Clinical Epidemiology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Can Contemporary Large Language Models Provide the Domain Knowledge Needed for Causal Inference? Evaluating Automated Causal Graph Discovery Through an ASCVD Case Study

Maryam Aziz, M Alan Brookhart

Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, USA

Correspondence: M Alan Brookhart, Department of Population Health Sciences, Duke University School of Medicine, DUMC 104023, Durham, NC, 27710, USA, Email alan.brookhart@duke.edu

Purpose: Directed acyclic graphs (DAGs) are critical in epidemiology and public health research for guiding study design and minimizing bias. Yet, developing DAGs for causal inference requires substantial domain knowledge. Given the vast amounts of training data for large language models (LLMs), this study assesses the effectiveness of prompt engineering for LLMs to generate DAGs that depict causal relationships in population health using OpenAI's GPT-4o and GPT-o1.

Methods: We consider a hypothetical study on statins vs no treatment for prevention of cardiovascular disease in a general adult population. We assessed four types of prompt engineering strategies: zero-shot, one-shot, instruction based, and chain of thought (CoT) prompts. Generated DAGs were assessed based on consistency, acyclicity, accuracy of sources, completeness (based on ASCVD risk score criteria), and adherence to the prompt.

Results: We found that all generated DAGs were acyclic, except for one run using the instruction-based prompt. Additionally, more than half of the DAGs included 6/7 of the ASCVD criteria, though race was absent from all. Overall, CoT resulted in the most complete DAGs and one-shot provided the most consistency across runs and adherence to the task in the prompt. The zero-shot prompt performed notably better on GPT-o1 compared to GPT-4o, consistently providing justifications and sources for variable inclusion.

Conclusion: While the findings suggest that LLMs have a baseline capacity to generate DAGs that adhere to basic epidemiological conventions, we also found several limitations including lack of justification, systematic omission of race, and frequent source hallucination, highlighting the need for human oversight and expertise. We conclude that contemporary LLMs cannot replace a domain expert's judgment but may serve as a brainstorming or pre-analysis tool for DAG development when guided by well-engineered prompts.

Keywords: artificial intelligence, cardiovascular disease, prompt engineering, directed acyclic graphs

Introduction

In epidemiology and statistics, there is a fundamental and well-known distinction between predictive modeling and causal inference: valid causal inference requires subject matter knowledge, prediction does not.

To correctly infer causal relations from nonexperimental data, one must make specific assumptions about the underlying system and data. These assumptions are untestable and their plausibility must be assessed through an understanding of the system being studied.[1–3] Frequently this subject-matter knowledge is depicted using causal directed acyclic graphs (DAGs), a fundamental tool in causal inference used to represent assumptions about the causal structure among variables.[4,5] Developing an accurate DAG requires domain knowledge to ensure that all relevant variables and their causal relationships are adequately depicted. DAGs are used to inform study design and support decisions about the specification of the necessary statistical models, such as propensity score models to control confounding[6] and statistical models needed to address censoring and missing data.[7] Two separate researchers working with the same data but using

Received: 2 July 2025
Accepted: 28 October 2025
Published: 30 October 2025

different DAGs may arrive at different conclusions about the magnitude and even direction of a particular causal effect. Only an experiment conducted in the same population could determine which approach was least biased.

Predictive modeling, on the other hand, does not rely on untestable assumptions. It is not concerned with a mechanistic understanding of a system being studied, only with statistical relationships among variables in the observed data.[8] Models are built on training data and evaluated in validation or test data that was not used to initially train the model. Subject matter knowledge may improve the performance of the predictive model by helping the researcher identify important predictive factors or specify models, but models can be developed without such knowledge. The problems addressed by predictive models are diverse and include estimating disease risk,[9] classifying images,[10] and making predictions about the therapeutic potential of novel compounds.[11] The models themselves can be incredibly large and complex, eg, large language models (LLMs) are predictive models based on deep neural networks that contain billions of parameters. Although models created for prediction can be used to estimate parameters in causal estimators, the selection of variables to include in these models still requires subject matter knowledge. This has limited the ability of automated procedures to assess causality as algorithms would need to be able to access domain knowledge.

