RWD270

Enhancing Causal Discovery in Chronic Diseases: The MAGIC Framework Using Multiple LLMs

Jihee Kim[1†], Minseol Jang[2, 3†], Miryoung Kim[4, 5], Hyun Jin Han[6], Kangjun Noh[1], Sumin Park[1], Kyungwoo Song[1, 7*], Hae Sun Suh[2, 3, 5*].

1 Department of Statistics and Data Science, Yonsei University, Seoul, Republic of Korea
2 Department of Regulatory Science, Graduate School, Kyung Hee University, Seoul, Republic of Korea
3 Institute of Regulatory Innovation through Science (IRIS), Kyung Hee University, Seoul, Republic of Korea
4 College of Pharmacy, Sunchon National University, Suncheon, Republic of Korea
5 College of Pharmacy, Kyung Hee University, Seoul, Republic of Korea
6 Vista Health Korea Ltd. Seoul, Republic of Korea
7 Department of Applied Statistics, Department of Statistics and Data Science, Yonsei University, Seoul, Republic of Korea

*** Corresponding author**

† These authors contributed equally to this work

Pharmaceutical Economics Big Data Analysis and Policy Lab

BACKGROUND

- Understanding causal relationships among chronic diseases is essential for identifying associations and minimizing bias
- Traditionally, directed acyclic graphs (DAGs) have relied on expert knowledge and literature review, limiting scalability and introducing potential bias
- With the recent advance of large language models (LLMs), it is now possible to explore knowledge-informed DAG construction

OBJECTIVES

- To evaluate the feasibility of LLM-based approaches
- To propose **MAGIC** (**M**ulti-LLM **A**ssisted **G**raph **I**nference and **C**orrection) that integrates statistical, clinical, and language-based feedback to improve DAG generation

METHODS

Two-part study: reference DAG development and comparative performance evaluation of causal discovery methods

Dataset

- **Source:** Korea National Health and Nutrition Examination Survey
- **Sample size:** 36,107 participants
- **Variables:** 13 chronic diseases
- **Evaluation metrics**
  - Skeleton accuracy: evaluate presence of edges (ignoring direction)
  - Orientation accuracy: evaluate correctness of edge directions
  - Structural Hamming Distance (SHD): minimum number of edge changes to match ground truth

Reference DAG: constructed through literature review of 138 peer-reviewed publications and expert consensus

MAGIC FRAMEWORK

Iteratively refines the causal graph through correction and validation

1. **Correction Prompt Generation**
   - Combine current graph structure with **statistical metrics** (phi, BDeu, disease duration using individual data from the Korea National Health and Nutrition Examination Survey) and **clinical knowledge** (via RAG)
   - Formulate prompts to guide LLM reasoning
2. **Graph Correction by Multiple LLMs**
   - Use 5 LLM models (GPT-4o, Claude 3.5 Sonnet, Gemini-2-Flash, DeepSeek R1, Llama 3.3 70B)
   - Each model suggests edge modifications (add, remove, reverse)
   - **Consensus-based voting** aggregates proposals

