Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AutoMeta: LLM Agent-based Automated Meta-Analysis in Public Health

Ke Zhu
Tsinghua University
Yuying Huang
Tsinghua University
Xingyue Tian
Tsinghua University
Yamin Sun
National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases
Tong Xia
tongxia@mail.tsinghua.edu.cn

Tsinghua University

Method Article

Keywords: Systematic reviews, Meta-analysis, Evidence synthesis, Large language models, AI agents, Automation

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11389292/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AutoMeta: LLM Agent-based Automated Meta-Analysis in Public Health

Ke Zhu[1], Yuying Huang[1], Xingyue Tian[1], Yamin Sun[2], Tong Xia[1*]

[1]Vanke School of Public Health, Tsinghua University, Beijing, China.
[2]National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, Beijing, China.

*Corresponding author(s). E-mail(s): tongxia@mail.tsinghua.edu.cn;

Abstract

Background: Meta-analysis supports public health decision-making, yet conducting and updating reviews remain time- and labour-intensive. Although large language model (LLM) agent systems have been proposed to automate literature reviews, heterogeneous exposure definitions, effect measures and covariate-adjustment models pose challenges for public health evidence synthesis.

Methods: We present AutoMeta, an LLM-based agentic workflow for public health meta-analysis that uses stage-specific instructions to guide planning and tool use across literature search, screening, data extraction and statistical synthesis, with structured outputs linking each stage. To address the heterogeneity of public health evidence, it expands exposure terminology, structures eligibility assessments, links extracted estimates to their analytical context, and uses tested functions for effect-size transformation and synthesis. For systematic evaluation,

we constructed AutoMetaBench from 27 systematic reviews spanning environmental health, nutrition and health behaviours, encompassing 421 included-study records, 4164 data fields and 95 effect units. We used this benchmark to assess performance at each stage and end to end.

Results: AutoMeta improved evidence coverage and synthesis accuracy while supporting rapid end-to-end execution. In stage-specific evaluations, it outperformed the respective comparators across all four stages, achieving 93.1% search recall, 94.8% recall among the top 50 ranked records, and 93.6% data extraction accuracy. Agreement with reference pooled effects and 95% confidence intervals reached 96.8% and 95.8%, respectively. Across 27 end-to-end tasks, AutoMeta reproduced reference pooled effects and both 95% confidence interval bounds within 5% relative error in 85.2% of tasks, compared with 7.4% for the single-agent comparator, while taking an average of only 26 minutes per task.

Conclusions: AutoMeta demonstrates substantial advantages in evidence synthesis and statistical analysis for public health meta-analysis. Its LLM-based agentic workflow improves synthesis efficiency and accuracy, while AutoMetaBench provides a dedicated benchmark for the systematic evaluation of automated methods. These contributions advance automated public health evidence synthesis and support timely evidence updates.

Keywords: Systematic reviews, Meta-analysis, Evidence synthesis, Large language models, AI agents, Automation

1 Background

In public health research, meta-analyses synthesize evidence across studies to quantify associations between exposures and health outcomes and evaluate intervention effects. Meta-analyses are positioned at the top of the evidence pyramid in evidence-based medicine [1]. Their findings inform public health guidelines, disease prevention

and policy decisions [2, 3]. However, producing these reviews still requires substantial expert judgement and manual work. A previous analysis estimated an average of approximately 67 weeks from review initiation to publication, with five authors per review [4]. As primary research accumulates, published reviews must incorporate new evidence to remain current and useful for decision-making [5, 6]. The growing demand for evidence synthesis and updates has therefore prompted efforts to develop automated methods that reduce manual workload while balancing efficiency and reliability [7].

Earlier machine-learning approaches supported study classification and screening prioritization [8], including active-learning frameworks such as ASReview [9]. Large language models (LLMs) are increasingly used to identify evidence and organize study data for systematic reviews and meta-analyses [10, 11]. These applications draw on their ability to follow instructions, interpret language and extract information from unstructured sources. Evaluations of screening and data extraction indicate that performance varies across tasks and evaluation settings [12–14]. LLM-based agents build on these capabilities by combining task planning, tool use and execution feedback to coordinate tasks across stages of evidence synthesis [15, 16]. For individual tasks, AutoBool uses retrieval performance as a reinforcement learning reward to train models to generate Boolean search queries [17]. CrossCritique verifies data reported in text, tables and figures through independent extraction by two models followed by review of disagreements [18]. Among multistage workflows, TrialMind constructs PubMed queries from clinical questions, generates eligibility criteria and connects these steps to the extraction of target results [19]. otto-SR assigns screening and extraction to separate agents that process study reports and data according to a predefined review protocol [20]. LUMEN coordinates multiple agents and invokes statistical software to perform meta-analysis, extending automation to quantitative synthesis [21].

However, existing tools have largely been developed and evaluated in clinical research settings [19–21]. Extending these tools to public health meta-analysis requires addressing methodological challenges arising from diverse study designs and complex evidence characteristics [22]. During literature searches, environmental and behavioural exposures may be described using different proxy measures, measurement approaches, and spatial and temporal scales. Searches relying solely on Medical Subject Headings (MeSH) may therefore fail to capture the full range of terminology used in these studies [22–24]. During screening, free-form generation of eligibility criteria, as in TrialMind, may omit or duplicate core requirements, affecting the completeness and consistency of eligibility assessments [19]. During data extraction, a single study may report multiple effect estimates for different adjustment models, subgroups, lag periods and sensitivity analyses. These values and their analytical context may be distributed across text, tables and figures, increasing the risk of mismatches with prespecified analysis targets [2, 23, 25]. Finally, public health studies report diverse effect measures, requiring calculation and transformation methods appropriate to the study design and available data. Valid statistical synthesis also depends on correctly handling effect estimates and their variances [2].

To address these challenges, we introduce AutoMeta, an agentic workflow for public health meta-analysis. Four agents plan stage-specific tasks and invoke appropriate tools to identify evidence, extract data and perform statistical synthesis. The search agent combines MeSH terms and free-text terms using PubMed field tags to reduce omissions caused by variation in terminology. The screening agent organizes eligibility assessments around four predefined dimensions: population, intervention or exposure, comparator and outcome. This structure supports comprehensive and consistent assessment of core requirements while retaining the evidence underlying each judgement. The extraction agent retrieves target values from text, tables and figures. It aligns and combines complementary information across these sources to link each

value to its corresponding adjustment model, subgroup and lag period. The resulting study-level data preserve both the analytical context and the provenance of each value. For statistical synthesis, we developed and tested a library of functions implementing established methods for calculating and transforming different effect measures. The meta-analysis agent selects appropriate functions according to the data type and target effect measure. This approach reduces the risk of computational errors arising from model-generated numerical results or ad hoc formulas.

To evaluate AutoMeta systematically, we constructed AutoMetaBench from 27 published systematic reviews spanning environmental health, nutrition and health behaviours, and conducted stage-specific comparisons and end-to-end evaluations. Stage-specific comparisons assessed literature search, screening, targeted data extraction and statistical synthesis using the inputs specified for each stage. For end-to-end evaluation, we selected one primary outcome from each review to construct 27 tasks. Each task began with a research question and proceeded through all four stages, using the actual outputs of preceding stages for subsequent processing. This design assessed evidence coverage, reproduction of reference results and execution efficiency across the complete workflow.

AutoMeta outperformed the respective comparators in all four stage-specific evaluations. In end-to-end comparisons with a single-agent system using the same model and tools, AutoMeta achieved a mean reference-study recall of 89.3%. It reproduced the reference pooled effect and both 95% confidence interval bounds in 85.2% of tasks, with each quantity within 5% relative error. Its mean total runtime was also lower than that of the single-agent comparator. These findings support AutoMeta's ability to combine evidence coverage, synthesis accuracy and execution efficiency throughout the workflow. Together, AutoMeta and AutoMetaBench provide a reusable workflow and evaluation resource for public health meta-analyses, supporting more efficient, verifiable evidence synthesis and subsequent updates.

2 Methods

We developed AutoMeta, an agentic workflow for meta-analysis in public health research (Fig. 1). The system accepts research questions expressed in natural language or predefined PICO/PECO elements as input. Benchmark construction and evaluation procedures are described in Section 3.

