Automation of systematic reviews of biomedical literature: a systematic review of studies indexed in PubMed

Barbara Tóth
Óbuda University  https://orcid.org/0000-0003-1827-6702

László Berek
Óbuda University  https://orcid.org/0000-0002-4126-1528

László Gulácsi
Óbuda University  https://orcid.org/0000-0002-9285-8746

Márta Péntek
Óbuda University  https://orcid.org/0000-0001-9636-6012

Zsombor Zrubka (✉ zsombor.zsombor@uni-obuda.hu)
Óbuda University  https://orcid.org/0000-0002-1992-6087

Systematic Review

Keywords: Systematic literature review, evidence synthesis, automation, artificial intelligence, machine learning, natural language processing, text mining

Posted Date: October 10th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3425312/v1

License: ☑️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background
The demand for high quality systematic literature reviews (SLRs) is growing for evidence-based medical decision making. SLRs are costly and require the scarce resource of highly skilled reviewers. Automation technology has been proposed to save workload and expedite the SLR workflow.

Objectives
We aimed to provide a comprehensive overview of SLR automation studies indexed in PubMed, focusing on the applicability of these technologies in real world practice.

Methods
In November 2022, we ran a combined search syntax of four published SLRs on SLR automation. Full-text English peer-reviewed articles were included if they reported Studies on SLR Automation Methods (SSAM), or Automated SLRs (ASLR). Bibliographic analyses and knowledge-discovery studies were excluded. Record screening was performed by single reviewers, the selection of full text papers was performed in duplicate. We summarized the publication details, automated review stages, automation goals, applied tools, data sources, methods, results and Google Scholar citations of SLR automation studies.

Results
From 5321 records screened by title and abstract, we included 123 full text articles, out of which 108 were SSAMs and 15 ASLRs. Automation was applied for search, record screening, full-text selection, data extraction, risk of bias assessment, evidence synthesis, assessment of evidence quality and reporting in 19 (15.4%), 89 (72.4%), 6 (4.9%), 13 (10.6%), 9 (7.3%), 2 (1.6%), 2 (1.6%), and 2 (1.6%) studies, respectively. Multiple SLR stages were automated by 11 (8.9%) studies. The performance of automated record screening varied largely across SLR topics. In published ASLRs we found examples of automated search, record screening, full-text selection and data extraction. In some ASLRs automation complemented fully manual reviews to increase sensitivity rather than to save workload. Reporting of automation details were often incomplete in ASLRs.

Conclusions
Automation techniques are being developed for all SLRs stages, but with limited real-world adoption. Most SLR automation tools target single SLR stages, with modest time savings for the entire SLR process and varying sensitivity and specificity across studies. Therefore, the real-world benefits of SLR automation remain uncertain. Standardizing the terminology, reporting, and metrics of study reports could enhance the adoption of SLR automation techniques in real-world practice.
Introduction

Systematic literature reviews (SLRs) and meta-analyses represent the highest level of evidence in evidence-based medicine, providing essential input to medical decision-making. While the number of published SLRs in PubMed was 80 per day in 2019, this number increased to 135 by 2021. The accelerated development of novel medical technologies such as software and digital devices, virtual reality, and chatbots will push further the demand for high-quality SLRs. Beyond medicine, systematic reviews are often performed in disciplines including engineering or the social sciences.

As the demand for SLRs grows, keeping them up-to-date is becoming increasingly challenging. The preparation of a SLR is a labour-intensive and time-consuming process requiring the scarce resource of highly skilled researchers. The typical lag for primary studies to be included in SLRs is 2.5–6.5 years, delaying the translation of results to medical decision-making. Although the Cochrane Handbook recommends that SLRs are updated biannually, 23% of SLRs can become outdated within 2 years due to the omission of new evidence that could impact their conclusions.

SLR automation using artificial intelligence (AI) has the potential to speed up the review process, reduce the workload of researchers, prevent human errors, and facilitate reproducibility by diminishing the role of human judgement. The feasibility of automation differs by stages of the SLR workflow, with search, record screening, full-text selection, data extraction, risk of bias assessment, evidence synthesis and reporting being the most prominent examples. Automated assessment of evidence quality is also under investigation.

Hence, recent SLR methodological guidelines have addressed the use of automation tools. The Cochrane Handbook acknowledges the use of AI tools when updating SLRs, or using AI as a second reviewer alongside a human reviewer. While the Handbook mentions active learning, it does not recommend its use on its own, and considers data extraction mainly as a manual process, despite citing some examples for automated data extraction. The latest PRISMA guideline also acknowledges the use of AI tools in record screening or priority ranking. It also sets out how to report the use of AI tools in the screening or risk of bias assessment stages of SLR reports, including the training of the tool and the method used to measure its validity. Automated risk of bias assessment is also a promising field for methodological innovation, but results are not yet convincing.

