Automation of Systematic Reviews with Large Language Models

Christian Cao[1†], Rohit Arora[2†], Paul Cento[3†], Katherine Manta[1], Elina Farahani[1], Matthew Cecere[1], Anabel Selemon[4], Jason Sang[3], Ling Xi Gong[2], Robert Kloosterman[1], Scott Jiang[5], Richard Saleh[1], Denis Margalik[1], James Lin[6], Jane Jomy[1], Jerry Xie[7], David Chen[1], Jaswanth Gorla[1], Sylvia Lee[8], Kelvin Zhang[7], Harriet Ware[9], Mairead Whelan[9], Bijan Teja[1,10], Alexander A. Leung[9], Lina Ghosn[11,12,13], Rahul K. Arora[9], Allen S. Detsky[1], Michael Noetel[14], David B. Emerson[15], Isabelle Boutron[11,12,13], David Moher[1,16,17], George Church[2], Niklas Bobrovitz[9]

[1]University of Toronto; [2]Harvard Medical School; [3]Independent Researcher; [4]McGill University; [5]University of British Columbia; [6]Massachusetts Institute of Technology; [7]University of Waterloo; [8]Mount Sinai Hospital; [9]University of Calgary; [10]St. Michael's Hospital; [11]Université Paris Cité; [12]Université Sorbonne Paris Nord; [13]Cochrane France; [14]The University of Queensland; [15]Vector Institute; [16]Ottawa Hospital Research Institute; [17]University of Ottawa.

Abstract

Systematic reviews (SRs) inform evidence-based decision making. Yet, they take over a year to complete, are prone to human error, and face challenges with reproducibility; limiting access to timely and reliable information. We developed *otto-SR*, an end-to-end agentic workflow using large language models (LLMs) to support and automate the SR workflow from initial search to analysis. We found that *otto-SR* outperformed traditional dual human workflows in SR screening (*otto-SR*: 96.7% sensitivity, 97.9% specificity; human: 81.7% sensitivity, 98.1% specificity) and data extraction (*otto-SR*: 93.1% accuracy; human: 79.7% accuracy). Using *otto-SR*, we reproduced and updated an entire issue of Cochrane reviews (n=12) in two days, representing approximately 12 work-years of traditional systematic review work. Across Cochrane reviews, *otto-SR* incorrectly excluded a median of 0 studies (IQR 0 to 0.25), and found a median of 2.0 (IQR 1 to 6.5) eligible studies likely missed by the original authors. Meta-analyses revealed that *otto-SR* generated newly statistically significant findings in 2 reviews and negated significance in 1 review. These findings demonstrate that LLMs can rapidly conduct and update systematic reviews with superhuman performance, laying the foundation for automated, scalable, and reliable evidence synthesis.

1 Introduction

Systematic reviews (SRs) are the foundation of evidence-based decision-making. However, SRs are incredibly resource-intensive, typically taking over 16 months and costing upwards of $100,000 to complete[1,2]. Delays in completing SRs can have major consequences for evidence-based practice, including failure to adopt effective therapies, or prolonged use of ineffective or harmful treatments initially supported by less rigorous evidence[3].

While several tools have been developed to accelerate SRs[4,5], none are capable of full automation with human-level accuracy. However, large language models (LLMs) offer new avenues to achieve automation with their ability to process and reason about natural language. We previously demonstrated that LLMs can achieve high screening performance[6]. Other recent work has demonstrated promise for LLMs in data extraction[7,8], though these studies rely on self-defined reference standards and evaluate on small datasets.

We introduce an LLM-based workflow (*otto-SR*) to support automated and human-in-the-loop SR workflows, from initial search to data analysis. Our framework uses GPT-4.1 (OpenAI) for screening articles and o3-mini-high (OpenAI) for data extraction, targeting tasks that typically consume the majority of human researcher time and effort. We evaluate our workflow on these core SR components, article screening and data extraction, with direct comparisons to traditional human workflows and other SR automation tools. To assess real-world utility, we reproduced and updated an entire issue of Cochrane reviews (n=12) using *otto-SR*, in under two days. *otto-SR* is designed to work alongside researchers, requiring only a protocol (objectives, eligibility criteria), search results, and defined extraction variables.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

† Equal contribution and correspondence to: christian.cao@mail.utoronto.ca; rohitarora@g.harvard.edu; paul.cento@gmail.com

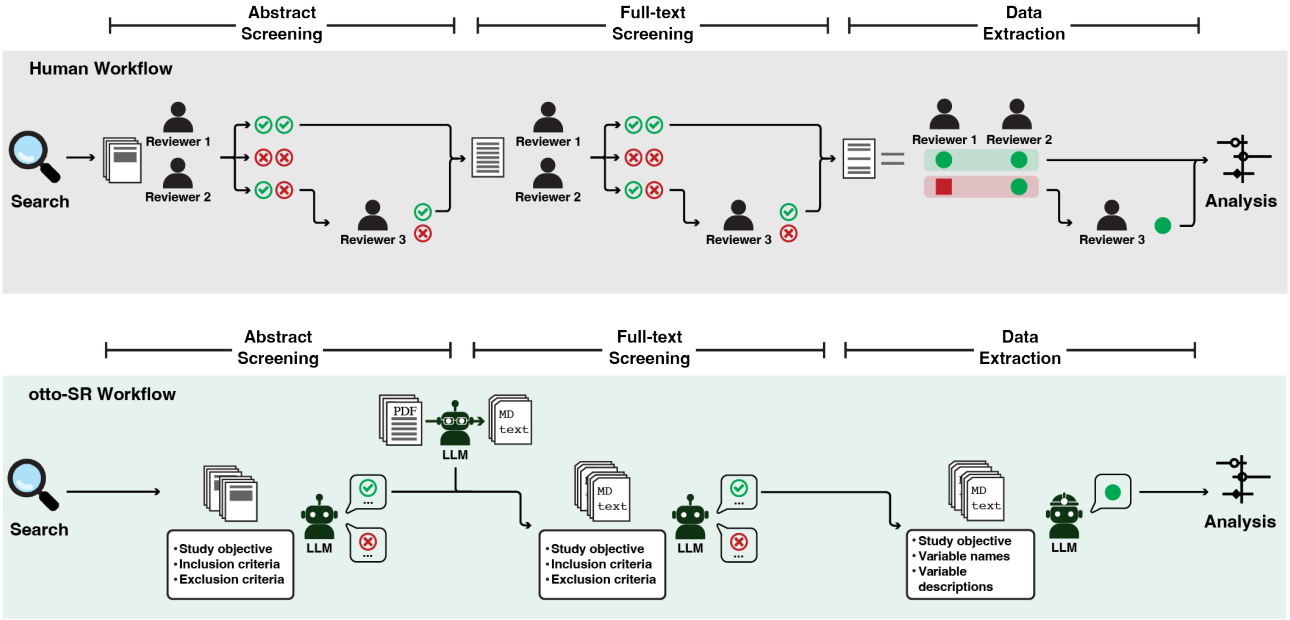


Figure 1: An automated systematic review workflow using LLMs. Infographic displaying the end-to-end SR process for humans (grey) and *otto-SR* (green). Abbreviations: Markdown file (MD)

2 An agentic workflow for systematic review automation

The gold-standard systematic review workflow begins with a comprehensive search to capture all potentially relevant citations[9]. 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 then undergo data extraction by two human reviewers independently, again adjudicated by a third reviewer when discrepancies arise. The complete human workflow is illustrated in **Figure 1** (top).

otto-SR is an end-to-end LLM-based workflow supporting both fully automated and human-in-the-loop systematic reviews. Citations identified from the original search are directly uploaded, in RIS format, to the *otto-SR* screening agent, which uses GPT-4.1 to screen abstract and full-text articles as a standalone reviewer. The resulting set of included articles is then fed into the *otto-SR* extraction agent, which performs data extraction with the o3-mini-high 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).