2.1 AutoMeta workflow

We represent each stage in AutoMeta as a task execution unit supported by a large language model (LLM). Each stage is configured with task instructions, available tools and output requirements suited to its role. The LLM provides semantic understanding, judgement and planning capabilities, while the agent coordinates model processing, tool calls and output generation. Task execution at each stage is expressed in a common form:

$$\begin{aligned} q &= (P,\ I/E,\ C,\ O), \\ y_s &= \mathrm{Agent}_s\left(q, x_s; \mathcal{I}_s, \mathcal{T}_s\right), \end{aligned} \tag{1}$$

Here, q denotes the research question, with P, I/E, C and O representing the target population, intervention or exposure, comparator and outcome, respectively [2, 22]. The stage index $s \in \{\mathrm{search}, \mathrm{screen}, \mathrm{extract}, \mathrm{meta}\}$ corresponds to literature search, study screening, data extraction and statistical synthesis, respectively. The terms x_s and y_s denote the stage input and structured output, respectively. The stage instructions and available tool set are denoted by $\mathcal{I}_s$ and $\mathcal{T}_s$, respectively. Agent_s represents the overall process required to complete the stage task and may involve one or more model and tool calls. This notation does not require stages to use the same number or sequence of calls.

Stage instructions specify the task objectives, processing constraints and output format. The shared research question q provides a consistent research scope and analytical objectives across stages. The stage input x_s supplies the materials required

for the current task. These include outputs from preceding stages and supplementary information, such as full texts of included studies, prespecified extraction targets or confirmed data. This information is assembled into a task context appropriate to the current processing step, guiding the model in interpreting inputs, making judgements or formulating plans. Search results, parsed content and other information returned by tools during execution can also inform the context for subsequent steps.

Tool calls connect the model's judgements or plans to concrete operations. The agent invokes tools from $\mathcal{T}_s$ as required by the task, supplying data and parameters that meet their input requirements. Tools perform operations such as database searches, document parsing or numerical computation. Their outputs can support subsequent model processing or be incorporated directly into stage results. The final output y_s records the results and necessary associated information according to stage requirements, supporting subsequent processing or presentation. Structured inputs and outputs allow stages to exchange and reuse the information they need while maintaining clear boundaries between their tasks.

AutoMeta is available as a web application with a unified interface through which researchers can enter research questions and initiate the meta-analysis workflow (Fig. 2).

2.2 Literature search with retrieval-guided query refinement

The literature search agent receives the shared research question q, identifies search concepts from the PICO/PECO definition and constructs executable queries. It then invokes the PubMed search tool to retrieve records and parses them into a structured format (Fig. 1). For questions about intervention effects or exposure associations, search concepts are organized around the target intervention or exposure, respectively, together with applicable population, outcome and other search criteria. Related terms within each concept are joined with OR to form concept blocks, which are

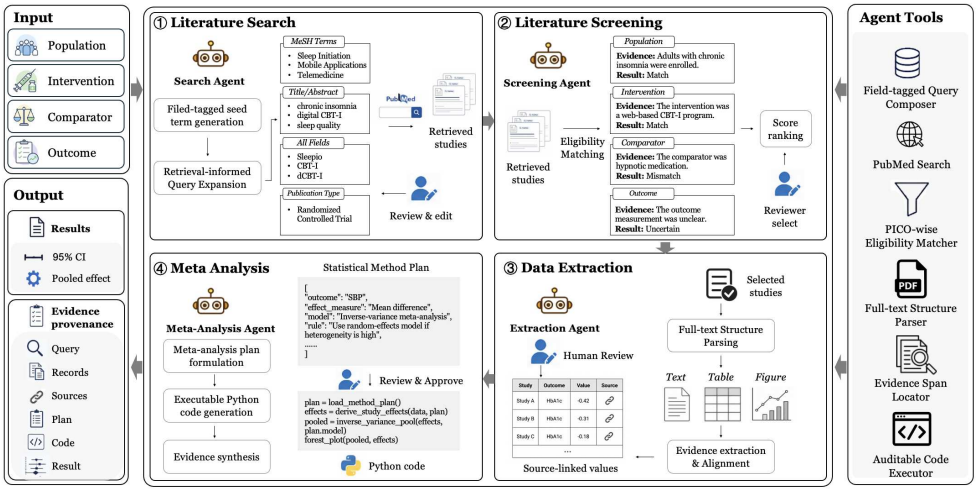


Fig. 1 Overview of the AutoMeta workflow. Research questions are organized using PICO (population, intervention, comparator and outcome), illustrated here, or PECO (population, exposure, comparator and outcome). Dedicated agents perform literature search, study screening, targeted data extraction and statistical synthesis. Supporting tools provide query construction and expansion, eligibility assessment, document parsing, evidence localization and statistical computation. Researchers can review and edit intermediate outputs and approve analysis plans. Outputs include pooled effect estimates, 95% confidence intervals and associated provenance records.

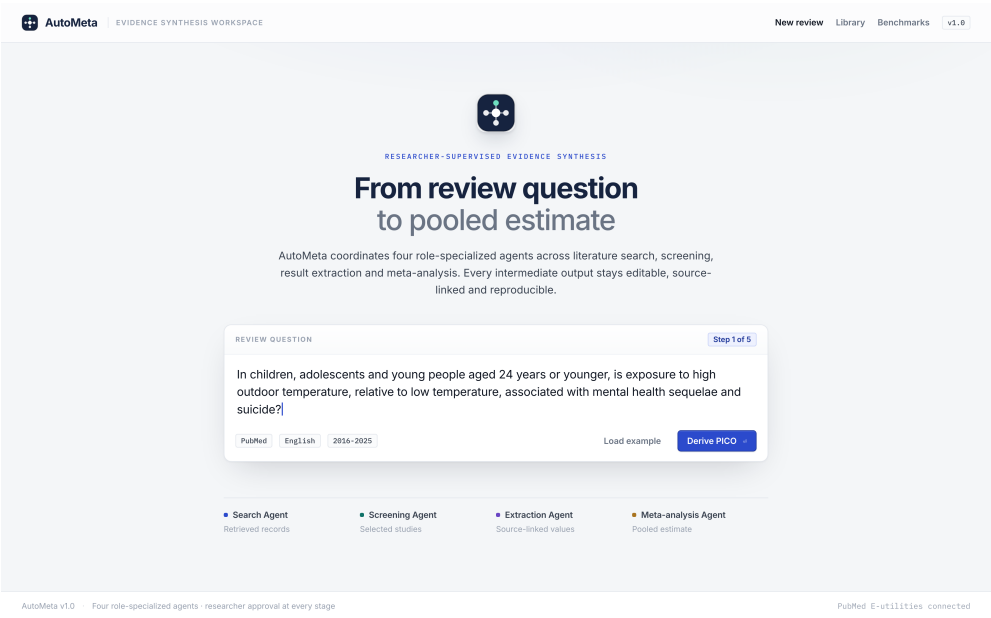


Fig. 2 AutoMeta web application. The unified interface allows researchers to enter a research question and initiate the meta-analysis workflow.

combined using Boolean operators appropriate to the research question [24]. The agent also assigns PubMed field tags to specify where terms are matched, such as `[Title/Abstract]` for titles and abstracts and `[Publication Type]` for applicable publication-type restrictions. PICO/PECO constrains the concepts to be searched, while field tags specify where those concepts are matched within bibliographic records, translating the requirements of intervention and exposure studies into explicit query structures.

