Automation of Systematic Reviews with Large Language Models

Christian Cao, MD[1, *], Rohit Arora, PhD[2, *], Paul Cento, BSc[3, *], Adil Budak, BSc[4], Katherine Manta, MSc, MD[1], Elina Farahani, MD[1], Matthew Cecere, MSc, MD[1], Anabel Selemon, MSc, BHSc[5], Jason Sang BSc[6], Ling Xi Gong[2], Robert Kloosterman, MSc, MD[1], Scott Jiang, BASc[7], Richard Saleh, MA, MD[1], Denis Margalik, MSc, MD[1], James Lin[8], Jane Jomy, MSc, MD[1], Jerry Xie, BSc[6], David Chen, MD[1], Jaswanth Gorla, MSc, MD[1], Sylvia Lee, MSc[9], Kelvin Zhang, BSc[6], Jennifer Kuang, BSc[6], Harriet Ware, MSc[10], Mairead Whelan, MSc[10], Bijan Teja, MD, PhD[11,12], Alexander A. Leung, MD, MPH[13], Rahul K. Arora, DPhil[10], Jennifer Pillay[14], Lisa Hartling[14], Michael Noetel, PhD[15], David B. Emerson, PhD[16], Allan S. Detsky, MD, PhD[1,17,21], Andrea C. Tricco PhD[18,19], George M. Church, PhD[2, †], David Moher, PhD[20,21,22, †], Niklas Bobrovitz, MD, DPhil[23, †]

1. Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
2. Department of Genetics, Harvard Medical School, Boston, MA, USA
3. Independent researcher
4. Department of Computer Science, Wilfrid Laurier University, Waterloo, ON, Canada
5. Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada
6. Department of Computer Science, University of Waterloo, Waterloo, ON, Canada
7. Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada
8. McGovern Institute for Brain Research, MIT, Cambridge, MA, USA
9. Icahn School of Medicine, Mount Sinai Hospital, New York, NY, United States of America
10. Centre for Health Informatics, Department of Community Health Sciences, University of Calgary, Calgary, AB, Canada
11. Department of Anesthesiology and Pain Medicine, University of Toronto, Toronto, ON, Canada
12. Department of Anesthesia and Critical Care Medicine, St. Michael's Hospital, Toronto, Ontario, Canada
13. Department of Medicine and Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, AB, Canada
14. Alberta Research Centre for Health Evidence, University of Alberta, Edmonton, AB, Canada
15. School of Psychology, The University of Queensland, St Lucia, Queensland, Australia
16. Vector Institute, Toronto, ON, Canada
17. Department of Medicine, Mount Sinai Hospital and University Health Network, Toronto, ON, Canada
18. Knowledge Translation Program, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada;
19. Epidemiology Division & Institute of Health Policy, Management, and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
20. Centre for Journalology, Methodological and Implementation Research Program, Ottawa Hospital Research Institute, Ottawa, Canada
21. School of Epidemiology and Public Health, University of Ottawa, Canada
22. Institute of Health Policy, Management & Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada.
23. Department of Emergency Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

*** equal contribution and correspondence:**
christian.cao@mail.utoronto.ca
rohitarora@g.harvard.edu
paul.cento@gmail.com
† **co-supervision of work**

Structured Abstract

Importance: Systematic reviews (SRs) inform evidence-based decision making. Yet, many take over a year to complete, are labor intensive, prone to human error, and face reproducibility challenges; thus limiting access to timely and reliable information.

Objective: To validate a large language model (LLM)-based workflow (otto-SR) to automate three of the most labour intensive tasks in performing SR's: article screening, data extraction, and risk of bias assessment; and to assess its feasibility in rapidly updating existing reviews.

Design, setting, and participants: We conducted a validation study in four phases, with direct benchmarking against graduate-level human researchers in phases 1 and 2. Phase 1: article screening performance was measured across 32,357 citations from 5 systematic reviews. The reference standard consisted of the original reviews' screening decisions after full-text screening. Phase 2: data extraction performance was measured across 4,495 data points from 495 studies in 7 reviews. Phase 3: risk of bias assessment (ROB2, Newcastle-Ottawa, QUADAS2) performance was measured across 345 studies from 12 reviews. Reference standards for Phase 2 and Phase 3 were created after blinded adjudication of the original review extraction and RoB assessments. Phase 4: otto-SR was used to reproduce and update the primary analysis from an issue of Cochrane reviews (n=12 reviews, 146,276 citations), with analytical comparisons to the original meta-analyzed findings. All discrepancies underwent dual human review.

Results: *otto-SR* showed high performance in phase 1 article screening (*otto-SR*: 96.7% sensitivity, 97.9% specificity; human: 81.7% sensitivity, 98.1% specificity) and phase 2 data extraction (*otto-SR*: 93.1% accuracy; human: 79.7% accuracy). In phase 3, *otto-SR* demonstrated high interrater reliability for risk of bias judgements (ROB2 0.98, Newcastle-Ottawa 0.95, QUADAS2 0.74; Gwet AC2). In phase 4, *otto-SR*, reproduced and updated the primary analysis from an issue of Cochrane reviews. Across Cochrane reviews, *otto-SR* incorrectly excluded a median of 0 studies (IQR 0 to 0.25), and found nearly twice as many eligible studies compared to the original authors (n= 114 vs. 64). Meta-analyses based on *otto-SR* generated screening and extraction outputs, subsequently verified through dual human review, yielded newly statistically significant effect estimates in 2 reviews and negated significance in 1 review.

Conclusions and relevance: LLMs have high performance in article screening, data extraction, and risk of bias assessments. They can rapidly reproduce and update existing systematic reviews, laying the foundation for automated, scalable, and reliable evidence synthesis.

Introduction

Systematic reviews (SRs) are a key element of evidence-based decision-making.[1] However, SRs are incredibly resource-intensive, and can take over 16 months and cost upwards of $100,000 to complete.[2,3] Delays in completing SRs can have major consequences for patients and members of the public, including failure to adopt effective therapies, or prolonged use of ineffective or harmful practices initially supported by less rigorous evidence.[4,5]

While several tools have been developed to accelerate SRs,[6,7] none are capable of full automation with human-level accuracy. However, large language models (LLMs) offer new avenues to progress towards automation with their ability to process and reason about natural language. We previously demonstrated that LLMs can achieve high performance in article screening for systematic reviews.[8] Other recent work has demonstrated promise for LLMs in data extraction,[9,10] and risk of bias assessments.[11,12] However, those studies have used small datasets, or relied on author-generated reference standards.

We introduce an LLM-based workflow (*otto-SR*) to support automated and human-in-the-loop (i.e., semi-automated with human oversight) SR workflows. Our framework uses GPT-4.1 (OpenAI) for screening articles, o3-mini-high (OpenAI) for data extraction, and GPT-5 (OpenAI) for risk of bias assessments. We evaluate our workflow on these core SR components in three phases: (1) article screening, (2) data extraction, and (3) risk of bias assessment, with direct comparisons to traditional human workflows and other SR automation tools. In a fourth phase (4), we assessed real-world utility by reproducing and updating the primary analyses from an issue of Cochrane SRs (n=12 reviews) using *otto-SR*.

Methods

An agentic workflow for systematic review automation

SR workflows begin with defining a research question and performing a comprehensive keyword search for relevant citations.[13] These citations undergo abstract and full-text screening by two human reviewers independently, with disagreements resolved by a third reviewer. The final set of relevant articles undergoes data extraction and risk of bias assessment by two human reviewers independently, again adjudicated by a third reviewer. The complete human workflow is illustrated in Figure 1 (top).

otto-SR is an end-to-end LLM-based workflow supporting both automated and human-in-the-loop systematic reviews. Citations identified from the original search are uploaded in RIS format to the *otto-SR* screening agent, which uses GPT-4.1 to screen and select for inclusion abstract and full-text articles as a standalone reviewer. The resulting set of included articles is passed to the *otto-SR* extraction agent, which performs data extraction with the o3-mini-high model, and the *otto-SR RoB agent*, which performs risk of bias assessments with the GPT-5 model. For full-text screening and data extraction, retrieved PDFs are processed by Gemini 2.0 flash and converted into structured Markdown (MD) files for downstream tasks. An overview of the *otto-SR* workflow is provided in Figure 1 (bottom).

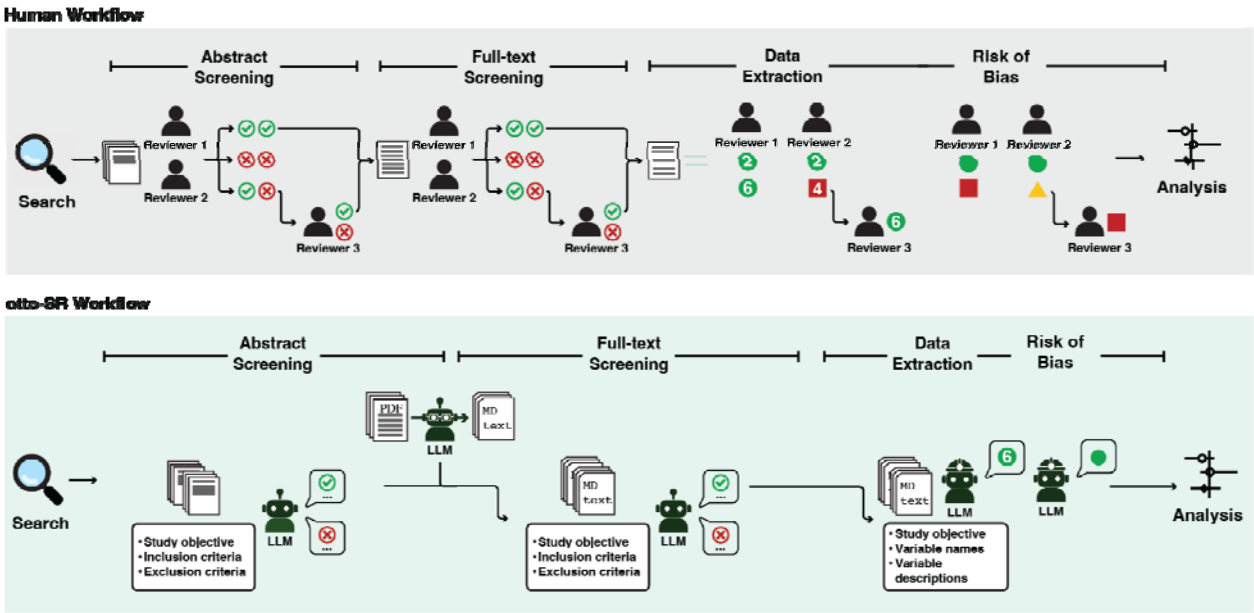


Figure 1. An automated systematic review workflow using LLMs

Infographic displaying the end-to-end SR process for humans (grey) and *otto-SR* (green). Inputs to *otto-SR* include the study protocol (objectives, eligibility criteria), search results, and defined extraction variables. Abbreviations: Markdown file (MD).

Phase 1: Evaluation of Systematic Review Screening Performance

We gathered five SRs via convenience sampling across four types of Oxford Centre for Evidence Based Medicine review questions: SeroTracker dataset[14] (prevalence), the Reinfection[15] and PA-Outcomes[16] datasets (intervention benefits), the PA-Testing[17] dataset (diagnostic test accuracy), and SVCF[18] dataset (prognosis); see Table 1 for review details.

Development of LLM Screening Agent

We developed a novel LLM-based screening agent using GPT-4.1 (OpenAI), adapted from our previously validated *ScreenPrompt* approach,[8] to screen articles at abstract and full-text stages (Supplementary Methods). The agent was prompted using the original, unaltered objectives and eligibility criteria from each respective review protocol (Supplementary Notes). Full-text article PDFs were converted into markdown (text-based formatting) with the Gemini 2.0 Flash model for full-text screening.

Screening Evaluation Methodology

We retrospectively evaluated screening performance using a diagnostic test accuracy study design. **Our reference standard was the final article inclusion or exclusion decisions of the original review authors after full-text screening**. “Included” articles represented the final set of articles included in each review, and “excluded” articles represented articles excluded from title, abstract, and full-text screening in each review.

We tested the *otto-SR* screening agent as a standalone reviewer across the full set of citations retrieved in all five SRs (n = 32,357 total citations) (Table 3,4). For comparison, we assessed the screening performance of three commercially available systematic review automation softwares (Elicit, Nested Knowledge, DistillerSR; evaluated January 2026), and postgraduate human researchers (n=4) with past SR experience, against a representative sample of citations from each SR (n=1,767 total citations) (Table 3,4; Supplementary Methods).

Human screening calibration

To validate the proficiency of our human reviewers in screening, we conducted a calibration exercise (n=400 citations). Our reviewers’ screening decisions were compared with those from an independent re-screening conducted by the original (SeroTracker) study authors, which was completed for a prior validation study.[19]

In that study,[19] the SeroTracker authors re-screened articles that had already been included or excluded as part of their living systematic review. The original SeroTracker screening decisions (prior to re-screening; generated through the initial review process) served as the reference standard for this calibration exercise, while the re-screening provided an external benchmark of expert screening performance. Our team of human reviewers had similar performance to the independent re-screening by the SeroTracker study authors (Our team: 80.2% sensitivity 97.7% specificity vs. re-screening Serotracker team: 81.3% sensitivity, 98.1% specificity) providing confidence that our reviewers had expert-level screening proficiency (Table 2).

Screening Data Analysis

We assessed the performance of the *otto-SR* screening agent, Elicit, Nested Knowledge, DistillerSR, and the external dual human reviewers by measuring accuracy, sensitivity, and specificity. We calculated 95% CIs for weighted (pooled-denominator) sensitivity and specificity using the Wilson method[20] with the binom package in R.

Phase 2: Evaluation of Systematic Review Data Extraction Performance

We gathered four datasets (SeroTracker,[14] PA-Outcomes,[16] PA-Testing,[17] Sepsis[21]) from the *BenchSR* database and three external SRs (CKD,[22] Process,[23] Psyc-meds[24]) via convenience sampling that provided publicly accessible raw data extraction information (see Table 5 for details of reviews). Variables for data extraction included descriptive information and outcome details (see Supplementary Notes).

Development of LLM-based Data Extraction Agent

We developed an extraction agent using the OpenAI o3-mini-high model.[25] In all cases, the *otto-SR* extraction agent was prompted with original author-defined variable descriptions. Full-text article PDFs were converted into markdown format with the Gemini 2.0 Flash model.

