Modeling multiple adverse pregnancy outcomes: Learning from diverse data sources

Saurabh Mathur[1], Veerendra P. Gadekar[2], Rashika Ramola[3], Avery Wang[3], Ramachandran Thiruvengadam[4], David M. Haas[5], Shinjini Bhatnagar[6], Nitya Wadhwa[6], Garbhini Study Group[6], Predrag Radivojac[3], Himanshu Sinha[2], Kristian Kersting[7], and Sriraam Natarajan[1]

[1] The University of Texas at Dallas, Richardson, Texas, USA
[2] Indian Institute of Technology Madras, Chennai, Tamil Nadu, India
[3] Northeastern University, Boston, Massachusetts, USA
[4] Pondicherry Institute of Medical Sciences, Puducherry, India
[5] Indiana University School of Medicine, Indianapolis, Indiana, USA
[6] Translational Health Sciences and Technology Institute, Faridabad, India
[7] Technische Universität Darmstadt, Darmstadt, Germany

Abstract. We consider the problem of modeling adverse pregnancy outcomes (APOs) from diverse data sets and aim to understand what is common between them and what is unique for each of these data sets. To this effect, we consider three different data sets (a clinical study from the US, EHRs from a US hospital, and a clinical study in India) and model three specific APOs - preterm birth, new hypertension, and preeclampsia. Since LLMs can efficiently summarize the scientific literature, we use them to generate initial hypotheses and use the different data sets to refine the hypotheses to create joint probabilistic models (as Bayesian networks). Our analyses show that there are eight relationships between risk factors common to all three populations and some unique relationships for specific populations.

Keywords: Bayesian Networks · Theory Refinement · LLMs.

1 Introduction

Adverse Pregnancy Outcomes (APOs) such as preterm birth (PTB) pose a significant challenge in maternal-child health, with approximately one in ten births occurring prematurely on a global scale. The implications of PTB extend beyond immediate neonatal mortality, influencing both short-term and long-term health outcomes [17]. However, the relationship between APOs and their risk factors can vary across geographical regions [9]. This makes integration and analysis of multiple data sets vital to understanding APOs and mitigating their risk.

We aim to model the differences and commonalities between data sets of APOs from different countries. Specifically, we aim to perform this analysis by inducing interpretable probabilistic models from three data sets from 2 countries, namely India (Garbh-Ini [1]) and the United States (nuMoM2b [13] and EHR

data from Regenstrief Institute). This would help advance our understanding of the multifaceted nature of APOs and potentially inform targeted interventions tailored to specific geographical regions.

Probabilistic graphical models such as Bayesian networks [19, 14] have long been used in AI for modeling interactions of multiple factors by learning joint distributions. In contrast to discriminative learning methods where the goal is to best predict an outcome, these generative models learn a joint distribution that can allow us to query comprehensively and understand the data in a more holistic manner. The biggest barrier to learning these models is the amount of data required which can be offset by using domain knowledge to construct an initial model and refining this model using the data.

Consequently, we employ the use of LLMs to generate an initial model (since LLMs can efficiently summarize the literature), refine the model with domain experts, and then use each of the data separately to refine the models for the respective populations. Once these different models are obtained, we perform meta-analyses of these models and summarize the findings. The common influence relationships that exist in all the data sets are between the risk factors BMI and HiBP and the three APOs new hypertension (NewHTN), preeclampsia (PreEc), and preterm birth (PTB). We also present the edges that are unique to each of these subpopulations (for instance, age is important in nuMoM2b but is not as influential in Garbh-Ini). Our hypothesis is that given such a unified yet diverse view, it is now possible to develop population-specific treatment plans for mitigating the APOs.

1.1 Data description

nuMoM2b: The nuMoM2b (Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be [13]) study focuses on identifying risk factors for APOs in the United States. It enrolled a diverse cohort of 10,038 nulliparous subjects across 8 US sites. Data collection occurred at the start of pregnancy and at subsequent visits throughout the pregnancy.

Electronic Health Records: Apart from the data from the nuMoM2b study, we also acquired Electronic Health Records (EHR) from the Regenstrief Institute. This data set includes non-nulliparous subjects but does not include information about family history of chronic conditions.

