Phenotype Aphril – Week 4
The Phinale of Our 5th Phenotype Phebruary: Building a Strong Phoundation for Reliable Real-World Evidence

Azza Shoaibi, PhD

Director, Observational Health Data Analytics, Johnson and Johnson

Co-lead of the phenotype development and evaluation workgroup

2026 Phenotype Aphril

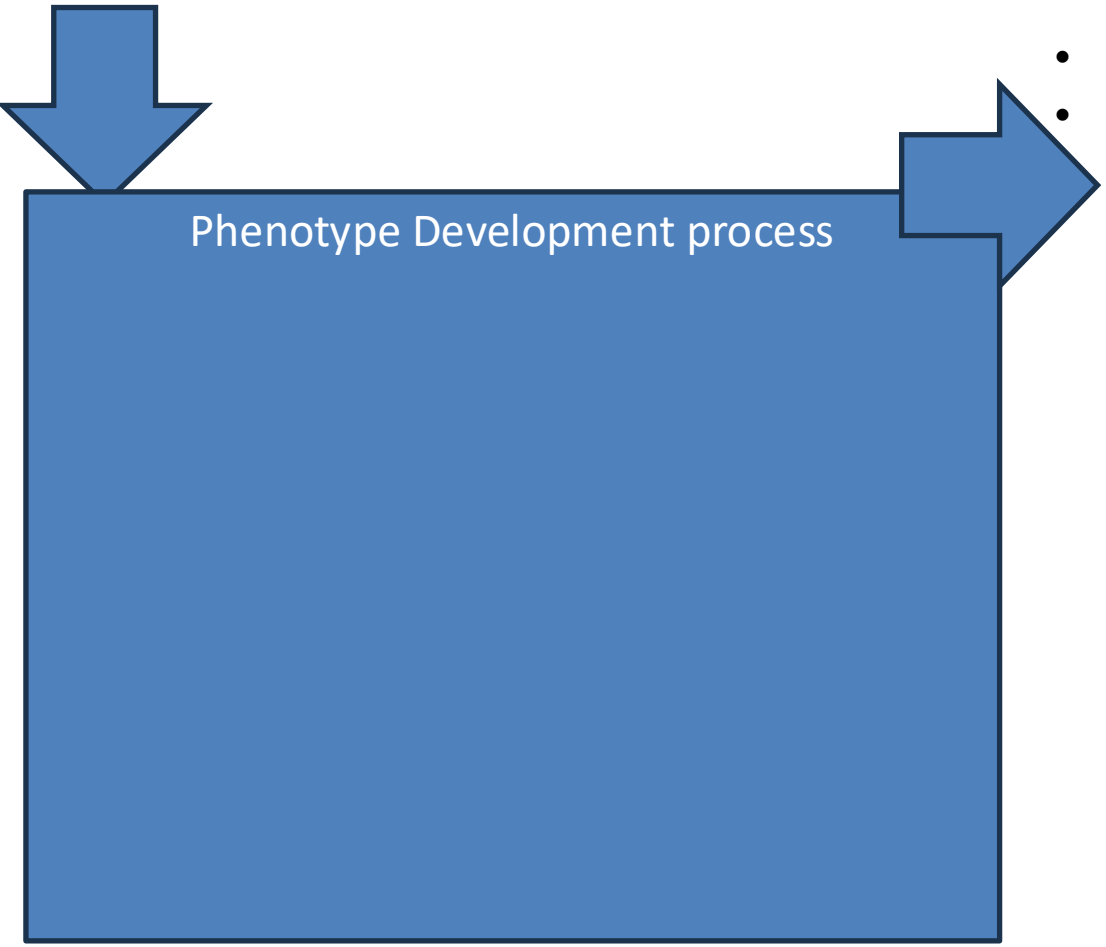
In partnership with the Generative AI and Foundational Models Workgroup

- Week 0: Phenotype Aphril kickoff
  - March 31 community call: Overview of phenotyping and opportunities for LLM
- Week 1: Phenotype development
  - April 7 community call: Demo – building cohorts in ATLAS
- Week 2: Phenotype evaluation
  - April 14 community call: Interactive – case adjudication using KEEPER
- Week 3: Iterative phenotype development
  - No community call: build your own AMI cohort definitions!
- Week 4: Iterative phenotype evaluation
  - April 28 community call: Review performance of AMI cohort definitions

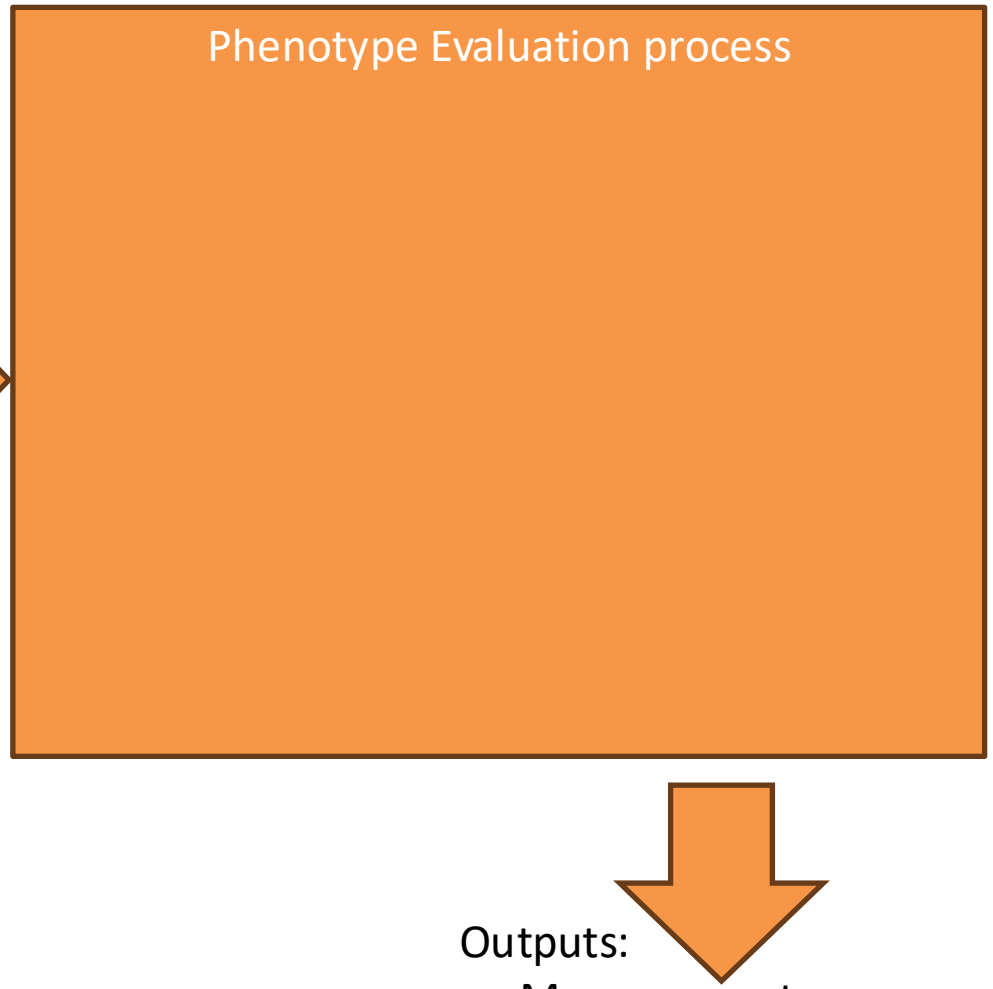
Phenotype process

Inputs:

- Clinical idea



- Cohort definition
- Database



Outputs:

- Measurement error quantification

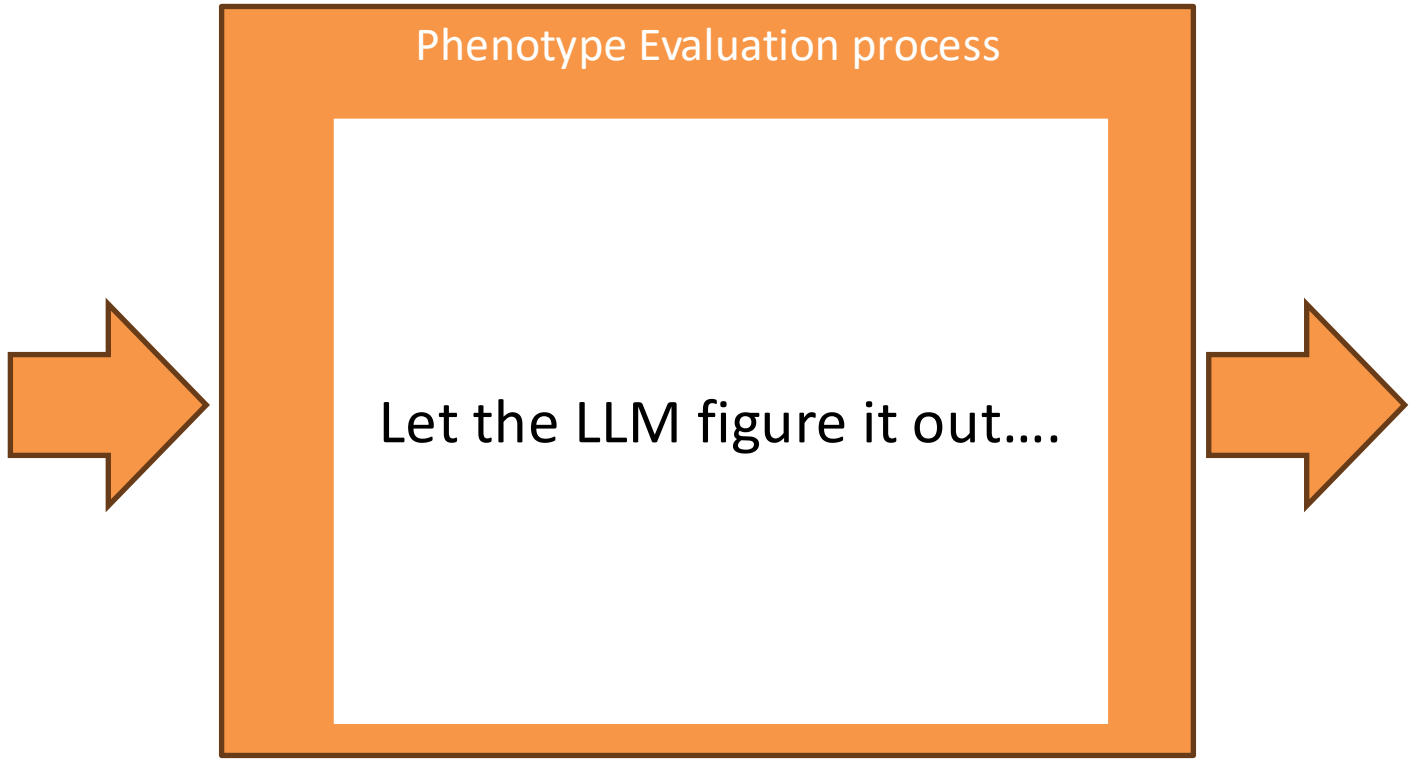
8 different MI definitions for evaluation

Cohort description	Source
Acute Myocardial infarction diagnosis or complications	Seed cohorts
Acute Myocardial infarction diagnosis	Seed cohorts
Acute Myocardial infarction diagnosis or complication, IP , primary	Seed cohorts
Acute Myocardial infarction diagnosis or complication, IP, 2dx	Seed cohorts
Acute myocardial infarction diagnosis, complications, diagnostic or therapeutic intervention, IP with no alternative diagnosis	Seed cohorts
All events of Myocardial infarction* , IP, ER	OHDSI Legend studies
Myocardial infarction diagnosis , IP	OHDSI covid
Myocardial infarction diagnosis , age >30, IP,ER, with treatment procedures or diagnostic procedures, another DX, no alternative diagnosis	AI submission

Phenotype Evaluation

Inputs:

- Clinical idea
- Cohort definition
- Database



Outputs:

- Measurement error quantification

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01433-4

Standardized patient profile review using large language models for case adjudication in observational research

Martijn J. Schuemie[1,2,3] ✉, Anna Ostropolets[1,4], Aleh Zhuk[1,5], Uladzislau Korsik[1,5], Seung In Seo[1,6], Marc A. Suchard[1,3], George Hripcsak[1,4] & Patrick B. Ryan[1,2,4]

Using administrative claims and electronic health records for observational studies is common but challenging due to data limitations. Researchers rely on phenotype algorithms, requiring labor-intensive chart reviews for validation. This study investigates whether case adjudication using the previously introduced Knowledge-Enhanced Electronic Profile Review (KEEPER) system with large language models (LLMs) is feasible and could serve as a viable alternative to manual chart review. The task involves adjudicating cases identified by a phenotype algorithm, with KEEPER extracting predefined findings such as symptoms, comorbidities, and treatments from structured data. LLMs then evaluate KEEPER outputs to determine whether a patient truly qualifies as a case. We tested four LLMs including GPT-4, hosted locally to ensure privacy. Using zero-shot prompting and iterative prompt optimization, we found LLM performance, across ten diseases, varied by prompt and model, with sensitivities from 78 to 98% and specificities from 48 to 98%, indicating promise for automating phenotype evaluation.

LLMs and humans show comparable performance (compared to human majority)

