Study Protocol: The "Minds Meet Machines" Challenge

Created Date: October 3, 2025 **Version:** 1.0.0

Section 1: Study Identification

1.1. Study Title:

- **Official Title:** A Multi-Arm, Blinded, Comparative Study Evaluating Human Expert versus Generative AI-Driven Workflows for Phenotype Concept Set Selection
- **Public Title:** The "Minds Meet Machines" (MMM) Challenge: Comparing Human and AI Performance in Selecting Medical Codes for Research.

1.2. Trial Acronym: MMM

1.3. Other Study ID Numbers:

- **Sponsor-Assigned ID:** OHDSI-GLOBAL-PhenotypeWorkgroup -20251009

1.4. Document Control

- **Version:** 1.0.0 (initial release; no modifications)
- **Created:** October 3, 2025
- **Maintainers:** Primary Investigators (PIs): Gowtham A Rao, Azza A Shoaibi

Section 2: Sponsor & Collaborators

2.1. Responsible Party: Azza A Shoaibi, Gowtham A Rao, Observational Health Data Sciences and Informatics (OHDSI)

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

2.3. Primary Investigators (PIs):

- Azza Shoaibi PhD (shoaibi.azza@gmail.com)
- Gowtham A Rao MD, PhD (rao@ohdsi.org)

2.5. Scientific advisors:

- Patrick Ryan PhD (ryan@ohdsi.org)
- Martijn Schuemie PhD (schuemie@ohdsi.org)
- Jack Murphy PhD (jackmurphy2351@gmail.com)

2.6. Honest Brokers (independent; not members of the PIs' team)

- Gaurav **Dravida** (dravida@ohdsi.org)
- Craig Sachson (sachson@ohdsi.org)
- Elisse Katzman (katzman@ohdsi.org)
- On site volunteers

Role Boundary: Honest brokers conduct **all human-facing activities** (announcements, inquiries, day-of check-in/enrollment, information notice distribution, grouping, logistics, custody of any optional signed forms, assignment of random IDs, export scrubbing, and delivery of de-identified datasets). The **PIs may support brokers operationally** (e.g., venue coordination, infrastructure readiness, materials preparation) and may engage the broader audience **only in non–individual participant-facing group settings** (e.g., general plenary remarks **after** human-facing activities conclude). PIs **do not** recruit, screen, enroll, or interact with participants individually.

Section 3: Study Description

3.1. Brief Summary (Lay Language): Observational health research requires identifying patients with specific conditions using lists of medical codes (concept sets). Creating these lists is traditionally slow and manual. This study, the "Minds Meet Machines" challenge, is a methodological evaluation comparing the accuracy and completeness of new Generative Artificial Intelligence (GenAI) tools against rigorous, human-led workflows for this task. The evaluation analyzes **only de-identified outputs**; the PIs have **no interaction** with participants.

3.2. Detailed Description (Rationale and Objectives): The MMM challenge is designed as a controlled evaluation using **de-identified outputs** to compare GenAI workflows to human-led workflows.

- **Primary Objective:** Evaluate performance (accuracy and completeness) of GenAI approaches vs. a rigorous, consensus-based human workflow for translating standardized clinical descriptions into phenotype concept sets.

- **Secondary Objectives:**

    - Quantify inter-human variability pre-reconciliation (based on de-identified outputs).
    - Analyze clinical impact with prevalence-weighted metrics.
    - Conduct a secondary analysis using selected source vocabularies (e.g., ICD-10-CM); primary analysis uses OHDSI standard concepts (e.g., SNOMED).

3.3. Study Phase: N/A (Methodological Study)

3.4. Study Type: Methodological evaluation of **de-identified outputs** (comparative design evaluating outputs from human-led vs. computer-led methodologies). The PIs analyze **de-identified outputs** provided by independent honest brokers (**sachson@ohdsi.org**, **dravida@ohdsi.org**). The PIs do **not** interact with participants and cannot readily ascertain identity. This qualifies as **Not Human Subjects Research (NHSR)**; we will request a formal

NHSR determination.

3.5. Study Purpose: Health Services Research / Methodological Evaluation.

Section 4: Study Design

4.1. Primary Purpose: Methodological Evaluation

4.2. Study Model: Parallel Assignment (multi-arm comparative design; quantitative analysis). The evaluation compares outputs of different methodologies applied to identical standardized inputs across 5–6 disease domains from several candidate domains that will be released in advance to the public.

