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

Stephen H. Bandeian*, MD, JD; J. Marc Overhage, MD, PhD**

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

****Fairbanks School of Public Health, Indiana University

Background

The OHDSI/OMOP Common Data Model (CDM) has been 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 "THEMIS", building on work initiated at AHRQ in 2007 and continued elsewhere.[1,2,3,4,5,6] THEMIS creates resources that help analysts with daily tasks while also developing a comprehensive framework. THEMIS aims to reduce redundancy, improve consistency, accelerate evidence generation, transparency, and systematic healthcare improvement.

THEMIS will proceed in sequential steps. Step 1 includes creating a simplified diagnostic hierarchy that balances detail against sample size, while also preventing inadvertent fragmentation of a single episode due to differences in the diagnoses reported during the episode. Currently, ODHSI lacks user-friendly, standardized tools to manage these issues, leading to inconsistent analytic practices, unnecessary complexity, and potential errors. Step 1 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 include 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-svc and diagnosis-medication tables will be used to identify services and medications commonly used for specific diagnoses. Additional components will be developed for the analysis of care processes.

Methods

To create a diagnosis taxonomy, we start with ICD10-CM and then map SNOMED concepts onto it. (Prior attempts to start with SNOMED, which is a directed acyclic graph, were unsuccessful due to its complexity.) The hierarchy's first two levels correspond to ICD-10 chapter and sub-chapter headings. The following three levels support analytics and the use of relationship tables:

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

These categories are selected from the set of truncated ICD10 codes (3–5 digits) and their labels, based on the criteria listed above. 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.

The next task is creating a diagnosis-diagnosis relationship table, which involves four steps:

Step 1: Construct diagnosis-based condition-eras from OMOP's condition occurrence table and the diagnosis hierarchy. Episode durations, especially for chronic conditions, will be extended based on clinically reasonable assumptions.

Step 2: Calculate diagnosis pair co-occurrence rates by self-joining the condition eras, counting 'Diagnosis A and Diagnosis B' pairs within clinically relevant intervals per individual. An incidence rate ratio (IRR) is computed comparing incidence in exposed individuals (those with the first diagnosis) to a baseline in unexposed individuals. Statistical significance will be tested using a Poisson-based Z-test.

Step 3: Apply statistical thresholds (elevated IRRs and significant Z-scores) to select a manageable set of candidate relationships for further validation.

Step 4: Validate and classify relationships between candidate diagnosis pairs using large language models (LLMs). The LLMs confirm clinically recognized relationships, identifying cases of direct or indirect causation, treatment-related relationships, or similar/confusable conditions, considering all directional permutations.

Results

Based on prior work (not using OMOP CDM), we expect to provide executable code and data from at least one OMOP instance with output structured as :

Planned October 2025 Diagnosis Relationship Table Using OMOP							
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

To demonstrate feasibility, we conducted an LLM-based validation and classification process using OpenAI's GPT-Turbo model and a diagnosis-to-diagnosis relationship table previously developed by AHRQ, utilizing methods similar to those described here. A small sample of results is shown below. Results from the 98-record test set are included in an attached Excel spreadsheet and appear reasonable.

Sample Output from LLM Classification of 2010 AHRQ Diagnosis Relationship Table				
CONCEPT_A	**CONCEPT_B**	**Observed Cases**	**Observed to Expected Ratio**	**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	

Conclusion

THEMIS builds on the robust foundation of the OMOP Common Data Model and the collaborative energy of the OHDSI community. These shared resources have enabled consistent, large-scale analytics across diverse data environments. THEMIS offers the opportunity to extend this infrastructure by introducing a clinically grounded diagnostic hierarchy and a set of relationships between concepts 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 without manual mapping, thereby improving both efficiency and analytical rigor. Future expansions—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.

THEMIS 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 THEMIS 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 top 10 conditions nationwide. BCBSA Association News. Available at: 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.

[6] 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.