### Minds Meet Machines (MMM) Challenge: Comprehensive Methodology and Statistical Analysis Plan

#### 1. Study Objectives and Hypotheses

**1.1. Primary Objective:** To evaluate the performance (accuracy) of GenAI-driven approaches compared to a rigorous, consensus-based, human-led workflow for the task of translating standardized clinical descriptions into phenotype concept sets, utilizing a dynamically created gold standard.

1.2. Primary Hypotheses:

The primary analysis focuses on the comparison of the primary performance metric, the Prevalence-Weighted F1 Score (F1W​), between the GenAI workflows and the human-led workflow.

Null Hypothesis (H0​):

There is no significant difference in the Prevalence-Weighted F1 Score (F1W​) between the concept sets generated by the GenAI-driven workflow(s) and the concept sets generated by the rigorous, human-led workflow.

Alternative Hypothesis (H1​):

There is a significant difference in the Prevalence-Weighted F1 Score (F1W​) between the concept sets generated by the GenAI-driven workflow(s) and the concept sets generated by the rigorous, human-led workflow.

**1.2. Secondary Objectives:**

* To characterize the baseline variability and agreement across all independent workflows (Human and AI) prior to adjudication.
* To quantify inter-human variability and the improvement gained through the human consensus process (Consensus Gain).
* To analyze the clinical impact of disagreements using prevalence-weighted scoring metrics.
* To conduct sensitivity analyses based on source vocabularies (e.g., ICD-10-CM).
* To conduct a qualitative analysis of human reasoning and consensus-building processes.

#### 2. Experimental Design Overview

**2.1. Study Design:** This is a methodological evaluation utilizing a parallel assignment, multi-arm, comparative experimental design with mixed methods (quantitative and qualitative analysis).

**2.2. Data Structure and Standardization:** The experiment compares the outputs of different methodologies applied to the same ~7 standardized pre-specified clinical ideas. The design is **paired**, meaning all experimental arms process the same inputs. Statistical analysis will account for this pairing to control for variability in difficulty between clinical ideas. The version of the OMOP vocabulary used is standardized across all workflows to the August 27th, 2025 version.

**2.3. Sample Size and Power:** The study includes N=7 standardized clinical ideas. Given this small sample size, the study has limited statistical power to detect small or moderate differences. Therefore, the analysis will focus primarily on descriptive comparisons and the magnitude of observed differences. Inferential statistics (p-values and Confidence Intervals) will be considered exploratory rather than confirmatory.

#### 3. Experimental Arms and Procedures

**3.1. Arm 1: Human Workflow (Control/Benchmark)** A rigorous "Split and Reconcile" model is employed to mitigate groupthink and establish a robust human benchmark.

* **Stratified Randomization:** Participants self-report expertise (Clinical and Informatics). Stratified randomization ensures a balanced mix of expertise across teams.
* **Independent Creation (Phase 1):** For each clinical idea, participants are divided into 2 or more independent sub-teams working in parallel using standard OHDSI tools (ATLAS, PHOEBE). The use of GenAI is prohibited.
* **Reconciliation:** The sub-teams convene and utilize a Modified Delphi (Delphi-like) method, facilitated by an honest broker, to negotiate a single, finalized concept set.
* **Data Capture:** Discussions during the creation and reconciliation phases are audio-recorded for qualitative analysis, authorized by informed consent.

**3.2. Arm 2: GenAI Workflows (Intervention)** Multiple distinct GenAI pipelines (denoted as ) participate.

* **Autonomous Execution:** Generation of concept sets must be autonomous.
* **Constraints:** "Human-in-the-loop" (HITL) intervention, post-editing of the AI-generated output, and sequential prompting (prompting may occur only once) are strictly prohibited.
* **Submission:** GenAI teams must submit their outputs by October 8th, 2025, 6:00 PM EST, along with documentation of their methodology.

#### 4. Evaluation Strategy and Adjudication

The evaluation focuses exclusively on the accuracy of the resolved **concept list**. The MMM Challenge employs a dynamic, post-hoc "On-the-Fly Adjudication" model to establish the True Gold Standard (TGS).

**4.1. The Adjudication Process and Blinding:**

* **Pooling:** The union of all concepts generated across all *N* experimental arms is compiled.
* **Delta Identification:**
  + The **Intersection** (Automatic Gold Standard - AGS): Concepts universally agreed upon.
  + The **Delta:** Concepts where any disagreement exists.
* **Blinded Adjudication:** Adjudicators (Outcomes Assessors) review the Delta, strictly blinded to the source (Human vs. AI) of each concept. To prevent bias, the interface displays only intrinsic concept information (Name, Vocabulary, hierarchy, prevalence) and excludes the "Agreement Level".
* **Adjudication Execution:** A designated clinical expert makes the final determination (Yes/No only) on whether to include each disputed concept in the TGS. Other participants may provide input, but the designated expert holds final decision authority. A neutral volunteer ("Honest Broker") may be present to ensure the adjudicator remains strictly true to the original clinical description, preventing definition drift.
* **TGS Definition:** The TGS is constructed by combining the AGS plus the concepts from the Delta approved by the blinded adjudicators.
* **Conflict Mitigation:** Experts adjudicating a specific domain cannot participate in the concept set creation for that domain.

