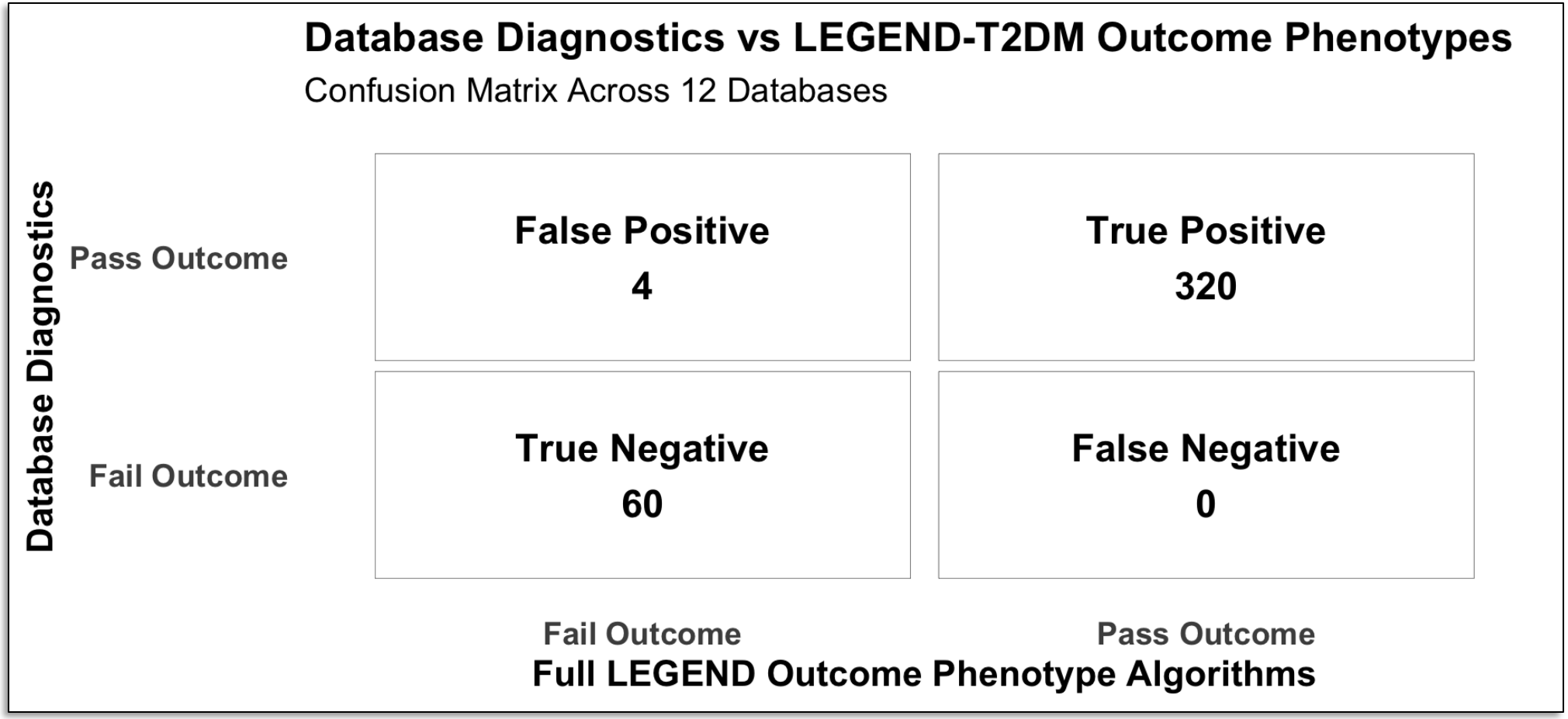
Validation Results



Sensitivity: 100.0%
Specificity: 10.0%

Positive predictive value (PPV): 80.8%
Negative predictive value (NPV): 100.0%

Validation Results



Sensitivity: 100.0%
Specificity: 93.75%

Positive predictive value (PPV): 98.7%
Negative predictive value (NPV): 100.0%

Key Takeaways

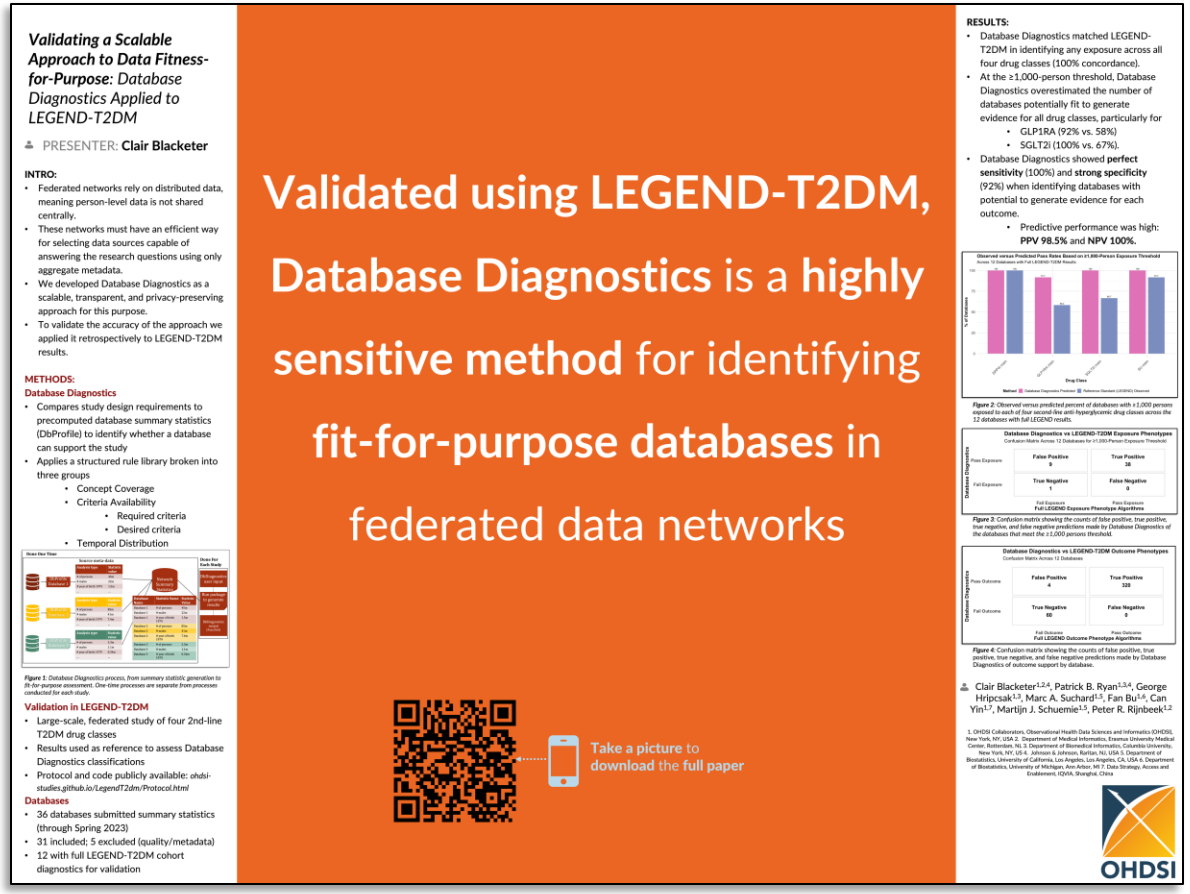
- Database Diagnostics showed **perfect sensitivity** (100%) and **strong specificity** (93.75%) when identifying databases with potential to generate evidence for each outcome, importantly revealing ***why* a database fails a given study design**
    - Predictive performance was high: **PPV 98.7%** and **NPV 100%.**
- Database Diagnostics **reliably identifies databases with any exposure of the targets and comparators**, but is less accurate at predicting which will meet the ≥1,000 person threshold when target/comparator definitions are complex

Meet me at the poster for more details!

#605

Blacketer@ohdsi.org

Evidencenetwork@ohdsi.org



Validating a Scalable Approach to Data Fitness-for-Purpose: *Database Diagnostics Applied to LEGEND-T2DM*

PRESENTER: **Clair Blacketer**

INTRO:

- Federated networks rely on distributed data, meaning person-level data is not shared centrally.
- These networks must have an efficient way for selecting data sources capable of answering the research questions using only aggregate metadata.
- We developed Database Diagnostics as a scalable, transparent, and privacy-preserving approach for this purpose.
- To validate the accuracy of the approach we applied it retrospectively to LEGEND-T2DM results.

METHODS:

Database Diagnostics

