nature communications

Article

https://doi.org/10.1038/s41467-025-62783-x

Evidence triangulator: using large language models to extract and synthesize causal evidence across study designs

Received: 8 October 2024

Accepted: 29 July 2025

Published online: 09 August 2025

Check for updates

Xuanyu Shi[1,2], Wenjing Zhao[1,2], Ting Chen[3], Chao Yang[4,5] & Jian Du[1,2,6]

Health strategies increasingly emphasize both behavioural and biomedical interventions, yet the complex and often contradictory guidance on diet, behavior, and health outcomes complicates evidence-based decision-making. Evidence triangulation across diverse study designs is essential for balancing biases and establishing causality, but scalable, automated methods for achieving this are lacking. In this study, we assess the performance of large language models in extracting both ontological and methodological information from scientific literature to automate evidence triangulation. A two-step extraction approach—focusing on exposure-outcome concepts first, followed by relation extraction—outperforms a one-step method, particularly in identifying the direction of effect (F1 = 0.86) and statistical significance (F1 = 0.96). Using salt intake and blood pressure as a case study, we calculate the Convergency of Evidence and Level of Convergency, finding a strong excitatory effect of salt on blood pressure (942 studies), and weak excitatory effect on cardiovascular diseases and deaths (124 studies). This approach complements traditional meta-analyses by integrating evidence across study designs, and enabling rapid, dynamic assessment of scientific controversies.

It is increasingly recognized that health strategies should prioritize both behavioral interventions and biomedical interventions (e.g., medications)[1]. 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 instance, 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[2]. Developing evidence-based prevention and 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 level of causality within a body of scientific evidence is a fundamental task, especially when research findings are inconsistent[3–5].

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[6]. Over the past four decades, meta-analysis has evolved to include increasingly complex methods for quantifying evidence, particularly

[1]Institute of Medical Technology, Peking University, Beijing, China. [2]National Institute of Health Data Science, Peking University, Beijing, China. [3]Business School, Dublin City University, Dublin, Ireland. [4]Renal Division, Department of Medicine, Peking University First Hospital, Peking University Institute of Nephrology, Beijing, China. [5]Center for Digital Health and Artificial Intelligence, Peking University First Hospital, Beijing, China. [6]State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University, Beijing, China. e-mail: dujian@bjmu.edu.cn

concerning the consistency of results from the same study design or the replicability of studies. In contrast, convergency, 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 convergency 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 convergency of scientific conclusions derived from different study designs[4,7,8], particularly in human behaviours[9]. These study designs have different and independent potential sources of bias[7] (see Table 1). Triangulation is a research strategy involving the use of at least two research methods to investigate and analyze the same research question, 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.) regarding the same cause-and-effect question (in fact, these study designs all aim to establish correlation) are consistent, the reliability of causality is stronger. At this point, correlation is moving towards causality[10]. When the results point to different directions, understanding the major source of bias instruct researchers future study designs[7].

However, current evidence triangulation studies primarily employ qualitative methods to explain the reliability of causality, lacking quantitative approaches. Researchers are accustomed to using retrospective description of relevant literature in the "Discussion" section of their 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[8]. A few pieces of empirical work on evidence triangulation involves a very high proportion of manual evidence screening and extraction for data elements[11]. Such retrospective, qualitative triangulation methods are susceptible to issues such as subjective selectivity of evidence and cognitive biases among different researchers.

Implementing a fully quantitative method for evidence triangulation requires a computable representation of research findings 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 no change) between an intervention and outcome, finer-grained information of research design among many lines of evidence need to be extracted. For evidence triangulation task, it is important to extract information such as measured outcomes, effect direction of intervention (increased vs. decreased), 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 machine learning techniques and Large Language Models (LLMs) to extract entities and relations from RCT reports[12–17]. However, these methods are predominantly based on the less specific framework of evidence-based medicine, which emphasizes Population-Intervention-Comparator-Outcome (PICO) related concepts, such as Trialstreamer and the EvidenceMap[18,19]. While some of these methods involve effect size and direction[20,21], 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 evidence triangulation process by taking advantage of computable knowledgebase in the form of a Subject-Predicate-Object semantic triple, such as SemMedDB[22]. However, the accuracy and recall rates of medical concepts and their relations extracted in SemMedDB are relatively low.

In this study, we try to examine the capabilities of LLMs in extracting ontological information such as intervention-outcome concepts, determining effect directions, as well as identifying methodological information such as study design. 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. We also aim to utilize the extracted data elements to assess the Convergency of Evidence (CoE, which represents the trending effect direction after triangulation) and the Level of Convergency (LoC, which denotes the strength of that converging direction). The overall rationale and pipeline of this work is shown in Fig. 1.

Results

Validation of model performance

To validate the performance of our model in named entity recognition (NER) and relation extraction (RE) in SDoH research, we employed a comprehensive evaluation approach using BERTScore, a state-of-the-art metric for assessing the semantic similarity between textual representations of exposure and outcome pairs[23]. BERTScore provides precision, recall, and F1-score metrics to quantify the semantic similarity between predicted and reference text sequences during validation. Given that LLM-extracted entities may not always perfectly align with human-extracted entities, we employed this similarity-based scoring method to more rigorously assess the extent to which the LLM's extractions correspond to the gold standard.

We compared the predicted associations generated by the model against the manually curated gold standard dataset. The following steps were undertaken:

- Similarity Assessment: BERTScore was calculated for each exposure and outcome pair to evaluate the semantic similarity between the model's predictions and the gold standard. For each PMID, precision, recall, and F1-score were computed, allowing for a nuanced understanding of the model's ability to capture relevant associations.

Table 1 | Strengths and inherent causal biases of different study designs

Study Design	Key Strengths	Main Potential Biases	Implications for Causal Inference
Randomized Controlled Trials (RCTs)	- Gold standard for causality due to randomization; - Rigorous control of intervention and outcome measurement	- Restricted generalizability; - Attrition bias (dropout); - Short-term follow-up limits capturing long-term effects	- High internal validity; - May not reflect real-world applicability if study population is unrepresentative
Observational Studies (OS)	- Large and diverse cohorts; - Longitudinal tracking of real-world behaviour; - Often captures rare outcomes or exposures	- Confounding bias due to non-random exposure assignment; - Selection bias; - Reverse causation	- Broad external validity; - Vulnerable to systematic biases if confounding variables and study populations are not carefully managed
Mendelian Randomization (MR)	- Reduces confounding and reverse causation via genetic instruments; - Biological plausibility tests for causal effects	- Horizontal pleiotropy (genetic variant affects multiple traits); - Population stratification; - Weak instrument bias	- Offers quasi-experimental conditions; - Violations of core assumptions can undermine reliability of the causal estimate

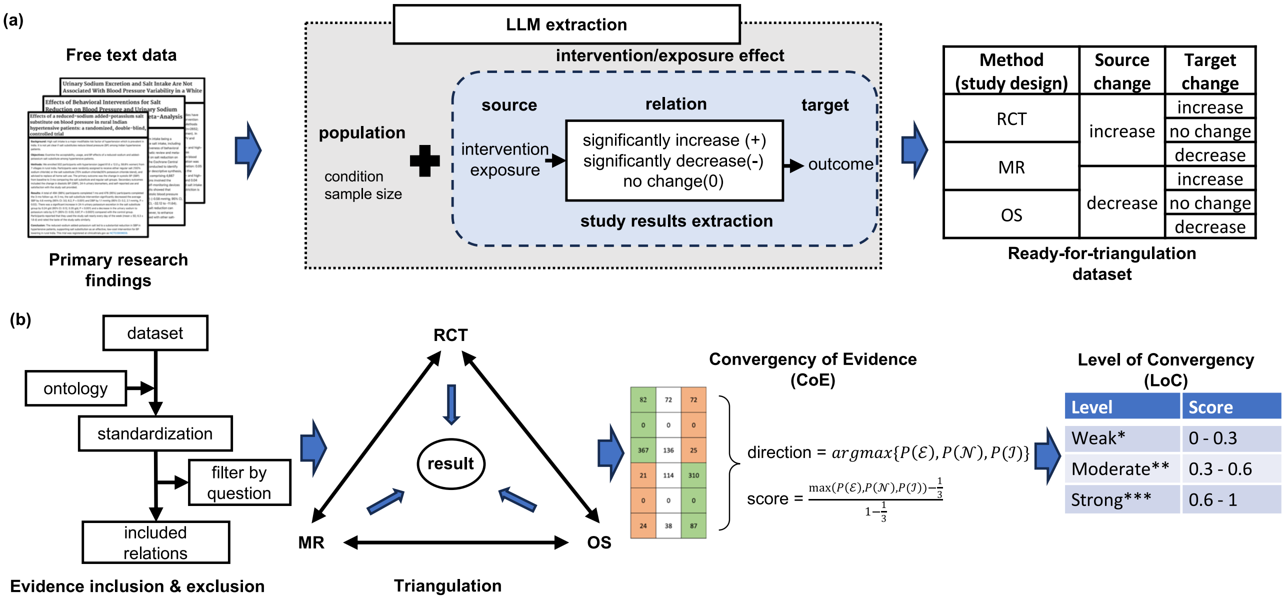

Fig. 1 | Overall workflow of automatic evidence triangulation using LLM. a The pipeline of using LLM to extract study designs, entities, and relations from textual titles and abstracts. **b** The framework of evidence triangulation; The Convergency of Evidence represents the integration algorithm of the supporting and opposing evidence behind the relation. The LoC score represents the reliability of causal relationship with scaled classification in three levels: weak (one star*), moderate (two stars**), and strong (three stars***). **Excitatory** ($\mathcal{E}$) Occurs when the outcome changes in the same direction as the exposure or intervention. For instance, if an increased salt intake increases blood pressure or a decreased salt intake decreases blood pressure, we label the effect as excitatory. **No Change** ($\mathcal{N}$) Refers to findings that indicate the exposure or intervention does not lead to a statistically significant change in the outcome (e.g., salt intake has no measurable impact on blood pressure levels). **Inhibitory** ($\mathcal{I}$) Occurs when the outcome changes in the opposite direction of the exposure or intervention. For example, if an increased salt intake decreases blood pressure, we label this effect as inhibitory.