The reliance on humans in causal inference may start to change with the emergence of LLMs that have been trained on vast corpora of text, including academic articles, white papers, web content, news stories, and social media posts. Given that the training data for these models encompasses a vast amount of domain knowledge, it is reasonable to wonder whether the models could be prompted to generate the DAGs needed for the design of studies and the specification of the component statistical models in causal estimators. Indeed, LLMs have shown ability to diagnose certain challenging diseases[12] and pass standardized tests,[13] tasks which depend on substantial domain knowledge.

To explore this idea, we attempted to prompt a LLM to develop and justify a DAG to control confounding in a hypothetical study of statin treatment (versus no treatment) on cardiovascular risk in a population of adults. We explored how different approaches to prompting the LLM affected various measures of the quality of the resulting output.

Materials and Methods

We assessed the effectiveness of prompt engineering in large language models (LLMs) for generating DAGs to depict causal relationships in population health using OpenAI's GPT-4o and GPT-o1.

Test Cases

We consider the case of a hypothetical study of HMG-CoA reductase inhibitors (statins) vs no treatment for prevention of cardiovascular disease in a general adult population. This scenario is a well-researched topic, allowing us to comprehensively assess GPT's ability to create DAGs in a scenario where there exists much domain knowledge in the LLM's training data.

Prompt Engineering

We assessed four types of prompt engineering strategies:

- Zero-shot, where the model generates responses based on its pre-trained information, without any specific examples or context provided.
- One-shot, where the model is provided with a set of training examples of the task (n = 1), providing more context to guide the model.[14]
- Chain of Thought, where the model is prompted to present its process step-by-step before arriving at the final answer, mimicking a thought process to solve the problem.[15]
- Instruction-Based, where the model is provided with step-by-step instructions for completing the task.

In each prompt, an identity is established, and context is given to the LLM through examples. The LLM is then prompted to develop a DAG for our specified clinical scenario. We tested all prompts on GPT-4o, and additionally we tested the zero-shot prompt on both GPT-4o and GPT-o1 to assess differences across model versions given that GPT-o1 is optimized for reasoning tasks.

In this study, GPT-4o and GPT-o1 were applied without any parameter updates and fine tuning. The test cases are run solely through textual interactions with the model. We used published exposure-outcome DAGs to provide context for the one-shot prompt,[16] and a published tutorial for the instruction-based prompt.[17] The prompts used in the study are provided in supplementary appendix 1.

Study Design

Each prompt was executed three times across separate instances of ChatGPT to ensure previous conversation history was not used. This repetitive testing helps identify any anomalies or variations in the model's performance. To evaluate the effectiveness of prompt engineering techniques, two researchers reviewed the generated DAGs based on consistency across runs, acyclicity, accuracy of sources, completeness, and adherence to the prompt. Following the individual assessments, the reviewers met to discuss their evaluations. Any differences were examined and discussed collaboratively. This discussion aimed to achieve a consensus for each evaluated output.

Results

We evaluated LLM output based on consistency across runs, acyclicity, accuracy of sources, completeness, and adherence to the prompt. Completeness was assessed based on the Atherosclerotic Cardiovascular Disease (ASCVD) risk score criteria, which evaluates a patient's 10-year risk of developing cardiovascular disease.[18] The score considers age, gender, race, total cholesterol (TC), high-density lipoprotein cholesterol (HDL), smoking habits, systolic and diastolic blood pressure, and diabetes. For our evaluation, we counted a DAG as complete if it included cholesterol levels (covering both total cholesterol and HDL) and hypertension (covering systolic and diastolic blood pressure), in addition to age, gender, race, diabetes, and smoking habits. All DAGs were independently evaluated by each author. Only one disagreement occurred, which was resolved through discussion.

Zero Shot Prompt

The zero-shot prompt produced acyclic DAGs across all runs (Figure 1). Runs 1 and 3 were nearly complete, covering 6/7 ASCVD criteria, while Run 2 covered 5/7 criteria. Race was absent from all generated DAGs and Run 2 also missed gender. Runs consistently included additional variables like diet and physical activity. The zero-shot prompt produced inconsistent results across runs. For instance, only Run 1 provided sources, with 8 out of 9 found to be relevant to the clinical scenario. In contrast, Run 2 lacked any justification, sources, or Python code. The DAGs produced by the zero-shot prompt primarily focused on confounders and only Run 2 included mediators and risk factors.