Data Extraction Evaluation Methodology

We retrospectively measured the performance of the *otto-SR* data extraction agent, Elicit, and Nested Knowledge in data extraction across all seven SRs (n=4,459 data points, 495 studies) (Table 5). The performance of DistillerSR and graduate-level human reviewers were assessed on a randomly sampled subset of articles from each review. The sample size was based on a McNemar test sample size approximation (n=1,453 data points, 156 studies) (Supplementary Methods). Graduate human reviewers performed data extraction in duplicate, with arbitration by a third reviewer.

Data extraction accuracy was determined through an LLM-as-a-judge framework to compare AI- or human extracted values (graduate-level reviewers) against the original author extractions (Supplementary Methods). However, given the known variability in dual human data extraction accuracy (reported rates: 65.8-85.5%),[26–31] original author-extracted values were not treated as a definitive gold standard. Instead, we applied a blinded adjudication process to resolve discrepancies between *otto-SR* extraction and the original authors. A panel of three blinded human reviewers compared randomized pairs of responses (*otto-SR* vs. original author) and selected the most accurate value (Supplementary Methods). **We used these final adjudicated judgements to construct a corrected reference standard** for all performance evaluations. This adjudication framework was adapted from established protocols in radiology, ophthalmology, and clinical trials.[32–35]

We additionally performed a post-hoc review of incorrect AI extraction to classify errors as either ‘parsing errors’ (i.e., errors with converting PDF files to markdown formatting), ‘inaccessible’ (i.e., data accessible only through author correspondence or supplementary material), or ‘true errors’ (i.e., cases where the original author values were incorrectly extracted).

Data Extraction Analysis

We assessed the performance of the *otto-SR* data extraction agent, Elicit, Nested Knowledge, DistillerSR, and external human reviewers by measuring the total accuracy at a variable level per study. If the human adjudicator classification was "inaccessible", the data point was removed from analysis. We calculated 95% CIs for weighted (pooled-denominator) accuracy using the Wilson method[20] with the binom package in R.

Phase 3: Evaluation of Systematic Review Risk of Bias Assessment Performance

We gathered 12 SR datasets via convenience sampling that reported publicly accessible risk of bias assessments using Cochrane RoB2[36] for RCTs (n=4 SRs, 214 studies),[37–40] Newcastle-Ottawa-Scale[41] for cohort studies (n=4 SRs, 75 studies),[42–45] and QUADAS2[46] for diagnostic test accuracy studies (n=4 SRs, 56 studies).[47–50]

Development of an LLM-based Risk of Bias Agent

We developed a risk of bias (ROB) agent using the OpenAI GPT-5 model. In all cases, the *otto-SR* ROB agent was prompted with the full critical appraisal tool instructions as published by the creators (ROB2, NOS, or QUADAS2), and review dataset metadata required for appraisal (e.g., outcomes assessed, objectives).

Risk of Bias Evaluation Methodology

We compared the performance of the otto-SR RoB agent with original review author appraisals for each validated tool (e.g., ROB2). As risk of bias is inherently subjective and prone to interpretive variability, we applied a blinded adjudication process involving independent reviewers to resolve disagreements between otto-SR and author assessments. **The resulting adjudicated judgements were used to construct a corrected reference standard.** Both *otto-SR* and original review author decisions were evaluated against this corrected reference standard.

Risk of Bias Analysis

We assessed the performance of the *otto-SR* RoB agent by measuring the total accuracy of risk of bias assessments (number of concordant appraisal assessments for each domain/question). We report confusion matrices (per domain; overall RoB) and inter-rater reliability metrics, including Gwet's AC2, which was selected for its robustness to category prevalence and class imbalance (Supplementary Methods). We calculated 95% CIs for weighted (pooled-denominator) accuracy using the Wilson method[20] with the binom package in R.

Phase 4: Evaluation of otto-SR for Reproducing and Updating Cochrane Reviews

To evaluate the reliability and generalizability of our automated SR workflow, we conducted a focused reproducibility assessment using one complete issue of the Cochrane Database of Systematic Reviews. Our aim was to reproduce each review's workflow end-to-end, from

literature search to meta-analysis, while using the *otto-SR* pipeline for screening, extraction, and RoB assessments.

We followed Cochrane methodology with one exception. Cochrane reviews typically include studies even if they do not report data for the review's primary outcome. This method allows for reporting of all outcomes for potential comparisons (e.g., all intervention and outcome combinations). In contrast, we focused on reproducing each review's predefined primary outcome. This constraint provided a clearer distinction for study eligibility.

We randomly selected the April 2024 issue of the Cochrane Database (Table 8, Supplementary Methods).[51–64] Of the 14 reviews in this issue, one review was excluded due to a lack of publicly available data, and a second review was excluded due to the absence of a reproducible search strategy.

Cochrane Database Searches

The original search strategy of each Cochrane review was reproduced using the exact terms and filters described in the review methods. Searches were limited to institutionally accessible databases. In cases where databases lacked precise date filtering (e.g., supporting month but not day-level granularity), we applied post-hoc filtering to approximate the original search window (Supplementary Data 2).

After each search, we cross-referenced our list of articles with those included in the original Cochrane reviews. Articles that were not retrievable using the original database search strategies were excluded from downstream screening, data extraction and analysis.

Cochrane Screening

All retrieved citations underwent abstract and full-text screening with the *otto-SR* screening agent, prompted with the original objectives, and inclusion and exclusion criteria from each Cochrane review (Supplementary Notes).

We retrieved additional abstract or DOI information from incomplete citation records through OpenAlex.[65] Citations without a retrievable abstract or DOI/trial identifier were excluded. Screening decisions were compared against the original Cochrane author decisions to calculate true positives, true negatives, false negatives, and false positives.

Cochrane Data Extraction

All studies included after full-text screening underwent data extraction using the *otto-SR* extraction agent which focused on the primary outcome defined in each Cochrane review. The extraction prompts used the original author-defined variable names and extraction logic (Supplementary Notes).

Data extraction also served as a final screening filter. Studies were programmatically excluded if they returned unavailable or unreportable values for the primary outcome (e.g., "na"), were

identified as duplicates, or involved only ineligible intervention-comparator pairs (e.g., Drug A vs. Drug B when only Drug A vs. placebo was eligible).

Cochrane Statistical Analysis

All meta-analyses were conducted using the metafor package in R (Code and Dataset Availability). We matched the original authors' specified meta-analytic model (e.g., random-effects, fixed-effect), effect size metric (risk ratio, odds ratio, rate ratio, mean difference, standardized mean difference), and continuity correction approach, where reported.

We conducted four meta-analytical comparisons: (1) ***Cochrane*** – meta-analysis using the original author-extracted data; (2) ***Matched*** – *otto-SR* results filtered to match the studies included by the original Cochrane authors; (3) ***Expanded*** – all eligible studies included by *otto-SR* under the original search date cut-off including those not in the original Cochrane review; (4) ***Updated*** – all eligible studies included by *otto-SR* using an updated search date extending to May 8 2025. We also derived 'corrected' values, for the 'matched,' 'expanded,' and 'updated' comparisons that served as the reference ground truth for analytical comparison.

Cochrane Data Correction and Comparison

Given potential data extraction errors by original Cochrane authors and *otto-SR*, we derived corrected values for each analytical comparison through dual human review. This included removal of false positive articles and inclusion of false negative articles, as well as correction of extracted data points. **All original Cochrane data, *otto-SR* extracted data, and corrected data (including descriptions of why data were corrected) are provided in Supplementary Data 1.** Details on our data correction process are provided in the Supplementary Methods.

Cochrane Risk of Bias assessment

We applied the otto-SR RoB agent to the articles included in the 'matched' analysis, restricting inclusion to reviews using ROB2, Newcastle Ottawa, or QUADAS2 tools. Details on our risk of bias assessment and correction process is provided in the Supplementary Methods.

Results

Phase 1: Evaluation of Screening Performance

At the abstract screening stage, the *otto-SR* screening agent achieved the highest sensitivity (weighted sensitivity 96.6% [total range, 94.1-100.0%]) (Fig 2, Table 3) followed by Elicit (88.5% [76.9-100%] sensitivity), dual human reviewers (87.3% [84.1-100%] sensitivity), DistillerSR (61.4% [50.3-93.8%] sensitivity), and Nested Knowledge (55.7% [17.2-95.6%] sensitivity). Dual human reviewers achieved the highest specificity in abstract screening (95.7% [92.5-98.7%] specificity), followed by DistillerSR (95.3% [91.6-100%] specificity), the *otto-SR* screening agent (93.9% [83.6-97.7%] specificity), Nested Knowledge (87.7% [75.0-98.0%] specificity), and Elicit (84.2% [65.7-95.9%] specificity).

After full-text screening, the *otto-SR* screening agent maintained the highest sensitivity (96.2% [92.3-100%] sensitivity), followed by Elicit (86.6% [79.12-100%] sensitivity), while human reviewers had a marked drop in sensitivity (63.3% [44.1-93.8%] sensitivity) (Fig 2, Table 4). This decline in human sensitivity was largely driven by poor performance on screening the "Reinfection" review (44.1% sensitivity, 95.3% specificity), likely due to complex inclusion criteria involving test-negative study designs, multiple interventions, and multiple time-specific outcomes. After removing this outlier review, human reviewers achieved a weighted sensitivity of 81.7% [76.4%-93.8%]. Nested Knowledge had the lowest full-text screening sensitivity of all four groups tested (53.2% [14.5-93.8%] sensitivity). Specificity remained high for both the *otto-SR* screening agent (96.9% [90.7-98.7%] specificity) and dual human reviewers (98.1% [96.7-100.0%] specificity). Elicit had the lowest specificity among all four groups tested (86.6% [82.6-98.0%] specificity).

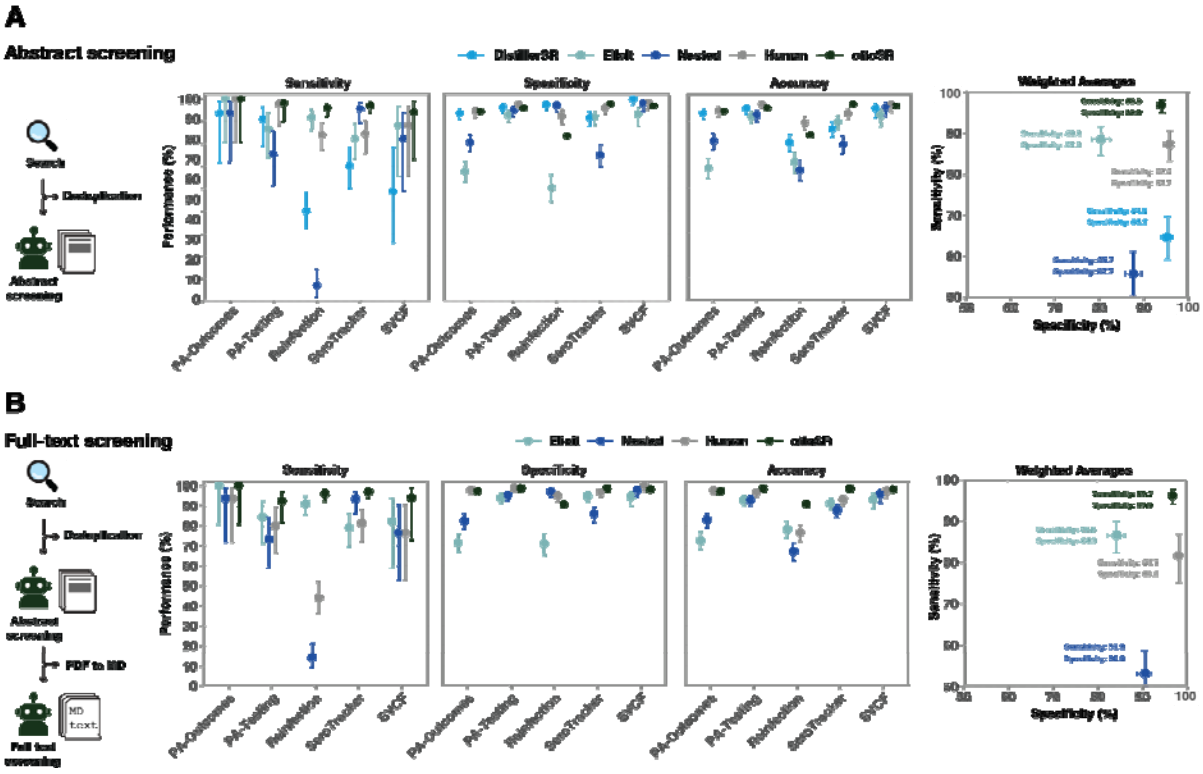


Figure 2. *otto-SR* screening agent performance on systematic review screening

A. diagram of *otto-SR* **abstract screening** agent (left), sensitivity, specificity of *otto-SR* screening agent, dual human reviewers, Elicit, Nested Knowledge, and DistillerSR for abstract screening evaluated across five reviews (middle). Weighted average for sensitivity and specificity across comparator groups (five reviews) (right). Error bars indicate 95% confidence intervals.

B. diagram of *otto-SR* **full-text screening** agent (left), sensitivity, specificity, of *otto-SR* screening agent, dual human reviewers, Elicit, and Nested Knowledge for full-text screening evaluated across five reviews (middle). Weighted average for sensitivity and specificity across otto-SR, Elicit, Nested Knowledge (five reviews) and dual human (four reviews) (right). Error bars indicate 95% confidence intervals. Abbreviations: Markdown file (MD).

Phase 2: Evaluation of Data Extraction Performance

Across all seven reviews, the *otto-SR* extraction agent achieved an average weighted accuracy of 93.1% (91.1.4-97.0%), outperforming dual human reviewers at 79.7% (69.1-91.0%), Elicit at 76.0% (60.7-85.8%), Nested Knowledge at 71.2% (39.1-83.8%), and DistillerSR at 46.1% (29.9-63.2%) (Fig 3a, Table 6). Notably, our human reviewers performed comparably to prior estimates of human data extraction accuracy (Fig 3b). When *otto-SR* extracted different values to the original authors, the blinded human reviewers sided with the *otto-SR* data extraction agent in 69.3% of cases (Fig 3c). In contrast, for discrepancies between original authors and our two human extractors, the blinded reviewer panel sided with the dual human extractors in 28.1% of cases.

In the 6.9% of cases where the *otto-SR* extraction agent was incorrect, post-hoc analysis revealed that 0.83% (37/4459) of data points were inaccessible to the model (supplementary files or data obtained through data request), 0.67% (30/4459) resulted from parsing errors, and 0.49% (22/4459) were cases where neither the *otto-SR* data extraction agent nor original author extraction was correct (Supplementary Fig 1).

