AUTOFOREST: Automatically Generating Forest Plots from Biomedical Studies with End-to-End Evidence Extraction and Synthesis

Massimiliano Pronesti[1,2], Angelo Miculescu[2], Mohsin Kapdi[2], Paul Flanagan[2], Oisín Redmond[2], Joao Bettencourt-Silva[1], Gurdeep S. Mannu[4], Spiros Denaxas[3,5], Rui Bebiano Da Providencia E Costa[3], Anya Belz[2], Yufang Hou[1,5]

[1]IBM Research [2]Dublin City University [3]UCL [4]University of Oxford
[5]IT:U Interdisciplinary Transformation University Austria

Correspondence: massimiliano.pronesti@ibm.com, yufang.hou@it-u.at

Abstract

Systematic reviews rely on forest plots to synthesise quantitative evidence across biomedical studies, but generating them remains a fragmented and labour-intensive process. Researchers must interpret complex clinical texts, manually extract outcome data from trials, define appropriate interventions and comparators, harmonise inconsistent study designs, and carry out meta-analytic computations—typically using specialised software that demands structured inputs and domain expertise. While recent work has demonstrated that large language models can extract study-level data from unstructured text, no existing system automates the complete pipeline from raw documents to synthesised forest plots. To address this gap, we introduce AUTOFOREST[1,2], the first end-to-end system that generates publication-ready forest plots directly from biomedical papers. Given one or more study papers, AUTOFOREST automatically suggests ICO (Intervention, Comparator, Outcome) elements, extracts outcome data, performs statistical synthesis, and renders the final forest plot. We describe the system architecture, user interface and demonstrate its effectiveness on real-world examples through a user study involving clinicians, showing how AUTOFOREST can accelerate evidence synthesis and substantially lower the barrier to conducting meta-analyses.

1 Introduction

Systematic reviews are essential to evidence-based medicine, combining results from multiple biomedical studies to answer clinical questions with increased statistical power and reduced uncertainty. A key tool in this process is the forest plot, which visualises the estimated effects and confidence intervals across studies, enabling transparent comparison and meta-analysis. Despite its importance, generating forest plots remains a manual and time-consuming task. As seen in Figure 1, researchers must first complete several upstream steps, including extracting outcome data from primary studies—often in unstructured PDF formats—defining the relevant intervention, comparator, and outcome (ICO) elements, and harmonising study designs and measurements. Following these steps, researchers use specialised software such as RevMan (The Cochrane Collaboration, 2020), which is used to conduct statistical synthesis using meta-analytic models and generate the final forest plot. RevMan requires users to manually enter data after converting it to a specific structured format (e.g. event counts and sample sizes for intervention and control groups); it offers no support for unstructured inputs or automation of the upstream steps.

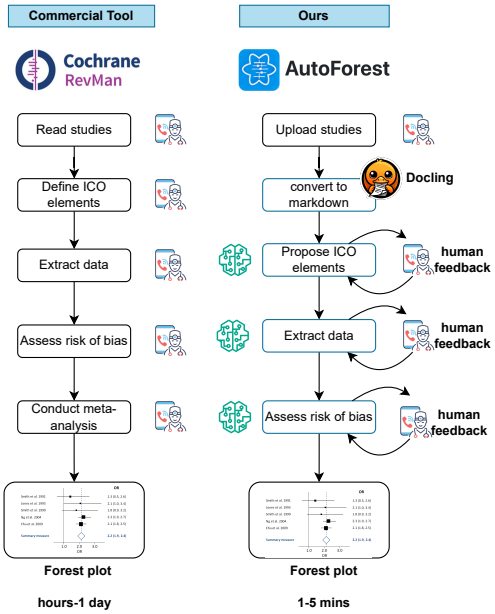

Figure 1: AUTOFOREST automatically generates forest plots in minutes instead of hours with minimal human effort, as opposed to existing commercial tools.

Recent research has made progress in automat-

[1]https://itu-nlp.github.io/projects/autoforest

[2]https://www.youtube.com/watch?v=R6ei97fOyXQ

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 128–137

ing individual components of this workflow. For example, large language models have been used to extract numerical results from clinical trial reports (Yun et al., 2024; Sun et al., 2024; Lai et al., 2025), and to support reasoning about study-level effects (Pronesti et al., 2025b) and risk of bias assessment (Ji et al., 2025; Wang et al., 2025; Pronesti et al., 2026). The approach by Pronesti et al. (2025b) introduces custom models paired with reinforcement learning to identify relevant numerical outcomes and estimate treatment effects across studies. However, it stops short of generating full meta-analytic visualisations, and still relies on manually specified ICO elements and downstream analysis tools. Other efforts employ interactive AI agents to streamline systematic review processes by assisting with study selection and summary generation, but they do not extract numerical effect sizes, perform statistical aggregation, or generate plots (Qiu et al., 2025).

To address these gaps, we introduce AUTOFOREST, the first end-to-end system that produces publication-ready forest plots directly from biomedical papers. Given one or more PDF documents, AUTOFOREST automatically suggests ICO elements, extracts outcome data using a numerical reasoning module, performs statistical synthesis using a random or fixed-effects model, and renders the resulting forest plot—all with minimal user input. Unlike traditional tools such as RevMan which require manual data extraction and entry, AUTOFOREST uses a unified pipeline to automatically extract data directly from unstructured text and perform statistical synthesis and bias assessment.