One Shot Prompt

The one-shot prompt produced acyclic DAGs across all runs; however, none of the DAGs fully covered the ASCVD criteria. All runs failed to include race and Runs 2 and 3 also missed cholesterol levels (Figure 2). Each run had additional variables including income/SES, physical activity, and BMI/obesity. All runs provided sources and python code for DAG generation, though the number of variables covered, and the accuracy of citations varied. Run 1 provided 10 sources as justification for 9/12 of the included variables. Upon title and abstract review, we found that 10/10 sources were relevant to the clinical scenario, with 2/10 having either the wrong name, year, or title in the citation. Run 2 provided 10 sources as justification for 7/11 of the variables included. We found that 4/10 of the sources provided did not exist. Of the remaining 6 sources 2/6 had either the wrong name, year, or title and 4/6 were correct and relevant to the clinical scenario. Lastly, Run 3 provided 11 sources to justify 10/12 of the variables in the DAG. Title and abstract review revealed that 9/11 of the sources were relevant to the clinical scenario, and 2/11 did not exist. The DAGs produced by the one-shot prompt all included variable types in addition to confounders (mediators or risk factors).

Instruction Based Prompt

The instruction-based prompt led to acyclic DAGs in 2/3 of the runs (Figure 3). DAG 3 includes a bidirectional arrow between statin use and cholesterol levels, creating a cyclic relationship. Since our prompt did not specify whether to include time varying structures, this bidirectional arrow may reflect a time-varying relationship between statin use and

Figure 1 Directed acyclic graph (DAG) generated using a zero-shot prompt in GPT-4o, illustrating confounding pathways between statin use and cardiovascular disease. Arrows represent hypothesized causal relationships identified by the LLM.

Figure 2 Directed acyclic graph (DAG) generated using a one-shot prompt in GPT-4o, illustrating confounding pathways between statin use and cardiovascular disease. Arrows represent hypothesized causal relationships identified by the LLM.

