EvidenceTriangulator: A Large Language Model Approach to Synthesizing Causal Evidence across Study Designs

Xuanyu Shi, MS[1,2], Wenjing Zhao, PhD[1,2], Chao Yang, MS[3], Jian Du[*], PhD[1,2]

[1]Institute of Medical Technology, Peking University, Beijing, China;
[2]National Institute of Health Data Science, Peking University, Beijing, China;
[3]Renal Division, Department of Medicine, Peking University First Hospital, Peking University Institute of Nephrology, Beijing, China

Abstract
In managing chronic diseases, the role of social determinants like lifestyle and diet is crucial. A comprehensive strategy combining biomedical and lifestyle changes is necessary for optimal health. However, the complexity of evidence from varied study designs on lifestyle interventions poses a challenge to decision-making. To tackle this challenge, our work focused on leveraging large language model to construct a dataset primed for evidence triangulation. This approach automates the process of gathering and preparing evidence for analysis, thereby simplifying the integration of reliable insights and reducing the dependency on labor-intensive manual curation. Our approach, validated by expert evaluations, demonstrates significant utility, especially illustrated through a case study on reduced salt intake and its effect on blood pressure. This highlights the potential of leveraging large language models to enhance evidence-based decision-making in health care.

Introduction

Social determinants of health (SDoH), especially lifestyle factors such as diet and exercise, are pivotal in managing major chronic diseases such as cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes. For instances, according to data from the Institute for Health Metrics and Evaluation (IHME), behavioral factors contribute significantly to ischemic heart disease and stroke, accounting for 69.2% and 47.4% of Disability-Adjusted Life Years (DALYs), respectively—the highest among all diseases. In particular, dietary factors contributed 57.1% and 30.6% of DALYs, respectively (1). It is increasingly recognized that health strategies should prioritize both SDoH-based behavioral interventions and biomedical interventions (e.g., medications) (2). Developing evidence-based intervention strategies encounters significant challenges due to the rapidly growing and piecemeal evidence, along with complex causal relationships from various study designs, including confounding and reverse causation. Evaluating the reliability of causal relationships within a body of scientific evidence is essential for evidence-based decision-making, especially when research findings are inconsistent (3, 4).

Meta-analysis (META) is an effective scientific method for quantitatively synthesizing research conclusions. Utilizing statistical techniques, it combines the results of different studies to obtain an overall quantitative estimate of the impact of specific interventions (e.g., salt restriction) on particular outcomes (e.g., blood pressure). It balances conflicting evidence quantitatively to achieve evidence-based decision-making based on synthesized scientific evidence. Since its introduction in the 1970s, meta-analysis has had a significant impact on various fields such as medicine, economics, sociology, and environmental science (5). Over the past four decades, meta-analysis has evolved to include increasingly complex methods for quantifying evidence, particularly concerning the consistency of results from the same study design or the replicability of studies. In contrast, convergence, reflecting the extent to which a given hypothesis is supported by different study designs, has not received the same attention (4). Currently, considering consistency and convergence is recognized as an important strategy for addressing the reproducibility crisis for the scientific community (4).

In recent years, the idea of "triangulation" has been introduced into the scientific community to measure the convergence of scientific conclusions derived from different study designs (4, 6, 7). These study designs have different and independent potential sources of bias (6). Triangulation is a research strategy involving the use of at least two research methods to investigate and analyze the same problem, mutually validating each other to enhance the robustness and reproducibility of conclusions. If conclusions derived from different research designs (such as observational studies (OS), mendelian randomization studies (MR), and randomized controlled trials (RCT), etc.)

[*] Correspondence to: dujian@bjmu.edu.cn

regarding the same scientific question are consistent, the reliability of causality is stronger. When the results point to different answers, understanding the major source of bias instruct researchers future study designs (6). However, currently, scholars primarily employ qualitative methods to explain the reliability of causal relationships through triangulation, lacking quantitative approaches. Qualitative methods mainly involve retrospective analysis of relevant literature in the "Discussion" section of papers, simply summarizing and discussing how many studies support the conclusions of the current study, how many do not, and reasons for lack of support, such as different experimental conditions (7). Current evidence triangulation in medicine involves a very high proportion of manual work (8). However, such retrospective, qualitative triangulation methods are susceptible to issues such as subjective selectivity of evidence and biases in understanding among different researchers.

