

Can AI support rapid evidence synthesis for policy? An evaluation of Elicit: The AI Research Assistant

Gemma Spiers

`gemma-frances.spiers@newcastle.ac.uk`

Newcastle University

Jane McDermott

University of Manchester

Kimberly Green

University of Manchester

Patience Kunonga

Newcastle University

Robert Barker

Newcastle University

Alex Hall

University of Manchester

Jabez Barnabas

Newcastle University

Hannah O'Keefe

Newcastle University

Barbara Hanratty

Newcastle University

Research Article

Keywords:

Posted Date: October 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9048478/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Can AI support rapid evidence synthesis for policy? An evaluation of Elicit: The AI Research Assistant

Dr Gemma Spiers

Professor Jane McDermott

Dr Kimberly Green

Miss Patience Kunonga

Dr Robert Barker

Dr Alex Hall

Mr Jabez Barnabas

Ms Hannah O'Keefe

Professor Barbara Hanratty

Acknowledgements

We gratefully acknowledge the input of colleagues at the Department of Health and Social Care, the NIHR Policy Research Programme, and the NIHR Healthy Weight Policy Research Unit for sharing their experience and expertise in the development and delivery of this work.

Funding statement

This research is funded through the National Institute for Health and Care Research (NIHR) Policy Research Unit in Healthy Ageing (funding reference NIHR206119). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Declarations

None.

Ethics

Ethical approval was not required for this study as no human participants were involved.

Contributions

GS and JM co-designed the study protocol. All authors contributed to the evaluation components detailed in the paper. GS led on writing the manuscript, and all authors reviewed the manuscript and contributed to the draft.

Abstract

Background: Advances in artificial intelligence (AI) have generated platforms that expedite review processes, potentially saving time and resources. As new platforms continue to develop, ongoing scrutiny of performance is critical. In this study, we evaluated Elicit, an AI assisted review tool. Two potential applications of this software were tested: a) as a tool to aid preliminary scoping of a topic, and b) as a rapid review tool.

Methods: Two test cases drawn from the authors' recent work were used as the benchmark to examine the performance of Elicit: a scoping of physical activity measures, undertaken to inform the development of a research question, and a rapid systematic review of interventions to promote physical activity in disadvantaged older populations. Both the scoping and rapid review were mimicked using the AI platform. For both tests, we used Elicit Pro. The process of using Elicit was documented and Elicit outputs evaluated according to an a-priori framework. Sensitivity and precision were calculated for both the Elicit and researcher-led reviews.

Findings: As a tool to aid preliminary scoping, Elicit produced a useful and broad ranging summary of evidence in a short timeframe. As a tool to deliver rapid evidence reviews in response to complex questions, Elicit made errors at all stages of the review process, and included a large volume of ineligible material: all of which may limit its ability to produce an accurate and impartial account of the evidence. Sensitivity and precision for the Elicit review were 8.3% and 0.8% respectively; and 100% and 0.7% for the human-review respectively.

Discussion: Elicit has potential value to policy researchers. It can usefully expediate preliminary scoping of topics, and it can support some individual components of rapid reviews. It requires input by researchers into decision-making and to check accuracy. The complexity of the rapid review topic in our study challenged the AI capabilities of Elicit. Researchers may need to invest time testing different prompts to optimise outputs. The potential time and resources saved by using Elicit in this way is unclear.

Can AI support rapid evidence synthesis for policy? An evaluation of Elicit: The AI Research Assistant

Background

Systematic and rapid reviews are a core tool to support health and social care policy decision making.¹ More recently, rapid reviews have gained popularity as a method to match the benefits of evidence syntheses to the time-limited needs of policy makers. The methods applied within rapid reviews mirror the same components of systematic reviews but with aspects streamlined for efficiency.^{2,3} Advances in artificial intelligence (AI) have generated platforms that further expedite review processes, potentially saving time and resources. Platforms such as Elicit (an AI assisted review tool which is semi-manual and allows investigators to override at every step of the review journey) are increasingly being used across research and policy organisations to support evidence synthesis.⁴

Recent evaluations have highlighted the value of Elicit in scoping material to inform review questions and act as the 'second' reviewer.⁵⁻⁹ The data extraction capabilities of Elicit have been shown to be comparable to those of humans.^{8,9} However, there is currently insufficient evidence to suggest that Elicit could fully replace reviews conducted by humans. Reported weaknesses include the use of limited search sources, a lack of repeatability with screening decisions, and failure to consistently apply eligibility criteria in screening decisions.⁵⁻⁷ More broadly, integrating AI into health and care policy research brings risks that require ongoing scrutiny. The replicability of methods is fundamental to robust scientific enquiry, yet AI models can lack transparency, specifically around how they source and process information.^{10,11}

AI-assisted evidence syntheses will inevitably be widely employed. The question is not *should they be used*, but *how can they be reliably and usefully applied by health and social care researchers*. Addressing this requires ongoing appraisal of new tools and platforms as they develop. The current evidence base for Elicit is small, with only five evaluations published to date, two of which were tests specific to its data extraction capabilities.⁵⁻⁹ Elicit also continues to refine and evolve as a platform. Further and continued scrutiny is therefore critical to fully understand its value and role in the health and social care research context. For policy researchers, where rapid reviews of evidence are often required with limited resources and in short time periods, Elicit potentially offers a valuable tool to aid the efficient delivery of evidence synthesis outputs.

