Elicit AI Data Extraction Guide for Clinical Papers

By Adrien Laurent, CEO at IntuitionLabs • 3/13/2026 • 45 min read

elicit ai | data extraction | systematic reviews | clinical research | evidence synthesis | llm tools | medical literature

Executive Summary

The rapid expansion of clinical research literature has made **structured data extraction** from published studies an essential task for evidence synthesis and decision-making ([1] systematicreviewsjournal.biomedcentral.com) ([2] pubmed.ncbi.nlm.nih.gov). Traditional systematic reviews, which compile such data manually, are laborious and error-prone; studies report average review timelines of over 1 year and significant costs ([3] systematicreviewsjournal.biomedcentral.com). Recent advances in natural language processing and AI – particularly large language models (LLMs) – promise to automate and accelerate this process. **Elicit** is a novel AI-based research assistant that uses LLMs to streamline systematic review tasks, including *semantic paper search*, *screening*, *data extraction*, and *report summarization* ([4] bmcmedresmethodol.biomedcentral.com) ([5] support.elicit.com). This report provides an **in-depth guide** to using Elicit for structured data extraction from clinical research papers.

We first review the **background** of systematic reviews and data extraction, highlighting key challenges and traditional solutions ([3] systematicreviewsjournal.biomedcentral.com) ([6] bmcmedinformdecismak.biomedcentral.com). Next, we describe Elicit's technology and features, including its semantic search engine and automated table-generation capabilities ([4] bmcmedresmethodol.biomedcentral.com) ([5] support.elicit.com). We then detail the **workflows** for using Elicit's Systematic Review (SR) module: from defining a research question and gathering relevant studies to designing extraction columns and executing full data extraction ([5] support.elicit.com) ([7] support.elicit.com). Throughout, we cite practical tips and insights from Elicit's documentation and user case studies.

Following the methodology, we present **empirical evidence and case studies** on Elicit's performance. For example, Elicit's internal case study in education research reported 99.4% data extraction accuracy ([8] elicit.com), and a CSIRO evaluation found near-zero missed data points. Independent studies using Elicit's outputs are emerging: for instance, a Cochrane collaboration (Bianchi *et al.*, 2025) systematically compared Elicit's extraction on 20 randomized trials to human extraction, finding that Elicit often captured "more" or "equal" information but sometimes missed detail (www.ovid.com). We contrast these results with other AI methods: for example, a recent proof-of-concept found another LLM (Claude 2) achieved 96.3% accuracy on 160 clinical trial data points ([10] pubmed.ncbi.nlm.nih.gov). We tabulate these findings to highlight accuracy comparisons.

Finally, we discuss **implications and future directions**. Elicit (and LLM tools in general) can greatly speed reviews (through automation of repetitive work) and potentially improve comprehensiveness, but they also require careful human oversight (www.ovid.com) ([11] bmcmedresmethodol.biomedcentral.com). Challenges include occasional hallucinations or omissions and dependency on the AI's knowledge cutoff. Ongoing development (e.g. better models, retrieval-augmented methods ([12] pmc.ncbi.nlm.nih.gov), and emerging reporting guidelines) should further improve reliability. In conclusion, Elicit offers a powerful, 10x-speedup approach to structured data extraction in systematic reviews ([13] support.elicit.com) ([14] bmcmedresmethodol.biomedcentral.com). Researchers should leverage Elicit's capabilities (guided by robust validation and transparency) to enhance efficiency and accuracy in synthesizing clinical evidence, while remaining mindful of its current limitations.

Introduction and Background

The **systematic review** is the cornerstone of evidence-based medicine, synthesizing data from multiple clinical studies to answer focused research questions. However, the sheer volume of published clinical papers makes traditional manual review processes increasingly untenable. In 2020, **over 100,000 systematic reviews** were registered on PROSPERO, up from hundreds a decade earlier ([3] systematicreviewsjournal.biomedcentral.com). Each systematic review can take **months to years** and cost on the order of $100–150k ([3] systematicreviewsjournal.biomedcentral.com). In particular, the **structured data extraction** phase – where a reviewer reads each included paper and fills pre-defined data fields (e.g. sample size, interventions, outcomes, effect sizes) – is a prime bottleneck. This step is *"cognitively demanding and time-consuming"* ([15]

pmc.ncbi.nlm.nih.gov) ([3] systematicreviewsjournal.biomedcentral.com), and errors are common (one analysis found systematic data extraction errors in many Cochrane reviews ([16] systematicreviewsjournal.biomedcentral.com)).

Recognizing these challenges, the research community has developed automated tools to assist data extraction. Early efforts like **ExaCT** (2010) used machine learning and rule-based NLP to identify key trial facts (eligibility, sample size, dosage, outcomes) from RCT publications ([6] bmcmedinformdecismak.biomedcentral.com). Similarly, tools like RobotReviewer leveraged classifiers to flag risk-of-bias text and data citations. More recently, transformer-based methods (e.g. BERT models) have shown promise: Panayi *et al.* (2023) pretrained a BERT with biomedical text and added CRF layers, achieving F1 scores of ~70–88% for entity recognition in oncology reviews ([17] systematicreviewsjournal.biomedcentral.com). Other studies explored distant supervision for extracting PICO elements from abstracts ([18] academic.oup.com).

However, most prior systems required extensive training data or only covered narrow extraction tasks ([1] systematicreviewsjournal.biomedcentral.com) ([6] bmcmedinformdecismak.biomedcentral.com).The emergence of **large language models (LLMs)** – general-purpose models like OpenAI's GPT or Anthropic's Claude with billions of parameters – has dramatically changed the landscape. These models can often perform extraction with minimal or zero-shot prompting, leveraging their broad pretraining on scientific corpora. For example, Gartlehner *et al.* (2024) used the Claude 2 LLM to extract 16 types of data from 10 open-access clinical trials and found *96.3% overall accuracy* ([10] pubmed.ncbi.nlm.nih.gov). Polak & Morgan (2024) showed that GPT-4 (ChatGPT) with careful prompt engineering ("ChatExtract" workflow) could achieve ~90% precision and recall for extracting materials properties ([19] www.nature.com) ([20] www.nature.com).

Elicit represents a commercial instantiation of this new paradigm. Developed by Ought, Elicit is an AI research assistant built around LLMs (originally GPT-3 and successors). Elicit integrates semantic search over academic databases with LLM-based generation, aiming to *automate many steps of research workflows* ([4] bmcmedresmethodol.biomedcentral.com) ([5] support.elicit.com). Notably, Elicit's **Systematic Review** workflow guides users through search, screening, and structured data extraction, producing tables of study data and even draft text summaries ([5] support.elicit.com) ([7] support.elicit.com). Early validation suggests dramatic speedups – Elicit's site claims up to *80% time saved* ([13] support.elicit.com) – and high accuracy: in one case, an independent body reported 99.4% data extraction accuracy ([8] elicit.com).

This report delves into **how to use Elicit for structured data extraction from clinical papers**. We review Elicit's functionality in detail, cite case studies and user experiments, and analyze its performance relative to manual methods and other AI tools. Our goal is to provide a comprehensive, evidence-based guide so that researchers can confidently integrate Elicit into their clinical literature reviews while understanding its strengths and limitations.

