Large Language Models and their role in evidence syntheses and writing

Riaz Qureshi, PhD, MSc
Assistant Professor
Dept. of Ophthalmology, School of Medicine
Dept. of Epidemiology, School of Public Health
University of Colorado Anschutz Medical Campus

Disclosures

1. **Faculty with Cochrane Eyes and Vision**

2. **Consulting for PICO Portal**

Steps in a systematic review

Emerging literature on ways to involve AI

Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation?

Riaz Qureshi, Daniel Shaughnessy, Kayden A. R. Gill, Karen A. Robinson, Tianjing Li & Eitan Agai

Systematic Reviews **12**, Article number: 72 (2023) | Cite this article

Appraising the Potential Uses and Harms of Large Language Models for Medical Systematic Reviews

Hye Sun Yun
Northeastern University
yun.hy@northeastern.edu

Iain J. Marshall
King's College London
iain.marshall@kcl.ac.uk

Thomas A. Trikalinos
Brown University

Byron C. Wallace
Northeastern University

Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions

by Ahmad Alshami [1], Moustafa Elsayed [2], Eslam Ali [3,4,*], Abdelrahman E. E. Eltoukhy [5,*] and Tarek Zayed [3]

Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses

Xufei Luo[1,2,3,4,5]; Fengxian Chen[6], PhD; Di Zhu[7], MPH; Ling Wang[7], MPH; Zijun Wang[1,2,3,4,5]; Hui Liu[1,2,3,4,5]; Meng Lyu[7], MPH; Ye Wang[7], MPH; Qi Wang[8,9], PhD; Yaolong Chen[1,2,3,4,5], MD, PhD

Enhancing Systematic Literature Reviews: Evaluating the Performance of LLM-Based Tools Across Key Systematic Literature Review Stages

Numaya Silva
Department of Industrial Management
University of Kelaniya
Dalugama, Sri Lanka
numaya11pre@gmail.com

Dilani Wickramaarachchi
Department of Industrial Management
University of Kelaniya
Dalugama, Sri Lanka
dilani@kln.ac.lk

Future of Evidence Synthesis: Automated, Living, and Interactive Systematic Reviews and Meta-analyses

Irbaz Bin Riaz, MD, MS, MBI, PhD; Syed Arsalan Ahmed Naqvi, Bashar Hasan, MD; and Mohammad Hassan Murad, MD, MPH

How Good Is ChatGPT for Medication Evidence Synthesis?

Hao LIU[a], Yifan PENG[b] and Chunhua WENG[a,1]

Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review

Judith-Lisa Lieberum[a], Markus Töws[b], Maria-Inti Metzendorf[c], Felix Heilmeyer[d], Waldemar Siemens[b], Christian Haverkamp[d], Daniel Böhringer[a], Joerg J. Meerpohl[b,e,1], Angelika Eisele-Metzger[b,e,*,1]

Opportunities, challenges and risks of using artificial intelligence for evidence synthesis

Waldemar Siemens [1], Erik von Elm,[2,3] Harald Binder,[4] Daniel Böhringer,[5] Angelika Eisele-Metzger,[1,2] Gerald Gartlehner [6,7] Piet Hanegraaf,[8] Maria-Inti Metzendorf [9] Jacob-Jan Mosselman [8] Artur Nowak,[10] Riaz Qureshi,[11] James Thomas,[12] Siw Waffenschmidt,[13] Valérie Labonté,[1,2] Joerg J Meerpohl[1,2]

Strengths of LLMs

(according to ChatGPT)

1. **Rapid Literature Review**: LLMs can quickly process and summarize large volumes of text, making them excellent tools for conducting rapid literature reviews. They can assist in identifying relevant studies, extracting data, and providing initial summaries of research findings.
2. **Understanding and Synthesizing Complex Information**: These models are capable of understanding complex and nuanced academic texts. They can synthesize information from multiple sources into coherent summaries, which is a critical aspect of systematic reviews.
3. **Language Translation and Inclusion**: LLMs can translate texts across languages, potentially broadening the scope of literature reviews to include non-English studies and thus enriching the systematic review process.
4. **Identification of Research Gaps and Trends**: By analyzing vast datasets, LLMs can help identify gaps in research and emerging trends, which can be particularly useful for systematic reviews looking to address unanswered questions or explore new areas.

Weaknesses of LLMs *(according to ChatGPT)*

1. **Quality and Bias Control**: LLMs may inadvertently propagate biases present in their training data or overlook the quality of sources. Ensuring the reliability and neutrality of synthesized reviews remains a challenge.
2. **Lack of Domain-specific Expertise**: While LLMs are knowledgeable across a broad range of subjects, they might lack the depth of understanding required for highly specialized fields. Expert human review is necessary to validate findings and interpretations.
3. **Difficulty in Assessing Methodological Quality**: LLMs may struggle to critically evaluate the methodological quality of studies, an essential component of systematic reviews. Human expertise is crucial for assessing study validity and risk of bias.
4. **Interpretation of Results and Contextual Understanding**: While LLMs can identify and summarize findings, they may not fully grasp the context or significance of these results within a specific research domain, requiring human oversight.

Sources for these thoughts…

1. Published Literature and Guidelines

- Recent peer-reviewed articles exploring AI applications in systematic reviews (e.g., Thomas et al. 2021; Marshall et al. 2023).
- Recommendations from authoritative guidelines like PRISMA 2020, Cochrane Handbook, and AMSTAR, which emphasize transparency, accuracy, and reproducibility in evidence synthesis.

2. Experience from Prior User Interactions

- Extensive conversations and iterative interactions with you and other users working in evidence synthesis, including developing systematic review methodologies, search strategies, data extraction processes, and narrative syntheses.

3. Technical Documentation and Limitations of LLMs

- OpenAI's technical reports, model documentation, and research papers that explicitly outline limitations, such as hallucinations, sensitivity to prompts, and potential misinterpretation of nuanced scientific content.

4. Practical Experience

- Direct practical experiences from prior interactions involving systematic reviews and data extraction tasks where users have highlighted both benefits (speed, drafting assistance) and pitfalls (accuracy issues, misunderstandings, limited critical appraisal capabilities).

In the process of integrating AI…

Potential gains?

- More efficient process
- Speed
- Accuracy
- Training
- Additional time for syntheses

Potential losses?

- Transparency
- Unrealistic expectations of speed
- Unrealistic expectations of accuracy
- Perpetuating biases
- Disconnect from the evidence

Courtesy of Karen Robinson

Outline of areas we will cover:

Focus on writing…

- Protocol development
  - Drafting methods
- Synthesizing the evidence
  - Drafting results
  - Getting started on synthesis
  - Revising text
  - Summarizing papers
- Guidance for use

<u>Many options for LLMs</u>

ChatGPT

Claude

Co-Pilot

Deep-Seek

Gemini

Llama

… and more

Considerations for this presentation:

- What do you need to see to trust the AI results?
- What is the biggest concern?
- How should we evaluate AI?
- What is the appropriate comparison?
- What are the best practices for the selection, use, and reporting of use of AI?
- What should we do differently?

Courtesy of Karen Robinson

PROTOCOL DEVELOPMENT
Helping with drafting (methods)

Drafting methods – searching

“Write me a section of text for a systematic review methods that describes how I will conduct the search and screening. We will include four bibliographic databases in our search (PubMed, EMBASE, CINAHL, and Web of Science) and two trial registries (clinicaltrials.gov, and International Clinical Trials Registry Platform). We will also have hand-searching of relevant journals for conference abstracts and utilize reference checking of included studies. The description of the search and screening should also include the following details: working with a librarian to develop our searches, using COVIDENCE as the platform for screening (including deduplication), using two independent methodologists to screen, and resolving discrepancies through discussion and use of a third reviewer if necessary.”