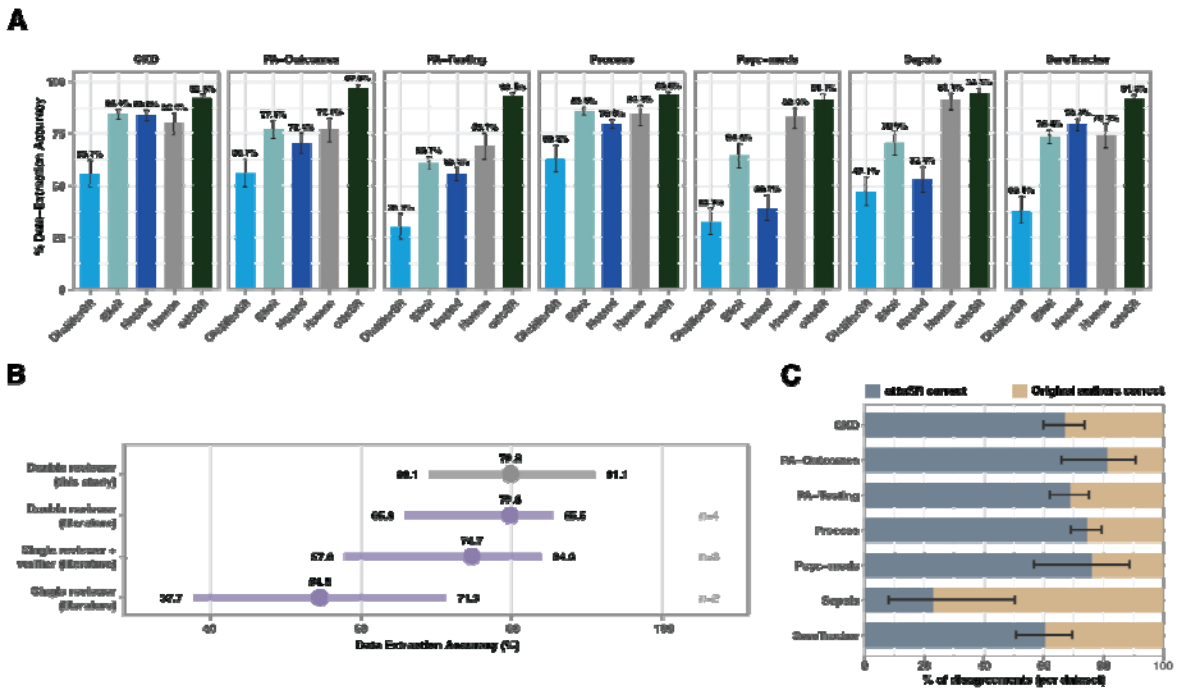


Figure 3. *otto-SR* extraction agent performance on systematic review data extraction
A. bar graph displaying data extraction accuracy of the *otto-SR* extraction agent (green) (4459 data points), Elicit (teal) (4459 data points), Nested Knowledge (dark blue) (4459 data points), DistillerSR (light blue) (1,453 data points), and dual human reviewers (grey) (1,453 data points) across 7 different systematic reviews. Error bars represent 95% confidence intervals.

B. dot plot depicting literature-derived human reviewer performance comparison (purple) against human reviewers in this study (grey). Dots represent mean value and upper and lower bars represent range.
C. bar graph depicting dual human adjudicator decisions for values marked as incongruent between original review and *otto-SR* extraction agent and dual human. Blue represents the newly conducted review being correct, while tan represents the original study authors being correct. Error bars represent 95% confidence intervals.

Phase 3: Evaluation of Risk of Bias Assessment Performance

Following blinded adjudication, both the *otto-SR* RoB agent and the original review author assessments were evaluated against a corrected reference standard. Against this reference, *otto-SR* achieved an average weighted accuracy of 96.6% (95% CI: 90.0–97.9) for RoB 2, 94.9% (93.9–97.8) for Newcastle–Ottawa Scale (NOS) cohort assessments, and 78.3% (71.4–84.3) for QUADAS-2 assessments (Figure 4; Table 7, Supplementary Fig. 2). Using the same corrected reference standard, the original review author assessments achieved weighted accuracies of 93.6% (87.1-95.6) for RoB 2, 91.4% (88.9-97.6) for NOS, and 78.3% (53.6-88.3) for QUADAS-2. When *otto-SR* generated different appraisal decisions to the original authors, the blinded human reviewers sided with the *otto-SR* RoB agent in 2.6% of cases (Fig 4). Of note, *otto-SR* had no major discrepancies (i.e., low vs. high ratings) in overall RoB2 domain assessments (Supplementary Fig 3). *otto-SR* also had high interrater reliability, with an average Gwet AC2 of 0.98, 0.95, and 0.74 for ROB2, Newcastle, and QUADAS2 tools, respectively, which was comparable to the agreement between the original review authors (Table 7).

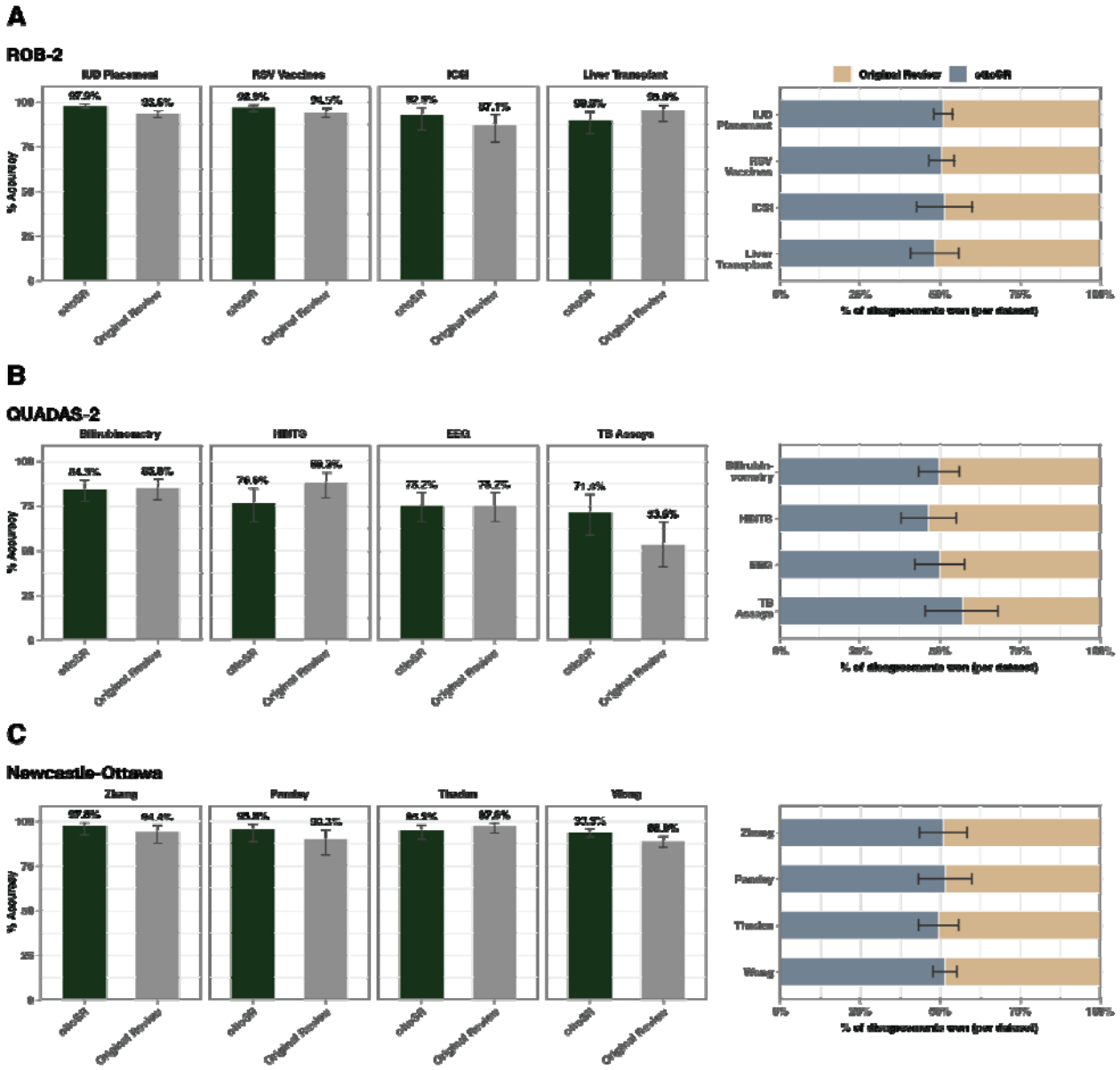


Figure 4. *otto-SR RoB agent* performance on systematic review risk of bias

A. Left: bar graph displaying per domain risk of bias accuracy of the *otto-SR* extraction agent (green) (1060 data points) and original review decisions, relative to corrected reference standard across four different systematic reviews with ROB2 risk of bias assessments. Error bars represent 95% confidence intervals.

Right: bar graph depicting human adjudicator decisions for values marked as incongruent between original review and *otto-SR RoB agent*. Blue bars indicate instances where the otto-SR risk of bias assessment was adjudicated as correct, while tan bars indicate instances where the original review risk of bias assessment was adjudicated as correct. Error bars represent 95% confidence intervals.

B. Left: bar graph displaying per domain risk of bias accuracy of the *otto-SR* extraction agent (green) (378 data points) and original review decisions, relative to corrected reference standard

across four different systematic reviews with QUADAS2 risk of bias assessments. Error bars represent 95% confidence intervals.
Right: bar graph depicting human adjudicator decisions for values marked as incongruent between original review and *otto-SR RoB agent*. Blue bars indicate instances where the otto-SR risk of bias assessment was adjudicated as correct, while tan bars indicate instances where the original review risk of bias assessment was adjudicated as correct. Error bars represent 95% confidence intervals.
C. Left: bar graph displaying per domain risk of bias accuracy of the *otto-SR* extraction agent (green) (666 data points) and original review decisions, relative to corrected reference standard across four different systematic reviews with Newcastle-Ottawa risk of bias assessments. Error bars represent 95% confidence intervals.
Right: bar graph depicting human adjudicator decisions for values marked as incongruent between original review and *otto-SR RoB agent*. Blue bars indicate instances where the otto-SR risk of bias assessment was adjudicated as correct, while tan bars indicate instances where the original review risk of bias assessment was adjudicated as correct. Error bars represent 95% confidence intervals.

Phase 4: Evaluation of otto-SR for Reproducing and Updating Cochrane Reviews
We combined the *otto-SR* screening and extraction agents into a single agentic workflow (Fig 5a) and used it to reproduce and update the primary analysis of 12 reviews from the April 2024 issue of the Cochrane Database of Systematic Reviews (Table 8). We reproduced the reported search strategies, updating searches to May 8, 2025, and identified 146,276 citations.

The *otto-SR* screening agent correctly identified all included studies (n=64, 100% sensitivity) across the 12 Cochrane reviews. *otto-SR* extraction results with missing primary outcome values, duplicate studies, or missing intervention-comparator groups were programmatically excluded (Methods). After this process, *otto-SR* incorrectly excluded a median of 0 articles (IQR 0 to 0.25) (Table 9). Incorrect exclusions were due to LLM-inaccessible supplementary data (n=2) or a failure to extract reported outcome values when present (n=2).

After filtering our results to match the original search date cutoffs, we identified 54 additional eligible studies ('***Expanded***') through *otto-SR* (median 2, IQR: 1 to 6.25 per review) that may have been missed in the original Cochrane reviews (Methods). *otto-SR* incorrectly included 10 false positive articles. However 9 of the 10 articles may have contained relevant data available through author data request.

Updating the search date to May 8, 2025 ('***Update***') identified another 14 new eligible studies (median 2.5, IQR 1 to 7.25 per review) (Table 9). The updated search identified two additional false positive studies, one of which may have contained relevant data through author data request.

Compared to the corrected per domain ROB2 assessments, *otto-SR* had high weighted accuracy (96.6%) and interrater concordance (0.96, Gwet AC2), with no major discrepancies (i.e., low vs. high) in overall RoB judgements (Supplementary Table 2, Supplementary Fig. 2,3)

In the '**Matched**' comparison group, *otto-SR* produced meta-analyzed effect estimates that had overlapping 95% CIs with both the original Cochrane data and corrected datasets across all reviews (Fig 5b, left; Table 10). In the '**Expanded**' analysis, two reviews (nutrition, depression) yielded new statistically significant effect estimates for the primary outcome (Fig 5b, middle), while the estimate from one review (alcohol) lost statistical significance compared to the original Cochrane estimates (Fig 5b, middle). These trends were consistent in the corrected '**Expanded**', corrected '**Update**', and *otto-SR* 'updated' analyses (Fig 5b, right).

One illustrative example comes from the nutrition review, where *otto-SR* identified 5 additional studies in the 'Expanded' comparison. This led to the new finding that preoperative immune-enhancing supplementation before gastric surgery is associated with a significant one-day reduction in mean hospital stay compared to usual care (*otto-SR:* MD -1.20 [95% CI -2.28 to -0.11], 9 studies vs Cochrane: MD -0.19 [-1.44 to 1.07], 4 studies). Detailed effect estimates and 95% CIs are provided for all groups and comparisons are provided in Table 10.