Overview of Elicit and Related Approaches

What is Elicit?

Elicit is an AI-driven research assistant designed for evidence-based tasks ([21] elicit.com) ([4] bmcmedresmethodol.biomedcentral.com). It leverages large language models (initially GPT-3, and presumably later GPT-4 or equivalent) in concert with semantic search over academic databases. Unlike general search engines, Elicit's search is "topic-based", using natural language understanding to find papers even without exact keyword matches ([4] bmcmedresmethodol.biomedcentral.com). Elicit has a multi-modal interface for systematic reviews: it supports **multi-tab searching** (combining broad questions with precise filters), **AI-assisted screening**, automated **data extraction**, and a final **research report generation** ([22] support.elicit.com) ([4] bmcmedresmethodol.biomedcentral.com).

Key features include:

- **Semantic Search:** Finds papers by concept similarity rather than exact keywords ([23] bmcmedresmethodol.biomedcentral.com). Elicit can query multiple sources (e.g. Semantic Scholar) in one integrated search ([23] bmcmedresmethodol.biomedcentral.com).

- **AI-Generated Summaries and Reports:** Elicit can summarize what it "thinks" the answer is, synthesizing findings from top papers ([4] bmcmedresmethodol.biomedcentral.com). The "Custom Report" feature automatically drafts overviews of the research question.
- **Screening Automation:** For systematic reviews, Elicit can auto-generate screening criteria from the research question ([5] support.elicit.com). It then scores and sorts papers by relevance to each criterion, allowing rapid inclusion/exclusion decisions.
- **Data Extraction Tables:** Crucially, Elicit offers an **automated extraction pipeline**: researchers define (or let Elicit suggest) data fields (columns), and Elicit populates a table with data from each paper, along with context (quotes and reasoning) ([7] support.elicit.com) ([24] support.elicit.com).
- **User Control and Transparency:** All Elicit outputs remain linked to source text. Users can click on any extracted cell to see the supporting quote from the paper (and an explanation), allowing verification ([25] support.elicit.com) ([26] support.elicit.com). This keeps a human in the loop for quality control.

By combining retrieval and generation (a **retrieval-augmented generation [RAG]** approach), Elicit aims to reduce the hallucination risk common in plain LLM outputs ([27] pmc.ncbi.nlm.nih.gov) ([23] bmcmedresmethodol.biomedcentral.com). It retrieves up-to-date literature and passes relevant snippets to the LLM during generation, thus grounding answers in the actual texts. As one design paper notes, RAG *"dynamically retrieves and integrates up-to-date, domain-specific information… reducing hallucinations and improving factual accuracy"* ([27] pmc.ncbi.nlm.nih.gov).

While many features of Elicit echo general LLM capabilities, its tight focus on systematic review tasks distinguishes it. For example, Elicit's screening and extraction are "process-based" – driven by the research question – rather than free-form Q&A ([4] bmcmedresmethodol.biomedcentral.com) ([5] support.elicit.com). It implicitly understands the PICO (Population, Intervention, Comparison, Outcome) schema and organizes results accordingly ([28] bmcmedresmethodol.biomedcentral.com). In summary, **Elicit is an AI-enhanced workflow engine** for evidence synthesis, whose latest Systematic Review module can "automate screening and data extraction" in reviews ([29] elicit.com) ([5] support.elicit.com). We next examine how to use it step by step.

Related Tools and Context

It is useful to situate Elicit within the broader ecosystem of automated data extraction tools. Early systems like **ExaCT** (2010) were tailored to clinical trial reports, using statistical classifiers and heuristics to extract predefined fields ([6] bmcmedinformdecismak.biomedcentral.com). ExaCT required substantial rule- or model-building for each field and user review of its suggestions. A decade later, tools like **SWIFT-Review** and others focused on prioritizing and semi-automating screening, while specialized extraction tools (e.g. **RobotReviewer**) underwrote risk-of-bias and metric extraction ([30] systematicreviewsjournal.biomedcentral.com).

More recent approaches have turned to deep learning. For example, several groups applied BERT-based models fine-tuned on clinical corpora to recognize trial attributes ([30] systematicreviewsjournal.biomedcentral.com). These models often achieve reasonable entity-recognition performance (e.g., BERT+CRF with F1 ~73% on oncology entities) ([17] systematicreviewsjournal.biomedcentral.com) but still require labeled training data for each task and may omit relations. In contrast, Elicit's approach is zero-shot or few-shot (the user provides high-level instructions rather than labeled examples).

Simultaneously, a number of studies have explored using cutting-edge LLMs for data extraction. Gartlehner *et al.* (2024) demonstrated that the Claude 2 LLM – when given extracted PDF text and asked directly – achieved *96.3% accuracy* on 160 key data items from randomized trials ([10] pubmed.ncbi.nlm.nih.gov). Polak & Morgan (2024) showed that GPT-4 with carefully designed prompts reached roughly *90% precision and recall* on materials science data extraction ([31] www.nature.com) ([20] www.nature.com). In parallel, experimental interfaces like **SciDaSynth** (CamPL:2025) are being developed to combine LLMs with interactive editing and visualization for data extraction ([12] pmc.ncbi.nlm.nih.gov) ([32] pmc.ncbi.nlm.nih.gov).

Collectively, these advances highlight that high-quality automated data extraction is now feasible in principle. However, most generic LLM approaches require manual prompting and post-editing. Elicit aims to integrate these capabilities into a

streamlined, domain-specific pipeline with built-in structure. In the following sections, we focus on **Elicit's procedures and performance** in extracting structured data from clinical and related research papers.

Using Elicit for Structured Data Extraction

Overview of the Systematic Review Workflow

Elicit's **Systematic Review (SR)** module provides a guided, multi-stage workflow covering literature search through final reporting ([5] support.elicit.com). The primary stages relevant to data extraction are:

1. **Define Research Question:** Begin by formulating a clear, answerable research question. Elicit's interface prompts you to input this question as free text (e.g. "What is the effect of Drug X vs placebo on outcome Y in population Z?"). Elicit will use this question as the basis for subsequent steps ([5] support.elicit.com). It can also suggest refinements or alternate phrasing to clarify the inquiry.

2. **Gather Papers:** Next, you assemble the pool of potentially relevant studies. Here Elicit offers **semantic search and multi-tab queries** ([33] support.elicit.com) ([23] bmcmedresmethodol.biomedcentral.com). You can open multiple search tabs and run variations (broad conceptual prompts or precise Boolean filters, including clinical trial filters and date ranges). Elicit will de-duplicate and consolidate results across tabs ([33] support.elicit.com). In addition to live searches, you can upload PDFs directly or import from your Elicit library. The goal is to converge on the complete set of candidate papers for the review.

3. **Screening:** In Stage 2, you specify inclusion/exclusion criteria. Elicit auto-generates a draft set of screening criteria based on your question ([22] support.elicit.com) (e.g. "population is adults with diabetes", "intervention includes drug therapy"). You can edit or add criteria as needed. Elicit then applies these criteria to your gathered papers in two phases: title/abstract screening, followed optionally by full-text screening ([34] support.elicit.com) ([11] bmcmedresmethodol.biomedcentral.com). During screening, each paper is scored on how well it matches the criteria, allowing you to quickly accept or reject, with the option to *override* the AI's decision ([34] support.elicit.com). Strict criteria (that force exclusion if not met) and adjustable thresholds allow control of sensitivity vs specificity. At the end of screening you have your **narrowed set of included studies** for extraction.