3 LLMs achieve state-of-the-art SR screening performance

We previously found that GPT4-preview could achieve high screening performance with effective prompting strategies[6]. Aiming to improve on these findings, we developed a screening agent leveraging GPT-4.1, a model which excels at instruction following[10,11], paired with optimized prompting strategies[6], to screen articles at abstract and full-text stages. The agent was prompted using the original, unaltered objectives and eligibility criteria from each respective review (**Supplementary Notes**). Full-text article PDFs were converted into markdown format with the Gemini 2.0 Flash model for full-text screening.

We evaluated the performance of the *otto-SR* screening agent on the complete original search across five published reviews (n=32,357 citations) covering four Oxford Centre for Evidence-Based Medicine (CEBM) question types: prevalence, diagnostic test accuracy, prognosis, intervention benefits (**Extended Data Table 1**). Dual human reviewers and Elicit (a commercial LLM-based SR automation software) were evaluated against a random representative sample of records for each review (n=1,767 citations) (Methods). The reference standard for inclusion/exclusion decisions was based on the original authors' final decisions after full-text screening.

To validate the proficiency of our human reviewers in screening, we conducted a calibration exercise (n=400 citations) where we compared the SR screening performance of our reviewers to the original study authors[12], who had independently re-screened the same set of articles. We found that the performance of our human

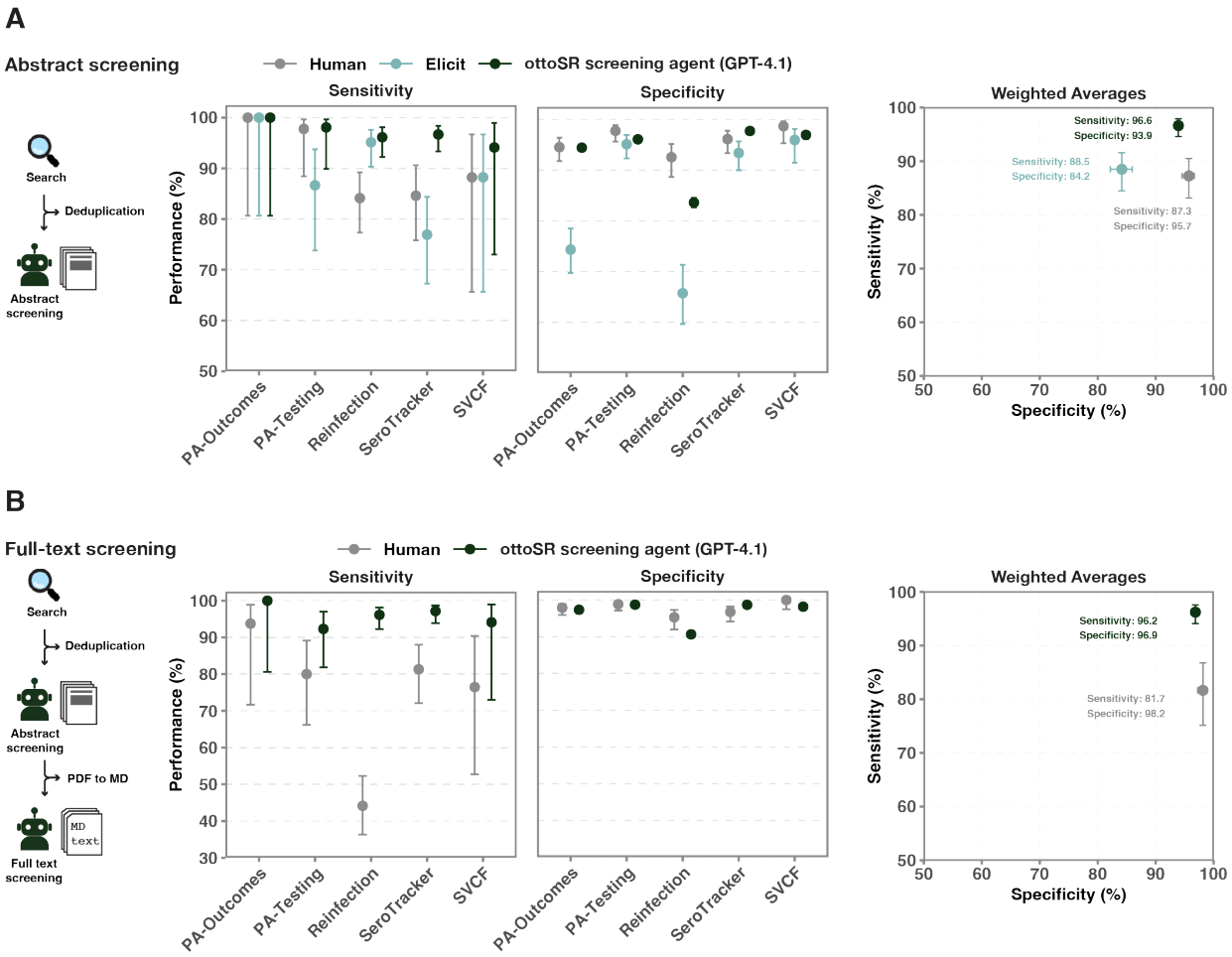


Figure 2: *otto-SR* screening agent (GPT-4.1) achieves superhuman screening sensitivity and specificity **A.** Diagram of *otto-SR* abstract screening agent (left), sensitivity, specificity of *otto-SR* screening agent, dual human reviewers, and Elicit, for abstract screening evaluated across five reviews (middle). Weighted averages for sensitivity and specificity across comparator groups (right). Error bars indicate 95% confidence intervals. **B.** diagram of *otto-SR* full-text screening agent (left), sensitivity, specificity, and accuracy of *otto-SR* screening agent evaluated across five reviews, and dual human reviewers for full-text screening evaluated across five reviews (middle). Weighted average for sensitivity and specificity across *otto-SR* (five reviews) and dual human (four reviews) (right). Error bars indicate 95% confidence intervals.

reviewers closely aligned with the original study authors (Our team: 80.2% sensitivity 97.7% specificity vs. original author team: 81.3% sensitivity, 98.1% specificity) providing confidence that our reviewers were reflective of expert-level screening (**Extended Data Table 2**).

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**, **Extended Data Table 3**). In comparison, Elicit (88.5% [76.9-100%] sensitivity) and dual human reviewers (87.3% [84.1-100%] sensitivity) had lower sensitivity. Dual human reviewers achieved the highest specificity in abstract screening (95.7% [92.5-98.7%] specificity), followed by the *otto-SR* screening agent (93.9% [83.6-97.7%] 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), while human reviewers had a marked drop in sensitivity (63.3% [44.1-93.8%] sensitivity) (**Fig. 2**, **Extended Data 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%]. 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 was not included in this comparison as it did not support

full-text screening.

Together, these findings suggest that the *otto-SR* screening agent can capture more relevant studies (true positives) than traditional dual human screening, while maintaining comparable specificity (minimizing false inclusions).

4 LLMs achieve state-of-the-art SR data extraction performance

Given the time-intensive nature of manual data extraction in SRs, we explored if advances in LLM reasoning could provide a path towards automation. To this end, we developed an extraction agent using the OpenAI o3-mini-high model[13], selected for its strong scientific reasoning, robust long-context retrieval, and cost. In all cases, the *otto-SR* extraction agent was prompted with original author-defined variable descriptions. Full-text article PDFs were also converted into markdown format with the Gemini 2.0 Flash model for data extraction.