Despite some positive experiences, the uptake of SLR automation tools is still limited. Trust in automated SLRs is based on the availability of high-quality summary studies of their results. Accordingly, several authors have systematically reviewed automation technologies in various stages of the SLR workflow. While aiming for a comprehensive summary, these studies differed in their focus, search strategies and number of included reports. The topics covered text mining for screening, data extraction, any automated SLR stage, or identifying high quality studies. Previous SLRs on SLR automation illustrated the challenge of developing search strategies to identify relevant research articles in the field. The large number of SLRs published on various information retrieval, text mining, and AI applications makes it challenging to identify automated SLRs, due to the large overlap in the terminology of these articles.

Due to the lack of specific search terms for articles on SLR automation, the use of general terms such as “automated SLR” carries the risk of low sensitivity, illustrated by the study of Dinter et al., which, despite
including automation studies in all stages of the SLR workflow and extending the electronic search with a manual snowball technique yielded fewer reports than earlier reviews focusing on a more specific aspect of SLR automation \(^2^8\). On the other hand, the risk of low specificity was demonstrated by the review of Adbelkader et al., which aimed to identify a special, yet clinically relevant subset of review automation use-cases \(^3^0\).

By combining the search strategies of previous reviews, the aim of this study was to provide a comprehensive overview on the scope of SLR automation across various stages of the SLR workflow, as well as the adoption of automation techniques in published SLRs among studies indexed in PubMed. Hence, we included both Studies on SLR Automation Methods (SSAM), and Automated SLRs (ASLRs) (i.e., studies that used automation techniques when answering a primary research question unrelated to SLR automation). We inquired what SLR stages were automated and what were the goals, the applied tools and methods, the data sources, and the key results of SLR automation. We also performed a citation analysis to assess the research impact of SLR automation studies.

**Methods**

We followed the PRISMA reporting guideline when conducting this review \(^2^5\).

**Automated systematic reviews**

To define SLRs, we used the general criteria proposed by Krnic-Martinic et al. \(^3^1\). As such, SLRs feature a well-defined research question, a reproducible search strategy, clear inclusion and exclusion criteria for relevant publications, reproducible selection and screening methods, critical appraisal of the quality or risk of bias for included studies, and reproducible data analysis or synthesis methods \(^3^1\). Throughout the review process, we considered as an SLR automation tool any method that aims to speed up, assist, or replace manual reviewer tasks that require human judgement with an algorithm-based solution, while aiming to yield comparable results achievable by human reviewers. Papers reporting on tools that can potentially assist the SLR workflow but are not developed or applied specifically for this purpose were excluded.

**Inclusion and exclusion criteria**

Using the definitions above, we included full-text English peer-reviewed articles of both SSAMs and ASLR with no limit on publication date.

We excluded bibliographic analyses, or text-based knowledge discovery studies or information retrieval studies from large corpora. These studies employ advanced analytical methods to generate new results, rather than reducing the workload for tasks that humans can achieve. Furthermore, we excluded narrative reviews and non-automated SLRs on SLR automation or SLR automation tools or methods.

**Search strategy**

We focused on published research in the medical field, so we limited our search on PubMed. The search was run on November 12, 2022. We extracted the search strategies of four published SLRs on SLR automation, \(^2^2\)\(^2^8\)\(^–^3^0\) identified during the planning of this review (Appendix Table S1). The four strategies were combined into a single search syntax using the Boolean „OR“ operator. We also run the four searches individually to count
duplicate records. Abdelkader et al. narrowed down their general search strategy by using terms that refer to the quality of the articles. For our search, these terms were removed to achieve higher sensitivity. The search syntax is provided in Appendix Table S2.

**Screening and selection of studies**

Screening of titles and abstracts was completed independently by three single researchers (BT, LB, ZZ) on the combined record set. Uncertain items were discussed. Full-text papers were then evaluated by two independent reviewers against the inclusion and exclusion criteria (BT, ZZ). In case of disagreement or if reviewers were not sure whether an article was suitable for inclusion, they discussed its eligibility, and a joint decision was taken.

**Data extraction**

Two reviewers (BT, ZZ) extracted data from each eligible article using a pre-designed spreadsheet. The second, senior reviewer compared and consolidated the extracted items. These encompassed publication meta-data, including details such as the first author’s name, publication year, article title, and the PubMed ID (PMID) for each article. Additionally, we collected information about the article type, categorizing them as either SSAM or ASLR. Furthermore, we identified the SLR stage where automation was applied, such as search, record screening, full-text selection, data extraction, risk of bias assessment, evidence synthesis, assessment of evidence quality, and reporting. Assigning automation methods to the appropriate SLR stages was challenging due to the diversity of approaches. In Table 1, we provide positive examples illustrating our decisions to categorize automation methods within specific automated SLR stages, as well as negative examples showing instances when a method was excluded or categorized elsewhere among the automated SLR stages. We considered only the laborious execution parts of the SLR workflow, omitting steps of the review planning phase. We also extracted details about the input text used, encompassing the title, abstract, full text, or metadata. In addition, we gathered information about the text representation methods employed, which ranged from basic techniques like bag-of-words or term frequency to more advanced methods such as vector representation and large language models. Moreover, if reported, we recorded the best performing machine learning models or algorithms used for text classification and task learning. We took note of the accessible corpora used for learning or testing, along with their weblinks if provided in the studies. Additionally, we recorded information about off-the-shelf or freeware automation software utilized in the studies, including any available weblinks. We noted if multiple packages were used from a single software environment (i.e., R, Python), without detailing the individual tools. Furthermore, we documented notable methodological details that had potential impact on results, such as experimentation with different feature sets or addressing feature imbalance. Finally, we noted key results related to performance metrics, including recall (sensitivity), precision (positive predictive value), workload-saving, time-saving, or any other significant metrics as reported by the authors.

