**TAXIS – Step 1: Simplified Diagnostic Hierarchies and Relationship Tables for the OHDSI Community**

**Stephen H. Bandeian\*, MD, JD; Shaun Grannis, MD, MS✝, J. Marc Overhage, MD, PhD🟄**

**\***Biomedical Informatics & Data Science, Johns Hopkins School of Medicine

**✝**Regenstrief Institute and Family Medicine, Indiana University School of Medicine

**🟄**Fairbanks School of Public Health and Regenstrief Institute, Indiana University

# Background

The OHDSI/OMOP Common Data Model (CDM) is highly successful, providing a strong foundation for data standardization. There is a valuable opportunity to enhance existing shared resources by developing more comprehensive and interoperable tools. These enhancements can reduce the need for custom solutions, minimize duplication, and improve consistency and integration. By building on this progress, the CDM can be further developed to support improved systematic outcome improvement efforts, in addition to single-topic research. We are launching "TAXIS", building on work initiated at AHRQ in 2007 and continued elsewhere.1,2,3,4,5,6 While PHOEBE supports concept-set construction via lexical/semantic/data-driven signals, TAXIS goes further by supplying directed diagnosis–diagnosis relationships and episode logic, improving phenotype refinement, causal covariate selection/negative controls, and pathway mapping. TAXIS will create resources to help analysts with daily tasks and develop a comprehensive framework. It aims to reduce redundancy, improve consistency, accelerate evidence generation, transparency, and systematic healthcare improvement.

TAXIS will proceed in sequential steps culminating in a comprehensive knowledge graph of clinical concept relationships. Step 1 involves creating a simplified diagnostic taxonomy that strikes a balance between detail and sample size, thereby avoiding inadvertent fragmentation of a single episode due to variations in the exact diagnoses reported during a condition episode. It will also establish a diagnosis-to-diagnosis relationship table, which represents a subset of a larger knowledge graph that encompasses relationships among diagnoses, services, and medications. This table, a simplified two-node representation, captures clinically meaningful connections between diagnoses. It supports the identification of potential causes of symptoms, disabilities, illnesses, and complications, offering insights that can inform prevention and care planning. Subsequent steps will create service and medication hierarchies, along with service–diagnosis and medication–diagnosis tables to identify symptoms, findings, and complications caused by services and medications. Additionally, diagnosis-service and diagnosis-medication tables will be used to identify services and medications commonly used for specific diagnoses.

# Methods

The diagnosis taxonomy was created starting with the ICD-10-CM hierarchy to which SNOMED concepts were mapped. The taxonomy's first two levels correspond to ICD-10 chapter and sub-chapter headings. The lower three levels are selected from the set of truncated ICD-10 codes (3–5 digits) and their labels, based on the criteria listed below. The fourth level also includes attributes indicating concept type (illness, injury, symptom, finding) and temporal nature (time-limited, e.g., acute infections, or ongoing/chronic, e.g., heart failure.

* diagnosis3: Combines similar or easily confusable conditions.
* diagnosis4: Identifies distinct illnesses or injuries with minimal detail.
* diagnosis5: Adds limited additional detail.

Using the condition groupings primarily at the 4th level to define condition episodes, statistical and generative AI methods are used to create a diagnosis-diagnosis relationship table in four steps. We do not use OMOP CONDITION\_ERA directly because eras operate at single concept\_id granularity. TAXIS episodes are built at the **taxonomy group** level (e.g., diagnosis4), aggregating clinically proximate concepts before era-like consolidation; this generalizes across coding variants while preserving episode continuity.

**Step 1:** Construct diagnosis-based condition episodes from the condition occurrence table using condition groups from the diagnosis taxonomy. Episode or phase durations, especially for chronic conditions, are defined based on clinically reasonable assumptions, such as when there is evidence of continued management or unchanged status or stage.

**Step 2:** Calculate diagnosis pair co-occurrence rates by self-joining the condition episodes, counting diagnosis pairs per individual. The IRR is defined as the exposed incidence rate divided by the unexposed rate, with expected exposed cases = (unexposed rate × exposed person-time)

**Step 3:** Apply the Poisson Wald Z on log(IRR) (two-sided; 95% CI from log(IRR) ± 1.96·SE) to select a manageable set of candidate relationships for further validation.

**Step 4:** Validate and classify relationships between diagnosis pairs using a GPT Turbo, a foundational large language model (LLM) using structured prompting. The LLM can confirm clinically recognized relationships classifies pairs into: (1) A causes B; (2) B causes A; (3–4) indirect causation; (5) common cause; (6–7) treatment-caused; (8) similar initial presentation; (9–10) subset/parent; (11) no clear relationship. ‘Co-occurring’ without mechanism maps to (11). Expert clinician review will then focus on relationships where statistical and LLM assessments are discordant.