4. **Define Extraction Fields (Columns):** Prior to auto-extraction, you need to set up the data table columns. Elicit will automatically suggest some relevant columns based on your question and the pilot papers, but you can customize these. For example, for a clinical drug trial SR, columns might include "Population size (N)", "Intervention dosage", "Control condition", "Primary outcome value", "Confidence Interval", etc. You can *add new columns, edit instructions, or duplicate existing ones* ([35] support.elicit.com). Elicit supports saving column presets for reuse. It is common to refine the column instructions by iterating on a small "pilot set" of ~10 papers. Elicit randomly selects 10 included papers to test-pilot the extraction; you review their extracted values (and supporting quotes) to gauge accuracy ([7] support.elicit.com). You might adjust the column prompt wording or format at this stage to improve consistency.

5. **Run Data Extraction:** Once injection of columns and prompts is finalized, you execute **full extraction** on the entire set of included papers ([36] support.elicit.com). Elicit applies each extraction column query across all PDFs. Importantly, it always shows you the supporting quote (text snippet from the paper) that led to each extracted cell ([25] support.elicit.com) ([26] support.elicit.com). This allows rapid verification: you can eyeball the highlighted passages if something looks wrong. The extraction may take time for large sets, but Elicit will continue processing in the background. When finished, the result is a structured table (viewable in-app or exportable as CSV) containing the extracted data and references.

6. **Review & Refine:** After extraction, you should manually review the table. Elicit marks each cell with reasoning text, so you can verify correctness. It's prudent to correct any errors by clicking into cells and editing the entries (or editing the source quote if misinterpreted). This hybrid AI-human approach ensures accuracy. You may also adjust columns or rerun extraction if needed.

7. **Generate Final Report (Optional):** After extraction, Elicit can generate a narrative summary report of your review based on the top-scoring papers ([37] support.elicit.com). This report's content comes from Elicit's natural language synthesis of the results and is an additional deliverable summarizing key points (though the core structured output is the data table).

Throughout, Elicit enforces **transparency**: each automated suggestion cites the original text ([26] support.elicit.com). The user remains responsible for final data quality. In essence, Elicit **locks in human oversight** at each step, ensuring that "auto" never means "unchecked."

This workflow is depicted schematically in Table 1 below, contrasting each step with traditional manual processes.

SR Step	Traditional Process	Elicit AI-Assisted Process
Define Research Question	Manual brainstorming; canonical PICO framework	Ask question in Elicit; AI suggests refinements for clarity ([5] support.elicit.com).
Literature Search	Keywords in PubMed, Embase, etc.; may miss semantically relevant studies ([23] bmcmedresmethodol.biomedcentral.com).	Multi-tab Elicit searches combining semantic and Boolean queries ([33] support.elicit.com); uses academic database sources (Semantic Scholar, etc.).
Screening	Human reviewers read titles/abstracts against inclusion criteria; laborious.	Elicit auto-generates screening criteria from query ([5] support.elicit.com), scores each paper, and lets user rapidly accept/reject ([34] support.elicit.com); supports adjustable strictness.
Develop Extraction Template	Decide data fields (e.g. Population, Intervention, Outcome, etc.) based on domain and objectives.	Elicit suggests columns from question context. User edits or adds columns with prompts ([35] support.elicit.com). Pilot on sample set (10 papers) to refine prompts and "save as preset" for reuse.
Data Extraction (Execution)	Reviewer manually reads each included full text and fills a spreadsheet of extracted values.	Click "Run Extraction": Elicit populates all column queries across papers ([36] support.elicit.com). Outputs supporting quotes for each entry ([25] support.elicit.com).
Verification	Often two reviewers double-check each cell; time-consuming but required.	The user quickly scans AI-filled table while checking highlighted source quotes. Edits any discrepancies immediately.
Time Efficiency	Manual extractions take many weeks in large reviews ([3] systematicreviewsjournal.biomedcentral.com).	Vendors claim ~10× faster with Elicit; e.g. one case ~80% time reduction ([13] support.elicit.com) (independent speed results are limited but promising).
Output	Standard SR data tables and PRISMA report developed independently.	Elicit provides exportable CSV and (for optional final step) an auto-generated research report summarizing findings ([37] support.elicit.com).

Table 1: Comparison of traditional systematic review steps vs. Elicit's AI-assisted workflow (adapted from Elicit documentation ([5] support.elicit.com) ([35] support.elicit.com) and the authors' experience). Citations indicate Elicit-specific features.

The workflow above highlights how Elicit transforms manual effort into streamlined steps. In the next sections, we delve into each phase in detail, with examples and citations.

Preparing a Systematic Review in Elicit

Defining the Research Question: A precise starting question is critical. You should articulate your population, intervention/comparator, and outcomes (PICO) in natural language. For example, a question might be "What is the effect of corticosteroids versus placebo on mortality in adult ICU patients with COVID-19 pneumonia?" Upon entering this question, Elicit will parse it to suggest screening criteria and data fields. (The goal is to explicitly encode the PICO elements.) If your question is broad or admits multiple interpretations, Elicit's interface may recommend refinements or alternate facets to explore ([38] support.elicit.com). Use this assistance to ensure your question focuses on exactly the desired concepts.

Gathering Relevant Studies: Once the question is set, you compile citations. In manual reviews, this involves running separate keyword queries on PubMed/Embase/ClinicalTrials.gov, then deduplicating. In Elicit, you use the *"Gather"* step. Here's how it works:

- **Multi-Tab Search:** Elicit lets you open multiple search tabs, each running a different query simultaneously ([33] support.elicit.com). For instance, Tab 1 might answer your broad question, while Tab 2 applies specific filters (year, study type). This flexibility avoids missing studies that might slip past one strategy.
- **Semantic and Boolean Queries:** On each tab, you can enter either a broad semantic query (free-text question) or precise Boolean strings. Elicit also supports built-in filters such as Clinical Trial tags and date ranges. As Elicit retrieves results, it de-duplicates citations across all tabs ([33] support.elicit.com).
- **Refinement and Iteration:** You might iterate several times. For each tab, run the search, inspect results, and then return to modify queries. This is akin to "test-driven" searching. The interface duplicates papers across tabs and highlights which tab they came from, facilitating completeness.

Elicit extracts metadata (journal, title, abstract, etc.) for found papers. It's important to recognize that Elicit's indices come from academic sources (e.g. Semantic Scholar ([23] bmcmedresmethodol.biomedcentral.com)), so the coverage may differ from conventional databases. To maximize sensitivity, one should try variations (e.g. synonyms, MeSH terms, author names). Elicit stops adding new papers when it exhausts results ("no more papers"), at which point you have your assembled corpus.