- Compares study design requirements to precomputed database summary statistics (DbProfile) to identify whether a database can support the study
- Applies a structured rule library broken into three groups
  - Concept Coverage
  - Criteria Availability
    - Required criteria
    - Desired criteria
  - Temporal Distribution

Validation in LEGEND-T2DM

- Large-scale, federated study of four 2nd-line T2DM drug classes
- Results used as reference to assess Database Diagnostics classifications
- Protocol and code publicly available: *ohdsi-studies.github.io/LegendT2dm/Protocol.html*

Databases

- 36 databases submitted summary statistics (through Spring 2023)
- 31 included; 5 excluded (quality/metadata)
- 12 with full LEGEND-T2DM cohort diagnostics for validation

Validated using LEGEND-T2DM, Database Diagnostics is a highly sensitive method for identifying fit-for-purpose databases in federated data networks

Take a picture to download the full paper

RESULTS:

- Database Diagnostics matched LEGEND-T2DM in identifying any exposure across all four drug classes (100% concordance).
- At the ≥1,000-person threshold, Database Diagnostics overestimated the number of databases potentially fit to generate evidence for all drug classes, particularly for
  - GLP1RA (92% vs. 58%)
  - SGLT2i (100% vs. 67%).
- Database Diagnostics showed **perfect sensitivity** (100%) and **strong specificity** (92%) when identifying databases with potential to generate evidence for each outcome.
  - Predictive performance was high: **PPV 98.5%** and **NPV 100%.**

OHDSI

Patrick B. Ryan, George Hripcsak, Marc A. Suchard, Fan Bu, Can Yin, Martijn J. Schuemie, Peter R. Rijnbeek