Search Strategy

To ensure a comprehensive and systematic identification of relevant literature, our search strategy will encompass multiple sources and methods. We will include four key bibliographic databases: PubMed, EMBASE, CINAHL, and Web of Science. Additionally, we will search two trial registries: ClinicalTrials.gov and the International Clinical Trials Registry Platform (ICTRP).

To enhance the sensitivity and specificity of our search strategy, we will collaborate closely with a professional librarian experienced in systematic review methodologies. Together, we will develop tailored search strings and strategies for each database, utilizing appropriate keywords and Medical Subject Headings (MeSH) terms to capture all pertinent studies.

Hand-Searching and Reference Checking

Beyond database searches, we will perform hand-searching of relevant journals to identify conference abstracts and other materials that might not be indexed in the databases. Furthermore, we will conduct reference checking of all included studies to capture any additional relevant articles that our initial search might have missed.

Screening Process

All search results will be imported into COVIDENCE, a systematic review management software, which will facilitate the deduplication process and streamline our screening workflow.

Initial Screening

Two independent methodologists will screen the titles and abstracts of all retrieved records. Each study will be assessed against predefined inclusion and exclusion criteria. To ensure consistency and reliability in the screening process, both methodologists will receive training and clear instructions on the criteria.

Full-Text Screening

For studies that appear to meet the inclusion criteria based on title and abstract, full texts will be obtained and reviewed. Again, two independent methodologists will perform this screening. Any discrepancies between the reviewers will be discussed and resolved through consensus. If consensus cannot be reached, a third reviewer will be consulted to make the final decision.

Run: ChatGPT 4.0, June 2024

Search Strategy and Screening

We will systematically search four bibliographic databases: PubMed, EMBASE, CINAHL, and Web of Science, as well as two clinical trial registries: ClinicalTrials.gov and the International Clinical Trials Registry Platform (ICTRP). A professional medical librarian will assist the research team in developing comprehensive and sensitive search strategies tailored to each database, employing controlled vocabulary (such as MeSH terms) and free-text terms pertinent to the research question. No language or date restrictions will be applied to ensure maximal inclusivity.

Additionally, we will manually hand-search relevant specialty journals and conference proceedings to identify potentially eligible conference abstracts and grey literature not captured by database searches. Furthermore, the reference lists of all included studies will be carefully examined (backward citation tracking) to locate additional relevant articles.

All search results will be imported into Covidence systematic review software (Veritas Health Innovation, Melbourne, Australia) for management, deduplication, and screening. Deduplication will first be performed automatically within Covidence, followed by manual verification by a review team member to ensure accuracy.

Two independent methodologists will then screen titles and abstracts of retrieved records based on predefined eligibility criteria. Studies identified as potentially relevant or whose relevance is uncertain will advance to full-text screening. Full-text articles will similarly be screened independently by two methodologists. Any discrepancies encountered during both screening phases will be resolved through discussion and consensus between the reviewers, and, when necessary, arbitration by a third senior reviewer.

Run: ChatGPT 4.5, April 2025

Drafting methods – search strategies

“Create a PubMed search query for a systematic review using keywords and MeSH terms to find relevant randomized controlled trials. The intervention concept should be related to “chocolate”, “flavonones”, and relevant synonyms. The population should be related to “hypertension”, “high blood pressure”, and include other synonyms.”

2023 – 34 Results

(("Chocolate"[MeSH Terms] OR "Chocolate"[All Fields]) OR ("Flavonones"[MeSH Terms] OR "Flavonones"[All Fields]) OR ("Cocoa"[MeSH Terms] OR "Cocoa"[All Fields])) **AND** *((("Hypertension"[MeSH Terms] OR "Hypertension"[All Fields]) OR ("High Blood Pressure"[MeSH Terms] OR "High Blood Pressure"[All Fields])) OR ("Arterial Hypertension"[MeSH Terms] OR "Arterial Hypertension"[All Fields]) OR ("Essential Hypertension"[MeSH Terms] OR "Essential Hypertension"[All Fields]))* **AND** *("Clinical Trial"[Publication Type] OR "Randomized Controlled Trial"[Publication Type])*.

Run: ChatGPT 3.0, March 2023

2024 – 285 Results

(((chocolate[MeSH Terms] OR chocolate[tiab] OR cocoa[tiab] OR cacao[tiab] OR flavonoids[MeSH Terms] OR flavonoid[tiab] OR flavonols[tiab] OR flavonol[tiab] OR epicatechin[tiab] OR procyanidins[tiab])) **AND** ((hypertension[MeSH Terms] OR hypertens*[tiab] OR high blood pressure[tiab] OR elevated blood pressure[tiab]))) **AND** ((randomized controlled trial[pt] OR controlled clinical trial[pt] OR randomized[tiab] OR placebo[tiab] OR trial[tiab] OR randomly[tiab]))

Run: ChatGPT 4.0, September 2024

2025 – 284 Results

("Chocolate"[MeSH] OR "Flavonoids"[MeSH] OR "Flavanones"[MeSH] OR chocolate*[tiab] OR cocoa[tiab] OR cacao[tiab] OR flavonoid*[tiab] OR flavanone*[tiab] OR polyphenol*[tiab]) **AND** ("Hypertension"[MeSH] OR hypertension[tiab] OR "high blood pressure"[tiab] OR "elevated blood pressure"[tiab]) **AND** ("Randomized Controlled Trial"[Publication Type] OR "randomized"[tiab] OR "randomised"[tiab] OR "random allocation"[MeSH])

Run: ChatGPT 4.5, April 2025

PICOT questions and search strategies formulation: A novel approach using artificial intelligence automation

Lucija Gosak PhD candidate, MSc, RN[1] | Gregor Štiglic Pl
Lisiane Pruinelli PhD, MSc, RN[4] | Dominika Vrbnjak PhD,

Regional anesthesia and acute pain

PDF

Evaluating the utility of large language models in generating search strings for systematic reviews in anesthesiology: a comparative analysis of top-ranked journals

Alessandro De Cassai [1, 2], Burhan Dost [3], Yunus Emre Karapinar [4], Müzeyyen Beldagli [5], Mirac Selcen Ozkal Yalin [4], Esra Turunc [3], Engin Ihsan Turan [6] and Nicolò Sella [2]

An Empirical Study Evaluating Cha Performance in Generating Search Stra Systematic Reviews

Yu, Fei University of North Carolina at Chapel Hill, USA | feifei@unc.edu
Kincaide, Heather University of North Carolina at Chapel Hill, USA | hkincaide@unc.edu
Carlson, Rebecca Beth University of North Carolina at Chapel Hill, USA | rcarlson@unc.edu

n et al

Improving Bibliographic Data Retrieval through Large Language Models

Jingjing Qiao

The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation

Yong Nam Gwon[1*], MD; Jae Heon Kim[1*], MD, PhD; Hyun Soo Chung[2], MD; Eun Jee Jung[2], MD; Joey C

Assessing the Quality of Biomedical Boolean Search Strings Generated by Prompted and Unprompted Models Using ChatGPT: A Pilot Study

Robyn B. Reed & Derek J. Barnett

"Through the looking glass: envisioning new library technologies" academic search using artificial intelligence tools

Jeanine Mary Williamson and Peter Fernandez

PubMed and beyond: biomedical literature search in the age of artificial intelligence

Check for updates

*Qiao Jin, Robert Leaman, and Zhiyong Lu**

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD, USA

oa

Drafting methods – data extraction

"Write me a section of text for a systematic review methods that describes how I will extract data and conduct risk of bias assessments for included studies. The description should specify that both steps will be done with two independent methodologists with discrepancies resolved by a third reviewer, if needed. Like screening, we will use COVIDENCE for data extraction using a standard, pre-designed extraction form that is piloted before use. For Risk of Bias, we will use the RoB 2.0 Excel form because our review is restricted to RCTs."

