IRB Research Protocol

**IRB Research Protocol**

**1. Title of Study**

Identifying and Validating Clinical Concept Relationships from the OMOP CDM

**2. Principal Investigator**

[PI name and contact]

**3. Study Personnel**

[Additional study personnel]

**4. Study Summary**

This study will derive, evaluate, and validate clinically meaningful relationships between clinical concepts using retrospective, de-identified data from an existing OMOP Common Data Model (CDM) database. Tools from the OHDSI (Observational Health Data Sciences and Informatics) ecosystem will be used to extract co-occurrence patterns and generate concept relationship tables.  Large language models (LLMs) will be applied to the relationships (not to the source data) to characterize the semantic nature of these relationships, with human validation of discordant classifications. No personally identifiable information (PII) or protected health information (PHI) will be accessed.

**5. Objectives and Hypotheses**

**Primary Objective:**  
To develop and validate concept relationship tables from OMOP CDM data using OHDSI tools and Generative AI.

**Secondary Objectives:**

* To assess co-occurrence statistics of concept codes within episode-based timeframes.
* To classify relationships (e.g., causal, associative, comorbid) using Generative AI-based semantic validation.
* To explore optimal granularity (e.g., ICD block, ICD-3, ICD-4) for concept grouping across analytic use cases.

**Hypothesis:**  
Aggregated, de-identified OMOP CDM data can be used to detect robust and clinically meaningful concept relationships that support the development of scalable analytic models.

**6. Background and Significance**

Efforts to build learning health systems and apply evidence-based medicine at scale are hindered by fragmented data and analytic methods that focus on one-off research questions. This study seeks to address that gap by systematically constructing clinical concept relationship tables that can inform analytic systems and LLMs. Prior work has shown that diagnosis co-occurrence and semantic validation methods can yield interpretable and scalable clinical insights.

**7. Study Design**

**Design:** Retrospective, secondary data analysis using de-identified OMOP CDM data.

**Data Source:** An existing OMOP CDM database hosted [details of hosting. A formal Data Use Agreement (DUA) is in place. The dataset is de-identified using the HIPAA Safe Harbor method and does not include any direct or quasi-identifiers.

**Analytic Steps:**

1. Define concept eras or episodes using ICD groupings and AHRQ Chronic Condition Indicator for diagnoses, for example.
2. Perform joins on the concept tables using bounded time windows.
3. Calculate co-occurrence statistics (e.g., lift, support) across concept pairs.
4. Filter concept pairs based on statistical thresholds.
5. Apply Generative AI to label the relationships semantically.  Only the concept description pairs will be sent to the Generative AI.
6. Validate a subset of the relationships with expert review.

**Output:** Tables of concept pairs with calculated metrics and semantic relationship labels (e.g., "causal", "comorbid", "no relationship") and direction.

**8. Subject Population**

**Human Subjects:** No direct interaction or intervention with human subjects.  
**Data Characteristics:** All data are retrospective, observational, and fully de-identified.  
**Inclusion/Exclusion:** Not applicable.

**9. Privacy and Confidentiality**

* Only de-identified structured data will be accessed.
* No identifiers (names, dates of birth, medical record numbers, zip codes, or encounter dates) will be examined.
* Data is stored on encrypted servers with role-based access control and audit logs.
* Only concept pair descriptions (no patient-level data) are sent to the Generative AI.

**10. Risks and Benefits**

**Risks:** The study presents minimal risk, with no use of PHI, PII, or direct patient interaction. The data used to derive the statistical relationships are de-identified, and appropriate security safeguards are in place. The results are aggregated at the population level and consist of data about the relationships between clinical concepts only.

**Benefits:** This research may improve understanding of the relationships between clinical concepts and support scalable methods for evidence generation from real-world data.

**11. Data Sharing and Dissemination**

Aggregated findings will be submitted to the OHDSI Symposium and peer-reviewed journals. Published materials (e.g., concept relationship tables) will exclude any individual-level data.