We demonstrate the system on real-world reviews and show that it can recover forest plots with quality comparable to expert-curated outputs. By automating the full workflow from document ingestion to quantitative synthesis, AUTOFOREST drastically reduces the time and expertise required to perform meta-analysis, making high-quality evidence synthesis more accessible and scalable.

2 Background and Related Work

Systematic Reviews are widely regarded as the gold standard in evidence-based medicine, providing rigorous syntheses of research to guide clinical decision-making (Murad et al., 2016). They aim to address the challenge of staying up-to-date with an ever-growing volume of medical literature by aggregating high-quality evidence for specific clinical questions (Higgins et al., 2024). However, producing a systematic review is time-consuming and costly: a 2019 study estimated that on average it takes 1–2 years and over $141,000 to complete one (Michelson and Reuter, 2019). Given the substantial resources required, there is growing interest in automating various steps of the process (Marshall and Wallace, 2019; Khraisha et al., 2024; Yun et al., 2024; Wang et al., 2024; Pronesti et al., 2025b).

Forest Plots constitute the cornerstone of quantitative synthesis in systematic reviews, visually summarising effect sizes and their associated confidence intervals across studies (Higgins et al., 2024). Each study in a forest plot is represented by a point estimate and confidence interval, facilitating immediate visual assessment of treatment effects and heterogeneity across studies. Despite their widespread use, generating forest plots is often challenging due to inconsistencies in study reporting, the need for manual data extraction from narrative texts and tabular data, and the statistical computations required to produce aggregated results (Pronesti et al., 2025a; Yun et al., 2024; Pronesti et al., 2025b).

Related Works. Current tools for meta-analytic synthesis, such as RevMan (The Cochrane Collaboration, 2020), primarily provide interfaces for manual data entry, requiring structured inputs and extensive human intervention at all upstream stages, including ICO (Intervention, Comparator, Outcome) identification, data extraction and risk of bias estimation. Commercial tools like Comprehensive Meta-Analysis (CMA)[3] offer streamlined GUI workflows for entering structured data and generating forest plots, but rely entirely on manual ICO definition and data entry (Bax et al., 2007; Borenstein, 2022). Web-based review platforms such as DistillerSR (DistillerSR Inc., 2023) provide AI-assisted screening and extraction capabilities, but still require manual validation and external tools for plotting (McMillan et al., 2020; DistillerSR Inc., 2023). EPPI-Reviewer (DistillerSR Inc., 2023) delivers integrated systematic review management with screening, extraction, and basic meta-analysis functions, yet again depends on user-driven ICO configuration and dataset curation (Thomas et al., 2010). In addition, risk of bias assessment (RoB)—an essential component of evidence synthesis—remains largely manual and delegated to separate tools. Established frameworks

[3] https://www.meta-analysis.com

Feature	RevMan	CMA	DistillerSR	EPPI-Reviewer	AutoForest (Ours)
Process raw documents	✗	✗	✓	✓	✓
ICO suggestion	✗	✗	✗	✗	✓
Automatic data extraction	✗	✗	✗	✗	✓
Automatic risk of bias assessment	✗	✗	✗	✗	✓
Forest plot generation	✓	✓	✗	✓	✓
Full pipeline automation	✗	✗	✗	✗	✓

Table 1: Comparison of AUTOFOREST with existing meta-analysis tools across key steps in the evidence synthesis workflow. ✓ indicates full support/automation, ✗ indicates no support.

such as RoB1 (Higgins et al., 2011) and its successor RoB2 (Sterne et al., 2019) provide structured criteria for evaluating bias in randomised trials, but their application is still guided by human reviewers, with current software offering little more than form-based interfaces for entering judgments rather than automating the reasoning process. A summary of these comparisons is provided in Table 1.

3 AUTOFOREST

AUTOFOREST is an interactive system that enables users to extract, validate, and visualise results from primary studies in the form of publication-ready forest plots. It is designed to support researchers conducting systematic reviews by automating time-consuming tasks such as evidence identification, numerical data extraction, risk of bias estimation, and meta-analytic aggregation, while ensuring interpretability and transparency of each step.

The architecture consists of a ReactJS frontend connected to a FastAPI backend. These components communicate through a thin REST API layer, and the backend delegates specific NLP and numerical tasks to model endpoints. Importantly, the system is model-agnostic: while we currently use large language models (LLMs) for evidence extraction and reasoning, the interface supports drop-in replacement of backend components, enabling experimentation with different models or pipelines.

3.1 Overall Workflow

Figure 2 depicts AUTOFOREST's interface. The user workflow is structured into five stages:

1. Upload studies in PDF format.
2. Define the comparison of interest as Intervention – Comparator – Outcome (ICO) triplets.
3. Extract numerical evidence from each study.
4. Assess risk of bias from each study for either RoB1 (Higgins et al., 2011) or RoB2 (Sterne et al., 2019) domains.
5. Generate and download a publication-ready forest plot.

Each of these stages is implemented as a distinct component in the UI and internally, allowing users to iteratively refine earlier decisions.