As researchers navigate the role of AI in evidence synthesis, an important consideration is how AI-driven review software responds to complex topics. Our experience as policy researchers points to an increasing complexity in review topics in response to the evidence needs of health and social care decision makers. For example, it is unusual for our current rapid reviews to consider just one type of intervention. More typical is a review that considers a range of interventions in the effort to understand what works to address a particular set of health goals. Furthermore, our scrutiny of intervention impact often spans a wide range of outcomes and outcome measures. Finally, our rapid reviews typically examine effectiveness by population subgroups to explore the health equity of interventions. If AI-supported tools such as Elicit are to support the delivery of evidence syntheses in health and social care policy, it is critical that they are able to do so for complex review topics.

In the work reported here, we aimed to evaluate the performance of Elicit as an evidence synthesis tool in the context of rapid, health and social care policy-focused research. We focus on two potential roles for this application:

- As an aid to develop and refine complex policy questions, which often involves scoping literature to inform research questions and approach;
- As a rapid review evidence synthesis tool.

Our objectives were to identify whether Elicit can be used as a) a research-scoping aid, and b) a rapid review evidence synthesis tool.

Methods

We undertook our evaluation of Elicit in two phases. Phase 1 addressed objective a, and phase 2 addressed objective b. For both tests, we used Elicit Pro (the largest of Elicit's three available individual licence packages) to test the platform at maximum capacity. Elicit Pro allows researchers to override its functions at any point. We used the version of Elicit Pro available in November 2025. For our evaluation, we limited researcher input (i.e. overriding Elicit's decisions) to only what was necessary to retain the focus on the original review topic. We specify the nature of this researcher input in the relevant sections below and in the supplementary materials. Our decision to limit researcher input was made to explore Elicit's automated capabilities in response to the review question posed. For both the phase 1 and 2 evaluations, we used two recently completed researcher-led reviews as the benchmark for comparison. We use the term 'researcher-led' to refer to evidence syntheses that are led and conducted by researchers as opposed to reviews that are predominantly conducted by AI software. The methods used for each evaluation are detailed below.

Phase 1: Can Elicit be used as a scoping aid?

In phase 1, we used a recent request from a governmental departmental policy team, which was scoped as a joint exercise between the NIHR Policy Research Unit in Healthy Ageing (HAPRU) and the NIHR Policy Research Unit in Healthy Weight (HWPRU). The policy requirement was to explore physical activity status assessment tools for primary care that can identify patients suitable for physical activity interventions. In response to this request, a brief literature review was undertaken as a form of scoping: this intended to inform the development of a research question on physical activity tools. In this context, 'scoping' does not refer to a *scoping review*, which is a type of evidence synthesis method.¹² Scoping in this context refers to the act of summarising literature on a broad topic to better understand coverage and gaps, and inform the need for a review or decide the review's focus. Such scoping typically relies on searching a single bibliographic database or repository for studies and reviews and produces a high-level summary of evidence. The scope is often informed by a broad framework rather than the strict criteria applied in systematic reviews.

To evaluate how well Elicit can scope and summarise literature to aid our thinking on a given policy question, we mimicked the unpublished brief literature review using the AI platform. This involved asking Elicit to summarise evidence and gaps on the topic (see supplementary materials for the prompt used). The process of using Elicit, and the approximate time spent by the researcher entering the review question and navigating the subsequent requests from the platform, was documented in an Excel spreadsheet.

The Elicit output was then appraised by topic experts in HAPRU and the HWPRU to explore depth, coverage and any important omissions. The purpose of this appraisal was to understand the Elicit output in terms of its utility to inform thinking on the topic of physical activity measures. We did not evaluate specific review methods for this phase, such as screening decisions and eligibility. This is because our parameters were broad and flexible as is typical when undertaking preliminary scoping.

Phase 2: Can Elicit be reliably used as a rapid review evidence synthesis tool in policy research?

In phase 2, we evaluated whether Elicit could be reliably used to undertake a rapid review of evidence. For this evaluation, we used a rapid review completed by HAPRU in 2025 about interventions targeting individuals and small groups to promote physical activity in older disadvantaged populations.¹³ To examine effectiveness in disadvantaged populations, we utilised studies that a) evaluated effectiveness in specific disadvantaged groups, or b) evaluated effectiveness in the general population but reported outcome data stratified by

levels of disadvantage. Disadvantaged populations were operationalised using the PROGRESS-Plus (Place; Race/ethnicity/culture/language; Occupation; Gender and sex; Religion, Education, Socioeconomic status, Social capital; plus additional context-specific characteristics (e.g., age, disability)) framework, a tool commonly used to provide focus on personal characteristics and social factors important to health equity.^{14,15}

The review focused on the effectiveness of interventions to promote physical activity, with a range of measures eligible (Supplementary materials Table 1). Physical function (e.g. gait speed, muscle function) and physical health outcomes not directly related to exercise intensity (e.g. cholesterol) were not eligible. Standard rapid review methods were used, including searches of grey and academic literature, screening, data extraction, study quality appraisal, an assessment of the certainty of evidence, and data synthesis (narrative, aided by forest plots of pooled data). Forty-seven studies were included in this review.

This review was chosen as a test case for our phase 2 evaluation as it represents a complex question. That is, the review covers multiple types of interventions across several dimensions of disadvantage (for example, socioeconomic status, ethnicity, gender etc), and with a range of eligible physical activity and process evaluation outcomes. The review was unpublished at the time of this evaluation.