Implementing a fully quantitative method for evidence triangulation requires a computable representation of research results and relevant metadata obtained from different study designs. Apart from determining the presence and direction of the effects (i.e., significant increase, significant decrease, and null) given an intervention and outcome, finer-grained information of research design among many lines of studies need to be extracted. For evidence triangulation task, it is important to extract information such as the duration and intensity of intervention, measured outcomes, intervention targets (prevention vs. treatment), characteristics of study populations (e.g., demographics), and other relevant contextual information.

Currently, there are natural language processing methods available for extracting conclusions from clinical research reports. This includes the utilization of Large Language Models (LLMs) to extract entities and relationships from RCT reports (9, 10). However, these methods are predominantly based on the broad framework of evidence-based medicine, which emphasizes Population-Intervention-Comparation-Outcome (PICO) related concepts, such as Trialstreamer and the EvidenceMap (11, 12). While some of these methods involve effect size and direction (13, 14), extracting and representing research design information from various sources of evidence, which is essential for triangulation, remains a subject for ongoing research. Most recently, there are attempts trying to accelerate the process by taking advantage of computable knowledgebase, such as SemMedDB and natural language processing tools (15). However, the accuracy and recall rates of medical concepts and their relationships extracted in SemMedDB are relatively low, posing significant challenges to triangulation based on a complete medical evidence system.

In this study, we try to examine the capabilities of LLMs in extracting intervention-outcome concepts, determining causal directions, as well as identifying study design information. Our objective is to develop an automatic approach to aggregate various lines of SDoH-related evidence across different study designs into a computable and comparable format that is ready for quantitative evidence triangulation.

Methods

Our procedure begins by collecting titles and abstracts from relevant literature. We then apply a LLM to systematically process these texts across various study designs, extracting key outcomes and methodological details. This leads to the aggregation of data into a coherent, transparent dataset that is ready for triangulation analysis (Fig. 1.).

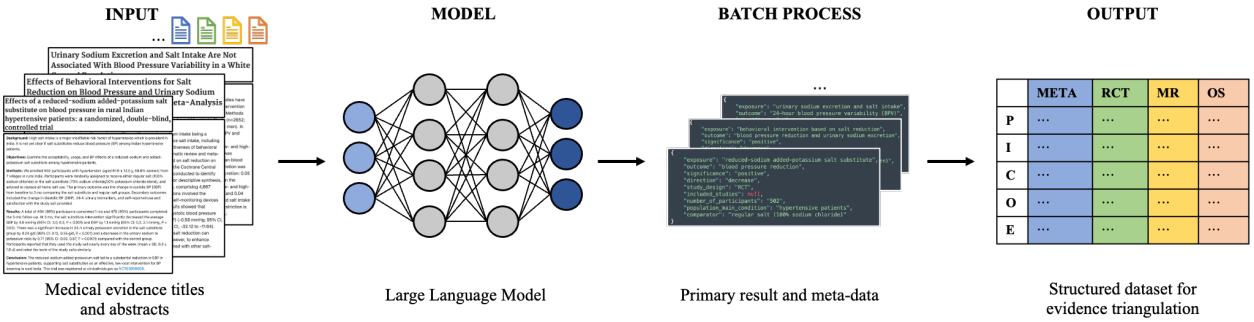

Figure 1. Overall workflow of automatic evidence triangulation using LLM

(1) Data source

Regarding data sources, the study utilizes literature categorized under publication types marked as meta-analysis, systematic reviews, observational studies, randomized controlled trials, clinical trials and related types available on

PubMed. The MeSH terms "cardiovascular diseases" and "Diet, Food, and Nutrition" are utilized as search terms, with MeSH major topic as the search field. The resulting search query is outlined below: "*(cardiovascular diseases[MeSH Major Topic]) AND (Diet Food,and Nutrition[MeSH Major Topic])*".For studies employing mendelian randomization (not a conventional publication type in PubMed), we additionally narrowed down the search to include only publication titles and abstracts containing the phrase "Mendelian randomization". In total, 4,268 articles were retrieved. This first dataset will be the corpus for validations for extracted results by LLM.