3.2 Document Upload and Parsing

Users begin by uploading primary studies in PDF format. These typically correspond to clinical trials or observational studies. Upon upload, each document is converted into a structured format using the `docling` library (Auer et al., 2024; Livathinos et al., 2025a,b), through which we render a side-by-side view of the original PDF and its Markdown representation (Figure 2, top right). This dual view allows users to cross-reference raw and structured content when inspecting extracted evidence.

3.3 ICO Selection

To define the scope of the meta-analysis, users must specify an Intervention (I), Comparator (C), and Outcome (O) triplet. Rather than relying on manual entry or simplistic frequency heuristics, AUTOFOREST employs a transformer-based language model to extract candidate ICO elements from each study (Prompt in Appendix B). Let $\mathcal{S} = \{s_1, \ldots, s_n\}$ denote the set of input studies. For each $s_i \in \mathcal{S}$, we define a candidate ICO set as

$$\mathcal{T}_i = \texttt{LLMExtract}(s_i) \subseteq \mathcal{I} \times \mathcal{C} \times \mathcal{O}$$

where $\mathcal{I}, \mathcal{C}, \mathcal{O}$ are the sets of possible interventions, comparators, and outcomes, respectively. Each $\mathcal{T}_i$ may contain multiple plausible triplets. To derive a consistent meta-analytic scope, we compute the set-theoretic intersection:

$$\mathcal{T}_{\text{shared}} = \cap_{i=1}^{n} \mathcal{T}_i$$

yielding the set of ICO triplets common to all studies. The user interface includes a "Get a suggestion" button that leverages $\mathcal{T}_{\text{shared}}$ to propose a candidate

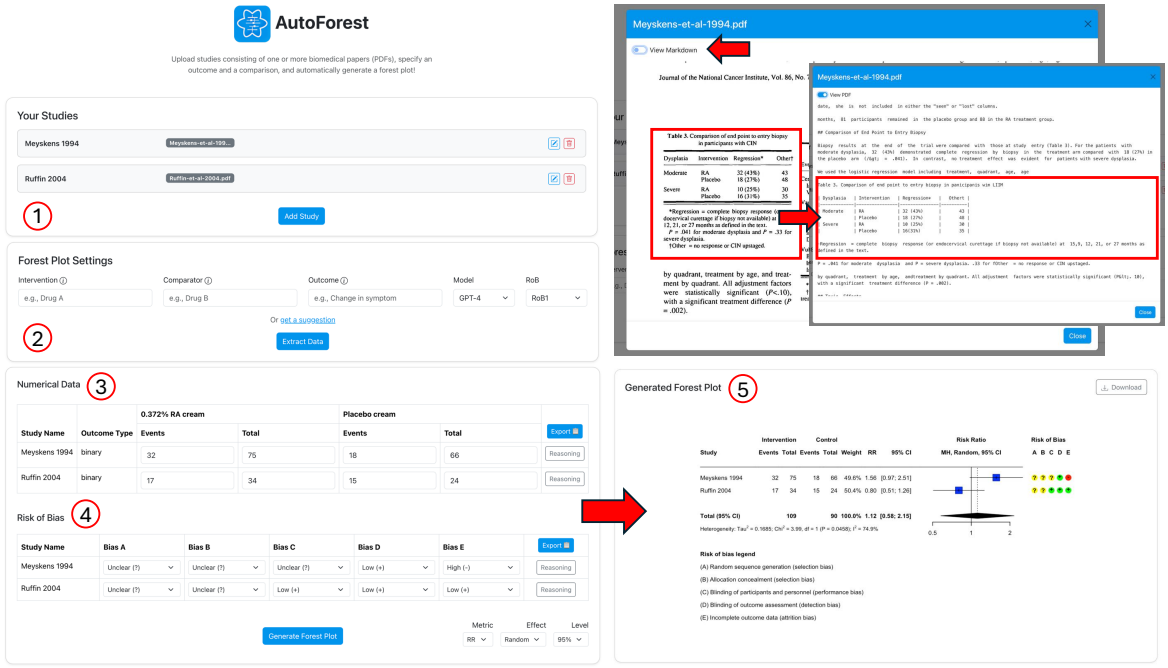

Figure 2: The interface of AUTOFOREST. The user can upload their studies (one or more documents) **(1)**, and generate ICO suggestions or manually input them **(2)**. Data is extracted from every study based on the predicted outcome type and provided ICO triplet **(3)**. Similarly, risk of bias is assessed based on the chosen RoB type (RoB1 or 2) **(4)**. A forest plot is then generated based on the desired effect model and metrics **(5)**.

ICO scope that is compatible across all included studies and therefore suitable for meta-analytic comparison. Users may override or refine the selection before initiating evidence extraction.

3.4 Numerical Evidence Extraction

Once an ICO triplet has been selected, AUTOFOREST automatically extracts the numerical evidence necessary to compute effect sizes. Following the methodology described by Pronesti et al. (2025b), we distinguish between binary and continuous outcomes and apply tailored extraction logic for each.

For binary outcomes (e.g., *number of deaths*), the system retrieves the number of events and the total sample size for each study arm.[4] For continuous outcomes (e.g., *mean blood pressure*), it extracts the means, standard deviations, and total sample sizes for both the intervention and comparator arms. These quantities are required to compute standard measures such as risk ratios or mean differences used in meta-analysis.