Procedure

We undertook our test in two attempts. In the first attempt, we entered the original review aim and parameters into the Elicit prompt box. At this stage, Elicit evaluates the entered text and asks the user to refine the prompt and/or add specific details of the review parameters (e.g. outcome, population type). To give Elicit as much detail as possible at this stage, we entered full details of the criteria along with the review aim. After Elicit chooses keywords, search strategy and search sources, there is an opportunity to refine the suggested criteria that Elicit generates based on the prompt. In response, we refined only one criterion: the use of any or no comparator in the study design, which was missing from the criteria generated by Elicit. All other criteria generated matched the researcher-led review criteria based on the detailed prompt entered at the initial stage. Elicit then generates screening, data extraction and a final report. The report generated by this attempt contained largely irrelevant material that did not answer the review question, and after reading further guidance from the Elicit website, we made a second attempt using a different approach.

In the second attempt, we reworded the prompt into a question (*What is the effectiveness of individual-level interventions to promote physical activity among disadvantaged and marginalised older adults?*). The review question was entered, but without specific details of the review parameters. We then opted to refine and manually add criteria separately when

prompted to do so by Elicit. As no details were added with the review question, we manually added prompts to reflect the review criteria. The criteria in both attempts one and two were broadly similar. The main differences were that in the second attempt, the outcome criteria were split into separate question prompts per outcome, and physical activity was specified as the primary outcome (Supplementary materials Table 2).

The process of using Elicit was documented in an Excel spreadsheet, including how Elicit analysed the question, and made requests for additional input from the researcher. The approximate time spent by the researcher entering the review question and navigating the subsequent requests from Elicit were also recorded.

Data and analysis

The data used in our evaluation were the Elicit report output and all associated data extraction and screening files produced by Elicit. These outputs were assessed using an evaluation framework (Supplementary materials table 3). The framework broadly aligns to the process of a researcher-led review: searching and retrieving records, record screening, data extraction and quality assessment, and data synthesis. For each component of the evaluative framework, assessments were undertaken by members of the research team and documented in Excel. Where necessary, the Elicit output was compared to the researcher-led review report. To allow some independence in the evaluation, the use of Elicit for each the phase 1 and 2 test cases were undertaken by researchers who were not involved in the original scoping exercise (phase 1) or rapid review (phase 2).

Findings

Phase 1: Can Elicit be used as a scoping aid?

From entering the initial prompt into Elicit to generating the report, the process of completing the review took six minutes. The researcher inputted into this process by refining the criteria (Supplementary materials table 2) to ensure they reflected studies of physical activity tools and measures used in primary care settings. This input took the researcher approximately two minutes. As the purpose of this test case was to explore Elicit's performance in broad level literature scoping, rather than taking a systematic approach to evidence synthesis, the researcher made no further input on the review methods.

The generated report was considered a useful summary by the field experts who judged the content. The evidence described was viewed as an indication of what sort of studies and gaps might be expected with further, systematic searches. Field experts noted that the

report contained the evidence they might expect on physical activity tools in primary care. The ability to generate the report in a short time frame was considered especially valuable in the context of time-limited policy research.

The lack of transparency in the methods underpinning the report was noted (we explore this further in the phase 2 test below). With little detail on how studies were identified and selected, the team noted that the report could only be considered a superficial overview of the evidence. Nevertheless, expert feedback was that this preliminary overview of evidence could serve as a valuable foundation for the scoping process.

Phase 2: Can Elicit be reliably used as a rapid review evidence synthesis tool?

From entering the initial prompt to generating a report, the process of completing the review took just over three hours. The time required for this work was substantially longer than the time required for the phase 1 review. We suspect this was partly because the researcher-led review used in this exercise utilised formal rapid review methods, and therefore additional time was needed to match the Elicit methods as closely as possible. Approximately one hour of this time was spent by the researcher manually responding to, and refining, the methods offered by Elicit. The majority of this time reflected the need to add and refine the review criteria (see supplementary materials table 2). Other input from the researcher included specifying the type of data to be extracted for one data extraction field proposed by Elicit. We did not provide input to other parts of the review, such as the search strategy, screening and selection of studies, or synthesis.

In the following section, we present the findings from our evaluation. These findings are based on the report generated from the second attempt to use Elicit (see *Methods*).

Review Searches

Table 1 summarises the search strategy produced by Elicit. For comparison, we also show the search strategy applied in the researcher-led review. Key differences include the absence of specific measures of physical activity in the Elicit strategy, the use of specific indexing terms in the researcher-led strategy, and the more nuanced terminology relating to population age, again in the researcher-led strategy. Search sources used by Elicit were Semantic Scholar, OpenAlex, PubMed, ClinicalTrials.gov, with the same strategy applied to all.

Elicit's screening begins with 500 documents identified by the search. However, it was not clear how these 500 documents were selected. Selection does not appear to be determined by date of publication, with a roughly even spread of publication dates across the 2015-2025

period specified: 2015 (7%), 2016 (6.4%), 2017 (9%), 2018 (10.2%), 2019 (10.6%), 2020 (11.4%), 2021 (8%), 2022 (12%), 2023 (8.6%), 2024 (10.8%), 2025 (6%).