Data Extraction

Data extraction will be conducted systematically to ensure accuracy and completeness. We will utilize COVIDENCE, a systematic review management software, for the data extraction process. A standard, pre-designed data extraction form will be developed and piloted before use to ensure it captures all necessary information and is user-friendly. Two independent methodologists will extract data from each included study. The data extraction form will capture key study characteristics, including but not limited to:

- Study identification details (author, year, title, journal)
- Study design and setting
- Participant characteristics (e.g., sample size, inclusion/exclusion criteria, demographics)
- Intervention and comparator details
- Outcomes measured and results
- Funding sources and potential conflicts of interest

To ensure consistency, the extraction form will be piloted on a subset of studies before full implementation. Any discrepancies between the two methodologists during data extraction will be resolved through discussion. If consensus cannot be reached, a third reviewer will adjudicate the differences.

Risk of Bias Assessment

Risk of bias in the included studies will be assessed using the RoB 2.0 tool, specifically designed for randomized controlled trials (RCTs). The RoB 2.0 tool will be used in its Excel format to facilitate systematic and structured assessment. Two independent methodologists will assess each included study's risk of bias across the following domains:

- Bias arising from the randomization process
- Bias due to deviations from intended interventions
- Bias due to missing outcome data
- Bias in measurement of the outcome
- Bias in selection of the reported result

Each domain will be rated as "low risk," "some concerns," or "high risk" of bias. Discrepancies between the reviewers will be resolved through discussion. If agreement cannot be reached, a third reviewer will be involved to make the final decision. By ensuring that both data extraction and risk of bias assessments are conducted independently by two methodologists with a third reviewer available for resolving disagreements, we aim to maintain the rigor and reliability of our systematic review.

Run: ChatGPT 4.0, June 2024

Data Extraction and Risk of Bias Assessment

Data extraction and risk of bias assessment will be conducted independently by two methodologists using Covidence (Veritas Health Innovation, Melbourne, Australia), with any discrepancies resolved through discussion and, if needed, adjudication by a third reviewer. We will develop and use a standardized data extraction form within Covidence, which will be piloted on a subset of included studies to ensure clarity, consistency, and comprehensiveness prior to full implementation. Extracted data will include study characteristics (e.g., author, year, setting, sample size), participant characteristics (e.g., age, gender, eligibility criteria), intervention and comparator details, outcomes of interest, duration of follow-up, and any reported adverse events.

For risk of bias assessment, we will restrict our evaluation to randomized controlled trials and apply the Cochrane Risk of Bias 2.0 (RoB 2.0) tool. The assessment will be conducted using the RoB 2.0 Excel tool, which evaluates potential sources of bias across five domains: randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result. Each domain, and the overall study, will be rated as “low risk,” “some concerns,” or “high risk” of bias. As with data extraction, all risk of bias assessments will be performed in duplicate, and disagreements will be resolved through discussion or by a third reviewer when necessary.

Run: ChatGPT 4.5, April 2025

Drafting methods – data analysis

"Write me a section of text for a systematic review methods that describes how I will conduct analyses. We will have qualitative analyses regardless, and if data are available we will have quantitative (i.e., meta-analyses). The section should describe the process for checking the clinical, methodological, and statistical heterogeneity of the included studies. The analysis should specify the effect measures that we will use for meta-analysis (outcomes of interest to us are change in blood pressure from baseline (over 6 or fewer months) and incidence of cardiovascular events (over 12 to 24 months)). The analysis section should also detail the type of model that we will use for each outcome."

Qualitative Analysis

Regardless of the availability of quantitative data, we will conduct a qualitative synthesis of the included studies. This will involve summarizing and interpreting the findings in a narrative format, focusing on the study characteristics, interventions, outcomes, and any notable patterns or themes. We will present the qualitative synthesis in tables and text to provide a comprehensive overview of the evidence base.

Quantitative Analysis

If sufficient data are available, we will conduct meta-analyses to quantitatively synthesize the results of the included studies. The process will involve the following steps:

Heterogeneity Assessment

Before performing meta-analyses, we will assess the clinical, methodological, and statistical heterogeneity of the included studies:

- **Clinical Heterogeneity**: We will examine variations in participant characteristics, interventions, comparators, and outcomes across studies.
- **Methodological Heterogeneity**: We will review differences in study design, risk of bias, and study quality.
- **Statistical Heterogeneity**: We will use the Chi-squared test (Cochran's Q) and the I^2 statistic to assess the presence and extent of statistical heterogeneity. An I^2 value greater than 50% may indicate substantial heterogeneity.

…

…

Effect Measures

For the meta-analyses, we will use the following effect measures for our outcomes of interest:

- **Change in Blood Pressure from Baseline (over 6 or fewer months)**: We will calculate the mean difference (MD) or standardized mean difference (SMD) with 95% confidence intervals (CIs) if different scales are used across studies.
- **Incidence of Cardiovascular Events (over 12 to 24 months)**: We will use the risk ratio (RR) with 95% CIs to summarize the effect sizes.

Meta-Analysis Models

Depending on the presence and extent of heterogeneity, we will choose appropriate models for the meta-analyses:

- **Fixed-Effect Model**: If heterogeneity is low ($I^2 \leq 50\%$), we will use a fixed-effect model, which assumes that the true effect size is the same across all included studies.
- **Random-Effects Model**: If heterogeneity is substantial ($I^2 > 50\%$), we will use a random-effects model, which accounts for variability both within and between studies and assumes that the effect sizes vary across studies.

Sensitivity and Subgroup Analyses

To explore the robustness of our findings, we will conduct sensitivity analyses by excluding studies at high risk of bias. Additionally, if data permit, we will perform subgroup analyses based on key variables such as population characteristics, intervention types, and study quality.

By following these rigorous analytical procedures, we aim to provide a thorough and reliable synthesis of the evidence on the change in blood pressure and the incidence of cardiovascular events associated with the interventions studied.

Data Synthesis and Analysis

We will conduct a qualitative synthesis of all included studies, describing key study characteristics, populations, interventions, comparators, and outcomes. This synthesis will summarize patterns in the findings, explore potential sources of variation across studies, and identify gaps in the evidence base.

If sufficient data are available and studies are sufficiently homogeneous in terms of clinical and methodological characteristics, we will conduct quantitative meta-analyses. Prior to pooling results, we will assess heterogeneity across studies at three levels:

- Clinical heterogeneity, by comparing study populations, intervention/comparator characteristics, and outcome definitions;
- Methodological heterogeneity, by examining study design features and risk of bias assessments;
- Statistical heterogeneity, by calculating the I^2 statistic and visually inspecting forest plots for consistency of effect estimates.

For continuous outcomes, such as change in systolic and/or diastolic blood pressure from baseline (measured over a period of six months or fewer), we will calculate mean differences (MDs) or standardized mean differences (SMDs) with 95% confidence intervals, depending on whether studies report outcomes using the same or different measurement scales.

For dichotomous outcomes, such as the incidence of cardiovascular events (evaluated over 12 to 24 months), we will calculate risk ratios (RRs) or odds ratios (ORs) with 95% confidence intervals, based on how the outcomes are reported across studies.