Extraction is carried out using a language model prompted with both the ICO definition and the full text of each study (Appendix B). For every study, the model proposes the relevant numerical values along with a free-text explanation of how those numbers were identified. All extracted values are directly editable in the interface and exportable in YAML format. This design supports a human-in-the-loop workflow, where researchers can inspect, validate, and override model outputs. The accompanying reasoning text helps identify potential misinterpretations and provides transparency of the model's decision-making process. Users may revise any extracted values, ensuring that the final extractions reflect expert judgement while benefiting from automation to reduce manual effort. The extraction step results in a structured evidence table where each row corresponds to a study and contains the extracted statistics and associated justifications.

3.5 Risk of Bias Assessment

Similarly to numerical data extraction, RoB assessment is performed by inputting the ICO triplet, the study text, and the risk domains under evaluation (prompt in Appendix B), offering support for both RoB1 (Higgins et al., 2011) and RoB2 (Sterne et al., 2019). The model produces domain-level judgments along with a domain-level explanation pointing to the underlying study text.

In addition, following the methodology described in Pronesti et al. (2026), we support a step-

[4]In the context of a forest plot, an arm refers to a group of participants in a clinical trial receiving a specific treatment.

level workflow for RoB2 in which the model is tasked with answering the signalling questions defined within each domain rather than directly assigning the final judgment. These question-level responses are subsequently processed using the official RoB2 decision algorithms ("macros" (Sterne et al., 2019); Prompt in Figure 10), allowing the system to derive the final risk-of-bias rating in a rule-based and fully transparent manner. This approach preserves model flexibility at the evidence-extraction stage while ensuring that the ultimate judgments remain consistent with RoB2 guidance.

3.6 Forest Plot Generation

After validating the evidence, users can generate a forest plot summarising the findings. AUTOFOREST supports standard effect size measures (e.g., Risk Ratio, Mean Difference) and meta-analysis models (Fixed or Random Effects). We use the `meta` package in R (Balduzzi et al., 2019) to ensure statistical correctness and compatibility with existing systematic review standards. The plot includes per-study point estimates and 95% confidence intervals, pooled effect estimate, heterogeneity statistics (e.g., I^2 for percentage of variation, τ^2 for variance of effects, p-values). The resulting forest plot is displayed in the interface and can be exported in publication-quality formats. Examples generated by the system are shown in Appendix A.

4 Evaluation

4.1 Experimental Setup

We evaluate AUTOFOREST through a combination of automatic evaluations and a controlled user study. Although the framework is model-agnostic and compatible with any LLM, our experiments use Claude Sonnet 4.5 (Anthropic, 2025), selected for its strong numerical reasoning and effectiveness in qualitative appraisal. Our goal is to assess the extent to which AUTOFOREST accelerates and improves the production of forest plots, both when used fully automatically and in human-in-the-loop mode. The evaluation centers on 32 forest plots drawn from 18 Cochrane systematic reviews, covering a total of 56 included studies. For each plot, the task consists of extracting numerical evidence from raw full-text documents for a given ICO triplet and producing a publication-ready forest-plot entry. The study is guided by three research questions:

- **RQ1**: Does AUTOFOREST reduce the time required to complete a forest plot compared to a manual workflow?
- **RQ2**: Does AUTOFOREST improve the accuracy of forest plots relative to those created manually?
- **RQ3**: Can AUTOFOREST help students achieve performance closer to that of domain experts?

4.2 User Study

We conducted a within-subjects user study involving four clinical domain experts and four graduate students familiar with the task. Participants compared three workflows: a manual baseline (Manual + RevMan), AUTOFOREST fully automated, and AUTOFOREST AI-assisted (human-in-the-loop). Participants were instructed to extract the numerical and bias data required for each forest-plot entry, log their start and end times for each task, and—when using AUTOFOREST—edit any incorrect numerical or RoB value. After completing all tasks, participants rated the usability and efficacy of the tool on a 1-5 Likert scale. Full instructions provided to participants are available at this URL.

4.3 Results

User Study Results (Table 2) demonstrate that AUTOFOREST significantly outperforms manual workflows across all metrics. For **RQ1**, the tool nearly halved the time required to complete a forest plot for both groups ($p < 0.001$). Regarding **RQ2**, the fully automated version ("AUTOFOREST only") achieved over 80% accuracy in data extraction, a substantial improvement over the manual baselines. Human-in-the-loop edits further refined these results, reaching a peak accuracy for experts of 90.2% for data extraction ($p = 0.013$) and 79.2% for RoB ($p = 0.050$). For **RQ3**, AUTOFOREST significantly narrowed the performance gap between experience levels; students using the tool achieved 86.4% accuracy, surpassing the manual performance of domain experts and nearly matching expert performance in the AI-assisted condition.

Qualitative feedback (Table 3) indicates high system utility, with participants particularly valuing the transparency provided by the "thought process". Overall, participants strongly endorsed the tool's potential for professional adoption (4.63/5), concluding that it could significantly reduce the human resources and time typically required to conduct systematic reviews (4.88/5).

