LLM-guided Causal Bayesian Network construction for pediatric patients on ECMO*

Saurabh Mathur[1**], Ranveer Singh[1**], Michael Skinner[1,2], Ethan Sanford[2], Neel Shah[3], Phillip Reeder[2], Lakshmi Raman[2], and Sriraam Natarajan[1]

[1] The University of Texas at Dallas, Richardson, TX, USA
[2] The University of Texas Southwestern Medical Center, Dallas, TX, USA
[3] Washington University in St. Louis, St. Louis, MO, USA

Abstract. Extracorporeal Membrane Oxygenation (ECMO) is a method for supporting patients with severe cardiac or respiratory failure. However, pediatric ECMO patients are at a higher risk of severe neurological injury (NI). Understanding underlying causal mechanisms is critical for clinical decision-making. To this effect, we explore using Large Language Models (LLMs) for the construction of Causal Bayesian networks. While LLMs can reproduce causal relationships reflected in their training data, they may also generate spurious associations. We address this by refining the LLM-generated BN using data from 71 patients and domain constraints elicited from our experts. Our empirical evaluation shows that our method can construct causal diagrams by combining domain knowledge with empirical patterns.

Keywords: Bayesian Networks · LLMs · Theory Refinement · ECMO

1 Introduction

Understanding the causal relationships between adverse health outcomes and their risk factors is crucial for developing effective preventative measures and targeted interventions. One important formalism for causal modeling is Causal Bayesian Networks (CBN [9]). The structure of CBNs is an acyclic graph consisting of a set of vertices, one for each variable, and a set of directed edges between pairs of variables, each edge denoting a causal relationship. We consider the task of constructing these causal graphs to model pediatric patients on extracorporeal membrane oxygenation (ECMO [6]). ECMO is a method of supporting critically ill patients with cardiac or pulmonary failure by using a heart-lung bypass until the heart or lungs recover function. For patients on ECMO, we aim to model the relationship between the presence of neurologic injuries, such as strokes, intracranial bleeding, and brain death, and their relevant risk factors.

Causal graph construction methods can be categorized into expert knowledge-based and data-driven methods. Expert knowledge-based methods construct

* We acknowledge support from NIH award R01NS133142.
** Equal contribution

causal graphs by eliciting the causal relationship from domain experts or from clinical guidelines such as the Quick Medical Reference (QMR [13]). This approach falls short when modeling exceedingly complex and less well-understood domains such as pediatric critical care. In contrast, data-driven methods aim to automatically construct these graphs based on patterns mined from empirical data [16]. Since causality cannot be established purely from observation, these methods make assumptions about the underlying causal relationships. For instance, data-driven methods commonly assume *causal sufficiency*, meaning that the causal relations between the set of variables under consideration are not mediated by an external or confounding factor. While this might be addressed by naively expanding the set of variables, the paucity of medical data limits the number of variables that could be considered. Data-driven causal discovery is further complicated by the temporal nature of the data, missingness due to data collection issues, and the naturally cyclic nature of human physiology [3]. Thus, this domain poses a number of challenges to causal learning methods.

We address these challenges in causal graph construction for critical care pediatric patients by taking a hybrid approach, combining data-driven methods with knowledge from two sources: critical care experts and a Large Language Model (LLM) trained on a large corpus.

2 Refinement based causal learning for pediatric ECMO

We focus on the task of constructing the structure of a CBN, i.e., a causal graph, to model the less well-understood and complex domain of pediatric patients on ECMO. This graph would help improve the understanding of the relationships between NI and its risk factors. We aim to construct the causal graph by combining domain knowledge with a small data set from patients on ECMO.

In particular, using the theory refinement framework [8], we construct BN graphs by eliciting an initial graph from an expert and iteratively adding, removing, or reversing directed edges between variables to maximize an empirical score. This method is based on hill-climbing search, approximately solving an exceedingly complex search problem by finding locally optimal solutions. A key issue is that the initial structure will greatly affect the efficacy of the final model. Moreover, without additional assumptions, it cannot always distinguish purely associational relationships from causal ones in the data.

We adapt the BN refinement procedure to causally model pediatric patients in the critical care setting by exploiting domain knowledge. First, using expert knowledge, we collect a set of Boolean variables summarizing clinically relevant information about each patient's ECMO run. Based on these variables, we enumerate all the causal relations that would be clinically impossible. For instance, NI cannot cause hypertension in the first 24 hours of ECMO. Second, we use a pretrained LLM as a source of approximate knowledge. These are deep generative models that have demonstrated impressive capabilities across a wide range of natural language processing tasks [18]. Since these models have been trained on a vast corpus that includes a large amount of medical literature, they offer a valuable source of approximate domain knowledge [7]. However, since these

are correlation-based models, they cannot distinguish between genuinely causal and purely associational relations [17]. As a result, their generated text might contain a combination of causal and non-causal statements. We use this text to construct an initial graph that would be edited through theory refinement.