As a proxy of potential research impact, we added the number of Google Scholar (GS) citations of the included studies, collected on 16th July 2023. Finally, from ASLRs we extracted the research aims, the number of records and included studies, key results, the automated SLR stage, and the applied SLR automation tools and their reported performance.
Table 1

Categorisation of the SLR stages, where automation was applied with positive and negative examples

<table>
<thead>
<tr>
<th>Automated SLR stage</th>
<th>Positive examples (i.e., automation methods categorised within the corresponding SLR stage)</th>
<th>Negative examples (i.e., methods excluded or categorised elsewhere)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search</td>
<td>• improving a search strategy (e.g., identifying relevant keywords to improve sensitivity and specificity) &lt;br&gt; • automating the construction of a search syntax</td>
<td>• converting of a search strategy from one database to another &lt;br&gt; • the setting up of search notifications for a living review &lt;br&gt; • deduplication of records</td>
</tr>
<tr>
<td>Record screening</td>
<td>• deciding on the potential eligibility of an article based on its title, abstract, or keywords (e.g., ranking, classification, matching text against inclusion criteria) &lt;br&gt; • updating the SLR with new potential articles using machine learning on annotated records from previous versions of the SLR</td>
<td>• tools that only support the transparency of the screening process or facilitate the communication between reviewers (e.g., review management software)</td>
</tr>
<tr>
<td>Full-text selection</td>
<td>• using full-text information to decide on the eligibility of an article based on any method</td>
<td>• methods that predict full-text eligibility without using information from the full-text report</td>
</tr>
<tr>
<td>Data extraction</td>
<td>• identifying / extracting data in an eligible article that is relevant to answering the research question (e.g., images, tables, effect sizes)</td>
<td>• extracting certain elements from the abstract/article to facilitate record screening, full-text selection or the assessment of risk-of-bias or reporting quality</td>
</tr>
<tr>
<td>Risk of bias assessment</td>
<td>• assessment of methodological adequacy, the presence / absence of methodological safeguards</td>
<td>• manual completion of risk of bias questionnaires</td>
</tr>
<tr>
<td>Evidence synthesis</td>
<td>• automating the selection and / or application of quantitative evidence synthesis methods (e.g., meta-analysis)</td>
<td>• standard statistical procedures, that replace manual computations &lt;br&gt; • novel descriptive or explorative or knowledge discovery methods (e.g., mapping, topic modelling, networking, visualization, trend analysis) &lt;br&gt; • bibliographic analyses</td>
</tr>
<tr>
<td>Assessment of evidence quality</td>
<td>• automating the assessment about the certainty / confidence in the evidence supporting the findings</td>
<td>• risk of bias or reporting quality assessments</td>
</tr>
<tr>
<td>Reporting</td>
<td>• e.g., textual summarization of results</td>
<td>• graphical representation of novel evidence synthesis methods (e.g., topic models, networks) &lt;br&gt; • graphs or tables from traditional evidence summaries (e.g., meta-analysis, funnel plot)</td>
</tr>
</tbody>
</table>

Data synthesis
We analysed data via descriptive methods. We summarized the number of eligible papers on automation methods and automated systematic reviews by publication year, and by the SLR stage descriptively, and provided a qualitative summary of reported time savings by each automated SLR stage. We also tabulated and qualitatively summarized the key characteristics of ASLRs.

Results

Results of the literature search

The four search strategies yielded 5484 hits, with only 163 duplicate records (3.0%), suggesting minimal overlap between previous SLRs on SLR automation. The combined search yielded 5321 results, out of which 411 potential eligible records were sent to full text screening. A further 288 articles were excluded during full text screening with reasons (Appendix Table S3). Finally, 123 articles were included (Fig. 2). We found 15 ASLR studies (12.2%), and 108 papers reporting SSAMs (87.8%). The extracted data from all included studies are summarized in Appendix Table S4.