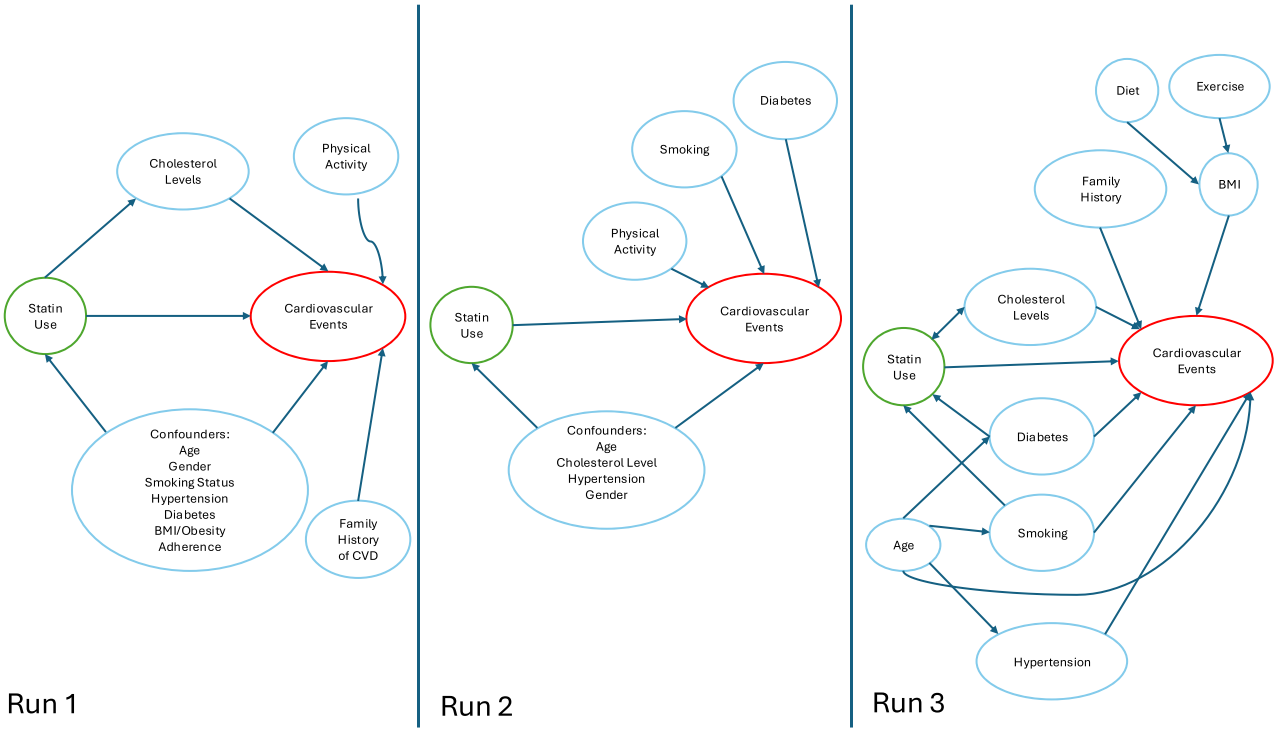

Figure 3 Directed acyclic graph (DAG) generated using an instruction-based prompt in GPT-4o, illustrating confounding pathways between statin use and cardiovascular disease. Arrows represent hypothesized causal relationships identified by the LLM.

cholesterol levels. However, acyclicity should have been maintained through temporally indexed nodes. Additionally, all DAGs failed to include race, and DAGs 2 and 3 also failed to include cholesterol levels and gender respectively. Runs consistently included additional variables like physical activity and family history. The instruction-based prompt was inconsistent across runs and did not adhere to the prompt. None of the runs provided sources and only Run 1 provided any justification and code for the DAG. The DAGs produced by the instruction-based prompt all included variable types in addition to confounders (mediators or risk factors).

Chain of Thought Prompt

All three runs of the chain of thought prompt produced acyclic and nearly complete DAGs, with all DAGs missing only race (Figure 4). Additional variables like physical activity and BMI were included in Runs 1 and 2. None of the runs provided sources, but each included justification for variables and python code to generate the DAG. Overall, the chain of thought prompt provided consistency across runs, with output following a systematic methodology breaking the task down into clear steps such as identifying variables, constructing relationships, and implementing the DAG. The DAGs produced by the chain of thought prompt primarily focused on confounders and only Run 3 included mediators and risk factors.

GPT-o1 Zero-Shot Prompt

The zero-shot prompt run on GPT-o1 produced acyclic DAGs across all runs; however, all runs failed to include race and cholesterol levels (Figure 5). Runs 1 and 3 included additional variables like obesity and family history. All runs provided python code for DAG generation, though the code produced by Run 3 caused an error and had to be manually revised to generate the DAG. All runs provided sources and justification for variable inclusion. Run 1 provided 9 sources as justification for 9/9 of the variables included in the DAG. Upon title and abstract review, we found that 9/9 sources were relevant to the clinical scenario, with 3/9 having either the wrong name, year, or title in the citation. Run 2 provided 6 sources as justification for 6/6 of the variables and included links to the sources. We found that 1/6 of the sources

https://doi.org/10.2147/CLEP.S550565

Figure 4 Directed acyclic graph (DAG) generated using a chain of thought prompt in GPT-4o, illustrating confounding pathways between statin use and cardiovascular disease. Arrows represent hypothesized causal relationships identified by the LLM.

Figure 5 Directed acyclic graph (DAG) generated using a zero-shot prompt in GPT-o1, illustrating confounding pathways between statin use and cardiovascular disease. Arrows represent hypothesized causal relationships identified by the LLM.

provided did not exist. Of the remaining 5 sources 3/5 had either the wrong link, name, year, or title and 2/5 were correct and relevant to the clinical scenario. Run 3 provided 9 sources as justification for 9/9 of the variables included in the DAG. 1/6 of the sources provided did not exist, 1/6 had the wrong name, year, or title, and the remaining 4/6 were correct and relevant to the clinical scenario. The DAGs produced by the zero-shot prompt using GPT-o1 primarily focused on confounders, though all runs included variable types in addition to confounders (mediators or risk factors). Table 1 summarizes the evaluation results across all prompt types and runs.