- Matching and Thresholding: We set a BERTScore threshold of 0.8 to identify matching pairs of exposures and outcomes between the predicted and gold standard data. Only those pairs exceeding this threshold were considered valid matches.
- Evaluation of Direction and Significance: For the matched pairs, we further evaluated the model's performance in predicting the direction (e.g., increase, decrease) and statistical significance of the associations. Precision, recall, and F1-score were calculated to quantify the model's performance in these dimensions.
- Error Analysis: We identified and reported falsely predicted associations in terms of direction and significance, providing insights into areas where the model may need further refinement. This part is provided in Supplementary Note 1.

Part 1: an expert-extracted dataset of relations between food&-nutrition and cardiovascular outcomes

In the one-step one-shot extraction, GPT-4o-mini achieved the highest F1 scores for exposure (0.86) and outcome (0.82) extraction, demonstrating strong overall performance (validation dataset #1). However, glm-4-airx had slightly higher precision in extracting the direction of the relationship, although all models showed moderate performance in this category. For significance extraction, deepseek-chat and GPT-4o-mini exhibited high F1 scores (0.86 and 0.87).

The two-step extraction method generally outperformed the one-step approach, particularly in handling complex indicators like direction and significance. Deepseek-chat was the most reliable model with an F1 score of 0.82 in direction and 0.96 in significance, especially in the two-step approach (Fig. 2a).

Part 2: External validation of human-extracted relationships between dietary factors and coronary heart disease

Additionally, we validated the approach in an external human-extracted dataset (validation dataset #2). In the one-step extraction approach, deepseek-chat achieved F1-scores of 0.76 for exposure, 0.88 for outcome, and 0.67 for association. Glm-4-airx performed consistently with F1-scores of 0.77 for exposure, 0.82 for outcome, and 0.79 for association. Qwen-plus showed competitive results with F1-scores of 0.76 for exposure, 0.84 for outcome, and 0.72 for association.

In the two-step extraction approach, deepseek-chat improved to F1-scores of 0.78 for exposure, 0.88 for outcome, and 0.75 for association. Glm-4-airx demonstrated balanced performance with F1-scores of 0.76 for exposure, 0.85 for outcome, and 0.86 for association. Qwen-plus maintained competitive performance with F1-scores of 0.74 for exposure, 0.88 for outcome, and 0.77 for association (Fig. 2b). Deepseek-chat and glm-4-airx emerged as the most reliable models across both extraction methods.

We also calculated Cost-Effective Ratio (CER, the ratio of mean F1 of all extraction fields divided by cost per million tokens) for each validation dataset. Figure 2 shows the model performance and CER for both datasets (Fig. 2c). The detailed performance of each model in both validation datasets can be found in the Source Data file. Overall, the mean F1 scores in both validations are higher for the two-step approach compared to the one-step approach. Deepseek achieved the best performance with the lowest cost. Thus, we used deepseek-chat for later case studies. Detailed extraction scenarios can be found in Supplementary Note 2. Extracted datasets by each model can be found in our data repository (see Data Availability).

We also compared SemRep[24] and achieved notably lower F1 scores on our validation dataset than any of the LLMs tested, demonstrating the robustness of our two-step LLM-based extraction (see Supplementary Note 3). Moreover, unlike rule-based or static knowledge graph approaches, this LLM-based framework adapts to new terminologies and better infers nuanced causal statements. A more detailed discussion and comprehensive comparison with other text-mining and knowledge graph methods is provided in the Supplementary Note 4.

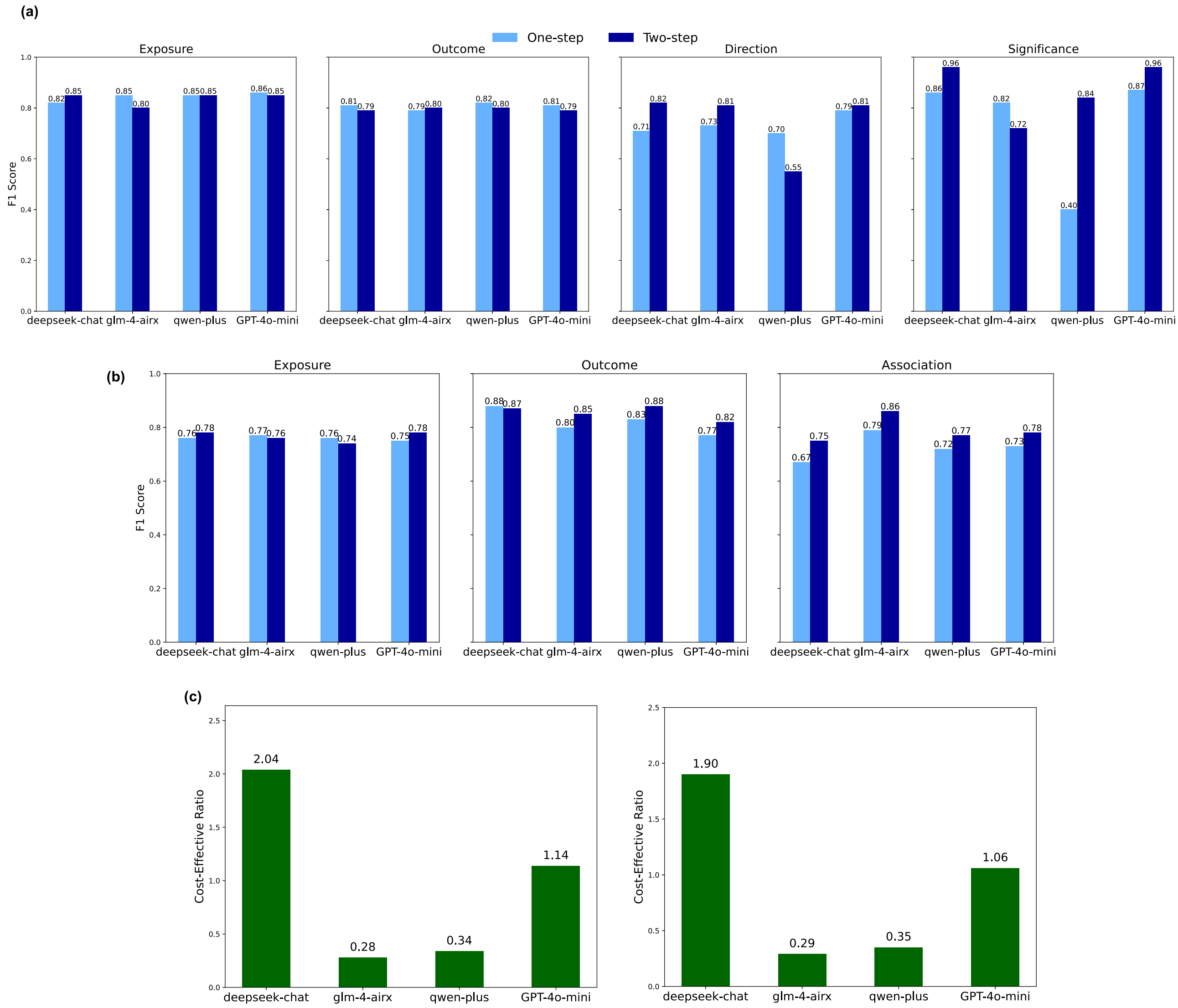

Fig. 2 | Model performance and cost-effective ratios in validation dataset 1&2. Pricing was recorded on 07/31/2024 which was the date for model performance calculation. **a** Shows model performance in the part 1 dataset. **b** Shows model performance in the part 2 dataset. **c** Shows the cost-effective ratio in both datasets.

Head-to-head comparisons with existing evidence synthesis approach

We performed head-to-head comparisons between our method and existing systematic reviews (as the golden standard) in two contentious cases: 1) the effect of salt on blood pressure, and 2) the effect of salt on cardiovascular diseases. These comparisons were made from multiple aspects, including coverage rate of evidence, distribution of the directions of effect, and the ability to track evidence in a timely manner.