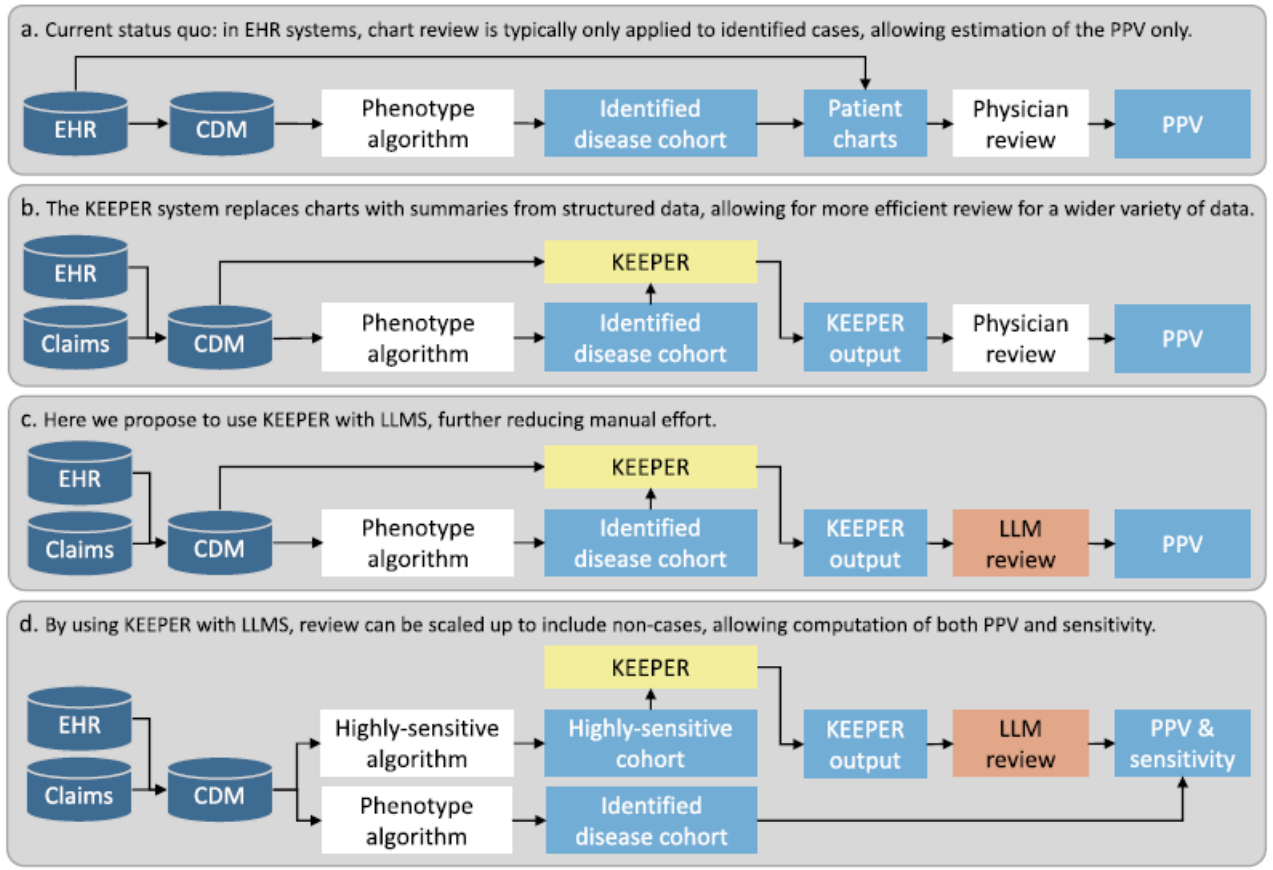
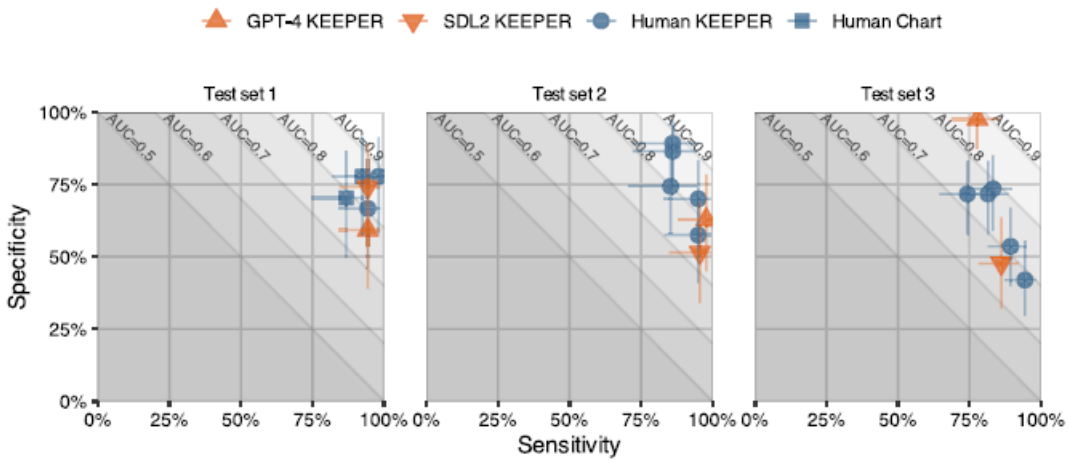
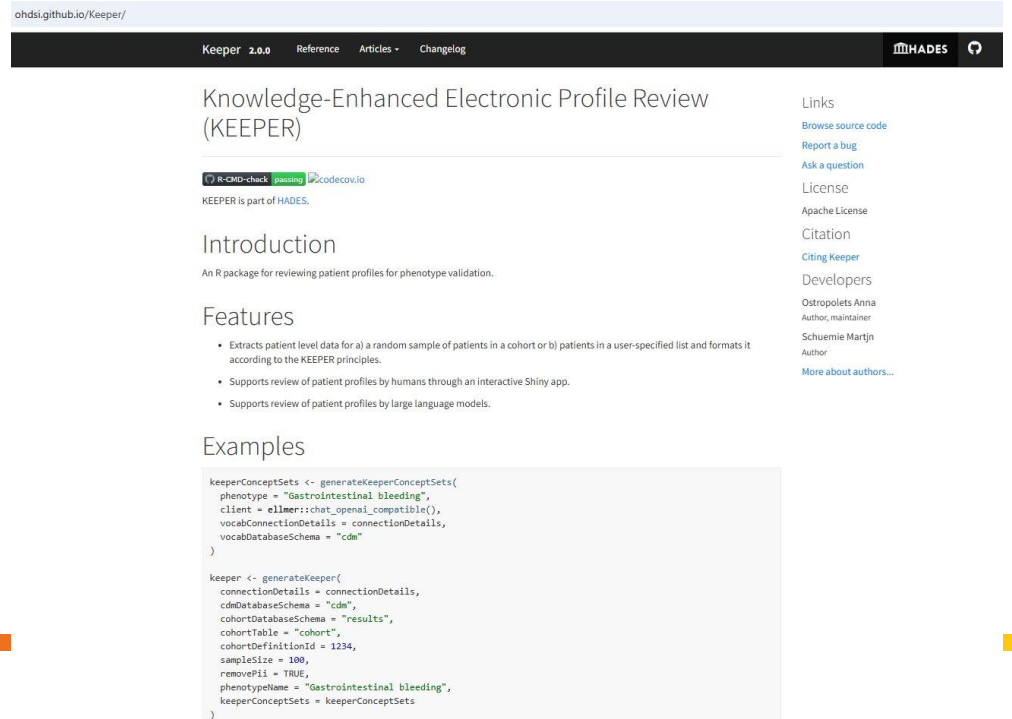


Fig. 3 | Sensitivity and specificity of reviewers for the three test sets. Points indicate sensitivity and specificity of each human or LLM reviewer against the gold standard. Error bars indicate 95% confidence intervals. For test set 1, the gold standard was created by external reviewers. For test set 2 and 3, the gold standard was the majority vote of human reviewers using a leave-one-out approach. Slanted lines denote iso-AUC contours, spaced 0.1 apart.