After executing the initial query, the agent reads the titles, abstracts and available indexing information of the retrieved publications and uses these materials together with the original research question to refine the query. It identifies alternative names and measurement terminology for relevant concepts, supplements or revises the initial terms, and assembles a final query with explicit field tags. This retrieval-informed query expansion incorporates expressions absent from the initial question formulation, addressing variation in the terminology used for environmental exposures, nutritional factors and health behaviours in public health literature [22]. The selection of expansion terms remains constrained by the prespecified research question, while query reconstruction retains the logical relationships between concepts and the matching fields assigned to terms. Initial query construction, feedback-based refinement, final retrieval and record structuring can be represented as

$$y_{\text{search}} = \text{Agent}_{\text{search}}\left(q; \mathcal{I}_{\text{search}}, \mathcal{T}_{\text{search}}\right) = (Q_f, R_f). \tag{2}$$

Here, Q_f denotes the final query actually executed, and R_f denotes the structured candidate set obtained from that query. The terms $\mathcal{I}_{\text{search}}$ and $\mathcal{T}_{\text{search}}$ denote the search-stage instructions and available tool set, respectively. $\text{Agent}_{\text{search}}$ represents the complete search-stage process, including query construction, retrieval-informed refinement, final retrieval and record parsing. This representation does not prescribe the number of model or tool calls.

Researchers can review and modify the refined query before the final search is run. The search agent invokes the PubMed search tool to execute the refined query and parses the returned records into a structured candidate set containing PubMed identifiers (PMIDs), titles, abstracts, authors, publication years, journals and publication types where available. The executed query Q_f and candidate set R_f jointly constitute the search-stage output y_{search}. Retaining the executed query supports reporting the search strategy as run, consistent with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for literature searches (PRISMA-S) [26]. The structured records are passed to the screening agent for subsequent eligibility assessment and prioritization for review.

2.3 PICO/PECO-based eligibility assessment and ranking

The screening agent receives the shared research question q and the structured candidate records R_f returned by the search stage. It assesses eligibility for each candidate across the PICO/PECO dimensions using information in the title and abstract, and produces a prioritized review queue (Fig. 1). For candidate record $r_i \in R_f$, let $\mathcal{D} = \{P, I/E, C, O\}$ denote the set of eligibility dimensions [2, 22]. The screening process and score aggregation can be represented as

$$\begin{aligned} y_{\text{screen}} &= \text{Agent}_{\text{screen}}\left(q, R_f; \mathcal{I}_{\text{screen}}\right), \\ S_i &= \sum_{d \in \mathcal{D}} m_{i,d}, \qquad m_{i,d} \in \{-1, 0, 1\}. \end{aligned} \tag{3}$$

Here, $\mathcal{I}_{\text{screen}}$ denotes the screening-stage instructions, and y_{screen} denotes the structured screening output. $\text{Agent}_{\text{screen}}$ represents the complete screening process, including dimension-specific eligibility assessment, score aggregation and ranking. The term $m_{i,d}$ denotes the matching score for candidate record r_i on dimension d, and S_i is the sum across the four dimensions.

Dimension-specific assessments use three states: `MATCH`, `UNCERTAIN` and `MISMATCH`, corresponding to scores of 1, 0 and -1, respectively. The agent assigns `MATCH` when information in the title or abstract supports eligibility on a given dimension, `MISMATCH` when there is evidence that the requirement is not met, and `UNCERTAIN` when the available information is insufficient to support either judgement. It separately assesses the target population, intervention or exposure, comparator and outcome, and records the text supporting each judgement. This structured matching approach evaluates eligibility against explicit requirements for each dimension, helping distinguish public health studies that address similar topics but differ in their populations, exposure definitions or outcome measurements. When relevant information is incompletely reported in titles and abstracts, `UNCERTAIN` preserves this uncertainty in eligibility assessment, distinguishing missing information from clear failure to meet the inclusion requirements.

The screening agent sums the dimension-specific scores with equal weight and ranks candidate records in descending order of the total score S_i. The screening output y_{screen} contains basic bibliographic information, dimension-specific judgements and supporting text, total scores and ranks. The ranking establishes priorities for review. Researchers confirm study selection through subsequent review and pass the selected studies to the extraction stage.

2.4 Targeted data extraction and evidence provenance

The extraction agent receives the shared research question q, the full texts of selected studies and prespecified extraction targets. It organizes information from the source reports that matches the target analytical conditions into a study-level data table (Fig. 1). Let F_i denote the full text of study i and g_{ij} its jth extraction target. Document parsing and target-specific extraction can be represented as

$$y_{\text{extract}} = \text{Agent}_{\text{extract}}\left(q, \{F_i\}, \{g_{ij}\}; \mathcal{I}_{\text{extract}}, \mathcal{T}_{\text{extract}}\right). \tag{4}$$

Here, g_{ij} specifies the required field and the applicable outcome, comparison, time point, subgroup and adjustment model. The output y_{extract} denotes the resulting structured study-level data table. The terms $\mathcal{I}_{\text{extract}}$ and $\mathcal{T}_{\text{extract}}$ denote the extraction-stage instructions and available tool set, respectively. $\text{Agent}_{\text{extract}}$ represents the complete extraction process, including document parsing and the reading, matching and integration of source information under the constraints of the research question and extraction targets.

The extraction agent first invokes Docling [27] to parse document layout and table structure, organizing the full text into addressable text passages, tables and figure elements while retaining their locations within the document. It uses multimodal reading to interpret information contained in figures. Guided by extraction target g_{ij}, the agent then extracts complementary information from the text, tables and figures, links each value to the corresponding outcome, comparison, time point, subgroup and adjustment model, and integrates information across sources when the analytical conditions match. For example, an effect estimate, its confidence interval, the sample size and the model specification may be reported at different locations. The agent combines these elements in a single extraction record when they refer to the same study and analytical target. Pairing values with their analytical context addresses a recurring difficulty in public health evidence extraction: identifying the intended result when multiple analyses are reported and numerical values and their defining conditions appear in different parts of the article [2, 25, 28].

When the extraction agent cannot locate supporting information for a target field in the available document content, it is instructed to return `NOT FOUND`. This records that supporting information was not located during the current extraction, without

inferring values unsupported by the source. The agent produces a structured study-level table y_{extract} containing study identifiers, target conditions, extracted values, statistical measure types and available source locations, including page numbers and references to tables, figures or text passages. Researchers can review and revise the table, after which the confirmed data are passed to the meta-analysis agent for analysis planning and statistical synthesis.

2.5 Statistical analysis planning and execution

The meta-analysis agent receives the shared research question q and the study-level data table x_{meta} confirmed by researchers. It develops a statistical analysis plan, generates an executable script that calls predefined statistical functions, and invokes a statistical execution tool to perform numerical calculations (Fig. 1). Analysis planning, script generation and execution can be represented as

$$\begin{aligned} y_{\text{meta}} &= \text{Agent}_{\text{meta}}\left(q, x_{\text{meta}}; \mathcal{I}_{\text{meta}}, \mathcal{T}_{\text{meta}}\right) \\ &= (\bar{\pi}, c, z), \\ z &= \text{Execute}(c, x_{\text{meta}}). \end{aligned} \tag{5}$$

Here, $\bar{\pi}$ denotes the analysis plan developed by the agent and confirmed after researcher review and any necessary revisions. The term c is an executable Python script that invokes predefined statistical functions, and z denotes the results returned by the statistical executor. The stage output y_{meta} comprises the confirmed analysis plan, execution script and statistical results. The terms $\mathcal{I}_{\text{meta}}$ and $\mathcal{T}_{\text{meta}}$ denote the meta-analysis-stage instructions and available tool set, respectively. $\text{Agent}_{\text{meta}}$ represents the complete stage process, with script generation and execution proceeding only after researchers confirm the analysis plan. Execute denotes execution of the script by the statistical tool.

When formulating the plan, the meta-analysis agent combines the research question with the analytical context of the input data to specify the target outcome, effect measure, comparison direction, required data transformations, synthesis model and requested outputs. For data requiring transformation, the agent specifies the target analytical scale, required inputs and applicability conditions in the plan and incorporates these requirements into subsequent calculations. This design translates differences in effect measures, reporting scales and comparison conditions across public health studies into explicit statistical processing requirements, aligning data preparation and synthesis with the prespecified analytical objective.

The statistical function library implements relevant methods described in the *Cochrane Handbook for Systematic Reviews of Interventions* [2]. Effect estimation, data transformation, pooling and uncertainty calculations supported by the workflow are encapsulated as callable functions with defined inputs, parameters and outputs. All functions in the library underwent testing and manual review to check consistency between their computational implementations and the underlying statistical methods.