We will use a random-effects meta-analysis model for all pooled estimates, to account for potential between-study variability and underlying heterogeneity in study populations and implementation. Meta-analyses will be conducted using appropriate statistical software (e.g., RevMan or R), and we will perform sensitivity analyses to explore the robustness of our findings. If meta-analysis is not possible for a given outcome due to insufficient or highly heterogeneous data, we will present the findings narratively and in structured summary tables.

Run: ChatGPT 4.5, April 2025

SYNTHESIZING THE EVIDENCE

LLMs for writing

Analysis of ChatGPT Tool to Assess the Potential of its Utility for Academic Writing in Biomedical Domain

Arun HS Kumar

Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses

Xufei Luo[1,2,3,4,5]; Fengxian Chen[6], PhD; Di Zhu[7], MPH; Ling Wang[7], MPH; Zijun Wang[1,2,3,4,5]; Hui Liu[1,2,3,4,5]; Meng Lyu[7], MPH; Ye Wang[7], MPH; Qi Wang[8,9], PhD; Yaolong Chen[1,2,3,4,5], MD, PhD

Automating Research Synthesis with Domain-Specific Large Language Model Fine-Tuning

TEO SUSNJAK, PETER HWANG, and NAPOLEON REYES, Department of Computer Science and IT, Massey University College of Sciences, Auckland, New Zealand
ANDRE L. C. BARCZAK, Bond University School of Information Technology, Gold Coast, Australia
TIMOTHY MCINTOSH, Cyberoo Pty Ltd., Surry Hills, Australia
SURANGIKA RANATHUNGA, Department of Computer Science and IT, Massey University College of Sciences, Auckland, New Zealand

Empowering Meta-Analysis: Leveraging Large Language Models for Scientific Synthesis

Jawad Ibn Ahad, *Apurba-NSU R&D Lab, ECE, North South University*, Dhaka, Bangladesh, jawad.ibn@northsouth.edu
Rafeed Mohammad Sultan, *Apurba-NSU R&D Lab, ECE, North South University*, Dhaka, Bangladesh, rafeed.sultan@northsouth.edu
Abraham Kaikobad, *Apurba-NSU R&D Lab, ECE, North South University*, Dhaka, Bangladesh, abraham.kaikobad@northsouth.edu
Fuad Rahman, *Apurba Technologies*, Sunnyvale, CA 94085, USA, fuad@apurbatech.com
Mohammad Ruhul Amin, *Computer and Information Science, Fordham University*, New York, USA, mamin17@fordham.edu
Nabeel Mohammed, *Apurba-NSU R&D Lab, ECE, North South University*, Dhaka, Bangladesh, nabeel.mohammed@northsouth.edu
Shafin Rahman, *Apurba-NSU R&D Lab, ECE, North South University*, Dhaka, Bangladesh, shafin.rahman@northsouth.edu

Are LLMs Good Literature Review Writers? Evaluating the Literature Review Writing Ability of Large Language Models

Xuemei Tang[1] Xufeng Duan[1] Zhenguang G. Cai[1,2]
[1]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
[2]Brain and Mind Institute, The Chinese University of Hong Kong

Accelerating Clinical Evidence Synthesis with Large Language Models

Zifeng Wang[1], Lang Cao[1], Benjamin Danek[1], Qiao Jin[2], Zhiyong Lu[2], Jimeng Sun[1,3#]
[1] Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL
[2] National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD
[3] Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Champaign, IL
#Corresponding authors. Emails: jimeng@illinois.edu

Harnessing Large Language Models in Medical Research and Scientific Writing: A Closer Look to The Future

Mohammad Abu-Jeyyab[1,2], Sallam Alrosan[3], Ibraheem M alkhawaldeh[2]

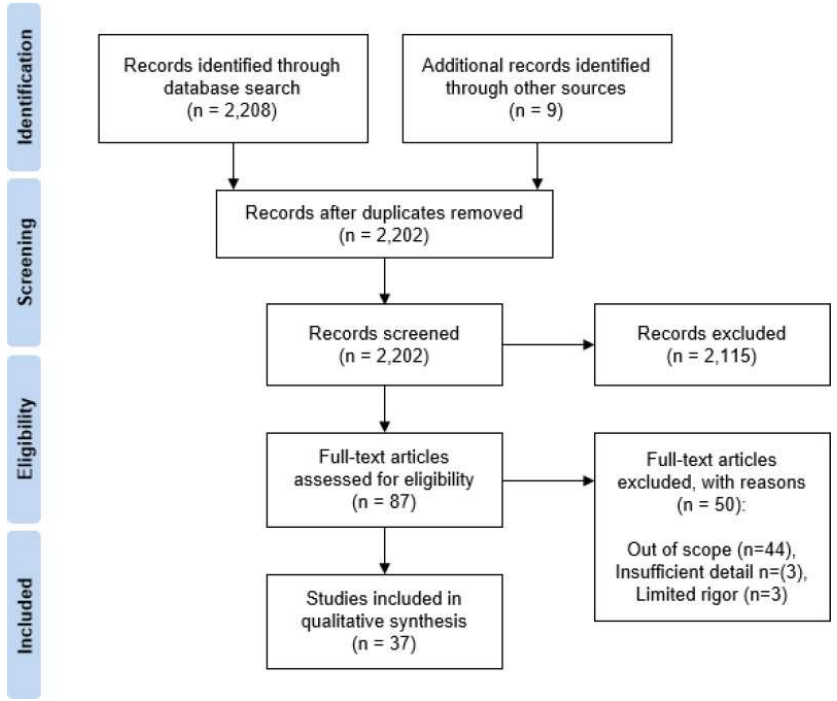
ARTICLE OPEN Check for updates

Evaluating large language models on medical evidence summarization

Liyan Tang[1], Zhaoyi Sun[2], Betina Idnay[3], Jordan G. Nestor[4], Ali Soroush[4], Pierre A. Elias[3], Ziyang Xu[5], Ying Ding[1], Greg Durrett[6], Justin F. Rousseau[7,8✉], Chunhua Weng[3✉] and Yifan Peng[2✉]

Getting started with writing – results

"Using the following completed PRISMA flow diagram, write a description of the results of the searching process that includes the total number of records retrieved, how many were removed as duplicates before screening, how many were screened and excluded at both title/abstract and full-text levels, the numbers excluded for specific reasons at full-text level, and the final number of studies included in the review."



Results of the Searching Process

The search strategy for this systematic review yielded a total of 2,208 records from the database search and an additional 9 records from other sources, resulting in a combined total of 2,217 records. After removing 15 duplicates, 2,202 records remained for screening.

Screening Process:

- **Title and Abstract Screening:** Out of the 2,202 records screened, 2,115 were excluded based on the title and abstract.
- **Full-Text Screening:** The remaining 87 full-text articles were assessed for eligibility. Of these, 50 articles were excluded for the following reasons:
  - Out of scope: 44 articles
  - Insufficient detail: 3 articles
  - Limited rigor: 3 articles
- **Final Inclusion:** A total of 37 studies were included in the qualitative synthesis of the review.

This systematic and thorough approach ensured that only the most relevant and high-quality studies were included in the final synthesis.

Run: ChatGPT 4.0, June 2024

Summarizing evidence

“Using the RESULTS from the five RCTs that I provided, summarize the evidence for the effects of dark chocolate supplementation on blood pressure at short term (3 or fewer months) and long term (12 or more months). Group the trials according to their time point and include any effect estimates for the change in blood pressure or the difference in blood pressure at the respective time points for each trial (where reported) and put the appropriate in-text citations (e.g., Curtis 2013) where relevant.”