In our analysis, we found that DAGs generated by all prompt types were acyclic, except for one run using an instruction-based prompt. Additionally, the DAGs included a significant portion of the ASCVD criteria (7/15 DAGs

Table 1 Evaluation Results Across Prompt Types and Runs

Technique	Criteria	Run 1	Run 2	Run 3
Zero-shot	Acyclic	✓[a]	✓	✓
	Complete	6/7 (no race)	5/7 (no race, gender)	6/7 (no race)
	Additional variables	Physical activity	Exercise, diet, family history	Diet/physical activity
	Sources provided	✓ (Last Name, Year)	X[b]	X
	Justification for variables	✓	X	✓
	Python code provided	✓	X (provided image of DAG)	✓
One-shot	Acyclic	✓	✓	✓
	Complete	6/7 (no race)	5/7 (no race, cholesterol levels)	5/7 (no race, cholesterol levels)
	Additional variables	Healthcare access, income level, obesity, diet, family history, physical activity	LDL cholesterol, exercise, BMI, income level, medication adherence, medical visits	Adherence to therapy, diet, SES, physical activity, family history, BMI
	Sources provided	✓ (author, year, title) for 9/12 variables	✓ (author, year, title) for 7/11 variables	✓ (author, title, year) for 10/12 variables
	Justification for variables	✓	✓	✓
	Python code provided	✓	✓	✓
Instruction based	Acyclic	✓	✓	X
	Complete	6/7 (no race)	5/7 (no race, cholesterol levels)	5/7 (no race, gender)
	Additional variables	BMI/obesity, adherence, physical activity, family history	Physical activity	Diet, family history, exercise
	Sources provided	X	X	X
	Justification for variables	✓	X	X
	Python code provided	✓	X (provided image of DAG)	X (provided image of DAG)
Chain of thought	Acyclic	✓	✓	✓
	Complete	6/7 (no race)	6/7 (no race)	6/7 (no race)
	Additional variables	Physical activity, BMI, SES	Obesity, diet and exercise	None
	Sources provided	X	X	X
	Justification for variables	✓	✓	✓
	Python code provided	✓	✓	✓

(*Continued*)

https://doi.org/10.2147/CLEP.S550565

Table 1 (Continued).

Technique	Criteria	Run 1	Run 2	Run 3
Zero-shot (o1)	Acyclic	✓	✓	✓
	Complete	5/7 (no race, cholesterol levels)	5/7 (no race, cholesterol levels)	5/7 (no race, cholesterol levels)
	Additional variables	Obesity, family history, SES	None	Obesity, family history, physical activity
	Sources provided	✓ (author, year, title) for 9/9 variables	✓ (author, year, title, link) for 6/6 variables	✓ (author, year, title) for 9/9 variables
	Justification for variables	✓	✓	✓
	Python code provided	✓	✓	✓ (incorrect)

Notes: [a]✓ - Meets criteria; [b]X – Does not meet criteria. (Source: Authors' analysis based on model output).

included 6/7 of all ASCVD criteria). However, race was absent from all DAGs. All prompt runs provided code or DAG images with the exposure and outcome denoted by color as specified in the prompt. The zero-shot (GPT-o1), one-shot, and chain of thought prompts provided the most consistency across runs, including justification and python code in all runs. However, the code provided from one run using a zero-shot prompt (GPT-o1) required manual correction to run. Only the zero-shot (GPT-o1) and one-shot prompt provided sources consistently, though the number of sources and the accuracy/existence of the sources varied significantly. In addition to prespecified evaluation criteria, we found that chain of thought and zero-shot output mostly focused on confounders while one-shot and instruction based included mediators and risk factors consistently. Overall, chain of thought resulted in the most complete DAGs and zero-shot (GPT-o1) and one-shot provided the most consistency across runs and adherence to the task in the prompt.