Once researchers confirm the analysis plan, the meta-analysis agent generates an execution script that selects the appropriate statistical functions, organizes their calling sequence and supplies the required data fields and analysis parameters. The agent then invokes the statistical execution tool to run the script. The functions calculate study-level effects, pooled effects, 95% confidence intervals and heterogeneity statistics. This design combines the agent's planning and tool orchestration with tested implementations of established statistical methods. Statistical results are produced through explicit computation, reducing the risk of errors from directly generating numerical answers or constructing calculation formulas ad hoc.

After execution, the meta-analysis agent reads the results returned by the statistical tool and organizes the outputs specified in the plan, including pooled effects, 95% confidence intervals, heterogeneity measures and forest plots. The confirmed analysis

plan, execution script and result files together form the stage output y_{meta}, preserving the links among input data, analysis settings and computed results so that researchers can review and rerun the analysis.

2.6 Use of generative AI in manuscript preparation

Codex (OpenAI) was used to assist with manuscript drafting and language editing. The authors reviewed the text and take full responsibility for its content.

3 Benchmark construction and experimental design

We evaluated AutoMeta using stage-specific comparisons and end-to-end tasks on AutoMetaBench, alongside comparisons of model performance and cost. Qwen3.7-Plus was the base large language model for AutoMeta.

3.1 Construction of AutoMetaBench

To evaluate automated methods across the stages of public health meta-analysis, we constructed AutoMetaBench from published systematic reviews and their reported meta-analysis results. We organized review questions, included studies, study-level data and pooled results into linked evaluation resources. AutoMetaBench comprises 27 reviews spanning environmental health (9 reviews), nutrition (8) and health behaviours (10). These reviews examine associations between exposures and health outcomes, as well as intervention effects. The benchmark contains 421 included-study records, 4,164 data fields and 95 effect units, providing reference data for study identification, targeted data extraction and statistical synthesis, respectively. Each effect unit represents a pooled result for a specific outcome and analytical target in a source review, comprising the pooled effect estimate, its 95% confidence interval bounds and the corresponding study-level input data.

For screening, we constructed a fixed pool of 3,000 candidate records for each review. We first included the review's reference studies, then added records retrieved in a previous AutoMeta search, and finally supplemented the pool with publications cited by the review and its included studies.

Extraction targets and reference values were drawn from study characteristics tables, results tables and forest plots in the source reviews. Each target was matched to its reference value by study, outcome, comparison and time point. The meta-analysis task used a reviewed study-level results table for each effect unit as input.

3.2 Stage-specific evaluation

All four stages were evaluated independently using their prescribed inputs. Within each task, methods received identical materials and were evaluated against the same reference standards, providing a common basis for comparing performance.

Literature search. We evaluated literature search on the 27 systematic reviews in AutoMetaBench, using their 421 included-study records as the reference set. We compared AutoMeta with AutoBool, a task-specific method optimized for Boolean queries [17], and TrialMind, a workflow for multistage clinical evidence synthesis [19]. A single-agent comparator used the same model backbone and tools as AutoMeta. Search performance was assessed using recall, the number of matched reference records and the number of retrieved candidate records. Recall was micro-averaged across reviews. Screening burden per true study (SBTS) quantified the number of retrieved records per reference study identified and was summarized as the arithmetic mean of per-review values. The 95% percentile bootstrap intervals for mean SBTS were based on 10,000 resamples of the per-review SBTS values.

Study screening. For each review, all methods ranked the same pool of 3,000 candidate records containing the reference studies. Comparators included MedCPT [29], a model for ranking biomedical literature by semantic relevance; TrialMind, which

supports evidence screening [19]; and a single-agent system using the same model backbone and tools as AutoMeta. We measured the proportion of reference studies appearing among the top 10, 30 and 50 ranked records (Recall@10, Recall@30 and Recall@50). Screening workload was assessed using k_{95}, the minimum screening depth required to recover at least 95% of the reference studies, and work saved over sampling at 95% recall (WSS@95) [30]. For the fixed candidate pool, WSS@95 was calculated as $0.95 - k_{95}/3000$. Values of k_{95} and WSS@95 were summarized as means across reviews with 95% confidence intervals.

Targeted data extraction. We provided each method with the full texts of included studies and prespecified extraction targets, and compared its outputs with reference values for 4,164 evaluation fields. Comparators included TrialMind, which supports targeted data extraction [19]; CrossCritique, which combines independent extraction by two models with cross-review of discrepancies [18]; and a single-agent comparator using the same model backbone and tools as AutoMeta. Structured outputs were compared with reference values field by field. Extraction performance was reported as the number of correctly extracted fields and the proportion of evaluated fields that were correct.

Statistical synthesis. We evaluated statistical synthesis across 95 effect units in AutoMetaBench, providing verified study-level results tables to AutoMeta and a single-agent comparator using the same model backbone and tools. We assessed agreement with the original reviews at three levels: effect direction, pooled effect estimates and 95% confidence intervals. Agreement in effect direction assessed whether a system recovered the same direction of association or intervention effect as the reference analysis. Each outcome was assessed separately against the results reported in the source reviews.

3.3 End-to-end evaluation

We evaluated AutoMeta's ability to conduct meta-analysis end to end, from a research question to statistical synthesis. Unlike the preceding stage-specific evaluations with predefined inputs, each task here began with a PICO/PECO research question and proceeded through the four stages. Literature retrieval, study screening, data extraction and statistical analysis produced a set of included studies, a pooled effect estimate and its 95% confidence interval bounds.

We selected one primary outcome from each of the 27 reviews in AutoMetaBench to construct 27 evaluation tasks. The included studies and corresponding synthesis results of each review served as reference standards. A researcher with relevant domain expertise used the AutoMeta web application to submit questions, operate the four-stage workflow and review its outputs. The single-agent comparator used the same underlying model, Qwen3.7-Plus, and the same tools, and both systems followed the same task execution rules. Total elapsed time was measured from question submission to final output and included researcher interactions, review and waiting time. Included-study recall was defined as the proportion of reference studies retained after screening and averaged with equal weight across reviews. Total elapsed time and included-study recall were summarized as means and standard deviations across the 27 tasks.

Statistical synthesis was evaluated using cumulative reproduction criteria. The first criterion required a valid pooled estimate and concordant effect direction, where applicable. The second additionally required the pooled estimate to have an original-scale relative error of no more than 5%. The third additionally required each 95% confidence interval bound to satisfy the same tolerance. All reproduction rates used 27 tasks per method as the denominator.

3.4 Model configurations and cost assessment

We compared performance across the four stages and application programming interface (API) costs for four model backbones using the same AutoMeta workflow: Qwen3-8B, Qwen-Turbo, Qwen3.7-Flash and Qwen3.7-Plus. Performance was assessed using search recall, Recall@50, extraction accuracy and pooled-effect accuracy. Costs were API charges per review, reported in renminbi (RMB) as mean ± standard deviation.

4 Results

4.1 AutoMeta improves reference study coverage while reducing downstream screening burden

AutoMeta retrieved 392 reference records, achieving 93.1% recall, 43.9 percentage points higher than AutoBool, the strongest comparator by recall (49.2%; Fig. 3a). At the same time, AutoMeta returned 9,398 candidate records, 59.7% fewer than AutoBool's 23,300. This comparison shows that AutoMeta covered more reference evidence with a smaller candidate set, reducing the number of records requiring downstream screening. Compared with TrialMind, AutoMeta achieved 56.3 percentage points higher recall (93.1% versus 36.8%) and 80.7% lower mean SBTS (24.24 versus 125.44; Fig. 3a,b).