Summarizing evidence (continued)

Short-term Effects (3 or Fewer Months)

1. **Dicks 2018:** In this 12-week study with hypertensive patients with type 2 diabetes, no significant change in blood pressure was observed with the intake of 2.5 g/day of flavanol-rich cocoa compared to the placebo. The study did not report specific numerical changes in BP, indicating no improvement in cardiometabolic parameters from cocoa consumption in the short term.

2. **Mozaffari-Khosravi 2016:** Over 8 weeks, consuming 40 g/day and 20 g/day of dark chocolate in patients with metabolic syndrome showed no significant effect on blood pressure. Specifically, the mean change of systolic blood pressure in the 40g, 20g, and control groups were 0.31 ± 1.81, 0.37 ± 1.65, and 0.26 ± 1.56 mmHg, respectively (P = 0.3). These figures for diastolic blood pressure were 0.08 ± 1.03, -0.02 ± 1.12, and 0.22 ± 1.03 mmHg, showing no significant changes.

3. **Muniyappa 2008:** This study evaluated the effects of a flavanol-rich cocoa drink (900 mg flavanols/day) over 2 weeks in individuals with essential hypertension. The study concluded that daily consumption did not significantly reduce blood pressure or improve insulin resistance. Specific effect estimates for BP changes were not provided in the summary, indicating a lack of significant BP reduction.

4. **Njike 2016:** Across all study participants, daily consumption of 5 or 10 g of cocoa-containing products for 8 weeks, compared to placebo, did not significantly lower 24-hour or office blood pressure. The study did not report specific numerical changes in BP for the overall group, suggesting no dose-response effect on blood pressure from cocoa consumption.

…

Summarizing evidence (continued)

Long-term Effects (12 or More Months)

1. **Curtis 2013:** After a 1-year intervention of consuming 27 g of flavonoid-enriched chocolate daily, no significant change in blood pressure was observed. Specifically, the flavonoid group saw no significant effect on CCA-IMT or blood pressure, with pulse pressure variability showing improvement (flavonoid: 20.11 ± 0.07 mm Hg/min; placebo: 0.10 ± 0.11 mm Hg/min; P = 0.04). For a subgroup with PWV data, net improvements were observed (flavonoid: -0.07 ± 0.38 m/s; placebo: 0.68 ± 0.25 m/s; P = 0.01), suggesting potential long-term benefits on arterial stiffness but not directly on BP.

Summary

- In short-term studies (up to 3 months), there is no consistent evidence that dark chocolate supplementation significantly lowers blood pressure in patients with type 2 diabetes, metabolic syndrome, or essential hypertension. The trials either did not report specific numerical changes in BP or found no significant effects. In the long-term (12 months), evidence from Curtis 2013 indicates potential benefits on arterial stiffness with no significant direct effects on blood pressure. These findings suggest that while dark chocolate may have some vascular benefits, its impact on blood pressure, especially in the short term, appears limited and may not be clinically significant across different populations.

Run: ChatGPT 4.0, June 2024

Revising to simplify language

"The following text is the results section from a Cochrane systematic review. It has a readability score of approximately 33, which corresponds with a grade level of approximately 12th grade (i.e., college reading level). Using the same or fewer words, I want you to revise the text so that it contains the same information but at a simpler reading level so that someone with a Grade 8-9 reading level would able to understand."

[Abstract results from Li 2019, CDSRs; 8(8):CD006460]


Assessing the Capability of Large Language Model Chatbots in Generating Plain Language Summaries

Himel Mondal [1], Gaurav Gupta [2], Pradosh Kumar Sarangi [3], Shreya Sharma [4], Pritam K. Choudhary [4], Ayesha Juhi [1], Anita Kumari [1], Shaikat Mondal [5]


The use of a large language model to create plain language summaries of evidence reviews in healthcare: A feasibility study

Colleen Ovelman[1,2] | Shannon Kugley[1] | Gerald Gartlehner[1,3] | Meera Viswanathan[1]

Original (335 words; 12th grade reading level; FK readability score = 33)

We included seven trials (five randomized controlled trials and two quasi-randomized controlled trials) conducted in six countries (China, India, Iran, Ireland, Spain, and the United States) with a total of 1177 amblyopic eyes. Three of these seven trials were from the original 2009 version of the review. We assessed two trials as having a low risk of bias across all domains, and the remaining five trials as having unclear or high risk of bias for some domains. As different occlusion modalities, atropine penalization regimens, and populations were used across the included trials, we did not conduct any meta-analysis due to clinical and statistical heterogeneity. Evidence from six trials (two at low risk of bias) suggests that atropine penalization is as effective as conventional occlusion in improving visual acuity. Similar improvement in visual acuity was reported at all time points at which it was assessed, ranging from five weeks (improvement of 1 line) to 10 years (improvement of greater than 3 lines). At six months, although most participants (363/522) come from a trial rated as at low risk of bias with a precise estimate (mean difference (MD) 0.03, 95% confidence interval (CI) 0.00 to 0.06), two other trials rated as at high risk of bias produced inconsistent estimates and wide confidence intervals (MD −0.02, 95% CI −0.11 to 0.07 and MD −0.14, 95% CI −0.23 to −0.05; moderate-certainty evidence). At 24 months, additional improvement was found in both groups, but there continued to be no meaningful difference between those receiving occlusion and those receiving atropine therapies (moderate-certainty evidence). We did not find any difference in ocular alignment, stereo acuity, or sound eye visual acuity between occlusion and atropine penalization groups (moderate-certainty evidence). Both treatments were well tolerated. Atropine was associated with better adherence (moderate-certainty evidence) and quality of life (moderate-certainty evidence), but also a higher reported risk of adverse events in terms of mild reduction in the visual acuity of the sound eye not requiring treatment and light sensitivity (high-certainty evidence). Skin, lid, or conjunctival irritation were more common among participants receiving patching than those receiving atropine (high-certainty evidence). Atropine penalization costs less than conventional occlusion.

Revised (251 words; 8th grade reading level; FK readability score = 67)

We looked at seven studies from six different countries that tested treatments on 1177 eyes with amblyopia, also known as lazy eye. Three of these studies were part of an earlier review done in 2009. We trusted two of the studies completely, but had some doubts about the others.

Because the studies were all a bit different, we didn't mix their results together. What we found from six of these studies is that using atropine drops works as well as the usual method of covering the eye to help improve vision. This was true for both short times, like five weeks, and long times, up to 10 years.

Six months in, the best results came from a study we could rely on, showing a small improvement. Two other studies weren't as reliable and their results weren't as clear. After two years, both ways of treating amblyopia showed more improvement, but neither was better than the other.

We also saw no difference in how well eyes lined up, how well they could see in 3D, or how well the good eye could see between the two treatments. Both treatments were safe and people didn't mind using them. People using atropine were more likely to stick with their treatment and were happier, but they had a slightly higher chance of their good eye not seeing as well and being more sensitive to light. People using eye patches had more irritation around their eye. Using atropine drops also cost less than using eye patches.