We evaluated the performance of the *otto-SR* extraction agent and Elicit in data extraction across seven published reviews (n=4,559 data points, 495 studies) (**Fig. 3A**, **Extended Data Table 5**). Dual human reviewers were assessed on a randomly sampled subset of articles from each review based on a McNemar test sample size approximation (n=1,453 data points, 156 studies) (Methods). Extracted variables included key descriptive and outcome data used by the original authors for downstream analysis (see Supplementary Notes).

Data extraction accuracy was determined through an LLM-as-a-judge framework to compare AI- or human extracted values against the original author extractions (Methods). However, given the known variability in dual human data extraction accuracy (reported rates: 65.8-85.5%)[14–19], original author-extracted values were not treated as a definitive gold standard (**Fig. 3B**). Instead, we applied a blinded adjudication process to resolve discrepancies between *otto-SR* extraction and the original authors. A panel of blinded human reviewers

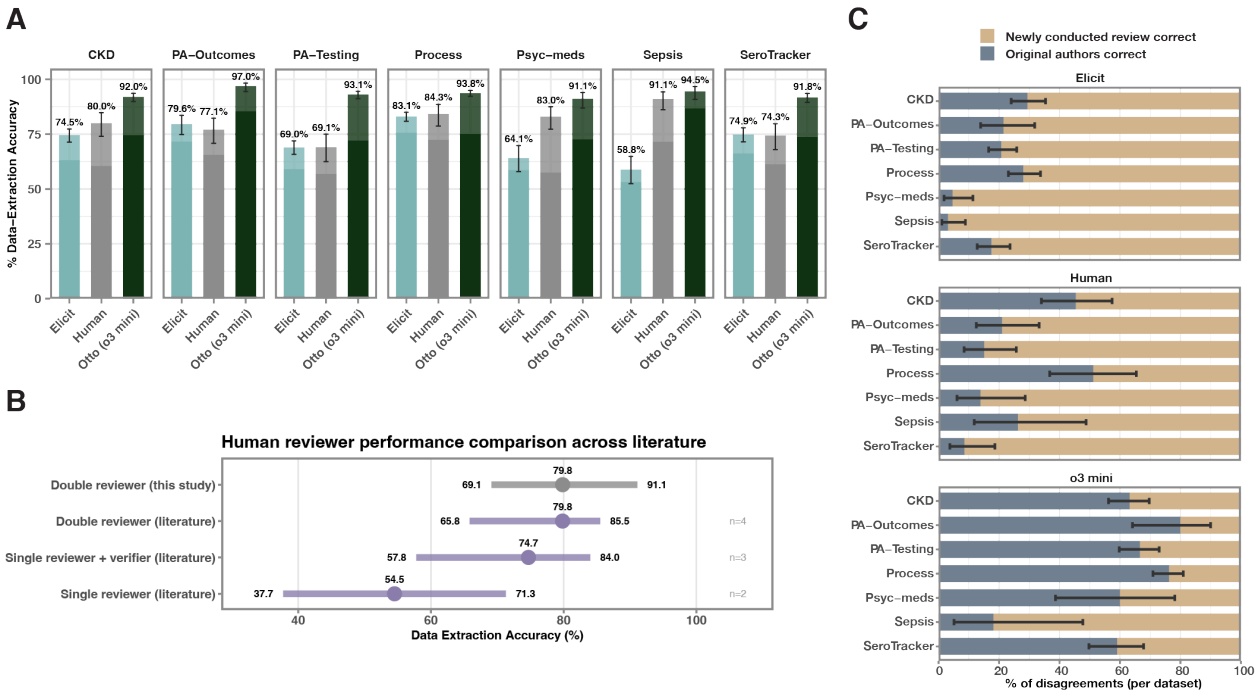


Figure 3: *otto-SR* extraction agent (o3-mini-high) performance on systematic review data extraction. A. Bar graph displaying data extraction accuracy of the *otto-SR* extraction agent (green) (4,459 data points), Elicit (teal) (4,459 data points), and dual human reviewers (grey) (1,453 data points) across 7 different systematic reviews. Error bars represent 95% confidence intervals. Shading represents pre- (lighter) and post- (darker) human adjudicated correction. **B.** Dot plot depicting literature-derived human reviewer performance comparison against human reviewers in this study. 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, Elicit, 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.

compared randomized pairs of responses (*otto-SR* vs. original author) and selected the most accurate value (see Methods). We used these judgements to construct a corrected gold standard for performance evaluations.

Across all seven reviews, the *otto-SR* extraction agent achieved an average weighted accuracy of 93.1% (91.1-97.0%), outperforming both dual human reviewers at 79.7% (69.1-91.0%), and Elicit at 74.8% (58.8-83.1%) (**Fig. 3A**, **Extended Data Figure 6**). When *otto-SR* extracted different values to the original authors, the blinded human reviewer panel 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 or Elicit, the blinded reviewer panel sided with the dual human extractors in 28.1% of cases, and Elicit in 22.4% of cases (**Fig. 3C**).

In the 6.9% of cases where the *otto-SR* extraction agent was incorrect, post-hoc analysis revealed that 0.83% (39/4459) 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 (**Extended Data Figure 1**).

5 An agentic workflow of LLMs can rapidly reproduce and update reviews

Given the high performance of our screening and extraction agents, we combined them into an agentic workflow, dubbed *otto-SR* (**Fig. 4A**). To evaluate the real-world applicability of *otto-SR*, we conducted a reproducibility assessment of a complete issue of SRs published in the Cochrane Database of Systematic Reviews.

We randomly selected the April 2024 issue of the Cochrane Database (**Extended Data Table 7**). 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 (**Extended Data Table 7**). For the 12 remaining reviews, we reproduced their reported search strategies, updating searches to May 8, 2025, and identified 146,276 citations. These citations were deduplicated and then screened at both the abstract and full-text stages using the *otto-SR* screening agent with original Cochrane review eligibility criteria (Supplementary Notes).

To ensure a focused and interpretable comparison, we diverged from Cochrane methodology in one key respect. Cochrane reviews typically include all studies, regardless of whether they report the review's primary outcome, to allow for all comparisons based on the available data (e.g., all intervention and outcome combinations). In contrast, we focused our analysis to reproduce each review's predefined primary outcome. This constraint provided a clearer distinction for study eligibility.

The *otto-SR* screening agent correctly identified all included studies (n=64) across the 12 Cochrane reviews. Citations passing screening then had primary outcome data extracted using the *otto-SR* extraction agent and original Cochrane study variable definitions (Supplementary Notes). *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) (**Extended Data Table 8**). 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 align with the original search cutoffs, we identified 54 additional eligible studies through *otto-SR* (median 2, IQR: 1 to 6.25 per review) that were likely missed in the original Cochrane reviews (Methods). *otto-SR* also incorrectly included 10 false positive articles after human review; however 9/10 may have contained relevant data available through additional author correspondence. Updating the search to May 8, 2025 identified another 14 new eligible studies (total n = 64, median 2.5, IQR 1 to 7.25 per review) (**Extended Data Table 8**). The updated search identified two additional false positive studies, one of which may have contained relevant data.

Extracted data was subsequently meta-analyzed using the same statistical methods as the original reviews, across three comparisons: (1) 'Matched' where *otto-SR* was restricted to the same set of articles as included in the original Cochrane analysis. (2) 'Expanded' which included all eligible studies identified by *otto-SR*, filtered to the original search cutoff date. (3) 'Update' which evaluated all articles with an updated May 8, 2025 search cutoff.