Screening Studies

With gathered studies in hand, the screening phase weeds out irrelevant papers. Elicit automates much of this:

- **Generate Screening Criteria:** Based on your research question, Elicit will propose inclusion/exclusion criteria (e.g. "Population includes adults", "Intervention involves Drug X", "Outcome includes event Y") ([5] support.elicit.com). You can edit these suggestions or add your own criteria (for example, excluding studies in animals or requiring RCT design).
- **Pilot Screening Phase:** Elicit often starts screening on titles/abstracts first. It presents a small batch of papers and asks you to label them include/exclude according to the criteria. This initial labeling helps Elicit calibrate its scoring.
- **Apply to All Papers:** Once the criteria are finalized, clicking *"Evaluate Screening"* runs the screening across your entire set. Elicit automatically excludes any papers failing any *strict* criteria, and ranks the rest by likelihood of meeting all criteria ([39] support.elicit.com). A screening score is assigned, and a summary column shows which criteria each paper satisfies. You can raise or lower the screening threshold slider to adjust how many are included for review.
- **Override and Review:** If Elicit includes a paper that you know should be out (or vice versa), you can click the paper to see detailed explanations for its decision. The right-hand sidebar allows overriding Elicit's decision for that paper. In practice, you might skim the lower-probability papers to catch misclassifications. The goal is to produce the final set of *included* papers that will go into data extraction.

Screening is usually done in duplicate in manual SRs for reliability, but with Elicit an efficient alternative is: one reviewer screens in Elicit and a second checks the included/excluded lists for any disagreements. *Note:* Bianchi *et al.* found that Elicit's screening of one example review identified 3 of 17 studies that the manual process missed ([40] bmcmedresmethodol.biomedcentral.com) ([11] bmcmedresmethodol.biomedcentral.com), illustrating that AI screening can complement human searching. However, they also found **variability** in Elicit's recall across runs – a warning that screening results may not be perfectly reproducible (we discuss reproducibility below).

Defining and Iterating on Extraction Columns

Once your study pool is finalized, you prepare the data table for extraction. In manual workflows, researchers would create a spreadsheet (e.g. in Excel) with columns like "Study ID", "Sample size (intervention)", "Sample size (control)", "Outcome measurement", and so on, based on the PICOS elements. With Elicit, you do this directly in-app using *data extraction columns*. The steps:

- **Initial Setup:** Click the *"Define extraction"* or equivalent step to begin. Elicit will randomly select 10 of your included papers as a **pilot set** for you to iterate on columns ([7] support.elicit.com).
- **Auto-Suggested Columns:** Elicit pre-populates some suggested columns based on your research question ([35] support.elicit.com). For instance, if your question mentions "population" and "outcome", it might suggest columns like "Population characteristics" or "Primary outcome value". These are just suggestions; you can rename or delete them.
- **Custom Columns:** You can click **"Add new"** to create any column. For each column, you define a clear *instruction*. For example, you might name a column "Baseline mean age" and instruct "Extract the mean age of participants at baseline for the treatment group." These instructions help guide the LLM. In each column's settings you can input or edit the prompt. You may need to experiment with wording to get the desired output.
- **Duplicate & Compare:** A useful trick is to duplicate a column with different instructions to see which works best ([41] support.elicit.com). For example, you could copy an "Outcome" column and ask one copy for the numeric value only, and another copy for a verbal conclusion. Then compare which outputs are more accurate on the pilot set.

- **Presets and Templates:** If you have existing column sets (from past projects or publicly shared Elicit templates), you can import those. You can also "Save as preset" any configured columns to reuse in future reviews ([41] support.elicit.com).
- **Inspect Pilot Extraction:** Each time you adjust columns, Elicit will immediately run extraction on the 10 pilot papers. You can then browse the resulting mini-table. For any cell, clicking it shows the exact quote from the paper that Elicit used. This allows you to assess both *completeness* (did it find the right info?) and *precision* (did it correctly interpret units, numbers, etc?). For example, if a pilot column "N_intervention" yields values that seem too high or low, you may refine the column prompt to specify exactly what to extract.

This iterative piloting of extraction columns is analogous to how one might *debug* a formula or script. The **iterative** process of seeing immediate feedback on a small subset ensures that by the time you run extraction on all papers, the prompts are well-tuned. Elicit's guidance wizard is especially valuable because designing optimal prompts for structured extraction can be non-intuitive for many users.

Defining Clinical Data Fields

In clinical or epidemiological domains, some standard extraction fields (columns) often include **population details** (e.g. gender, age range, N), **study method** (design, setting), **intervention details** (drug dose, frequency), **outcome settings** (definition of endpoint, follow-up time), and **results** (effect sizes, confidence intervals). Elicit can be directed to extract nearly any such field with an appropriate instruction. Some illustrative column instructions might be:

- *"Population N (treatment group)"* – extract the number of participants in the intervention arm.
- *"Intervention specification (dose and schedule)"* – capture the dosage regimen of the intervention.
- *"Primary efficacy outcome value and unit"* – get the numeric result for the primary outcome (e.g. "RR=0.85").
- *"95% confidence interval for primary outcome"* – pick out the confidence limits.
- *"Any reported adverse events count"* – extract if AE rates are given.
- *"Follow-up duration (months)"* – find how long outcomes were measured.

Elicit will typically locate these in methods or results sections. It also handles qualitative fields: for instance, you could have a column "Conclusion Summary" with a prompt like *"Summarize the authors' main conclusion about the question."* and Elicit would put a short summary text per study.

The flexibility of Elicit's columns is a major strength: if your review has custom data needs (say, coding types of interventions), you define columns accordingly. But the **salient tip** is to make instructions as clear and concise as possible. The LLM excels at extracting exactly what the instruction asks for, and can follow chain-of-thought queries (like asking to check consistency or units).

Finally, remember to **order your columns logically** – typically, identification first (Study ID, year, authors), then population, intervention, outcome specifics, and finally numeric results. This mirrors the standard formats used in meta-analysis tables.

Running the Full Extraction and Reviewing Results

After polishing the columns on the pilot set, you proceed to extraction across **all** included studies:

1. **Run Full Extraction:** Click *"Next"* or *"Run full extraction"*. Elicit will ask for confirmation of how many papers will be processed (deducted from your allowance). Once confirmed, the tool applies each column's prompts to each paper in your library ([36] support.elicit.com). This can take some time for large reviews (Elicit may run in the background).
2. **Examine the Data Table:** When finished, Elicit displays the complete extraction table. Each row is a paper, and each column is the field you defined. Crucially, each cell value is highlighted with a small "speech bubble" or icon – clicking it reveals the excerpt from the paper that yielded this value ([25] support.elicit.com). For example, if the column is "Mean baseline age", clicking the cell should show the sentence from the Methods or Results stating that age. This transparency is key for validation.

3. **Validate and Edit:** Go through the table systematically. If an extracted value looks wrong, click into that cell and review the quote. Often errors arise from ambiguity (e.g. extracting "5" but it meant "5 mg"), or multiple possible numbers. In such cases, you may either correct the number manually, or refine that column's prompt (e.g. adding unit constraints or context) and rerun extraction if needed. Because Elicit retained the original context, you could even see if units were mentioned nearby. If an extraction is partially correct, such as "42±5" instead of "42", you can either manually trim it or adjust the extraction instructions.

4. **Consistency Checks:** Elicit's table can be exported as CSV. Many users then perform sanity checks outside: for instance, flagging any columns with missing data, or sorting by numeric values to spot outliers. If certain columns are missing for entire papers, it may indicate those data weren't reported or that the column should be modified to handle variations. Correction can be iterative: e.g., if "Sample size" was sometimes extracted from the wrong study arm, refine the prompt.