4.3. Masking/Blinding: Single blind (Outcomes Assessor). Adjudicators (PIs' evaluation team) are strictly blinded to the source (Human vs. AI arm) of each concept. Honest brokers prepare blinded adjudication inputs with **only concept-level attributes** (e.g., name, vocabulary, hierarchy, prevalence). No contributor/team metadata is provided to adjudicators. Brokers compile/deliver the adjudication bundle (or host a read-only app/spreadsheet). Adjudicators never receive contributor identities or underlying source files.

4.4. Allocation: Self-selection or broker-managed grouping. Participants self-organize or are grouped solely by brokers using non-identifiable inputs. PIs neither collect characteristics nor assign/randomize participants.

4.5. Arms and Methodologies:

Arm	Arm Description	Methodology	Methodology Description
Arm 1: Human Workflow (Control/Benchmark)	"Split and Reconcile" model. Human-led process facilitated by the brokers. Brokers collect final de-identified concept sets and non-identifiable metrics; they transmit blinded outputs to the PIs. The PIs receive no identifiers and no re-identification key.	Participants use standard OHDSI tools (ATLAS, PHOEBE, ATHENA) provided by the brokers. External LLM/GenAI use by the human arm is prohibited.	Independent creation by two voluntary sub-teams followed by a broker-facilitated consensus reconciliation phase.
Arm 2: Generative AI Workflows (alternative)	Distinct GenAI pipelines (including anonymous). Outputs are collected and blinded by brokers.	Fully automated GenAI pipeline(s)	Autonomous generation of concept sets using the specified pipelines. **No** HITL or post-edit. **Single-prompt** only (sequential prompting

Arm	Arm Description	Methodology	Methodology Description
			prohibited).

Standardized Inputs: Identical inputs are disseminated **only by the brokers** via OHDSI forum posts and/or broker email broadcasts. PIs do **not** contact participants from study start date till study end date. Inputs consist of a "fully specified" clinical description (clinical context, presentation, inclusion/exclusion criteria, research utility; see appendix). A clinical expert PI may review input content; **brokers** disseminate it.

Section 5: Outcome Measures

Focus: **accuracy** and **completeness** of each arm's concept list (structure can be optimized post-hoc).

5.1. Primary Outcome Measure:

Title	Description	Time Frame
Prevalence-Weighted F1 Score	Harmonic mean of precision/recall vs. the True Gold Standard (TGS); disagreements weighted by concept prevalence across OHDSI network. Analysis uses OHDSI standard concepts (e.g., SNOMED). TGS is established post-hoc via blinded adjudication (Appendix C.1).	Post (2–4 weeks post-workshop)

5.2. Secondary Outcome Measures:

Title	Description	Time Frame
Unweighted F1 Score	Standard F1 vs. TGS with equal weighting.	Post
Inter-Human Variability	F1 between two human sub-teams before reconciliation (de-identified outputs).	Post
Human–GenAI Variability	F1 between reconciled human list and each GenAI list before adjudication.	Post
Source Vocabulary F1	F1 based on selected source	Post

Title	Description	Time Frame
	codes (e.g., ICD-10-CM) via OHDSI mappings to resolve TGS and generated lists into target source codes.	

Section 6: Eligibility Criteria (Human Participants)

*Note: Screening/onboarding are conducted **solely** by honest brokers. PIs do not interact with participants.*

6.1. Sex/Gender: All

6.2. Age Limits: 18+ (adult professionals)

6.3. Accepts Healthy Volunteers: Yes (participants are research professionals/experts; no health status assessment).

6.4. Inclusion Criteria (and Day-of Enrollment Workflow):

- **Door Greeting & Verbal Agreement: Gaurav Dravida** (or designated broker) greets participants at the door, provides the **Information Notice** (Appendix J), and obtains **verbal agreement** to participate.
- **Optional Written Form (Broker Custody):** Brokers may provide an optional written form. Participants **may** add identifying info and **may** sign; **forms go into a broker-only box**. PIs never see or access these forms.
- **Random ID Assignment (Broker-Only):** Brokers **draw a random number** and hand it to each participant as their **study ID**. Participants keep this ID and **must use it** in ATLAS and the adjudication/evaluation interface. PIs do **not** know IDs and receive no linkage.
- **No Linkage Between Paper & ID:** There is **no link** between any identifying info on the optional signed paper (in the box) and the randomly assigned ID.
- **Replacement IDs:** If a participant **loses/misplaces** an ID **prior to the experiment start**, brokers may issue a **replacement ID**. After start, no replacements are issued.
- **Pairing Guidance (Split Model):** Brokers may encourage voluntary pairing of a more experienced participant with a less experienced one. Pairing is by self-perception and voluntary.
- **Adults (≥18 years).**