Given potential data extraction errors by original Cochrane authors and *otto-SR*, we derived corrected values for each comparison through dual human review. This also included removal of false positive articles and addition of false negative articles. For each review, we also generated corresponding Cochrane meta-analyses

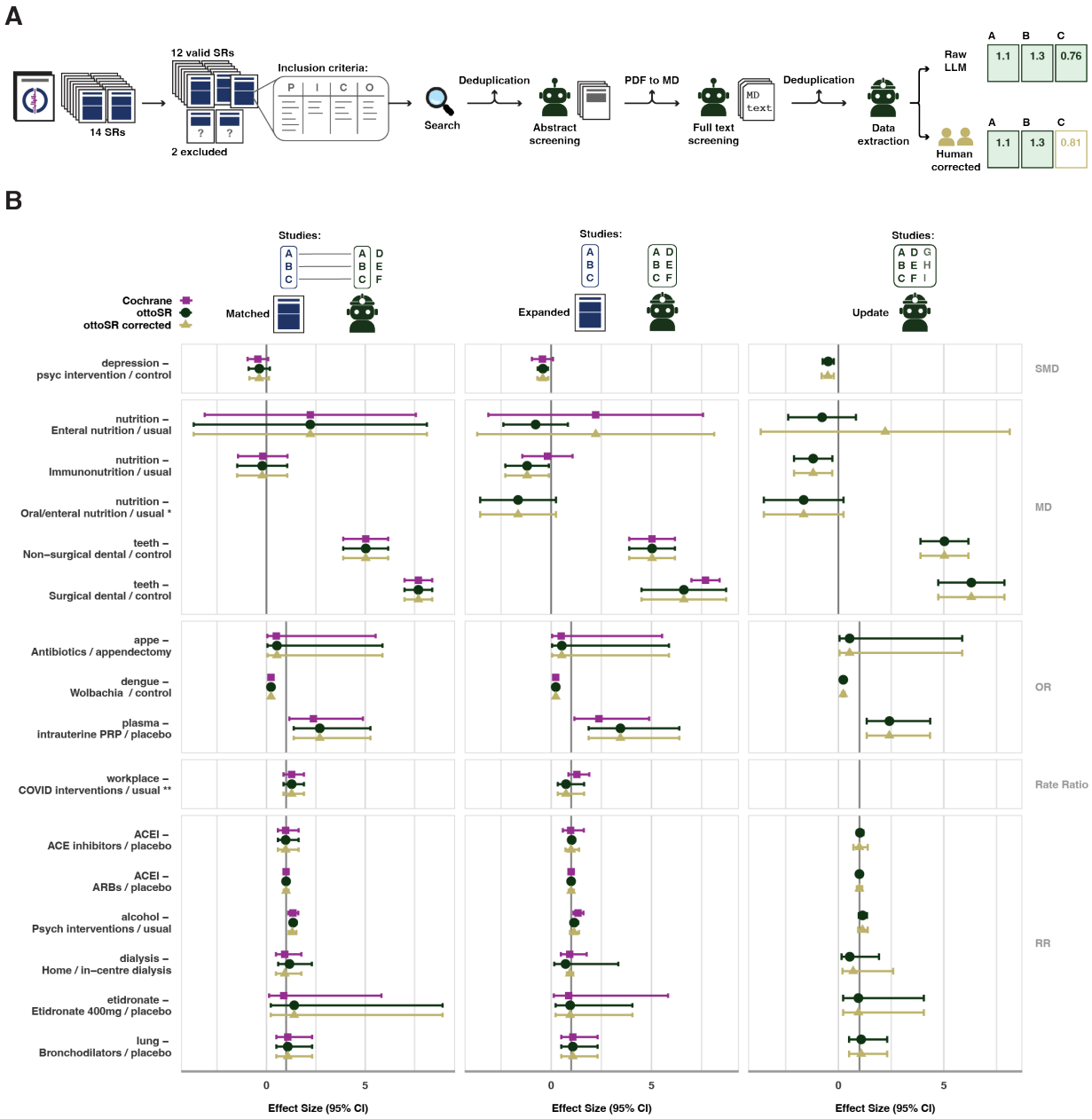


Figure 4: Evaluating *otto-SR* for automation of systematic reviews. A. Infographic depicting use of *otto-SR* for systematic review automation in a complete edition of the Cochrane Database of Systematic Reviews (n = 12). **B.** Forest plots depicting differences between *otto-SR* (green), original Cochrane study authors (purple), and corrected standard (gold). 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.

using the original author-extracted data. All original Cochrane data, *otto-SR* extracted data, and corrected data (including notes) are provided in Supplementary Data 1.

In the 'Matched' comparison group, *otto-SR* produced meta-analyzed effect estimates which had overlapping 95% CIs with both the original Cochrane data and corrected datasets across all reviews (**Fig. 4B**, left; **Extended Data Table 9**). In the 'Expanded' analysis, two reviews (nutrition, depression) yielded new statistically significant effect estimates (**Fig. 4B**, middle), while the estimate from one review (alcohol) lost statistical significance compared to the original Cochrane estimates (**Fig. 4B**, middle). These trends were consistent in the corrected 'Expanded', *otto-SR* 'updated', and the corrected 'Update' analyses (**Fig. 4B**, right).

One illustrative example comes from the nutrition review, where *otto-SR* identified 5 additional studies. This led to the new finding that preoperative immune-enhancing supplementation before gastric surgery is associated with a 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; 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 **Extended Data Table 9**.

6 Discussion

Systematic review workflows are often hindered by the time- and labor-intensive demands of screening and data extraction. In this study, we demonstrate that *otto-SR*, an end-to-end SR automation pipeline, powered by GPT 4.1 and o3-mini-high, can accelerate these steps without compromising performance.

Our findings highlight several opportunities for LLMs to be implemented in systematic reviews. First, workflows like *otto-SR* can be used to update existing systematic reviews by leveraging their original published protocols. This provides a unique advantage: it enables direct comparison of screening and extraction results against the original review, facilitating validation and assessments of reproducibility. Second, the ability to rapidly process articles opens the door to truly living systematic reviews–where updates could be performed monthly, weekly, or even daily–ensuring constant access to the most current evidence. Third, *otto-SR* may be used to generate *de novo* reviews, provided that researchers develop clear, detailed protocols akin to those pre-registered in PROSPERO. In all cases, structured and clear methodology is essential for ensuring interpretability, reproducibility, and high-quality automation.

Our Cochrane reproducibility assessment highlighted common reproducibility challenges. All 12 reviews had issues with search reproducibility and 2 reviews lacked methodological clarity. These findings align with prior work by Rethlefsen *et al*[20], 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[20–24]. To address these challenges, we suggest that published SRs include the following : (i) the 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) the complete code used for analyses. Current reporting guidelines for systematic reviews (PRISMA) endorse most, but not all of these points[25]. While Cochrane reviews routinely provide such materials, most other SRs do not, limiting reproducibility and external validation.

The performance of LLMs in conducting evidence synthesis, as demonstrated in this study, also highlights an opportunity to reconsider 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.

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. Thirdly, 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[26,27], LLM judgements may still introduce occasional 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[28–31]. This approach represented a methodological improvement over prior work that relied on self-created ground truths[7]. Sixth, the GPT-4.1 knowledge cut-off was updated to June 2024. While our screening and data extraction benchmarks primarily involved closed-source data, it's possible that the April 2024 Cochrane reviews were included in the model's pretraining corpus. This may have introduced minor inclusion bias favoring studies already included in the original Cochrane review; though it would not affect newly identified studies. Finally, with respect to the Cochrane reproducibility assessment, it's possible that content specific experts would make 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 discrepancies.

7 Conclusion

In conclusion, our study marks a major advancement in the development of SR automation tools using LLMs. The immediate applications of this include: rapid updates and truly 'living' reviews, mass assessments of reproducibility across the SR literature, and faster *de novo* reviews. Future research should focus on developing comprehensive and complete benchmarks of SRs to better support and refine automation efforts. We also encourage research into the capabilities of LLMs for other SR workflow tasks, such as search term generation and risk of bias assessment. The implementation of fully autonomous SRs could accelerate the synthesis of up-to-date evidence, save thousands of hours of manual work, and provide significant benefits in medicine and other fields.