Garbh-Ini: The Garbh-Ini study [1] conducted in a single site within Haryana, India, aims to characterize PTB and identify associated risk factors. It enrolled 8,050 subjects both nulliparous and non-nulliparous, and collected data at the start of pregnancy and at subsequent visits throughout the pregnancy.

2 Background

Bayesian Networks (BNs [19]) are a class of Probabilistic Graphical Models (PGMs [14]) that factorize the joint distribution over a set of variables using a Directed Acyclic Graph (DAG) and local conditional probability distributions (CPDs). The DAG has a node corresponding to each variable and a directed edge between nodes represents influence. For example, an edge Age $\rightarrow$ PTB would

imply that the age of the subject at pregnancy influences our belief about the likelihood of preterm birth. The local CPDs quantify the influence in terms of probability values. Formally, a BN $\mathcal{M}$ over a set of n variables $V = \{X_1, \ldots, X_n\}$ is defined as the tuple $\langle \mathcal{G}, \theta \rangle$ where $\mathcal{G}$ is the DAG representing the structure of the BN and θ is the set of parameters for the local CPDs. The joint probability distribution over V defined by the BN is

$$P(X_1, \ldots, X_n) = \prod_{X \in V} P_\theta(X \mid \mathrm{Pa}_X) \tag{1}$$

where Pa_X is the set of parents of the BN node corresponding to variable X. BNs can reason under uncertainty and answer probabilistic queries about the variables. Additionally, since BNs consist of directed influences between variables and local conditional probabilities, they are easy to interpret.

The structure of the BN encodes conditional independence relations (CIs) between variables; each variable X is independent of its nondescendents given its parents Pa_X. These two properties – reasoning under uncertainty and interpretability make BNs a good fit for high-stakes domains such as healthcare that require models that can reason about complex relationships between variables while being able to develop trust with domain experts.

In this work, we induce BNs from each of the 3 data sets and compare the influence relations between APOs and their risk factors. However, inducing the structure of a BN directly from data is a data-hungry and computationally hard problem [7]. One approach to mitigate this problem is Theory Refinement [16]. This approach involves constructing an initial BN structure from domain knowledge and then refining this BN using data. Specifically, the BN is refined by performing local operations such as adding an edge, deleting an edge, and reversing an edge to maximize a given heuristic score. Commonly used scores include the Minimal Description Length(MDL [15]) and Bayesian-Dirichlet Scores(BD [6]). The MDL score can be adapted to exploit local structure [12] in the form of context-specific independence relations (CSIs [3]) if the local conditional distributions of the BN are represented as decision trees. While prior works obtain the initial BN from a domain expert, we aim to construct the initial BN by extracting approximate domain knowledge from a deep generative model.

Large Language Models as approximate knowledge sources: LLMs [26] are a class of deep generative models for text data. They consist of two Artificial Neural Networks (ANNs) called an encoder and a decoder. These encoder and decoder ANNs are used to encode input prompt text from a user and to generate response text from the encoded prompt respectively. Examples of such LLMs include General Purpose Transformer (GPT [5]) and Gemini [23]. These models are fit using large amounts of textual data and have been shown to generate realistic text. However, they cannot reason about the information embedded in them [25]. As a result, prior work has tried to extract knowledge from existing LLMs and inject the knowledge into models that can perform reasoning [20, 18, 10]. Inspired by these directions, we extract knowledge in the form of influence relations from an LLM, use this knowledge to instantiate a BN, and then refine the BN using clinical data.

3 Methodology

We aim to find the relationships between variables common across the three data sets, and the ones unique to each data set. We formalize this task as the following problem

Given: Data sets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ over a set of variables $\boldsymbol{X}$, and an LLM $\mathcal{O}$
To Do: Find a set of pairwise influences that are supported in all 3 data sets and the sets of influences supported only in particular data sets.

We address the problem of identifying consistent and dataset-specific relationships by learning three BN structures and then comparing them. We identify two types of edges, common edges, present in all refined BNs, which represent pairwise influences consistently supported by all data sets; and dataset-specific edges, unique to a specific BN, which represent pairwise influences supported only in the corresponding data set.