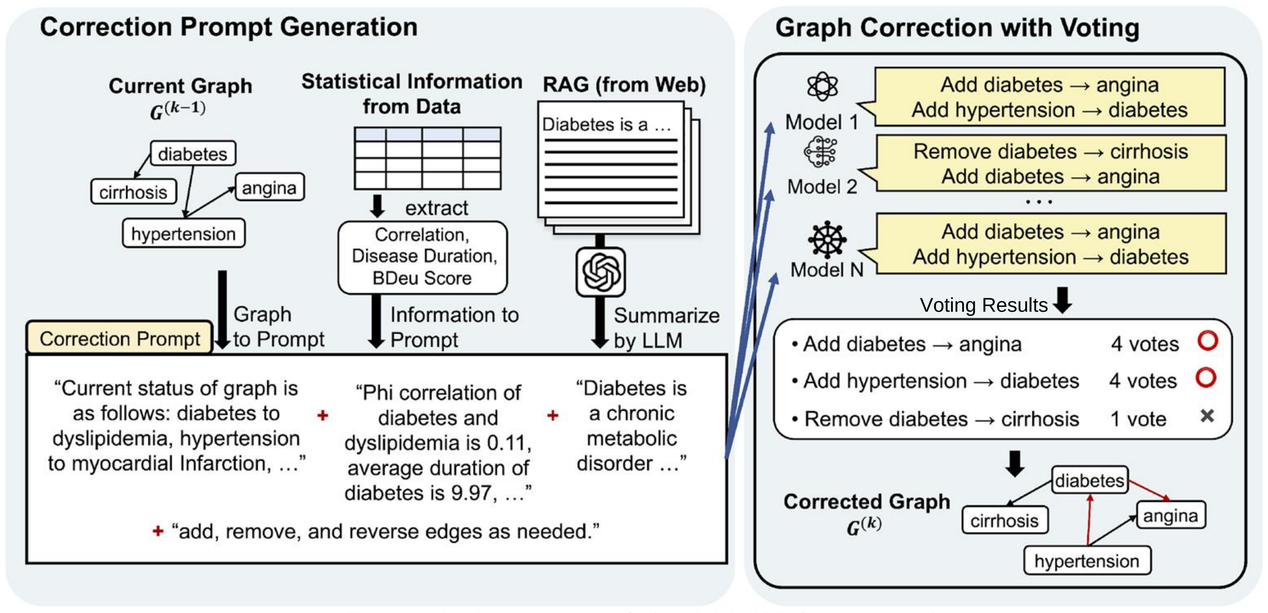


Figure 1. Overview of the MAGIC framework

RESULTS

- After five rounds of expert review and discussion, MAGIC was deemed clinically plausible and methodologically robust.
- MAGIC achieved the best overall performance with skeleton precision 0.941, recall 0.640, F1 score 0.762; orientation precision 0.735, recall 0.500, F1-score 0.595; SHD 27 after the third iteration.
- Performance converged at the third iteration, indicating stable final results.

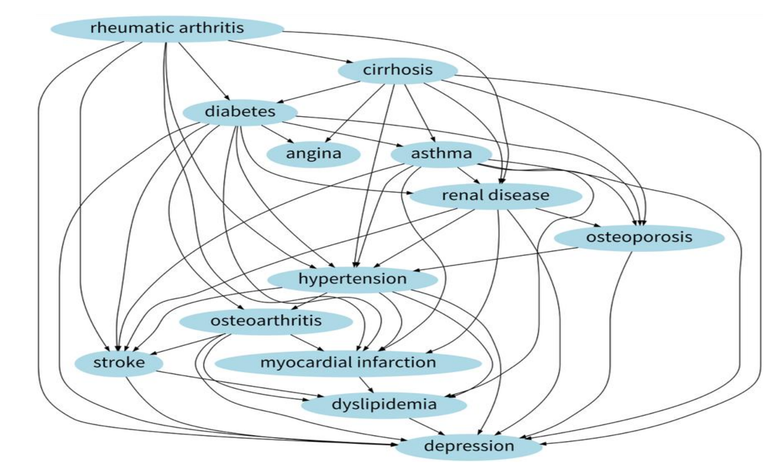
Table 1. Performance comparison of causal discovery methods

Model	Skeleton Precision (%)	Skeleton Recall (%)	Skeleton F1 Score (%)	Orientation Precision (%)	Orientation Recall (%)	Orientation F1 Score (%)	SHD
PC	0.875	0.420	0.568	0.391	0.180	0.247	44
GES	0.826	0.380	0.521	0.278	0.100	0.147	49
LiNGAM	0.905	0.380	0.535	0.571	0.240	0.338	40
LLM_Pairwise	**0.963**	0.520	0.675	**0.778**	0.420	0.546	30
LLM_BFS	0.867	0.260	0.400	0.467	0.140	0.215	45
MAGIC (Ours)	0.941	**0.640**	**0.762**	0.735	**0.500**	**0.595**	**27**

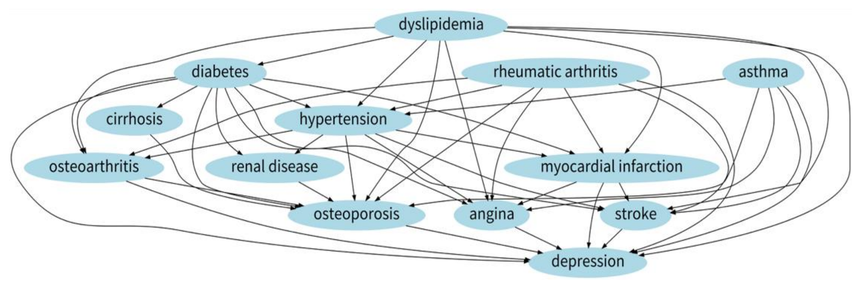
Table 2. MAGIC performance evolution across five self-correction iterations

Iteration	Skeleton Precision (%)	Skeleton Recall (%)	Skeleton F1 Score (%)	Orientation Precision (%)	Orientation Recall (%)	Orientation F1 Score (%)	SHD
1	**0.966**	0.560	0.709	**0.759**	0.440	0.557	29
2	0.939	0.620	0.747	0.727	0.480	0.578	28
3	0.941	**0.640**	**0.762**	0.735	**0.500**	**0.595**	**27**
4	0.941	**0.640**	**0.762**	0.735	**0.500**	**0.595**	**27**
5	0.941	**0.640**	**0.762**	0.735	**0.500**	**0.595**	**27**

Figure 2. DAG Comparison: Reference vs MAGIC



(a) Reference DAG



(b) MAGIC Output DAG

CONCLUSIONS

- MAGIC demonstrates the potential of LLM-guided, feedback-enhanced causal discovery for scalable and reliable causal graph construction.
- By integrating real-world data, clinical context, and multi-model consensus, this approach offers a reproducible and interpretable framework for complex chronic disease research and supports broader applications in healthcare and epidemiology.

ACKNOWLEDGMENT

This research was supported by a grant(RS-2024-00331719, 21153MFDS601) from Ministry of Food and Drug Safety in 2025.