Characteristics of the included studies

Date of publication

The first included paper was published in 2006. It investigated whether automation could reduce the SLR workload. The study suggested that 20% - 50% time could be saved with a 95% recall level during abstract screening by using a bag of words model and a voting perceptron machine learning classifier. Since 2014, the number of studies increased rapidly with 56.1% (69/123) of included papers published from 2019 onwards. We found automation examples for all stages of the SLR workflow (Fig. 3).

Search

Nineteen included papers (15.4%) aimed to automate or improve database searches. The first included paper from 2011 applied text-mining to construct a search syntax for PubMed, using the Apache Lucene platform. Eleven papers used a plethora of text-mining tools to aid search syntax building, such as Anne O'Tate, AntConc, Apache Lucene, BiblioShiny, Carrot2, CitNetExplorer, EndNote, Keyword-Analyzer, Leximancer, Lingo3G, Lingo4G, MeSH on Demand, MetaMap, Microsoft Academic, PubReMiner, Systematic Review Accelerator, TerMine, Text Analyzer, Tm for R, VOSviewer, Voyant, Yale MeSH Analyzer as well as in-house solutions. Two papers introduced curated article collections, such as Cochrane CENTRAL, and the Realtime Data Synthesis and Analysis (REDASA) COVID-19 dataset, which were assembled using various automation techniques. Other tools included an automated extension of PubMed searches to the ClinicalTrials.gov database, a Boolean query refiner, a support vector machine (SVM) classifier as alternative to PubMed search filters for review updating, a strategy using the Patient, Intervention, Comparator, and Outcome framework (PICO) terms in the title field only, an automated full-text retrieval and targeted search replacing database screening, and a Microsoft Excel-based convenience tool to build Boolean queries.

Record screening
The most popular SLR automation approach was record screening based on titles and abstracts (N = 89, 72.4%). Within this approach, automated classification (N = 32/89, 36.0%) was the most frequently reported strategy. In automatic classification, a subset of manually screened records are used to train a machine learning classifier, which proposes records that should undergo manual full-text selection. The second most prevalent strategy was active learning (N = 24/89, 27.0%) \cite{41, 49, 58, 64, 69, 72, 74, 77, 80, 84, 85, 90, 94, 105, 107, 110, 116, 118, 120, 123, 131, 134, 145}. In active learning, a small seed group of relevant records are used for initial training. Records are manually screened by the order of relevance predicted by the model. Using the results, the model is periodically re-trained until finding relevant records becomes unlikely. In the third most used strategy, review updates, all included papers and excluded records of a published review are used for training, and the aim is to predict the inclusion of a record from new search results in the updated review (N = 12/89, 13.5%) \cite{34, 46, 81, 83, 89, 93, 112, 124, 125, 129, 152, 153}. The priority ranking strategy (N = 10/89, 11.2%) \cite{33, 36, 38, 65, 66, 75, 91, 132, 134, 139} was used least often. This strategy predicts the priority of records after single training round. By screening relevant records early, subsequent phases of the SLR can advance faster. Other studies applied a combination of strategies \cite{53, 98}, used alternative methods such as filtering \cite{18} or similarity of Medline elements \cite{59}, reported the automation software without detailing the strategy \cite{87, 114, 146}, used convenience tools to speed up screening \cite{100, 119}, or omitted record screening and applied topic modelling directly to full-text selection \cite{113}.

SVM was by far the most prevalent machine learning method, usually used in ensemble models (N = 24/89, 27.0%) \cite{33, 35–38, 41, 46, 49, 53, 57, 58, 60, 64, 69, 74, 78, 80, 84, 92, 96, 105, 106, 126, 129, 145}, followed by naive Bayes (N = 7, 7.9%) \cite{40, 42, 47, 48, 129, 130, 146}, and logistic regression (N = 7, 7.9%) \cite{52, 62, 85, 93, 94, 106, 124}. More recent developments included the use of similarity-based metrics \cite{59, 72, 81, 125}, and advanced neural networks, including a feed-forward neural network \cite{102}, bidirectional long-short term memory network (BiLSTM) \cite{85, 110}, deep learning \cite{118}, and networks integrated in large language models (e.g., bidirectional encoder representations for transformers, BERT) \cite{117, 144}. Studies in which the machine learning model was not specified (N = 30/89, 33.7%) often reported the use of an off-the-shelf automation software (N = 27/89, 30.3%).