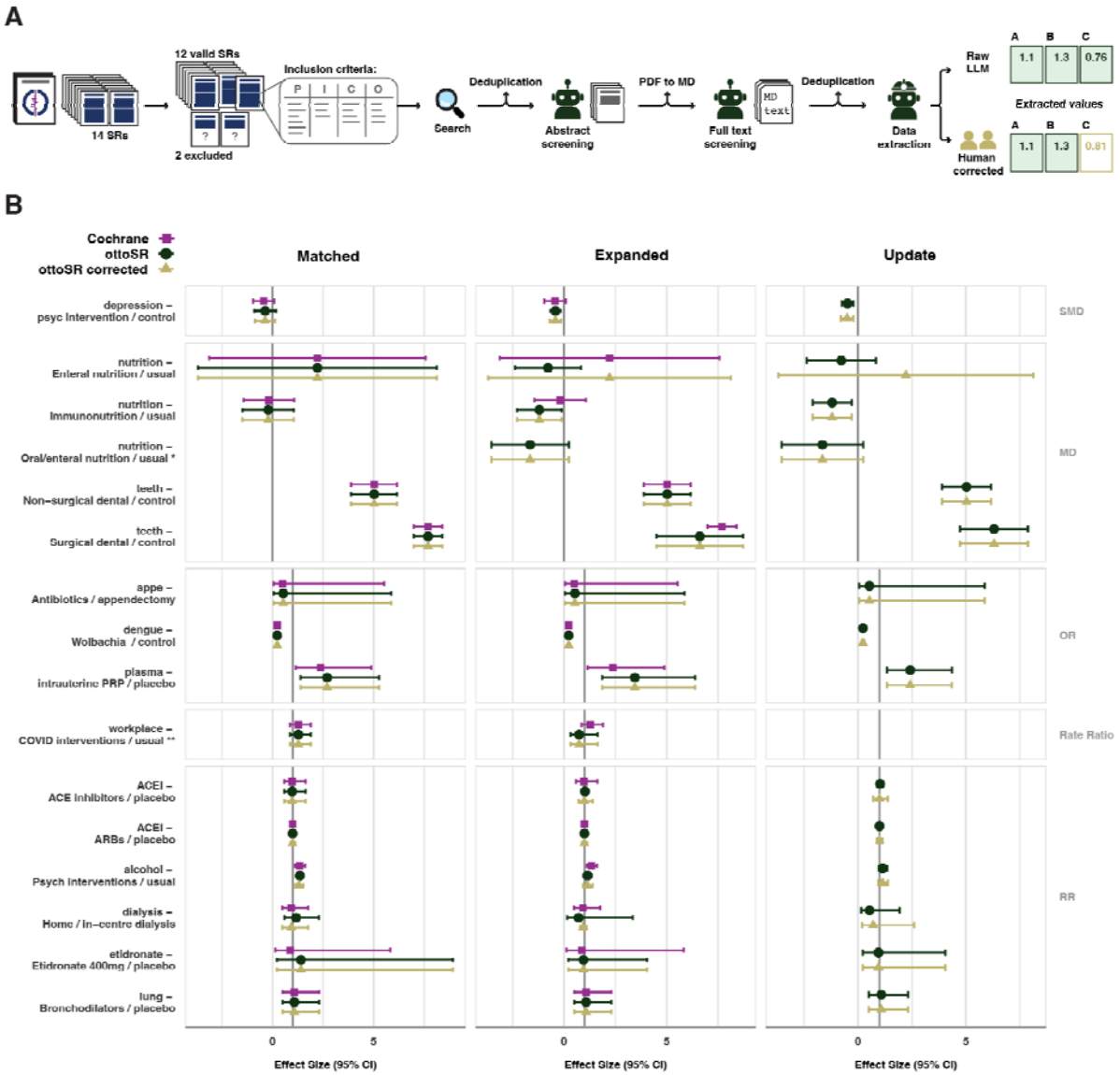


Fig 5. Evaluating *otto-SR* for automation of systematic reviews

A. Infographic depicting use of *otto-SR* for SR automation in a complete edition of the Cochrane Database of Systematic Reviews (n = 12 SRs, n=64 studies).

B. Forest plots depicting differences between *otto-SR* (green), original Cochrane study authors (purple), and corrected standard (gold). The 'corrected' estimates were derived through dual human review of data extraction values to establish a ground truth. Each row is representative of meta-analyzed estimates derived in a systematic review. Error bars represent 95% confidence interval, MD = Mean Difference, OR = Odds Ratio, RR = Risk Ratio, SMD = Standardized Mean Difference. The **matched** comparison (left) shows estimates derived from articles only included in the original Cochrane reviews. The **expanded** comparison (middle) displays estimates derived from additional articles identified by *otto-SR* falling within the original search dates. The **update** plot (right) displays estimates derived from all articles found by *otto-SR* in a May 8 2025 search. * *otto-SR* discovered a new treatment group, mixed oral / enteral nutrition, which was not found in the original Cochrane review, consequently no matched

analysis was conducted. ** workplace citations were provided by original study authors due to challenges with the electronic search, consequently no updated search was performed.

Discussion

In this study, we demonstrate that *otto-SR*, an LLM-based SR automation pipeline, can accelerate key article screening, data extraction, and risk of bias assessment steps with high performance. We also demonstrate the feasibility of applying an LLM workflow to reproduce and update the primary analysis from multiple Cochrane reviews.

Our work has several strengths. First, we evaluated an LLM workflow for multiple phases of the SR process. Second, our phase 4 analysis examined downstream changes in effect estimates when reproducing and updating Cochrane reviews using automated screening, data extraction, and risk of bias assessment. Many prior studies have evaluated individual components in isolation, and do not assess how missed studies or extraction errors propagate into final meta-analyses.[66–70] Third, otto-SR was applied using unaltered review objectives and protocols, demonstrating the potential for generalizable use of LLM-based workflows. This contrasts with prior work relying on review-specific prompt optimization,[66,71] which may limit external applicability. Finally, we tested *otto-SR* on a diverse set of reviews including reviews of prevalence, prognosis, diagnostic test accuracy, and intervention effectiveness, thus increasing the generalizability of the findings.

Our work has several limitations. First, although our analysis was validated across a wide range of SRs, further research is needed to assess the generalizability to other clinical questions and qualitative reviews. Second, our LLM parser may have incorrectly extracted information from PDF articles. In future studies, enhanced vision capabilities of new models may better support screening/extraction efforts. Third, our workflow was limited to the main text of articles and did not extract data from supplementary tables or figures. While we could have manually incorporated these materials, we intentionally avoided human intervention to test the end-to-end capabilities of our automated workflow. Fourth, due to the vast number of data points (and their unstandardized nature), we employed an LLM-as-a-judge framework to assess data extraction accuracy. Although this approach has been previously validated,[72,73] LLM judgements may still introduce errors. Fifth, we encountered instances where the original author's decisions for data extraction appeared inaccurate. However, we addressed this by performing randomized and blinded adjudication, adopting best practices seen in radiology, ophthalmology, and clinical trials.[32–35] This approach represented a methodological improvement over prior work that relied on self-created ground truths.[9] Sixth, the GPT-4.1 knowledge cut-off was updated to June 2024 (released April 2025). However, our screening and data extraction benchmarks primarily involved closed-source data, and the April 2024 Cochrane reviews were closed access during the time of model pretraining (Cochrane articles are closed-access for a year after publication); mitigating concerns about potential pretraining data overlap. Finally, with respect to the Cochrane reproducibility assessment, it’s possible that content specific experts would have made different article inclusion and data extraction decisions. However, we closely adhered to each review’s protocol, contacted study authors to clarify methodological uncertainties, and documented all discrepancies.

Our findings highlight several opportunities for LLMs in SRs. First, workflows like *otto-SR* may be used to update existing SRs by leveraging their original published protocols. This provides a unique advantage: it enables validation of screening and extraction using the original review. Second, the ability to rapidly screen articles, extract articles, and assess risk of bias would allow LLMs to support living SRs for which updates could be performed monthly, weekly, or even daily to ensure constant access to the best available cumulative evidence. This livingness reflects knowledge - it is always evolving and not fixed. Third, after further prospective validation, LLMs may be used to generate *de novo* reviews, provided that researchers develop clear research questions with detailed protocols akin to those registered in PROSPERO. In all cases, structured, transparent, and clear methodology is essential for ensuring interpretability, reproducibility, and high-quality automation.

The Phase 4 Cochrane reproducibility assessment identified common reproducibility challenges. All 12 reviews had issues with search reproducibility or methodological uncertainty (Table 8). These findings align with prior work by Rethlefsen et al.[74] who found that only 1% of reviews have a fully reproducible search strategy. Previous studies have also shown that reproducibility failures can occur at every stage of the SR process.[74–79] Furthermore, there are few SR reproducibility experiments; the results here (n=12 reviews) and by Hamilton et al.[80] (n=10 reviews) are the largest to date. To address these challenges, published SRs should include: (i) a complete search strategy; (ii) raw search files (e.g., RIS file); (iii) raw data extraction outputs with data dictionaries; (iv) list of data procured via author correspondence; and (v) code used for analyses. Current reporting guidelines for systematic reviews (PRISMA) endorse most, but not all of these points.[81] While Cochrane reviews routinely provide such materials, most other SRs do not, limiting reproducibility and external validation. Tools like *otto-SR* can support automatic generation and reporting of these files, reducing the burden on authors.

The high performance of LLMs in conducting evidence synthesis, as demonstrated in this study, may warrant consideration in how scientific content is published. While most research is written for human readers, the rise of LLM-supported evidence curation highlights the value of making studies machine-readable. Using structured formats like markdown or html (as now offered by Arxiv), and providing raw numerical data from figures could support this new paradigm.

Conclusion

These findings represent an advancement in the development of SR automation tools. Downstream applications include rapid SR updates, truly ‘living’ reviews, and mass assessments of reproducibility across the SR literature. Future research should focus on developing comprehensive and complete benchmarks of SRs to better support and refine automation efforts, and evaluating the performance of LLMs on prospective de novo reviews. We also encourage research into the capabilities of LLMs for other SR workflow tasks, such as search term generation and statistical analysis. The implementation of more autonomously completed SRs could accelerate the synthesis of up-to-date evidence, saving thousands of hours of manual work, and transform how scientists, clinicians, and patients make decisions.

References

1. L'Abbé KA, Detsky AS, O'Rourke K. Meta-Analysis in Clinical Research. Ann Intern Med 1987;107(2):224–33.

2. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017;7(2):e012545.

3. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun 2019;16:100443.

4. Prasad V, Cifu A. Medical reversal: why we must raise the bar before adopting new technologies. Yale J Biol Med 2011;84(4):471–8.

5. Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative Meta-Analysis of Therapeutic Trials for Myocardial Infarction. N Engl J Med 1992;327(4):248–54.

6. Fabiano N, Gupta A, Bhambra N, Luu B, Wong S, Maaz M, et al. How to optimize the systematic review process using AI tools. JCPP Adv 2024;4(2):e12234.

7. Ge L, Agrawal R, Singer M, Kannapiran P, De Castro Molina JA, Teow KL, et al. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. Syst Rev 2024;13(1):269.

8. Cao C, Sang J, Arora R, Chen D, Kloosterman R, Cecere M, et al. Development of Prompt Templates for Large Language Model–Driven Screening in Systematic Reviews. Ann Intern Med 2025;ANNALS-24-02189.

9. Lai H, Liu J, Bai C, Liu H, Pan B, Luo X, et al. Language models for data extraction and risk of bias assessment in complementary medicine. Npj Digit Med 2025;8(1):74.

10. Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, Crotty K, et al. Data extraction for evidence synthesis using a large language model: A proof□of□concept study. Res Synth Methods 2024;15(4):576–89.

11. Lai H, Ge L, Sun M, Pan B, Huang J, Hou L, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA Netw Open 2024;7(5):e2412687.

12. Huang J, Lai H, Zhao W, Xia D, Bai C, Sun M, et al. Large Language Model–Assisted Risk-of-Bias Assessment in Randomized Controlled Trials Using the Revised Risk-of-Bias Tool: Evaluation Study. J Med Internet Res 2025;27:e70450.

13. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev [Internet] 2019 [cited 2024 Jun 1];Available from: https://doi.wiley.com/10.1002/14651858.ED000142

14. Bobrovitz N, Arora RK, Cao C, Boucher E, Liu M, Donnici C, et al. Global seroprevalence of SARS-CoV-2 antibodies: A systematic review and meta-analysis. PLOS ONE 2021;16(6):e0252617.

15. Bobrovitz N, Ware H, Ma X, Li Z, Hosseini R, Cao C, et al. Protective effectiveness of previous SARS-CoV-2 infection and hybrid immunity against the omicron variant and severe disease: a systematic review and meta-regression. Lancet Infect Dis 2023;23(5):556–67.

16. Samnani S, Cenzer I, Kline GA, Lee SJ, Hundemer GL, McClurg C, et al. Time to Benefit of Surgery vs Targeted Medical Therapy for Patients With Primary Aldosteronism: A Meta-analysis. J Clin Endocrinol Metab 2024;109(3):e1280–9.

17. Leung AA, Symonds CJ, Hundemer GL, Ronksley PE, Lorenzetti DL, Pasieka JL, et al. Performance of Confirmatory Tests for Diagnosing Primary Aldosteronism: a Systematic Review and Meta-Analysis. Hypertension 2022;79(8):1835–44.

18. Mascarenhas D, Weisz D, Jasani B, Persad N, Main E. Premedication for rapid sequence intubation in neonates - a network meta-analysis. PROSPERO 2022 CRD42022384259 [Internet]. PROSPEROAvailable from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022384259

19. Perlman□Arrow S, Loo N, Bobrovitz N, Yan T, Arora RK. A real□world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Res Synth Methods 2023;14(4):608–21.

20. Wilson EB. Probable Inference, the Law of Succession, and Statistical Inference. J Am Stat Assoc 1927;22(158):209–12.

21. Teja B, Berube M, Pereira TV, Law AC, Schanock C, Pang B, et al. Effectiveness of Fludrocortisone Plus Hydrocortisone versus Hydrocortisone Alone in Septic Shock: A Systematic Review and Network Meta-Analysis of Randomized Controlled Trials. Am J Respir Crit Care Med 2024;209(10):1219–28.

22. Cleary F, Prieto-Merino D, Nitsch D. A systematic review of statistical methodology used to evaluate progression of chronic kidney disease using electronic healthcare records. PLOS ONE 2022;17(7):e0264167.

23. Antonacci G, Lennox L, Barlow J, Evans L, Reed J. Process mapping in healthcare: a systematic review. BMC Health Serv Res 2021;21(1):342.

24. Kopcalic K, Arcaro J, Pinto A, Ali S, Barbui C, Curatoli C, et al. Antidepressants versus placebo for generalised anxiety disorder (GAD). Cochrane Database Syst Rev [Internet] 2025 [cited 2025 Jun 8];2025(2). Available from: http://doi.wiley.com/10.1002/14651858.CD012942.pub2

25. OpenAI. OpenAI o3-mini System Card [Internet]. 2025;Available from: https://cdn.openai.com/o3-mini-system-card-feb10.pdf

26. Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. Systematic review data extraction: cross-sectional study showed that experience did not increase

accuracy. J Clin Epidemiol 2010;63(3):289–98.

27. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol 2006;59(7):697–703.

28. Carroll C, Scope A, Kaltenthaler E. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. BMC Res Notes 2013;6(1):539.