Case 1: examining the effect of salt on blood pressure. To validate our automated evidence triangulation approach for assessing the salt–blood pressure association in hypertensive populations, we compared our results with a series of Cochrane Database of Systematic Reviews (CDSR) from 2002 to 2020[25–29]. The CDSR has long been considered a gold-standard reference, with multiple iterations evaluating how salt reduction affects various outcomes including blood pressure. In particular, we focused on studies included in the CDSR that reported salt reduction interventions in hypertensive populations, aiming to determine how well our automatic system captured similar evidence for predefined and targeted population and whether our conclusions aligned with those of traditional meta-analyses.

Coverage rate of CDSR included evidence in our method

From the CDSR, we manually extracted included studies from the forest plots provided across 5 versions of the review, identifying a total of 202 unique relations in 101 primary studies that specifically targeted hypertensive population. Of these 101 primary studies, our automated pipeline identified 80 (79.2%) as containing salt–blood pressure relations in our broader evidence pool, and 67 (66.3%) when strictly focusing on hypertensive populations using matching module. The slight discrepancy is likely due to our reliance on titles and abstracts: some CDSR-included studies may have only specified hypertension in the full text, which our method does not presently parse. Nonetheless, the high coverage indicates that our approach successfully captured most of the key evidence scrutinized by the traditional systematic reviews. Note that we did not refer to the query strategy used in CDSR, and we only queried the PubMed bibliographic database.

Triangulation-ready dataset comparison to CDSR

We generated a triangulation-ready dataset by querying PubMed for RCTs, MRs, and OSs published up to 2024. We developed a module to check whether the extracted exposures and effect directions are consistent with the target research question, ie, 'salt' and 'blood pressure', in this case. Furthermore, we excluded any extracted relations where the "significance" field was marked as "not found," ensuring that our triangulation-ready dataset included only those findings derived directly from study results rather than background or methods sections.

When we examined the reported direction of effect in these studies, we categorized each as 1) "supportive" if the reported effect size was greater than 0 with a confidence interval not crossing 0 (indicating reduced blood pressure with lower salt intake) or 2) "non-supportive" if the effect size was ≤0 or the confidence interval overlapped 0. These criteria mirrored the CDSR's classification. Among the 202 CDSR relations, 49.0% (99/202) were deemed supportive—a moderate endorsement of salt reduction for hypertensive population. The latest CDSR analysis suggests that existing evidence provides only limited support for a causal relationship between reduced salt intake and lowered blood pressure. We propose that this modest observed effect may stem from the inclusion of a significant proportion of non-supportive studies in the evidence synthesis.

In total, 1,954 extracted relations from 942 studies were identified using our method by February 2025. Restricting to studies published up to 2018 produced 1,505 extracted relations from 719 studies, enabling a direct comparison to the CDSR's 2018 cutoff (actual cutoff was April 2018). In our triangulation dataset, 62.8% (945/1505) of evidence is supportive, while in the comparable evidence set included in CDSR, only 49% held a supportive stance.

Convergency of Evidence (CoE)

To quantitatively assess the extent of consensus regarding the effect of salt on blood pressure, we applied our CoE metric. By 2018, the CoE reached 0.441 and remained stable at this level from 2000, indicating a moderate, albeit not overwhelming, consensus that the relation between salt intake and blood pressure is excitatory. This intermediate level of agreement aligns with the traditional meta-analytic findings in the CDSR, suggesting that both methods converged on a similar stance regarding the direction of effect.

After 2018, the addition of new designed studies—particularly four supportive MR analyses—boosted the overall CoE to 0.688 by February 2025. Surpassing two-thirds in the CoE score signifies a more decisive consensus in favor of the excitatory relation, reflecting the impact that more robust or genetically informed evidence (e.g., MR studies) can have on solidifying scientific understanding. See Fig. 3 for evidence matrices and CoE calculation, and Fig. 4 for yearly trend of CoE.

The substantial overlap with CDSR included studies and parallel evolution of conclusions demonstrates that our LLM pipeline—despite relying solely on results reported in abstracts—closely aligns with established systematic review processes.

This automated approach enables the rapid integration of new studies, providing a near real-time snapshot of how scientific consensus shifts as new evidence emerges. This dynamic feature is particularly valuable in understanding and addressing controversies, such as the impact of reduced consumption of salt on blood pressure, where scientific opinion can evolve over time.

While lacking the rigor of full-text scrutiny inherent to systematic reviews/meta-analyses, our pipeline provides efficient preliminary assessments to determine when a systematic review and meta-analysis is warranted. By identifying instances where the CoE remains low or contested, the approach can guide researchers to more rigorously investigate potential sources of heterogeneity or bias.

In summary, comparing our automated system to the authoritative CDSR shows that our approach not only covers most of the core evidence but also reaches broadly consistent conclusions regarding the excitatory effect between salt reduction and blood pressure lowering in hypertensive populations. The iterative integration of new studies—especially from emerging research designs like MR—further strengthens the robustness of these conclusions. It highlights the potential of automated evidence synthesis for validating and updating the scientific consensus on contentious health problems.

Case 2: examining the effect of salt on cardiovascular diseases. To further validate our approach, we extended its application to evaluate the associations between salt intake and cardiovascular diseases (CVD) or mortality. We utilized a manually curated dataset of 82 publications (14 systematic reviews and 68 primary studies) published in 2014, each examining the impact of sodium intake on cerebro-cardiovascular disease or mortality[30,31]. The positions of these publications were originally labeled as "supportive," "contradictory," or "inconclusive" regarding the hypothesis that salt reduction benefits population health[31]; for our analysis, "contradictory" and "inconclusive" were merged into a single "non-supportive" category.

Among the 50 primary studies included in the 14 systematic reviews, 21 (42%) were supportive. When comparing the conclusions of the reviews to the findings of the included studies, we found that 53.7% (101/188) of the time, the direction of effect in the primary studies aligned with the reviews' conclusions, while 46.3% (87/188) of the time they did not. Specifically, reviews classified as supportive included similarly supportive studies 45.0% of the time, whereas non-supportive reviews included non-supportive studies 61.0% of the time. Despite this relatively low rate of alignment, some reviews still reached a supportive conclusion overall. To visualize this potential bias, we plotted a modified evidence inclusion network with stance accordance between reviews and their included primary studies (see Supplementary Fig. 1).

Using our triangulation method, after matching for exposures (salt) and outcomes (CVD or mortality), we generated a dataset of 124 studies with 201 relations by 2014. Considering studies with only one type of relation, 49.0% (55/113) are supportive. We observed an overall CoE of 0.237 by 2014, suggesting a weaker excitatory effect at that time. However, our temporal visualization of CoE indicates that since 2005, the direction and strength of the evidence have steadily converged on an excitatory relationship, suggesting that as more evidence accumulated, the conclusion that salt reduction benefits cardiovascular outcomes became more stable (Fig. 5).

Thus, while readers of these 14 systematic reviews might perceive conflicting results, our automated approach reveals an emerging consensus. Even though the body of evidence has continued to grow in multiple directions since 2005, the net effect has stabilized in favor of salt reduction conferring cardiovascular benefits.

Search queries for evidence triangulation of both comparisons are provided in Supplementary Note 5.

According to analyses on two cases, evidence demonstrates strong consensus regarding the beneficial impact of salt reduction on blood pressure lowering, reflected by high CoE. In contrast, salt reduction's protective effects against cardiovascular disease and all-cause mortality remain subjects of ongoing debate in the literature, manifesting as moderate CoE.

Discussion

This study illustrates the potential of using LLMs to automate the extraction and triangulation of SDoH-related evidence across diverse study designs. Our approach utilizes a two-step extraction method, which first sequences exposure-outcome concept extraction followed by relation extraction, exhibiting better performance over the one-step method. This strategy essentially functions as a triple extractor, capturing entities and the relations between them, similar to the SemRep system that extracts Subject-PREDICATE-Object semantic triples from biomedical text using a rule-based approach. The related ASQ platform has provided a user-friendly way to query SemRep-extracted triples along with associated evidence sentences, contributing to a novel approach for evidence triangulation. However, the extraction performance of SemRep is limited due to its reliance on a

		increase (+)	no change (0)	decrease (-)
RCT increased	SD_1	91	87	77
MR increased	SD_2	2	0	0
OS increased	SD_3	519	189	39
RCT decreased	SD_4	22	137	386
MR decreased	SD_5	0	0	2
OS decreased	SD_6	41	58	108

$$\begin{aligned}\mathrm{P}(\mathcal{E}) = \left(\tfrac{1}{6}\right)\times\Big[&\left(\frac{91\times0.48}{(91\times0.48)+(87\times0.34)+(77\times0.19)}\right)_{SD_1}+\left(\frac{2}{2+0.1+0.1}\right)_{SD_2}+\left(\frac{519\times0.71}{(519\times0.71)+(189\times0.24)+(39\times0.05)}\right)_{SD_3}+\\ &\left(\frac{386\times0.80}{(386\times0.80)+(137\times0.18)+(22\times0.03)}\right)_{SD_4}+\left(\frac{2}{2+0.1+0.1}\right)_{SD_5}+\left(\frac{108\times0.44}{(108\times0.44)+(58\times0.31)+(41\times0.24)}\right)_{SD_6}\Big] = 0.792\end{aligned}$$