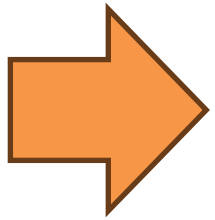
KEEPER-LLM

- Also use LLM to draft KEEPER input concept sets
- Added Shiny app for human review
- (Slightly) modify system prompt for newer LLMs:
  - Reasoning (no need for chain-of-thought prompting)
  - Enforcing structured output



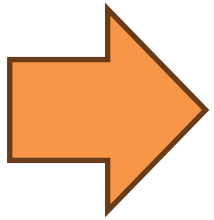
Phenotype Evaluation

Inputs:
- Clinical idea
- Cohort definition
- Database



Phenotype Evaluation process

KEEPER-LLM:
1. Create structured patient profiles (KEEPER) for persons in and out of cohort definition (week 2)
2. LLM adjudicates profiles to determine case/non-case status and confidence
3. Compare patients in cohort definition to estimate sensitivity, specificity, PPV, NPV



Outputs:
- Measurement error quantification

1. Generating Keeper profiles for review

1) Strat with clinical description for MI

2) Construct the inputs and draft Input Concept Sets

- **Diagnosis of interest:** 'Acute myocardial infarction'
- **Alternative diagnoses:** Aortic aneurysm, panic attack, esophageal reflux…

consider concepts related to both MI and its differential diagnoses

- **Symptoms**
- **Diagnostics (**Procedures, Measurements)
- **Therapeutic interventions (**Drugs: Procedures)
- **Complications**
- use the ellmer package to connect to an LLM from your provider of choice, including Anthropic, Google, OpenAI, or a local LLM

3) Keeper extracts data based on the input concept sets: If a concept belonging to an input concept set is found in a patient's records, Keeper will extract it along with its date relative to the index date

Adjudicating Person 4081

Knowledge-Enhanced Electronic Profile Review (KEEPER)

Profile | Timeline

Demographics

Age: 64
Sex: MALE
Observation period: day -156 - day 956
Race: White
Ethnicity: Not Hispanic or Latino

Presentation

Acute myocardial infarction (Claim, Secondary diagnosis)
Abnormal results of cardiovascular function studies (Claim, Secondary diagnosis)
Coronary arteriosclerosis (Claim, Primary diagnosis)
Essential hypertension (Claim, Secondary diagnosis)
Hyperlipidemia (Claim, Primary diagnosis)
Obesity (Claim, Secondary diagnosis)

Visits

Inpatient Visit (days 22 to 27)
Laboratory Visit - Independent Laboratory (day 30)
Outpatient Visit - Internal Medicine (day -5)
Outpatient Visit - Physician / Diagnostic Radiology (day 8)
Outpatient Visit - Thoracic Surgery (Cardiothoracic Vascular Surgery) Physician (day 0)
Pharmacy visit - General Surgery (day -2, day 16, day 20)
Pharmacy visit - Interventional Cardiology (day 4)
Pharmacy visit - Nephrology (day -17, day 13)

Symptoms

Dyspnea (day -5)

Phenotype:

Acute myocardial infarction

Decision

Decision
- Case
- Non-case

Certainty
- High
- Low

Correct index day

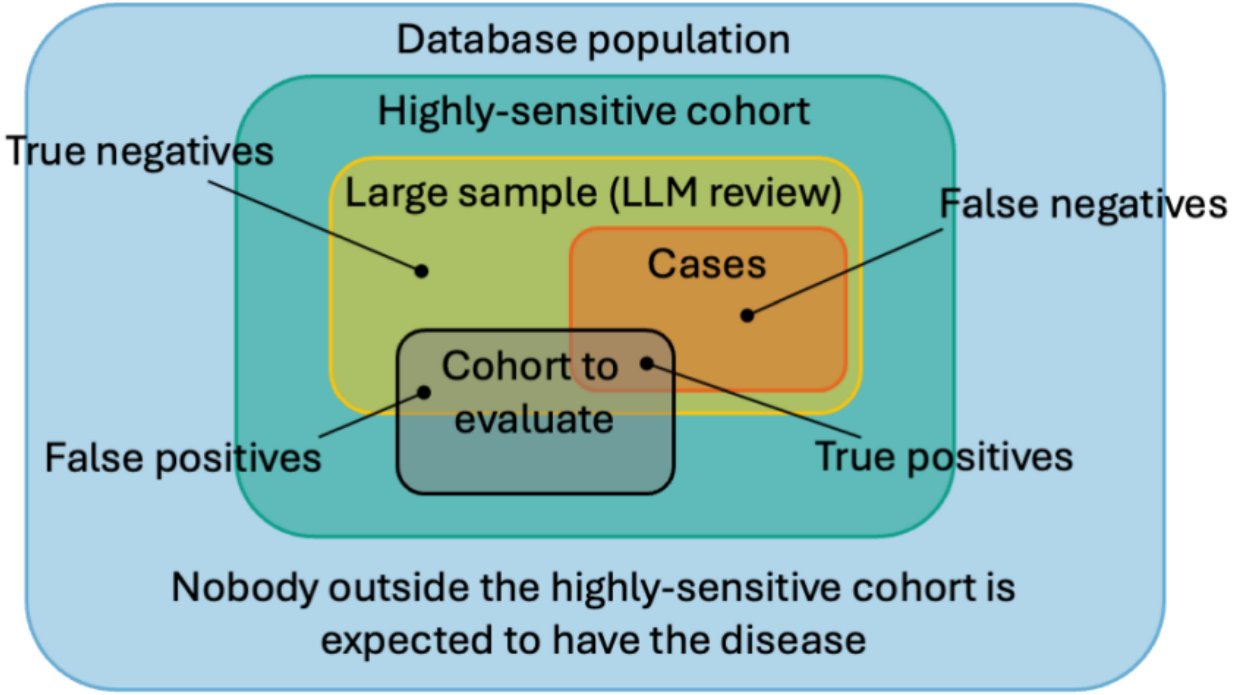
Patient selection

< 408 / 10000 >

Color legend

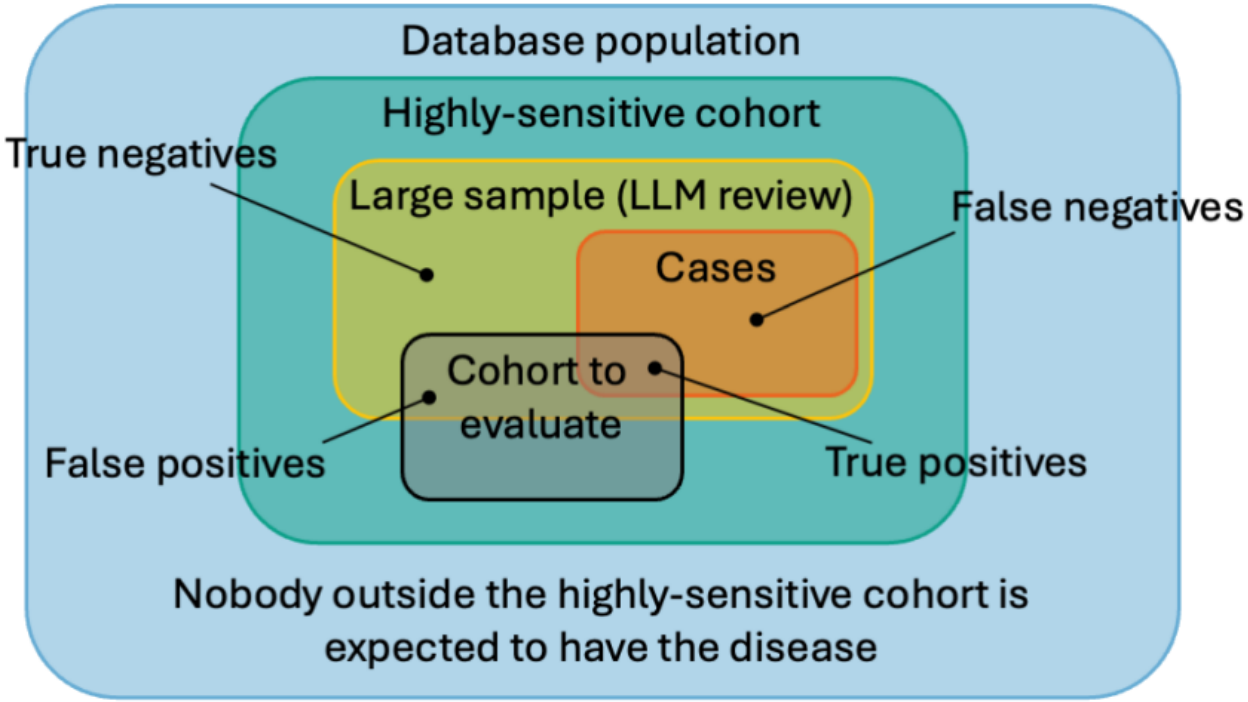
- Disease of interest
- Both
- Alternative diagnoses