8 Methods

8.1 Article Screening Datasets

To identify putative screening datasets, we leveraged the previously published BenchSR database of published SRs[6]. In brief, this consisted of 10 distinct SRs spanning nine unique clinical domains and contained study information (review objectives, inclusion/exclusion criteria) and the complete set of labeled 'included' and 'excluded' citations from the original search of each SR.

From the BenchSR database, we performed stratified random sampling across the four Oxford CEBM review questions, selecting SRs for each type. Our sample included various datasets: the SeroTracker dataset[32] for reviews of prevalence (calibration set adapted from Perlman-Arrow *et al.*[12]), the Reinfection[33] and PA-Outcomes[34] datasets for reviews of intervention benefits, the PA-Testing[35] dataset for reviews of diagnostic test accuracy, and SVCF[36] dataset for reviews of prognosis (**Extended Data Table 1**).

8.2 LLM Screening Methodology

We developed a novel LLM-based screening system adapted from our previously validated ScreenPrompt approach[6]. Our LLM based screening agent uses the GPT-4.1 model with a 32,768-token output limit and default parameters (temperature=1, frequency_penalty=0, presence_penalty=0, top_p=1).

For full-text screening, we implemented a PDF parsing pipeline using the Gemini 2.0 Flash model to convert full-text documents into structured Markdown inputs given a simple prompt (Supplementary Methods), which was then processed by GPT 4.1 for full-text screening. Full-text PDF articles were programmatically retrieved through OpenAlex[37], a comprehensive corpus of over 240 million scholarly works sourced from open platforms such as Crossref, MAG, DataCite, HAL, PubMed, Institutional Repositories. Articles that were not available through OpenAlex but were included after abstract screening were retrieved via institutional access.

In all instances, GPT-4.1 was prompted using the original, unaltered objectives and eligibility criteria from each respective review (Supplementary Notes).

8.3 Article Screening Benchmarking

We evaluated screening performance using a diagnostic test accuracy (DTA) 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 each SR (n = 32,357) (**Extended Data Table 3, 4**). Screening was conducted in two stages: first, an abstract-level screen was applied to all citations; those marked as "included" then underwent full-text screening to determine the final set of included articles.

For comparison, we assessed the performance of Elicit (evaluated on April 12, 2025), a commercially available systematic review automation software, and a panel of human reviewers against a representative sample of citations from each SR. To assess sensitivity, we included all articles deemed 'included' in each SR (i.e., entire inclusion set) where possible. For specificity, we determined a minimum specificity sample size of 139 'excluded' articles with Cochran's sample size[38], based on an expected specificity of 90%, 5% margin of error, and 95% confidence level. These excluded articles were randomly sampled from each review's pool of non-included citations (**Extended Data Table 3, 4**).

To evaluate the performance of Elicit in screening, we uploaded all PDF articles for each sample into the platform. Elicit was tested on a sample of citations, rather than the full dataset, due to its 500-record screening limit per review. The inclusion criteria provided to Elicit was identical to the inclusion criteria provided to the *otto-SR* screening agent. As Elicit does not natively support exclusion criteria, we tested both inclusion criteria alone and inclusion criteria combined with inverse exclusion terms. Performance was higher using inclusion criteria alone (**Extended Data Table 1**), so this approach was adopted. Elicit automatically retrieved titles and abstracts, followed by screening using a default inclusion score threshold of 2.5. No full-text screening workflow was available. All evaluations were conducted using Elicit's paid "Pro" plan.

To evaluate the performance of humans in data extraction, we assembled a panel of four graduate-level researchers with past SR experience (1 BSc, 3 MSc; all current MD students) to perform dual screening[9]. All screening was performed independently and in duplicate. Conflicts during screening were resolved by a third independent reviewer. Screening followed a standard end-to-end workflow: all citations were screened at the title/abstract stage, and citations deemed eligible by reviewers were advanced to full-text screening.

8.4 Human Calibration

To verify screening proficiency, human reviewers first completed a calibration exercise using a set of citations from SeroTracker, a comprehensive systematic review on SARS-CoV-2[32,39]. In SeroTracker, the original study authors conducted a study to assess internal consistency through re-screening a previously screened dataset using the same eligibility criteria[12]. Our reviewers screened this same dataset, and we compared their performance against the original SeroTracker authors' results.The performance of our reviewers was comparable to the original SeroTracker authors. Details are found in **Extended Data Table 2**.

8.5 Screening Data Analysis

We assessed the performance of the *otto-SR* screening agent, Elicit, and dual human reviewers by analyzing accuracy, sensitivity, specificity; and reported true positives, true negatives, false positives, and false negatives. We calculated 95% CIs for weighted (pooled-denominator) sensitivity and specificity using the Wilson method[40] with the binom package in R.

8.6 Data Extraction Datasets

We utilized four datasets (SeroTracker, PA-Outcomes, PA-Testing, Sepsis) from the BenchSR database that contained raw data extraction results provided by the original authors. In addition, we identified three external SRs (CKD[41], Process[42], Psyc-meds[43]) that also provided publicly accessible raw data extraction information (**Extended Data Table 5**). Variables assessed for extraction included key descriptive and outcome data used by the original authors for downstream analysis (see Supplementary Notes).

8.7 LLM Data Extraction Methodology

We developed a novel LLM-based data extraction agent using prompting best practices. Our data extraction agent uses the o3-mini-high model from OpenAI, high reasoning effort and a 100,000-token output limit. The same markdown extracted from full-article PDF as used in the full-text screens were passed as inputs to o3-mini-high for extraction. In all cases, the extraction agent was prompted with author-defined variables and corresponding descriptions (Supplementary Notes).

8.8 Data Extraction Benchmarking

We evaluated the performance of the *otto-SR* data extraction agent, Elicit, and dual human reviewers for data extraction across all datasets. The variable definitions used for extraction are provided in the Supplementary Notes.

For the *otto-SR* data extraction agent and Elicit, we used the complete set of articles with available data extraction results. Due to its extensive size (n=2,736 included articles), the SeroTracker dataset was randomly downsampled to 100 articles for evaluation. For the Psyc-meds dataset, only studies with published data were included (**Extended Data Table 5**).

To evaluate the performance of Elicit in data extraction (evaluated on March 22, 2025), we uploaded all articles into the Elicit data extraction platform and used the same variable descriptions provided to the *otto-SR* extraction agent (Supplementary notes). All extractions were performed with the 'high accuracy' feature, accessed through the Elicit 'Pro' paid plan. In cases where Elicit encountered an error or failed to extract data, we retried up to a maximum of 5 times.

To evaluate the performance of humans in screening, we assembled a panel of seven graduate-level researchers with past SR experience (3 BSc, 4 MSc; all current MD students). Human data extraction was performed independently and in duplicate. Discrepancies were resolved by a third human reviewer. For human data extraction, we determined sample size using a McNemar test for sample size approximation. Using an estimated human accuracy of 80% (reported rates: 65.8-85.5%)[14–19], LLM accuracy of 90%, and 95% confidence, we determined a minimum number of 204 variables per study to be extracted. Article counts for each testing dataset are provided in **Extended Data Table 5**.

Due to the unstandardized nature of data extraction results (e.g., SeroTracker review - name of immunoassay used), we used an LLM-as-a-judge framework to programmatically determine data extraction accuracy. In this setup, the o3-mini-high LLM was used to compare each AI- or human-extracted value to the original author value and determine if the two were equivalent. This evaluation method has been validated in prior LLM benchmarking efforts, including the LLM Chatbot arena[26] and OpenAI's HealthBench[27].