$$\begin{aligned}\mathrm{P}(\mathcal{N}) = \left(\tfrac{1}{6}\right)\times\Big[&\left(\frac{87\times0.34}{(91\times0.48)+(87\times0.34)+(77\times0.19)}\right)_{SD_1}+\left(\frac{0.1}{2+0.1+0.1}\right)_{SD_2}+\left(\frac{189\times0.24}{(519\times0.71)+(189\times0.24)+(39\times0.05)}\right)_{SD_3}+\\ &\left(\frac{137\times0.18}{(386\times0.80)+(137\times0.18)+(22\times0.03)}\right)_{SD_4}+\left(\frac{0.1}{2+0.1+0.1}\right)_{SD_5}+\left(\frac{58\times0.31}{(108\times0.44)+(58\times0.31)+(41\times0.24)}\right)_{SD_6}\Big] = 0.142\end{aligned}$$

$$\begin{aligned}\mathrm{P}(\mathcal{J}) = \left(\tfrac{1}{6}\right)\times\Big[&\left(\frac{77\times0.19}{(91\times0.48)+(87\times0.34)+(77\times0.19)}\right)_{SD_1}+\left(\frac{0.1}{2+0.1+0.1}\right)_{SD_2}+\left(\frac{39\times0.05}{(519\times0.71)+(189\times0.24)+(39\times0.05)}\right)_{SD_3}+\\ &\left(\frac{22\times0.03}{(386\times0.80)+(137\times0.18)+(22\times0.03)}\right)_{SD_4}+\left(\frac{0.1}{2+0.1+0.1}\right)_{SD_5}+\left(\frac{41\times0.24}{(108\times0.44)+(58\times0.31)+(41\times0.24)}\right)_{SD_6}\Big] = 0.066\end{aligned}$$

direction = Excitatory

$$\textbf{CoE score} = \frac{0.792-(\frac{1}{3})}{1-(\frac{1}{3})} = \textbf{0.688 (strong LoC)}$$

Fig. 3 | Triangulation result for Case 1: the effect of salt on blood pressure. Evidence matrices integrating multiple study designs (RCTs, MRs, and OSs) and detailed evidence triangulation algorithms assessing the effect of salt intake on blood pressure, including related studies published through February 2025.

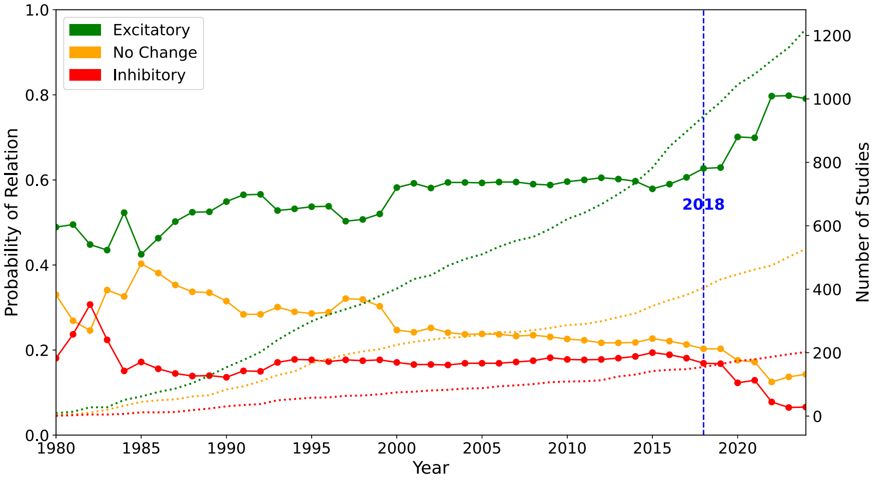

Fig. 4 | Convergency of Evidence (CoE) for the cumulative body of related studies assessing the effect of salt on blood pressure, shown by publication year. The blue dashed line at 2018 indicates the cutoff year for inclusion in the latest Cochrane Database of Systematic Reviews (CDSR). Greens lines represent 'excitatory', yellow lines represent 'no change', and red lines represent 'inhibitory'. Solid lines with circle markers represent the cumulative probability of relations, and dotted lines represent the cumulative number of studies.

rule-based method developed two decades ago, despite ongoing updates and extensions with relation classification approaches[24,32,33]. Through data evaluations using CoE and LoC with a focused case study on the impact of salt intake on blood pressure, we demonstrated that LLMs can remarkably simplify the synthesis of medical evidence, enhancing the efficiency of evidence-based decision-making. After triangulating evidence from different study designs, the relationship between salt intake levels and blood pressure tends to indicate an excitatory effect direction. However, the LoC for this effect just reaches 'strong', suggesting that there are still some contradictory research findings. This result is consistent with the evidence status of salt controversy discussed in recent years[31,34,35].

This study also distinguishes our approach from others that utilize LLMs for meta-analysis. While decision-making should ideally be grounded in causal relationships between interventions and outcomes, predictions can be based on correlative relationships. Our objective was not to automate meta-analysis, which focuses on evidence derived from the same study design, such as RCTs. Recent proof-of-concept studies have shown that LLMs like Claude 2, Bing AI, and GPT-4 can improve the efficiency and accuracy of data extraction for evidence syntheses[36–38]. However, these studies often involve limited datasets, either based on a single case[37] or a small sample size of RCTs[36,38]. Although key data elements necessary for evidence triangulation, such as primary outcomes and effect estimates, can be accurately extracted in over 80% of RCTs, the performance of LLMs remains suboptimal with larger datasets. Additionally, other studies have evaluated the sensitivity and specificity of LLMs like GPT-3.5 Turbo in tasks such as title and abstract screening. The findings suggest that while these models offer promise, they are not yet sufficient to replace manual screening entirely[39]. On the other hand, GPT-4 Turbo-assisted citation screening has shown potential as a reliable and time-efficient alternative to systematic review processes[40]. With these technological advances, Mayo Clinic has proposed an AI-empowered integrated framework for living, interactive systematic reviews and meta-analyses, enabling continuous, real-time evidence updates[41,42].

In contrast, our focus is on utilizing LLMs to perform convergency analysis among results obtained from different study designs, known as triangulation analysis. The key difference between meta-analysis

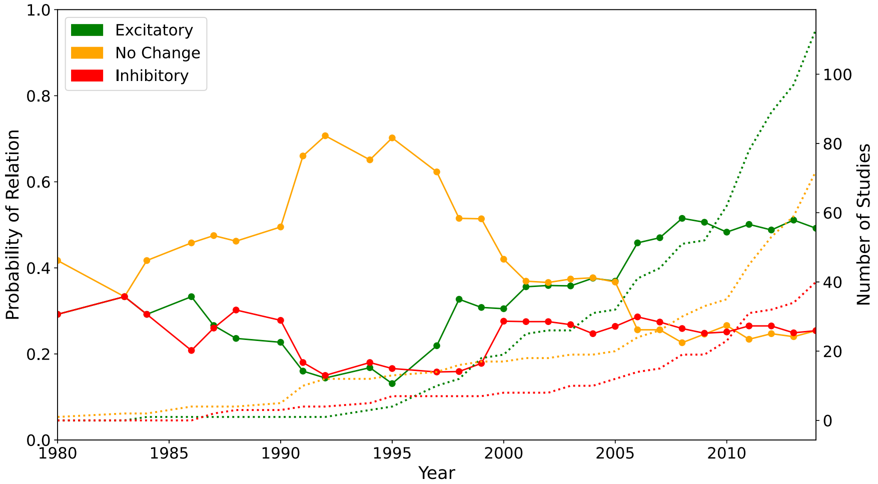

Fig. 5 | Convergency of Evidence (CoE) for the cumulative body of related studies assessing the effect of salt on cardiovascular diseases or mortality (Case 2), shown by publication year. For comparison with existing systematic reviews, only data up to 2014 are shown here. Greens lines represent 'excitatory', yellow lines represent 'no change', and red lines represent 'inhibitory'. Solid lines with circle markers represent the cumulative probability of relations, and dotted lines represent the cumulative number of studies.

and triangulation analysis lies in their focus: while meta-analysis assesses consistency within a single study design, triangulation analysis examines the convergency of conclusions across diverse study designs. Although there are no widely accepted quantitative methods for assessing convergency, insights can be drawn from convergency analysis, originated from neurobiological studies, primarily using a vote-counting approach across different study designs[43–45]. Lastly, in causal graphical models, the concept of a causal relationship is uniform across different types of variables. Whether the graph pertains to biological or economic phenomena, the underlying principles of causality remain the same[46]. This perspective also applies to the method in this study, that evidence should be and could be triangulated with CoE and LoC to conclude a causal relationship, whether in biomedical, economic, and environment field. As a result of constantly evolving research, the CoE ratings may change as more research findings becomes available. This is particularly the case for exposure-outcome pairings with low LoC due to contradictory results. Our approach is to harmonize confusion and help consumers make informed decisions about diet, exercise, and other activities that can affect their long-term health, as well as help researchers shape future clinical studies.