Group	Method	Data Extraction		RoB		
		Acc (%)	Edit rate (%)	Acc (%)	Edit rate (%)	Time (min) ↓
Students	Manual (RevMan)	36.9	–	40.0	–	53.8
	AUTOFOREST only	83.3	–	62.5	–	–
	AUTOFOREST + edits	**86.4**	2.8	**65.1**	14.3	**26.5**
Experts	Manual (RevMan)	45.8	–	69.4	–	70.4
	AUTOFOREST only	82.5	–	63.7	–	–
	AUTOFOREST + edits	**90.2**	10.4	**79.2**	12.5	**29.8**

Wilcoxon signed-rank tests (full sample, $N = 8$): Data Extraction $p = 0.013$, RoB $p = 0.050$, Time $p < 0.001$.

Table 2: Accuracy and change rate for data extraction and RoB tasks performed by students and domain experts using RevMan and AUTOFOREST. "AUTOFOREST only" indicates performance without human verification.

Statement	Rating
• The system was easy to use.	4.63 ± 0.48
• The thought process in the data extraction step was useful to understand how each numerical value was derived and to verify or correct the automatically extracted data.	4.25 ± 0.66
• The thought process in the risk of bias step was useful to understand the rationale behind each judgement and to verify or correct the automatically assigned risk of bias assessments.	4.13 ± 0.60
• The tool could be used in a professional context with minor improvements	4.63 ± 0.48
• The tool could significantly reduce the time and number of professionals required to conduct a systematic review.	4.88 ± 0.33

Table 3: Post-study usability ratings for the AI-assisted conditions (1 = Strongly Disagree, 5 = Strongly Agree).

ICO Suggestions Let $I_{\text{SR}} \in \mathcal{I}, C_{\text{SR}} \in \mathcal{C}, O_{\text{SR}} \in \mathcal{O}$ denote the sets of interventions, comparators, and outcomes addressed in the systematic review, and let $I_{\text{sugg}}, C_{\text{sugg}}, O_{\text{sugg}}$ be the corresponding sets suggested by the model. For each element type $X \in \{I, C, O\}$, we compute:

- the proportion of correct suggestions (precision) $P_X = \frac{|X_{\text{sugg}} \cap X_{\text{SR}}|}{|X_{\text{sugg}}|}$
- the proportion of incorrect or hallucinated suggestions, $H_X = \frac{|X_{\text{sugg}} \setminus X_{\text{SR}}|}{|X_{\text{sugg}}|}$
- the proportion of elements in the systematic review that the model correctly captured (recall), $R_X = \frac{|X_{\text{sugg}} \cap X_{\text{SR}}|}{|X_{\text{SR}}|}$

Results are reported in Table 4. We observe that the tool achieves high precision and strong coverage of the systematic review's ICO elements, indicating accurate suggestions with limited omissions.

Markdown Conversion We assess the fidelity of the document-to-structure conversion layer used by AUTOFOREST through 4 quantitative metrics capturing table detection, table parsing quality and numerical cell accuracy on the 206 tables contained in the 56 studies used for the user study (Table 5).

All metrics show that the conversion layer maintains table integrity and numerical fidelity with high accuracy, ensuring reliable downstream extraction.

Element	$P_X \uparrow$	$H_X \downarrow$	$R_X \uparrow$
I	94.1	5.9	94.1
C	83.3	16.7	91.6
O	94.3	5.7	90.6

Table 4: Evaluation metrics for suggested ICO elements. P_X: precision, H_X: hallucination rate, R_X: recall

Metric	Acc (%)
Table detection	98.1
Table Structure (TEDS)	93.4
Numerical Cell	97.8
Table captions	98.1

Table 5: Document-to-structure conversion metrics.

RoB2 vs. RoB2+Macros Table 6 reports the accuracy of the standard RoB2 workflow compared to the variant augmented with macro-based reasoning. The macros lead to a clear improvement, increasing overall domain-level accuracy from 61.6% to 70.8%. This confirms that incorporating structured reasoning templates helps the model produce more consistent and interpretable risk-of-bias judgments, in line with the findings of Pronesti et al. (2026).

Method	Acc (%)
RoB2 (standard)	61.6
RoB2 (macros)	**70.8**

Table 6: Comparison of RoB2 evaluation with and without macro-based reasoning.

5 Conclusions

This paper introduced AUTOFOREST, an end-to-end system designed to automate the generation of forest plots directly from unstructured biomedical study documents. The traditional process of creating these essential visualisations is labour-intensive, often taking several hours and requiring domain expertise. AUTOFOREST addresses this bottleneck by integrating document parsing, automated ICO suggestions, evidence extraction, risk of bias assessment and reasoning to substantially reduce the time required to perform meta-analysis. Our user study suggests that AUTOFOREST particularly when used in a human-in-the-loop capacity, can accelerate evidence synthesis substantially in comparison to the standard manual workflow. Crucially, our results indicate that AI assistance may help bridge the gap between novice and expert performance; students using AUTOFOREST outperformed the manual baseline of domain experts and approached the accuracy of experts using the tool.

6 Limitations

While AUTOFOREST demonstrates a significant reduction in the time and effort required to generate forest plots, we acknowledge several limitations in the current work.