8.9 Data Extraction Correction

Prior research has shown wide variability in the accuracy of dual human extraction, with reported rates ranging from 65.8-85.5%[14–19]. As such, original author-provided values did not represent a reliable ground truth. To address this, we conducted a blinded correction process for cases where the LLM-as-a-judge flagged discrepancies between *otto-SR* extracted values and original review author extracted values. A panel of three independent, experienced graduate-level human reviewers validated outputs. Reviewers were presented with two anonymized and randomized responses (LLM and original author) and asked to select one of four options: Option A correct, Option B correct, Both correct, or Neither correct. Each disagreement was evaluated in parallel and resolved by a third independent arbitrator. The final adjudicated results were used to construct corrected ground truth datasets. To evaluate the accuracy of *otto-SR*, Elicit, and dual human reviewers, we then applied the LLM-as-a-judge framework to compare each system's outputs against this corrected dataset. This adjudication framework was adapted from established protocols in radiology, ophthalmology, and clinical trials[28–31].

We note a potential limitation in our validation process: when *otto-SR* and the original authors produced identical values, we assumed these were correct without further adjudication. Consequently, if both sources made the same systematic error, it would go undetected. This approach could potentially bias our evaluation against alternative models (e.g., Elicit or dual human reviewers) that disagreed with both reference sources. To evaluate this limitation, we performed spot checks on a random 10% sample of extractions where *otto-SR* and the original authors agreed, finding no errors or inconsistencies.

We additionally performed a post-hoc review of incorrect AI outputs to classify errors as either 'parsing errors' (i.e., errors with the PDF parsing pipeline), 'inaccessible' (i.e., data accessible only through author correspondence or supplementary material), or 'true errors' (i.e., cases where the original author values were correctly extracted).

8.10 Extraction Data Analysis

We assessed the performance of the *otto-SR* data extraction agent, Elicit, and human reviewers by analyzing the total accuracy at a variable level per study. If the human adjudicator classification was "inaccessible", the

data point was removed from analysis for *otto-SR*, Elicit, and human reviewers. We calculated 95% CIs for weighted (pooled-denominator) accuracy using the Wilson method[40] with the binom package in R.

8.11 Cochrane Reproducibility

To evaluate the reliability and generalizability of our automated systematic review workflow, we conducted a focused reproducibility assessment using an entire issue of the Cochrane Database of Systematic Reviews. Our aim was to approximate each review's workflow end-to-end, from literature search through to data extraction and meta-analysis, using the *otto-SR* pipeline.

We selected the April 2024 issue through random sampling. Of the 14 reviews published, two were excluded: one due to the absence of downloadable data, and another due to an irreproducible search strategy (authors provided a search strategy to populate the Cochrane specialized register, but not for the review itself). This left 12 eligible reviews spanning a range of clinical domains (**Extended Data Table 7**). The Cochrane database was chosen for its rigorous and standardized reporting practices, public data availability, and detailed methodological documentation.

8.12 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 from the original search were excluded from downstream screening, data extraction and analysis.

8.13 Cochrane Screening

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

To ensure a focused and interpretable comparison, we deviated from Cochrane's inclusion practice in one key respect. Cochrane reviews typically include all studies reporting the eligible population and intervention of interest. This approach allows authors to explore all comparisons based on the available data (e.g., all intervention and outcome combinations, including those not pre-specified). While valuable for comprehensive synthesis, the generation of comparisons after screening can make study eligibility unclear.

Instead of focusing on all possible comparisons, we focused our analysis to reproduce each Cochrane review's primary analytical comparison. This constraint allowed for unambiguous inclusion criteria, where studies had to meet specific predefined interventions, comparators, and outcome criteria. Citations without a retrievable abstract or DOI/trial identifier were excluded. Screening decisions were compared against Cochrane author decisions to calculate true positives, false negatives, false positives, and true negatives.

8.14 Cochrane Data Extraction

For all studies passing full-text screening, outcome data was extracted using the *otto-SR* extraction agent, focusing exclusively on the primary outcome defined in each Cochrane review. To ensure consistency, we used the original author-defined variable names and extraction logic (Supplementary Notes).

Data extraction also served as a secondary filter. While *otto-SR* achieved high specificity (~97%), for a review of 10,000 citations, this would equate to nearly 300 false positive articles. To counteract this, studies were programmatically excluded if they returned unavailable or unreportable values for the primary outcome (e.g., "na" values), were identified as duplicates, or involved ineligible intervention-comparator pairs (e.g., Drug A vs. Drug B when only Drug A vs. placebo was eligible). This secondary filtering step helped remove residual false positives from the screening phase, though it may have introduced occasional misclassifications (then labeled as false negatives).

8.15 Analysis

All meta-analyses were conducted using the metafor package in R (Code and Dataset Availability). To ensure fair comparison, we matched the original authors' specified meta-analytic model (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 Cochrane primary analysis study set. (3) Expanded – all eligible studies included by *otto-SR* under the original search cut-off. (4) Updated – all eligible studies included by *otto-SR* from an updated search extending to May 8 2025. We also derived 'corrected' values (see below), for the 'matched,' 'expanded,' and 'updated' comparisons that served as the reference ground truth for analytical comparison.

8.16 Cochrane Data Correction and Comparison

To address known concerns about the reliability of original author-extracted data, we conducted an adjudication process to derive corrected data extraction and screening information for the 'Matched,' 'Expanded,' and 'Updated' analyses. For our 'Matched' analysis, a panel of two human reviewers compared data extraction and screening decisions from the original Cochrane authors and the *otto-SR* extraction agent, selecting the correct value through re-assessment of the source article. In our 'Expanded' and 'Updated' analysis, where Cochrane data was not available, a panel of two human reviewers compared *otto-SR* data extraction and screening decisions against original study articles, selecting the correct value through re-assessment of the source article. Final articles included in the corrected analysis consisted of all *otto-SR* true positive articles, and any Cochrane true positive articles missed by *otto-SR*. The extracted values in this final dataset reflected the most accurate, reviewer-verified information and served as the reference standard for Cochrane and *otto-SR* performance comparisons (Supplementary Data 1 for raw and corrected values, reviewer notes, and error classifications).

To ensure consistency and transparency, we applied standardized rules for study eligibility across analysis sets. First, articles had to be retrievable through our reproduced search; unretrievable citations were excluded from all *otto-SR*-based analyses. Second, for author data requests, we included studies in the Cochrane analysis only if authors explicitly stated that they contacted study authors and specified which studies and outcomes were supplemented. If a data request was suspected for a study, but the review only reported vague references to data requests without further specification, the study was considered unverifiable and excluded from both the Cochrane and *otto-SR*-corrected analyses. For instance, in the ACEi review, the authors appeared to assign zero mortality events in studies that did not report mortality or adverse events; but did not clearly state which studies had data requests. For such cases, we excluded those studies from the Cochrane and *otto-SR* corrected analyses to avoid introducing speculative data. Suspected data requests occurred in three reviews (4 studies, Nutrition; 15 studies, ACEi; 4 studies, Depression). A high-level summary of methodological issues is provided in **Extended Data Table 7**. Detailed notes for the exclusion of studies are provided in Supplementary Data 1.

Studies with supplementary data (not extractable by *otto-SR*) were included in the Cochrane and corrected analyses, thereby penalizing the model. If the data was inaccessible to *otto-SR* due to format limitations (e.g., embedded figures), the study was excluded from *otto-SR* analyses but retained in the Cochrane and corrected sets. These criteria aimed to balance reproducibility, verifiability, and the practical constraints of automation.

9 Acknowledgements

We thank Guravneet Gill for assistance with database access; Zion Chan and Li Li 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.