Discussion

Our study investigated the capability of LLMs to generate the DAGs that are needed to estimate causal effects in non-experimental settings. We compared the performance of different approaches to prompt engineering for DAG creation and examined the performance of a "reasoning model" (GPT-o1), which solves the task by breaking it down into several sequential sub-tasks. Our findings suggest that while contemporary LLMs can extract relevant elements of domain knowledge and generate valid DAG structures, their outputs remain incomplete and inconsistent. This work adds to a growing body of literature on the use of AI/LLMs for causal discovery.[19] Past research has focused on using LLMs for iterative causal graph discovery by querying the LLM to identify pairwise causal relationships, which introduces in computational and scalability challenges.[20–22] Our approach instead focuses on advancing prompt engineering to guide the LLM in generating a comprehensive DAG directly, avoiding the need for iterative prompting.

Using multiple approaches to prompt engineering, we found that LLMs could consistently generate graphs that are acyclic, as required, and incorporate key variables that are present in a commonly used cardiovascular disease risk score. The chain-of-thought and one-shot prompts yielded the best results overall. They generated more complete DAGs, included justification for the directed edges in the graph, and produced more consistent output across runs. This aligns with recent work on prompt engineering that demonstrated that structured prompting techniques significantly enhance LLM performance on complex reasoning tasks.[14,15] As LLMs improve and become optimized for reasoning capabilities, their ability to generate accurate DAGs with more minimal prompting may improve. The zero-shot prompt performed notably better on GPT-o1 compared to GPT-4o, consistently providing justifications and sources for variable inclusion.

We observed that the LLMs were often able to propose plausible multi-node causal pathways. For example, a one-shot run identified two 4-node paths (socioeconomic status - > diet -> LDL levels -> cardiovascular risk and socioeconomic status - > physical activity -> BMI -> cardiovascular risk). However, these more complex causal paths were not consistently identified by the models. This reinforces the notion that while LLMs may be valuable tools for generating preliminary causal structures, expert oversight is essential to refine and validate these pathways to ensure plausibility and completeness.

Across all prompt engineering strategies, race was consistently omitted, despite being a component of the cardiovascular risk score. The systematic exclusion of race has important ramifications that would likely yield biased results due to obscured causal pathways. This is especially relevant in ASCVD research, where well-documented health disparities are present. Human expertise is, therefore, essential to ensure that key social determinants of health are not systematically excluded from causal diagrams proposed by LLMs. The exclusion of race raises concerns about the applicability of LLM-generated DAGs for causal inference in health disparities research, where accounting for social determinants of health is critical.

Another major limitation was the reliability of sources. The one-shot and GPT-o1 zero-shot prompts attempted to provide references, but many citations were either hallucinated or inaccurate. This finding aligns with prior research indicating that LLMs often fabricate sources while maintaining syntactically plausible citations, an issue that has been observed in attempts to use LLMs in systematic review,[23] and discussed more generally as a challenge of using LLMs for academic research.[24] These hallucinations present a significant barrier to integrating LLMs into scientific workflows without additional validation mechanisms.

Our findings should be considered in light of the study's limitations. We acknowledge that the study was limited to a single clinical scenario (statins vs no treatment on cardiovascular disease prevention), impacting the generalizability of the results. While this is a widely studied scenario allowing for a comprehensive evaluation of the generated DAGs, the results may vary for less studied treatment contrasts where the LLM training corpus contains less information. Furthermore, our analysis did not explicitly evaluate the generated DAGs for fairness or bias propagation from social and historical biases potentially embedded in the model's training data which may compromise the equity of downstream applications. Additionally, the low number of reviewers (N=2) introduces the potential for subjectivity in the evaluation.

Several avenues for further research could improve the utility of LLMs for causal inference. First, expanding testing to less well-studied scenarios would assess whether LLMs can generalize beyond deeply researched topics like statin use and cardiovascular disease. Second, refining prompt engineering strategies may improve output consistency and completeness. Future work could also investigate integrating external structured knowledge sources, such as curated medical ontologies (eg, UMLS, MeSH) to enhance the factual reliability of LLM outputs. Relatedly, further research is needed on mitigating hallucinated references, including exploring hierarchical knowledge bases as a potential solution. Additionally, comparing results across different LLM architectures and fine-tuned models optimized for biomedical reasoning (eg, BioBERT, Med-PaLM 2) could provide insights into model-specific strengths and weaknesses. Finally, future work should address fairness in LLM-generated causal models to ensure the responsible application of AI within epidemiological research.