6.5. Exclusion Criteria:

- Inability/unwillingness to proceed after reviewing the Information Notice.
- Children, prisoners, other vulnerable populations (as screened by brokers).

- Conflict of Interest: Clinical experts designated as adjudicators (PIs' evaluation team) for a domain (Phase 2) are **prohibited** from participating in concept set creation (Phase 1) for the same domain.

6.6. Online Participants (MS Teams)

- **Broker-Hosted Only:** Online participation occurs in a **broker-hosted MS Teams session**.
- **No Screen Sharing by Participants:** Participants **do not** share screens. This prevents inadvertent disclosure of identities or non-research information.
- **No PI Contact:** PIs do **not** interact with online participants; any questions are handled by brokers via chat/voice.
- **Same Rules Apply:** Brokers assign/randomize IDs to online participants and deliver inputs; PIs remain blinded.

Section 7: Recruitment & Location Information

7.1. Announcements, Recruitment & Day-of Enrollment Strategy: Public announcements are issued **in advance** by the honest brokers (e.g., OHDSI forum posts and broker emails) to raise awareness. **All enrollment/check-in occurs on the workshop day—October 9, 2025—and is conducted solely by the brokers.** Any pre-event inquiries or sign-ups (if used) are broker-handled and **not** shared with the PIs. **At no time** do PIs recruit, reply to participant inquiries, screen, or enroll participants during the study period.

7.2. Study Start Date (Workshop Date): October 9, 2025

7.3. Primary Completion Date: October 31, 2025 (estimated end of adjudication)

7.4. Study Completion Date: Data collection (broker-managed) involves only concept sets via OHDSI ATLAS. Brokers ensure no personal identifiers are captured/linked to outputs provided to the PIs. Data collection completes **October 9, 2025**.

7.5. Locations:

- Facility: Hyatt Regency, 2 Albany Street, New Brunswick, NJ, 08901
- **Participant-Facing Supervision:** Brokers supervise all participant spaces. PIs will **not** enter participant rooms/sessions during participant-facing activity.
- **Contact (Broker Only):** dravida@ohdsi.org; other on-site volunteers.
- **Recruitment (Broker Only):** dravida@ohdsi.org; sachson@ohdsi.org; other on-site volunteers.

Section 8: Oversight & Administrative Information

8.1. IRB/Ethics Committee Approval: PIs analyze **non-identifiable outputs** provided by **independent honest brokers** (see Appendix H). **All** human-facing procedures (announcements, inquiries, check-in/enrollment, information notice distribution, random ID assignment, logistics, data collection, de-identification) are broker-managed. PIs may support

brokers operationally (e.g., infrastructure) and may give **non–participant-facing group remarks after** human-facing activities conclude. PIs will **not** be present in participant-facing sessions. PIs will have **no access** to identifiers or any linking key. **No re-identification** will be attempted.

8.2. Human Subjects Protection Review Board Status: Pending NHSR Determination.

8.3. Information Notice: No informed consent is obtained by the PIs. Brokers provide the Information Notice (Appendix J). Proceeding indicates agreement. Optional signatures, if collected, are **retained solely by brokers** and are **not shared** with PIs.

8.4. Data Monitoring Committee (DMC): Not required; methodological minimal-risk; no health or sensitive data shared.

8.5. Plan to Share IPD (Individual Participant Data): PIs receive/analyze only **de-identified** data provided by brokers.

- **Allowed:** de-identified concept sets, blinded arm labels, aggregate non-identifiable metrics (e.g., timing/accuracy).
- **Prohibited:** names, emails, employer/organization, job titles, IP addresses, user logs, and **any open-ended free-text fields**. Brokers scrub ATLAS exports and related files of any user-identifying metadata before transmission.