#### 5. Statistical Analysis Plan (SAP)

The study design necessitates a structured, two-phased approach to the Statistical Analysis Plan (SAP). The analysis first seeks to understand the variability of the inputs (Phase A) and then proceeds to measure the accuracy of the outputs against the TGS (Phase B).

5.1. Phase A: Pre-Adjudication Analysis (Before TGS Creation)

This analysis occurs immediately after Phase 1 (independent creation) and before human reconciliation or adjudication. It utilizes the raw outputs from all independent workflows (all human sub-teams and all AI pipelines) to address Secondary Objective 1.

**Objective:** To quantify and visualize the similarities and differences across all independent experimental workflows in a blinded manner.

**Statistical Analysis Plan:** All analyses and visualizations will utilize anonymized labels (e.g., Workflow A, Workflow B) to maintain blinding.

**5.1.1. Data Preparation and Structure:**

* A binary concept presence/absence matrix will be constructed (Rows: Union of all concepts; Columns: Independent workflows).
* This matrix will be annotated with OHDSI network Concept Prevalence (RecordCount).

**5.1.2. Descriptive Set Characteristics (Quantifying Difference):**

* **Set Size:** The total number of concepts included by each workflow. Descriptive statistics (mean, median, range) will be reported.
* **Uniqueness:** The count and proportion of concepts proposed by only a single workflow.

**5.1.3. Quantitative Similarity Metrics (Quantifying Similarity):**

* **Jaccard Index:** The primary metric for set similarity, defined as the size of the intersection divided by the size of the union of two sets: .
* The Jaccard Index will be calculated for every pair of workflows.

**5.1.4. Visualization of Similarity Patterns:**

* **Similarity Heatmap and Clustering:** A heatmap will visualize the pairwise Jaccard Indices. Hierarchical clustering will be applied to identify clusters of workflows that produce similar outputs and identify methodological outliers.
* **Upset Plot:** An Upset plot will visualize the intersections between the concept sets, quantifying concepts unique to specific workflows and shared by various combinations.

**5.1.5. Concept Agreement Analysis and Clinical Context:**

* **Match Score Distribution:** We will calculate the "Match Score" for each concept (the number of workflows that included the concept). A histogram will visualize the distribution of these scores.
* **Agreement vs. Prevalence:** A scatter plot visualizing Match Score (Y-axis) vs. Concept Prevalence (X-axis, log scale) will be generated. This identifies clinically significant areas of difference (i.e., high prevalence concepts with low Match Scores).

5.2. Phase B: Post-Adjudication Analysis (After TGS Creation)

This analysis occurs after the TGS has been established through blinded adjudication. It compares the performance of the finalized Human Workflow (reconciled output) and the AI pipelines against the TGS to address the Primary Objective and remaining Secondary Objectives.

**5.2.1. Primary Objective Analysis**

**Primary Outcome Measure:** **Prevalence-Weighted F1 Score ()**: The harmonic mean of precision and recall against the TGS, weighted by concept prevalence ().

Metric Calculation: We calculate weighted counts for True Positives (WTP), False Positives (WFP), and False Negatives (WFN) relative to the TGS:

WTP=∑i∈TP​Vi​ ; WFP=∑i∈FP​Vi​ ; WFN=∑i∈FN​Vi​

Weighted Precision (PW​) and Weighted Recall (RW​):

PW​=WTP+WFPWTP​ ; RW​=WTP+WFNWTP​

Statistical Analysis Plan:

The significance level (Alpha) for all exploratory hypothesis testing will be set at 0.05.

* **Descriptive Statistics:** Descriptive statistics (mean, median, SD, range, 95% CI) for the (and related metrics: Precision, Recall, weighted/unweighted) will be reported for the Human Workflow (reconciled output) and each AI pipeline.
* **Comparative Analysis:** The difference in scores between each AI workflow () and the Human workflow () will be calculated for each clinical idea. Paired differences will be summarized (mean difference, 95% CI of the mean difference).
* **Exploratory Inferential Analysis (Paired Comparisons):** To test the primary hypothesis (), the normality of the paired differences will be assessed (e.g., Shapiro-Wilk test). A **Paired t-test** (if normal) or **Wilcoxon Signed-Rank Test** (if non-normal) will be used. Given the small sample size (N=7), the Wilcoxon Signed-Rank Test is generally preferred if there is any doubt about the normality assumption.
* **Exploratory Non-Inferiority Assessment:** We will explore if GenAI approaches are non-inferior to the human workflow. The non-inferiority margin () is pre-specified as 0.05. This margin represents the maximum acceptable reduction in that would still render the AI approach practically comparable. Non-inferiority is suggested if the lower bound of the 95% CI for the mean difference () is greater than (i.e., > -0.05).
* **Multiple Comparisons Adjustment (Exploratory):** As multiple AI pipelines are compared against the single Human control, **Dunnett's Test** will be used to adjust the CIs and p-values for the exploratory inferential analyses (both difference and non-inferiority), controlling the family-wise error rate.