To provide a specific example of the relationship between a particular disease and dietary factors, we further selected salt and hypertension as the intervention-outcome pair and retrieved relevant publications. Consistent with the aforementioned limitations on publication types, we refined the search terms to include MeSH terms related to salt and hypertension. The constructed search query is as follows: *("sodium chloride"[MesH Major Topic] OR "salts"[MeSHMajor Topic]) AND ("hypertension"[MesH Major Topic] OR ("bloodpressure"[MeSH Major Topic] OR "blood pressure determination[MesH Major Topic] OR "arterial pressure"[MesH Major Topic]))*. Ultimately, we retrieved a total of 289 articles. This second dataset will be used to exhibit the formation of automatic-extracted ready-for-triangulation evidence dataset in the results section.

(2) LLM-based study results extraction

For the task of extracting precise and insightful results from health-related documents, we employed GPT-4 (model version: gpt-4-0125-preview), a GPT-based (Generative Pre-trained Transformer) LLM model introduced by OpenAI (16, 17), renowned for its high performance in information extraction within health domain(18, 19).

The specific extraction tasks for the model are designed as following:

- **Identification of study design**
  The initial step involves using GPT-4 to categorize the study design present in medical abstracts. The designs considered include RCT, MR, OS, and META.
- **Extraction for meta-analyses and systematic reviews**
  For abstracts identified as META, we ask GPT-4 to extract the number of included studies and their respective study designs. This step is crucial for understanding the strength and diversity of evidence in these comprehensive analyses.
- **Primary result identification**
  Next, we ask GPT-4 to identify the primary result from each abstract. This involves recognizing the main findings that the study reports, which is essential for summarizing the study's major contribution to the field.
- **Intervention/Exposure and outcome extraction**
  Following the identification of the primary results, the model extracts key entities including intervention or exposure and the corresponding primary outcome.
- **Relationship and statistical significance**
  First the model extracts the direction of the relationship from the intervention/exposure to the outcome. The model assesses whether the intervention/exposure increases, decreases or an effect was not found. Then we ask GPT-4 to extract statistical significance of the identified relationship, ensuring the ability to distinguishing positive results from negative results.
- **Population, Participant Number and Comparator Group information**
  Adhering to the standard representation medical evidence, we ask the model to extract information on the population condition under study, the number of participants, and details of the comparator group if applicable.

This prompt is to follow a logical progression from study-level information (study design), to more specific study result extraction (intervention/exposure, primary outcome, relationship direction, statistical significance), then contextual details (population, participant number, comparator). Fig.2. shows a graphical illustration of the overflow and logics of the designed prompt. For each abstract, GPT-4 first determines the study design. If the abstract pertains to a meta-analysis, the model then identifies the number and types of included studies. Subsequently, it locates the primary result, extracts relevant details about the intervention/exposure and outcome, and assesses the direction and significance of the relationship. Information about the study population, the number of participants, and comparator group details are also extracted, providing a comprehensive overview of each study's evidence.