Concept lists are not attributed to individuals. Brokers submit lists via ATLAS, assign sequential IDs, and provide only these sequential IDs plus de-identified lists to the PIs. No linkage to personal identifiers is possible. Aggregate, de-identified data may later be described in group-level reports.

8.6. Public Posting (Protocol, Information Notice, De-identified Results): Public posting will occur at: **https://github.com/ohdsi-studies/MindMeetsMachines** (GitHub repository **managed by the PIs**). Only de-identified materials (as allowed) will be posted.

Appendices: Detailed Methodology and Procedures

Appendix A: Experimental Arm Procedures

A.1. GenAI Arm Procedures

- **Pre-Workshop Submission:** Brokers send inputs in advance. GenAI outputs + methodology documentation are due to **brokers by email** on or before **October 8, 2025** (e.g., sachson@ohdsi.org, dravida@ohdsi.org).
- **Autonomous Execution Requirement:** Execution must be autonomous. **No** HITL or post-editing. **Single-prompt** only.
- **Verification:** No A/V recording is collected/reviewed by PIs. Verification relies on submitted documentation.

A.2. Human Arm Procedures (The "Split and Reconcile" Model)

- **Team Formation & Characterization:**
    - PIs do **not** collect any survey data or participant characteristics.
    - Allocation: self-selection or broker-managed grouping (Section 4.4).
    - Bias mitigation: grouping managed by brokers; no PI judgment.
- **Independent Creation:** Two sub-teams (broker-divided) work in parallel in a broker-provisioned ATLAS instance. Use standard OMOP concepts primarily; source codes only in edge cases (not used in primary analysis). **Human arm may not use external LLM/GenAI tools.**
- **Reconciliation:** Broker-facilitated Modified Delphi (Appendix C.3) to produce a single reconciled list.
- **Final Output:** Reconciled list is the official Human Arm output. Brokers enforce naming conventions, attribute sequential team numbers, and mask participant attribution before transfer to PIs.

Appendix B: Workshop Logistics and Timeline (Broker-Run Activities)

- **Introduction & Standardization (30 min):** Brokers review logistics + Information Notice (Appendix J).
- **Phase 1: Human Baseline Generation (60–90 min):** Split-and-Reconcile. Brokers manage timing; ATLAS locked at hard stop.
- **Critical Window (45 min):**
    - **Data Sprint (Broker):** Extract human/AI sets; resolve to standard concepts; **blind** all sets; prepare adjudication bundle.
    - **Methodology Review (Broker-Hosted):** Optional AI leads' overview to participants; occurs **after** concept creation; **no PIs present**.
- **Conclusion (15 min):** Brokers close workshop.

Appendix C: Evaluation Methodology Details (PI Activities)

C.1. Establishing the Gold Standard (Post-Hoc Adjudication) Brokers provide the union of all concepts. PIs adjudicate:

- **Intersection:** universally agreed concepts.
- **Delta:** any concept with disagreement.
- **Blinded Review:** Brokers present the Delta via app/spreadsheet; PIs have no source identities.
- **Interface:** Only concept-intrinsic info (name, vocabulary, class/hierarchy, prevalence). **No** "agreement level" shown.

- **TGS:** Intersection + broker-presented Delta approvals.

C.2. Vocabulary Standardization Strategy Primary analysis on OHDSI standard concepts (e.g., SNOMED). Secondary analysis (if any) on selected source codes via OHDSI mappings.

C.3. Consensus Methodology (Modified Delphi) Broker-facilitated during reconciliation; PIs' internal discussion during adjudication; final judgment by lead clinical expert.

Appendix D: Statistical Analysis Plan (SAP)

- **Primary Analysis:** The primary endpoint is the Prevalence-Weighted F1 score for each arm against the TGS, calculated within the OHDSI standard vocabulary substrate. Descriptive statistics (mean, median, standard deviation, range, 95% Confidence Intervals) will be reported for the Human Arm and each distinct GenAI pipeline across all clinical ideas.

Here's the formula for the F1 score:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

$$\textbf{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$$

$$\textbf{Recall} \text{ (also called Sensitivity)} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

General formula for a Weighted F1 score:

$$\text{Weighted F1} = \sum_{i=1}^{n} w_i \times F_{1,i}$$

The OHDSI network **prevalence** concept count dataset acts as the weight Wi. The **F1 score** is computed for each arm against the TGS (target group set) within the OHDSI vocabulary.

