Adapting Hypergraph-based Knowledge Graph Models to the OMOP Common Data Model

She Yuting

Department of Mathematics
National University of Singapore

Honours Year Project Presentation
AY2025/2026

Supervisor: Han Fei

Motivation

Clinical data landscape

- **Electronic Health Records (EHRs)**: patient-level, temporal, sparse and noisy.
- **Biomedical Knowledge Graphs (KGs)**: population-level, curated relations (e.g. drug–disease).

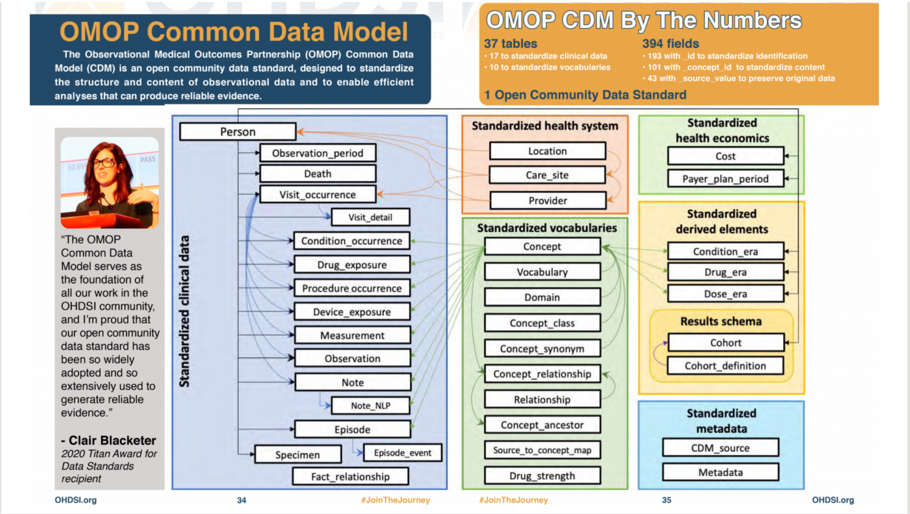
Problem

- EHR-only models may overfit local patterns, ignore external knowledge.
- KG-only models lack patient context, cannot account for comorbidities or contraindications.

Goal

- Unify EHR and KG using **hypergraph-based neural networks**, in a way that fits the **OMOP Common Data Model (CDM)**.

What is OMOP CDM?



Overview of the Project

Based on: HypKG (Xie et al., 2025) – Hypergraph-based Knowledge Graph Contextualization for Precision Healthcare.

My project:

- Study the mathematics behind **hypergraph neural networks** and attention-based message passing.
- Build a **SynPUF1k** hypergraph that respects **OMOP CDM** structure.
- Adapt an AllSet / HypKG-style model to this setting and run **mortality prediction** experiments.
- Compare behaviour with traditional tabular ML and discuss implications for precision healthcare.

Background: Hypergraphs

Graphs vs Hypergraphs

- Graph edge: connects two nodes.
- Hyperedge: can connect *any number* of nodes.

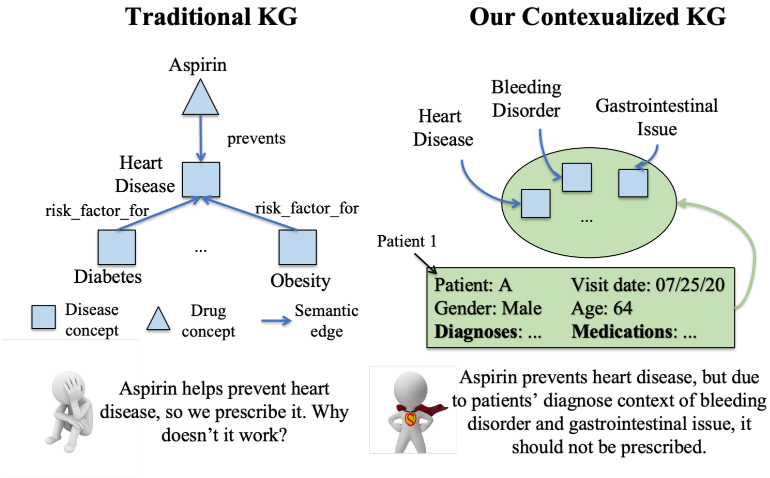
In this work

- Nodes V: medical concepts (diagnoses, drugs, procedures, lab concepts, …).
- Hyperedges E: patients (or visits) aggregating many concepts.

Why hypergraphs?

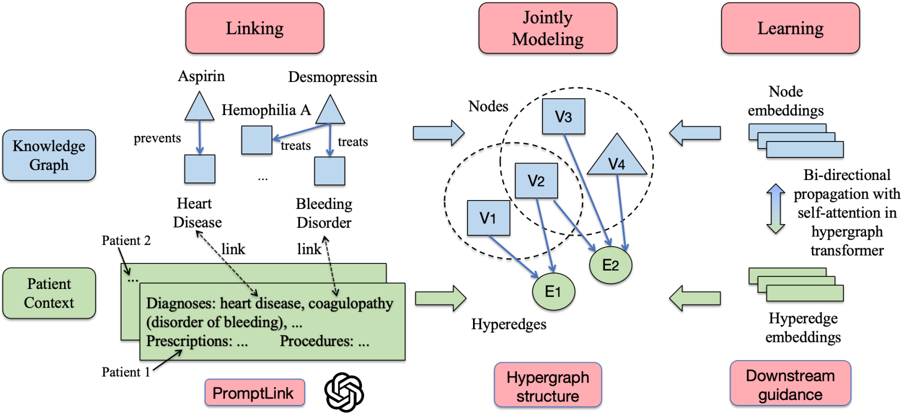
- High-order co-occurrence: e.g. "hypertension + diabetes + CKD + certain drugs".
- Naturally model multi-entity clinical events and multimorbidity patterns.