Relative to the single agent with the same backbone and tools, AutoMeta increased the number of retrieved reference records from 191 to 392, an increase of 201 records. Mean SBTS fell from 52.66 to 24.24, a 54.0% reduction (Fig. 3a,b). The joint illustration of recall and SBTS clarifies these two dimensions (Fig. 3c). Higher positions indicate greater reference coverage, whereas positions farther left indicate a lower screening burden per reference study. AutoMeta's aggregate results improved along

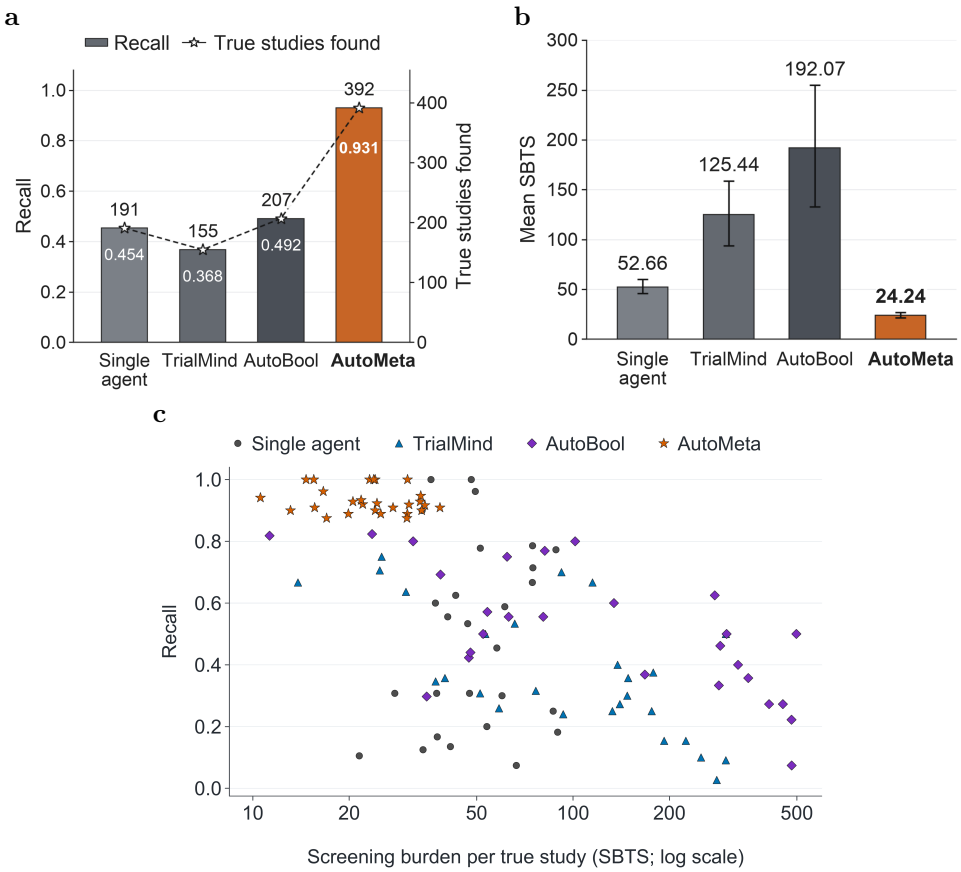


Fig. 3 Search coverage and screening burden. a, Micro-averaged recall (bars) and matched reference records (stars) across 27 reviews containing 421 reference records. **b**, Arithmetic mean of per-review screening burden per true study (SBTS). **c**, Joint recall–SBTS illustration with a logarithmic horizontal axis. Higher recall and lower SBTS indicate better performance. Error bars in b are 95% percentile bootstrap intervals for the mean SBTS, based on 10,000 resamples of the per-review SBTS values.

both dimensions relative to all comparators, indicating a more favourable combination of evidence coverage and screening burden.

4.2 AutoMeta prioritizes reference studies and substantially reduces screening workload

AutoMeta achieved the highest recall of reference studies at all three screening depths (Fig. 4a). Its Recall@10 reached 57.7%, compared with 27.1% for TrialMind,

the strongest comparator across the Recall@k metrics. This gain of 30.6 percentage points, corresponding to approximately 2.13 times TrialMind's recall, indicates that researchers could identify a larger proportion of reference studies within a small initial screening budget. When screening extended to the first 30 and 50 records, AutoMeta's recall reached 86.2% and 94.8%, respectively, compared with 57.8% and 66.7% for TrialMind.

AutoMeta achieved the lowest mean k_{95}, at 138.48 records. Compared with the single-agent system using the same model backbone, which was the strongest comparator on this metric at 422.19 records, AutoMeta reduced the mean screening depth by 283.71 records, or 67.2% (Fig. 4b). AutoMeta therefore reached the same target recall with approximately one-third of the screening workload, supporting the value of a specialized screening workflow for improving review efficiency. Although TrialMind achieved higher Recall@50 than the single-agent system (66.7% versus 48.7%), its mean k_{95} was higher. This indicates that stronger coverage near the top of the list does not necessarily translate into a lower screening workload at a high recall target. AutoMeta maintained an advantage on both measures.

Correspondingly, AutoMeta achieved a WSS@95 of 90.4%, exceeding the single-agent system (80.9%) and TrialMind (72.0%) by 9.5 and 18.4 percentage points, respectively (Fig. 4c). The mean screening depth needed to reach 95% reference-study recall corresponded to only approximately 4.6% of the candidate pool, allowing researchers to recover most reference studies within a small portion of the ranked list and substantially reducing the number of records requiring review.

a

Metric	Single agent	MedCPT	TrialMind	AutoMeta	Fold change
Recall@10	0.084	0.026	0.271	**0.577**	**2.13×**
Recall@30	0.317	0.118	0.578	**0.862**	**1.49×**
Recall@50	0.487	0.185	0.667	**0.948**	**1.42×**

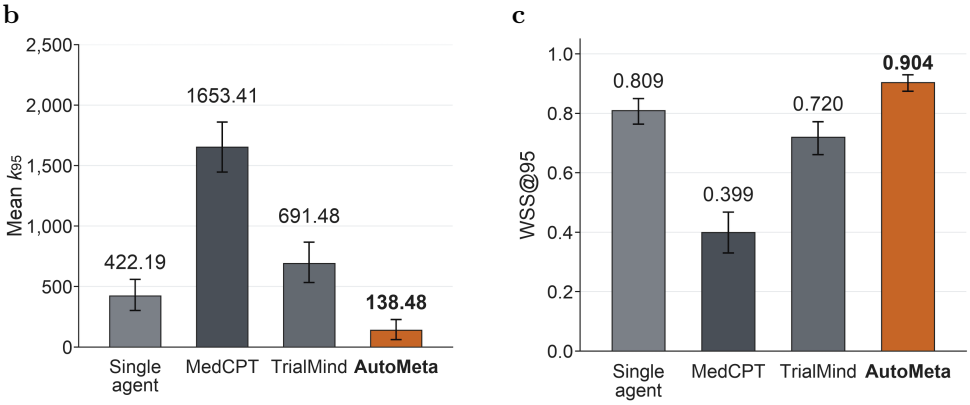


Fig. 4 Reference study prioritization and screening workload. a, Recall among the top 10, 30 and 50 ranked records. Fold changes compare AutoMeta with TrialMind. **b**, Mean minimum screening depth required to recover at least 95% of reference studies (k_{95}). **c**, Work saved over sampling at 95% recall (WSS@95). For each review, all methods ranked the same pool of 3,000 candidate records containing the reference studies. WSS@95 was calculated as $0.95 - k_{95}/3000$. Bars in b and c show means across reviews; error bars indicate 95% confidence intervals. Lower k_{95} and higher WSS@95 indicate lower screening workload.

4.3 AutoMeta supports accurate extraction of results from public health studies

AutoMeta correctly extracted 3,896 of 4,164 fields, achieving an accuracy of 93.6%, the highest among all methods evaluated (Fig. 5). Compared with TrialMind, the best-performing comparator, AutoMeta correctly extracted 384 additional fields, increasing accuracy by 9.2 percentage points. It also correctly extracted 516 more fields than CrossCritique, a gain of 12.4 percentage points, and 1,149 more fields than the single-agent comparator, a gain of 27.6 percentage points. These results show that AutoMeta can accurately extract prespecified target data from study reports.