As an input to machine learning models, most often bag-of-words (BOW) text representations were applied (N = 30/89, 33.7%) \cite{32, 33, 35, 36, 38, 40, 42, 46, 47, 49, 57, 58, 62, 65, 69, 74, 83–85, 93, 94, 96, 98, 105, 110, 124–126, 145, 151}, followed by term-frequency / inverse document frequency (TF-IDF) (N = 16/89, 18.0%) \cite{37, 41, 48, 60, 65, 66, 79–81, 92, 96, 113, 125, 128, 129, 138}, topic models (N = 10/89, 11.2%) \cite{34, 44, 60, 64, 66, 80, 81, 85, 113, 123, 125}, keywords (N = 9, 10.1%) \cite{35, 66, 80, 108, 110, 134, 135, 138, 152}, standardized terms such as Medical Subject Headings (MeSH) (N = 6/89, 6.7%) \cite{33, 57, 62, 66, 72, 138}, or semantic annotation to the Unified Medical Language System (UMLS) (N = 6/89, 6.7%) \cite{33, 41, 42, 72, 79, 123}, named entity recognition \cite{85, 130, 144}, various word or document vector representations (N = 10, 11.2%) \cite{64, 96, 98, 106, 110, 116, 118, 123, 125, 135}, or various BERT models (N = 5, 5.6%) \cite{102, 117, 144, 149, 153}. As raw input, most studies used PubMed records including title, abstract, MeSH terms, and in a few instances, bibliographic details. Few studies used full-text input (N = 5, 5.6%) \cite{79, 91, 108, 113, 134} and database records from ClinicalTrials.gov or Cochrane (N = 4/89, 4.5%) \cite{81, 125, 143, 144}. We note that some studies were conducted on published SLR databases, such as the EPPI Centre database \cite{41, 49, 60, 64, 74} or those from the Oregon Drug Effectiveness Review Project (DERP) \cite{32, 34, 36, 38, 40, 59, 65, 69, 72, 85, 96, 98, 141}. Links to public SLR resources were extracted and provided in Table S4.
The off-the-shelf or freeware screening automation software were Abstrackr, EPPI Reviewer, RobotAnalyst, Distiller SR, Rayyan, Systematic Review Accelerator, RCT Tagger, SWIFT Review, SyRF, ASR (Automated Systematic Review), ASReview, Aggregator, ATCER, Cochrane RCT Classifier, Covidence, Curious Snake, DoCTER, GAP Screener, MetaPreg, Research Screener, revtools, RobotAnalyst, and TeMMPo. The detailed description of these tools is beyond the scope of this study. The weblinks to these tools were extracted from the references and are provided in Table S4.

The great variety of applied automation strategies, reported performance metrics, and applied datasets prevented a level performance comparison of automated record screening tools. A key observation is that, although the mean performance of automation tools improved over time, their performance varied greatly across different research topics covered by SLRs. On 15 SRLSs of the Oregon DERP dataset, the mean workload saved over sampling at 95% recall (WSS@95) of automation tools increased from 23.4% in 2006 (range 0.31%-70.5%) through 33.5% in 2010 (range 8.5% − 62.5%) to 48.4% in 2016 (range 13.7%-82.6%) and 41.0% in 2017 (range 5.8%-81.6%). On the same dataset, the WSS@95 of Rayyan was 49%±18%.

The variability of performance was illustrated by the post-hoc analysis of results using a PICO-based term recognition strategy in study titles. The single keyword “Parkinson's,” appearing in most records of a SLR deteriorated the specificity of the automated screener leading to only 11% workload savings. When omitting terms related to participants, the workload savings increased to 57% in the same dataset. In contrast, the original strategy yielded 78% workload savings in an SLR focused on phenytoin use for seizure prophylaxis in brain injury.

The time saving achieved by automated record screening also varied. The median time saving was 29.8 hours per study (range 11.7–198 hours) across 10 SLRs, with a mean time saving of 32.5 seconds per record (range: 18.1–43.5 seconds). Another study reported median 26 hours time saving across 16 SLRs (range: 9−42 hours), with a mean time saving of 22.6 seconds per record in a subset of 10 SLRs (range: 9.6–27.0 seconds). Other studies reported 23.5, 44.7, and 61, 64, and 92 hours time savings per SLR. In the study of Hamel et al., the median time saving increased from 29.8 to 36 hours when the averted workload of full-text selection was also taken into account. Time savings were also affected by the learning curve of reviewers. In a SLR involving 10599 records, manual screening of all records took 61 hours (20.7 second per record), while screening the first 1809 records to train the automation tool took 16.3 hours (32.4 second per record). The time saving per record was 15.2 second.

Full-text selection

Six papers (4.9%) focused on automated full-text selection. Most studies searched keywords using text-mining tools. The first paper, an ASLR from 2016 used Linux bash to search keywords in full-text PDF files. Another study comparing automation with duplicate human reviewers used QDA Miner. An environmental health SLR used the segmenteR R package to extract terms from specified article sections. A large environmental health ASLR used Distiller SR. Two studies aiming to dramatically speed up the SLR process applied a
convenience tool for navigation and full-text management in a reference management software (Systematic Review Accelerator)\textsuperscript{18,119}.

Time saving was reported in one study: 30.5 hours were saved on the automated full-text selection of 555 articles (198 seconds per article)\textsuperscript{108}.