29. Tang L, Wang R, Doi SAR, Furuya-Kanamori L, Lin L, Qin Z, et al. Double data extraction was insufficient for minimizing errors in evidence synthesis: a randomized controlled trial [Internet]. 2023 [cited 2025 Jun 8];Available from: http://medrxiv.org/lookup/doi/10.1101/2023.10.16.23297056

30. Xu C, Yu T, Furuya-Kanamori L, Lin L, Zorzela L, Zhou X, et al. Validity of data extraction in evidence synthesis practice of adverse events: reproducibility study. BMJ 2022;e069155.

31. Li T, Saldanha IJ, Jap J, Smith BT, Canner J, Hutfless SM, et al. A randomized trial provided new evidence on the accuracy and efficiency of traditional vs. electronically annotated abstraction approaches in systematic reviews. J Clin Epidemiol 2019;115:77–89.

32. Plesner LL, Müller FC, Brejnebøl MW, Krag CH, Laustrup LC, Rasmussen F, et al. Using AI to Identify Unremarkable Chest Radiographs for Automatic Reporting. Radiology 2024;312(2):e240272.

33. Nguyen HQ, Lam K, Le LT, Pham HH, Tran DQ, Nguyen DB, et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Sci Data 2022;9(1):429.

34. Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India. JAMA Ophthalmol 2019;137(9):987.

35. Godolphin PJ, Bath PM, Algra A, Berge E, Brown MM, Chalmers J, et al. Outcome Assessment by Central Adjudicators Versus Site Investigators in Stroke Trials: A Systematic Review and Meta-Analysis. Stroke 2019;50(8):2187–96.

36. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;l4898.

37. Kamath MS, Vogiatzi P, Sunkara SK, Woodward B. Oocyte activation for women following intracytoplasmic sperm injection (ICSI). Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Jan 29];2024(12). Available from: http://doi.wiley.com/10.1002/14651858.CD014040.pub2

38. Zapata LB, Snyder E, Nguyen AT, Kapp N, Ti A, Whiteman MK, et al. Misoprostol for intrauterine device placement. Cochrane Database Syst Rev [Internet] 2025 [cited 2026 Jan 29];2025(9). Available from: http://doi.wiley.com/10.1002/14651858.CD015584.pub2

39. Tingle SJ, Dobbins JJ, Thompson ER, Figueiredo RS, Mahendran B, Pandanaboyana S, et al. Machine perfusion in liver transplantation. Cochrane Database Syst Rev [Internet] 2023 [cited 2026 Jan 29];2023(9). Available from: http://doi.wiley.com/10.1002/14651858.CD014685.pub2

40. Saif-Ur-Rahman K, King C, Whelan SO, Blair M, Donohue S, Madden C, et al. Efficacy and safety of respiratory syncytial virus vaccines. Cochrane Database Syst Rev [Internet] 2025 [cited 2026 Jan 29];2025(9). Available from: http://doi.wiley.com/10.1002/14651858.CD016131

41. Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses [Internet]. Newctle.-Ott. Scale NOS Assess. Qual. Nonrandomised Stud. Meta-Anal.2026;Available from: https://ohri.ca/en/who-we-are/core-facilities-and-platforms/ottawa-methods-centre/newcastle-ottawa-scale

42. Zhang T, Mei Q, Zhang Z, Walline JH, Liu Y, Zhu H, et al. Risk for newly diagnosed diabetes after COVID-19: a systematic review and meta-analysis. BMC Med 2022;20(1):444.

43. Wong C, Mohamad Asfia SKB, Myles PS, Cunningham J, Greenhalgh EM, Dean E, et al. Smoking and Complications After Cancer Surgery: A Systematic Review and Meta-Analysis. JAMA Netw Open 2025;8(3):e250295.

44. Paluch AE, Bajpai S, Bassett DR, Carnethon MR, Ekelund U, Evenson KR, et al. Daily steps and all-cause mortality: a meta-analysis of 15 international cohorts. Lancet Public Health 2022;7(3):e219–28.

45. Pandey A, Salahuddin U, Garg S, Ayers C, Kulinski J, Anand V, et al. Continuous Dose-Response Association Between Sedentary Time and Risk for Cardiovascular Disease: A Meta-analysis. JAMA Cardiol 2016;1(5):575.

46. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med 2011;155(8):529–36.

47. Okwundu CI, Olowoyeye A, Uthman OA, Smith J, Wiysonge CS, Bhutani VK, et al. Transcutaneous bilirubinometry versus total serum bilirubin measurement for newborns. Cochrane Database Syst Rev [Internet] 2023 [cited 2026 Jan 29];2023(5). Available from: http://doi.wiley.com/10.1002/14651858.CD012660.pub2

48. Inbaraj LR, Daniel J, Sathya Narayanan MK, Srinivasalu VA, Bhaskar A, Scandrett K, et al. Truenat MTB assays for pulmonary tuberculosis and rifampicin resistance in adults and adolescents. Cochrane Database Syst Rev [Internet] 2025 [cited 2026 Jan 29];2025(3). Available from: http://doi.wiley.com/10.1002/14651858.CD015543.pub2

49. Gottlieb M, Peksa GD, Carlson JN. Head impulse, nystagmus, and test of skew examination for diagnosing central causes of acute vestibular syndrome. Cochrane Database Syst Rev [Internet] 2023 [cited 2026 Jan 29];2023(11). Available from: http://doi.wiley.com/10.1002/14651858.CD015089.pub2

50. Rakshasbhuvankar AA, Nagarajan L, Zhelev Z, Rao SC. Amplitude-integrated electroencephalography compared with conventional video-electroencephalography for detection of neonatal seizures. Cochrane Database Syst Rev [Internet] 2025 [cited 2026 Jan 29];2025(8). Available from: http://doi.wiley.com/10.1002/14651858.CD013546.pub2

51. Natale P, Palmer SC, Navaneethan SD, Craig JC, Strippoli GF. Angiotensin-converting-enzyme inhibitors and angiotensin receptor blockers for preventing the progression of diabetic kidney disease. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(5). Available from: http://doi.wiley.com/10.1002/14651858.CD006257.pub2

52. Doleman B, Fonnes S, Lund JN, Boyd-Carson H, Javanmard-Emamghissi H, Moug S, et al. Appendectomy versus antibiotic treatment for acute appendicitis. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(4). Available from: http://doi.wiley.com/10.1002/14651858.CD015038.pub2

53. Vaidakis D, Papapanou M, Siristatidis CS. Autologous platelet-rich plasma for assisted reproduction. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(4). Available from: http://doi.wiley.com/10.1002/14651858.CD013875.pub2

54. Minozzi S, Ambrosi L, Saulle R, Uhm SS, Terplan M, Sinclair JM, et al. Psychosocial and medication interventions to stop or reduce alcohol consumption during pregnancy. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(4). Available from: http://doi.wiley.com/10.1002/14651858.CD015042.pub2

55. Sandall J, Fernandez Turienzo C, Devane D, Soltani H, Gillespie P, Gates S, et al. Midwife continuity of care models versus other models of care for childbearing women. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(5). Available from: http://doi.wiley.com/10.1002/14651858.CD004667.pub6

56. Owens D, Watkinson S, Harrison JE, Turner S, Worthington HV. Orthodontic treatment for prominent lower front teeth (Class III malocclusion) in children. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(4). Available from: http://doi.wiley.com/10.1002/14651858.CD003451.pub3

57. Fox T, Sguassero Y, Chaplin M, Rose W, Doum D, Arevalo-Rodriguez I, et al. *Wolbachia* -carrying *Aedes* mosquitoes for preventing dengue infection. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(4). Available from: http://doi.wiley.com/10.1002/14651858.CD015636.pub2

58. Constantin AM, Noertjojo K, Sommer I, Pizarro AB, Persad E, Durao S, et al. Workplace interventions to reduce the risk of SARS-CoV-2 infection outside of healthcare settings. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(4). Available from: http://doi.wiley.com/10.1002/14651858.CD015112.pub3

59. Wells GA, Hsieh SC, Peterson J, Zheng C, Kelly SE, Shea B, et al. Etidronate for the primary and secondary prevention of osteoporotic fractures in postmenopausal women. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(4). Available from: http://doi.wiley.com/10.1002/14651858.CD003376.pub4

60. Ng G, Bruschettini M, Ibrahim J, Da Silva O. Inhaled bronchodilators for the prevention and treatment of chronic lung disease in preterm infants. Cochrane Database Syst Rev

[Internet] 2024 [cited 2026 Feb 18];2024(4). Available from: http://doi.wiley.com/10.1002/14651858.CD003214.pub4

61. Cheetham MS, Ethier I, Krishnasamy R, Cho Y, Palmer SC, Johnson DW, et al. Home versus in-centre haemodialysis for people with kidney failure. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(5). Available from: http://doi.wiley.com/10.1002/14651858.CD009535.pub3

62. Ernst M, Folkerts AK, Gollan R, Lieker E, Caro-Valenzuela J, Adams A, et al. Physical exercise for people with Parkinson's disease: a systematic review and network meta-analysis. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(4). Available from: http://doi.wiley.com/10.1002/14651858.CD013856.pub3

63. Sowerbutts AM, Burden S, Sremanakova J, French C, Knight SR, Harrison EM. Preoperative nutrition therapy in people undergoing gastrointestinal surgery. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(4). Available from: http://doi.wiley.com/10.1002/14651858.CD008879.pub3

64. Ski CF, Taylor RS, McGuigan K, Long L, Lambert JD, Richards SH, et al. Psychological interventions for depression and anxiety in patients with coronary heart disease, heart failure or atrial fibrillation. Cochrane Database Syst Rev [Internet] 2024 [cited 2026 Feb 18];2024(5). Available from: http://doi.wiley.com/10.1002/14651858.CD013508.pub3

65. Priem J, Piwowar H, Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts [Internet]. 2022 [cited 2026 Jan 30];Available from: http://arxiv.org/abs/2205.01833

66. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. 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. Res Synth Methods 2024;15(4):616–26.

67. Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. Syst Rev 2024;13(1):158.

68. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev 2024;13(1):219.

69. Konet A, Thomas I, Gartlehner G, Kahwati L, Hilscher R, Kugley S, et al. Performance of two large language models for data extraction in evidence synthesis. Res Synth Methods 2024;15(5):818–24.

70. Gartlehner G, Kugley S, Crotty K, Viswanathan M, Dobrescu A, Nussbaumer-Streit B, et al. Artificial Intelligence–Assisted Data Extraction With a Large Language Model: A Study Within Reviews. Ann Intern Med 2025;178(12):1763–71.

71. Delgado-Chaves FM, Jennings MJ, Atalaia A, Wolff J, Horvath R, Mamdouh ZM, et al. Transforming literature screening: The emerging role of large language models in systematic reviews. Proc Natl Acad Sci 2025;122(2):e2411962122.

72. Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena [Internet]. 2023 [cited 2025 Jun 8];Available from: https://arxiv.org/abs/2306.05685

73. Arora R, Wei J, Hicks R, Bowman P, Quinonero-Candela J, Tsimpourlas T, et al. HealthBench: Evaluating Large Language Models Towards Improved Human Health [Internet]. 2025 [cited 2025 Jun 8];Available from: https://openai.com/index/healthbench/

74. Rethlefsen ML, Brigham TJ, Price C, Moher D, Bouter LM, Kirkham JJ, et al. Systematic review search strategies are poorly reported and not reproducible: a cross-sectional metaresearch study. J Clin Epidemiol 2024;166:111229.

75. Ioannidis JPA. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta□analyses. Milbank Q 2016;94(3):485–514.

76. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol 2017;17(1):152.

77. Bertizzolo L, Bossuyt P, Atal I, Ravaud P, Dechartres A. Disagreements in risk of bias assessment for randomised controlled trials included in more than one Cochrane systematic reviews: a research on research study using cross-sectional design. BMJ Open 2019;9(4):e028382.

78. Shah K, Egan G, Huan L (Nichoe), Kirkham J, Reid E, Tejani AM. Outcome reporting bias in Cochrane systematic reviews: a cross-sectional analysis. BMJ Open 2020;10(3):e032497.

79. Nguyen PY, McKenzie JE, Alqaidoom Z, Hamilton DG, Moher D, Page MJ. Reproducibility of meta-analytic results in systematic reviews of interventions: meta-research study. BMJ Med 2025;4(1):e002024.

80. Hamilton DG, McKenzie JE, Nguyen PY, Rethlefsen ML, McDonald S, Brennan SE, et al. Evaluation of the replicability of systematic reviews with meta-analyses of the effects of health interventions. Res Synth Methods 2026;1–19.

81. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;n71.

Acknowledgements

We thank Guravneet Gill for assistance with database access; Zion Chan, Li Li, Lina Ghosn, and Isabelle Boutron for discussions during project development; and Luke Son for technical insights. We acknowledge all systematic review authors whose published data and methodologies enabled this reproducibility assessment.

Author Information

These authors contributed equally: C.C, R.A., P.C. Author contribution statements: C.C, R.A, N.B contributed to the conception and design of the work. C.C, R.A, K.M, M.C, E.F, A.S, R.K, R.S, D.M, J.L, J.J, D.C, J.G, S.L contributed to data acquisition, cleaning, human comparisons, and human arbitration. C.C, R.A, P.C, J.S, S.J, J.X, K.Z generated code for evaluations and benchmarking. C.C, R.A, P.C, L.X.G analyzed study data. N.B, G.M.C, D.M, A.T, D.B.E, R.K.A, L.H, J.P, A.S.D, M.N, A.A.L, B.T, M.W, H.W contributed to project supervision and provided feedback on the study. C.C, R.A, P.C, prepared the original draft of the manuscript with input from all co-authors. All authors were responsible for review and editing of the manuscript. All authors debated, discussed, edited, and approved the final version of the manuscript.

Funding and Ethics Declaration

There was no direct funding support for this manuscript.

N.B reports grants from the Public Health Agency of Canada through Canada's COVID-19 Immunity Task Force, the World Health Organization Health Emergencies Programme, the Robert Koch Institute, the Canadian Medical Association Joule Innovation Fund, the Canadian Association of Emergency Physicians and Alberta Health Services Emergency Strategic Clinical Network. Disclosures for G.M.C. can be found at http://arep.med.harvard.edu/gmc/tech.html. R.K.A. is employed at OpenAI and owns stock as part of the standard compensation package. R.A. reports grants from the CIHR Institute of Genetics.