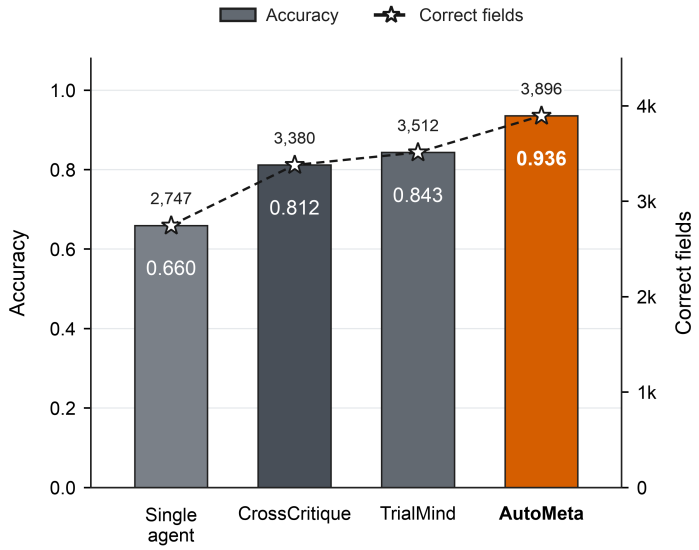


Fig. 5 Accuracy of targeted data extraction. Bars show field-level accuracy (left axis), and stars show the number of correctly extracted fields (right axis).

4.4 AutoMeta enables accurate quantification and statistical synthesis of public health evidence

Across 95 effect units, AutoMeta achieved directional agreement of 96.8%, compared with 78.9% for the single-agent comparator, an improvement of 17.9 percentage points (Fig. 6). When evaluating pooled effect estimates, AutoMeta maintained a match rate of 96.8%, whereas the single-agent comparator achieved 73.7%, widening the gap to 23.2 percentage points and showing that AutoMeta's advantage extended to the calculation of specific effect values.

For uncertainty estimation, AutoMeta achieved a 95% confidence interval match rate of 95.8%, substantially higher than the comparator's 54.7%, with the gap increasing further to 41.1 percentage points. As evaluation progressed from effect direction to effect magnitude and its uncertainty, the single-agent comparator's match rate declined, whereas AutoMeta maintained agreement above 95% at all three levels.

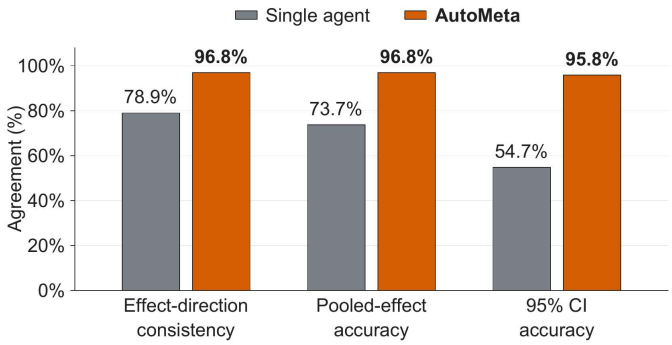


Fig. 6 Agreement of statistical synthesis results with reference analyses. Bars show effect-direction consistency, pooled-effect accuracy and 95% confidence interval accuracy across 95 effect units. AutoMeta and the single-agent comparator received the same verified study-level inputs and used the same model backbone and tools. Each outcome was assessed separately against the results reported in the source reviews.

AutoMeta can therefore reproduce reference pooled effects and their 95% confidence interval bounds from verified study-level data.

4.5 AutoMeta performance and API costs across model backbones

Qwen3.7-Plus achieved the highest scores on all four metrics (Table 1), while Qwen3.7-Flash maintained similar performance at substantially lower cost: its scores were 4.7–5.2 percentage points below those of Plus, and its mean API cost per review was RMB 2.25, approximately 90.0% lower than the RMB 22.40 for Plus. Among the three API models, Qwen-Turbo had the lowest mean cost; however, Flash cost only RMB 0.22 more per review and scored higher on all four metrics. These results inform model selection within the evaluated AutoMeta workflow: Plus may be preferred when performance is the priority, whereas Flash offers a favourable balance between performance and API costs.

Table 1 Performance and API costs of AutoMeta across model backbones.

Backbone	Search recall	Recall@50	Extraction accuracy	Pooled-effect accuracy	Cost, RMB (mean ± SD)
Qwen3-8B	0.637	0.477	0.690	0.774	-
Qwen-Turbo	0.746	0.672	0.765	0.879	2.03 ± 0.62
Qwen3.7-Flash	0.881	0.900	0.889	0.916	2.25 ± 0.66
Qwen3.7-Plus	**0.931**	**0.948**	**0.936**	**0.968**	22.40 ± 6.75

Note. Costs are API charges per review, reported in RMB as mean ± standard deviation (SD). Higher values indicate better performance on all four metrics; bold values identify the highest score in each performance column.

4.6 AutoMeta supports efficient end-to-end execution of public health meta-analyses

Across the 27 end-to-end tasks, AutoMeta required less time on average to complete the end-to-end workflow than the single-agent comparator. AutoMeta completed the workflow in 26.00 ± 4.54 min, compared with 30.00 ± 7.89 min for the single-agent comparator (mean ± standard deviation; Fig. 7a). This corresponded to a 4.00-min reduction in mean total elapsed time (13.3%).

AutoMeta achieved a mean included-study recall of 89.33 ± 5.69%, compared with 21.20 ± 8.05% for the single-agent comparator, a difference of 68.13 percentage points corresponding to approximately 4.21 times the comparator's recall (Fig. 7b). AutoMeta thus retained a more complete set of reference studies through search and screening.

Under the cumulative reproduction criteria, AutoMeta met the applicable direction criterion in 92.59% of tasks, compared with 62.96% for the single-agent comparator. With an additional requirement of no more than 5% relative error in the pooled effect, cumulative reproduction rates were 92.59% and 11.11%, respectively. Further requiring both 95% confidence interval bounds to satisfy the same tolerance yielded cumulative reproduction rates of 85.19% and 7.41%, respectively, a difference of 77.78 percentage points (Fig. 7c). AutoMeta reproduced the reference pooled effects and

their 95% confidence interval bounds in a larger proportion of end-to-end tasks than the single-agent comparator.