**Data extraction**

Thirteen studies (10.6\%) involved an automated data extraction tool. The first paper published in 2010 introduced ExaCT, a rule-based tool to extract clinical trial characteristics\textsuperscript{39}. The efficiency of ExaCT was prospectively compared with that of human reviewers, and showed modest time savings\textsuperscript{121}. Further four papers applied text mining to create structured summaries of relevant pieces of information from full text documents. Out of these, three studies used in-house packages including UMLS semantic annotation\textsuperscript{63}, keyword search\textsuperscript{134}, and PICO entity recognition using BERT\textsuperscript{149}. The fourth tool, developed for public health purposes, Dextr\textsuperscript{148} combined vector embedding text representation and deep learning. Further approaches included PECO tagging in a rapid evidence mapping study using SWIFT Review\textsuperscript{91}, extraction of geographic locations from the manuscript\textsuperscript{133}, extraction of endpoints as comparative claim sentences\textsuperscript{55}, data extraction from ClinicalTrials.gov for meta-analyses\textsuperscript{97}, and convenience tools to highlight relevant sentences\textsuperscript{130}, or extract data from graphs\textsuperscript{88}. Finally, development of the REDASA Covid-19 dataset involved human experts in the loop, web-crawling and a natural language processing search engine to provide a real-time curated open dataset for evidence syntheses to aid pandemic response\textsuperscript{127}.

Using automated data extraction, the mean time savings per included study were 454\textsuperscript{148}, 691\textsuperscript{121}, and 1440\textsuperscript{97} seconds. The synthesized outcomes per study ranged between 5\textsuperscript{148} and 24\textsuperscript{97}. The time saving depended on the applied automation strategy. In a study by Gates et al\textsuperscript{121}, when automated data extraction was used to expedite a second reviewer, the time saving was 3.7 hours on a SLR involving 75 studies. However, when automation replaced the second reviewer, the time saving increased to 14.4 hours. Mean time saving was 352 seconds per graph when using a convenience data extraction tool\textsuperscript{88}.

**Risk of bias assessment**

Nine (7.3\%) studies looked into the automation of risk of bias assessment. The first studies were published in early 2016 introducing RobotReviewer\textsuperscript{67} and an alternative prototype tool, Systematic Review Assistant\textsuperscript{68}. Both tools were trained on the Cochrane Database for Systematic Reviews. Following the Cochrane Risk of Bias tool for randomised controlled trials (RCTs), RobotReviewer provides an overall assessment of risk of bias, and extracts supporting sentences from PDF files of full-text reports\textsuperscript{67}. RobotReviewer was used in further five studies\textsuperscript{18,99,103,119,137}. One paper assessed the risk of bias in preclinical animal studies, comparing various techniques including recurrent neural networks with attention, convolutional neural networks and BERT\textsuperscript{150}. Tangentially related to risk of bias assessment, an environmental health study automatically ranked papers based on their data quality\textsuperscript{134}.

Using RobotReviewer, the mean time saving on automated risk of bias assessment per study was 69 seconds in 52 RCTs (755 vs 824 seconds)\textsuperscript{99}. In another SLR risk of bias assessment using seven domains of the
The Cochrane Collaboration’s RoB tool required 23 hours and 40 minutes for 16 studies (5340 seconds per study), while RobotReviewer finished in 2 hours and 12 minutes assessing four risk of bias domains (495 seconds per study), saving 4845 seconds per study \(^{119}\).

**Evidence synthesis**

We identified two papers on automated evidence synthesis, both published in 2022. One of them applied a full SLR automation workbench involving automated data extraction followed by combined script for effect size calculation and meta-analysis (MetaPreg) \(^{130}\). The other paper introduced the DIAeT tool for generating qualitative evidence summary sentences from clinical trials \(^{142}\).

**Assessment of evidence quality**

We identified two papers focusing on the automated assessment of evidence quality using a semi-automated quality assessment tool (SAQAT). SAQAT is based on a Bayesian network classifier that assigns probabilities to overall GRADE (Grades of Recommendation, Assessment, Development, and Evaluation) categories using a set of standardised questions. Both papers were published in 2015 \(^{23,24}\).

**Reporting**

We identified one study from 2022, where automated report generation was part of an integrated SLR automation workflow using MetaPreg, an integrated SLR automation platform focusing on medicines during pregnancy \(^{130}\).