Table 1. Search strategies used by Elicit and in researcher-led review on the same topic

Researcher-led review search strategy	Elicit search strategy output based on prompt
1 ((promot* or encourag* or boost* or elevat* or develop* or increas*) adj1 (physical activit* or exercis* or fitness or train*)).ti,kf.	("individual intervention" OR "individual-level intervention" OR "personal intervention")
2 Exercise/	AND
3 Exercise Therapy/	("physical activity" OR exercise OR "physical exercise")
4 1 or 2 or 3	AND
5 ((ag* adj3 ("50" or "60" or "65" or fifty* or sixty*)) or senior* or older adult* or aged person* or advanced age or senescen* or old ag* or geriatric* or senium or frail elder* or elderly or middle aged or older person or older people).kf,tw.	(disadvantaged OR marginalised OR marginalized OR "socioeconomically disadvantaged" OR "low income" OR vulnerable OR underserved)
6 exp aged/ or middle aged/	AND
7 5 or 6	("older adult" OR elderly OR senior OR aged)
8 ((physical activit* or physical exercise* or PA) adj2 (score? or level? or calculat* or moderate or vigorous or ratio? or increas* or index* or indices or measur* or pedomet* or acceleromet* or MVPA or PASE or intensi* or METs or "heart rate" or VO? max or "calori* consumption*" or "metabolic equivalent*" or IPAQ or CHAMPS or SQUASH or BPAQ or actigraph* or PASB-Q)).kf,tw.	AND
9 Accelerometry/	("50 years" OR "age 50" OR "over 50")
10 Actigraphy/	
11 8 or 9 or 10	
12 4 and 7 and 11	
13 Humans/ not (Animals/ and Humans/)	
14 12 and 13	
15 health promotion/ or healthy people programs/ or weight reduction programs/	
16 ((intervention* adj8 (physical or activi* or exercis* or fitness or train* or app or health* or promot*)) or program*).mp.	
17 15 or 16	
18 14 and 17	
19 limit 18 to yr="2015 -Current"	

Study selection: screening

Elicit provided an Excel spreadsheet detailing the 500 studies screened and the decision in response to each criterion. Judgements for each criterion were coded as *yes*, *no* or *maybe*.

A final decision (include or exclude) was based on a score summed across all criteria judgements from 0-5. Of the 500 studies screened, Elicit deemed 243 studies eligible. These 243 studies all scored at least 3.9. The excluded studies were all 3.5 or below.

The report is based on 'the 40 sources that had the highest screening scores out of the 243 sources' (see also *Included Studies*, below). At the time of this evaluation, the Elicit report was limited to 40 studies due to the capabilities of the LLM used. Of these 40, eligibility scores were 4.9 (n=8), 4.7 (n=4), and 4.6 (n=28). However, a further 30 studies not in the report also scored 4.6. No details were provided on how Elicit selected the particular 28/58 studies that scored 4.6 for the report.

Of the 40 studies included in the report, two were judged by Elicit to have met all screening criteria. The remaining 38 studies were given judgements of *maybe* on between one and four criteria. All of the 40 report studies were judged to meet the following criteria: individual level intervention, study design, age of population, age-specific results and physical activity focus. The remaining criteria (e.g. publication date, disadvantaged population) were judged as *maybe*.

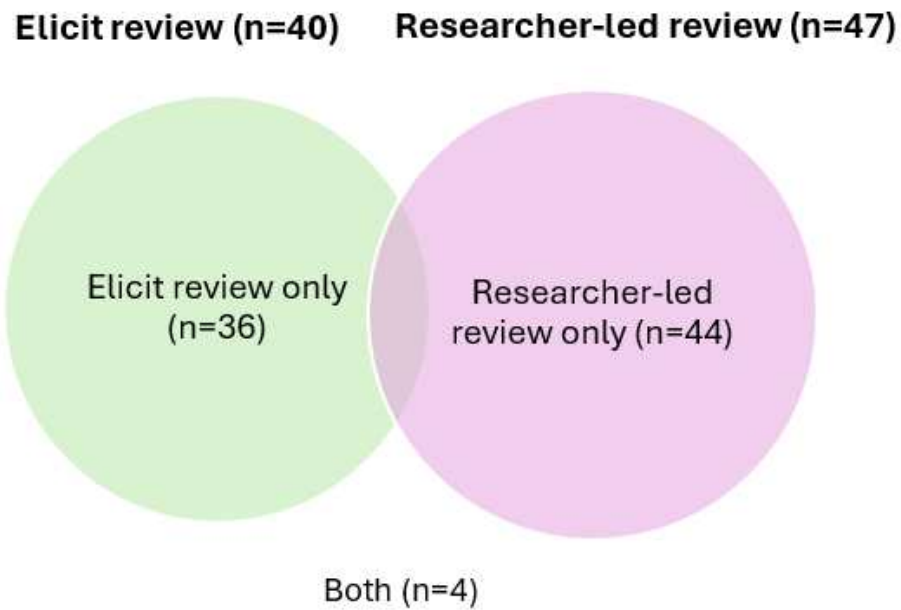
Of the 30 studies that were given a score of 4.6 but not included in the report, none met all criteria. However, similar to the studies included in the report, all of these 30 studies were judged to meet the following criteria: individual level intervention, study design, age of population, age-specific results and physical activity focus. This means that nine of the 40 included studies had at least three 'maybe' judgements (all of which scored 4.6), but 17 of the 30 studies not in the report (which also scored 4.6) had two or fewer 'maybes.' In other words, some of the studies used in the report appeared to have more uncertainty in the screening decision than others that were not used in the report.