First, the scope of our evaluation warrants a cautious interpretation of findings. While the user study provides valuable insights, we only conducted the experiment with a sample of 8 participants. A larger, more diverse cohort, with varied levels of expertise conducting systematic reviews, would be necessary to establish broader usability. Moreover, the current evaluation focuses on standard parallel-group trial designs; complex designs such as multi-arm or crossover trials have not yet been explicitly tested in this paper.

Second, AUTOFOREST is designed to accelerate a specific segment of the evidence synthesis pipeline and does not address the full spectrum of tasks in a systematic review, such as literature searching, screening or study selection.

Fianlly, while the system provides textual explanations to support its data extractions and risk of bias judgments, a current limitation is the absence of a visual feature to directly highlight this evidence in the source document. Implementing such a traceability feature would make checking the model's work significantly easier for the user and represents a valuable direction for future research.

References

Anthropic. 2025. Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, Lokesh Mishra, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2024. Docling technical report. *Preprint*, arXiv:2408.09869.

Sara Balduzzi, Gerta Rücker, and Guido Schwarzer. 2019. How to perform a meta-analysis with R: a practical tutorial. *Evidence-Based Mental Health*, (22):153–160.

Leon Bax, Lu Yu, Nobuko Ikeda, and Karel G M Moons. 2007. A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Medical Research Methodology*, 7(1):40.

Michael Borenstein. 2022. Comprehensive meta-analysis software. In Matthias Egger, Julian P T Higgins, and George Davey Smith, editors, *Systematic Reviews in Health Research: Meta-analysis in Context*, pages 535–548. Wiley.

DistillerSR Inc. 2023. DistillerSR. Version 2.35. https://www.distillersr.com/. Accessed June–July 2025.

Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch, editors. 2024. *Cochrane Handbook for Systematic Reviews of Interventions, version 6.5 (updated August 2024)*. Cochrane. Available from https://training.cochrane.org/handbook.

Julian PT Higgins, Douglas G Altman, Peter C Gøtzsche, Peter Jüni, David Moher, Andrew D Oxman, Jelena Savović, Kenneth F Schulz, Laura Weeks, and Jonathan AC Sterne. 2011. The cochrane collaboration's tool for assessing risk of bias in randomised trials. *bmj*, 343.

Changkai Ji, Bowen Zhao, Zhuoyao Wang, Yingwen Wang, Yuejie Zhang, Ying Cheng, Rui Feng, and Xiaobo Zhang. 2025. RoBGuard: Enhancing LLMs to assess risk of bias in clinical trial documents. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1258–1277, Abu Dhabi, UAE. Association for Computational Linguistics.

Qusai Khraisha, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. 2024. 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. *Research Synthesis Methods*, 15(4):616–626.

Honghao Lai, Jiayi Liu, Chunyang Bai, Hui Liu, Bei Pan, Xufei Luo, Liangying Hou, Weilong Zhao, Danni Xia, Jinhui Tian, Yaolong Chen, Lu Zhang, Janne Estill, Jie Liu, Xing Liao, Nannan Shi, Xin Sun, Hongcai Shang, Zhaoxiang Bian, and 17 others. 2025. Language models for data extraction and risk of bias assessment in complementary medicine. *npj Digital Medicine*, 8(1):74.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2025a. Docling: An efficient open-source toolkit for ai-driven document conversion. *Preprint*, arXiv:2501.17887.

Nikolaos Livathinos, Christoph Auer, Ahmed Nassar, Rafael Teixeira de Lima, Maksym Lysak, Brown Ebouky, Cesar Berrospi, Michele Dolfi, Panagiotis Vagenas, Matteo Omenetti, Kasper Dinkla, Yusik Kim, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, Tim Strohmeyer, A. Said Gurbuz, and Peter W. J. Staar. 2025b. Advanced layout analysis models for docling. *Preprint*, arXiv:2509.11720.

Iain J. Marshall and Byron C. Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1):163.

Sandra McMillan, Caroline Hamel, Sharon E Kelly, Kanokporn Thavorn, David B Rice, George A Wells, and Bonnie Hutton. 2020. An evaluation of distillerSR's machine learning-based prioritization tool for title/abstract screening–impact on reviewer-relevant outcomes. *BMC Medical Research Methodology*, 20(1):1–11.

Matthew Michelson and Katja Reuter. 2019. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary Clinical Trials Communications*, 16:100443.

M. Hassan Murad, Noor Asi, Mouaz Alsawas, and Fares Alahdab. 2016. New evidence pyramid. *Evidence-Based Medicine*, 21(4):125–127.

Massimiliano Pronesti, Anya Belz, and Yufang Hou. 2026. Beyond outcome verification: Verifiable Process Reward Models for structured reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, California, USA. Association for Computational Linguistics.

Massimiliano Pronesti, Joao H Bettencourt-Silva, Paul Flanagan, Alessandra Pascale, Oisín Redmond, Anya Belz, and Yufang Hou. 2025a. Query-driven document-level scientific evidence extraction from biomedical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28034–28051, Vienna, Austria. Association for Computational Linguistics.