Two general internal medicine-trained informaticians independently reviewed LLM classifications with substantial agreement. [Steve: What do we want to say in response to the reviewer who asked about pancreatic pseudocysts causing pancreatitis, which, of course, can rarely be caused by compressing the pancreatic duct. Is this an example where prompts and relationship categories can be refined? Do we say that we are focused on predominant or common relationships? Ignore it?]

Table – Selected classification results using GPT Turbo, Additional results in supplemental data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CONCEPT\_A** | **CONCEPT\_B** | **Observed Cases** | **Observed to Expected Ratio (IRR)** | **Relationship** |
| Acute Pancreatitis | Acute posthemrg anemia | 1,455 | 1.0 | A causes B |
| Thrombosis - portal vein | 161 | 7.7 |
| Pancreatic disorders nos | 393 | 10.2 |
| Abdominal pain | 74 | 2.0 |
| Pancreatic pseudocyst | 2,193 | 20.6 |
| Peritonitis - nec/nos | 1,011 | 4.2 |
| Substance abuse | 1,664 | 2.2 | B causes A |
| Biliary tract obstruction | 2,549 | 19.4 |
| Hypercalcemia | 528 | 1.2 |
| Magnesium disorders | 1,289 | 2.2 |

# Results

We will provide executable code and processes that produce the relationships in Table 2, along with data from at least one OHDSI instance.

Table -- The pairwise relationship table produced will contain the diagnosis groupings (e.g., diagnosis4) pairs, summary data derived from the CDM (observed and expected cases and exposure days), statistical association metrics (IRR and Z-score), and relationship type and direction based on LLM validation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CONCEPT\_A** | **CONCEPT\_B** | **Observed Cases** | **Expected Cases** | **Exposure Days** | **IRR** | **Z-score** | **Relationship** |
| Pneumonia | Sepsis | 150 | 25 | 2,100 | 6.0 | 25.0 | A causes B |
| Cough | 275 | 70 | 2,100 | 3.9 | 24.5 |
| Acute bronchitis | 230 | 80 | 2,100 | 2.9 | 16.8 | Similar conditions |
| Ankle pain | 95 | 40 | 17,100 | 2.4 | 8.7 | None |

# Conclusion

TAXIS builds on the robust foundation of the OMOP CDM and the collaborative energy of the OHDSI community. It offers the opportunity to extend this infrastructure by introducing a clinically grounded diagnostic taxonomy and a set of relationships between concepts (diagnoses in Step 1) that support more reproducible, scalable, and efficient analytics.

For OHDSI users, these tools simplify cohort building, improve phenotype accuracy, and streamline confounding control. For instance, users studying sepsis can easily identify related upstream and downstream conditions using the TAXIS relationships without manual mapping, thereby improving both efficiency and analytical rigor. Subsequent steps, including linking diagnoses to services and mapping service-to-service relationships, will unlock insights into care pathways, service variation, and potential improvements in care delivery. These steps will enable more dynamic modeling of real-world clinical practice.

TAXIS reflects the strength of the OHDSI community: building shared, rigorous, and reusable tools. Through continued collaboration, this foundation will evolve into a knowledge layer that enhances research and accelerates its real-world impact. This first step in the TAXIS project is just the beginning of what we can achieve together.

# References

1. Bandeian S, Clinical Analytic Model, Council on Health Care Economics and Policy, Princeton Conference XV, 2008, https://heller.brandeis.edu/council/pdfs/2008/Steve-Bandeian.pdf.
2. Bandeian S, Population Health Management, Informatics, and the Clinical Analytic Model, Johns Hopkins Informatics Grand Rounds, Feb 2014.
3. Blue Cross Blue Shield Association. Blue Cross Blue Shield Health Index identifies the top 10 conditions nationwide. BCBSA Association News. Available at[: https://www.bcbs.com/about-us/association-news/blue-cross-blue-shield-health-indexidentifies-top-10-conditions-nationwide.](https://www.bcbs.com/about-us/association-news/blue-cross-blue-shield-health-index-identifies-top-10-conditions-nationwide) Accessed June 27, 2025.
4. Bandeian S, Using a Longitudinal Patient History Sourced from Claims Data to Analyze and Predict Potentially Avoidable Utilization, Costs, and Adverse Outcomes, Johns Hopkins CHSOR Seminar, Dec 2019.
5. Bandeian S, Tompkins CP, Davison A. A Future Health Care Analytic System: Part 1—What the Destination Looks Like. In:

Kiel JM, Kim GR, Ball MJ, eds. Healthcare Information Management Systems: Cases, Strategies, and Solutions. 5th ed. Cham, Switzerland: Springer International Publishing; 2022:404.

1. Bandeian S, Tompkins CP, Davison A. A Future Health Care Analytic System: Part 2— What is Needed and ‘Getting It Done’. In: Kiel JM, Kim GR, Ball MJ, eds. Healthcare Information Management Systems: Cases, Strategies, and Solutions.

5th ed. Cham, Switzerland: Springer International Publishing; 2022:419.