# The "Minds Meet Machines" (MMM) Challenge

## 1. Project Identification

1.1. Project Title: The "Minds Meet Machines" (MMM) Initiative: Evaluating Human and AI Workflows to inform future for OHDSI Concept Set Development.

1.2. Study Date: October 9, 2025

1.3. Sponsor: Observational Health Data Sciences and Informatics (OHDSI) Community.

1.4. IRB of Record: N/A.

## 2. Personnel

**2.1. Project Team:**

* Christopher Mecoli, MD, MHS
* Gowtham A Rao, MD, PhD
* Azza Shoaibi, PhD

**2.2. Scientific Advisors:**

* Patrick Ryan, PhD
* Martijn Schuemie, PhD

## 3. Background and Rationale

3.1. Background:

The development of accurate concept sets (medical code lists) is foundational for observational research that uses real world data. The current process is manual, time-intensive, and subject to variability. Generative Artificial Intelligence (GenAI) offers potential tools to improve the efficiency and consistency of this workflow.

The primary intent is to serve as a management tool to evaluate and improve the internal methodologies used by the OHDSI collaborative for phenotype development. The goal is to generate evidence to test and refine these workflows and assist the OHDSI community in determining if LLM’s are a viable path to pursue for future phenotype development at this time.

## 4. Aims and Objectives

**4.1. Primary Aim:**

To evaluate the performance (accuracy and completeness) of GenAI-driven approaches compared to rigorous, human-led workflows for translating standardized clinical descriptions into concept sets, in order to improve OHDSI phenotype development processes.

**4.2. Secondary Objectives:**

1. To quantify inter-human variability in the current workflow and the improvement gained through the reconciliation process (Consensus Gain).
2. To conduct a qualitative analysis of human reasoning and consensus-building processes to identify best practices and inefficiencies.

## 5. Methods

5.1. Design:

A multi-arm, blinded, comparative evaluation utilizing mixed methods (quantitative and qualitative analysis). The project utilizes a paired design, comparing the outputs of different methodologies (Human vs. AI) applied to the same standardized inputs. The evaluation involves three main phases: Concept Set Generation (Phase 1), Initial Comparative Analysis (Phase 2), and Adjudication (Phase 3).

5.2. Setting and Participants:

The evaluation will occur during a hybrid workshop at the OHDSI Symposium on October 9th, 2025 (New Brunswick, NJ, and virtual). Participants are adult professionals (researchers, clinicians, informaticians) affiliated with OHDSI.

5.3. Standardized Inputs and Materials:

All arms will receive identical, standardized, "fully specified" clinical descriptions in advance (publicly released October 5th 2025). The version of the OMOP vocabulary used for concept generation will be standardized across all arms to the August 27th, 2025 version.

**Table 1: Clinical Ideas for Phenotyping**

|  |  |  |
| --- | --- | --- |
| **id** | **name** | **clinical description** |
| C01 | Systemic Lupus Erythematosus | <https://github.com/ohdsi-studies/MindMeetsMachines/tree/main/C01> |
| C02 | Rheumatoid  arthritis | <https://github.com/ohdsi-studies/MindMeetsMachines/tree/main/C02> |
| C03 | Diabetic Macular Edema | <https://github.com/ohdsi-studies/MindMeetsMachines/tree/main/C03> |
| C04 | Deep  Venous Thrombosis | <https://github.com/ohdsi-studies/MindMeetsMachines/tree/main/C04> |
| C05 | Ovarian cancer | <https://github.com/ohdsi-studies/MindMeetsMachines/tree/main/C05> |
| C06 | Uveitis | <https://github.com/ohdsi-studies/MindMeetsMachines/tree/main/C06> |
| C07 | Systemic  sclerosis | <https://github.com/ohdsi-studies/MindMeetsMachines/tree/main/C07> |

**5.4. Arms and Procedures:**

* **Arm 1: Human Workflow (Current Practice)**
  + **Methodology:** "Split and Reconcile" Model (Modified Delphi). Independent creation by randomized sub-teams followed by a consensus reconciliation phase. Standard OHDSI tools (e.g., ATLAS) are used; GenAI is prohibited.
* **Arm 2: Generative AI Workflows**
  + **Methodology:** Autonomous generation of concept sets by multiple distinct GenAI pipelines (K≈2-5). Submissions are due by October 8th, 2025, 6:00 PM EST. "Human-in-the-loop" (HITL) intervention or post-editing of the AI output is strictly prohibited.

**5.5. Stratified Randomization (Human Arm):**

1. **Self-Identification:** Participants will complete a data collection form to self-identify their expertise based on the following rubric for two domains: (A) Clinical Expertise and (B) Informatics/Tooling Proficiency.
   * *Low/Novice:* Limited prior experience in the domain; requires guidance.
   * *Medium/Familiar:* Some experience in the domain; comfortable with standard tasks but may require assistance with complex scenarios.
   * *Expert:* Extensive experience and proficiency in the domain; capable of leading development and troubleshooting complex issues.
2. **Anonymization:** Participants will be assigned a unique numerical identifier (placeholder).
3. **Stratification and Assignment:** Stratified randomization will be used to assign participants to teams in Phase 1 (Generation). This ensures a balanced mix of expertise across teams. Participants are randomly assigned to a disease area so they may or may not have background in the disease area they are assigned to.