2. Creating a reference set for MI (by adjudicating)



1) Build a highly-Sensitive Cohort. Anybody who has one concept/code of MI. OR Anybody who has at least two of ‘highly specific concepts’ of Keeper input concepts
2) Running Keeper on the highly-Sensitive to extract Keeper profiles on 10 k sample and use an LLM to adjudicate these cases- store the reference set for MI

3. Evaluating Target Cohorts

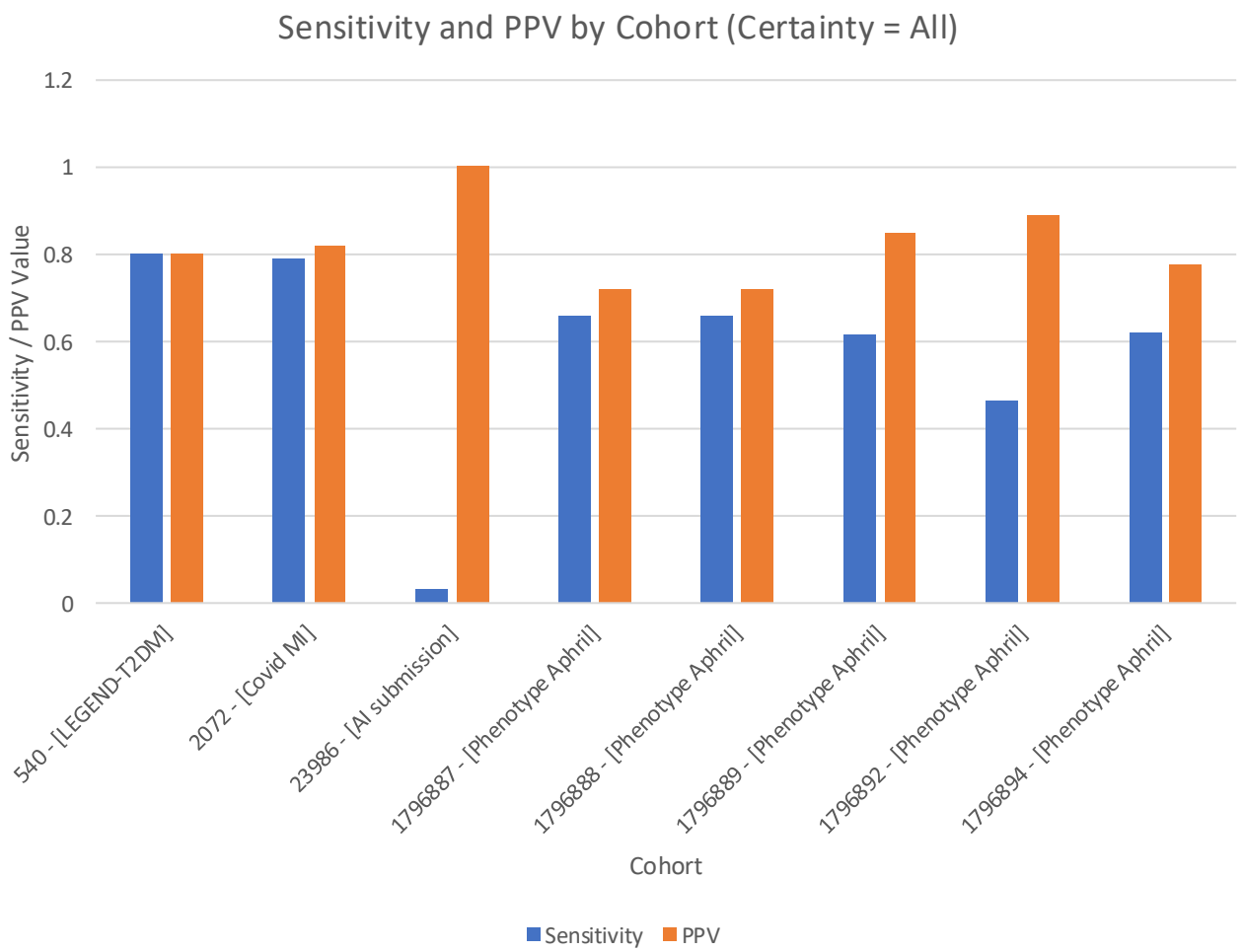


1) Generate the target cohort in the same database
2) Estimate measurement error

PPV= tp/(tp+fp)

Sensitivity= tp/(tp+fn)

Results



Cohort description	Cohort ID
Acute Myocardial infarction diagnosis or complications	1796888
Acute Myocardial infarction diagnosis	1796887
Acute Myocardial infarction diagnosis or complication, IP , primary	1796889
Acute Myocardial infarction diagnosis or complication, IP, 2dx	1796892
Acute myocardial infarction diagnosis, complications, diagnostic or therapeutic intervention in IP with no alternative diagnosis	1796894
All events of Myocardial infarction* , IP, ER	540
Myocardial infarction diagnosis , IP	
Myocardial infarction diagnosis , age >30, IP,ER, with treatment procedures or diagnostic procedures, another DX, no alternative diagnosis	23986