Conclusion

While our findings indicate that current LLMs can assist in structuring DAGs and summarizing aspects domain knowledge, they are not reliable enough to replace human subject matter experts. Current LLMs could serve as a useful adjunct, assisting research teams with DAG development by identifying overlooked causal connections. However, with contemporary LLMs, rigorous human validation remains essential to ensure methodological soundness and prevent the propagation of biases and use of incomplete causal frameworks. Looking ahead, these findings support using LLMs within a human-in-the-loop workflow that pairs automated DAG development with checks to verify sources and ensure the graphs consider fairness and equity.

Ethics Approval and Informed Consent

This study did not involve human or animal participants, samples, or data. Nevertheless, the research complied with institutional ethical standards and best practices for transparency.

Consent for Publication

Not applicable. This study does not include any images, videos, recordings, or identifiable data requiring consent for publication.

Acknowledgments

The abstract of this paper was presented at the International Society for Pharmacoepidemiology 41st Annual Meeting as a poster presentation with interim findings. The poster's abstract was published in 'Poster Abstracts' in Pharmacoepidemiology and Drug Safety.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

The authors received no external funding for this work.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology*. 2001;12(3):313–320. doi:10.1097/00001648-200105000-00011
2. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002;155(2):176–184. doi:10.1093/aje/155.2.176
3. Pearl J. An introduction to causal inference. *Int J Biostat*. 2010;6(2):Article7. doi:10.2202/1557-4679.1203
4. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37–48. doi:10.1097/00001648-199901000-00008
5. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669–688. doi:10.1093/biomet/82.4.669
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55. doi:10.1093/biomet/70.1.41
7. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. *Int J Epidemiol*. 2015;44(4):1452–1459. doi:10.1093/ije/dyu272
8. Shmueli G. To explain or to predict? *Statistical Science*. 2010;25(3):289–310,22. doi:10.1214/10-STS330
9. Lund JL, Kuo TM, Brookhart MA, et al. Development and validation of a 5-year mortality prediction model using regularized regression and medicare data. *Pharmacoepidemiol Drug Saf*. 2019;28(5):584–592. doi:10.1002/pds.4769
10. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit Med*. 2023;6(1):74. doi:10.1038/s41746-023-00811-0
11. Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. *Cell*. 2020;180(4):688–702e13. doi:10.1016/j.cell.2020.01.021
12. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*. 2023;330(1):78–80. doi:10.1001/jama.2023.8288
13. Abbas A, Rehman MS, Rehman SS. Comparing the performance of popular large language models on the National Board of Medical Examiners Sample Questions. *Cureus*. 2024;16(3):e55991. doi:10.7759/cureus.55991
14. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Presented at: Proceedings of the 34th International Conference on Neural Information Processing Systems; Vancouver, BC, Canada; 2020.
15. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Presented at: Proceedings of the 36th International Conference on Neural Information Processing Systems; New Orleans, LA, USA; 2022.
16. Etminan M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. *Chest*. 2020;158(1, Supplement):S21–S28. doi:10.1016/j.chest.2020.03.011
17. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. *J Clin Epidemiol*. 2022;142:264–267. doi:10.1016/j.jclinepi.2021.08.001
18. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2014;129(25 Suppl 2):S49–73. doi:10.1161/01.cir.0000437741.48606.98
19. Wan G, Lu Y, Wu Y, Hu M, Li S. Large language models for causal discovery: current landscape and future directions. *arXiv [csCL]*. 2025;2025.
20. Willig M, Zečević M, Dhami DS, Kersting K. Can foundation models talk causality? *arXiv [csAI]*. 2022;2022.
21. Long S, Schuster T, Piché A. Can large language models build causal graphs? *arXiv [csCL]*. 2024;2024.
22. Kıcıman E, Ness R, Sharma A, Tan C. Causal reasoning and large language models: opening a new frontier for causality. *arXiv [csAI]*. 2024;2024.
23. Chelli M, Descamps J, Lavoue V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. *J Med Internet Res*. 2024;26:e53164. doi:10.2196/53164
24. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224–226. doi:10.1038/d41586-023-00288-7