The screening data file included a summary of the overall reasoning for a screening judgement including where there is uncertainty. This offers some transparency in the judgements and the include/exclude decision applied. However, it is not clear whether or how Elicit weighed criteria to determine the original scores. For records that obtained the same score of 4.6, it is also not clear how Elicit judged which to use in the report.

Study selection: included studies

Of the 40 studies in the Elicit review, three were included in the researcher-led review (Figure 1). One further study in the Elicit report was eligible and identified in the researcher review but ringfenced for a separate synthesis and is thus not counted within the 47 studies attributed to the report. The remaining 36 studies in the Elicit report did not meet the eligibility criteria (Table 2).

Figure 1. Proportional Venn diagram comparing studies included in Elicit and researcher-led reviews*



**One study was identified by the researcher-led review but later ringfenced for a separate synthesis and thus not included in the researcher-led report. This paper is included within the '4' common studies for transparency, but is not counted within the 47 studies attributed to the researcher-led report.*

Table 2. Reasons why studies included in the Elicit review did not meet the eligibility criteria

Criteria	Number of Elicit studies that did not meet criteria (n)*
The intervention was not evaluated in, or data were not stratified by, an eligible PROGRESS+ population	16
Ineligible intervention	11
Ineligible outcome	9
Protocol	4
Conference abstract	4
Ineligible population age	1
Systematic review	1

* Numbers do not sum due to some records having multiple reasons for ineligibility.

Of the 47 studies in the researcher-led review, three were included in the Elicit report. Of the remaining 44 studies, ten were retrieved during Elicit's screening of 500 titles and abstracts. Of these ten, nine were assessed by Elicit as ineligible, with one study judged eligible (scored 4.3 by Elicit) but not included in the report. The reasons for ineligibility given for the nine studies were: failure to meet the PROGRESS+ criteria, or not published in a high-income country. However, these judgements were incorrect. The studies judged not to meet the PROGRESS+ criteria were in fact eligible as they reported effectiveness data by PROGRESS+ criteria (sex/gender, rural place of residence, education). The study judged not to be from an OECD high income country was published in Singapore, which is classed as a high-income country by the OECD. The remaining 34/44 studies included in the researcher-led review were not identified by the Elicit searches.

To understand why 34 (out of 44) studies in the researcher-led review were not identified by the Elicit searches, we explored the features of the studies, including study design, PROGRESS+ population, sample size, number of study arms, outcome measures, length of evaluation follow up and journal and country of publication. However, we did not identify any clear patterns that would explain why these studies were not identified by Elicit. The completeness of the search sources may only account for the omission of some of these papers. For example, these 34 studies are published in 27 unique journals, of which 24 (89%) are in PubMed, 22 (81%) are in Semantic Scholar and 11 (41%) are in Open Alex.

The 40 studies included in the Elicit report were published between 2015 and 2024. The geographical span of studies was similar to that in the researcher-led review, although the majority of the studies were not eligible. We could detect no evidence to suggest that the included studies were biased towards any particular group or setting.

Sensitivity, which measures the proportion of relevant studies identified from all relevant studies in existence, was low for Elicit (Table 3). In contrast, sensitivity was high for the researcher-led review. A sensitivity calculation requires an estimate of the number of relevant studies in existence. As we found no eligible studies in the Elicit review that the researcher-led review did not identify, we used the number of studies identified by the researcher-led review as the estimate of total relevant studies. This accounts for the high sensitivity figure attributed to the researcher-led review. However, we acknowledge that the researcher-led review will not be exhaustive, and the true sensitivity figure will be less than 100%. Precision, which measures the proportion of relevant studies from the total retrieved, was similar for both the Elicit and researcher-led review. Detailed formula and data for these calculations are in the Supplementary Materials.

Table 3. Sensitivity and precision

	Sensitivity	Precision
Elicit review	8.3%	0.8%
Researcher-led review	100%	0.7%

Data extraction and quality assessment

Elicit provided an Excel file of data extracted from all studies it judged to be eligible (n=243), which in this test case, exceeded the 40 studies included in the final report. When using Elicit, there was an opportunity to amend the data extraction field headings. In this test case, we amended the 'study population' field heading to request that data be extracted about the PROGRESS+ factors considered in the study. When compared with the researcher-led review, there were similarities in some data extraction fields, such as study characteristics, overview of the interventions, outcomes and adherence/attrition. In addition to the Excel file, data tables were included in the report.

Descriptive information about each study was detailed and could be easily exported as a table if required. However, the completeness and clarity of the outcome data on effectiveness that Elicit extracted was variable, in both the Excel file and report. Effect estimates were extracted for some but not all studies. One field provided in the Excel file pinpoints an associated table in the publication for outcome data, although these fields were not complete for every study. It is not clear if the absence of data in the file meant that no associated data were reported in the paper.

A quality appraisal is not yet available within Elicit, therefore this component of the evaluation was omitted.

Synthesis

The synthesis was a narrative summary. The summary was wide ranging, however, much of the generated content did not answer the review question. The results section comprised five sections: summary of effectiveness, secondary outcomes, costs data, differential effects, and barriers and facilitators to using the intervention.

