SHORT COMMUNICATION

# The Use of Artificial Intelligence in Dermatology Systematic Reviews: A Comparative Analysis of Elicit Against Human Reviewers

Jui VYAS[1], Jeffrey R. JOHNS[2], Emily FORREST[3], Mari Ann HILLIAR[4], Sam SALEK[5] and Andrew Y. FINLAY[2]
*[1]Centre for Medical Education, School of Medicine, Cardiff University, Cardiff, CF14 4XN, United Kingdom, [2]Division of Infection and Immunity, School of Medicine, Cardiff University, Cardiff, CF14 4XN, United Kingdom, [3]School of Medicine, Cardiff University, Cardiff, CF14 4XN, United Kingdom, [4]Health Library, University Library Service, Cardiff University, Cardiff, CF14 4YU, United Kingdom, and [5]School of Health, Medicine and Life Sciences, University of Hertfordshire, Hatfield, AL10 9AB, United Kingdom. Email: vyasjj@cardiff.ac.uk*

Submitted Nov 21, 2025. Accepted after revision Jan 22, 2026
Published Mar 23, 2026. DOI: 10.2340/actadv.v106.adv-2025-0213 Acta Derm Venereol 2026; 106: adv-2025-0213.

Systematic reviews (SRs) are crucial to support evidence-based medicine but are highly labour-intensive to conduct given the need to review all available literature, extract relevant data, analyse results and form conclusions (1, 2). Elicit, a journal article discovery tool powered by artificial intelligence (AI), is designed specifically to streamline the process of literature review, evidence synthesis and data extraction from scholarly sources; this could potentially streamline initial screening and data extraction in SRs. Elicit allows the use of "natural language questions" instead of keywords to quickly find and analyse articles (3). Our aim was to compare the efficiency and accuracy of Elicit against a published manual search and extraction SR (4) of the Dermatology Life Quality Index (DLQI) used as the primary outcome in clinical trials. Two aspects were examined: article searching and data extraction.

## METHODS

In our previously published SR (4), searches of 7 online databases for randomized controlled trial (RCT) articles using DLQI found 1,375 articles that were then manually reviewed to determine if DLQI was a primary outcome. An Elicit "concept search" was performed with the text "Using the dermatology life quality index DLQI in people with a variety of skin conditions as the outcome measure or endpoint in RCTs using placebo or control groups", and articles were assessed as to whether they met the initial primary outcome definition. A second Elicit search was performed using keywords "DLQI", "dermatology life quality index" and "randomised controlled trials" to search for papers relating to the DLQI and RCTs. The results were then compared with the original bibliographic database search.

To assess data extraction accuracy, all 24 PDF files from the original SR (4) manuscripts were uploaded to Elicit. Default Elicit output columns (participant count, methodology, outcome measured) were used, and separate customized columns were created, e.g. "topical", "systemic" and "randomised". Answers were recorded and compared with the original search. Where answers given by Elicit were incorrect or differed from the original search, the justification given by Elicit and further evidence supported by the paper were explored by clicking on the text answers. This was repeated for each column (see Table SI). For JADAD risk of bias scoring (5, 6), Elicit was unable to determine components directly; thus, results from answers given by Elicit concerning randomization, blinding and fate (of participants) were assessed, and the JADAD score was calculated manually. More details are given in Appendix SI.

## RESULTS

### *Traditional database vs Elicit searches*

The Elicit "concept search" produced only 59 results. This included, unexpectedly, three duplicates that were in the Semantic Scholar database (7). Only one of these 59 matched the 24 publications included in our original SR, perhaps because the abstract stated "Primary endpoint was the mean change in Dermatology Life Quality Index (DLQI) at 6 months". The second Elicit search using keywords produced 27 unique publications, but none matched the 24 publications included in our original SR. For comparison, a direct search for "dermatology life quality index" in the Semantic Scholar database (not using Elicit) found 4,320 articles (since 2006, the earliest date for Semantic Scholar). The same search in PubMed returned 3,357 articles.

### *Human vs Elicit data extraction*

Table SI shows a comparison of all data extracted by Elicit from PDFs of the 24 articles included in the original SR. Generally, agreement was good, but 13 results (across all 24 studies) given by Elicit were considered incorrect by our team. In 5/24 studies (20.8%), Elicit failed to correctly detect locations where studies were performed, defaulting instead to the country of the author's institution. Elicit gave incorrect information for randomization, fate (of participants) and blinding, resulting in incorrect JADAD scores in 5/24 (20.8%) cases. For the study disease, 100% correct information was found.

## DISCUSSION

*Context specificity*

There are fundamental challenges in using an AI tool, such as Elicit, for SRs. Unlike traditional bibliographic database searches using scripts, Elicit searches are not reproducible over time, as it is not currently possible to limit the searches by date, because the source material on which the search is performed is continually updated, and AI search strategies are constantly evolving. The inability of the search to be consistently reproducible breaks a foundation of research – to be reproducible (8).

