Supplementary material has been published as submitted. It has not been copyedited, or typeset by Acta Dermato-Venereologica

APPENDIX S1
RATIONALE FOR AI USE IN SYSTEMATIC REVIEWS

AI tools are rapidly developing and many are providing support for performing systematic reviews (SR). A scoping review of 47 publications identified the reliability and validity of available tools to support automation of SRs (1). However, the accuracy between AI vs manual extraction remains a barrier for widespread acceptance for AI to be used as a primary source. Affengruber et al. (2) reported on tools for title and abstract screening from 103 published studies to investigate available methods to improve the efficiency of SRs, however, few studies evaluated methods or tools supporting other SR production tasks, such as protocol preparation, full-text retrieval and certainty of evidence assessment, particularly in real-world workflows. Recently, Clark et al. (2025)(3) performed a SR of generative artificial intelligence (GenAI) use in evidence synthesis. They concluded "current evidence does not support GenAI use in evidence synthesis without human involvement or oversight". However, like our present study, they found that "for most tasks other than searching, GenAI may have a role in assisting humans with evidence synthesis."

Many traditional bibliographic databases, however, have been built around human based keyword categorization and indexing methods. Medline, for example, uses Medical Subject Headings (MeSH) controlled vocabulary indexing, which involves human expertise. This results in reliable, reproducible searches based on meaningful "technical" subjects. Most of these traditional databases are now also using AI to perform or aid these tasks. In contrast, keywords in large language model (LLM) AI searches are merely "tokens" used to generate text or "results" by probabilistic methods (statistical regularities in language, relationships between words, syntax, and concepts, and long-range dependencies e.g. subject–verb agreement discourse, using maximum likelihood estimation and minimizing cross-entropy loss) predicting the next token based on patterns learned from data, making it appear focused and contextual.

Thus, although LLM AI's use high-dimensional functional approximation (4) (learning complex functions with many inputs that maps them to outputs, when both the inputs and the function itself live in very high-dimensional spaces), learned representations (4) (embeddings, dense numerical vectors that encode the meaning, relationships, and context of things) and "attention" mechanisms (5) (dynamically focusing on the most relevant parts of the input when producing an output as opposed to classic pattern matching) that model context and relationships and learnt patterns, they don't "know" facts the way humans do, do not retrieve exact memorized answers (except in rare cases) and can only generalize patterns to new inputs. Thus they appear to be able to answer novel questions, produce answers that they have never seen and make mistakes confidently ("hallucinate").

Bibliographic databases often use "root and stemming" (6) based on natural language processing (NLP), reducing words to their root form or stem. AI's ability to accurately reproduce the "root and stemming" approach is not well demonstrated. "Lemmatization" in AI (a more advanced version of stemming, taking into account the word's context and grammar) shows promise in providing only valid words with high accuracy but has high computational burden. Automatic Term Mapping (ATM) in PubMed includes stemming and variant recognition, processes that are algorithmic and handled automatically by PubMed's search engine and still appear to be superior to those in other AI tools.

An important concept relating to stemming is conflation (7) (reducing different word forms to a common root, or stem, so they are treated as the same)(8). It involves treating different words or phrases as semantic matches because they refer to the same central idea. PubMed and Medline use conflations primarily through ATM, synonyms, and MeSH mapping (not crude stemming). Elicit does not appear to use stemming-based conflation in the way traditional search engines like PubMed use it and instead Elicit uses semantic similarity to capture word variants and related concepts. This may be less powerful and result in lower accuracy results.

A recent study (9) compared results from an umbrella review conducted independently of AI with results of Elicit (released March 2025) AI-based searching using the three criteria, repeatability, reliability and

accuracy. It was found that Elicit could serve as a valuable complementary tool for researchers when designing or writing SRs, however, it should be used with caution, and certain principles needed to be followed to maintain methodological rigour and integrity (2). Elicit uses the Semantic Scholar database (10) as its primary source thus excluding many publishers, although other sources are being added. This database is therefore only a small subset of the combination of MEDLINE, The Cochrane Library, Embase, Web of Science, Scopus, CINAHL (EBSCO) and PsycINFO databases used in our SR. Elicit searches are therefore currently not comprehensive.