However, learning a BN structure from data is a difficult problem. Firstly, BNs are highly expressive models, and finding the structure that maximizes the data likelihood would result in an overly complex BN that overfits that training data. To address this, BN structures are learned by minimizing a cost function that includes implicit [6] or explicit [15] regularization. Secondly, even with a cost function (say $\text{Cost}(\mathcal{M}, \mathcal{D})$), learning the structure of a BN requires us to solve the following combinatorial optimization problem:

$$\arg\min_{\mathcal{M}} \text{Cost}(\mathcal{M}, \mathcal{D}) \tag{2}$$

This problem requires a search over a superexponential number of BN graph structures. Not only is searching over such a large space computationally intractable (NP-Hard to be specific), but it also requires a large amount of data to be able to determine the optimal structure[7].

One way to make this problem tractable is to exploit domain knowledge. We encode domain knowledge in three ways. Firstly, we encode domain knowledge through the choice of scoring function. Specifically, we use the MDL scoring function that prefers concise structures over more complex structures through an explicit penalty term. Secondly, we use domain knowledge about relations between the variables to construct an initial BN structure. While such BN structures are generally elicited from domain experts, we obtain the initial structure by querying an LLM. We further restrict the search space by using domain knowledge to identify and exclude temporally impossible edges. For instance, the edge PTB $\rightarrow$ BMI is invalid as preterm birth cannot influence Body Mass Index (BMI) measured at the pregnancy's start. By incorporating domain knowledge, we restrict the search from an exhaustive exploration of all BN structures to a local search over the structures in the neighborhood of an initial structure obtained from an LLM.

3.1 BN refinement using the MDL Score

We refine the initial BN structure for each data set by minimizing the MDL score. The MDL score for a BN (denoted by $\mathcal{M}$) with respect to a data set

(denoted by $\mathcal{D}$) is the sum of the description length of the data encoded using the BN model ($\mathrm{DL}(D \mid \mathcal{M})$) and the description length of the BN model itself ($\mathrm{DL}(\mathcal{M})$). Concretely, the MDL score is given as

$$\mathrm{MDL}(\mathcal{M}; \mathcal{D}) = \mathrm{DL}(D \mid \mathcal{M}) + \mathrm{DL}(\mathcal{M}) \tag{3}$$

The first term is the description length of the encoded data and captures the number of bits required to encode the data points using the probabilities estimated by the BN model. The second term is the description length of the model and captures the complexity of the BN itself. Since Huffman coding allows data points to be encoded using their probabilities, $\mathrm{DL}(\mathcal{D} \mid \mathcal{M})$ is approximated by the negative log-likelihood of the data set under the BN. The description length (DL) of a BN, denoted by DL(M), captures the complexity of the model. It consists of two components, the description length of the graphical structure of the BN $\mathcal{G}$ and that of the parameters of the local conditional distributions θ.

3.2 Encoding the BN model

Description Length of the Graphical Structure. This term represents the space required to encode the BN's structure G. Each node's description includes the number of parents and their names. Since each node can be encoded in $\log n$ units of space, the description length of the structure is $\sum_{X \in V}(1 + |\mathrm{Pa}_X|) \log n$.

Description Length of the Parameters. This term represents the space required to encode the parameters, θ. These parameters define the local CPDs over each node given its parents. There are two ways to encode these distributions, as tables and as trees. Conditional Probability Tables (CPTs) explicitly enumerate the conditional probability values corresponding to each parent configuration. Each entry in a CPT can be encoded as an ordered list of fixed-width floating-point values, each of which can be encoded in space $\frac{1}{2} \log N$, where N is the size of the data set. The resulting description length for all the CPTs of the BN is $\sum_{X \in V}(|X| - 1)|\mathrm{Pa}_X|(\frac{1}{2} \log N)$.