The summary of effectiveness signposted to impact on physical activity (*'The majority of studies with effectiveness data reported positive effects on physical activity outcomes'*). Even so, most of the studies detailed in this section did not report physical activity outcomes, with non-eligible outcomes such as mobility and gait speed described. The sections on costs and secondary outcomes did not answer the review question, describing outcomes that were not within the review purview. The summary of differential effects referred to population sub-groups, however, most of these sub-groups were not eligible on the PROGRESS+ criteria that we specified. There was a brief description of effectiveness of an intervention by socioeconomic status, but for a non-eligible outcome (mobility disability) and a non-eligible intervention. Very little of the synthesis linked effectiveness to disadvantaged groups, as per the review question.

A section in the report described implementation factors, including identifying the intervention components that facilitated or hindered success. The abstract also identified the components required of interventions to achieve effectiveness. However, it was not clear what analysis was undertaken to reach these claims. For example, synthesis approaches such as Intervention Component Analysis, and analysis of qualitative data on barriers and facilitators, can be used to identify intervention features linked to effectiveness. No details were provided as to whether these approaches were applied here.

Finally, four tables were included the report, detailing study characteristics, study populations, intervention details and brief outcome data. The tables reporting descriptive information were useful and mirror that offered in a typical review output. However, three of the four tables were incomplete in terms of studies included. That is, not all 40 studies in the report were included in tables where this would be expected. For example, the study population table included 32 of the 40 studies. The intervention characteristics table included 21 of the 40 studies. No details were provided as to why the remaining studies were not included in these tables. The table of physical activity outcome data included only a limited number of studies. This likely reflects that a proportion of studies did not report eligible physical activity outcomes, despite this specified as a criterion (primary outcome) of the review.

Report Conclusions

The conclusions for each the Elicit and researcher-led review differ markedly. This may be due to the largely different material drawn on in each review. Key points of divergence in the summary of findings and conclusion between the two reviews are:

- The conclusion in the Elicit review rested on a number of outcomes that were not eligible (i.e. non-physical activity outcomes).
- The only reference to specific disadvantaged groups or PROGRESS+ factors in the Elicit summary was in relation to socioeconomic status and income, which cited non-eligible evidence. In the Elicit review, there was no mention of intervention effectiveness by gender and ethnicity, both of which feature in the summary and conclusions of the researcher-led review.
- There was no explicit reference to evidence gaps in the Elicit review conclusion and summary, whereas the researcher-led review identified this as a major limitation of the evidence.
- The quality appraisal used in the researcher-led review meant that conclusions were caveated by the quality and certainty of the evidence. As Elicit did not apply an appraisal, the conclusions do not make reference to the quality of the evidence.

Outputs

Elicit provided three outputs: a report, and two excel files containing screening results and data extraction. The report provided was concise and included data tables and references. There was no flow diagram detailing searches and screening results, although it would be possible to produce one based on the screening data. Whilst the summaries of evidence within the report were clear and coherent, they were in this case largely based on ineligible evidence. Statements of evidence were not cited in the version we evaluated, and some conclusions required us to manually trace the material through searching in the document for keywords.

Discussion

Our evaluation of Elicit demonstrates the platform offers potential benefits but is not yet capable of fully replicating a researcher-led rapid review of evidence. As a tool to expediate scoping of a broad literature, our findings suggest it is capable of doing so. The report produced by Elicit offered an indicative summary of evidence that can serve as a basis on which to refine our focus. Typically, preliminary scoping of a topic does not require strict parameters in what is considered in-scope. The accuracy of eligibility decisions made by Elicit is therefore less important in this scenario.

Our evaluation pointed to a number of risks associated with using Elicit to produce a rapid review of a complex topic. At the outset, the search function lacked the sophistication required to adequately identify relevant material. In this test case, 34 eligible studies were not identified by Elicit's search strategy. Due to the lack of detail and transparency in the search process, it was not possible to determine why these eligible studies were not identified. Screening decisions lacked transparency and consistency, which risks introducing bias and including irrelevant material. An assessment of study quality is not yet available within Elicit, which risks placing too much weight on evidence from poor quality studies.

The Elicit report produced a wide-ranging summary of evidence, but much of the content was superfluous to need and did not answer the review question. This most likely reflects the ineligibility of most included studies. The transparency of the synthesis was poor: claims were not cited with evidence for traceability, and it was unclear how conclusions had been reached. Such transparency is crucial for any review, whether it uses AI or not.¹⁸ The Responsible use of AI in Systematic Evidence Syntheses (RAISE) framework argues that researchers should be held fully accountable for the findings of reviews.¹⁸ To do so without traceable claims is challenging. A recent update to Elicit now includes citations within the report text, which should overcome issues of traceability and transparency.

All of these limitations undermine the core principle of robust evidence syntheses (whether full or rapid): minimising bias at each step to produce an impartial survey of evidence. Despite these shortcomings, there may still be a role for Elicit to support individual components of a rapid evidence synthesis.

For less complex review questions with fewer criteria, Elicit may have the potential to support full text screening, with the Pro version allowing users to upload up to 500 records. Our experience suggests that researchers should be involved to check screening decisions, especially where the review reflects a complex topic like the one used here. For rapid reviews, Cochrane guidance suggests double screening a proportion of records to test agreement before proceeding with single screening.² The same principle may apply here, with a human checking a proportion of Elicit's decisions for accuracy before deciding whether more (or all) should be verified.