Use of AI tools as second reviewers for data extraction in SRs was explored by comparing the performance of Elicit and ChatGPT against human reviewers **(11)**. Precision, recall and F1-score (harmonic mean of precision and recall) used to assess overall extraction quality were 92%, 92% and 92% for Elicit and 91%, 89% and 90% for ChatGPT. Although Elicit and ChatGPT "demonstrated high and similar performance in data extraction compared to human reviewers, particularly for standardized variables", error analysis "revealed confabulations" (incorrect data) in 4% of data points. They proposed using AI-assisted extraction to replace the second human extractors, "with the second human instead focusing on reconciling discrepancies between AI and the primary human extractor".
AI tools such as RoboReviewer (12) or Dimensions (13) produce more "template" like outputs based on keyword searches, rather than attempting to summarise queries. They appear to consistently extract data from specific locations within documents (patents, clinical trials or academic publications) and specific contexts i.e. other words in the sentence construct that indicate what these data refer to. The resulting outputs are thus much more focused, accurate and homogenous, meeting a specific need for this sort of data.

An in-depth study evaluated supervised (semi-automated) machine learning methods for predicting articles from title/abstract screening relevant for full-text review in a SR (14). Combining the predictions of the 10 best-performing algorithms improved the performance to 95% sensitivity and 64% specificity in a validation dataset with relatively short computation time,

Risk of Bias analysis is an important feature of SRs. Mahuli et al. (15) explored using ChatGPT for conducting Risk of Bias analysis and data extraction from a RCT. They suggested that AI can identify items "missed by the human eye" and "shows promise in reducing workload and time, but careful implementation and validation are necessary".

FURTHER DETAILS OF METHODS

Article searching with manual publication search: In our previously published SR (16), MEDLINE (Ovid), The Cochrane Library, Embase, Web of Science, Scopus, CINAHL (EBSCO) and PsycINFO online databases from 1 January 1994 (year of DLQI publication)(17) to 16 November 2021 were searched using bespoke scripts, independently by two authors, and results merged. Search terms included 'DLQI' and 'dermatology life quality index'. Database-specific 'article type/study type' keywords, language keywords (English) and age-selection keywords were also used to search the required types of study to be included, e.g. Medical Subject Headings terms for randomised controlled trials (RCTs). Articles matching the search criteria were exported to an EndNote20® (Clarivate, Philadelphia, USA) database, with duplicates excluded during the import through Endnote filtering rules. The 1375 unique articles found were manually reviewed by the reviewers to determine if they used the DLQI as a primary outcome in the RCT's, using a definition of primary outcome based on accepted medical definitions, literature usage and expert opinions. The published SR gives further details (16).

In the original manual extraction methodology (16), two independent reviewers extracted data from the included publications stored in an EndNote database into parallel REDCap database tables, and an adjudicator resolved any disagreements. For categorising the intervention used in each study, the intervention column of Elicit was customised to specified answers such as "topical" "systemic" or "non-pharmacological"

for which Elicit was able to respond with categorised answers. However, where Elicit was unable to match key terms, including "fate" and "blinding", synonyms or context was given by our team in searches to aid Elicit. For example, "completion" or "double-blind" was used for Elicit to understand the context of responses needed. If Elicit was still unable to read terms after two tries, the answers and reasoning given were recorded and discussed with the rest of the team.

COMMENTS

Elicit summarises what it has "read", however, these summaries may be incorrect, necessitating rechecking the original text. Thus, there is little time saving between AI finding and summarising appropriate text requested based on query terms (natural language or otherwise) and humans searching and reading the original text; we can read the original and extract what we need based on objective criteria. Humans understand where key data occurs within an article, whereas Elicit often produces erroneous results, as in our study, by using inappropriate parts of a manuscript, e.g. using author list to seek study location. In a traditional SR, everything extracted into data tables has been read, analysed and content extracted independently by human reviewers (usually by two or more), then compared (harmonised), and independently adjudicated. In this study it was difficult to identify treatment methodology as topical, systemic or non-pharmacological as Elicit often detected these words in the introduction or discussion, having no relevance to study methodology. Elicit may over-generalise; it produces generalised statements that may in fact only be true within a specific context.