To summarize, we consider knowledge in the form of expert selected variables, constraints on their causal relationships based on temporal order, and an initial hypothesis about the causal graph based on approximate domain knowledge from an LLM. Using this knowledge, we construct a causal graph using the data driven theory refinement procedure to eliminate non-causal edges from the LLM-generated causal graph. Concretely, our method is a three-phase procedure inspired by the Greedy Equivalence Search algorithm [2]. We first populate a causal graph by prompting an LLM, and then refine this graph by eliminating edges unsupported by the data via a refinement procedure limited to deletion operations. Finally, we discover new edges by performing full, unrestricted refinement on the resulting graph[4].

3 Empirical evaluation

Setup. We evaluate our method using a small data set obtained from the Children's Medical Center of Dallas. It consists of time-series data from the ECMO runs of 71 pediatric patients (ages 0–19 years, with an average age of 4.32 years). Our domain experts, Drs. Raman, Shah, and Sanford defined the causal modeling domain in terms of eight Boolean variables. This set includes a variable indicating whether the patient suffered a neurological injury (NI) based on tests performed at the end of the ECMO run, and seven variables indicating the presence of seven risk factors of NI, based on existing clinical research [12,5,15]: *High VIS, Hypotension, Hypertension, Low Platelet, High Lactate, Low pH, and Relative pCO2*. The presence of these risk factors early in the ECMO run has been associated with adverse outcomes. As a result, we define these variables based on the data from the first 24 hours post-cannulation on ECMO. We computed the values for each risk factor variable for each patient by testing if its corresponding event (e.g., Low pH) occurs at least once within the patient's data. Additionally, we compute *Relative pCO2* as a large change in the pCO2 value after initiating EMCO relative to the value one hour prior to cannulation[11].

We used 4 pre-trained LLMs as approximate knowledge sources: GPT-4o, DeepSeek, LLaMA, and Gemini. We compare the LLM-generated graphs and the graphs obtained from our refinement procedure to 3 data-driven baselines: search-and-score (**SS**) [4], **PC** [14], and **FCI** [14]. For each automatically constructed graph, we quantify the difference from an expert-constructed graph using 3 metrics: the number of Spurious Edges (**SE**), the Structural Hamming Distance (**SHD** [1]), and the Structural Intervention Distance (**SID**[10]). SHD is the number of edge additions, deletions, and reversals required to transform each graph into the expert graph, while SID quantifies the divergence in causal conclusions derived from the graphs.

[4] Supplementary material: `https://github.com/saurabhmathur96/LLM-guided-CBNs`

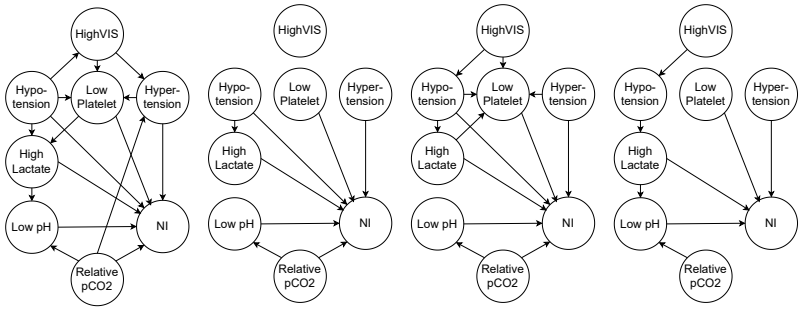

Fig. 1. Causal graphs constructed by pooling output of all the LLMs, after deletion-only refinement, after full-refinement, and the expert-constructed graph (left to right)

Results. Tables 1 and 2 present quantitative results of the evaluation, aggregated over 10 bootstrap samples. **(1)** LLMs construct graphs that are semantically closer to the expert graph than data-driven graphs; they, however, contain more spurious edges. **(2)** Deletion-only refinement correctly eliminates spurious edges; it also eliminates some weak causal edges. **(3)** Full refinement yields graphs with more spurious edges; these edges are supported by empirical data and a partial causal structure, making them more plausible than the LLM-generated ones. Our domain experts' visual inspection of these graphs validated the plausibility of these edges. For instance, consider the graphs in Figure 1. The LLM's graph misses the edge HighVIS → Hypotension; the full refinement procedure correctly recovers that edge from data and adds it to the graph. Additionally, our method added the edge Relative pCO2 → NI; in retrospect, the experts agree that this may be supported by the literature.