Elicit produced detailed data extraction of study characteristics (such as country and study population) in an Excel file in a short timeframe. This information had a good level of detail and could be easily exported. Using Elicit to extract descriptive data may be valuable for researchers undertaking rapid reviews, although second checking a proportion may be required to assess accuracy as per rapid review guidelines. The time saved by using Elicit in this way is uncertain and requires further evaluation. Extracted outcome data was less

helpful, with effect estimates reported for some but not all studies, and some data summaries lacking clarity. We did not check data extraction accuracy, but recent tests by others suggest Elicit performs well in this respect.^{8,9} The inconsistent outcome data we observed here may reflect the complexity of the topic. Greater researcher input may be required depending on the complexity of the data.

Using Elicit for complex review topics: considerations for researchers

In this test case, the complexity of the question may have challenged the AI capabilities of Elicit. To support Elicit, we ensured our criteria were specific, giving clear details across all of the PICOS components (Supplementary materials). Despite this, 36 studies in the Elicit report that did not meet the review eligibility criteria. Reasons for ineligibility mostly reflected the nature of the intervention and outcome and whether the evaluation was conducted with consideration of effectiveness for disadvantaged groups. In contrast, screening decisions relating to population age appeared unproblematic.

Identifying eligible studies may have been particularly challenging for Elicit when processing documentary text about the intervention. This is likely because eligible and ineligible interventions can appear similar on terminology. For example, an intervention comprising physical activity (potentially described as a 'physical activity intervention') would not be eligible if it was designed to achieve (and thus evaluated) non-physical activity outcomes, such as reducing frailty. Similarly, an intervention with no physical activity component (for example, a buddying intervention) would be eligible if physical activity promotion was the goal and such outcomes were reported. The nuance of the text in these examples highlights the added value of a human to determine eligibility beyond keywords and phrases alone.

Whilst we considered this a complex review topic, such complexity is not unusual in the field of health and care policy research. Interventions and models of care are often considered complex due to the range of components, aims, outcomes, staffing and settings involved.¹⁹ Furthermore, our focus on equity, by considering effectiveness for disadvantaged groups, will become the norm rather than the exception in future evidence syntheses of what works. This is because funders are increasingly interested in interventions and models of care that have a low risk of widening health inequalities.²⁰ In the future it is therefore critical that AI platforms supporting reviews are able to robustly respond to such complex requests.

An important consideration here is what researchers can do to enhance the performance of Elicit. AI platforms work from the instructions they receive. A crucial benefit of the Elicit software is that it allows researchers to override its functions. For complex review topics, researchers may need to invest time testing and developing the prompts to optimise Elicit's functions.

Strengths and limitations of this work

Our approach used a detailed, a-priori evaluative framework covering the methodological components of traditional rapid reviews. This has generated a comprehensive assessment of the capabilities of Elicit. Our use of a complex review topic as a test-case adds to existing knowledge by exploring how Elicit responds to questions representative of current health and care policy research. A limitation of our study design was that it was not possible to blind evaluators. That is, some of those undertaking the phase 1 and 2 evaluations were also involved in delivering the test case scoping and rapid review. To balance this, researchers who were not involved in the original work also undertook the evaluation assessments. Further, the researchers responsible for engaging with Elicit and generating the reports for each phase 1 (PK) and 2 (KG) were not involved in the delivery of the original work. This approach offers some degree of independence in our assessments in the absence of a fully blinded evaluation. The researcher-led reviews were used as a benchmark to compare the performance of Elicit. However, we acknowledge that researcher-led reviews are not perfect and by no means an exhaustive account of all evidence.

As detailed in the findings section, Elicit classed a total of 243 studies as eligible after screening, but only 40 of these studies were included in the generated report. We manually checked the eligibility of these 40 studies against the review criteria, but not the eligibility of the remaining 203 studies. This was beyond the scope of our work and means we cannot make a clear judgement on whether these studies would have made a difference to the overall report. However, the screening scores of the 243 studies suggests that most did not meet all of the inclusion criteria according to Elicit's classification (see *Findings*). It is therefore unlikely these studies would have enhanced the accuracy and relevance of the generated report if they were included and synthesised.

Finally, we intended to check the accuracy of claims made in the Elicit report against the studies from which such claims are made. However, we did not complete this task because much of the report content was irrelevant to the review question and there were no clear descriptions of eligible evidence. Furthermore, without citations of studies in the generated report, it was also challenging to trace claims to the associated publications. We checked the extracted data forms for the type of data extracted and whether such information mirrored that in a traditional review. However, we did not set out to check the accuracy of the extracted data as this would have been beyond the scope of this rapid evaluation. Therefore, we cannot comment on the accuracy of the data extracted by Elicit.

Implications of future versions of Elicit

At the time of completing this evaluation (December 2025), a new version of Elicit was released. The updated software included two changes to functionality that have implications for this evaluation. First, researchers can enable 'strict criteria': this means ensuring that only studies that meet all criteria are included, mirroring the approach of a traditional systematic review. The option to apply strict criteria controls, if applied to our evaluation here, would likely have removed all of the irrelevant material that was included in the report. This function would unlikely impact the omission of eligible material demonstrated in our evaluation, as this appeared to be a result of the searches not identifying 34 out of 44 eligible studies.