5. **Quality Control:** At minimum, a senior reviewer should cross-check a subset of rows and columns against the original papers for consistency. The cited support text in each cell speeds this dramatically. According to Elicit documentation, this step is 100% human-driven: Elicit itself only *proposes* entries ([25] support.elicit.com). In one validation study, Elicit's results were compared to human extraction by co-reviewers (www.ovid.com); while complete agreement was not reached, the AI often captured more details than a single extractor. Thus best practice is to **triangulate AI outputs with human judgment**.

In summary, the user's role after extraction is akin to proofreading a manufactured table: Elicit does the bulk data-finding automatically, and the reviewer ensures no relevant data was missed or misinterpreted. This greatly accelerates the process: even if Elicit's extraction is only 90% correct, the reviewer needs to fix 10% of cells rather than extract 100% manually.

Key insight: *Elicit always shows the "why" behind each extracted datum*. This openness means reviewers are not working in the dark. It also allows bulk exporting: once verified, the final table (with references) can be downloaded for analysis or inclusion in meta-analytic software. Elicit's built-in Research Report generator then creates a narrative summary from the top studies (limited to the top 80 by default) ([42] support.elicit.com), but this is ancillary to the structured data.

Data Analysis and Performance Evidence

The utility of Elicit hinges on its **accuracy and efficiency**. Here we survey quantitative findings from case studies and research comparing Elicit (or similar LLM tools) to human performance in structured extraction. Wherever possible, we cite evidence from published evaluations.

Elicit's Reported Accuracy

Official data (Elicit's own case studies) suggest very high accuracy. For instance, VDI/VDE (a German engineering organization) conducted a large systematic review using Elicit. They reported that **1502 of 1511** required data points were correctly extracted automatically – an accuracy of **99.4%** ([8] elicit.com). Similarly, researchers at CSIRO (Australia) benchmarked Elicit against manual extraction and found **near-zero false negatives**, meaning Elicit missed essentially no key information compared to human reviewers ([43] elicit.com). Elicit's team also mentions **94–99%** accuracy rates in replicated reviews tested internally ([44] elicit.com). (These figures likely reflect aggregate performance on fields like numerical outcomes.)

These case-study numbers are **promising**, though they come from internal testing contexts and must be interpreted cautiously. Still, they indicate that *for well-defined fields*, Elicit can reach near-human reliability.

Independent Comparative Study

Importantly, independent academic studies are beginning to evaluate Elicit's extraction rigorously. Bianchi *et al.* (2025) conducted a comparative study (Cochrane Evidence Synthesis and Methods) on Elicit's extraction of RCT data (www.ovid.com) (www.ovid.com). They took 20 randomized trials (German nursing care interventions) with a known human-extracted dataset (in FIT-NursingCare). The findings were:

- **29.3%** of data points: Elicit actually extracted *more* information than the human reviewer (e.g. capturing data the human had omitted) (www.ovid.com).
- **20.7%**: Elicit's extraction was exactly *equal* to the human extraction.
- **45.7%**: Elicit's output was *partially equal*, meaning it got some but not all of the needed content (often leaving out sub-details).
- **4.3%**: Elicit's output *deviated*, i.e., it was incorrect or inconsistent with the human data (www.ovid.com).

Overall, this suggests that Elicit covered at least some relevant data in nearly all fields, missing nothing substantial in only ~4% of fields. In some categories (e.g. study identifiers), Elicit matched nearly 100% of entries, while in more complex fields like interventions or subgroup details, there was more partial completion. The authors concluded that combining Elicit with human review could improve efficiency: in 29.3% of cases Elicit "got more" and in ~20% equal, reducing human workload. However, they caution that "extracting more" isn't always better (sometimes Elicit picks up unneeded data), and that any AI output should be verified (www.ovid.com) ([11] bmcmedresmethodol.biomedcentral.com).

This study also compared Elicit to ChatGPT (another LLM) for the same task, finding that while ChatGPT performed moderately well on some tasks, Elicit slightly outperformed it numerically (www.ovid.com). Importantly, Bianchi *et al.* noted there was **no prior independent evaluation** of Elicit's accuracy, and that these results should be taken as a proof-of-concept (www.ovid.com). Their broader message is that **AI can augment but not fully replace human extraction yet**; even experts extract differently, so "ground truth" is not absolute. Overall, this independent test shows Elicit's extraction ability is at least comparable to experienced human reviewers for many variables (www.ovid.com), validating its potential.

Comparisons with Other AI Approaches

Aside from Elicit-focused evaluations, we consider performance of similar LLM-based extraction in the literature. The proof-of-concept by Gartlehner *et al.* (2024) used Claude 2 to extract 16 pre-specified data items from 10 RCTs ([10] pubmed.ncbi.nlm.nih.gov). Remarkably, they reported **96.3% accuracy** across 160 data points. Claude's main errors were omissions (e.g. missing a data point that was present) but overall it almost matched human correctness. Similarly, Poser *et al.* (2026) built an ensemble ("consensus") of multiple LLMs to extract structured data from 30 multiple sclerosis patient reports. After prompt tuning, their **LLM consensus achieved a true error rate of only ~1.48%**, essentially on par with expert clinicians (~2%) ([48] pmc.ncbi.nlm.nih.gov). These results underscore that modern LLMs are highly capable in medical extraction tasks, provided suitable prompting and verification.

Table 2 summarizes key quantitative findings from several studies. The "Accuracy/Metric" column lists either accuracy percentages or F1 scores as reported. All studies involve extracting defined data fields from clinical or scientific texts.

Study (Year)	Data/Domain	Method/Tool	Performance Metric	Key Findings and Notes
Bianchi *et al.* (2025) (www.ovid.com)	20 RCTs (nursing interventions)	Elicit (LLM-based) vs. human	Data fields: *29.3% "more"*; *20.7% equal*; *45.7% partial*; *4.3% deviating* (www.ovid.com)	Elicit often extracted as much or more than human; some partial captures. AI aids but needs human checks.
Gartlehner *et al.* (2024) ([10] pubmed.ncbi.nlm.nih.gov)	10 RCTs (pain management trials)	Claude 2 (LLM, GPT-like)	*96.3% overall accuracy* ([10] pubmed.ncbi.nlm.nih.gov)	High accuracy across 160 data elements; only 6 errors. Demonstrated strong feasibility of LLM extraction in evidence reviews.
Panayi *et al.* (2023) ([17] systematicreviewsjournal.biomedcentral.com)	Oncology reviews (Cancer)	BERT+CRF (pretrained on biomedical)	F1 (entity recog.): *~73%* (avg across tasks) ([17] systematicreviewsjournal.biomedcentral.com)	With domain pretraining, achieved moderate-high F1 for entity recognition; relation extraction F1 >90% for common relations.
Polak & Morgan (2024) ([31] www.nature.com) ([20] www.nature.com)	Materials science papers (properties)	GPT-4 (ChatGPT) + ChatExtract prompts	*~90.8% precision, 87.7% recall* (bulk modulus); *91.6% precision, 83.6% recall* (metallic glass) ([31] www.nature.com) ([20] www.nature.com)	Using conversational LLM with follow-up queries yielded near-90% precision/recall in zero-shot extraction tasks. Suggests LLM power after engineering.
Poser *et al.* (2026) ([49] pmc.ncbi.nlm.nih.gov)	Clinical reports (multiple sclerosis)	Multi-LLM consensus	*~1.48% error rate* (LLM consensus, true-error)	Combining OpenAI, Anthropic and Google LLMs and optimizing prompts brought errors to ~1.5%, comparable to physician

Study (Year)	Data/Domain	Method/Tool	Performance Metric	Key Findings and Notes
				performance ([49] pmc.ncbi.nlm.nih.gov).
Elicit (case study)	Systematic review (education)	Elicit (LLM-based)	*99.4% extraction accuracy* ([8] elicit.com)	Internal case: 1502/1511 data points correctly auto-extracted.