Run: ChatGPT 4.0, June 2024

Guidance for LLM use in writing

Ethical Use of Artificial Intelligence for Scientific Writing: Current Trends

Journal of Human Lactation
2024, Vol. 40(2) 211–215
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08903344241235160
journals.sagepub.com/home/jhl
Sage

Ellen Chetwynd, PhD MPH BSN IBCLC[1]

Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers

Tanisha Mishra[1], Edward Sutanto[2,3], Rini Rossanti[4], Nayana Pant[5], Anum Ashraf[6], Akshay Raut[7], Germaine Uwabareze[8], Ajayi Oluwatomiwa[9] & Bushra Zeeshan[10,11]

Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows

Ryan Watkins[1]

EDUCATION

Ten simple rules for using large language models in science, version 1.0

Gabriel Reuben Smith[1]*, Carolina Bello[1], Lalasia Bialic-Murphy[1], Emily Clark[1], Camille S. Delavaux[1], Camille Fournier de Lauriere[1], Johan van den Hoogen[1], Thomas Lauber[1], Haozhi Ma[1], Daniel S. Maynard[1,2], Matthew Mirman[3], Lidong Mo[1], Dominic Rebindaine[1], Josephine Elena Reek[1], Leland K. Werden[1], Zhaofei Wu[1,4], Gayoung Yang[1,5], Qingzhou Zhao[1,6], Constantin M. Zohner[1], Thomas W. Crowther[1]

Reporting guideline for the use of CHatbots and other gEnerative artificial intelligence tools in mEdical Research: the CHEER Statement

Xufei Luo,[1], Yih Chung Tham,[2*], Mauro Giuffrè,[3], Robert Ranisch,[4], Mohammad Daher,[5], Kyle Lam,[6], Alexander Viktor Eriksen,[7], Susan L Norris,[8], Che-Wei Hsu,[9,58], Akihiko Ozaki,[10], Jean-Christophe Bélisle-Pipon,[11], Qingyu Chen,[12], Fabio Ynoe de Moraes,[13], Brian D. Earp,[14], Sahil Khanna,[15], Lorenzo Righetto,[16], Renne Rodrigues,[17], Yousif Subhi,[18], Kuan-Pin Su,[19,20], Emir Begagić,[21], Egor Chumakov,[22], Sophie Curbo,[23], Aybars Kivrak,[24], Ery Ayelen Ko,[25], Myeong Soo Lee,[26], Dengxiong Li,[27], Andrey Litvin,[28], Peng Liu,[29], Sebastian Porsdam Mann,[30,31], José Darío Martínez-Ezquerro,[32,33], Surapaneni Krishna Mohan,[34], Philip Moons,[35,36,37], Alejandro Quiroga-Garza,[38], Riaz Qureshi,[39], Ximing Xu,[40], Stephen R Ali,[41], Nash Anderson,[42], Hiroj Bagde,[43], Charlotte Blease,[44], Randy D'Amico,[45], Hannah Decker,[46], Adrian Egli,[47], Shijian Feng,[48], Sheng Li,[49], Nav Persaud,[50], Murali Ramanathan,[51], Gemma Sharp,[52], Ye Wang,[53], Wah Yang,[54], Qing-xin Yu,[55], Zhaoxiang Bian,[56], Yaolong Chen,[1*], Janne Estill,[1,57*]

ICASR

International Collaboration for the Automation of Systematic Reviews

Home | About | Resources | Events | Papers | 2024 ICASR meeting

Ten simple rules

Areas of concern…

1. Follow the rules of the target journal
2. Outline relevant risks before LLM use
3. Avoid plagiarism
4. Respect confidentiality
5. Verify the truthfulness of content generated by the LLM

Use an LLM to…

6. … perform a more inclusive search
7. … summarize content
8. … refine written English in formal and informal communication
9. … improve scientific coding
10. … jump-start your scientific writing

RULE 0. TRANSPARENCY – Models, timing, prompts, uses

What comes next?

- Useful as tools, but not perfect
- Need for pre-specification and transparency
- Guidelines and recommendation for use

ChatGPT4.0; DALL-E Image Generator

Thank you!

riaz.qureshi@cuanschutz.edu

ASSESSING AND EXTRACTING THE EVIDENCE

LLMs for data extraction

Data extraction for evidence synthesis using a large language model: A proof-of-concept study

Gerald Gartlehner[1,2] | Leila Kahwati[1] | Rainer Hilscher[1] | Ian Thomas[1] | Shannon Kugley[1] | Karen Crotty[1] | Meera Viswanathan[1] | Barbara Nussbaumer-Streit[2] | Graham Booth[1] | Nathaniel Erskine[3] | Amanda Konet[1] | Robert Chew[1]

RESEARCH ARTICLE

ChatGPT-4o can serve as the second rater for data extraction in systematic reviews

Mette Motzfeldt Jensen[1,2]*, Mathias Brix Danielsen[1,2], Johannes Riis[1,2], Karoline Assifuah Kristjansen[3,4], Stig Andersen[1,2], Yoshiro Okubo[5], Martin Grønbech Jørgensen[1,2]

Performance of two large language models for data extraction in evidence synthesis

Amanda Konet[1] | Ian Thomas[1] | Gerald Gartlehner[1,2] | Leila Kahwati[1] | Rainer Hilscher[1] | Shannon Kugley[1] | Karen Crotty[1] | Meera Viswanathan[1] | Robert Chew[1]

Collaborative large language models for automated data extraction in living systematic reviews

Muhammad Ali Khan, MBBS[1], Umair Ayub, PhD[1], Syed Arsalan Ahmed Naqvi, MD[1], Kaneez Zahra Rubab Khakwani, MD[2], Zaryab bin Riaz Sipra, MD[3], Ammad Raina, DO[4], Sihan Zhou, BS[1], Huan He, PhD[5], Amir Saeidi, MS[1,6], Bashar Hasan, MD[7], Robert Bryan Rumble, MSc[8], Danielle S. Bitterman, MD[9], Jeremy L. Warner, MD, MS[10,11,12,13], Jia Zou, PhD[6], Amye J. Tevaarwerk, MD[14], Konstantinos Leventakos, MD, PhD[7], Kenneth L. Kehl, MD, MPH[15], Jeanne M. Palmer, MD[1], Mohammad Hassan Murad, MD[7], Chitta Baral, PhD[6], Irbaz bin Riaz, MD, PhD*[,1,16]

AI-driven evidence synthesis: data extraction of randomized controlled trials with large language models

Jiayi Liu, MM[a,b], Honghao Lai, MD[a,b], Weilong Zhao, MM[a,b], Jiajie Huang, MSN[c], Danni Xia, MM[a,b], Hui Liu, MD[d], Xufei Luo, MD[d,e,f], Bingyi Wang, MM[d], Bei Pan, MD[d], Liangying Hou, MD[d,g], Yaolong Chen, MD[d,f,h], Long Ge, MD[a,b,e,*], on behalf of ADVANCED Working Group

How good are large language models for automated data extraction from randomized trials?

Zhuanlan Sun[1], Ruilin Zhang[2], Suhail A. Doi[3], Luis Furuya-Kanamori[4], Tianqi Yu[5], Lifeng Lin[6], Chang Xu[7,8]

Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages

Qusai Khraisha[1,2] | Sophie Put[3] | Johanna Kappenberg[2] | Azza Warraitch[1,2] | Kristin Hadfield[1,2]

Exploring the Use of a Large Language Model for Data Extraction in Systematic Reviews: a Rapid Feasibility Study

Lena Schmidt[1,*], Kaitlyn Hair[2], Sergio Graziosi[3], Fiona Campbell[1], Claudia Kapp[4], Alireza Khanteymoori[5], Dawn Craig[1], Mark Engelbert[6,7] and James Thomas[3]

From promise to practice: challenges and pitfalls in the evaluation of large language models for data extraction in evidence synthesis

Gerald Gartlehner,[1,2] Leila Kahwati,[1] Barbara Nussbaumer-Streit,[2] Karen Crotty,[1] Rainer Hilscher,[1] Shannon Kugley,[1] Meera Viswanathan,[1] Ian Thomas,[1] Amanda Konet,[1] Graham Booth,[1] Robert Chew[1]