Massimiliano Pronesti, Michela Lorandi, Paul Flanagan, Oisín Redmond, Anya Belz, and Yufang Hou. 2025b. Enhancing study-level inference from clinical trial papers via reinforcement learning-based numeric reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30357–30373, Suzhou, China. Association for Computational Linguistics.

Rui Qiu, Shijie Chen, Yu Su, Po-Yin Yen, and Han Wei Shen. 2025. Completing a systematic review in hours instead of months with interactive AI agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31559–31593, Vienna, Austria. Association for Computational Linguistics.

Jonathan AC Sterne, Jelena Savović, Matthew J. Page, Roy G. Elbers, Natalie S. Blencowe, Isabelle Boutron, Christopher J. Cates, He Cheng, Mark S. Corbett, Sandra M. Eldridge, Miguel A. Hernán, Sally Hopewell, Asbjørn Hróbjartsson, Diana R. Junqueira, Peter Jüni, Jamie J. Kirkham, Toby Lasserson, Tianjing Li, Ann McAleenan, and 8 others. 2019. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*, 366:l4898.

Zhuanlan Sun, Ruilin Zhang, Suhail A Doi, Luis Furuya-Kanamori, Tianqi Yu, Lifeng Lin, and Chang Xu. 2024. How good are large language models for automated data extraction from randomized trials? *medRxiv*, pages 2024–02.

The Cochrane Collaboration. 2020. Review manager (revman) [computer program]. https://training.cochrane.org/online-learning/core-software-cochrane-reviews/revman/revman-5-download. Version 5.4.

James Thomas, Julie Brunton, and Silvia Graziosi. 2010. EPPI-reviewer 4.0: software for research synthesis. Technical report, EPPI-Centre, UCL.

Jianyou Wang, Weili Cao, Longtian Bao, Youze Zheng, Gil Pasternak, Kaicheng Wang, Xiaoyue Wang, Ramamohan Paturi, and Leon Bergen. 2025. Measuring risk of bias in biomedical reports: The RoBBR benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3220–3248, Suzhou, China. Association for Computational Linguistics.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2024. Accelerating clinical evidence synthesis with large language models. *NPJ Digital Medicine*.

Hye Sun Yun, David Pogrebitskiy, Iain James Marshall, and Byron C Wallace. 2024. Automatically extracting numerical results from randomized controlled trials with large language models. In *Machine Learning for Healthcare Conference*. PMLR.

A Qualitative Comparison of Domain Experts with AUTOFOREST

We present a qualitative comparison of a forest plot manually created by a domain expert (Figure 3) with the one generated by AUTOFOREST with no human verification (Figure 4) and the ground truth from the original systematic review (Figure 5).

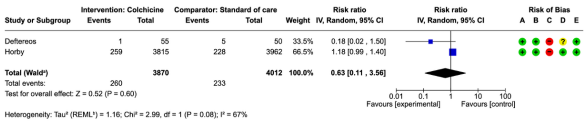

Figure 3: Forest plot created by a domain expert with RevMan. Only three of the eight numeric entries are correct (37.5% Acc). The expert was unable to locate the relevant numerical data for one study and mistakenly used data from a different result section. The risk of bias table presents mistakes for bias B,C,D (50% Acc).

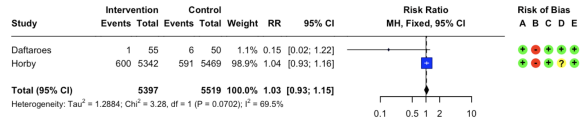

Figure 4: Forest plot generated by AUTOFOREST. Out of eight numeric entries, only two are incorrect, demonstrating a substantial improvement in accuracy over the manual expert process (75% Acc). The risk of bias map is all correct but bias D for the first study, also demonstrating a big leap with the manual version (90% Acc).

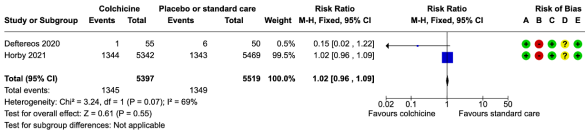

Figure 5: Ground truth forest plot.

In addition, we show a comparison of forest plot generated by AUTOFOREST (Figure 6, top) and further verified by a domain expert (Figure 6, middle) with the corresponding ground truth from the original systematic review (Figure 6, bottom).

These four visualisations illustrate the differences in accuracy across manual and automated forest plot construction. The domain expert, using RevMan, made several mistakes both in data extraction and risk of bias assessment. Most errors were due to difficulties locating the appropriate numerical values in the full text and selecting data from the wrong result section, or in the inner difficulty posed by qualitative appraisal in bias judgment. In contrast, AUTOFOREST provided a very accurate

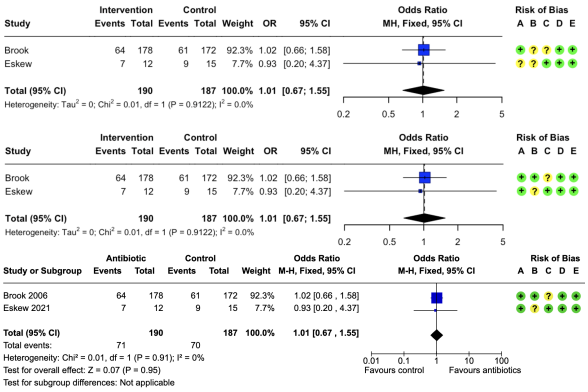

Figure 6: **Top**: Forest plot generated by AUTOFOREST; **middle:** Forest plot edited by domain expert, scoring 100% in both data extraction and risk of bias (RoB); **bottom:** ground truth from the original systematic review.

starting point, suggesting that automated extraction can reliably support domain expert's workflow. When domain experts were provided with AUTOFOREST's output as a starting point, they were able to produce a fully accurate plot. This suggests that AUTOFOREST can serve as a useful aid in manual review workflows.