Moreover, we argue that the strength of triangulation lies in its ability to address issues related to ranking and the exclusion of studies based on perceived quality. Rather than being a method for assessing the weight of evidence, triangulation is a strategy for integrating evidence from diverse approaches. It is not merely about gathering all available evidence from different sources, nor is it solely concerned with evaluating heterogeneity or effect measure modification. Instead, triangulation involves a thoughtful and explicit consideration of the various biases that can lead studies in different directions. In evidence syntheses, inclusivity is prioritized because even clearly flawed studies may introduce biases that are predictable in the direction of effect. These biased studies can be contrasted with others that offer a more consistent and unbiased estimate on the direction of effect[47]. In fact, triangulation synthesizes the direction of effects, rather than the effect size reported in individual studies. Also, it was suggested that observational studies and RCTs should not be combined in the same meta-analysis[48], and extra caution is needed when interpreting the results of meta-analyses that combine effect values from both types of studies[49]. In contrast, our approach combines structure instead of value, providing a solution that minimizes the bias introduced by integrating results from different study designs.

In general, if individual study evidence does not properly control for confounding variables, it can lead to an overestimation of effect sizes. When these overestimated effect sizes are subsequently combined in a meta-analysis, the effect size is further exaggerated. Existing studies have suggested that when causal modeling is guided by causal directed acyclic graphs, evidence from a study that has not properly controlled for confounding variables should be excluded and not included in a meta-analysis. The triangulation approach, which synthesizes structure rather than data, can help avoid this type of bias[50,51].

Limitations

This study faced several limitations, including difficulties in accurately classifying study designs and interpreting associations due to data inconsistencies. The extracted entities were not mapped to standard biomedical vocabularies like SNOMED CT[52] or UMLS[53], leading to potential misalignment and incorrect relation pairing, which could affect the final LoC. Furthermore, not all relevant study designs were included, limiting the comprehensiveness of the conclusions. The reliance on expert annotations also introduced subjective bias, potentially affecting the generalizability of the findings. While the two-step extraction approach showed improved performance, it requires further refinement to handle the complexity and variability of biomedical data effectively.

To maximize the utility of LLMs in evidence triangulation, future work should focus on addressing these limitations through continuous model fine-tuning and the development of more objective evaluation methods. A critical challenge remains in harmonizing evidence, including standardizing study populations and cause-and-effect entities across different study designs. The study aims to use LLMs to extract and align these elements through concept similarity measures like BERTScore, rather than relying on superficial string-based matches. Future studies will also introduce biomedical ontologies to better map the hierarchical structure of cause-and-effect concepts, leading to a more standardized and comprehensive approach to 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. The workflow ends with a quantitative evidence triangulation algorithm to discover the LoC behind a relationship between a SDoH factor and a health outcome.

Data sources

Validation dataset #1. 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, 4268 articles were retrieved. This first dataset consists of 100 randomly selected studies from the corpus, used to validate the results extracted by LLM. The extracted results dataset and validation dataset are provided (see Data Availability). Entities and relations are manually annotated by 4 domain experts and clinicians (see Acknowledgement for details).

Validation dataset #2. The dataset used for external validation consists of 291 human-extracted relations between dietary factors and coronary heart disease, derived from the Nurses' Health Study[54]. It includes a wide range of dietary exposures, such as specific nutrients and food items, and their associations with cardiovascular outcomes. This data was meticulously curated and visualized in a knowledge graph, capturing both positive and negative associations, as well as effect size (hazard ratio, risk ratio, odds ratio, etc.) which can be served as a critical external foundation for testing the two-step extraction approach. The extracted results dataset and validation dataset are provided (see Data Availability).

LLM-based study results extraction

For the task of extracting precise and insightful results from health-related documents, we employed medium-tier LLMs, which includes deepseek-chat[55], glm-4-airx[56], qwen-plus[57], and GPT-4-mini[58]. While these models were not the top performers in all metrics, they were selected as a compromise, balancing both extraction performance and economic accessibility, such as cost per token. This balance makes them suitable for large-scale extraction tasks where both accuracy and cost-effectiveness are critical. The cost-effectiveness analysis can be found in the Results section.

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

Methodological information:

- Identification of study design
  The initial step involves using LLM to categorize the study design present in medical abstracts. The designs considered include RCT, MR and OS.

Ontological information:

- Primary result identification
  Next, we ask LLM 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. The model also identifies the direction (increased or decreased) of intervention/exposure for later relation alignment.
- Relation and statistical significance
  First the model extracts the direction of the relation 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 LLM to extract statistical significance of the identified relation, 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, relation direction, statistical significance), then contextual details (population, participant number, comparator). Figure 6 shows a graphical illustration of the overflow and logics of the designed prompt. For each abstract, LLM first determines the study design. Subsequently, it locates the primary result, extracts relevant details about the intervention/exposure and outcome, and assesses the direction and statistical significance of the relation. 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.

To enhance the accuracy and robustness of entity and relation extraction, we implemented and compared a two-step extraction pipeline with a direct one-step extraction approach (Fig. 7). The one-step extraction method simultaneously identifies and extracts both entities (e.g., exposures and outcomes) and their relations directly from the text. In contrast, the two-step extraction process separates these tasks: the first step involves using NER prompt to identify and extract entities from the text, such as specific dietary factors and cardiovascular outcomes. In the second step, these extracted entities are then used to identify and extract the relations among them using RE prompt. This sequential approach allows for more precise entity recognition before relation extraction, potentially reducing errors and improving overall extraction accuracy. Full prompts and code implementations for both the one-step and two-step extraction methods can be found at the GitHub repository, providing a comprehensive guide for replicating these processes.

Convergency of evidence

Originally developed for neurobiological studies, ResearchMap provides a graph-based representation of empirical evidence, enabling the systematic quantification of "convergency" and "consistency" across diverse study designs. It uses a Bayesian scoring method called the Cumulative Evidence Index (CEI) to evaluate the strength of causal relationships. By drawing from its scoring strategy, our work adapts similar principles to develop the Convergency of Evidence (CoE) and Level of Convergency (LoC) metrics in a public health setting[43–45].

To systematically evaluate the direction of effect for exposure-outcome associations derived from multiple lines of evidence across various study designs, our proposed framework quantifies the greatest likelihood of three possible relations, i.e., excitatory, no change, and inhibitory, reported in the evidence.

1. Excitatory (ε)
   Occurs when the outcome changes in the same direction as the exposure or intervention. For instance, if an increased salt intake increases blood pressure or a decreased salt intake decreases blood pressure, we label the direction of effect as excitatory.

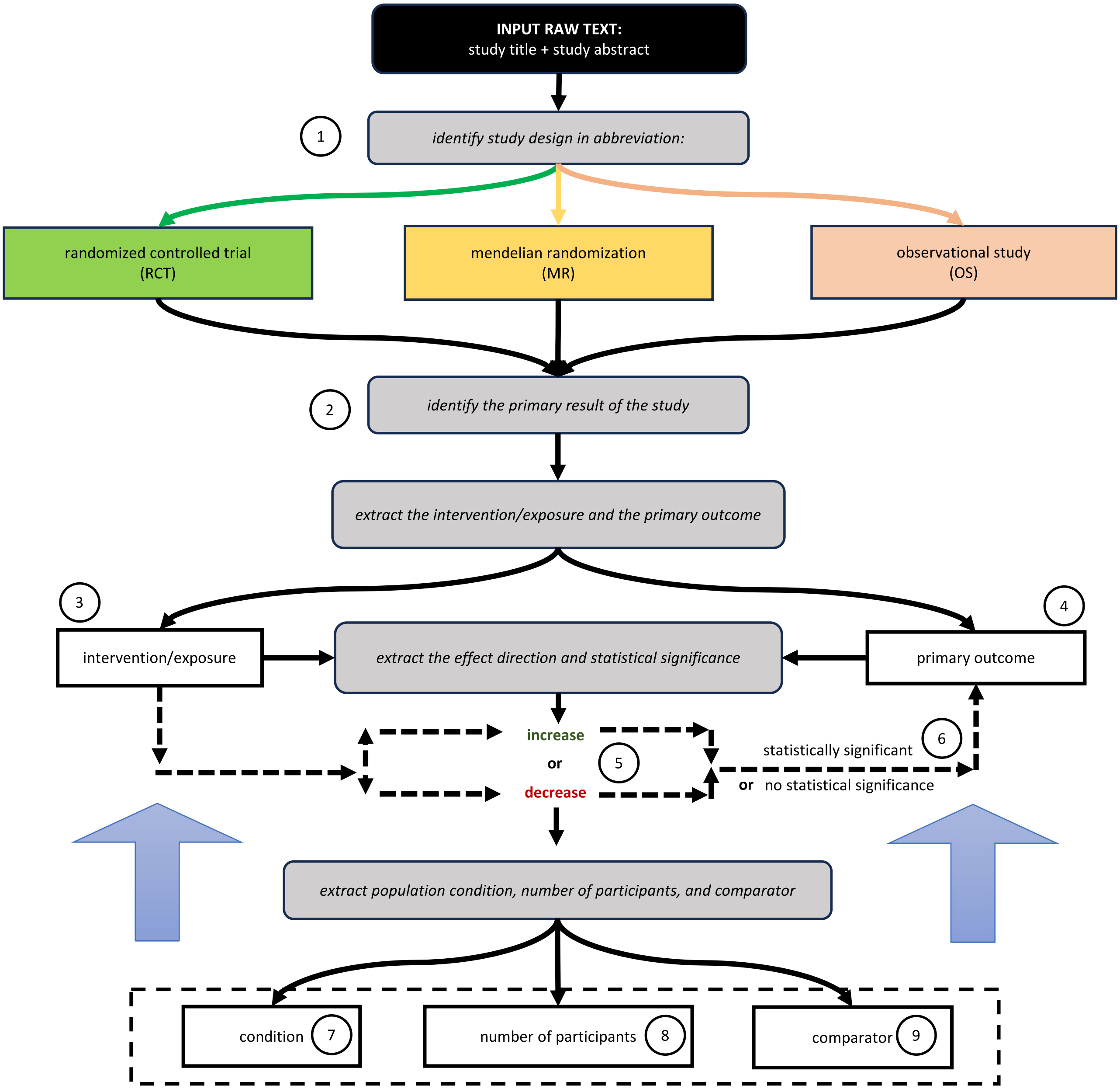
INPUT RAW TEXT:
study title + study abstract
1
identify study design in abbreviation:
randomized controlled trial (RCT)
mendelian randomization (MR)
observational study (OS)
2
identify the primary result of the study
extract the intervention/exposure and the primary outcome
3
intervention/exposure
extract the effect direction and statistical significance
4
primary outcome
increase
or
decrease
5
statistically significant
6
or no statistical significance
extract population condition, number of participants, and comparator
condition
7
number of participants
8
comparator
9

Fig. 6 | 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 fields by the model.