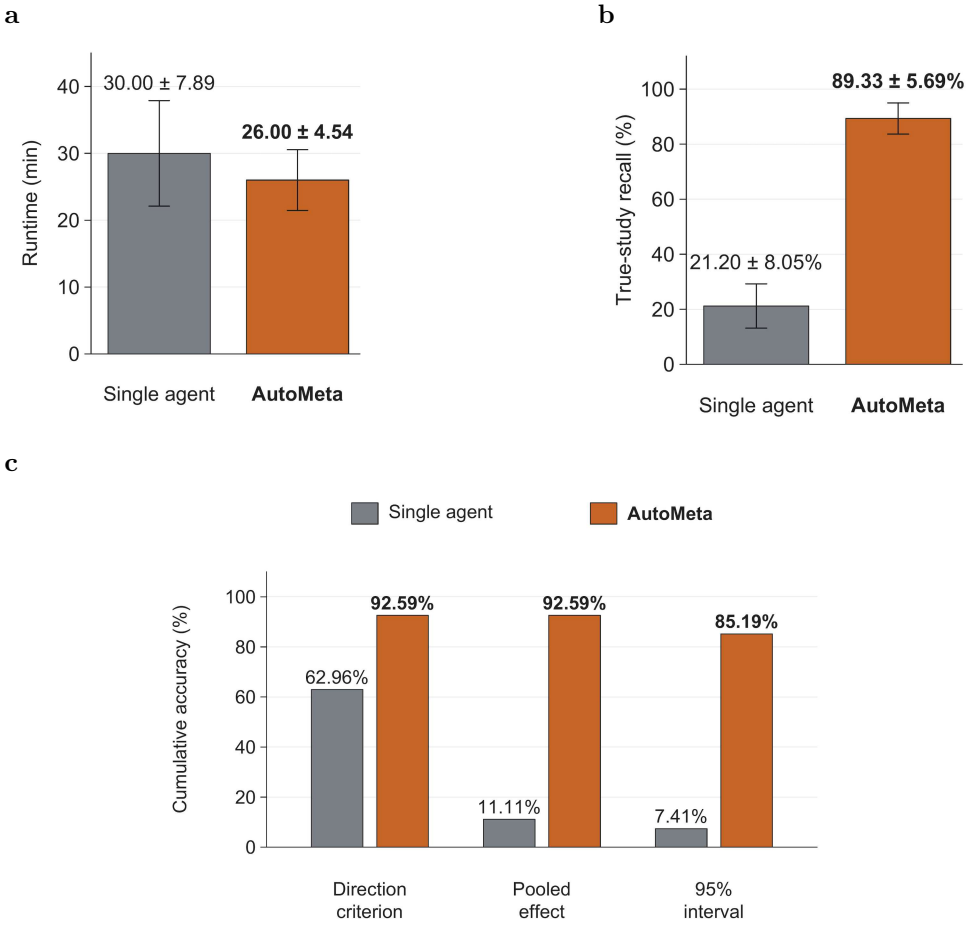


Fig. 7 End-to-end performance across 27 public health meta-analysis tasks. a, Mean total elapsed time. **b**, Mean post-screening recall of reference studies. Error bars in **a** and **b** show the standard deviation across 27 tasks; recall standard deviations are expressed in percentage points. **c**, Cumulative reproduction rates. The first criterion requires a valid pooled estimate and concordant direction. The second additionally requires the pooled estimate to have an original-scale relative error of at most 5%. The third additionally requires each 95% confidence interval bound to meet the same tolerance. All rates use 27 tasks per method as the denominator.

In an illustrative analysis of heat exposure and mental health among children and adolescents [23], AutoMeta synthesized 10 effect estimates from eight studies. The

pooled risk ratio was 1.14 (95% confidence interval, 1.08–1.20; Fig. 8b). The web application displayed the execution trace (Fig. 8a), forest plot and statistical diagnostics, and retained data provenance, the analysis plan and executable code. These outputs illustrate the complete progression from a research question to a verifiable synthesis.

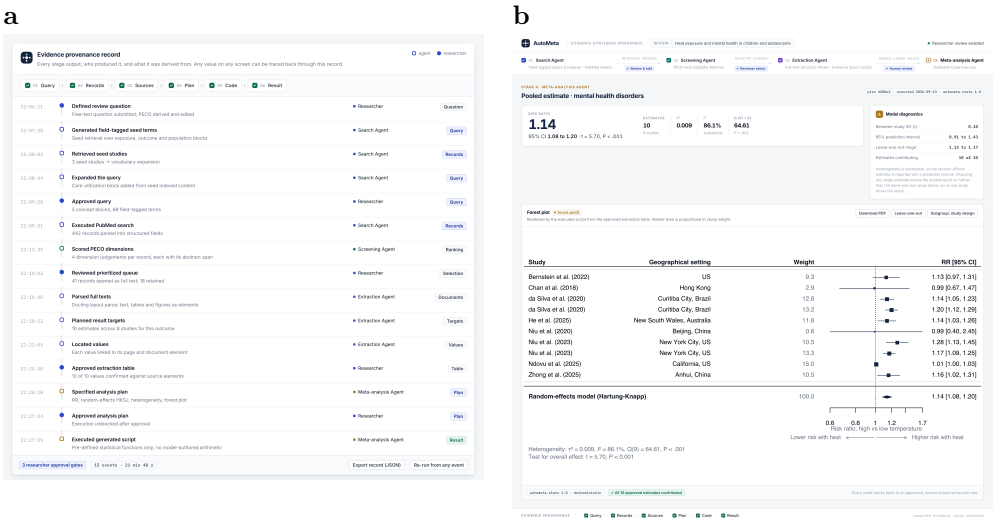


Fig. 8 End-to-end execution and final synthesis in the case study. a, Execution trace linking the four stages, responsible agents, researcher decisions and outputs. **b**, Final synthesis of 10 effect estimates from eight studies, yielding a pooled risk ratio of 1.14 (95% confidence interval, 1.08–1.20), with a forest plot and statistical diagnostics.

5 Discussion

In this study, we developed AutoMeta, an agentic workflow for public health meta-analysis, and constructed AutoMetaBench for its evaluation. Stage-specific and end-to-end evaluations on AutoMetaBench demonstrate that AutoMeta can integrate evidence identification and quantitative synthesis across the full workflow from a research question to statistical synthesis, highlighting its value for public health meta-analysis.

AutoMeta's advantages in search and screening were reflected in more complete coverage of reference studies and fewer records requiring review to achieve high recall

(Figs. 3 and 4). Environmental exposures, nutritional factors and health behaviours are described in diverse ways in public health research, while populations, exposure definitions and outcome criteria also vary considerably [22, 24]. These differences make it harder for automated tools to identify relevant publications and assess eligibility, increasing the risk of missing studies or mistaking topical similarity for eligibility. AutoMeta addresses these challenges by using retrieval feedback to expand terminology, together with field tags and PICO/PECO definitions to structure searches and eligibility assessments. On the same public health benchmark, TrialMind [19], a workflow developed for clinical evidence synthesis, achieved 36.8% recall, compared with 93.1% for AutoMeta (Fig. 3). Reaching 95% recall required review of approximately 138 records on average, 67.2% fewer than for the single-agent comparator (Fig. 4). These findings show that AutoMeta concentrates reference studies near the top of the review queue, reducing screening requirements while retaining broad coverage of reference evidence.

Accurate handling of complex quantitative results is another major strength of AutoMeta. Its targeted data extraction accuracy reached 93.6%, exceeding that of the evaluated comparators (Fig. 5). With the model backbone and tools held constant, AutoMeta also achieved higher agreement with reference pooled effects and 95% confidence intervals than the single-agent comparator (Fig. 6). These gains are particularly relevant to public health meta-analysis: a single study may report results for multiple subgroups, time points and adjustment models, and selecting the wrong effect estimate can shift the synthesis away from the prespecified research question [2, 25]. Guided by explicit extraction targets, AutoMeta extracts and aligns numerical data across tables, text and figures. It then develops a statistical analysis plan and generates code that calls tested statistical functions for effect-size transformation and synthesis, keeping target identification and statistical execution aligned with the same analytical requirements. The larger gain in confidence interval reproduction

than in point-estimate reproduction further highlights its strength in statistical tasks involving uncertainty estimation.

The advantages in evidence coverage, result reproduction and execution efficiency persisted when evaluation was extended to the complete workflow, from the research question to pooled results. AutoMeta met the cumulative reproduction criteria for direction, where applicable, the pooled effect and both 95% confidence interval bounds in 85.2% of tasks, compared with 7.4% for the single-agent comparator. Mean total elapsed time decreased from 30 minutes for the comparator to 26 minutes for AutoMeta (Fig. 7). These differences were observed with each stage using the actual outputs of preceding stages, so the final reproduction rate captured the combined performance of study identification, data extraction and statistical synthesis. With the model backbone and available tools held constant, this comparison supports the value of the specialized workflow in the evaluated tasks.

This study also has several limitations. LLMs may make errors when assessing eligibility, matching data to extraction targets or selecting analyses, so key methodological decisions still require researcher judgement [31]. Searches are currently restricted to PubMed and may omit relevant studies indexed elsewhere [32]; the workflow also does not yet include risk-of-bias assessment or certainty-of-evidence rating [2, 33]. Applicability to other topics, study designs and settings remains to be tested. In addition, total elapsed time includes researcher interaction, review and waiting, and cannot be directly translated into savings in person-hours. The benefits of the system in routine research therefore need to be assessed in user studies.

Future work will expand the number of systematic reviews, included studies and evaluation tasks in AutoMetaBench, incorporating data from a broader range of public health and health research domains to improve the benchmark's coverage and representativeness. We will also conduct prospective end-to-end evaluations within researchers' ongoing meta-analysis projects, involving users with different professional

backgrounds and levels of experience. These evaluations will assess the quality of the final synthesis, researcher time spent on the task, total elapsed time and system usability. Planned extensions include searches across multiple databases, integration of risk-of-bias tools and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework [33–35], and evaluation of more capable or more efficient model backbones. The aim is to translate the performance gains observed in the benchmark into demonstrable improvements in synthesis quality and efficiency in routine research, enabling more timely synthesis and updating of public health evidence to inform disease prevention, health promotion and decision-making [5].