10 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, I.B, D.B.E, R.K.A, L.G, 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.

11 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, and 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 review, LLC. R.A is a non equity holding founder/advisor in otto review, LLC.

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

12 Code and Dataset Availability

All datasets and code used for data analysis will be made available on publication.

13 References

1. 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 Feb;7(2):e012545.
2. 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 Dec;16:100443.
3. Prasad V, Cifu A. Medical reversal: why we must raise the bar before adopting new technologies. *Yale J Biol Med*. 2011 Dec;84(4):471–8.
4. 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 Jun;4(2):e12234.
5. 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 Oct 25;13(1):269.
6. 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 Feb 25;ANNALS-24-02189.
7. 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 Jan 31;8(1):74.
8. 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 Jul;15(4):576–89.
9. 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 Editorial Unit, editor. *Cochrane Database Syst Rev* [Internet]. 2019 Oct 3 [cited 2024 Jun 1]; Available from: https://doi.wiley.com/10.1002/14651858.ED000142
10. OpenAI. Introducing GPT-4.1 in the API [Internet]. OpenAI. [cited 2025 Jun 8]. Available from: https://openai.com/index/gpt-4-1/
11. Sirdeshmukh V, Deshpande K, Mols J, Jin L, Cardona EY, Lee D, et al. MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs [Internet]. *arXiv*; 2025 [cited 2025 Jun 8]. Available from: https://arxiv.org/abs/2501.17399
12. 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 Jul;14(4):608–21.
13. OpenAI. OpenAI o3-mini System Card [Internet]. OpenAI; 2025. Available from: https://cdn.openai.com/o3-mini-system-card-feb10.pdf
14. 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 Mar;63(3):289–98.
15. 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 Jul;59(7):697–703.
16. 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 Dec;6(1):539.
17. 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
18. 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 May 10;e069155.
19. 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 Nov;115:77–89.
20. 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 Feb;166:111229.
21. Ioannidis JPA. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. Milbank Q. 2016 Sep;94(3):485–514.
22. 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 Nov 28;17(1):152.
23. 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 Apr;9(4):e028382.
24. 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 Mar;10(3):e032497.
25. 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 Mar 29;n71.
26. 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]. arXiv; 2023 [cited 2025 Jun 8]. Available from: https://arxiv.org/abs/2306.05685
27. 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]. OpenAI; 2025 [cited 2025 Jun 8]. Available from: https://openai.com/index/healthbench/
28. 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 Aug 1;312(2):e240272.
29. 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 Jul 20;9(1):429.
30. 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 Sep 1;137(9):987.
31. 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 Aug;50(8):2187–96.
32. 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. Khudyakov YE, editor. *PLOS ONE*. 2021 Jun 23;16(6):e0252617.
33. 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 May;23(5):556–67.
34. 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 Feb 20;109(3):e1280–9.
35. 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 Aug;79(8):1835–44.
36. 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]. *PROSPERO*. Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022384259
37. Priem J, Piwowar H, Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts [Internet]. *arXiv*; 2022 [cited 2025 Jun 8]. Available from: https://arxiv.org/abs/2205.01833
38. Cochran WG. Sampling techniques, 3rd edition. John Wiley; 2002.
39. Bergeri I, Whelan MG, Ware H, Subissi L, Nardone A, Lewis HC, et al. Global SARS-CoV-2 seroprevalence from January 2020 to April 2022: A systematic review and meta-analysis of standardized population-based studies. Suthar AB, editor. *PLOS Med*. 2022 Nov 10;19(11):e1004107.
40. Wilson EB. Probable Inference, the Law of Succession, and Statistical Inference. *J Am Stat Assoc*. 1927 Jun;22(158):209–12.
41. 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. Aoun M, editor. *PLOS ONE*. 2022 Jul 29;17(7):e0264167.
42. Antonacci G, Lennox L, Barlow J, Evans L, Reed J. Process mapping in healthcare: a systematic review. *BMC Health Serv Res*. 2021 Apr 14;21(1):342.
43. Kopcalic K, Arcaro J, Pinto A, Ali S, Barbui C, Curatoli C, et al. Antidepressants versus placebo for generalised anxiety disorder (GAD). Cochrane Central Editorial Service, editor. *Cochrane Database Syst Rev* [Internet]. 2025 Jan 30 [cited 2025 Jun 8];2025(2). Available from: http://doi.wiley.com/10.1002/14651858.CD012942.pub2

Automation of Systematic Reviews with Large Language Models

Extended Data Tables and Figures

Extended Data Table 1 – Descriptive overview of datasets used in article screening

CEBM Type	Dataset	Clinical Domain (Web of Science)	Population	Intervention/ Exposure	Study Types Included	Citations Identified from Electronic Search after deduplication (Total N) [Human/Elicit Sample N]	Citations Included After Full-Text Screening ('Included' N) [Human/Elicit Sample 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	10,000 [400]	210 [91]
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	8,000 [400]	52 [45]
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	2,257 [167]	17 [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.	6,724 [400]	181 [145]
	PA-Outcomes	Endocrinology & Metabolism Surgery	Patients with primary aldosteronism	Surgical adrenalectomy and medical treatment with a mineralocorticoid receptor antagonist	Randomized controlled trials, observational studies	5,376 [400]	16 [16]
Intervention Harms *What are the COMMON/ RARE harms?*	-	-	-	-	-	-	-
Screening *Is this (early detection) test worthwhile?*	-	-	-	-	-	-	-

Extended Data 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.

Extended Data Table 3 – Abstract Screening Performance of *otto-SR*, dual human reviewers, and Elicit

Dataset	Total Records (Include /Exclude) (n)	Model	True Positives (n)	True Negatives (n)	False positives (n)	False negatives (n)	Sensitivity (95% CI)	Specificity (95% CI)
SeroTracker	10,000 (210/9,790)	otto-SR	204	9,567	223	6	96.7 (93.3-98.4)	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)
	393 (91/302)	Elicit	70	282	20	21	76.9 (67.3-84.4)	93.4 (90.0-95.7)
Reinfection	6180 (181/5,999)	otto-SR	174	5,016	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)
	395 (144/251)	Elicit	137	165	86	7	95.1 (90.3-97.6)	65.7 (59.7-71.3)
PA-Testing	7757 (52/7,705)	otto-SR	51	7,403	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)
	396 (45/348)	Elicit	39	331	17	6	86.7 (73.8-93.7)	95.1 (92.3-96.9)
PA-Outcomes	4309 (16/4,293)	otto-SR	16	4,054	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)
	394 (16/378)	Elicit	16	281	97	0	100.0 (80.6-100.0)	74.3 (69.7-78.5)
SVCF	1954 (17/1,937)	otto-SR	15	148	2	2	94.1 (73.0-99.0)	97.0 (96.1-97.6)
	167 (17/150)	Dual human	16	1,878	59	1	88.2 (65.7-96.7)	98.7 (95.3-99.6)
	165 (17/148)	Elicit	15	142	6	2	88.2 (65.7-96.7)	95.9 (91.4-98.1)

Extended Data Table 4 – Full-text screening performance of *otto-SR* and dual human reviewers

Dataset	Total Records (Include /Exclude) (n)	Model	True Positives (n)	True Negatives (n)	False positives (n)	False negatives (n)	Sensitivity (95% CI)	Specificity (95% CI)
Serotracker	10,000 (210/9,790)	otto-SR	204	9,663	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)
Reinfection	6180 (181/5,999)	otto-SR	174	5,439	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)
PA-Testing	7757 (52/7,705)	otto-SR	48	7,608	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)
PA-Outcomes	4309 (16/4,293)	otto-SR	16	4,179	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)
SVCF	1954 (17/1,937)	otto-SR	16	1,902	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)