C.C. P.C. J.S. are founders of and hold equity in Otto Science Institute. N.B, D.M, A.C.T, A.B hold equity in Otto Science Institute. Of note, Otto Science institute was founded after the time of initial manuscript writing.

No funding source had any role in the design of this study, its execution, analyses, interpretation of the data, or decision to submit results.

Code and Dataset Availability

Datasets will be made available upon reasonable request.

Code used for data analysis and figure generation will be made available on publication. The otto-SR platform source code and model weights will not be open-sourced due to intellectual property, commercial viability, and safety considerations associated with unmonitored use. To support reproducibility, researchers can reproduce all reported results by running the same versioned otto-SR workflow via controlled access at ottosr.com (contact research@ottosr.com).

Tables

Table 1 - Phase 1: Descriptive overview of datasets used in article screening

Centre for Evidence-Based Medicine Question Type	Dataset	Clinical Domain (Web of Science)	Population	Intervention/ Exposure	Study Types Included	Citations Identified from Electronic Search after deduplication (Total N)	Citations Included After Full-Text Screening ('Included' N)
Prevalence *How common is the problem?*	SeroTracker *Test Sample	Infectious Diseases	Humans of any age	Prevalence of SARS-CoV-2 antibodies	Cross-sectional, repeated cross sectional, and cohort study designs	10000	210
Diagnostic Test Accuracy *Is this diagnostic or monitoring test accurate?*	PA-Testing	Endocrinology & Metabolism	Patients with primary aldosteronism	Confirmatory tests used to diagnose primary aldosteronism	Case-control, diagnostic testing	8000	52
Prognosis *What will happen if we do not add a therapy?*	SVCF	Pediatrics Cardiovascular System & Cardiology	Preterm infants <32 weeks gestational age	Low SVC flow identified by Doppler assessment in the first 48 hours after birth	Randomized controlled trials, cohort or case-control studies	2257	17
Intervention Benefits *Does this intervention help?*	Reinfection	Infectious Diseases	Humans of any age, in any geographical setting.	Individuals with previously confirmed SARS-CoV-2 infection that have documented vaccination status	Test-negative case-control, traditional case-control, cross-sectional, cohort, non-randomized controlled trials, and randomized controlled trials.	6724	181

	PA-Outcomes	Endocrinology & Metabolism Surgery	Patients with primary aldosteronism	Surgical adrenalectomy and medical treatment with a mineralocorticoid receptor antagonist	Randomized controlled trials, observational studies	5376	16
Intervention Harms *What are the COMMON/RARE harms?*	-	-	-	-	-	-	-
Screening *Is this (early detection) test worthwhile?*	-	-	-	-	-	-	-

Table 2 - Screening calibration: Performance of human reviewers relative to original study authors

Screening Stage	Group	True positives, n	True negatives, n	False positives, n	False negatives, n	Sensitivity	Specificity
Dual Abstract	SeroTracker Human	77	297	12	4	84.6%	96.1%
	Team 1*	75	299	10	16	82.4%	96.8%
	Team 2*	78	296	13	13	85.7%	95.8%
Dual full-text	SeroTracker Human	74	299	10	17	81.3%	96.8%
	Team 1*	72	310	8	19	79.1%	97.4%
	Team 2*	74	303	6	17	81.3%	98.1%

*Team 1 and Team 2 represent our panel of human reviewers, who screened records independently and in duplicate. Conflicts were arbitrated by a third reviewer

Table 3 - Phase 1: Abstract screening performance of *otto-SR*, dual human reviewers, Elicit, NestedKnowledge, and DistillerSR

Dataset	Records, total n (Include/ Exclude, n)	Model	True positives, n	True negatives, n	False positives, n	False negatives, n	Sensitivity (95% CI)	Specificity (95% CI)
SeroTracker	10000 (210/9790)	otto-SR	204	9567	223	6	97.1 (93.9-98.7)	97.7 (97.4-98.0)
	400 (91/309)	Dual human	77	297	12	14	84.6 (75.8-90.6)	96.1 (93.3-97.8)
	390 (91/299)	Elicit	75	275	24	16	82.4 (73.3-88.9)	92.0 (88.3-94.5)
	400 (91/309)	Nested	87	232	77	4	95.6 (89.2-98.3)	75.1 (70.0-79.6)
	400 (91/309)	Distiller	64	283	26	27	70.3 (60.3-78.7)	91.6 (88.0-94.2)
Reinfection	6180 (181/5999)	otto-SR	174	5016	983	7	96.1 (92.2-98.1)	83.6 (82.7-84.5)
	400 (145/255)	Dual human	122	236	19	23	84.1 (77.3-89.2)	92.5 (88.7-95.2)

	398 (145/253)	Elicit	133	153	100	12	91.7 (86.1-95.2)	60.5 (54.3-66.3)
	400 (145/255)	Nested	25	248	7	120	17.2 (12.0-24.2)	97.3 (94.4-98.7)
	400 (145/255)	Distiller	73	249	6	72	50.3 (42.3-58.4)	97.6 (95.0-98.9)
PA-Testing	7757 (52/7705)	otto-SR	51	7403	302	1	98.1 (89.9-99.7)	96.1 (95.6-96.5)
	400 (45/355)	Dual human	44	347	8	1	97.8 (88.4-99.6)	97.7 (95.6-98.9)
	393 (45/348)	Elicit	39	323	25	6	86.7 (73.8-93.7)	92.8 (89.6-95.1)
	400 (45/355)	Nested	34	337	18	11	75.6 (61.3-85.8)	94.9 (92.1-96.8)
	400 (45/355)	Distiller	41	342	13	4	91.1 (79.3-96.5)	96.3 (93.8-97.8)
PA-Outcomes	4309 (16/4293)	otto-SR	16	4054	239	0	100.0 (80.6-100.0)	94.4 (93.7-95.1)

	400 (16/384)	Dual human	16	363	21	0	100.0 (80.6-100.0)	94.5 (91.8-96.4)
	392 (16/376)	Elicit	16	255	121	0	100.0 (80.6-100.0)	67.8 (62.9-72.3)
	400 (16/384)	Nested	15	310	74	1	93.8 (71.7-98.9)	80.7 (76.5-84.4)
	400 (16/384)	Distiller	15	360	24	1	93.8 (71.7-98.9)	93.8 (90.9-95.8)
SVCF	1954 (17/1937)	otto-SR	16	1878	59	1	94.1 (73.0-99.0)	97.0 (96.1-97.6)
	167 (17/150)	Dual human	15	148	2	2	88.2 (65.7-96.7)	98.7 (95.3-99.6)
	165 (17/148)	Elicit	15	138	10	2	88.2 (65.7-96.7)	93.2 (88.0-96.3)
	166 (17/149)	Nested	14	146	3	3	82.4 (59.0-93.8)	98.0 (94.2-99.3)
	166 (17/149)	Distiller	10	149	0	7	58.8 (36.0-78.4)	100.0 (97.5-100.0)

Table 4 - Phase 1: Full-text screening performance of *otto-SR* and dual human reviewers

Dataset	Records, total n (Include/ Exclude, n)	Model	True positives, n	True negatives, n	False positives, n	False negatives, n	Sensitivity (95% CI)	Specificity (95% CI)
Serotracker	10000 (210/9790)	otto-SR	204	9663	127	6	97.1 (93.9-98.7)	98.7 (98.5-98.9)
	400 (91/309)	Dual human	74	299	10	17	81.3 (72.1-88.0)	96.8 (94.1-98.2)
	390 (91/299)	Elicit	72	284	15	19	79.1 (69.7-86.2)	95.0 (91.9-96.9)
	400 (91/309)	Nested	85	266	43	6	93.4 (86.4-96.9)	86.1 (81.8-89.5)
Reinfection	6180 (181/5999)	otto-SR	174	5439	560	7	96.1 (92.2-98.1)	90.7 (89.9-91.4)
	400 (145/255)	Dual human	64	243	12	81	44.1 (36.3-52.3)	95.3 (92.0-97.3)
	398 (145/253)	Elicit	132	180	73	13	91.0 (85.3-94.7)	71.1 (65.3-76.4)

	400 (145/255)	Nested	21	248	7	124	14.5 (9.7-21.1)	97.3 (94.4-98.7)
PA-Testing	7757 (52/7705)	otto-SR	48	7608	97	4	92.3 (81.8-97.0)	98.7 (98.5-99.0)
	400 (45/355)	Dual human	36	351	4	9	80.0 (66.2-89.1)	98.9 (97.1-99.6)
	393 (45/348)	Elicit	38	327	21	7	84.4 (71.2-92.3)	94.0 (91.0-96.0)
	400 (45/355)	Nested	33	338	17	12	73.3 (59.0-84.0)	95.2 (92.5-97.0)
PA-Outcomes	4309 (16/4293)	otto-SR	16	4179	114	0	100.0 (80.6-100.0)	97.3 (96.8-97.8)
	400 (16/384)	Dual human	15	376	8	1	93.8 (71.7-98.9)	97.9 (95.9-98.9)
	392 (16/376)	Elicit	16	269	107	0	100.0 (80.6-100.0)	71.5 (66.8-75.9)
	400 (16/384)	Nested	15	317	67	1	93.8 (71.7-98.9)	82.6 (78.4-86.0)

SVCF	1954 (17/1937)	otto-SR	16	1902	35	1	94.1 (73.0-99.0)	98.2 (97.5-98.7)
	167 (17/150)	Dual human	13	150	0	4	76.5 (52.7-90.4)	100.0 (97.5-100.0)
	165 (17/148)	Elicit	14	140	8	3	82.4 (59.0-93.8)	94.6 (89.7-97.2)
	166 (17/149)	Nested	13	146	3	4	76.5 (52.7-90.4)	98.0 (94.2-99.3)

Screening decisions following abstract screening then underwent full-text screening.

Table 5 - Phase 2: Descriptive overview of datasets used in data extraction benchmark

Dataset	Clinical Domain (Web of Science)	Population	Intervention/ Exposure	Study Types Included	Total articles [Sample N]*	Total data-points [Sample N]*
SeroTracker *Test sample	Infectious Diseases	Humans of any age	Prevalence of SARS-CoV-2 antibodies	Cross-sectional, repeated cross sectional, and cohort study designs	100 [30]	700 [210]
PA-Testing	Endocrinology & Metabolism	Patients with primary aldosteronism	Confirmatory tests used to diagnose primary aldosteronism	Case-control, diagnostic testing	51 [12]	867 [204]
PA-Outcomes	Endocrinology & Metabolism Surgery	Patients with primary aldosteronism	Surgical adrenalectomy and medical treatment with a mineralocorticoid receptor antagonist	Randomized controlled trials, observational studies	16 [10]	336 [210]
Sepsis	General & Internal Medicine	Individuals aged ≥16 years with septic shock (sepsis and use of at least one vasopressor)	Treatment with hydrocortisone alone, hydrocortisone-fludrocortisone, placebo or usual care.	Randomized controlled trials	16 [13]	256 [208]
CKD	Urology & Nephrology	Empirical literature studying nature or burden of CKD progression	N/A	Retrospective cohort studies, case-control studies, cross-sectional studies	80 [21]	800 [210]

Process	Health Care Sciences & Services	Empirical literature studying process mapping in healthcare	N/A	Methodological studies	104 [17]	1248 [204]
Psyc-meds	Pharmacology & Pharmacy Psychiatry	Adults with a diagnosis of generalized anxiety disorder	Antidepressant monotherapy	Randomized controlled trials	28 [23]	252 [207]

*Sample N refers to the number of studies that underwent data extraction by dual human reviewers and DistillerSR. The sample size was based on a McNemar test sample size approximation.

Table 6 - Phase 2: Accuracy of otto-SR, dual human reviewers, and Elicit in data extraction

Dataset	Model	Corrected Numerator	Corrected Denominator	Accuracy
SeroTracker	Human	156	210	74.3
	Elicit	509	693	73.4
	Otto	636	693	91.8
	Nested	549	693	79.2
	Distiller	80	210	38.1
PA-Outcomes	Human	158	205	77.1
	Elicit	253	328	77.1
	Otto	318	328	97.0
	Nested	231	328	70.4
	Distiller	115	205	56.1
PA-Testing	Human	141	204	69.1
	Elicit	526	867	60.7

	Otto	807	867	93.1
	Nested	481	867	55.5
	Distiller	61	204	29.9
Sepsis	Human	174	191	91.1
	Elicit	168	238	70.6
	Otto	225	238	94.5
	Nested	126	238	52.9
	Distiller	90	191	47.1
Process	Human	172	204	84.3
	Elicit	1071	1248	85.8
	Otto	1170	1248	93.8
	Nested	994	1248	79.6
	Distiller	129	204	63.2
CKD	Human	168	210	80.0

	Elicit	675	800	84.4
	Otto	736	800	92.0
	Nested	670	800	83.8
	Distiller	117	210	55.7
Psyc-meds	Human	171	206	83.0
	Elicit	160	248	64.5
	Otto	226	248	91.1
	Nested	97	248	39.1
	Distiller	67	206	32.5

Table 7 - Phase 3: Descriptive overview and risk of bias performance of otto-SR compared to corrected ground truth

Dataset	RoB tool	Comparison	Corrected Numerator	Corrected Denominator	Cohen's Kappa	Gwet AC2
RSV Vaccines	ROB2	otto-SR - corrected	315	325	0.89	0.99
		Original reviewer - corrected	307	325	0.76	0.98
Liver Transplant	ROB2	otto-SR - corrected	81	90	0.39	0.95
		Original reviewer - corrected	86	90	0.77	0.99
Misoprosotol IUD	ROB2	otto-SR - corrected	563	575	0.86	0.99
		Original reviewer - corrected	538	575	0.25	0.98
Oocyte activation	ROB2	otto-SR - corrected	65	70	0.79	0.97
		Original reviewer - corrected	61	70	0.56	0.94
COVID Diabetes	Newcastle	otto-SR - corrected	88	90	0.74	0.98
		Original reviewer - corrected	85	90	-0.03	0.94
Sedentary CVD	Newcastle	otto-SR - corrected	69	72	0.65	0.95
		Original reviewer - corrected	65	72	0.18	0.89
FUBC Mortality	Newcastle	otto-SR - corrected	120	126	0.84	0.93

		Original reviewer - corrected	123	126	0.92	0.97
Smoking surgery complications	Newcastle	otto-SR - corrected	355	378	0.65	0.93
		Original reviewer - corrected	336	378	0.47	0.86
Amplitude-integrated EEG	QUADAS2	otto-SR - corrected	79	105	0.55	0.66
		Original reviewer - corrected	79	105	0.48	0.73
HINTS	QUADAS2	otto-SR - corrected	59	77	0.38	0.78
		Original reviewer - corrected	68	77	0.52	0.91
Truenat MTB assay	QUADAS2	otto-SR - corrected	40	56	0.51	0.64
		Original reviewer - corrected	30	56	0.28	0.7
Transcutaneous bilirubinometry	QUADAS2	otto-SR - corrected	118	140	0.57	0.87
		Original reviewer - corrected	119	140	0.14	0.87

Table 8 - Phase 4: Descriptive overview of reviews included in the Cochrane April 2024 issue.