B Prompts

The prompts used for the ICO suggestion, data extraction and risk of bias steps (with and without macros) are reported in Figure 7, 8, 9, and 10 respectively.

Prompt for ICO suggestion

Article: {article}

From the text below, suggest the most likely values for intervention, comparator, and outcome as bullet lists under each heading. Make sure every bullet is a self-explanatory word or short phrase that can be used as an ICO element in a meta-analysis.

Follow this format:

```
Intervention:
- intervention 1
- intervention 2 ...

Comparator:
- comparator 1
- comparator 2
- ...

Outcome:
- outcome 1
- outcome 2
...
```

Output:

Figure 7: Prompt for ICO suggestion.

Prompt for numerical data extraction

Articles: {articles}

Question: Based on the given trial articles, what is the outcome type and corresponding numerical data for the following Comparison and Outcome?

Comparison: {comparison}
Outcome: {outcome}

First, determine and output the outcome_type as either: binary or continuous

Then, provide the extracted data in format as follows:
If the outcome is binary, use this format:

outcome_type: binary
intervention:
events: NUMBER total: NUMBER
comparator:
events: NUMBER total: NUMBER

If the outcome is continuous, use this format:

outcome_type: continuous
intervention:
mean: NUMBER standard_deviation: NUMBER group_size: NUMBER
comparator:
mean: NUMBER standard_deviation: NUMBER group_size: NUMBER

Use post-intervention data when both pre and post are available. If multiple timepoints are reported, choose the one closest to the timepoint of interest, or the latest available. You must first think about the question and output your thought process using <think></think> tags followed by the requested output format.

Figure 8: Prompt for numerical data extraction.

Prompt for RoB

Articles: {articles}

Question: Based on the given trial article, what is the risk of bias for the following domains, given Intervention, Comparator and Outcome?

Intervention: {intervention}
comparator: {comparator}
Outcome: {outcome}

Domains to assess: {domains}

For each domain, provide your assessment as either: {tool_values}.
Notice: you must stick to one of these **exact** values.
You must first think about the question and output your thought process using <think></think> tags followed by the requested output format.
Use this format:

<think>
Your thought process here
</think>
A: RISK_LEVEL
B: RISK_LEVEL
C: RISK_LEVEL
D: RISK_LEVEL
E: RISK_LEVEL

Figure 9: Prompt for Risk of Bias Assessment.

Prompt for RoB2 with macros

Articles: {articles}
Intervention: {intervention}
comparator: {comparator}
Outcome: {outcome}

Task:
1. Identify whether the study uses ITT or PP.
2. Answer all bias-assessment questions (A–E).
3. Put reasoning inside <think></think>.
4. Then output a JSON object using:
answers ∈ {Y, N, PN, PY, NI, NA}
each entry as { "q": "<id>", "ans": "<value>" }

Domain B rule:
- If treatment_type = ITT → answer ITT questions; PP questions = NA.
- If treatment_type = PP → answer PP questions; ITT questions = NA.

Questions:

A — Randomisation
1.1 allocation sequence random?
1.2 allocation concealed?
1.3 baseline differences problematic?

B-ITT — Deviations from intended intervention
2.1 participants aware?
2.2 carers aware?
2.3 deviations due to trial context?
2.4 deviations likely affected outcome?
2.5 deviations balanced?
2.6 appropriate analysis for assignment?
2.7 could failure to analyze by assignment impact results?

B-PP — Deviations from intended intervention
2.1 participants aware?
2.2 carers aware?
2.3 non-protocol interventions balanced?
2.4 failures in implementation?
2.5 non-adherence affected outcome?
2.6 appropriate analysis for adherence?

C — Missing outcome data
3.1 nearly all data available?
3.2 evidence result not biased by missing data?
3.3 could missingness depend on true value?
3.4 likely missingness depended on true value?

D — Measurement of outcome
4.1 measurement inappropriate?
4.2 measurement differ across groups?
4.3 assessors aware?
4.4 could awareness influence assessment?
4.5 likely assessment influenced?

E — Selective reporting
5.1 analysis pre-specified?
5.2 result selected from multiple possible measures?
5.3 result selected from multiple analyses?

Required Output:

<think>
Your reasoning here.
</think>
{
"treatment_type": "...",
"A": [{ "q": "...", "ans": "..." }, ...],
"B": [{ "q": "...", "ans": "..." }, ...],
"C": [{ "q": "...", "ans": "..." }, ...],
"D": [{ "q": "...", "ans": "..." }, ...],
"E": [{ "q": "...", "ans": "..." } , ...]
}
Do not add backticks or any other text outside the specified format.

Figure 10: Prompt for RoB2 with macros.