**Automating multiple stages of the SLR workflow**

While most papers focused on a single SLR stage, eleven studies (8.9%) automated multiple stages. Using the Systematic Review Accelerator, a team was able to complete the SLR process within a two-week timeframe by automating multiple SLR stages including search, record screening, full text selection and risk of bias assessment \(^{18,119}\). In one of these studies, time savings were documented versus a manual work. The SLR involved 586 records and 16 studies. The full manual review took 126 hours (out of which 25 hours was spent on task learning), and automation was applied on SLR stages taking 41 hours and 33 minutes to complete (out of which learning time was 6 hours 5 minutes). For the same SLR stages, automation took 11 hours and 48 minutes (including 1 hour and 18 minutes for learning the tasks), saving 30 hours, which amounted to 23.8% of the total completion time. Another team also automated multiple steps of the SLR using MetaPreg and finished a SLR in 14 days, saving 10.7 workdays compared to a conventional SLR approach \(^{130}\). Others combined multiple open-access tools including SWIFT Review, R, and Python packages to automate the record screening, full-text selection and data extraction of a SLR on the toxic effects of nanomaterials \(^{134}\). Some studies combined two stages from either search, screening, full-text selection or data extraction. These studies included two ASLRs \(^{113,132}\), studies on alternative SLR approaches, such as Rapid Evidence Mapping \(^{91}\) and Potential Technologies Review \(^{98}\), and the REDASA Covid-19 dataset \(^{127}\). A study used automated record screening before evaluating a text mining algorithm for full text selection \(^{108}\), and another automated record screening in connection with PICO named entry recognition for data extraction \(^{149}\).

**Google scholar citations**
The average number of citations per article was 122.3 (range: 0-9015, median: 22). The most cited paper (published in 2016) introduced Rayyan, a leading SLR platform (N = 9015) \(^{69}\), followed by an ASLR on mindfulness for smoking cessation (N = 526) \(^{136}\), a study introducing Curious snake, a freeware active learning-based screening automation tool (N = 323) \(^{41}\), the seminal study from Cohen et al, introducing an automated classifier tool and WSS@95, a key performance metric for screening automation (N = 320) \(^{32}\), and an ASLR on leptospirosis transmission (N = 304) \(^{61}\). Further nine SSAMs \(^{18,39,40,49,53,64,65,67,100}\), and two ASLRs \(^{50,147}\) received over 100 citations. From the nine highly cited SSAMs four introduced automation tools, such as the revtools R package for screening \(^{100}\), the SWIFT Review text mining tool \(^{65}\), ExaCT for automatic extraction of clinical trial data \(^{39}\), and RobotReviewer for automated assessment of risk of bias in clinical trials \(^{67}\), and five reported methodological innovation, such as completing a SLR in 2 weeks \(^{18}\), reducing workload in extreme reviews with 1 million records \(^{53}\), certainty-based screening in active learning \(^{49}\), topic detection based on paragraph vectors in active learning \(^{64}\), and an improved automated classification algorithm \(^{40}\).

**Summary of automated systematic reviews**

The topics of ASLRs were usually broad, with on average 17952 records (range 962-52219) and 691 included studies (range 13-6305). From the 15 ASLRs, four (26.7%) reviews automated the search \(^{113,43,61,136}\), eleven (73.3%) the screening \(^{50,83,89,95,110,113,120,132,146,147}\), two (13.3%) the full text selection \(^{71,132}\) and one (6.7%) the data extraction phase \(^{133}\). One study did not report the software \(^{110}\), six used open source software \(^{43,61,71,83,113,133}\), and eight studies used off-the shelf tools \(^{50,89,95,120,132,136,146,147}\). Three studies (20.0%) reported recall with values between 96%-100% \(^{89,95,132}\). Workload saved on screening could be obtained from eight (53.3%) studies \(^{50,71,83,89,95,110,113,132}\) with values ranging between 31.7%-100%. Some studies used automated screening to extend manual searches, thereby increasing the sensitivity of the reviews at the cost of minimal extra screening effort \(^{95,120}\). Details of the ASLRs are provided in Table 2.
Table 2
Characteristics of automated SLRs (N = 15)

<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of records&lt;sup&gt;a&lt;/sup&gt; / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oertelt-Prigione, 2011&lt;sup&gt;43&lt;/sup&gt;</td>
<td>Compare gender-related aspects of studies in stroke and myocardial infarction</td>
<td>962 / 405</td>
<td>Search</td>
<td>Text-mining was used to aid PubMed search. No further details were reported.</td>
<td>Apache Lucene</td>
<td>R: na / WLS: na</td>
</tr>
<tr>
<td>Mytton, 2014&lt;sup&gt;50&lt;/sup&gt;</td>
<td>To identify qualitative studies on facilitators and barriers of engagement in parenting programs</td>
<td>12249 / 26</td>
<td>Screening</td>
<td>Automatic term recognition was trained on 7246 citations screened by a single reviewer, and then applied on all records (n = 12249, i.e.: 444/11805). After confirming eligibility, 37 citations were selected via automatic term recognition for full-text assessment.</td>
<td>EPPI-Reviewer 4</td>
<td>R: na / WLS: 37.2%</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of records after removing duplicates

<sup>b</sup> WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

<sup>c</sup> the authors reported that all manual tasks were performed by two independent reviewers
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of records&lt;sup&gt;a&lt;/sup&gt; / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trypsteen, 2016&lt;sup&gt;71&lt;/sup&gt;</td>
<td>Map the use of droplet digital PCR (ddPCR) in HIV virus quantification.</td>
<td>2565 / 19</td>
<td>Full text selection</td>
<td>After database search, 2206 full text PDF files were collected, and searched for the presence of relevant keywords. The resulting 42 papers were manually examined for eligibility.</td>
<td>Linux Bash</td>
<td>R: na / WLS: 100% (manual screening was omitted)</td>
</tr>
</tbody>
</table>