2. No Change ($\mathcal{N}$)
   Refers to findings that indicate the exposure or intervention does not lead to a statistically significant change in the outcome (e.g., salt intake has no measurable impact on blood pressure levels).
3. Inhibitory ($\mathcal{J}$)
   Occurs when the outcome changes in the opposite direction of the exposure or intervention. For example, if an increased salt intake decreases blood pressure, we label this effect as inhibitory.

To reduce bias by study size, we enhanced our triangulation algorithm to better differentiate weight of evidence by incorporating the number of participants from RCT and OS. We intentionally excluded participant counts from MR, as MR relies on genetic variants rather than traditional enrollment metrics, making participant numbers less directly applicable within this specific framework. This framework incorporates participant contributions by applying a participant weight to the relations counts and utilizes Laplace smoothing (or named Lidstone smoothing) to address zero-frequency problems.

We also excluded relations within the 'significance' field labeled as 'not found' to ensure that only 'empirical evidence' stated in study results with statistical test information is included to the final evidence triangulation measurement. Relations extracted from statement lacking statistical significance cannot be so-called evidence, as they are often hypothetical claims found in background or method section, rather than actual findings.

Probability estimation for the direction of effect. Let P($\mathcal{E}$), P($\mathcal{N}$), and P($\mathcal{J}$) represent the probabilities of the observed excitatory, no change, and inhibitory effects, respectively. These probabilities are computed by aggregating weighted contributions across six distinct study design $\left(SD_i, where\ i=1,2,\ldots,6\right)$ to present convergency. Below is the detailed process of the proposed algorithm.

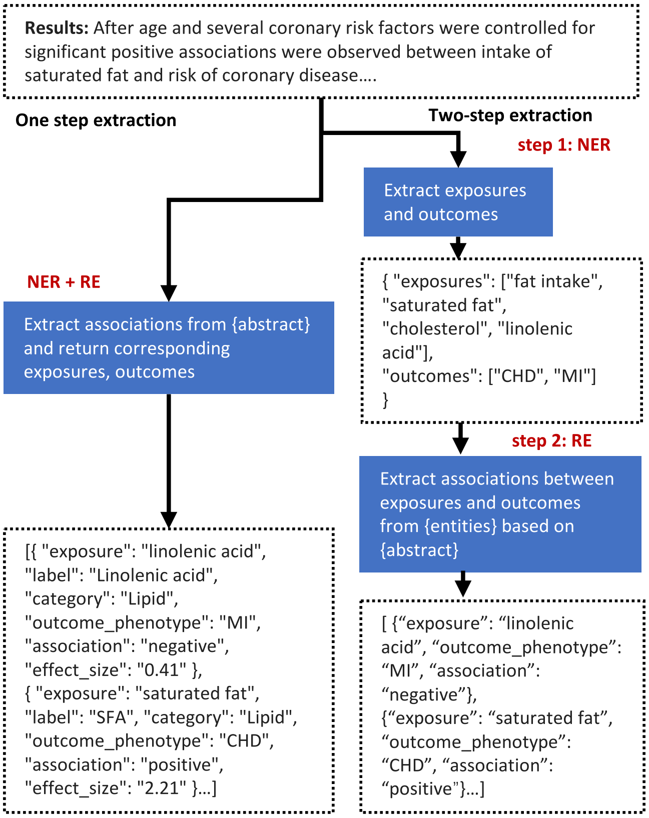
Results: After age and several coronary risk factors were controlled for significant positive associations were observed between intake of saturated fat and risk of coronary disease….
One step extraction
Two-step extraction
step 1: NER
Extract exposures and outcomes
NER + RE
Extract associations from {abstract} and return corresponding exposures, outcomes
{ "exposures": ["fat intake", "saturated fat", "cholesterol", "linolenic acid"], "outcomes": ["CHD", "MI"] }
step 2: RE
Extract associations between exposures and outcomes from {entities} based on {abstract}
[{ "exposure": "linolenic acid", "label": "Linolenic acid", "category": "Lipid", "outcome_phenotype": "MI", "association": "negative", "effect_size": "0.41" }, { "exposure": "saturated fat", "label": "SFA", "category": "Lipid", "outcome_phenotype": "CHD", "association": "positive", "effect_size": "2.21" }…]
[{"exposure": "linolenic acid", "outcome_phenotype": "MI", "association": "negative"}, {"exposure": "saturated fat", "outcome_phenotype": "CHD", "association": "positive"}…]

Fig. 7 | An illustration of LLM prompts in one-step and two-step NER and RE from study results. On the left is the one-step extraction approach, which extracts both entities and relations simultaneously. On the right is the two-step extraction approach, which first identifies entities then determines the relations.

Each combination SD_i represents a specific pairing of study design and exposure direction:

- SD_1 : RCT – increasing level of intervention
- SD_2 : MR – increasing level of exposure
- SD_3 : OS – increasing level of exposure
- SD_4 : RCT – decreasing level of intervention
- SD_5 : MR – decreasing level exposure
- SD_6 : OS- decreasing level exposure

For each SD_i, the relation counts are adjusted by the number of participants in each study as the weight (PW_{iX}) and smoothed using Laplace smoothing with a smoothing factor of $\alpha = 0.1$ to prevent zero-probability relations.

Probability of an excitatory effect. First, define the **adjusted count** $\widetilde{C_{i\mathcal{X}}}$ for each study design SD_i and relation type $\mathcal{X} \in \{\mathcal{E}, \mathcal{N}, \mathcal{J}\}$ as:

$$\widetilde{C_{\rangle\mathcal{X}}} = \begin{cases} C_{i\mathcal{X}} \times PW_{i\mathcal{E}} + \alpha, \text{if } C_{i\mathcal{X}} \times PW_{i\mathcal{E}} = 0, \\ C_{i\mathcal{X}} \times PW_{i\mathcal{E}} \qquad \mathit{otherwise} \end{cases} \tag{1}$$

Then the probability for $\mathrm{P}(\mathcal{E})$ becomes

$$\mathrm{P}(\mathcal{E}) = \frac{1}{6}\sum_{i=1}^{6} \frac{\widetilde{C_i\mathcal{E}}}{\widetilde{C_i\mathcal{E}} + \widetilde{C_i\mathcal{N}} + \widetilde{C_i\mathcal{J}}} \tag{2}$$

$$\mathrm{P}(\mathcal{E}) = \frac{1}{6}\sum_{i=1}^{6} \frac{C_{i\mathcal{E}} \times PW_{i\mathcal{E}} + \alpha}{C_{i\mathcal{E}} \times PW_{i\mathcal{E}} + \alpha + C_{i\mathcal{N}} \times PW_{i\mathcal{N}} + \alpha + C_{i\mathcal{J}} \times PW_{i\mathcal{J}} + \alpha} \tag{3}$$

Where:

- $C_{i\mathcal{E}}$ is the count of excitatory relations in study design SD_i
- $PW_{i\mathcal{E}}$ is the participant weight for the excitatory relation in SD_i (1 for MR)
- $\alpha = 0.1$ is the Laplace smoothing factor, used to handle zero-frequency occurrences

Probability of no change.

$$\mathrm{P}(\mathcal{N}) = \frac{1}{6}\sum_{i=1}^{6} \frac{C_{i\mathcal{N}} \times PW_{i\mathcal{N}} + \alpha}{C_{i\mathcal{E}} \times PW_{i\mathcal{E}} + \alpha + C_{i\mathcal{N}} \times PW_{i\mathcal{N}} + \alpha + C_{i\mathcal{J}} \times PW_{i\mathcal{J}} + \alpha} \tag{4}$$

Probability of an inhibitory effect.

$$\mathrm{P}(\mathcal{J}) = \frac{1}{6}\sum_{i=1}^{6} \frac{C_{i\mathcal{J}} \times PW_{i\mathcal{J}} + \alpha}{C_{i\mathcal{E}} \times PW_{i\mathcal{E}} + \alpha + C_{i\mathcal{N}} \times PW_{i\mathcal{N}} + \alpha + C_{i\mathcal{J}} \times PW_{i\mathcal{J}} + \alpha} \tag{5}$$

Participant size adjustment. To mitigate disproportionate weighting from studies with extremely large populations, we restricted to the 5th–95th percentile range. This adjustment reduces bias introduced by large-scale studies, such as PMID 24854239[59], which surveyed approximately 96 million participants from an entire province in China. If weighted directly by such a large population, the probability of the corresponding relation could disproportionately dominate the overall analysis, leading to skewed interpretations. Limiting the participant weight between 5th-95th ensures that no single study can disproportionately influence the calculated probabilities.