Dataset	Population	Intervention	Study Types Included	Primary Outcome	Methodological issues in search	Methodological issues in data extraction	Date of Electronic search
ACEI[51]	Adults with diabetes and kidney disease	ACEi alone ARB alone ACEi + ARB	Randomized controlled trials (RCTs)	All-cause mortality	Authors report using the Cochrane Kidney and Transplant Register of Studies, although the search used for this register was not provided. The search strategy provided appeared to be used to derive the Kidney and Transplant Register (31,444 citations)	Authors extracted '0' mortality events, even in the absence of any complications or survival reporting As the absence of reporting does not suggest non-events, we opted to exclude studies that did not report on any complications or survival/mortality. Did not clarify the timing for all-cause mortality. Assumed latest time point.	17 March 2024 - search
Appe[52]	Adults with suspected acute uncomplicated or simple appendicitis	Antibiotic treatment Appendectomy (surgery)	Parallel-group RCTs	All-cause mortality (latest time point)	1. Clinicaltrials.gov searched using only "appendicitis" as formal strategy was not provided 2. ICTRP was not searched as search terms were not provided.		September 6 2021 - protocol July 19 2022
Plasma[53]	Women and couples undergoing IVF or ICSCI	Intrauterine infusion or injection of platelet-rich plasma	RCTs	Number of live births or ongoing pregnancy	1. Cochrane Gynaecology and Fertility specialised register was searched through CENTRAL not ProCite		January 9 2023 - search

Alcohol[54]	Individuals who consume alcohol during pregnancy	Any medication or psychosocial intervention to reduce alcohol consumption or achieve abstinence	Parallel, individually randomised RCTs, cluster□RCTs, quasi□randomised studies	Number abstinent from alcohol after treatment	1. Cochrane Drugs and Alcohol Group Specialised Register was searched through CENTRAL not CRSlive. Search did not include XDI term as not a valid parameter in CENTRAL.		January 8 2024 - search
Midwife[55]	Excluded - search provided was for generating the Cochrane group specialized registrar (~150k citations), but did not provide search terms used to query this register. Given the size of the registrar and no specified query terms, this study was excluded.						
Teeth[56]	Children and adolescents receiving orthodontic treatment to correct class III malocclusion	Non-surgical orthodontic interventions Surgically-anchored orthodontic interventions	RCTs	Overjet (prominence of lower front teeth) (follow-up within 9-15 months)	1. Cochrane Oral Health Trials Register was searched through CENTRAL not ProCite		July 16 2023 search
Dengue[57]	Adults and children living in Dengue infection prevalent areas	Wolbachia carrying Aedes species (mosquito) deployment	RCTs, cluster-RCTs	Number of virologically confirmed dengue infections	1. WOS search of core collection was performed independently of CABI 2. Search of LILACs was done in the LILACs Plus collection		January 24 2024 Search
Workplace [58]	Adults in (non-healthcare) workplace or occupational settings	Any intervention that attempted reduce exposure to SARS-CoV-2	RCTs, non□randomised studies of interventions	Incidence of symptomatic SARS-CoV-2 infection events	Author search was not reproducible (confirmed with original authors). Original authors kindly provided the		April 2023 - search

					raw RIS files from the initial search. Consequently, no update was performed.		
Etidronate [59]	Postmenopausal women with osteoporosis (including those with higher risk of osteoporotic fracture)	Etidronate 400mg Etidronate 200mg	RCTs	Number of hip fractures (latest time point)	1. Only conducted 2012 and 2023 search 2. Embase was searched using code oemezd and EBM Reviews - Cochrane Central Register of Controlled Trials was searched using code cctz 3. Omitted line 144 of OVID search: “remove duplicates from 143” As all deduplication was done through the *otto-SR* pipeline		February 1 2023 - search
Lung[60]	Preterm infants at risk for chronic lung disease (CLD) or identified as having CLD	Inhaled bronchodilators via any modularity of inhaled administration	RCTs, quasi-randomized controlled trials	Mortality within trial period (total)	1. Medline EBSCO host, did not use search term: Publication date limitation: 2016 to 2023 as nonspecific to date of their search. 2. ICTRP search result returned fewer results when		May 12 2023 - search

					exactly copied.		
Dialysis[61]	Adults with kidney failure	In-centre hemodialysis (ICHD) Home hemodialysis (HD)	RCTs, quasi-randomized controlled trials, non□randomised studies of interventions	Cardiovascular death (total)	1. Missing Cochrane Kidney and Transplant Register of Studies (study authors did not provide search terms)		October 9 2022 search
Parkinson's[62]	Excluded - no publicly available downloadable data at time of manuscript writing (March 2025)						
Nutrition[63]	People undergoing non-emergency gastrointestinal tract-related surgery	Any pre-operative nutritional therapy intervention	RCTs	Length of hospital stay	1. British Nursing Index Archive was unable to be searched 2. AMED: The subject heading 'Enteral nutrition' is invalid in this database. Was replaced by 'enteral feeding' The subject heading 'Fatty Acids, Omega-3' is invalid in this database and was removed from search. The AMED search, after using these substitutions, returned zero search results.		March 2023 search
Depression [64]	Adults with heart disease	Any psychological intervention for	RCTs	Depression symptom score	1. Ovid medline version not specified, used	The authors report "we measured depression and anxiety as change in	July 5/7 2022 search

		depression		(latest time point)	OVID MEDLINE(R) ALL 1946 to May 07 2025, did not use ovid term: “limit 80 to yr="2009□2022"” as nonspecific to date of their search. 2. Embase: did not use term: “limit 87 to yr="2009□2022"” as nonspecific to date of their search. 3. Pyschinfo: did not use term: “limit 74 to yr="2009□2022"” as nonspecific to date of their search. exp "Fibrillation (Heart)"/ was invalid, was substituted with exp "Heart Fibrillation"/ 4. CINAHL COMPLETE: did not use published date filter in search.	symptoms (mean score).” However, authors extracted the symptom score at the latest time point (did not assess change in score as reported). We opted to match this logic, instructing the LLM to extract the symptom score at the latest time point.	

Table 9 - Phase 4: Descriptive Overview of Studies included and excluded in the Cochrane reproducibility analysis.

Dataset	Citations	Cochrane included*	*otto-SR* included	*otto-SR* excluded	*otto-SR* included	*otto-SR* included

	Identified from Electronic Search Search up to May 2025	N study	'Matched' N study	'matched' *Cochrane studies incorrectly excluded by otto-SR* N study	'Expanded' *Total articles included with Cochrane search* N study (N false positive)	'Update' *Total articles included in updated search* N study (N false positive)
ACEI	31,444	ACEi + ARB: 0 ACEi alone: 8 ARB alone: 7 **5 studies were not recovered in electronic search* **15 studies were excluded due to lack of explicit mortality reporting* **1 study (Krairittichai 2009) was incorrectly included in the ARB alone vs. placebo analysis. interventions were ACEi + ARB vs. ACEi alone*	ACEi + ARB: 0 ACEi alone: 7 ARB alone: 6	ACEi alone: 1 ARB alone: 1 **Muirhead 1999: Otto incorrectly excluded (contributed estimates to both ACEi alone and ARB alone comparisons).*	ACEi + ARB: 1 ACEi alone: 13 (1 false positive) ARB alone: 7	ACEi + ARB: 1 ACEi alone: 13 (1 false positive) ARB alone: 7
Appe	3,045	5	5	0	6	6
Plasma	572	6	6	0	9	14
Alcohol	7,194	3	3	0	12 (3 false positive)	12 (3 false positive)

Midwife	Excluded - search not reproducible					
Teeth	4,813	Non-surgical: 5 Surgical: 2	Non-surgical: 5 Surgical: 2	0 0	Non-surgical: 5 Surgical: 3	Non-surgical: 5 Surgical: 4
Dengue	2,155	1	1	0	1	1
Workplace	24,641	1	1	0	3	3
Etidronate	2,832	400mg: 2 200mg: 0 **2 studies (200mg) were incorrectly included:* *Ishida 2004: retracted May 2025* *Iwamato 2001: Cochrane authors incorrectly included. Placebo group violated the protocol requirement for an active-control design*	400mg: 2 200mg: 0	0 0	400mg: 4 (1 false positive) 200mg: 0	400mg: 4 (1 false positive) 200mg: 0
Lung	5,751	1	1	0	1	1
Dialysis	9,088	2	1	1 *Marshall 2021: Otto incorrectly excluded. Cardiovascular mortality data was in supplementary data (inaccessible to LLM)*	4 (1 false positive)	5 (1 false positive)
Parkinson's	Excluded - no downloadable data					

Nutrition	4,937	Immune: 4 Enteral: 1 **7 studies were not recovered after reproducing* electronic search **1 study excluded due to lack of mean length of hospital stay reporting*	Immune: 4 Enteral: 1	0 0	Immune: 9 Enteral: 2 (1 false positive) Oral or enteral: 1 (new comparison)	Immune: 10 Enteral: 2 (1 false positive) Oral or enteral: 1 (new comparison)
Depression	49,804	16 **1 study was not recovered in electronic search* **4 studies were excluded due to lack of mean depression score reporting*	15	1 *Humphries 2021: Otto incorrectly excluded. Depression symptom score was in supplementary data (inaccessible to LLM)*	40 (3 false positive)	49 (5 false positive)

Table 10 - Phase 4: Meta-analyzed results between original cochrane authors, *otto-SR*, and corrected ground-truth across Cochrane reviews

Dataset	Analysis Comparison	Model	Effect Estimate (95% CI)	Type	I2	Intervention events	Intervention Denominator	Control Events	Control Denominator	Num Studies
Matched										
ACEi	ACEi alone vs. control	Cochrane	0.97 (0.58 to 1.63)	RR	10.66	346	2823	343	2867	8
		otto-SR	0.97 (0.58 to 1.63)	RR	10.66	346	2808	343	2845	7
		Corrected	0.97 (0.58 to 1.63)	RR	10.66	346	2837	343	2876	8
	ARB alone vs. control	Cochrane	0.99 (0.87 to 1.13)	RR	0	271	2128	272	1913	7
		otto-SR	0.99 (0.89 to 1.12)	RR	0	268	2066	269	1882	6
		Corrected	0.99 (0.87 to 1.13)	RR	0	271	2128	272	1913	7
Appe	Antibiotics vs. appendectomy	Cochrane	0.5 (0.04 to 5.53)	OR	0	1	1189	2	1188	5
		otto-SR	0.53 (0.05 to 5.87)	OR	0	1	1202	2	1217	5
		Corrected	0.53 (0.05 to 5.87)	OR	0	1	1202	2	1217	5
Plasma	Intrauterine PRP vs. control	Cochrane	2.38 (1.16 to 4.89)	OR	54.39	83	283	40	281	6
		otto-SR	2.7 (1.38 to 5.26)	OR	47.93	90	283	41	281	6

		Corrected	2.7 (1.38 to 5.26)	OR	47.93	90	283	41	281	6
Alcohol	Any psychosocial intervention vs. treatment as usual	Cochrane	1.34 (1.11 to 1.61)	RR	0	159	220	81	158	3
		otto-SR	1.35 (1.19 to 1.53)	RR	0	165	228	85	165	3
		Corrected	1.31 (1.12 to 1.52)	RR	0	165	325	85	214	3
Teeth	Non-surgical treatment vs. untreated control	Cochrane	5.03 (3.89 to 6.17)	MD	86.65		99		85	5
		otto-SR	5.03 (3.89 to 6.17)	MD	86.65		99		85	5
		Corrected	5.03 (3.89 to 6.17)	MD	86.65		99		85	5
	Surgical treatment vs. untreated control	Cochrane	7.69 (6.99 to 8.4)	MD	0		20		10	2
		otto-SR	7.69 (6.99 to 8.4)	MD	0		20		10	2
		Corrected	7.69 (6.99 to 8.4)	MD	0		20		10	2
Dengue	Wolbachia intervention vs. existing intervention	Cochrane	0.23 (0.18 to 0.3)	OR	0	67	2905	318	3401	1
		otto-SR	0.23 (0.18 to 0.3)	OR	0	67	2905	318	3401	1
		Corrected	0.23 (0.18 to 0.3)	OR	0	67	2905	318	3401	1
Workplace	Any workplace intervention vs. existing	Cochrane	1.28 (0.86 to 1.9)	Rate Ratio	0					1
		otto-SR	1.28 (0.86 to	Rate Ratio	0					1

	intervention		1.9)							
		Corrected	1.28 (0.86 to 1.9)	Rate Ratio	0					1
Etidronate	Etidronate 400mg vs. placebo	Cochrane	1.4 (0.22 to 8.92)	RR	(Fixed effects)	4	262	2	263	2
		otto-SR	1.4 (0.22 to 8.92)	RR	(Fixed effects)	4	262	2	263	2
		Corrected	0.87 (0.13 to 5.82)	RR	(Fixed effects)	2	144	2	139	2
Lung	Inhaled bronchodilator vs. placebo	Cochrane	1.08 (0.5 to 2.31)	RR	0	12	87	11	86	1
		otto-SR	1.08 (0.5 to 2.31)	RR	0	12	87	11	86	1
		Corrected	1.08 (0.5 to 2.31)	RR	0	12	87	11	86	1
Dialysis	HHD vs. ICHD	Cochrane	0.92 (0.48 to 1.77)	RR	0	170	1294	4090	29606	2
		otto-SR	1.17 (0.59 to 2.3)	RR	0	14	58	12	58	1
		Corrected	0.92 (0.48 to 1.77)	RR	0	170	1294	4090	29606	2
Nutrition	Pre-operative enteral nutrition vs. control	Cochrane	2.22 (-3.13 to 7.57)	MD	0		13		13	1
		otto-SR	2.22 (-3.68 to 8.12)	MD	0		11		9	1
		Corrected	2.22 (-3.68 to 8.12)	MD	0		11		9	1
	Pre-operative	Cochrane	-0.19 (-1.44 to 1.07)	MD	24.77		192		192	4

	immune enhancing nutrition vs. control	otto-SR	-0.21 (-1.48 to 1.05)	MD	24.64		183		182	4
		Corrected	-0.22 (-1.48 to 1.05)	MD	24.56		180		182	4
Depression	Any psychosocial intervention vs. control	Cochrane	-0.43 (-0.96 to 0.09)	SMD	96.549		907		968	16
		otto-SR	-0.36 (-0.9 to 0.18)	SMD	96.4712		837		876	15
		Corrected	-0.37 (-0.86 to 0.13)	SMD	96.2996		926		988	16
Expanded										
ACEi	ACEi alone vs. control	Cochrane	0.97 (0.58 to 1.63)	RR	10.66	346	2823	343	2867	8
		otto-SR	1.02 (0.88 to 1.18)	RR	0	368	3424	366	3419	13
		Corrected	0.99 (0.71 to 1.39)	RR	4.288	353	3212	347	3212	13
	ARB alone vs. control	Cochrane	0.99 (0.87 to 1.13)	RR	0	271	2128	272	1913	7
		otto-SR	0.99 (0.89 to 1.12)	RR	0	268	2086	269	1902	7
		Corrected	0.99 (0.87 to 1.13)	RR	0	271	2148	272	1933	8
Appe	Antibiotics vs. appendectomy	Cochrane	0.5 (0.04 to 5.53)	OR	0	1	1189	2	1188	5
		otto-SR	0.53 (0.05 to 5.87)	OR	0	1	1217	2	1231	6
		Corrected	0.53 (0.05 to 5.87)	OR	0	1	1218	2	1231	6