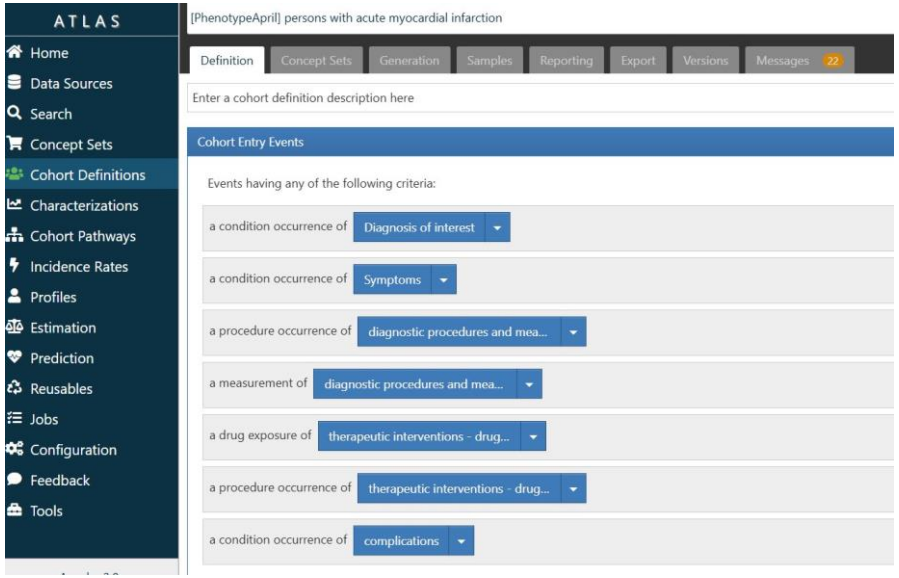
Few learnings around sources of errors

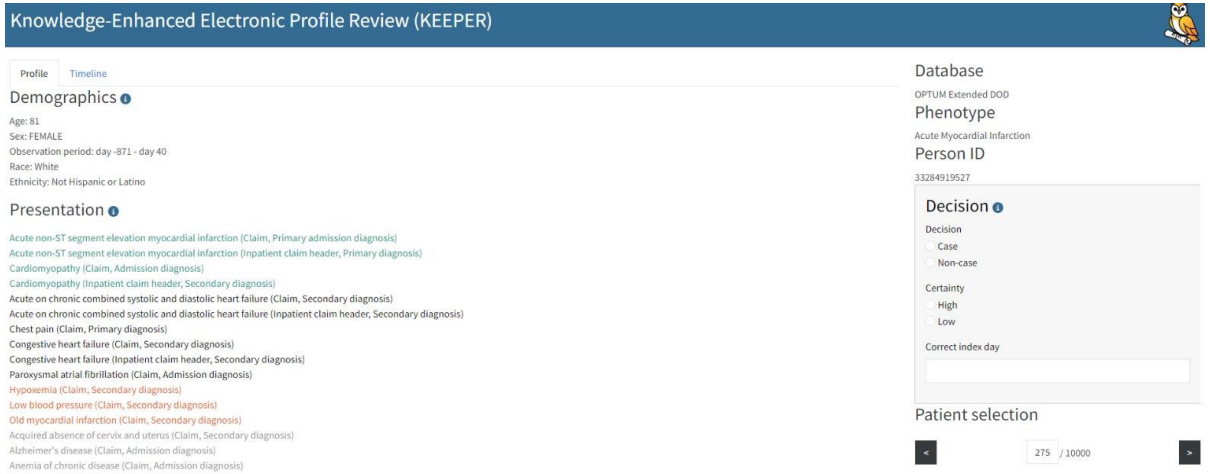
- **Old MI**: Distinct patterns differentiate acute incident MI from prevalent MI, extending beyond the presence of diagnosis codes or inpatient setting (e.g., carry-over codes).
- **Rule-out MI**: Some patterns distinguish rule-out MI from confirmed MI, reflecting differences in clinical course and management.
- **Can't reply on cardiac enzymes**: Key clinical discriminators (e.g., troponin elevation) are not consistently observed, even among cases with substantial supporting evidence for MI.
- **Was there a true injury versus ischemia?**: Uncertainty and misclassification may stem from ambiguity around true myocardial injury versus diagnostic suspicion of ischemia.
- **Multiple codes on the same day:** Multiple MI diagnoses and competing alternative diagnoses can co-occur within patient records.

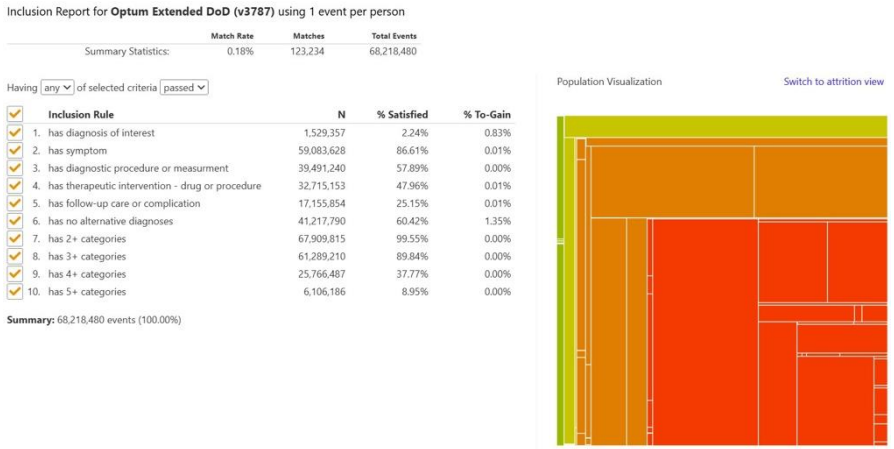
2026 Phenotype Aphril

Clinical elements associated with disease status

- **Diagnosis of interest:** 'Acute myocardial infarction'
- **Symptoms:** Chest pain, Difficulty breathing, Sweating, Nausea, …
- **Diagnostics:**
  - Procedures: Electrocardiography, angiocardiography, …
  - Measurements: Troponin, creatine kinase, …
- **Therapeutic interventions:**
  - Drugs: beta blocker, aspirin, heparin, clopidogrel, …
  - Procedures: percutaneous coronary intervention, coronary artery bypass grafting, insertion of cardiac pacemaker…
- **Complications:** heart failure, cerebral hemorrhage, acute kidney injury, acute pulmonary edema, hepatic failure…
- **Alternative diagnoses:** Aortic aneurysm, panic attack, esophageal reflux…