The participant weight (PW) is defined as:

$$PW_{iX} = \frac{\mathrm{AdjNumPart}_{iX}}{\sum Y \mathrm{AdjNumPart}_{iY}} \tag{6}$$

- PW_{iX}: Participant weight for outcome X in study design i
- AdjNumPartiX: Adjusted number of participants:
  - For non-MR studies: Number of participants is restricted to the 5th-95th percentile range.
  - For MR studies: Original number of participants is retained (unchanged).
- The denominator sums over all outcomes $Y (Y \in \{\text{Increase}, \text{NoChange}, \text{Decrease}\})$ within the same study design.

To quantify the strength of convergence toward a specific directional relation, we introduce the Convergency of Evidence (CoE) and Level of Convergency (LoC). This metric assesses the dominance of the most probable relation relative to an uninformed baseline of $\frac{1}{3}$ (the probability of a relation before any study conducted (no prior knowledge)).

$$\mathrm{CoE} = \frac{\max(P(\mathcal{E}), P(\mathcal{N}), P(\mathcal{J})) - \frac{1}{3}}{1 - \frac{1}{3}} \tag{7}$$

The LoC value ranges from 0 (indicating no convergency toward any relation) to 1 (indicating complete convergence on a single relation). The LoC is further categorized to reflect the strength of the evidence:

- Weak Convergency: $0 \le \mathrm{LoC} \le 0.3$
- Moderate Convergency: $0.3 < \mathrm{LoC} \le 0.6$
- Strong Convergency: $0.6 < \mathrm{LoC} \le 1.0$

Explanation of the rationale for choosing 0.3 and 0.6 as the LoC thresholds

1. Heuristic Partitioning
   The numeric cutoffs at 0.3 and 0.6 serve as a practical, rule-of-thumb way to divide the 0 to 1 scale into "weak," "moderate," and "strong" convergence. By setting three tiers, researchers can quickly gauge how close the evidence is to a consensus—whether it's only slightly better than random chance, moderately conclusive, or approaching unanimity.
2. Adaptation from ResearchMap's Cumulative Evidence Index
   The concept of "strong" vs. "weak" evidence broadly parallels frameworks introduced in the neurobiological research tool ResearchMap, which uses a similar Bayesian measure of convergence. Although ResearchMap does not dictate specific numeric cutoffs, our thresholds reflect a similar logic—where 0 to 0.3 indicates minimal convergence and above 0.6 signals robust alignment of results.
3. Flexibility for Domain-Specific Needs
   These thresholds are not intended as absolute. Researchers focused on highly conservative domains (e.g., drug safety) might lower the threshold for "strong" to 0.5, while more exploratory fields might place it at 0.8. Thus, the 0.3 and 0.6 cutoffs are primarily a pragmatic starting point for identifying levels of consensus in most general settings, yet can be fine-tuned to align with specific fields or study distributions.

Dominant directional relation. The relation with the highest probability is identified as the dominant directional relation:

$$direction = argmax P(\mathcal{E}), P(\mathcal{N}), P(\mathcal{J}) \tag{8}$$

Note that we categorize each study by design type (MR, RCT, OS) but do not apply design-specific weighting when computing CoE/LoC, ensuring a transparent and standardized triangulation process. This approach prioritizes identifying whether different lines of evidence align or conflict rather than assessing their relative strength, avoiding additional assumptions about study hierarchy. While RCTs are often considered stronger than observational studies, our method treats all designs equally, balancing the inherent biases of each study design.

Advantages of proposed method. To track which newly published studies notably influence the CoE score, we also developed a specialized module. Specifically, for each year in which the CoE shifts by more than 0.1 or the predominant direction of effect changes, the module automatically flags and outputs the largest studies within each relation category (excitatory, inhibitory, or no change). This provides researchers a quick overview of the high-impact studies from that year, allowing them to more easily interpret and validate why the overall CoE has changed. Examples of this module's output are provided in the Supplementary Note 6.

Our approach brings together three key advantages—breadth, continuity, and adaptability—that traditional meta-analyses or other design-specific methods typically cannot provide:

1. Breadth of Evidence
   Rather than relying solely on RCTs or a specific study design, the CoE/LoC framework draws in findings from all relevant sources—observational studies, mendelian randomization, or smaller-scale trials. This inclusivity guards against overlooking pivotal data (e.g., contradictory or null results) that might not fit stringent meta-analysis inclusion criteria. By weaving multiple lines of evidence into a unified assessment, the method produces a more well-rounded understanding of the causal landscape.
2. Temporal Evolution
   Meta-analyses and systematic reviews often act as snapshots, valid only for the range of studies available at a given time. In contrast, our dynamic pipeline updates its convergency score whenever new studies appear, thereby capturing real-time shifts in the evidence base. In the case study of the association between salt intake and blood pressure, the gradual aggregation of multiple designs has steadily reinforced an excitatory effect—providing a granular, year-by-year progression that showcases how and when consensus grows. This evolving picture lets clinicians, guideline committees, and policymakers see trends emerging, rather than waiting for a new static meta-analysis cycle.
3. Adaptability to Contradictory Findings
   Traditional meta-analyses often reduce complexities to a single pooled estimate and may downweigh or exclude heterogeneous results, especially if the analytic model (e.g., fixed-effect or random-effect) cannot reconcile significant between-study discrepancies. Our scoring algorithm, by contrast, intrinsically accommodates and highlights discrepancies among diverse study designs, helping detect signals that warrant closer scrutiny. For example, if observational data suggest a relationship is weaker or nonexistent while mendelian randomization findings strongly support it, the CoE/LoE framework visibly reflects that tension—giving clinicians actionable insight into ongoing controversies.

Overall, this dynamic, all-inclusive methodology does more than simply tally results: it follows how evidence matures, intensifies, or—if contradictory studies accumulate—weakens over time. For decision-makers in evidence-based medicine, having this evolving synthesis can be highly practical, providing a quick overview of the status of body of evidence before conducting a detailed meta-analysis or systematic review. It points to where the field is converging, how fast lines of research might challenge prevailing wisdom. Consequently, clinicians and policymakers can respond more quickly to emerging consensus, or pivot sooner if strong counter-evidence surfaces.

Statistics & Reproducibility

No traditional analyses are used in this study, thus no statistical method was used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. However, we proposed a new analysis algorithm that involves data extraction from unstructured data for: study exposure, outcome, direction of relation, statistical significance, population, study design.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Source Data of tables are provided as an attachment with this paper. And other related data including example input data for information extraction, extracted output data by different LLMs, and organized dataset for the two cases of evidence triangulation can be found in the Figshare repository, https://doi.org/10.6084/m9.figshare.28514210. Source data are provided with this paper.

Code availability

The codes generated during and analyzed during the current study are available in the GitHub repository, https://github.com/xuanyshi/llm-evidence-triangulation. Cite this article if any data or code re-used: https://doi.org/10.5281/zenodo.15781463.