LLMs for Risk of Bias

Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment

Bashar Hasan [1,2] Samer Saadi [1,2] Noora S Rajjoub,[1] Moustafa Hegazi,[1,2] Mohammad Al-Kordi,[1,2] Farah Fleti,[1,2] Magdoleen Farah,[1,2] Irbaz B Riaz,[3] Imon Banerjee,[4,5] Zhen Wang [1,6] Mohammad Hassan Murad [1,2]

JAMA Network | Open

Original Investigation | Statistics and Research Methods

Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models

Honghao Lai, MM; Long Ge, MD; Mingyao Sun, MSN; Bei Pan, MD; Jiajie Huang, MSN; Liangying Hou, MD; Qiuyu Yang, MD; Jiayi Liu, MM; Jianing Liu, MSN; Ziying Ye, MM; Danni Xia, MM; Weilong Zhao, MM; Xiaoman Wang, MD; Ming Liu, MD; Jhalok Ronjan Talukdar, PhD; Jinhui Tian, MD; Kehu Yang, MD; Janne Estill, PhD

Exploring the potential of Claude 2 for risk of bias assessment: Using a large language model to assess randomized controlled trials with RoB 2

AUTHORS

Angelika Eisele-Metzger*[1,2], Judith-Lisa Lieberum*[3], Markus Toews[1], Waldemar Siemens[1], Felix Heilmeyer[4], Christian Haverkamp[4], Daniel Boehringer[3], Joerg J Meerpohl[1,2]

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials

Simon Šuster[1] | Timothy Baldwin[2] | Karin Verspoor[1,3]

Streamlining Systematic Reviews: Harnessing Large Language Models for Quality Assessment and Risk-of-Bias Evaluation

MSR80
AI-ENABLED RISK OF BIAS ASSESSMENT OF RCTS IN SYSTEMATIC REVIEWS: A CASE STUDY

Langham J,[1] Reason T,[1] Malcolm B,[2] Klijn S,[3] Gimblett A[1]
[1]Estima Scientific Ltd, London, UK, [2]Bristol Myers Squibb, Middlesex, UK, [3]Bristol Myers Squibb, Lawrence Township, NJ, USA

Pilot study on large language models for risk-of-bias assessments in systematic reviews: A(I) new type of bias?

Joseph Barsby [1,2] Samuel Hume [1,3] Hamish AL Lemmey [1,4] Joseph Cutteridge [5,6] Regent Lee [7] Katarzyna D Bera [3,7]

RoBGuard: Enhancing LLMs to Assess Risk of Bias in Clinical Trial Documents

Changkai Ji[1], Bowen Zhao[1], Zhuoyao Wang[1], Yingwen Wang[2], Yuejie Zhang[1], Ying Cheng[1], Rui Feng[1,2*], Xiaobo Zhang[2*],

Summarizing study characteristics

“I want a table that presents the characteristics of each study. I want you to create a table with each study in its own row. The top row should be a header row that specifies the names of the eight columns. For the columns, in order of left to right, I want the following: "Study ID", "Intervention", "Comparator", "Follow-up", "Eligibility criteria", "Age", "Sex", "Conclusions". For any item in the table that is copied and pasted, put the extracted information in quotes. If any characteristic is not reported, say "Not reported". Use the following instructions for each item:

- For the "Study ID", use the last name of the first listed author and the year of publication.
- For the "Intervention", copy and paste the description of the intervention that participants receive. Also describe the dose and timing of the intervention.
- For the "Comparator", copy and paste the description of the comparison that participants receive. Also describe the dose and timing of the comparator.
- For the "Follow-up", extract the longest point of follow-up at which participants were assessed.
- For the "Eligibility criteria", copy and paste the description of the inclusion and exclusion criteria for participating in the trial.
- For the "Age", extract the mean and standard deviation for the age of participants in each trial arm at baseline.
- For the "Sex", extract the number of men and women in each trial arm at baseline.
- For the "Conclusions", copy and paste the main conclusion from the abstract.”

Run: ScholarGPT (ChatGPT 4.0), June 2024

Study ID	Intervention	Comparator	Follow-up	Eligibility criteria	Age	Sex	Conclusions
Curtis 2013	"27 g flavonoid-enriched chocolate/d [850 mg flavan-3-ols (90 mg epicatechin) + 100 mg isoflavones (aglycone equivalents)/d]"	"Matched placebo"	"1 year"	"Postmenopausal women with T2DM, aged ≤75 y, receiving established therapy for T2DM, excluding those with recent smoking history, significant vascular disease or cancer, hormone replacement therapy use, poor diabetes control, and elevated systolic BP."	"51 to 74 years"	"Not reported "	"Flavonoids may augment existing therapeutic strategies to reduce cardiovascular disease risk in postmenopausal T2DM patients, with clinically relevant improvements in arterial stiffness observed; equol producers were particularly responsive."
Dicks 2018	"Capsules with 2.5 g/day of a flavanol-rich cocoa"	"Cocoa-free capsules"	"12 weeks"	"Hypertensive patients with T2D, stable pharmacological treatment, with good adjustment for glucose metabolism, lipids, and BP, excluding those treated with insulin, any changes in chronic medication in the previous three months, history of cardiovascular events, malabsorption disorders, smoking, pregnancy or lactation, present/former alcohol or drug abuse, and excessive flavanol-rich foods consumption."	"64.2 ± 1.5 years"	"18 men, 17 women"	"Regular intake of a usual serving size of flavanol-rich cocoa does not improve cardiometabolic parameters in stably treated patients with T2D and hypertension."
Mozaffari 2016	"40 g/d DC (40G), 20 g/d DC (20G)"	"No DC as the control group (CG)"	"8 weeks"	"Patients with MetS, according to the NCEP, ATP III, aged 30-60 y, excluding those with cardiovascular, hepatic, and renal diseases, allergic reactions to cocoa components, pregnancy or lactation, antioxidant supplements consumption."	"51.38 ± 6.95 years"	"45 men and 49 women"	"Daily intake of 40 g of DC with 76% purity for an 8-week period had no effect on body weight, BMI, BP, and oxidative stress in patients with MetS."
Muniyappa 2008	"Flavanol-rich cocoa drink (150 mL twice a day, ≈900 mg flavanols/d)"	"Flavanol-poor placebo (≈28 mg flavanols/d)"	"2 weeks"	"Adults with mild-to-moderate hypertension, not taking any medication or nutritional supplements except for antihypertensive agents, excluding those with diabetes, liver disease, pulmonary disease, renal insufficiency, coronary heart disease, heart failure, peripheral vascular disease, coagulopathy, allergies to cocoa, or severe systemic diseases."	"21 to 65 years"	"Not reported "	"Daily consumption of flavanol-rich cocoa for 2 weeks is not sufficient to reduce blood pressure or improve insulin resistance in human subjects with essential hypertension."
Njike 2016	"5 vs. 10 g of cocoa powder in cocoa-containing products for 8 weeks"	"Nutrient-matched products not containing cocoa"	"8 weeks"	"Adults with stage 1 hypertension, on no more than one blood pressure medication, BMI ≤40 kg/m2, excluding those using lipid-lowering medication or aspirin unless stable on medication, severe hypertension, allergy to cocoa products, regular use of vitamin C, vitamin E, fish oil, flax seed oil, omega-3 fatty acids, Coenzyme Q10, fiber supplements, garlic pills, arginine, red yeast rice, and/or any antioxidant."	"Average age 53.6 years"	"63 women and 59 men"	"Including cocoa in the diet of patients with stage 1 hypertension seems to exert differential beneficial effects on cardiometabolic risk factors in certain sub-groups of patients."