OHDSI AI-Driven Pherotype Development Workflow

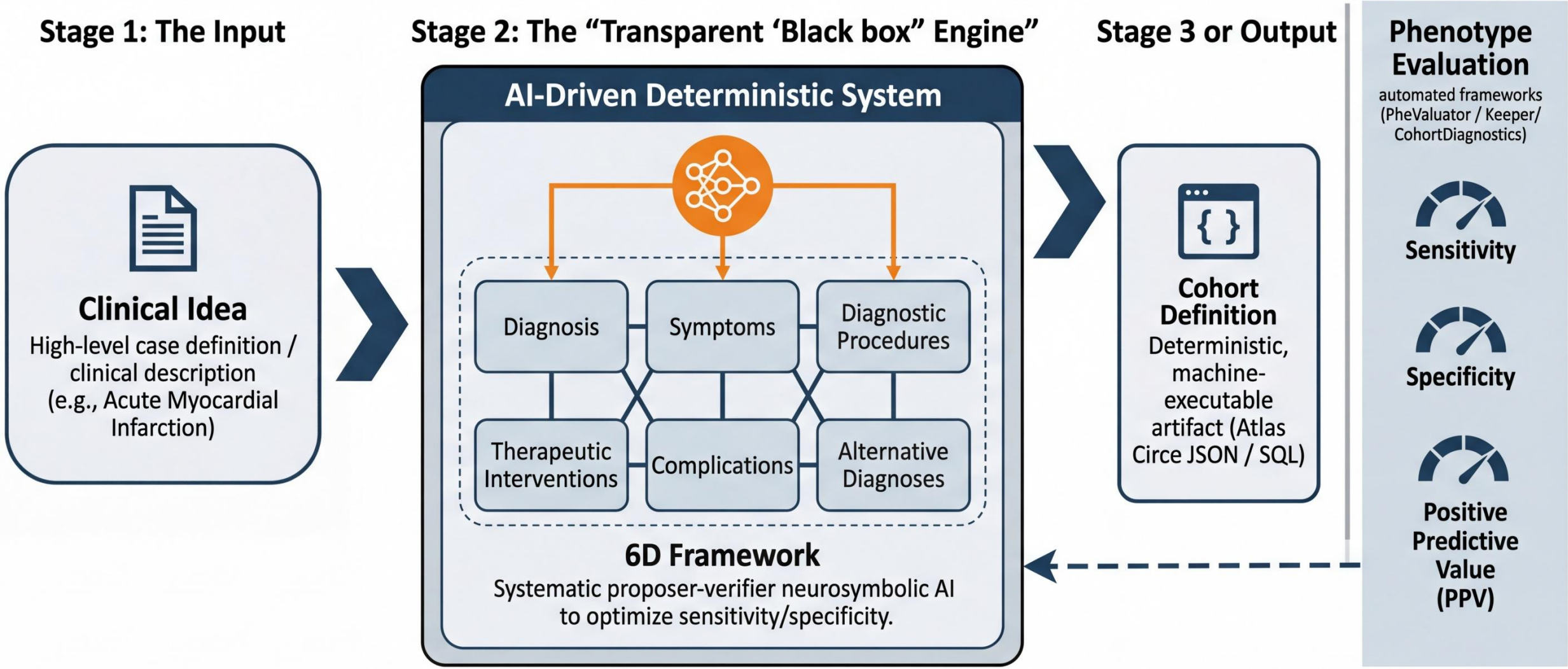
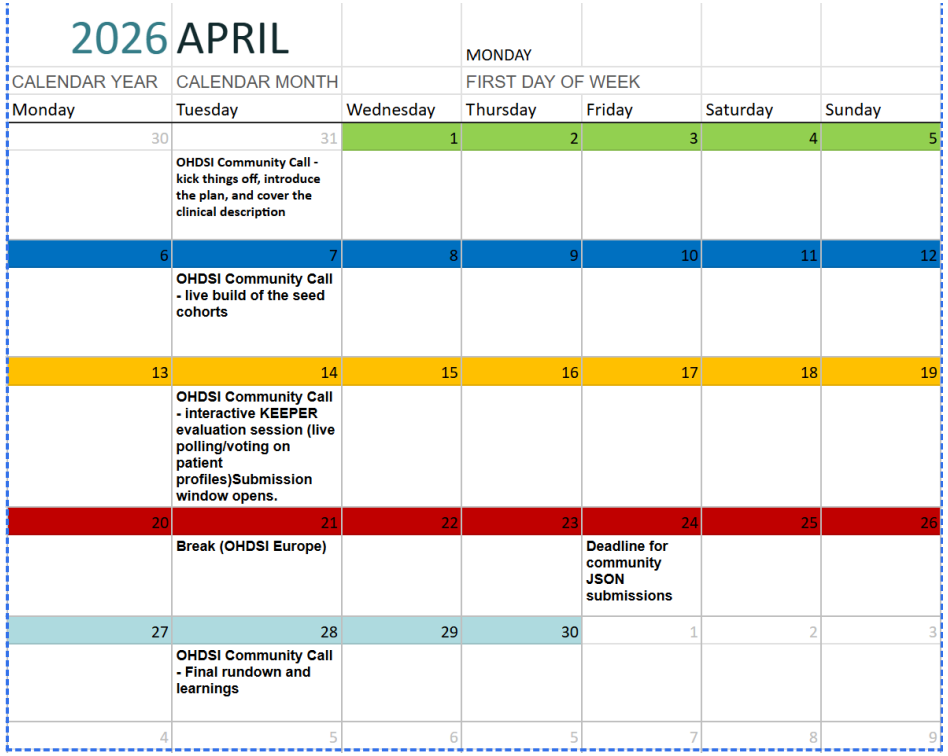


Figure 1: Systematic Workflow for Phenotype Development. This conceptual architecture illustrates the transition from manual, subjective heuristic coding to a standardized, automated pipeline. A raw clinical idea **(Input)** is processed by an **AI-driven system** that **deconstructs the phenotype intent** using a **universal six-dimensional framework. This engine generates a deterministic, machine-executable cohort definition (Output).** Crucially, the development phase is decoupled from evaluation, relying on automated probabilistic tools to objectively measure performance metrics (Sensitivity, Specificity, PPV) and ensure reproducibility.

What we achieved- and what we didn't (YET)



2026 APRIL — MONDAY

CALENDAR YEAR | CALENDAR MONTH | FIRST DAY OF WEEK

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
30	31	1	2	3	4	5
	OHDSI Community Call - kick things off, introduce the plan, and cover the clinical description					
6	7	8	9	10	11	12
	OHDSI Community Call - live build of the seed cohorts					
13	14	15	16	17	18	19
	OHDSI Community Call - interactive KEEPER evaluation session (live polling/voting on patient profiles)Submission window opens.					
20	21	22	23	24	25	26
	Break (OHDSI Europe)			Deadline for community JSON submissions		
27	28	29	30	1	2	3
	OHDSI Community Call - Final rundown and learnings					
4	5	6	5	7	8	9

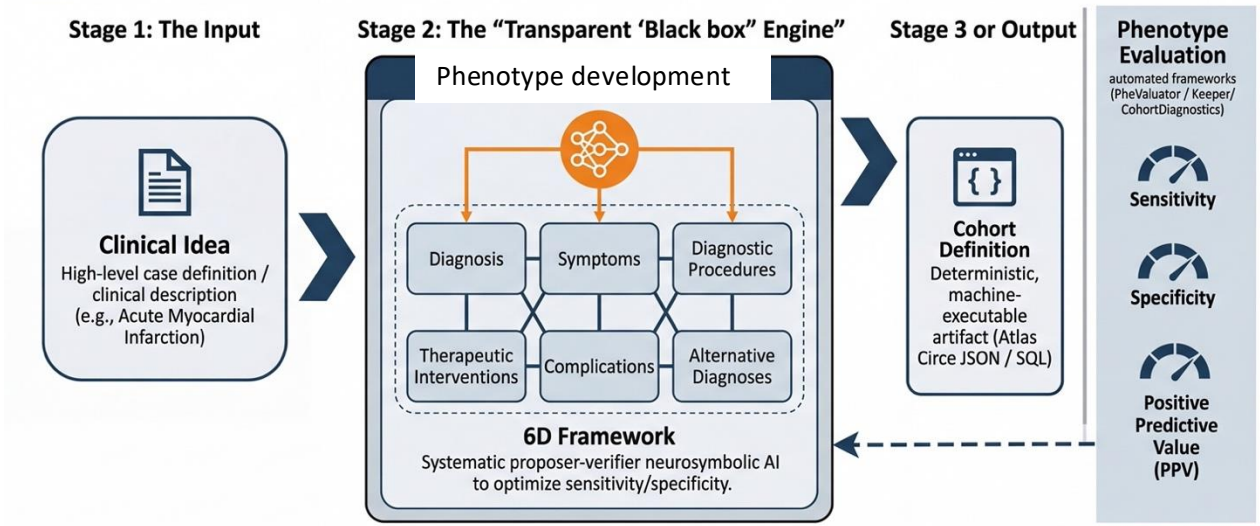
In partnership with Generative AI and Foundational Models Workgroup