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

Extended Data Table 5 – Descriptive overview of datasets used in data extraction benchmark

Dataset	Clinical Domain (Web of Science)	Population	Intervention/ Exposure	Study Types Included	Total articles [Human sample N]	Total data-points [Human 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]

Extended Data Table 6 – Accuracy of otto-SR, dual human reviewers, and Elicit in data extraction

Dataset	Model	Raw numerator	Total Data Points	Corrected Numerator	Corrected Denominator	Accuracy
SeroTracker	Human	129	210	156	210	74.3
	Elicit	463	700	524	700	74.9
	Otto	517	700	636	693	91.8
PA-Outcomes	Human	138	210	158	205	77.1
	Elicit	241	336	261	328	79.6
	Otto	287	336	318	328	97.0
PA-Testing	Human	116	204	141	204	69.1
	Elicit	513	867	598	867	69.0
	Otto	625	867	807	867	93.1
Sepsis	Human	149	208	174	191	91.1
	Elicit	136	256	140	238	58.8
	Otto	222	256	225	238	94.5
Process	Human	148	204	172	204	84.3
	Elicit	944	1248	1037	1248	83.1
	Otto	938	1248	1170	1248	93.8
CKD	Human	127	210	168	210	80.0
	Elicit	505	800	596	800	74.5
	Otto	596	800	736	800	92.0
Psyc-meds	Human	119	207	171	206	83.0
	Elicit	148	252	159	248	64.1
	Otto	183	252	226	248	91.1

Extended Data Table 7 – 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	Time taken: Electronic search to publication (Months)
ACEI	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
Appe	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	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	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	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	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	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	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 raw RIS files from the initial search. Consequently, no update was performed.		April 2023 - search
Etidronate	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	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 exactly copied.		May 12 2023
Dialysis	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	Excluded - no publicly available downloadable data						
Nutrition	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	Adults with heart disease	Any psychological intervention for depression	RCTs	Depression symptom score (latest time point)	1. Ovid Medline version not specified, used 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. Psychinfo: 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.	The authors report “we measured depression and anxiety as change in 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.	July 5/7 2022 search

Extended Data Table 8 – Descriptive overview of studies included and excluded in the Cochrane reproducibility analysis

Dataset	Citations Identified from Electronic Search Search up to May 2025	Cochrane included* N study	*otto-SR* included 'Matched' N study	*otto-SR* excluded 'matched' *Cochrane studies incorrectly excluded by otto-SR* N study	*otto-SR* included 'Expanded' *Total articles included with Cochrane search* N study (N false positive)	*otto-SR* included '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	4	5

				Marshall 2021: Otto incorrectly excluded. Cardiovascular mortality data was in supplementary data (inaccessible to LLM)	(1 false positive)	(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)

Extended Data Table 9 – 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 intervention	Cochrane	1.28 (0.86 to 1.9)	Rate Ratio	0	-	-	-	-	1
		otto-SR	1.28 (0.86 to 1.9)	Rate Ratio	0	-	-	-	-	1
		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 immune enhancing nutrition vs. control	Cochrane	-0.19 (-1.44 to 1.07)	MD	24.77	-	192	-	192	4
		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

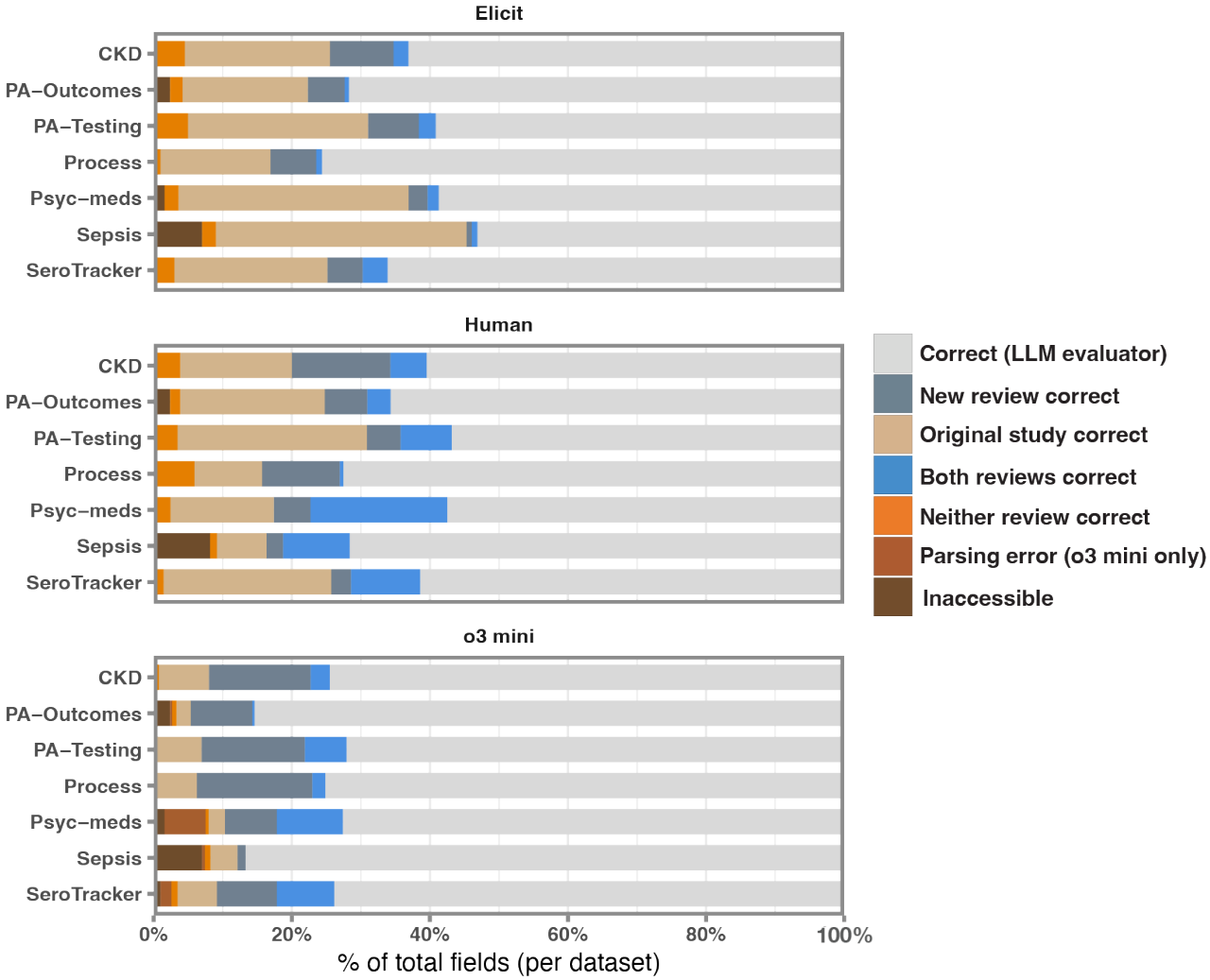
Expanded										
ACEi	ACEi alone vs. control	Cochrane	0.97 (0.58 to 1.63)	RR	10.66	346	2823	343	2867	8
		otto-SR	1.02 (0.88 to 1.18)	RR	0	368	3424	366	3419	13
		Corrected	0.99 (0.71 to 1.39)	RR	4.288	353	3212	347	3212	13
	ARB alone vs. control	Cochrane	0.99 (0.87 to 1.13)	RR	0	271	2128	272	1913	7
		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 0.3)	OR	0	67	2905	318	3401	1

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 vs. appendectomy	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	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. placebo	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	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.



Extended Data Figure 1: Data extraction answer classification. Bar graph displaying answer classification resulting from dual human adjudication across o3-mini-high, dual human extraction, and Elicit across all 7 evaluated systematic reviews.