References

1. Ali, M. K., Sudharsanan, N. & Thirumurthy, H. Behaviour change in the era of biomedical advances. *Nat. Hum. Behav.* **7**, 1417–1419 (2023).
2. Roth, G. A. et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *J. Am. Coll. Cardiol.* **76**, 2982–3021 (2020).
3. Hernán, M. A., Hsu, J. & Healy, B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE* **32**, 42–49 (2019).
4. Munafò, M. R. & Davey Smith, G. Robust research needs many lines of evidence. *Nature* **553**, 399–401 (2018).
5. Editorial, Assessing the evidence of risk. *Nat. Med.* **28**, 1967–1967 (2022).
6. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018).
7. Lawlor, D. A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *Int. J. Epidemiol.* **45**, 1866–1886 (2016).
8. Munafò, M. R., Higgins, J. P. T. & Smith, G. D. Triangulating Evidence through the Inclusion of Genetically Informed Designs. *Cold Spring Harb. Persp. Med.* **11** https://doi.org/10.1101/cshperspect.a040659 (2021).
9. Bailey, D. H. et al. Causal inference on human behaviour. *Nat. Hum. Behav.* **8**, 1448–1459 (2024).
10. Pearl, J. & Mackenzie, D. *The book of why: the new science of cause and effect*. (Basic books, 2018).
11. Sae-Jie, W. et al. Triangulating evidence from observational and Mendelian randomization studies of ketone bodies for cognitive performance. *BMC Med.* **21**, 340 (2023).
12. Joseph, S. A. et al. FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (*2024)
13. Wadhwa, S., DeYoung, J., Nye, B., Amir, S. & Wallace, B. C. Jointly extracting interventions, outcomes, and findings from RCT reports with LLMs, *Machine Learning for Healthcare Conference*, 754–771 (2023)
14. Ghosh, M. et al. AlpaPICO: Extraction of PICO frames from clinical trial documents using LLMs. *Methods* **226**, 78–88 (2024).
15. Mutinda, F., Liew, K., Yada, S., Wakamiya, S. & Aramaki, E. in *Proceedings of the first Workshop on Information Extraction from Scientific Publications*. 26-31.
16. Hu, Y., Keloth, V. K., Raja, K., Chen, Y. & Xu, H. Towards precise PICO extraction from abstracts of randomized controlled trials using a section-specific learning approach. *Bioinformatics* **39**, btad542 (2023).
17. Zhang, G. et al. A span-based model for extracting overlapping PICO entities from randomized controlled trial publications. *J. Am. Med. Inform. Assoc.* **31**, 1163–1171 (2024).
18. Marshall, I. J. et al. Trialstreamer: A living, automatically updated database of clinical trial reports. *J. Am. Med. Inform. Assoc.* **27**, 1903–1912 (2020).
19. Kang, T. et al. EvidenceMap: a three-level knowledge representation for medical evidence computation and comprehension. *J. Am. Med. Inform. Assoc.* **30**, 1022–1031 (2023).
20. Mayer, T., Marro, S., Cabrio, E. & Villata, S. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artif. Intell. Med.* **118**, 102098 (2021).
21. Whitton, J. & Hunter, A. Automated tabulation of clinical trial results: A joint entity and relation extraction approach with transformer-based language representations. *Artif. Intell. Med.* **144**, 102661 (2023).
22. Liu, Y. & Gaunt, T. R. Triangulating evidence in health sciences with Annotated Semantic Queries. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btae519 (2024).
23. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
24. Kilicoglu, H., Rosemblat, G., Fiszman, M. & Shin, D. Broad-coverage biomedical relation extraction with SemRep. *BMC Bioinforma.* **21**, 1–28 (2020).
25. Jürgens, G. & Graudal, N. Effects of low sodium diet versus high sodium diet on blood pressure, renin, aldosterone, catecholamines, cholesterols, and triglyceride. *The Cochrane database of systematic reviews*, CD004022 (2003).
26. Jürgens, G. & Graudal, N. A. Effects of low sodium diet versus high sodium diet on blood pressure, renin, aldosterone, catecholamines, cholesterols, and triglyceride. *Cochrane Database of Systematic Reviews*, CD004022.pub2 (2004).
27. Graudal, N. A., Hubeck-Graudal, T. & Jurgens, G. Effects of low sodium diet versus high sodium diet on blood pressure, renin, aldosterone, catecholamines, cholesterol, and triglyceride. *The Cochrane database of systematic reviews*, CD004022.pub3 (2011).
28. Graudal, N. A., Hubeck-Graudal, T. & Jurgens, G. Effects of low sodium diet versus high sodium diet on blood pressure, renin, aldosterone, catecholamines, cholesterol, and triglyceride. *Cochrane Database of Systematic Reviews*, CD004022.pub4 (2017).
29. Graudal, N. A., Hubeck-Graudal, T. & Jurgens, G. Effects of low sodium diet versus high sodium diet on blood pressure, renin, aldosterone, catecholamines, cholesterol, and triglyceride. *Cochrane Database of Systematic Reviews*, CD004022.pub5 (2020).
30. Fu, Y., Clarke, C. V., Van Moer, M. & Schneider, J. Exploring evidence selection with the inclusion network. *Quant. Sci. Stud.* **5**, 219–245 (2024).
31. Trinquart, L., Johns, D. M. & Galea, S. Why do we think we know what we know? A metaknowledge analysis of the salt controversy. *Int. J. Epidemiol.* **45**, 251–260 (2016).
32. Rindflesch, T. C. & Fiszman, M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J. Biomed. Inform.* **36**, 462–477 (2003).
33. Ming, S., Zhang, R. & Kilicoglu, H. Enhancing the coverage of SemRep using a relation classification approach. *J. Biomed. Inform.* **155**, 104658 (2024).
34. Hsiao, T. K., Fu, Y. & Schneider, J. Visualizing evidence-based disagreement over time: the landscape of a public health controversy 2002-2014. *Proc. Assoc. Info. Science and Technology. Association for Information Science and Technology* **57** https://doi.org/10.1002/pra2.315 (2020).
35. Cook, N. R., He, F. J., MacGregor, G. A. & Graudal, N. Sodium and health—concordance and controversy. *Bmj* **369**, m2440 (2020).
36. Gartlehner, G. et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Res. Synth. methods* **15**, 576–589 (2024).
37. Hill, J. E., Harris, C. & Clegg, A. Methods for using Bing's AI-powered search engine for data extraction for a systematic review. *Res. Synth. Methods* **15**, 347–353 (2024).
38. Konet, A. et al. Performance of two large language models for data extraction in evidence synthesis. *Res. Synth. methods* **15**, 818–824 (2024).
39. Tran, V.-T. et al. Sensitivity and Specificity of Using GPT-3.5 Turbo Models for Title and Abstract Screening in Systematic Reviews and Meta-analyses. *Ann. Intern. Med.* **177**, 791–799 (2024).

40. Oami, T., Okada, Y. & Nakada, T.-a. and Performance of a Large Language Model in Screening Citations. *JAMA Netw. Open* **7**, e2420496–e2420496 (2024).
41. Riaz, I. B., Naqvi, S. A. A., Hasan, B. & Murad, M. H. Future of Evidence Synthesis: Automated, Living, and Interactive Systematic Reviews and Meta-Analyses. *Mayo Clinic Proceedings: Digital Health*. **2**, 361–365 (2024).
42. Riaz, I. B. et al. First-line systemic treatment options for metastatic castration-sensitive prostate cancer: a living systematic review and network meta-analysis. *JAMA Oncol*. **9**, 635–645 (2023).
43. Matiasz, N. J. et al. ResearchMaps. org for integrating and planning research. *PloS one* **13**, e0195271 (2018).
44. Matiasz, N. J., Wood, J., Wang, W., Silva, A. J. & Hsu, W. Experiment Selection in Meta-Analytic Piecemeal Causal Discovery. *IEEE Access* **9**, 97929–97941 (2021).
45. Matiasz, N. J., Wood, J. & Silva, A. J. Quantifying convergence and consistency. *Eur. J. Neurosci*. **60**, 6391–6394 (2024).
46. Matiasz, N. J. *Planning Experiments with Causal Graphs*. (University of California, Los Angeles, 2018).
47. National Academies of Sciences, E. & Medicine. *Triangulation in Environmental Epidemiology for EPA Human Health Assessments. Proceedings of a Workshop*. (The National Academies Press, 2022).
48. Bun, R.-S., Scheer, J., Guillo, S., Tubach, F. & Dechartres, A. Meta-analyses frequently pooled different study types together: a meta-epidemiological study. *J. Clin. Epidemiol*. **118**, 18–28 (2020).
49. Yao, M. et al. Including non-randomised studies of interventions in meta-analyses of randomised controlled trials changed the estimates in more than a third of the studies: Evidence from an empirical analysis. *J. Clin. Epidemiol*. **183**, 111815 (2025).
50. Dijk, S. W., Caulley, L. M., Hunink, M. & Labrecque, J. From complexity to clarity: how directed acyclic graphs enhance the study design of systematic reviews and meta-analyses. *Eur. J. Epidemiol*. **39**, 27–33 (2024).
51. McIntyre, K. J., Tassiopoulos, K. N., Jeffrey, C., Stranges, S. & Martin, J. Using causal diagrams within the Grading of Recommendations, Assessment, Development and Evaluation framework to evaluate confounding adjustment in observational studies. *J. Clin. Epidemiol*. **175**, 111532 (2024).
52. Donnelly, K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud. health Technol. Inform*. **121**, 279 (2006).
53. Bodenreider, O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids Res*. **32**, D267–D270 (2004).
54. Milanlouei, S. et al. A systematic comprehensive longitudinal evaluation of dietary factors associated with acute myocardial infarction and fatal coronary heart disease. *Nat. Commun*. **11**, 6074 (2020).
55. Zhu, Q. et al. DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence. *arXiv preprint arXiv:2406.11931* (2024).
56. GLM, T. et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793* (2024).
57. Bai, J. et al. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).
58. Achiam, J. et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
59. Bi, Z. et al. Hypertension prevalence, awareness, treatment, and control and sodium intake in Shandong Province, China: baseline results from Shandong–Ministry of Health Action on Salt Reduction and Hypertension (SMASH). *Preventing Chronic Dis*. **11**, E88 (2014).

Acknowledgements

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; 82330107 to J.D.). We are grateful to the following experts for their invaluable contributions in data extraction 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. We also thank our graduate student Guanghui Deng from Peking University for manual data extraction from CDSR.

Author contributions

Conceptualization: J.D., X.S. Methodology: J.D., X.S., C.Y., T.C. Formal analysis: X.S. Data Curation: X.S., W.Z., T.C. Writing—original draft: X.S., J.D. Writing—review & editing: J.D. Funding acquisition: J.D.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-62783-x.

Correspondence and requests for materials should be addressed to Jian Du.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.