Method	**SE**	**SHD**	**SID**
SS	4.3 ± 2	9.7 ± 2	17.9 ± 5
PC	0.8 ± 0.9	8.1 ± 1.4	16.5 ± 4
FCI	0.1 ± 0.3	8 ± 0.4	14.9 ± 0.3
LLM Union	9	9	5
+ Refine (del.)	3 ± 0.9	5.4 ± 1.1	5.1 ± 1.4
+ Refine (full)	4.8 ± 1.1	6.6 ± 1.3	6.2 ± 3.5

Table 1. Data-driven vs LLM-based methods.

Limitations and Future work. We considered the problem of modeling neurological injuries in the pediatric population on ECMO. We take the first step towards tackling this hard problem by exploiting exact and approximate domain knowledge to augment sparse data to construct a causal graph. There are three main directions for future work. Firstly, this study is based on data from only 71 patients from a single center, of which 23.9 % suffered a neurological injury. Future work should consider larger data sets from multiple centers. Secondly, the 8 variables considered in the study were chosen based on the latest clinical research. However, this might not strictly satisfy causal sufficiency since the domain is still being actively researched. Future work should include additional factors such as ECMO pump speed and medical conditions that might mediate the causal relationships between these variables. Finally, future work should extend this method to causal relations across time.

LLM	LLM output		Refine (del.)		Refine (full)	
	SE	SHD	SE	SHD	SE	SHD
GPT	6	6	2.6 ± 0.5	4.9 ± 0.7	4.7 ± 0.9	6.5 ± 1.1
Gemini	6	8	1 ± 0.9	7.1 ± 1.4	3.2 ± 1.3	8.7 ± 2.1
LLaMA	4	5	0.1 ± 0.3	6.1 ± 0.8	2.7 ± 1.0	8.1 ± 1.2
DeepSeek	7	7	1.6 ± 0.7	3.1 ± 0.9	4.2 ± 1.4	5.9 ± 1.7
LLM Union	9	9	3 ± 0.9	5.4 ± 1.1	4.8 ± 1.1	6.6 ± 1.3

Table 2. Effect of CBN refinement on LLM-elicited causal graphs.

References

1. Acid, S., De Campos, L.M.: Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs. JAIR **18**, 445–490 (May 2003)
2. Chickering, D.M.: Optimal Structure Identification With Greedy Search. JMLR (2002)
3. Claassen, J.A., et al.: Regulation of cerebral blood flow in humans: physiology and clinical implications of autoregulation. Physiological reviews (2021)
4. Koller, D., Friedman, N.: Probabilistic Graphical Models. MIT press (2009)
5. LaRovere, K.L., Vonberg, F.W., et al.: Patterns of head computed tomography abnormalities during pediatric extracorporeal membrane oxygenation and association with outcomes. Pediatric neurology **73** (2017)
6. Lin, J.C.: Extracorporeal membrane oxygenation for severe pediatric respiratory failure. Respiratory care **62**(6), 732–750 (2017)
7. Mathur, S., et al.: Modeling Multiple Adverse Pregnancy Outcomes: Learning from Diverse Data Sources. In: AIME (2024)
8. Mooney, R.J., Shavlik, J.W.: A Recap of Early Work on Theory and Knowledge Refinement. In: AAAI Spring Symposium (2021)
9. Pearl, J.: Causality. Cambridge University Press (2009)
10. Peters, J., Bühlmann, P.: Structural Intervention Distance for Evaluating Causal Graphs. Neural Computation **27**(3), 771–799 (2015)
11. Shah, N., et al.: Early Changes in Arterial Partial Pressure of Carbon Dioxide and Blood Pressure After Starting Extracorporeal Membrane Oxygenation in Children: Extracorporeal Life Support Organization Database Study of Neurologic Complications. Pediatric Critical Care Medicine **24**(7), 541–550 (2023)
12. Shah, N., et al.: Neurologic Statistical Prognostication and Risk Assessment for Kids on Extracorporeal Membrane Oxygenation — Neuro SPARK. ASAIO Journal (2024)
13. Shwe, M.A., et al.: Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base. Methods of information in Medicine **30**(04) (1991)
14. Spirtes, P., et al.: Causation, Prediction, and Search. MIT press (2001)
15. Wood, S., Iacobelli, R., Kopfer, S., Lindblad, C., Thelin, E.P., Fletcher-Sandersjöö, A., Broman, L.M.: Predictors of intracranial hemorrhage in neonatal patients on extracorporeal membrane oxygenation. Scientific Reports **13**(1), 19249 (2023)
16. Zanga, A., et al.: A Survey on Causal Discovery: Theory and Practice. IJAR (2022)
17. Zecevic, M., et al.: Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. TMLR (2023)
18. Zhao, W.X., et al.: A Survey of Large Language Models. arXiv preprint arXiv:2303.18223 (2023)