We demonstrated a non-circular, structured evaluation approach (example) that is

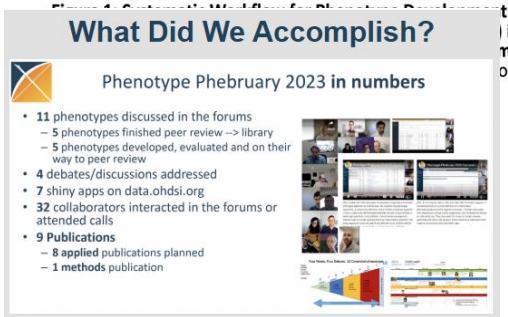
1. Anchored by prior clinical knowledge
2. Enriched with patient-level data, insights, and patterns
3. Utilized knowledge embedded in existing OHDSI assets (vocabulary structures)
4. Scalable through LLM models

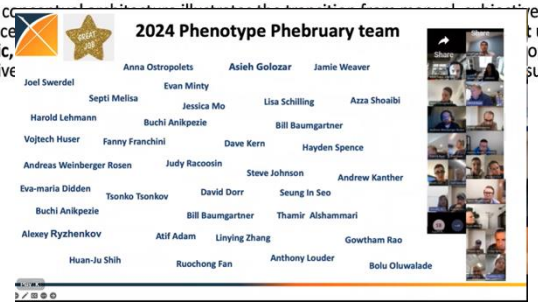
What we still need to work on:

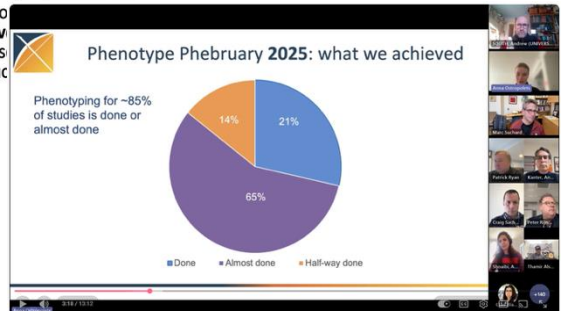
1. **Insufficient volume of community submission**
2. **Limited hands-on iteration with the full end-to-end process**

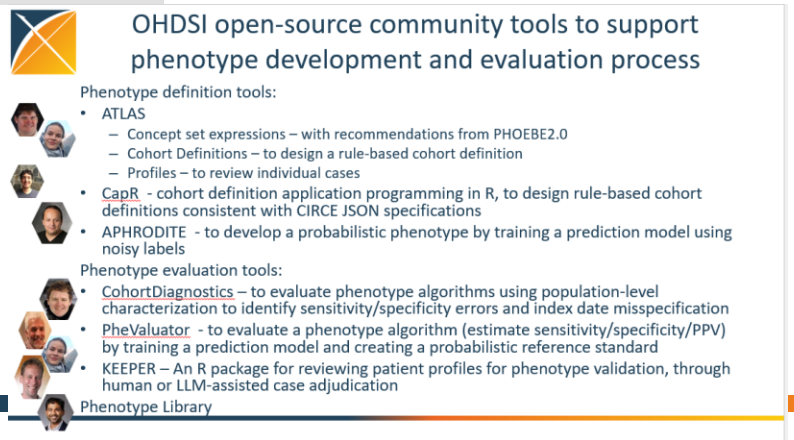

Stage 1: The Input
Clinical Idea
High-level case definition / clinical description (e.g., Acute Myocardial Infarction)
Stage 2: The "Transparent 'Black box" Engine"
Phenotype development
Diagnosis
Symptoms
Diagnostic Procedures
Therapeutic Interventions
Complications
Alternative Diagnoses
6D Framework
Systematic proposer-verifier neurosymbolic AI to optimize sensitivity/specificity.
Stage 3 or Output
Cohort Definition
Deterministic, machine-executable artifact (Atlas Circe JSON / SQL)
Phenotype Evaluation
automated frameworks (PheValuator / Keeper/ CohortDiagnostics)
Sensitivity
Specificity
Positive Predictive Value (PPV)


Phenotype Phebruary Daily Updates
OHDSI
Watch on YouTube


What Did We Accomplish?
Phenotype Phebruary 2023 in numbers
11 phenotypes discussed in the forums
5 phenotypes finished peer review --> library
5 phenotypes developed, evaluated and on their way to peer review
4 debates/discussions addressed
7 shiny apps on data.ohdsi.org
32 collaborators interacted in the forums or attended calls
9 Publications
8 applied publications planned
1 methods publication


2024 Phenotype Phebruary team
GREAT JOB
Anna Ostropolets
Asieh Golozar
Jamie Weaver
Joel Swerdel
Evan Minty
Septi Melisa
Jessica Mo
Lisa Schilling
Azza Shoaibi
Harold Lehmann
Buchi Anikpezie
Bill Baumgartner
Vojtech Huser
Fanny Franchini
Dave Kern
Hayden Spence
Andreas Weinberger Rosen
Judy Racoosin
Steve Johnson
Andrew Kanther
Eva-maria Didden
Tsonko Tsonkov
David Dorr
Seung In Seo
Buchi Anikpezie
Bill Baumgartner
Thamir Alshammari
Alexey Ryzhenkov
Atif Adam
Linying Zhang
Gowtham Rao
Huan-Ju Shih
Ruochong Fan
Anthony Louder
Bolu Oluwalade


Phenotype Phebruary 2025: what we achieved
Phenotyping for ~85% of studies is done or almost done
14%
21%
65%
Done
Almost done
Half-way done


OHDSI open-source community tools to support phenotype development and evaluation process
Phenotype definition tools:
ATLAS
Concept set expressions – with recommendations from PHOEBE2.0
Cohort Definitions – to design a rule-based cohort definition
Profiles – to review individual cases
CapR - cohort definition application programming in R, to design rule-based cohort definitions consistent with CIRCE JSON specifications
APHRODITE - to develop a probabilistic phenotype by training a prediction model using noisy labels
Phenotype evaluation tools:
CohortDiagnostics – to evaluate phenotype algorithms using population-level characterization to identify sensitivity/specificity errors and index date misspecification
PheValuator - to evaluate a phenotype algorithm (estimate sensitivity/specificity/PPV) by training a prediction model and creating a probabilistic reference standard
KEEPER – An R package for reviewing patient profiles for phenotype validation, through human or LLM-assisted case adjudication
Phenotype Library

This is just an example! We are on a long journey

- As black-box models and “advanced agents” proliferate, rigorous and systematic evaluation becomes essential.
- Meaningful evaluation requires well-defined reference sets to enable iteration, benchmarking, and continuous improvement.
- Establishing reference datasets and gold standards is inherently challenging and demands sustained, community-driven collaboration.
- Join OHDSI AI and the phenotype development and evaluation work groups to make a difference!!!

Lets discuss

- We told you about our fun with agents, tell us about yours ?!
- Any special learning from clinical workgroups ?
- Our journey into OHDSI phenotype reference set?