**5.6. Data Collection:**

1. **Concept Sets:** The primary concept lists generated by Arm 1 (reconciled) and Arm 2 (all pipelines). The required output format is the OMOP OHDSI Atlas Concept Set expression (JSON).
2. **Expertise Data:** De-identified survey data (from the data collection form) used for stratification.
3. **Qualitative Data:** All deliberations during the Human Arm's Independent Creation and Reconciliation phases will be audio-recorded to analyze reasoning patterns and process efficiencies.
4. **Prevalence Data:** A master file containing OHDSI network record counts (RecordCount) for standard concepts will be utilized for analysis and prioritization.

**5.7. Adjudication and Blinding (Phase 3):**

A True Gold Standard (TGS) will be established post-hoc.

* **Consolidation:** All generated concept sets (N arms) are extracted and resolved into a master list of unique concept IDs.
* **Automatic Gold Standard (AGS):** Concepts present in the intersection of all N submitted concept sets (total agreement) are automatically accepted into the TGS.
* **The Delta (Adjudication Pool):** Concepts where any disagreement exists (present in 1 to N-1 sets) are compiled for manual adjudication.
* **Prioritization:** The Delta is prioritized using a weighted ranking system to focus expert time. Prioritization is based on (1) maximal disagreement (Match Score closest to N/2) and (2) highest Concept Prevalence (RecordCount).
* **Blinding:** Clinical experts (Outcomes Assessors) will adjudicate the Delta. They are strictly blinded to the source (Human vs. AI) of each concept and the degree of consensus (Match Score). The review file will only contain concept details and prevalence.
* **Process:** The designated clinical expert makes the final determination (Yes/No only) on whether to include each disputed concept in the TGS. The other participants are allowed to provide input. A neutral volunteer ("Honest Broker") may be present to ensure the adjudicator remains strictly true to the original clinical description, preventing definition drift.
* **Conflict Mitigation:** Experts serving as adjudicators for a specific domain cannot participate in the concept set creation (Phase 1) or influence the creation of the concept set. They may offer technical assistance (e.g., user interface and study logistics) but not clinical assistance related to concept set building.

## 6. Measures and Analysis

**6.1. Primary Outcome Measure:**

Prevalence-Weighted F1 Score (F1W​) against the TGS. This metric weights the impact of each concept based on its prevalence (Pi​, Record Count) in the OHDSI data network.

* Calculate weighted counts: ; ;
* Calculate Weighted Precision () and Weighted Recall (): ;

**6.2. Secondary Outcome Measures:**

* Inter-Human Variability (F1 score between human sub-teams before reconciliation).
* Consensus Gain (Comparison of independent subgroup F1 vs. TGS against the reconciled human F1 vs. TGS).
* Unweighted F1 Score, Precision (Weighted/Unweighted), and Recall (Weighted/Unweighted).
* Thematic analysis of transcribed audio recordings.

**6.3. Analysis Plan:**

Descriptive statistics (mean, median, standard deviation, range, 95% Confidence Intervals) for the F1W​ and other metrics will be reported for the Human Arm and each distinct AI pipeline. The primary analysis will focus on descriptive comparisons. The difference in F1W​ scores between each AI arm and the Human control arm will be calculated and summarized for each clinical idea. Thematic analysis will be employed on audio transcripts to identify patterns in human reasoning and factors influencing consensus.

## 7. Ethical Considerations and Data Management

7.1. Ethics review considerations

This project is an internal Quality Improvement designed to aid the OHDSI Phenotyping workgroup in allocating future resources to process development. The OHDSI Phenotyping methodology has limited usefulness outside of OHDSI. No PHI will be utilized, and no risks to participants is anticipated. Based on the description put for the by the National Bioethics Advisory Commission (NBAC) in its December 19, 2000 draft document, the event organizers have made the determination that this project is a Quality Improvement project and does not constitute Human Subject Research under DHS guidelines.

If collected data is deemed of interest outside of the OHDSI community, IRB approval or acknowledgment of QI/NHSR status will be obtained prior to publication in any academic journal.

7.2. Informed consent to record and photograph participants:

Although this project is not Human Subject Research, the methodology involves recording of participants' conversations and public ally releasing the output of their work. Because of this, participation by OHDSI members will be voluntary, and consent to participate and talent release will be documented. All participants will also have provided explicit consent to photography as part of conference attendance.

7.3. Confidentiality and Data Management:

Participants are identified only by their assigned numerical placeholder. Survey responses regarding expertise will be fully de-identified post-randomization and reported only in aggregate. Concept set outputs are attributed to teams, not individuals. Audio recordings and transcripts will be stored securely; published reports will only include anonymized data.