Study ID	Intervention	Comparator	Follow-up	Eligibility criteria	Age	Sex	Conclusions
Curtis 2013	"Participants consumed flavonoid-enriched chocolate daily for 1 year given as a split dose (2 × 13.5 g/day). The daily dose provided 90 mg epicatechin (850 mg total flavan-3-ols) and 100 mg isoflavones (aglycone equivalents from a daidzein-rich extract)."	"Participants consumed placebo chocolate daily, matched for appearance, taste, and macronutrient content."	1 year	"Postmenopausal women with type 2 diabetes mellitus (T2DM), aged ≤75 years, receiving established (≥12 months) therapy for T2DM (statins ≥40 mg simvastatin or ≥10 mg atorvastatin). Exclusion criteria included current or recent history of smoking, significant history of vascular disease or cancer, hormone replacement therapy use, poor diabetes control (glycated hemoglobin ≥10%), and elevated systolic BP (≥160 mm Hg)."	Mean age: 62.1 ± 0.7 (flavonoid), 63.0 ± 0.8 (placebo)	Not reported	"Although the 1-year intervention did not change CCA-IMT or BP, clinically relevant improvements in arterial stiffness were observed; equol producers were particularly responsive. Flavonoids may augment existing therapeutic strategies to reduce cardiovascular disease risk in postmenopausal T2DM patients, and longer studies are needed to examine the effects on atherosclerosis progression."
Dicks 2018	"Participants received five capsules daily, each providing 0.5 g ACTICOA™ cocoa powder (2.5 g/day total), recommended to be taken in the morning (three capsules) and evening (two capsules), for 12 weeks."	"Participants received placebo capsules containing pure microcrystalline cellulose (identical regimen as intervention group)."	12 weeks	"Patients with type 2 diabetes and hypertension, diabetes duration ≥1 year, good glycemic control (HbA1c 48–58 mmol/mol). Excluded: treatment with insulin, changes in medication last 3 months, cardiovascular history, malabsorption disorders, smoking, pregnancy, lactation, alcohol/drug abuse, vitamin/antioxidant supplements, excessive flavanol-rich food intake."	Mean age: 65.6 ± 2.6 (cocoa), 62.8 ± 1.6 (placebo)	Cocoa: 7 men, 10 women; Placebo: 11 men, 7 women	"Regular intake of a usual serving size of flavanol-rich cocoa does not improve cardiometabolic parameters in stably treated patients with T2D and hypertension. As medication modulates partly the same targets as cocoa flavanols, future studies should focus on preventive effects of cocoa against diabetes and cardiometabolic diseases in individuals with preexisting abnormalities not requiring pharmacological treatment."

Summarizing study characteristics

“I want you to extract some additional data from study full-text pdfs related to the outcomes and analyses. Give me the results in a table with each item as its own column and the study as the row. Use the following instructions for each item:

- For the "Study ID", use the last name of the first listed author and the year of publication.
- For the “Numbers lost to follow-up", use information from the start of the results or the study participant flow diagram to give the numbers of participants lost and specific reasons, if reported.
- For the “Primary outcome", copy and paste the description of the primary outcome that was assessed for participants. If no outcome is designated as primary, use the first listed outcome.
- For “Blood Pressure: Baseline (Systolic)”, extract the mean and standard deviation for the systolic blood pressure at baseline for both groups and specify the group.
- For “Blood Pressure: Baseline (Diastolic)”, extract the mean and standard deviation for the diastolic blood pressure at baseline for both groups and specify the group.
- For “Blood Pressure: Endpoint (Systolic)”, extract the mean and standard deviation for the systolic blood pressure at the longest follow-up time point for both groups and specify the group.
- For “Blood Pressure: Endpoint (Diastolic)”, extract the mean and standard deviation for the diastolic blood pressure at the longest follow-up time point for both groups and specify the group.
- For the “Time point”, specify the time point of the outcome data that was extracted for the endpoints.”

Study ID	Numbers lost to follow-up	Primary outcome	Blood Pressure: Baseline (Systolic)	Blood Pressure: Baseline (Diastolic)	Blood Pressure: Endpoint (Systolic)	Blood Pressure: Endpoint (Diastolic)	Time point
Dicks 2018	"All 42 participants finished the study. However, one subject of each group was excluded due to changes in chronic medication (levothyroxine, cortisone) which might have affected our outcome markers, changes in BW of ≥5% and due to a compliance <80%. Moreover, one participant of the cocoa group was excluded due to not being in fasted state at the second visit."	"Blood pressure (the primary outcome measure)"	Cocoa group: 139.1 ± 3.2 mmHg; Placebo group: 141.6 ± 4.2 mmHg	Cocoa group: 78.1 ± 2.9 mmHg; Placebo group: 79.1 ± 1.8 mmHg	Cocoa group: 138.5 ± 3.7 mmHg; Placebo group: 140.4 ± 4.1 mmHg	Cocoa group: 78.2 ± 2.4 mmHg; Placebo group: 78.2 ± 2.6 mmHg	12 weeks

Table 4. Data on blood pressure and on laboratory investigation.

	Cocoa group (n = 17)			Placebo group (n = 18)			P Baseline
	Baseline	Week 12	p	Baseline	Week 12	p	
Blood pressure							
Systolic (mmHg) [$]	139.1 ± 3.2	138.5 ± 3.7	ns [c]	141.6 ± 4.2	140.4 ± 4.1	ns [c]	ns [a]
Diastolic (mmHg) [$]	78.1 ± 2.9	78.2 ± 2.4	ns [c]	79.1 ± 1.8	78.2 ± 2.6	ns [c]	ns [a]
Glucose metabolism							
Fasting blood glucose (mmol/l)	7.6 ± 0.3	7.5 ± 0.2	ns [c]	7.6 ± 0.3	7.8 ± 0.2	ns [c]	ns [a]
HbA_{1c} (mmol/mol)	46.5 (43.2; 49.7)	46.5(41.0; 50.8)	ns [d]	47.5(44.3; 55.2)	48.6(43.2; 53.0)	ns [d]	ns [b]
Insulin (pmol/l)	99.6 ± 11.0	83.1 ± 9.0	ns [c]	89.6 ± 10.1	91.8 ± 7.7	ns [c]	ns [a]
HOMA-IR	4.7 ± 0.5	3.8 ± 0.4	ns [c]	4.4 ± 0.6	4.5 ± 0.4	ns [c]	ns [a]
Lipid status							
Total cholesterol (mmol/l)	5.0 ± 0.2	4.9 ± 0.2	ns [c]	4.7 ± 0.2	4.6 ± 0.2	ns [c]	ns [a]
LDL-cholesterol (mmol/l)	3.0 ± 0.2	2.9 ± 0.2	ns [c]	2.8 ± 0.2	2.9 ± 0.2	ns [c]	ns [a]
HDL-cholesterol (mmol/l) [#]	1.3(1.2; 1.5)	1.4(1.2; 1.8)	ns [c]	1.3(1.1; 1.4)	1.2(1.2; 1.4)	ns [c]	ns [a]
LDL/HDL cholesterol ratio	2.3 ± 0.2	2.1 ± 0.2	ns [c]	2.3 ± 0.2	2.3 ± 0.2	ns [c]	ns [a]
Triglycerides (mmol/l) [#]	1.3(0.9; 1.9)	1.4(0.9; 1.8)	ns [c]	1.8(1.3; 2.3)	1.5(1.1; 2.0)	ns [c]	ns [a]
Creatinine (µmol/l)	61.0 ± 3.8	61.0 ± 3.8	ns [c]	61.0 ± 3.1	61.0 ± 3.1	ns [c]	ns [a]