Table 2: Performance of AI-based data extraction in select studies (in clinical or related domains). Accuracy metrics vary by task. Sources are cited in square brackets.

These data indicate that AI methods – especially LLMs – are approaching *professional* levels of extraction quality if used carefully. Elicit's claimed accuracy (~99%) and independent findings (≤4% major misses) align with other LLM results (95–97%+) in controlled tests ([10] pubmed.ncbi.nlm.nih.gov) (www.ovid.com). It should be noted, however, that extraction quality can vary by domain complexity. For example, capturing a single number from text is generally easier than summarizing an entire methodology. The moderate partial-match rates in Bianchi *et al.* reveal that Elicit sometimes splits hairs or omits detail. Thus, researchers should calibrate their expectations: *AI can handle routine, clearly-stated data extraction very well, but may struggle with nuanced interpretation (e.g. multi-part outcomes or graphical data)*.

Efficiency Gains

While accuracy is often the focus, one of Elicit's major appeals is **speed**. Manually extracting data from even a dozen papers can take days or weeks. In contrast, Elicit automates this once the workflow is set up. The vendor claims *10× faster* in extraction and *up to 80% time savings* overall ([13] support.elicit.com). No independent study has yet measured actual time savings, but anecdotal user reports are consistent: staging the extraction columns and validation takes a fraction of the time of reading whole papers. For example, after initial setup, one can get Elicit to fill a 20×10 table almost instantly, whereas manual entry would take many hours.

Importantly, any time saved must be weighed against the time to learn and configure the AI workflow. New users often need 1–2 hours to master defining good prompts. However, this front-loaded cost is quickly amortized for large projects. Once comfortable, a researcher can apply Elicit to multiple tasks (the presets become templates). A preliminary finding by Bernard *et al.* suggests that Elicit's AI assistance enabled the discovery of additional studies beyond manual search, implying *broadening evidence capture* ([40] bmcmedresmethodol.biomedcentral.com). This expansion of coverage – *"more evidence considered due to automation"* as Elicit advertises ([50] elicit.com) – is an underappreciated efficiency gain, though more formal quantification is pending.

Limitations and Reliability

No AI extraction tool is perfect, and Elicit is no exception. The primary concerns include:

- **Hallucinations or Misinterpretations:** LLMs occasionally fabricate or misunderstand text. Elicit mitigates this via RAG (using actual quotes) but errors still occur (the 4.3% "deviating" in Bianchi *et al.* (www.ovid.com)). For example, if a paper mentions multiple doses, Elicit might pick the wrong one by ambiguity. The "partial/incomplete" 45.7% category in the Cochrane study mostly reflected this kind of issue. Users must be vigilant that the quote matches the intended field.
- **Reproducibility and Stability:** Bernard *et al.* noted that repeated Elicit searches can yield variable results if done at different times ([40] bmcmedresmethodol.biomedcentral.com). This stochasticity comes from AI randomness and changing corpora. Their study found the number of hits varied across repeated queries. For screening, this means Elicit might include slightly different papers on rerun. Thus, for full rigor one should run the pipeline once and fix the included set (or at least document Elicit's steps thoroughly for PRISMA-AI compliance ([51] bmcmedresmethodol.biomedcentral.com)).

- **Scope of Text:** Elicit's extraction is currently limited to *English-language* content and often depends on abstracts/full-text available through its integrations (Semantic Scholar etc.). It may miss data embedded in figures or scanned text, although recent updates claim partial support for table-of-text and figure reading ([35] support.elicit.com) ([52] pmc.ncbi.nlm.nih.gov). Any data solely in non-digital form (like in older PDF tables) could escape.

- **Need for Human Oversight:** Elicit is a tool, not an autonomous agent. It requires careful user validation. The patterns of error in studies suggest that treating AI outputs as tentative is prudent. Organizations like Cochrane recommend dual extraction by humans; with Elicit this becomes: one runs Elicit, another reviews Elicit's table (www.ovid.com) ([11] bmcmedresmethodol.biomedcentral.com). In practice, this hybrid workflow retains the check that two people looked at each data point – one by AI merits, one by human review.

In short, while Elicit dramatically reduces manual drudgery, it demands methodological safeguards. As one source advises, *"the use of an AI tool should only be at certain stages… not for automating the entire process"* ([53] bmcmedresmethodol.biomedcentral.com). Researchers should pre-specify which data will be extracted (e.g. following Cochrane Handbook or Joanna Briggs protocols), record all AI usage, and verify unusual values manually. Given these caveats, the evidence suggests that Elicit can greatly accelerate data extraction **without sacrificing accuracy**, provided its outputs are treated with appropriate scrutiny.

Case Studies and Perspectives

Elicit's adoption and evaluation has begun in diverse contexts. We summarize several relevant cases.

- **German Education Policy (VDI/VDE):** In the development of education guidelines, researchers used Elicit to automate a systematic review. They reported *1502/1511* data points correctly extracted (99.4% accuracy) ([8] elicit.com). This remarkable result, albeit from an internal case, indicates that for well-formatted data (likely numeric and textual items in reports), Elicit can be extremely reliable.

- **CSIRO (Australia):** A published mention (CSIRO) found that Elicit had near-zero false negatives when compared to manual extraction ([43] elicit.com). In other words, anything a human found, Elicit also found (or more). The implication is that Elicit is *sensitive* and casts a wide net. However, it may include irrelevant bits (false positives), which is where human curation is needed.

- **Bernard *et al.* (France, 2025) – Screening Example:** The BMC study by Bernard *et al.* evaluated Elicit in an umbrella review of smart living environments for aging. They used Elicit to replicate their prior review search. Results: Elicit found 3 additional studies not caught previously ([40] bmcmedresmethodol.biomedcentral.com), but only 17.6% of the original included studies were identified . This highlights that Elicit's semantic search can compensate for human oversight to some extent (finding new relevant records) but also misses many if used alone. As a result, they conclude Elicit is a helpful *supplement*, not a replacement.

- **Bianchi *et al.* (Switzerland, 2025) – Extraction Study:** As noted above, this systematic comparison of Elicit vs human on 20 RCTs found that Elicit often captured the same or more detail as the human reviewer (www.ovid.com). It also demonstrated that Elicit's performance may be higher in some fields (e.g. identifying basic study parameters) than others (complex multi-arm interventions). The authors suggest that Elicit can boost completeness and speed during extraction, but final analysis still requires human judgment on data validity (www.ovid.com) ([11] bmcmedresmethodol.biomedcentral.com).