Conceptual View of HypKG



Idea: Static KG relations may not hold under patient-specific context (age, comorbidities, medication history). HypKG introduces a hypergraph layer that contextualises KG knowledge via patient-specific hyperedges.

HypKG Pipeline



Pipeline:

1. Entity linking (EHR text / codes $\rightarrow$ KG entities).
2. Hypergraph construction (patients / visits as hyperedges).
3. Hypergraph transformer: alternating node–edge attention.

Datasets

- **MIMIC-III**: ICU EHR data (hospital-level), rich temporal features.
- **PROMOTE**: Stroke rehabilitation cohort, functional outcome prediction.
- **SynPUF1k**: Synthetic claims-like OMOP CDM dataset (1k patients).

Why SynPUF1k?

- OMOP-compliant, fully synthetic $\Rightarrow$ safe for open experimentation.
- Mirrors the structure of real claims databases.
- Excellent testbed for adapting hypergraph models to standardized CDM.

Hypergraph Construction on SynPUF1k

Data source: headerless, tab-delimited CSVs for `person`, `visit_occurrence`, `condition_occurrence`, `drug_exposure`, `procedure_occurrence`, `device_exposure`, `measurement`, `observation`, `death`.

Steps:

1. Use DuckDB to assign OMOP column names and query tables.
2. Extract all standard concept IDs from clinical domains.
3. Each unique concept $\to$ node v.
4. Each patient $p \to$ hyperedge e_p containing all concepts that occurred for p.
5. Label $y_p = 1$ if p appears in `death`, otherwise $y_p = 0$.

Evaluation Metrics: AUROC, AUPR, Macro–F1

AUROC (Area Under ROC Curve)

- Measures the model's ability to *rank* positive vs. negative cases.
- Threshold-independent: uses all possible decision boundaries.
- Interpreted as: probability that a randomly chosen positive is ranked higher than a negative. (True positive against false positive)
- **Limitation:** insensitive to class imbalance.

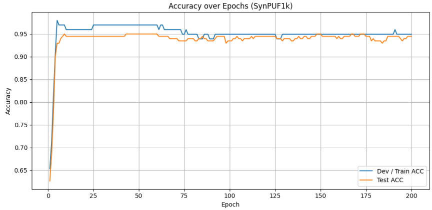
AUPR (Area Under Precision–Recall Curve)

- Focuses on the positive (often rare) class.
- Precision: proportion of predicted positives that are correct.
- Recall: proportion of true positives found.
- **More informative than AUROC** for imbalanced problems (e.g. mortality prediction).

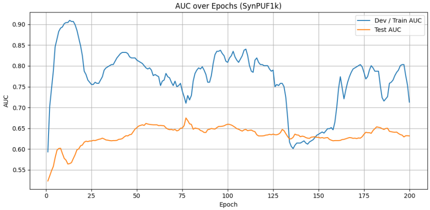
Macro–F1 Score

- F1 = harmonic mean of precision and recall.
- Macro–F1 = average F1 over positive and negative classes.
- Gives **equal weight** to rare and common classes.

Training Dynamics on SynPUF1k (1/2)

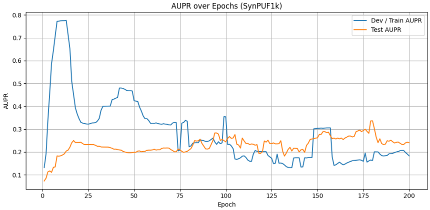


(a) Accuracy over epochs.

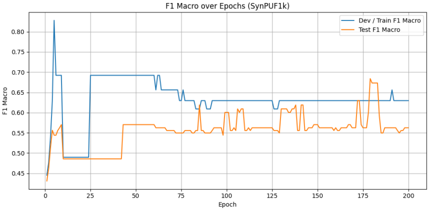


(b) AUC over epochs.

Training Dynamics on SynPUF1k (2/2)



(a) AUPR over epochs.



(b) Macro F1 over epochs.

Final Metrics on SynPUF1k

Metric	Dev (final)	Test (final)
Accuracy (ACC_G)	0.949	0.944
AUC (AUC_G)	0.713	0.632
AUPR (AUPR_G)	0.183	0.241
Macro F1 (F1_MACRO_G)	0.630	0.563

Interpretation:

- High accuracy partly driven by class imbalance.
- AUC and AUPR are clearly above random, indicating useful signal in the hypergraph.
- Dev–test gap is modest $\Rightarrow$ reasonable generalisation.

Hypergraph vs Traditional ML on SynPUF1k

Model	Higher-order structure	AUC (test)
Logistic regression	No (linear)	≈ 0.5–0.6
Tree ensembles (RF/GBM)	Limited (implicit)	up to 0.7–0.9 on rich
HypKG hypergraph	Yes (explicit attention)	0.63 (SynPUF1k

Hypergraph model:

- Encodes concept–patient incidence directly.
- Captures higher-order co-occurrence patterns.
- Naturally compatible with KG initialisation and regularisation.

Limitations and Future Work

Limitations

- Scalability to very large real-world hypergraphs.
- Temporal structure is collapsed; no explicit time modelling yet.
- SynPUF1k experiments used random node embeddings (no real KG embeddings).

Future Directions

- Incorporate KG embeddings (e.g. UMLS, DrugBank) and regularisation.
- Model temporal dynamics via temporal hypergraphs or sequence modules.
- Apply pipeline on real OMOP-based EHR data with privacy-preserving infrastructure.

Thank you!