**5.2.2. Secondary Objectives Analysis**

* **Inter-Human Variability and Consensus Gain (Objective 2):**
  + *Inter-Human Variability:* Calculate the average pairwise F1 score between all combinations of the *n* independent human sub-teams (using Phase 1 outputs).
  + *Consensus Gain:* Use a paired test (Wilcoxon Signed-Rank Test preferred) to compare the F1 scores (vs TGS) of the independent human sub-teams against the F1 score (vs TGS) of the reconciled human output.
* **Clinical Impact of Disagreements (Objective 3):**
  + Calculate **Spearman's Rank Correlation Coefficient** between the scores and the Unweighted F1 scores across all workflows. A low correlation indicates that performance is highly sensitive to concept prevalence.
* **Source Vocabulary Analysis (Objective 4):**
  + *Sensitivity Analysis:* The primary descriptive analysis will be repeated using the Source Vocabulary F1 Score (calculated against the TGS based on source codes).
* **Qualitative Analysis of Human Reasoning (Objective 5):**
  + *Thematic Analysis:* Transcribed audio recordings from the Human Reconciliation and Adjudication phases will be analyzed using thematic analysis to identify patterns in reasoning and consensus-building. Transcripts will be independently coded by two researchers, with assessment of inter-coder reliability (e.g., Cohen's Kappa).

# Appendix F1 score:

The F1 score is a statistical metric used to evaluate the performance of a classification model or, in the context of information retrieval, the accuracy of a generated set of results compared to a known "gold standard."1

It is designed to provide a single score that balances the trade-off between two crucial aspects of performance: **Precision** and **Recall**.2

Here is a detailed breakdown of the F1 score and its components.

**1. The Components: Precision and Recall**

To understand the F1 score, one must first understand the underlying categories of evaluation results:

* **True Positive (TP):** An item was correctly identified as positive (it was included, and it should have been).3
* **False Positive (FP):** An item was incorrectly identified as positive (it was included, but it should not have been).4
* **False Negative (FN):** An item was incorrectly identified as negative (it was not included, but it should have been).

Based on these categories, we define Precision and Recall:

**Precision (Positive Predictive Value)**

Precision measures the *exactness* of the results.5 It answers the question: "Out of all the items identified as positive, what proportion were correct?" High precision means few False Positives.

**Recall (Sensitivity)**

Recall measures the *completeness* of the results.6 It answers the question: "Out of all the items that are actually positive, what proportion were successfully identified?" High recall means few False Negatives.

**2. The F1 Score Calculation**

There is often a trade-off between precision and recall.

* If you attempt to capture every possible positive item (aiming for high recall), you will likely include many incorrect items (resulting in low precision).
* If you are very strict about only including items you are certain about (aiming for high precision), you will likely miss some correct items (resulting in low recall).7

The F1 score combines these two metrics into a single value by calculating their **harmonic mean**.8

Interpretation:

The F1 score ranges from 0 to 1.9

* **1:** Indicates perfect precision and perfect recall.
* **0:** Indicates that either precision or recall (or both) is zero.10

**3. Rationale for Using the F1 Score**

The F1 score is favored over simple accuracy (the total number of correct predictions) because it provides a better measure of performance, especially when data is imbalanced.11

Furthermore, the use of the *harmonic mean* (rather than a simple average) is significant because it penalizes extreme imbalances between precision and recall.12 For an F1 score to be high, both precision and recall must be high.13

For example, if a model has a Precision of 1.0 (perfect) but a Recall of 0.2, the simple average is 0.6, but the F1 score is 0.33. The F1 score effectively highlights that the overall performance is poor because the recall is low.

**4. Context in the "Minds Meet Machines" Study**

In the "Minds Meet Machines" (MMM) Challenge protocol, the F1 score is used to evaluate how accurately the Human and AI workflows generate medical concept sets compared to the True Gold Standard (TGS).

Crucially, the study uses a specialized version as its primary outcome: the **Prevalence-Weighted F1 Score ()**.

In clinical research, not all concepts are equally important. Missing a concept that occurs frequently in a database (high prevalence) has a greater impact than missing a rare concept. The addresses this by weighting the calculation:

* Errors (False Positives or False Negatives) on high-prevalence concepts are penalized more heavily than errors on low-prevalence concepts.
* This ensures the evaluation metric is clinically relevant, prioritizing accuracy on the concepts that matter most in real-world data.