Local conditional distributions can be represented as trees to exploit local structure [12] in the form of CSIs [3]. The description length of such a tree-structured local conditional distribution over a variable X given its parents is $B(|X|-1)(\frac{1}{2} \log N) + \sum_{l=1}^{d} \log(|\mathrm{Pa}_X| - A_l)$. Here, B is the number of leaf nodes, d is the depth of the tree and A_l is the number of internal nodes at level l.

3.3 Computing the CSI-aware MDL Score

To account for CSIs in the MDL score we use the Classification and Regression Trees (CART [4]) algorithm. At each node, we fit a decision tree to predict the node's value from its parents. This decision tree serves as the tree-structured CPD for computing the MDL score. The overall MDL score is given by the

following equation:

$$\begin{aligned}\mathrm{MDL}(M;\mathcal{D}) = &-\sum_{\boldsymbol{x}\in\mathcal{D}} \log P_{\mathcal{M}}(\boldsymbol{x}) + \sum_{X\in V}(1+|\mathrm{Pa}_X|)\log n \\ &+ \sum_{X\in V} B_X(|X|-1)(\frac{1}{2}\log N) + \sum_{l=1}^{d}\log(|\mathrm{Pa}_X| - A_{X_l})\end{aligned} \tag{4}$$

where B_X and A_{X_l} are the number of leaf nodes and the number of internal nodes at level l for the decision tree fit for node X respectively.

4 Experimental evaluation

We consider 3 APOs, namely, New Hypertension (NewHTN), Preeclampsia (PreEc), and Pre-term birth (PTB), and study their relationship with 5 risk factors from prior work [8]. Specifically, the risk factors include Family History of diabetes (Hist), Age at the start of pregnancy (Age), Body Mass Index at the start of pregnancy (BMI), presence of Hypertension at the start of pregnancy (HiBP), and Parity. Of these variables, Parity does not apply to nuMoM2b as the study selected nulliparous subjects (Parity = 0) and Hist was not available in the EHR data. We removed data points that had missing values for any of the considered variables. Table 1 summarizes the variables, their discrete values, and the corresponding proportions in each of the three data sets.

We obtained a set of edges from Gemini to construct an initial BN structure and then refined this structure for each of the three data sets. Figure 1 shows the initial BN obtained from Gemini, the edges common to all the refined BNs, and the edges unique to each of the three data sets[8]. Apart from these, the edges $\{\text{Age} \to \text{Parity}, \text{Parity} \to \text{PTB}, \text{Parity} \to \text{PreEc}\}$ were present in both the data sets that had the Parity variables available (Garbh-Ini and EHR).

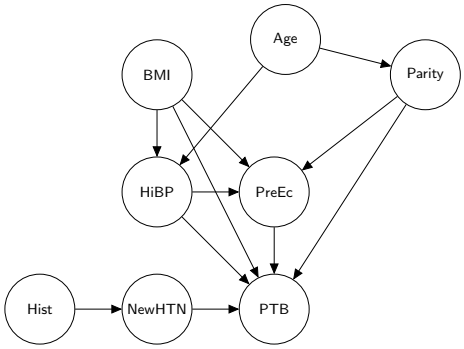
The edges common to all three refined BNs reflect existing domain knowledge. High BMI is known to increase the risk of Hypertensive disorders of pregnancy such as preeclampsia and new hypertension [21, 2]. Hypertensive disorders of pregnancy are known to increase the risk of preterm birth [24]. Finally, hypertension at the start of pregnancy (HiBP) and new hypertension are known risk factors for preeclampsia [11].

The edge from BMI to Parity in the BN learned from the EHR data might reflect the fact that high obesity negatively influences fertility [22]. This edge is supported by the EHR data which has the largest proportion of high BMI subjects. While BMI is expected to rise with an increase in Age, the influence relation is unique to the BN learned from the Garbh-Ini data set.