While the process of data extraction by humans may include errors, the use of multiple reviewers, with agreement between them concerning data extracted is now the benchmark in systematic reviewing. Cochrane Collaboration (18,19) and Joanna Briggs Institute (20) processes may be lengthy and expensive requiring high levels of expertise/training. Human errors can be reduced by quality control methods such as engagement of a second reviewer. Corrections cannot be made in the same way using only AI, which lacks judgement and experience and can be influenced both internally and externally, leading to inaccuracies. There are not enough studies that validate each AI tool for either searching or extracting data for SRs. Such validation studies quickly become obsolete as AI systems are rapidly evolving. However, Elicit seems useful to find a theoretical framework or topic for research and very useful for finding references to support a research result, discussion or conclusion. It may also prove to be useful as a second reviewer for extractions, with sufficient human oversight.

REFERENCES FOR APPENDIX S1

1. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022; 144: 22-42.
2. Affengruber L, van der Maten MM, Spiero I, Nussbaumer-Streit B, Mahmic-Kaknjo M, Ellen ME, et al. An exploration of available methods and tools to improve the efficiency of systematic review production: a scoping review. BMC Med Res Methodol 2024; 24: 210.
3. Clark J, Barton B, Albarqouni L, Byambasuren O, Jowsey T, Keogh J, et al. Generative artificial intelligence use in evidence synthesis: A systematic review. Research Synthesis Methods 2025; 16: 601-619.
4. Y. Bengio, A. Courville, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 2013; 35: 1798-1828.
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. Advances in Neural Information Processing Systems 2017; 30: 5998-6008.
6. Thakare AD, Laddha S, Pawar A. Hybrid Intelligent Systems for Information Retrieval. 1st ed. New York: Chapman and Hall/CRC; 2022.
7. Jennifer P, Muthukumaravel A. Conflation Methods in Stemming Algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE) 2019; 8: 1176-1180.

8. Göksel G, Arslan A, Dinçer BT. A selective approach to stemming for minimizing the risk of failure in information retrieval systems.
9. 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.
10. Gusenbauer M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics 2019; 118: 177-214.
11. 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 Evidence Synthesis and Methods 2025; 3: e70036.
12. Marshall IJ, Kuiper J, Banner E, Wallace BC. Automating Biomedical Evidence Synthesis: RobotReviewer. Proc Conf Assoc Comput Linguist Meet 2017; 2017: 7-12.
13. Singh VK, Singh P, Karmakar M, Leta J, Mayr P. The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. Scientometrics 2021; 126: 5113-5142.
14. Kebede MM, Le Cornet C, Fortner RT. In-depth evaluation of machine learning methods for semi-automating article screening in a systematic review of mechanistic literature. Research Synthesis Methods 2023; 14: 156-172.
15. Mahuli SA, Rai A, Mahuli AV, Kumar A. Application ChatGPT in conducting systematic reviews and meta-analyses. Br Dent J 2023; 235: 90-92.
16. 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.
17. Finlay AY, Khan GK. Dermatology Life Quality Index (DLQI)--a simple practical measure for routine clinical use. Clin Exp Dermatol 1994; 19: 210-216.
18. Cochrane. Cochrane Handbook for Systematic Reviews of Interventions (current version) https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current. London, UK: Cochrane, 2025.
19. Cochrane. Methodological Expectations of Cochrane Intervention Reviews (MECIR) Manual https://www.cochrane.org/authors/handbooks-and-manuals/mecir-manual. London: Cochrane, 2025.
20. Joanna Briggs Institute. Critical Appraisal Tools https://jbi.global/critical-appraisal-tools. Joanna Briggs Institute, 2025.