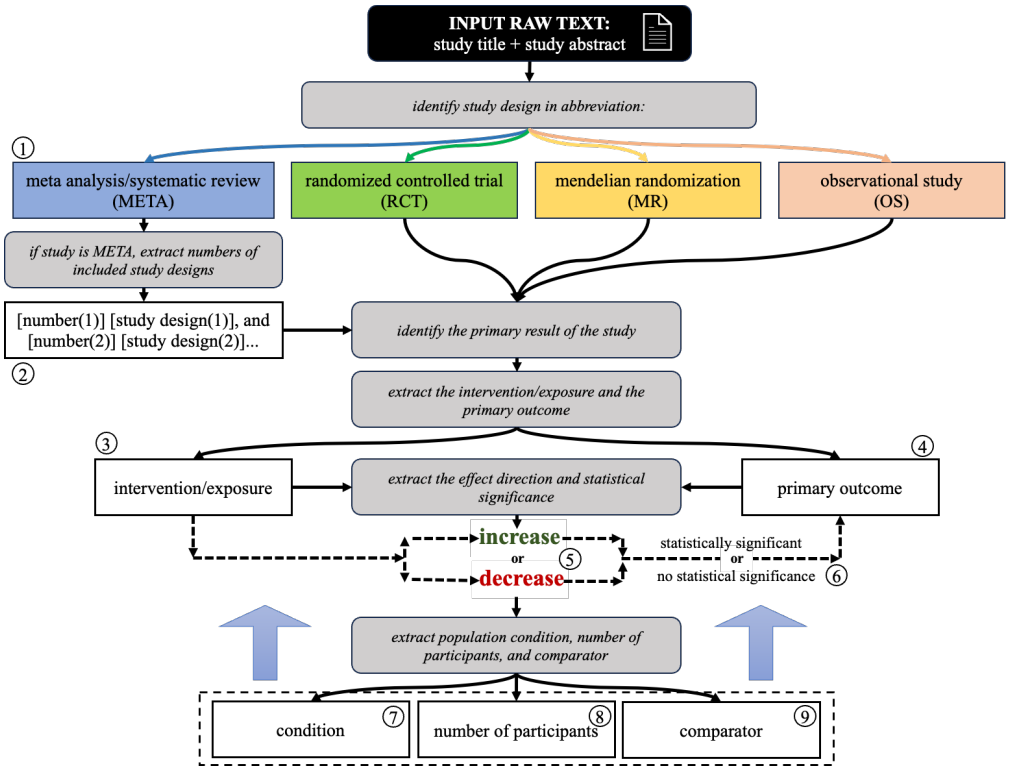

Figure 2. A flowchart describing the overall logic of using LLM to extract medical evidence in structured format. Gray boxes represent each part of the prompt in each step. Circled numbers (1-9) represent the extracted information by the model.

Results

(1) Human evaluation of model extraction

To assess the efficacy and precision of our model in organizing medical evidence into a structured format, we engaged five experts from relevant domains to review the extraction outcomes. Expert backgrounds include oncology, cardiovascular disease, clinical pharmacy, public health, and pediatric clinical and big data. A hundred publications were randomly selected for this evaluation process. The experts utilized a Likert Scale (a scale that measures opinions from "strongly agree" to "strongly disagree.") to rate the LLM's performance on the extraction tasks. Before the evaluation, we provided the experts with an overview of the study's goals and rationale to ensure a comprehensive understanding.

Each of the five experts was tasked with evaluating the performance across five criteria:

- Intervention/exposure, comparator and primary outcome accuracy
- Study design accuracy
- Population accuracy
- Relationship reasoning
- Statistical significance reasoning

To guarantee the reliability of our evaluation, we had two different experts reviewing each extraction result. We calculated the average scores from both reviewers for each publication and then determined the overall mean scores for entity accuracy. For relationship evaluation, we analyzed the distribution of correct, incorrect, or missed extractions. Experts also provided written feedback memos on any issues they identified in the extracted results. These feedback memos were summarized and analyzed.

In the evaluation feedback (Table 1) (30 missing ratings), the model performed notably well in extracting study information, with high accuracy in identifying exposure, comparator, and outcomes (mean score 4.69), and study design (mean score 4.84). The model frequently achieved perfect scores for accurate information extraction, especially in METAs, with minimal instances of inaccuracies. In terms of identifying study populations, the model achieved a high mean score of 4.68, effectively pinpointing nearly all relevant entities. For reasoning about relationships and

statistical significance, the model demonstrated satisfactory accurate reasoning, with a substantial proportion of correct instances compared to incorrect or absent reasoning.

Table 1. Expert human evaluation results

SCORE	INDICATOR	DESCRIPTION	COUNT
		Exposure, Comparator, and Outcomes Accuracy	**mean=4.69**
5	Completely accurate	Accurately extracted all relevant information	153/200 (76.5%)
4	Mostly accurate	Extracted most of the relevant information with minor omissions	31/200 (15.5%)
3	Partially accurate	Extracted significant relevant information with notable omissions	13/200 (6.5%)
2	Poorly accurate	Extracted key aspects but missed a substantial amount of detail	0/200 (0.0%)
1	Inaccurate	Failed to extract key information	1/200 (0.5%)
		Study Design Accuracy	**mean=4.84**
5	Completely accurate	Correctly identified design and extracted study counts for META	181/200 (90.5%)
4	Mostly accurate	Correctly identified design but errors in study counts or designs for META	0/200 (0.0%)
3	Partially accurate	Correctly identified design but not classified into desired categories	4/200 (2.0%)
2	Poorly accurate	Failed to identify valid design and included studies for META	2/200 (1.0%)
1	Inaccurate	Incorrectly identified design and study counts for META	5/200 (2.5%)
		Study Population Accuracy	**mean=4.68**
5	Almost all identified	Almost all relevant entities correctly identified	153/200 (76.5%)
4	Mostly identified	Most relevant entities correctly identified	7/200 (3.5%)
3	Partially identified	A significant number of entities correctly identified	13/200 (6.5%)
2	Few identified	Few entities correctly identified	4/200 (2.0%)
1	Very few identified	Very few entities correctly identified	3/200 (1.5%)
Categorical indicators		**Relationship Direction Reasoning**	
		Correct	145/200 (72.5%)
		Incorrect	34/200 (17.0%)
		Not reasoned	21/200 (10.5%)
		Statistical Significance Reasoning	
		Correct	157/200 (78.5%)
		Incorrect	22/200 (11.0%)
		Not reasoned	21/200 (10.5%)