In our study, Elicit, using Large Language Models, produced 59 results, whereas keyword searching produced 27 results. For comparison, a direct search for "dermatology life quality index" in the Semantic Scholar database (not using Elicit) found 4,320 articles (since 2006, the earliest date for Semantic Scholar). The same search in PubMed returned 3,357 articles, indicating that more complex natural language searching in Elicit effectively filters the number of articles returned, supporting the utility of these algorithms for application in SRs. Although Elicit produced fewer responses to review, in our study, only one of the 59 publications and none of the 27 fulfilled our inclusion criteria.

An example of a challenging keyword search in this study was "primary outcome", which in clinical trials has a very specific medical definition (9). AI tends to equate this with "most important", "main" or "first", which misses the term specificity in its technical context, i.e., aligning directly with the primary study aim or used to determine the sample size and power of the study. In our study, Elicit searches included many articles that were not RCTs, were non-peer reviewed, or were study protocols, SRs or meta-analyses.

*Content accuracy*

Elicit did not always give accurate answers due to limited "understanding" of the clinical context – it is difficult for Elicit to understand the meaning of words such as "participant count" from context alone, resulting in incorrect answers. The need for human validation prolonged the extraction time, leading researchers to spend more time extracting correct information and cross-validating. Furthermore, the extraction to custom columns for responses that are not within the default list is still cumbersome and slow, although presumably this will be improved in the future. Extracting information for JADAD scoring (5, 6) was possible, but cumbersome, and not always accurate.

Elicit gave incorrect answers for JADAD scoring in 5/24 publications. This may result from inherent limitations of Elicit with keyword searching as outlined above. Overall, Elicit was faster at extracting information in comparison to our manual reading of individual articles. The extracted data closely matched the source, but sometimes the results were incorrect. Elicit enables users to explore the answers it gives by highlighting evidence in the source and rationalizing its answers, allowing human discretion in determining the accuracy of the AI decision. This feature was frequently used in this study.

Elicit users should read the explanations given during data extraction and compare them to the original articles to confirm extraction accuracy.

A study strength is that we performed a comparison with an already published SR. However, due to the challenging definition of "primary outcome" (4, 9), manual selection was required to identify studies using DLQI as the primary outcome. This study did not compare any other AI tools, as only Elicit was designed specifically to help researchers, academics and professionals streamline the process of literature review, evidence synthesis and data extraction from scholarly sources. AI tools are changing very rapidly, and their accuracy and speed may improve between submission and publication of this manuscript. However, this study provides insight into the trajectory of change in the development of AI tools for the execution of SRs.

## ACKNOWLEDGEMENTS

*Funding sources:* Funding was provided by the Division of Infection and Immunity, School of Medicine, Cardiff University, Cardiff, UK

*IRB approval status:* No IRB approval was required for this study.

*Conflict of interest:* Andrew Y Finlay is joint copyright owner of the DLQI. Cardiff University receives royalties from some use of the DLQI: AYF receives a proportion of these under standard university policy. Sam Salek has received an unrestricted educational grant from GSK, is a consultant for Novo Nordisk and produces educational materials for Abbvie. Jui Vyas participated in an Advisory Board for Amgen, has received payment or honoraria from L'Oreal and support from UCB pharma for attending meetings. Faraz Ali has received honorariums from Abbvie, Janssen, LEO pharmaceuticals, Lilly pharmaceuticals, L'Oreal, Novartis and UCB. Faraz Ali and Jeffrey Johns' department receives income from royalties from the Dermatology Life Quality Index (DLQI) and related instruments. Emily Forrest and Mari Ann Hilliar have no conflicts of interest to declare.

## REFERENCES

1. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017; 7: e012545. https://doi.org/10.1136/bmjopen-2016-012545
2. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022; 144: 22–42. https://doi.org/10.1016/j.jclinepi.2021.12.005

3. Bernard N, Sagawa Y Jr, Bier N, Lihoreau T, Pazart L, Tannou T. Using artificial intelligence for systematic review: the example of elicit. BMC Med Res Methodol 2025; 25: 75. https://doi.org/10.1186/s12874-025-02528-y
4. Johns JR, Vyas J, Ali FM, Ingram JR, Salek S, Finlay AY. The Dermatology Life Quality Index as the primary outcome in randomized clinical trials: a systematic review. Br J Dermatol 2024; 191: 497–507. https://doi.org/10.1093/bjd/ljae228
5. Stephen HH, Douglas MJ. Appendix: Jadad scale for reporting randomized controlled trials. Evidence-based Obstetric Anesthesia: Blackwell Publishing Ltd; 2005.
6. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996; 17: 1–12. https://doi.org/10.1016/0197-2456(95)00134-4
7. Gusenbauer M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics 2019; 118: 177–214. https://doi.org/10.1007/s11192-018-2958-5
8. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005; 2: e124. https://doi.org/10.1371/journal.pmed.0020124
9. Vetter TR, Mascha EJ. Defining the primary outcomes and justifying secondary outcomes of a study: usually, the fewer, the better. Anesth Analg 2017; 125: 678–681. https://doi.org/10.1213/ANE.0000000000002224