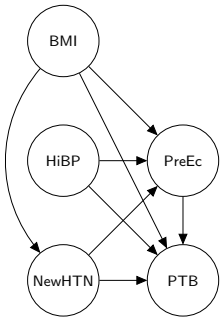
5 Discussion

A few important differences between the populations need to be pointed out. First, while the nuMoM2b study studied nulliparous subjects (first-time moth-

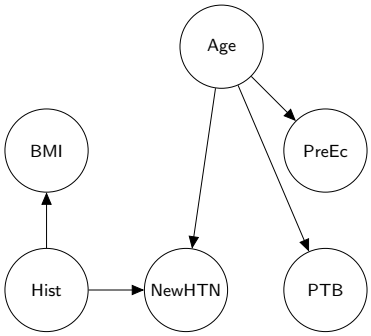
[8] The code for the experiments, the LLM prompt, and the list of temporally impossible edges is available at `https://github.com/saurabhmathur96/BN-Refinement`



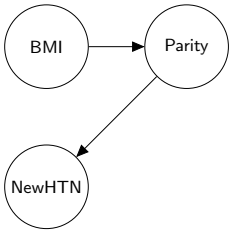
(a) Initial BN obtained from an LLM



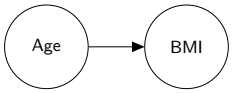
(b) Edges common to all the refined BNs



(c) Edges unique to nuMoM2b



(d) Edges unique to EHR



(e) Edges unique to Garbh-Ini

Fig. 1: The initial BN structure obtained from an LLM (a), edges common to the BNs refined on all 3 data sets (b), edges unique to nuMoM2b (c), EHR (d) and (e)

Variable	Value	nuMoM2b	Garbh-Ini	EHR
Age	≤ 21	21.03%	31.54%	9.87%
	21-35	72.36%	67.78%	75.27%
	>35	6.61%	0.67%	14.86%
BMI	≤ 18	3.39%	19.64%	1.12%
	18-25	51.29%	67.13%	31.53%
	>25	45.31%	13.22%	67.35%
Parity	=0	100%	48.50%	6.3%
	0-2	N/A	47.12%	67.36%
	>2	N/A	4.38%	26.34%
Hist	TRUE	20.55%	8.10%	N/A
HiBP	TRUE	2.84%	2.10%	9.37%
PReEc	TRUE	5.85%	3.80%	7.54%
NewHTN	TRUE	16.12%	3.40%	11.09%
PTB	TRUE	8.11%	12.80%	9.41%
Total		9,368	4,159	16,487

Table 1: Variable-value proportions for each of the three data sets

ers), there were no such restrictions in the other two datasets. Second, the common risk factors and APOs were chosen across the different data for the purposes of this study. Consequently, APOs such as gestational diabetes were not considered as they were computed differently in the Garbh-Ini study. Thus, some of the relationships such as the influence of family history might include some hidden confounders (such as gestational diabetes). Exploring these issues remains an open problem. Finally, a variable such as race, a social construct, which plays an important role in a diverse dataset such as the EHR is not considered due to its absence in the single-state study in India.

Nonetheless, several common themes emerged. The influence of HiBP and BMI is quite significant across populations and data sets. It is clear that in nuMoM2b participants, age has a direct influence on PTB while in Garbh-Ini participants, age directly influences BMI (potentially through multiple pregnancies). It is important to understand the key differences in the data itself and these models provide a way of doing that. Future research could explore several avenues, including incorporating more data sets, identifying hidden confounders, understanding the similarities and differences in population, and extending these analyses to more global data sets. Finally, integrating multi-omic data, such as gene expression and proteomics data, alongside clinical data from diverse sources could offer deeper insights into the molecular pathways underlying APOs.

Acknowledgements: SM, RR, AW, DH, PR and SN acknowledge the support by the NIH grant R01HD101246. VG, RT and HS thank the Centre for Integrative Biology and Systems Medicine (IBSE) and Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI), IIT Madras for their support. KK acknowledges support by the European Union grant agreement no. 101120763 - TANGO.