The expert evaluation feedback memo highlights several key points: 1) Full texts of the articles are needed to evaluate if some of the extracted results are accurate. 2) It notes discrepancies in outcome measures, with most indicators showing decline and variations in significance and direction across different outcomes. 3) Some studies were identified as potentially out of scope, such as animal experiments and descriptive articles. 4) Additionally, there was confusion about the direction of health actions and outcomes.

Regarding the mentioned feedbacks, we look forward to improving our future model by: 1) Acquire full texts using PubMed Central to provide a more comprehensive corpus. 2) Using a finer-grained representation of study results, i.e., one-to-one relationships among multiple health actions and outcomes. 3) Adding a forehanded task to identify if the study matches the scope of design. 4) Extracting the directions of health actions, e.g., lower or higher; and the directions of outcome measure, e.g., improvement or worsen.

(2) A case of evidence triangulation of salt on blood pressure

In our analysis, we explored the effect of salt intake on blood pressure using a dataset derived from 289 studies. This dataset included 36 METAs, 124 RCTs, 117 OSs, and 12 studies where the design was not identified. We employed the proposed LLM pipeline to extract structured information based on the PICO framework along with primary efficacy results and other relevant metadata from these publications. A sample of the extracted evidence is presented in Table 2. It is important to note that no MR studies were found to support any direction of the relationship in this dataset.

The evidence from METAs and RCTs mainly supports the intervention of reducing salt intake. In contrast, OSs typically involved higher salt intakes as exposures. Findings from METAs consistently indicate that lowering salt intake is linked with reduced blood pressure levels. Similarly, OSs suggest that higher salt intake is associated with an increased risk of hypertension-related conditions. In summary, the evidence across different study designs points towards a beneficial effect of lower salt consumption on blood pressure, providing a clear narrative on the impact of dietary salt on cardiovascular health.

Table 2. Example of automatic-extracted ready-for-triangulation evidence dataset of salt-on-hypertension

Design	Meta		RCT		MR	OS	
Journal	BMJ	Cochrane	Hypertension	Hypertension	null	J Intern Med	Medicine
Pmid	32094151	34164803	31079530	28993451	null	33210391	32702808
Year	2020	2021	2019	2017	null	2021	2020
Included studies	133 RCTs	21 RCTs	null	null	null	null	null
Exposure	reduction in dietary sodium	reduced dietary salt	dietary sodium reduction	sodium reduction and the DASH diet	null	dietary salt intake	salty snacks dietary pattern
Outcome	blood pressure levels	blood pressure and albuminuria	blood pressure and skin capillary density	blood pressure	null	risk of new-onset atrial fibrillation (AF)	gestational hypertension
Direction	decrease	decrease	decrease	decrease	null	increase	increase
Significance	positive	positive	positive	positive	null	positive	positive
Condition	null	chronic kidney disease (CKD)	untreated black hypertensives	pre-/stage 1 hypertension	null	community-dwelling adults	nulliparous pregnant Chinese women
Comparator	usual sodium intake	higher dietary salt intake	placebo	typical American diet	null	third quintile of sodium excretion	null
Population	12,197	1,197	64	412	null	473,080	2,580

Conclusion and discussion

To conclude, this study demonstrates the potential using LLM to automate the extraction and triangulation of SDoH-related evidence from diverse study designs. Through expert evaluations and a focused case study on the impact of salt intake on blood pressure, we illustrated that LLM can significantly simplify the synthesis of medical evidence, enhancing the efficiency of evidence-based decision-making. Despite its promise, the study faced challenges in accurately classifying study designs and interpreting outcomes due to inconsistencies in the data. Additionally, the reliance on expert evaluations introduces subjective bias. Moving forward, addressing these limitations through continuous model training and exploring more objective evaluation methods is crucial for maximizing the utility of LLMs in evidence triangulation.