- **Panayi *et al.* (2023) – ML Extraction in Oncology:** Though not about Elicit specifically, this study shows the viability of transformer-based extraction in clinical reviews. Their BERT+CRF model for oncology trials achieved solid F1 scores (~70–88%) ([17] systematicreviewsjournal.biomedcentral.com). They emphasize that pretraining on biomedical text improved accuracy significantly. This context implies that Elicit – which presumably is powered by models continuously updated on academic text – likely benefits similarly from domain-specific knowledge.

- **Polak & Morgan (2024) – Materials Science LLM Extraction:** While in a different field, their ChatGPT-4 "ChatExtract" workflow (zero-shot with prompt engineering) attained ~90% precision/recall ([31] www.nature.com). The methods (iterative question-answering, redundancy) could be conceptually applied to clinical data extraction in Elicit's design. The high performance reinforces that, with thoughtful prompting, ChatGPT-level models can extract numeric data nearly as well as LLMs specifically trained for data tasks.

These cases illustrate multiple perspectives:

- **User Perspective:** Elicit drastically cuts workload. In all reported uses, users emphasize saving time. For example, the Elicit website claims users finish tasks 10× faster ([13] support.elicit.com). Anecdotally, many early adopters (PhD students, clinicians) describe it as having the effect of many "100%-efficient junior researchers" working in parallel. But users also stress the importance of checking AI outputs; none claim to abandon human review completely.

- **Developer Perspective:** Elicit's designers focus on trust, transparency, and integration. Features like showing source quotes, iterative column tuning, and shareable workflows reflect best practices in **human-AI interaction**. The SciDaSynth designers even recommend "highlight [ing] uncertain or missing information" in future tools, echoing requests Elicit users make ([54] pmc.ncbi.nlm.nih.gov). Elicit's recent addition of "High-accuracy mode" (mentioned in the researchgate screenshot description ([55] www.researchgate.net)) suggests ongoing work to balance sensitivity and precision.

- **Research Community Perspective:** The emergence of Elicit has stimulated debate in evidence-synthesis circles. Cochrane and PRISMA groups are developing guidelines for AI in reviews ([53] bmcmedresmethodol.biomedcentral.com). A Cochrane commentary notes that while initial results are promising, rigorous validation is still lacking (www.ovid.com) ([11] bmcmedresmethodol.biomedcentral.com). Critics caution about overreliance and reproducibility. On the other hand, advocates see Elicit as a harbinger of next-generation systematic reviews: more living (constantly updated) and more comprehensive, thanks to AI assistance.

Discussion of Implications and Future Directions

Our review of Elicit's usage and evaluations suggests several **key implications** and directions for the future:

- **Acceleration of Evidence Synthesis:** Elicit exemplifies how AI can *dramatically speed up* systematic reviews. If extraction can be done in hours instead of months, researchers can tackle larger and more up-to-date syntheses. This could expand the scope of evidence-based practice, making living systematic reviews (updated continuously) more feasible ([13] support.elicit.com). Healthcare policy bodies (like WHO or NICE) might increasingly rely on such AI-accelerated workflows to keep guidelines current.

- **Quality and Consistency:** By standardizing extraction protocols through AI prompts, Elicit can improve consistency in data capture (less operator variability). However, slight deviations in phrasing or typographical errors can still trip up the model. Future work may focus on robust pre- and post-processing (e.g. normalizing synonyms, helping the AI disambiguate) to further reduce errors. The SciDaSynth design (with semantic grouping of extracted values ([56] pmc.ncbi.nlm.nih.gov)) points toward interfaces that combine AI with data visualization to catch anomalies. Elicit or similar tools may evolve to include such cross-document analyses, highlighting inconsistent units or outlier values.

- **Integration with Knowledge Graphs and Databases:** Structured extraction yields tabular data that can feed into larger knowledge bases. Imagine connecting Elicit to clinical trial registries or medical databases: results from published trials could automatically populate searchable databases of outcomes. Some of Elicit's data could flow into the same ecosystem as ClinicalTrials.gov. Already, Elicit's structured outputs are amenable to meta-analysis software. Future work could enable one-click uploading of extraction results to RevMan or R meta-analysis packages.

- **LLM Improvements and Customization:** The underpinning LLM engines are rapidly improving. GPT-4 (and successors) have better reasoning and retrieval faculties. Elicit may incorporate newer models to boost accuracy. Moreover, domain-adapted LLMs (like BioGPT or customized ChatGPTs for health) might become available. One can envision an Elicit that runs a domain-specific LLM fine-tuned on clinical trial corpora. This could decrease hallucinations and improve extraction of subtle medical language. Reinforcement-learning-from-human-feedback (RLHF) specialized for data extraction might also refine its answers.

- **Multipronged AI Strategies:** As seen in Poser's "LLM consensus" approach ([49] pmc.ncbi.nlm.nih.gov), combining outputs from multiple models can increase reliability. Elicit might in future take such an ensemble approach: e.g., running each query through GPT-based and Claude-based backends and reconciling differences. This could potentially lower the error rates further (currently ~4% major errors (www.ovid.com), but consensus might reduce that to negligible levels).

- **Better Evaluation Metrics:** The community will need standardized benchmarks to assess these tools. Just as there are GLUE benchmarks for NLP, we might see shared extraction tasks for systematic review data (perhaps as extensions of the CADIMA or Cochrane AI projects). Metrics beyond plain accuracy (like completeness, precision-recall on key fields) should be reported. The work by Panayi *et al.* ([17] systematicreviewsjournal.biomedcentral.com) scoring F1 may serve as a template: future studies should use comparable metrics so that performance can be meaningfully aggregated across studies.

- **Ethics and Transparency:** From a policy standpoint, any AI-facilitated review must be transparent about its use. Updates to PRISMA (for example, forthcoming PRISMA-AI) will likely require authors to disclose that Elicit or similar tools were used ([53] bmcmedresmethodol.biomedcentral.com). Journals and guideline committees may begin to ask authors to supply extraction tables with AI-flagged fields and human edits. Importantly, AI tools should not become black boxes in the literature – the supporting quotes and rationales must be published or archived alongside the data.

- **Education and Training:** Finally, the uptake of Elicit implies new skills for researchers. Training programs for evidence synthesis will likely include modules on “how to use LLM tools responsibly”. Librarians and review methodologists must become adept at guiding AI-driven reviews. This raises equity issues as well – not all institutions have access to paid versions of Elicit, and knowledge-sharing (e.g. through saved Q&A prompts) will be vital.

Looking forward, the **future of structured data extraction** in clinical research is bright but nuanced. Tools like Elicit are early steps toward an integrated AI-assisted research environment. Further improvements may close the loop: e.g. linking extracted data to statistical analysis pipelines or dynamic knowledge bases, enabling “living” meta-analyses that update as new studies arrive.

Conclusion

This report has examined the methodology and evidence for using **Elicit** to extract structured data from clinical research papers. We have shown that Elicit supports a full systematic review pipeline, from question formulation through to final report, with special emphasis on automating data extraction into tables. Practical guidance was provided on how to define research questions, gather literature, set up extraction columns, and verify outputs using Elicit’s interface ([5] support.elicit.com) ([7] support.elicit.com).