Bibliography

[1] Bhatnagar, S., Majumder, P.P., Salunke, D.M., for Advanced Research on Birth Outcomes—DBT India Initiative (GARBH-Ini), I.G.: A pregnancy cohort to study multidimensional correlates of preterm birth in india: study design, implementation, and baseline characteristics of the participants. American Journal of Epidemiology **188**(4), 621–631 (2019)

[2] Bohiltea, R.E., Zugravu, C.A., Nemescu, D., Turcan, N., Paulet, F.P., Gherghiceanu, F., Ducu, I., Cirstoiu, M.M.: Impact of obesity on the prognosis of hypertensive disorders in pregnancy. Experimental and Therapeutic Medicine **20**(3), 2423–2428 (2020)

[3] Boutilier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in bayesian networks. In: Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence. pp. 115–123 (1996)

[4] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Wadsworth (1984)

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

[6] Buntine, W.: Theory refinement on bayesian networks. In: UAI, pp. 52–60. Elsevier (1991)

[7] Chickering, M., Heckerman, D., Meek, C.: Large-sample learning of bayesian networks is np-hard. Journal of Machine Learning Research **5**, 1287–1330 (2004)

[8] Chu, H., Ramola, R., Jain, S., Haas, D.M., Natarajan, S., Radivojac, P.: Using association rules to understand the risk of adverse pregnancy outcomes in a diverse population. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2023: Kohala Coast, Hawaii, USA, 3–7 January 2023. pp. 209–220. World Scientific (2022)

[9] Dhaded, S.M., Saleem, S., Goudar, S.S., Tikmani, S.S., Hwang, K., Guruprasad, G., Aradhya, G.H., Kusagur, V.B., Patil, L.G.C., Yogeshkumar, S., et al.: The causes of preterm neonatal deaths in india and pakistan (purpose): a prospective cohort study. The Lancet Global Health **10**(11), e1575–e1581 (2022)

[10] Dietterich, T.: What's Wrong with Large Language Models and What We Should Be Building Instead. (2024)

[11] Dimitriadis, E., Rolnik, D.L., Zhou, W., Estrada-Gutierrez, G., Koga, K., Francisco, R.P.V., Whitehead, C., Hyett, J., da Silva Costa, F., Nicolaides, K., Menkhorst, E.: Pre-eclampsia. Nature Reviews Disease Primers **9**(1), 1–22 (Feb 2023). https://doi.org/10.1038/s41572-023-00417-6

[12] Friedman, N., Goldszmidt, M.: Learning bayesian networks with local structure. In: Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence. pp. 252–262 (1996)

[13] Haas, D.M., Parker, C.B., Wing, D.A., Parry, S., Grobman, W.A., Mercer, B.M., Simhan, H.N., Hoffman, M.K., Silver, R.M., Wadhwa, P., et al.: A description of the methods of the nulliparous pregnancy outcomes study: monitoring mothers-to-be (numom2b). American journal of obstetrics and gynecology **212**(4), 539–e1 (2015)

[14] Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)

[15] Lam, W., Bacchus, F.: Using causal information and local measures to learn bayesian networks. In: UAI. pp. 243–250. Elsevier (1993)
[16] Mooney, R.J., Shavlik, J.W.: A recap of early work on theory and knowledge refinement. In: AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (2021)
[17] Ohuma, E.O., Moller, A.B., Bradley, E., Chakwera, S., Hussain-Alkhateeb, L., Lewin, A., Okwaraji, Y.B., Mahanani, W.R., Johansson, E.W., Lavin, T., et al.: National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. The Lancet **402**(10409), 1261–1271 (2023)
[18] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X.: Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering (2024)
[19] Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann (1988)
[20] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.: Language models as knowledge bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2463–2473 (2019)
[21] Roush, G.C.: Obesity-induced hypertension: Heavy on the accelerator (2019)
[22] Silvestris, E., De Pergola, G., Rosania, R., Loverro, G.: Obesity as disruptor of the female fertility. Reproductive Biology and Endocrinology **16**, 1–13 (2018)
[23] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
[24] Wilson, D.A., Mateus, J., Ash, E., Turan, T.N., Hunt, K.J., Malek, A.M.: The association of hypertensive disorders of pregnancy with infant mortality, preterm delivery, and small for gestational age. In: Healthcare. vol. 12, p. 597. Multidisciplinary Digital Publishing Institute (2024)
[25] Zhang, H., Li, L.H., Meng, T., Chang, K.W., Van den Broeck, G.: On the paradox of learning to reason from data. In: IJCAI (2023)
[26] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)