6 Conclusions

AutoMeta integrates semantic interpretation of public health research questions, traceable data extraction and statistical computation using dedicated tools into a verifiable four-stage workflow. Stage-specific and end-to-end evaluations show that this workflow improves study identification and synthesis performance while reducing mean total elapsed time on the evaluated tasks. Together, AutoMeta and AutoMetaBench provide a reusable foundation for conducting and evaluating public health meta-analyses, with the potential to improve synthesis efficiency and support timely evidence updates.

List of abbreviations

AI Artificial intelligence

API Application programming interface

CI Confidence interval

GRADE Grading of Recommendations Assessment, Development and Evaluation

LLM Large language model

MeSH Medical Subject Headings

PECO Population, exposure, comparator and outcome

PICO Population, intervention, comparator and outcome

PMID PubMed identifier

PRISMA-S Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for literature searches

RMB Renminbi

RR Risk ratio

SBTS Screening burden per true study

SD Standard deviation

WSS@95 Work saved over sampling at 95% recall

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

AutoMetaBench, the benchmark dataset for evaluating literature search, study screening, targeted data extraction and statistical synthesis, is publicly available at https://huggingface.co/datasets/zhuk26/AutoMetaBench [36]. The AutoMeta source code is publicly available at https://github.com/zhuke26/AutoMeta [37], enabling users to deploy and run the system locally. The AutoMeta web application is available at https://autometa.site [38].

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific funding.

Authors' contributions

K.Z. designed the study and developed the methodology. X.T. and Y.H. collected and analysed the data. T.X. and Y.S. supervised the project and provided critical feedback and guidance throughout the study. K.Z. drafted the manuscript, and T.X. revised and finalised it. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

[1] Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evidence based medicine. 2016;21(4):125–127.

[2] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions. Cochrane; 2024. Version 6.5, updated August 2024. Available from: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current.

[3] Alonso-Coello P, Schünemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. BMJ. 2016;353:i2016. https://doi.org/10.1136/bmj.i2016.

[4] 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. https://doi.org/10.1136/bmjopen-2016-012545.

[5] Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic review: 1. Introduction—the why, what, when, and how. J Clin Epidemiol. 2017;91:23–30. https://doi.org/10.1016/j.jclinepi.2017.08.010.

[6] Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med. 2007;147(4):224–233. https://doi.org/10.7326/0003-4819-147-4-200708210-00179.

[7] O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015;4(1):5. https://doi.org/10.1186/2046-4053-4-5.

[8] Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8(1):163. https://doi.org/10.1186/s13643-019-1074-9.

[9] van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021;3(2):125–133. https://doi.org/10.1038/s42256-020-00287-7.

[10] Scherbakov D, Hubig N, Jansari V, Bakumenko A, Lenert LA. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. J Am Med Inform Assoc. 2025;32(6):1071–1086. https://doi.org/10.1093/jamia/ocaf063.

[11] Luo X, Chen F, Zhu D, Wang L, Wang Z, Liu H, et al. Potential roles of large language models in the production of systematic reviews and meta-analyses. J Med Internet Res. 2024;26:e56780. https://doi.org/10.2196/56780.

[12] Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. J Med Internet Res. 2024;26:e48996. https://doi.org/10.2196/48996.

[13] 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):616–626. https://doi.org/10.1002/jrsm.1715.

[14] Clark J, Barton B, Albarqouni L, Byambasuren O, Jowsey T, Keogh J, et al. Generative artificial intelligence use in evidence synthesis: A systematic review. Res Synth Methods. 2025;16(4):601–619. https://doi.org/10.1017/rsm.2025.16.

[15] Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A survey on large language model based autonomous agents. Front Comput Sci. 2024;18(6):186345. https://doi.org/10.1007/s11704-024-40231-1.

[16] Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. ReAct: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations; 2023. Available from: https://arxiv.org/abs/2210.03629v3.

[17] Wang S, Scells H, Koopman B, Zuccon G. AutoBool: Reinforcement-Learned LLM for Effective Automatic Systematic Reviews Boolean Query Generation. In: Demberg V, Inui K, Marquez L, editors. Proceedings of the 19th Conference of

the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2026. p. 1468–1493. Available from: https://aclanthology.org/2026.eacl-long.68/.

[18] Khan MA, Ayub U, Naqvi SAA, Khakwani KZR, Sipra ZbR, Raina A, et al. Collaborative large language models for automated data extraction in living systematic reviews. J Am Med Inform Assoc. 2025;32(4):638–647. https://doi.org/10.1093/jamia/ocae325.

[19] Wang Z, Cao L, Danek B, Jin Q, Lu Z, Sun J. Accelerating clinical evidence synthesis with large language models. npj Digit Med. 2025;8(1):509. https://doi.org/10.1038/s41746-025-01840-7.

[20] Cao C, Arora R, Cento P, Budak A, Manta K, Farahani E, et al.: Automation of systematic reviews with large language models. Preprint. Version 4, 4 May 2026; first posted 13 June 2025. medRxiv preprint. Available from: https://www.medrxiv.org/content/10.1101/2025.06.13.25329541v4.

[21] Huang YH, Lin YS.: LUMEN: Cost-Transparent Multi-Agent Pipeline for Automated Systematic Review and Meta-Analysis. Preprint, version 1. arXiv preprint arXiv:2606.28362. Available from: https://arxiv.org/abs/2606.28362v1.

[22] Morgan RL, Whaley P, Thayer KA, Schünemann HJ. Identifying the PECO: a framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. Environ Int. 2018;121:1027–1031. https://doi.org/10.1016/j.envint.2018.07.015.

[23] Lai KY, Bauermeister S, Sarkar C. Associations of heat exposure with mental health and suicide in children and adolescents: a systematic review and meta-analysis. npj Ment Health Res. 2026;5(1):7. https://doi.org/10.1038/

s44184-026-00190-w.

[24] Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Metzendorf MI, et al. Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions. Cochrane; 2025. Version 6.5.1, updated March 2025. Available from: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-04.

[25] Li L, Mathrani A, Susnjak T. What level of automation is "good enough"? A benchmark of large language models for meta-analysis data extraction. Res Synth Methods. 2026;17(4):671–692. https://doi.org/10.1017/rsm.2025.10066.

[26] Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10(1):39. https://doi.org/10.1186/s13643-020-01542-z.

[27] Auer C, Lysak M, Nassar A, Dolfi M, Livathinos N, Vagenas P, et al. Docling technical report. arXiv preprint arXiv:240809869. 2024;.

[28] Kataoka Y, Takayama T, Yoshimura K, So R, Tsujimoto Y, Yamagishi Y, et al. Automating the data extraction process for systematic reviews using GPT-4o and o3. Res Synth Methods. 2026;17(1):42–62. https://doi.org/10.1017/rsm.2025.10030.

[29] Jin Q, Kim W, Chen Q, Comeau DC, Yeganova L, Wilbur WJ, et al. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. Bioinformatics. 2023;39(11):btad651. https://doi.org/10.1093/bioinformatics/btad651.

[30] Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association. 2006;13(2):206–219. https://doi.org/10.1197/jamia.M1929.

[31] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. 2025;43(2):1–55. https://doi.org/10.1145/3703155.

[32] Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev. 2017;6(1):245. https://doi.org/10.1186/s13643-017-0644-y.

[33] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026.

[34] Sterne JA, 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;366:l4898. https://doi.org/10.1136/bmj.l4898.

[35] Higgins JP, Morgan RL, Rooney AA, Taylor KW, Thayer KA, Silva RA, et al. A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E). Environ Int. 2024;186:108602. https://doi.org/10.1016/j.envint.2024.108602.

[36] zhuk26. AutoMetaBench [Internet]. Hugging Face; 2026 [cited 15 October 2026]. Available from: https://huggingface.co/datasets/zhuk26/AutoMetaBench.

[37] zhuke26. AutoMeta source code [Internet]. GitHub; n.d. [cited 15 October 2026]. Available from: https://github.com/zhuke26/AutoMeta.

[38] AutoMeta web application [Internet]; n.d. [cited 15 October 2026]. Available from: https://autometa.site.