Drawing on case studies and research, we highlighted that Elicit can achieve high accuracy in data extraction – often comparable to trained humans (for example, >95% of data fields correct in internal tests ([8] elicit.com) and independent comparisons showing the same or more information extracted in ~50% of cases (www.ovid.com)). Other LLM-based studies (e.g. Claude 2 and GPT-4) similarly demonstrate 90–96% accuracy on clinical data tasks ([10] pubmed.ncbi.nlm.nih.gov) ([31] www.nature.com). These results indicate that large language models, when guided properly, can substantially reduce the workload of systematic reviews.

We have emphasized that **human oversight remains essential**. Researchers must carefully check Elicit’s outputs against source texts ([25] support.elicit.com). Any AI extraction should be viewed as a highly-efficient second reader rather than a final arbiter. We also noted limitations (variable reproducibility ([40] bmcmedresmethodol.biomedcentral.com), occasional errors) and the need for adherence to emerging best-practice guidelines ([53] bmcmedresmethodol.biomedcentral.com). Despite these caveats, Elicit’s potential is clear. It allows evidence syntheses to be **ten times faster** while maintaining rigor, unlocking the ability to review more literature in less time ([13] support.elicit.com). In an era of information overload, this acceleration could prove transformative.

In conclusion, to effectively use Elicit for structured data extraction:

- **Input clear, focused queries and let Elicit suggest fields.** Use its semantic search and screening to gather papers comprehensively ([5] support.elicit.com).
- **Define and iteratively refine your data columns.** Use the pilot extraction feature to ensure each column prompt yields the intended data ([35] support.elicit.com). Leverage Elicit’s suggested columns as a starting point ([35] support.elicit.com).
- **Run the full extraction and verify thoroughly.** Check each extracted cell against the original text, correcting as needed. Treat Elicit’s output as draft data that must be reviewed.
- **Document your AI use.** For transparency, note in your methods which tasks Elicit automated and include any PRISMA-AI checklist items. The in-app citations provide a starting point for traceability.

With these guidelines, researchers can harness Elicit’s AI capabilities to greatly speed up the extraction of structured clinical data, while maintaining accuracy. The evidence suggests that, within a properly managed workflow, Elicit is a powerful ally in systematic reviews. As LLM technology continues to advance, such AI tools will likely become an indispensable component of evidence synthesis, enabling healthcare decisions to be based on broader and faster analyses of the literature than ever before.

External Sources

[1] https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-023-02351-w#:~:Owing...

[2] https://pubmed.ncbi.nlm.nih.gov/38432227/#:~:Data%...

[3] https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-023-02351-w#:~:makes...

[4] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02528-y#:~:Among...

[5] https://support.elicit.com/en/articles/7927169#:~:Elici...

[6] https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-10-56#:~:Clini...

[7] https://support.elicit.com/en/articles/4267649#:~:From%...

[8] https://elicit.com/solutions/literature-review#:~:VDI%2...

[9] https://www.ovid.com/journals/cesm/fulltext/10.1002/cesm.70033\~data-extractions-using-a-large-language-model-elicit-and#:~:Data%...

[10] https://pubmed.ncbi.nlm.nih.gov/38432227/#:~:used%...

[11] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02528-y#:~:compl...

[12] https://pmc.ncbi.nlm.nih.gov/articles/PMC12581027/#:~:retri...

[13] https://support.elicit.com/en/articles/7927169#:~:With%...

[14] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02528-y#:~:the%2...

[15] https://pmc.ncbi.nlm.nih.gov/articles/PMC12581027/#:~:1...

[16] https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-023-02351-w#:~:estim...

[17] https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-023-02351-w#:~:For%2...

[18] https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btad542/7260503#:~:trial...

[19] https://www.nature.com/articles/s41467-024-45914-8#:~:Altho...

[20] https://www.nature.com/articles/s41467-024-45914-8#:~:%28i,...

[21] https://elicit.com/?run=17be867037aebb4608590aa818d6fa10&workflow=table-of-papers#:~:,you%...

[22] https://support.elicit.com/en/articles/7927169#:~:1,pap...

[23] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02528-y#:~:power...

[24] https://support.elicit.com/en/articles/7927169#:~:Data%...

[25] https://support.elicit.com/en/articles/4267649#:~:Save%...

[26] https://support.elicit.com/en/articles/7927169#:~:confi...

[27] https://pmc.ncbi.nlm.nih.gov/articles/PMC12581027/#:~:paper...

[28] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02528-y#:~:infor...

[29] https://elicit.com/?run=17be867037aebb4608590aa818d6fa10&workflow=table-of-papers#:~:Elici...

[30] https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-023-02351-w#:~:ExaCT...

[31] https://www.nature.com/articles/s41467-024-45914-8#:~:overc...

[32] https://pmc.ncbi.nlm.nih.gov/articles/PMC12581027/#:~:the%2...

[33] https://support.elicit.com/en/articles/7927169#:~:Conso...

[34] https://support.elicit.com/en/articles/7927169#:~:Scree...

[35] https://support.elicit.com/en/articles/4267649#:~:Elici...

[36] https://support.elicit.com/en/articles/4267649#:~:Step%...

[37] https://support.elicit.com/en/articles/4267649#:~:Elici...

[38] https://support.elicit.com/en/articles/7927169#:~:Defin...

[39] https://support.elicit.com/en/articles/7927169#:~:By%20...

[40] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02528-y#:~:The%2...

[41] https://support.elicit.com/en/articles/4267649#:~:Click...

[42] https://support.elicit.com/en/articles/4267649#:~:Once%...

[43] https://elicit.com/solutions/literature-review#:~:Resea...

[44] https://elicit.com/solutions/literature-review#:~:94%E2...

[45] https://www.ovid.com/journals/cesm/fulltext/10.1002/cesm.70033\~data-extractions-using-a-large-language-model-elicit-and#:~:Our%2...

[46] https://www.ovid.com/journals/cesm/fulltext/10.1002/cesm.70033\~data-extractions-using-a-large-language-model-elicit-and#:~:their...

[47] https://www.ovid.com/journals/cesm/fulltext/10.1002/cesm.70033\~data-extractions-using-a-large-language-model-elicit-and#:~:2%20B...

[48] https://pmc.ncbi.nlm.nih.gov/articles/PMC12946029/#:~:Throu...

[49] https://pmc.ncbi.nlm.nih.gov/articles/PMC12946029/#:~:Resul...

[50] https://elicit.com/?run=17be867037aebb4608590aa818d6fa10&workflow=table-of-papers#:~:10x...

[51] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02528-y#:~:not%2...

[52] https://pmc.ncbi.nlm.nih.gov/articles/PMC12581027/#:~:extra...

[53] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02528-y#:~:Given...

[54] https://pmc.ncbi.nlm.nih.gov/articles/PMC12581027/#:~:match...

[55] https://www.researchgate.net/figure/Screenshot-of-the-Elicit-interface-showing-the-data-extraction-setup-The-interface_fig1_392328691#:~:Image...

[56] https://pmc.ncbi.nlm.nih.gov/articles/PMC12581027/#:~:match...

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.