Plasma	Intrauterine PRP vs. control	Cochrane	2.38 (1.16 to 4.89)	OR	54.39	83	283	40	281	6
		otto-SR	3.45 (1.86 to 6.39)	OR	60.51	181	557	58	558	9
		Corrected	3.45 (1.86 to 6.39)	OR	60.51	181	557	58	558	9
Alcohol	Any psychosocial intervention vs. treatment as usual	Cochrane	1.34 (1.11 to 1.61)	RR	0	159	220	81	158	3
		otto-SR	1.14 (0.97 to 1.35)	RR	55.53	545	1023	415	940	12
		Corrected	1.14 (0.94 to 1.39)	RR	57.63	483	995	370	875	9
Teeth	Non-surgical treatment vs. untreated control	Cochrane	5.03 (3.89 to 6.17)	MD	86.65		99		85	5
		otto-SR	5.03 (3.89 to 6.17)	MD	86.65		99		85	5
		Corrected	5.03 (3.89 to 6.17)	MD	86.65		99		85	5
	Surgical treatment vs. untreated control	Cochrane	7.69 (6.99 to 8.4)	MD	0		20		10	2
		otto-SR	6.61 (4.5 to 8.72)	MD	92.72		36		26	3
		Corrected	6.61 (4.5 to 8.72)	MD	92.72		36		26	3
Dengue	Wolbachia intervention vs. existing intervention	Cochrane	0.23 (0.18 to 0.3)	OR	0	67	2905	318	3401	1
		otto-SR	0.23 (0.18 to 0.3)	OR	0	67	2905	318	3401	1
		Corrected	0.23 (0.18 to	OR	0	67	2905	318	3401	1

			0.3)							
Workplace	Any workplace intervention vs. existing intervention	Cochrane	1.28 (0.86 to 1.9)	Rate Ratio	0					1
		otto-SR	0.74 (0.33 to 1.64)	Rate Ratio	79.79					3
		Corrected	0.74 (0.33 to 1.64)	Rate Ratio	79.79					3
Etidronate	Etidronate 400mg vs. placebo	Cochrane	0.87 (0.13 to 5.82)	RR	(Fixed effects)	2	144	2	139	2
		otto-SR	0.95 (0.22 to 4.05)	RR	(Fixed effects)	5	335	4	336	4
		Corrected	0.95 (0.22 to 4.05)	RR	(Fixed effects)	5	295	4	296	3
Lung	Inhaled bronchodilator vs. placebo	Cochrane	1.08 (0.5 to 2.31)	RR	0	12	87	11	86	1
		otto-SR	1.08 (0.5 to 2.31)	RR	0	12	87	11	86	1
		Corrected	1.08 (0.5 to 2.31)	RR	0	12	87	11	86	1
Dialysis	HHD vs. ICHD	Cochrane	0.92 (0.48 to 1.77)	RR	0	170	1294	4090	29606	2
		otto-SR	0.72 (0.15 to 3.34)	RR	74.78	216	3322	3168	22209	4
		Corrected	0.93 (0.78 to 1.1)	RR	0	171	1357	4092	29778	3
Nutrition	Pre-operative enteral nutrition vs. control	Cochrane	2.22 (-3.13 to 7.57)	MD	0		13		13	1
		otto-SR	-0.78 (-2.38 to 0.83)	MD	11.02		44		42	2

		Corrected	2.22 (-3.68 to 8.12)	MD	0		11		9	1
	Pre-operative immune enhancing nutrition vs. control	Cochrane	-0.19 (-1.44 to 1.07)	MD	24.77		192		192	4
		otto-SR	-1.2 (-2.28 to -0.11)	MD	60.56		450		447	9
		Corrected	-1.2 (-2.28 to -0.12)	MD	60.43		447		447	9
	Pre-operative oral or enteral nutrition vs. control	otto-SR	-1.65 (-3.54 to 0.24)	MD	0		40		40	1
		Corrected	-1.65 (-3.54 to 0.24)	MD	0		40		40	1
Depression	Any psychosocial intervention vs. control	Cochrane	-0.43 (-0.96 to 0.09)	SMD	96.549		907		968	16
		otto-SR	-0.41 (-0.67 to -0.16)	SMD	94.978		2782		2668	40
		Corrected	-0.42 (-0.69 to -0.15)	SMD	94.9152		2455		2352	38
Update										
ACEi	ACEi alone vs. control	otto-SR	1.02 (0.88 to 1.18)	RR	0	368	3424	366	3419	13
		Corrected	0.99 (0.71 to 1.39)	RR	4.288	353	3212	347	3212	13
	ARB alone vs. control	otto-SR	0.99 (0.89 to 1.12)	RR	0	268	2086	269	1902	7
		Corrected	0.99 (0.87 to 1.13)	RR	0	271	2148	272	1933	8
Appe	Antibiotics	otto-SR	0.53 (0.05 to	OR	0	1	1217	2	1231	6

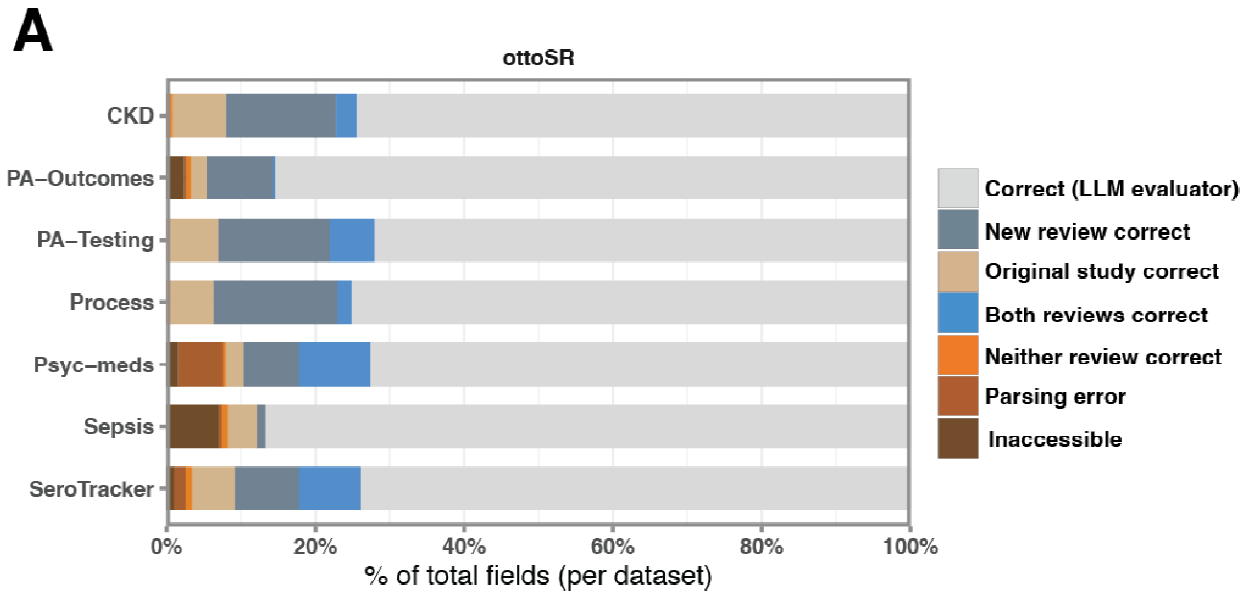
	vs. appendectomy		5.87)							
		Corrected	0.53 (0.05 to 5.87)	OR	0	1	1218	2	1231	6
Plasma	Intrauterine PRP vs. control	otto-SR	2.42 (1.34 to 4.35)	OR	75.33	324	823	187	809	14
		Corrected	2.41 (1.34 to 4.35)	OR	75.35	324	831	187	817	14
Alcohol	Any psychosocial intervention vs. treatment as usual	otto-SR	1.14 (0.97 to 1.35)	RR	55.53	545	1023	415	940	12
		Corrected	1.14 (0.94 to 1.39)	RR	57.63	483	995	370	875	9
Teeth	Non-surgical treatment vs. untreated control	otto-SR	5.03 (3.89 to 6.17)	MD	86.65		99		85	5
		Corrected	5.03 (3.89 to 6.17)	MD	86.65		99		85	5
	Surgical treatment vs. untreated control	otto-SR	6.3 (4.73 to 7.87)	MD	94.09		53		39	4
		Corrected	6.3 (4.73 to 7.87)	MD	94.09		53		39	4
Dengue	Wolbachia intervention vs. existing intervention	otto-SR	0.23 (0.18 to 0.3)	OR	0	67	2905	318	3401	1
		Corrected	0.23 (0.18 to 0.3)	OR	0	67	2905	318	3401	1
Workplace	Any workplace intervention vs. existing intervention	otto-SR	0.74 (0.33 to 1.64)	Rate Ratio	79.79					3
		Corrected	0.74 (0.33 to 1.64)	Rate Ratio	79.79					3
Etidronate	Etidronate 400mg vs.	otto-SR	0.95 (0.22 to 4.05)	RR	(Fixed effects)	5	335	4	336	4

	placebo	Corrected	0.95 (0.22 to 4.05)	RR	(Fixed effects)	5	295	4	296	3
Lung	Inhaled bronchodilator vs. placebo	otto-SR	1.08 (0.5 to 2.31)	RR	0	12	87	11	86	1
		Corrected	1.08 (0.5 to 2.31)	RR	0	12	87	11	86	1
Dialysis	HHD vs. ICHD	otto-SR	0.54 (0.15 to 1.93)	RR	80.02	221	3467	3238	22641	5
		Corrected	0.7 (0.19 to 2.59)	RR	81.38	176	1502	4162	30210	4
Nutrition	Pre-operative enteral nutrition vs. control	otto-SR	-0.78 (-2.38 to 0.83)	MD	11.02		44		42	2
		Corrected	2.22 (-3.68 to 8.12)	MD	0		11		9	1
	Pre-operative immune enhancing nutrition vs. control	otto-SR	-1.2 (-2.11 to -0.29)	MD	59.9		506		503	10
		Corrected	-1.2 (-2.11 to -0.29)	MD	59.74		503		503	10
	Pre-operative oral or enteral nutrition vs. control	otto-SR	-1.65 (-3.54 to 0.24)	MD	0		40		40	1
		Corrected	-1.65 (-3.54 to 0.24)	MD	0		40		40	1
Depression	Any psychosocial intervention vs. control	otto-SR	-0.49 (-0.75 to -0.22)	SMD	97.1878		5218		5023	49
		Corrected	-0.51 (-0.79 to -0.22)	SMD	97.3672		4876		4706	45

PRP = platelet rich plasma. HHD = home hemodialysis. ICHD = in-centre hemodialysis.

Supplementary figure 1. Data extraction answer classification

bar graph displaying answer classification resulting from dual human adjudication across all 7 evaluated systematic reviews.



Supplementary figure 2. Confusion matrix of per domain risk of bias assessments
Confusion matrix displaying per domain risk of bias assessments from the otto-SR RoB agent relative to original review risk of bias assessment decisions for (A) ROB2, (B) QUADAS-2, (C) Newcastle-Ottawa, and (D) ROB2 in the phase 4 Cochrane reproduction; relative to a corrected reference standard.

A

ROB-2

otto-SR ↓ / Corrected →	Low	Some Concerns	High	Total
Low	917	8	2	927
Some Concerns	24	105	2	131
High	0	0	2	2
Total	941	113	6	

B

QUADAS-2

otto-SR ↓ / Corrected →	Low	Unclear	High	Total
Low	232	10	5	247
Unclear	25	20	5	50
High	30	7	44	81
Total	287	37	54	

C

Newcastle-Ottawa

otto-SR ↓ / Corrected →	Star ⭐	No Star	Total
Star ⭐	581	22	603
No Star	12	51	63
Total	593	73	

D

ROB-2 - Cochrane (Phase 4)

otto-SR ↓ / Corrected →	Low	Some Concerns	High	Total
Low	75	6	1	82
Some Concerns	9	46	2	57
High	0	5	6	11
Total	84	57	9	

Supplementary figure 3. Confusion matrix of overall RoB assessments

Confusion matrix displaying overall RoB risk of bias assessments from the otto-SR RoB agent and original review risk of bias assessment decisions, relative to a corrected reference standard for (A) Phase 3 ROB2, and (B) Phase 4 Cochrane reproducibility. Matrices were only generated for the ROB2 tool, as other tools only provided assessments on a per domain/question basis.

A

ROB-2 Overall

otto-SR ↓ / Corrected →	Low	Some Concerns	High	Total
Low	118	2	0	120
Some Concerns	13	75	2	90
High	0	0	2	2
Total	131	77	4	

B

ROB-2 Overall - Cochrane (Phase 4)

otto-SR ↓ / Corrected →	Low	Some Concerns	High	Total
Low	3	0	0	3
Some Concerns	0	14	3	17
High	0	5	5	10
Total	3	19	8	