Decision-making should be grounded in causal relationships between interventions and outcomes, while prediction can rely solely on correlative relationships. In this study, our objective is not for an automatic meta-analysis focusing solely on evidence derived from the same study design (such as RCTs). Instead, we aim to utilize LLMs to enable a convergence analysis between results obtained from different types of study designs, known as triangulation analysis. We argue that the primary distinction between meta-analysis and triangulation analysis lies in the consideration of evidence from various study designs. The former assesses the consistency of evidence within a single study design, while the latter examines the convergence of conclusions across diverse study designs. Although there are currently no widely accepted quantitative methods for assessing convergence, insights may be drawn from *Cumulative Evidence Index (CEI)* of convergence in neurobiological experimental results, primarily utilizing a vote-counting approach across different study designs (20, 21).

Acknowledgement

This study was funded by the National Key R&D Program for Young Scientists (Project number 2022YFF0712000 to JD) and the National Natural Science Foundation of China (Project number 72074006 to JD). We declare no conflicts of interest.

We are grateful to the following experts for their invaluable contributions and insightful feedback on this study: Dr. Guohua He from Sun Yat-sen University First Affiliated Hospital, Dr. Na He from Peking University Third Hospital,

Dr. Zhenhua Lu from Peking University Cancer Hospital, Dr. Weihua Hu from Peking University, and Dr. Mingming Zhao from Peking University Third Hospital.

Authors contribution

XS was responsible for manuscript drafting, data collection, data analysis, literature review, results evaluation, and the formation of the conclusion. WZ was involved in data collection. CY was involved in method design. JD was involved in research conceptualization, and responsible for primary supervision.

References

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. Journal of the American college of cardiology. 2020;76(25):2982-3021.
2. Ali MK, Sudharsanan N, Thirumurthy H. Behaviour change in the era of biomedical advances. Nature human behaviour. 2023;7(9):1417-9.
3. Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. CHANCE. 2019;32(1):42-9.
4. Munafò MR, Davey Smith G. Robust research needs many lines of evidence. Nature. 2018;553(7689):399-401.
5. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature. 2018;555(7695):175-82.
6. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. International journal of epidemiology. 2016;45(6):1866-86.
7. Munafò MR, Higgins JPT, Smith GD. Triangulating Evidence through the Inclusion of Genetically Informed Designs. Cold Spring Harbor perspectives in medicine. 2021;11(8).
8. Sae-Jie W, Supasai S, Kivimaki M, Price JF, Wong A, Kumari M, et al. Triangulating evidence from observational and Mendelian randomization studies of ketone bodies for cognitive performance. BMC medicine. 2023;21(1):340.
9. Joseph SA, Chen L, Trienes J, Göke HL, Coers M, Xu W, et al. FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence. arXiv preprint arXiv:240211456. 2024.
10. Wadhwa S, DeYoung J, Nye B, Amir S, Wallace BC, editors. Jointly extracting interventions, outcomes, and findings from RCT reports with LLMs. Machine Learning for Healthcare Conference; 2023: PMLR.
11. Marshall IJ, Nye B, Kuiper J, Noel-Storr A, Marshall R, Maclean R, et al. Trialstreamer: A living, automatically updated database of clinical trial reports. Journal of the American Medical Informatics Association : JAMIA. 2020;27(12):1903-12.
12. Kang T, Sun Y, Kim JH, Ta C, Perotte A, Schiffer K, et al. EvidenceMap: a three-level knowledge representation for medical evidence computation and comprehension. Journal of the American Medical Informatics Association : JAMIA. 2023;30(6):1022-31.
13. Mayer T, Marro S, Cabrio E, Villata S. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. Artificial intelligence in medicine. 2021;118:102098.
14. Whitton J, Hunter A. Automated tabulation of clinical trial results: A joint entity and relation extraction approach with transformer-based language representations. Artificial intelligence in medicine. 2023;144:102661.
15. Liu Y, Gaunt TR. Triangulating evidence in health sciences with Annotated Semantic Queries. medRxiv. 2022:2022.04. 12.22273803.
16. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.
17. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.
18. Lee P, Goldberg C, Kohane I. The AI revolution in medicine: GPT-4 and beyond: Pearson; 2023.
19. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digital Medicine. 2021;4(1):93.
20. Matiasz N, Wood J, Silva A. Quantifying convergence and consistency. Authorea Preprints. 2023.
21. Matiasz NJ, Wood J, Doshi P, Speier W, Beckemeyer B, Wang W, et al. ResearchMaps. org for integrating and planning research. PloS one. 2018;13(5):e0195271.