Second, Elicit has now increased the maximum number of studies included in the report (from 40 to 80 studies). Our evaluation showed that Elicit identified 10 eligible studies but did not include these in the report (see *Findings*). All but one of these studies were deemed ineligible by Elicit. Even so, the change to the number of studies permitted in the report may have resulted in the inclusion of the one eligible study. It was not possible for us to redo our evaluation with these software updates. However, we reran the review in January 2026 using the same criteria (with no changes to the strictness function) to assess whether an 80-study report limit resulted in a change in the inclusion of eligible studies.

In the updated version of the report, 80 studies were included, but only 8% (n=6) were eligible (Supplementary Materials Table 4). Compared to the previous version of the Elicit report, an additional three eligible studies were included. All three studies were identified and included in the researcher-led review. These three studies were identified by the previous version of Elicit, but two were excluded, and one was deemed eligible but not included in the report. As per the previous version of Elicit, 34 eligible studies from the researcher-led review were not identified by the Elicit search. Overall, the updated version of Elicit does not appear to improve the accuracy and relevance of the report in terms of eligible evidence.

Implications for policy research

Elicit is a valuable tool to expediate preliminary scoping of a topic and could reasonably undertake some components of a rapid review alongside a researcher. The time saving element of Elicit for aiding components of rapid reviews remains uncertain given the need for second checking by a researcher. This is important in the context of rapid reviews, where time-saving, and curtailed use of second researchers for checking, underpin their use and delivery.² Further scrutiny of whether the monthly cost of Elicit for these purposes represents value for money is needed.

Conclusion

For rapid reviews of complex topics, Elicit does not yet appear capable of replacing researcher-led reviews to achieve the same, robust outputs. Many of the limitations identified undermine the fundamental goal of systematic and rapid reviews to minimise bias. Elicit could offer valuable time savings when used as a pre-review scoping tool. The software could also support some individual components of reviews jointly with a researcher. Ongoing evaluation is needed to understand the extent to which this offers savings to time and costs.

References

1. Petticrew M, Roberts H. *Systematic Reviews in the Social Sciences: A Practical Guide*; 2006.
2. Garritty C, Hamel C, Trivella M, et al. Updated recommendations for the Cochrane rapid review methods guidance for rapid reviews of effectiveness. *BMJ* 2024; **384**: e076335.
3. Tricco AC, Antony J, Zarin W, et al. A scoping review of rapid review methods. *BMC Medicine* 2015; **13**(1): 224.
4. Elicit. Elicit: the AI Research Assistant. www.elicit.com, 2025.
5. Bernard N, Sagawa Jr Y, Bier N, Lihoreau T, Pazart L, Tannou T. Using artificial intelligence for systematic review: the example of elicit. *BMC Medical Research Methodology* 2025; **25**(1): 75.
6. Lau O, Golder S. Comparison of Elicit AI and Traditional Literature Searching in Systematic Reviews using Four Case Studies. *medRxiv* 2025: 2025.06.17.25329772.
7. Whitfield S, Hofmann MA. Elicit: AI literature review research assistant. *Public Services Quarterly* 2023; **19**(3): 201-7.
8. Helms Andersen T, Marcussen TM, Termannsen AD, Lawaetz TWH, Nørgaard O. Using Artificial Intelligence Tools as Second Reviewers for Data Extraction in Systematic Reviews: A Performance Comparison of Two AI Tools Against Human Reviewers. *Cochrane Evid Synth Methods* 2025; **3**(4): e70036.
9. Hilkenmeier F, Pelzer M, Stierle C, Fink-Lamotte J. Evaluating the AI Tool “Elicit” as a Semi-Automated Second Reviewer for Data Extraction in Systematic Reviews: A Proof-of-Concept. *Social Science Computer Review* 2025: 08944393251404052.
10. Scotti KL, Young S, Gainey MA, Lan H. Artificial Intelligence and Automation in Evidence Synthesis: An Investigation of Methods Employed in Cochrane, Campbell Collaboration, and Environmental Evidence Reviews. *Cochrane Evidence Synthesis and Methods* 2025; **3**(5): e70046.
11. Siemens W, von Elm E, Binder H, et al. Opportunities, challenges and risks of using artificial intelligence for evidence synthesis. *BMJ Evidence-Based Medicine* 2025; **30**(6): 381.
12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 2005; **8**(1): 19-32.
13. National Institute for Health and Care Research (NIHR) Policy Research Unit in Healthy Ageing. Movement for All: Closing Physical Activity Gaps for Marginalised and Disadvantaged Adults in England. *National Institute for Health and Care Research (NIHR) Policy Research Unit in Healthy Ageing* 2025.
14. Cochrane Methods Equity Group. PROGRESS-Plus. 2025. <https://methods.cochrane.org/equity/projects/evidence-equity/progress-plus> (accessed 15.01.2026).
15. O'Neill J, Tabish H, Welch V, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. *Journal of Clinical Epidemiology* 2014; **67**(1): 56-64.
16. Haupka N, Culbert JH, Schneidermann A, Jahn N, Mayr P. Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar. *arXiv* 2025.
17. Otio. Semantic Scholar Review: Features & Benefits. <https://otio.ai/blog/semantic-scholar>, 2025.
18. Thomas J, Flemmyng E, Noel-Storr A. Responsible use of AI in evidence Synthesis (RAISE): recommendations and guidance. 2025.
19. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ* 2021; **374**: n2061.

20. National Institute of Health & Care Research. Inclusive research design to become an NIHR condition of funding. <https://www.nihr.ac.uk/inclusive-research-design-become-nihr-condition-funding>, 2025.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterials.docx](#)