- **Hypothesis Testing (Non-Inferiority):** We will test the hypothesis that GenAI approaches are non-inferior to the reconciled Human workflow. A non-inferiority margin will be established *a priori*. Paired t-tests or Wilcoxon signed-rank tests (depending on the distribution of the scores) will be used to compare the performance of each AI arm to the Human arm across the paired clinical ideas.

- **Inter-Human Variability Analysis:** To address Secondary Objective 1, we will calculate the F1 score between the two independent human sub-teams (Team A vs. Team B) before reconciliation. This will be compared to the F1 score of the reconciled human output vs. the TGS to quantify the improvement gained through the consensus process.

- **Sensitivity Analysis:** The secondary analysis using source vocabularies (Appendix C.2) will serve as a sensitivity analysis to assess the impact of vocabulary mapping quality on the results.

- **Subgroup Analysis:** Performance metrics will be analyzed stratified by phenotype/ clinical domain.

Appendix F: Study Limitations

This study design has several limitations that must be acknowledged:

- **The "Missed Concept" Blind Spot:** The on-the-fly adjudication methodology establishes the gold standard based on the union of generated concepts. This design cannot identify or account for concepts that were missed by *all* experimental arms (Human and AI).

- **Time Constraints:** The execution of the human workflow within a constrained workshop setting may not fully reflect the time typically invested in real-world phenotype development.

- **Vocabulary Mapping Confounding:** Both the Primary and Secondary analyses are dependent on the quality, completeness, and bi-directionality of the OHDSI vocabulary mappings. Imperfections in the mapping may introduce confounding variables, potentially penalizing an arm (human or AI) for errors inherent in the vocabulary infrastructure rather than the methodology itself.

- **Generalizability:** The results are specific to the clinical ideas selected, the expertise of the human participants, and the specific GenAI pipelines included. The phenotype selected may not cover the full complexity amd variations of concept set selection of the full spectrum of clinical domains and terms.

- **Small sample size:** 5 phenotypes may not be sufficient to evaluate the performance gains or losses of GenAi arms.

- Expert clinical opinion and adjudication may not be consistent and not an ideal gold standard.

Appendix G: Decision Archaeology

(This section summarizes the rationale for key decisions made during the planning phase, based on the peer-review critique.)

- **The Structural Proficiency Score (SPS)**

    - *Rejection Rationale:* Accuracy of the final concept *list* is the first-order priority; structure is secondary and can be algorithmically optimized. Additionally, not all AI tools produce OHDSI expressions.

- **Pre-Curated "Pristine" Gold Standard (A Priori Model)**
    - *Rejection Rationale:* Risk of flawed curation and the "Ceiling Effect" (penalizing the discovery of valid new concepts). Dynamic adjudication was deemed superior.
- **Single-Team Human Process**
    - *Rejection Rationale:* Susceptible to "groupthink" and insufficiently rigorous. The "Split and Reconcile" model ensures independent work followed by peer review.
- **Clinical Experts Participating in Concept Generation in their Specialty**
    - *Rejection Rationale:* Conflict of Interest and Bias. Experts cannot participate in construction and then serve as objective adjudicators for the same concept set. A strict separation of duties was mandated.

Appendix H: Non-Engagement Statement

The PIs are **not engaged** in HSR under 45 CFR 46. Activities are limited to analysis of de-identified outputs.

1. No PI interaction/intervention with individuals.
2. Honest brokers conduct all human-facing activities.
3. PIs receive only de-identified outputs; no identifiers; no key.
4. No PII/free text/user logs provided to PIs.
5. Data are coded; key retained by brokers; DUA prohibits release.
6. No re-identification; purge + notify if any identifiers are inadvertently received.

Appendix I: Data Use Agreement (DUA) Clauses (Bulleted Summary)

(allowed/prohibited scope; no key sharing; no re-ID; security; incident handling; retention/destruction—retained as previously drafted)

Appendix J: Broker Information Notice (Draft)

To be distributed by brokers. (No signatures collected by PIs; optional signatures, if any, are broker-only and never linked to random IDs.)

Contacts (Broker Only): dravida@ohdsi.org | sachson@ohdsi.org **Public Posting (Protocol/Notice/De-identified Results, if allowed):** https://github.com/ohdsi-studies/MindMeetsMachines