---

<sup>a</sup> Number of records after removing duplicates

<sup>b</sup> WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

<sup>c</sup> the authors reported that all manual tasks were performed by two independent reviewers
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of records&lt;sup&gt;a&lt;/sup&gt; / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xiong, 2018&lt;sup&gt;83&lt;/sup&gt;</td>
<td>Meta-analysis on the relative risk of atrial fibrillation in diabetes mellitus.</td>
<td>4177 / 29</td>
<td>Screening</td>
<td>Search in title (n = 139, i/e: 26/113), manual selection of relevant seed studies. Then search in all fields (n = 4177), followed by K-means clustering and maximum entropy classification on similarity to seed studies. Records in most similar cluster (n = 416, i/e: 38/378) were manually screened. Studies for meta-analysis (n = 29) were selected manually. Manual screening in pairs (n = 4177, i/e: 45/4132) also yielded 29 studies for meta-analysis.</td>
<td>[R]</td>
<td>R: na / WLS: 87%</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of records after removing duplicates

<sup>b</sup> WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

<sup>c</sup> the authors reported that all manual tasks were performed by two independent reviewers
Currie, 2019 89

Systematic review and meta-analysis of chemotherapy-induced peripheral neuropathy (CIPN).

<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of recordsa / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Currie, 2019 89</td>
<td>Systematic review and meta-analysis of chemotherapy-induced peripheral neuropathy (CIPN).</td>
<td>33,814 / 180</td>
<td>Original review</td>
<td>Screening</td>
<td>Using the original review’s duplicate manual screening results as training set (n = 33,814, i/e: 6,506/27,308), a ML classifier was run on records from updated search (n = 11,880). Model selection / evaluation were performed on randomly selected 10% / 10% records screened manually in duplicate. The classifier with best precision at cut-off for 0.95 recall was selected. Then relevant chemotherapy terms for CIPN were sought by text-mining in titles/abstracts (n = 6,108, i/e: 928/5,180) to select included records for full-text selection.</td>
<td>SyRF (retrieved from reference)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Updated review</td>
<td>11,880 / 157</td>
<td></td>
<td></td>
<td>Further 85% workload saving on full text selection due to text-mining:</td>
</tr>
</tbody>
</table>

---

**a** Number of records after removing duplicates

**b** WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

**c** the authors reported that all manual tasks were performed by two independent reviewers
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of recordsa / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odintsova, 2019</td>
<td>A comprehensive overview of reviews on the genetics of human aggression, and primary genome-wide association studies (GWASs).</td>
<td>Reviews: 1686 + 13572 / 18 + 4</td>
<td>Screening</td>
<td>Using a manually annotated dataset (n = 2955, i/e: 152/2803) the ASR software was trained on samples with different i/e ratios (n = 500). The model with greatest precision at recall ≤ 0.03 was applied to classify the retrieved records for reviews (n = 1713, i/e: 1081 / 695), GWAS studies (n = 356, i/e: 243/113) and records from a broad search (n = 13572 after removing duplicates, i/e: 6469/7103).</td>
<td>ASR (Automated Systematic Review)</td>
<td>Reviews: R: 100% / WLS: 31.7%</td>
</tr>
<tr>
<td>Li, 2020</td>
<td>Review of satellite Earth observation (EO) or geographic information system (GIS) data</td>
<td>Screening</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>R: na / WLS: 85.7%</td>
</tr>
</tbody>
</table>

---

a Number of records after removing duplicates

WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript).

c the authors reported that all manual tasks were performed by two independent reviewers.
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of records$^a$ / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
<th>Automation evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>identifying landscape factors that affect dengue fever transmission.</td>
<td>Records (n = 7696) were filtered using text scoring manual weights on pre-selected keywords to select initial training set (n = 2034), followed by active learning in 5 cycles, using an initial training dataset from text scoring (n = 45, i.e.: 15/30). A word2vec CBOW model with BiLSTM algorithm was used (deep active learning). All records designated as potentially relevant were screened manually (n = 1056, i.e.: 131/925). In consecutive training cycles relevant records were combined with randomly selected irrelevant records from text scoring in 1:2 ratio, until all records were classified.</td>
<td>Records (n = 1056) were manually screened records. No relevant records were found manually among the 925 records classified as irrelevant by the algorithm.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Number of records after removing duplicates

$^b$ WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

$^c$ the authors reported that all manual tasks were performed by two independent reviewers
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of records(^a) / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thiabaud, 2020 (^113)</td>
<td>To review the sociobehavioral factors influencing HIV prevalence and incidence in Malawi.</td>
<td>16942 / 27</td>
<td>Full text selection</td>
<td>Pdf files were automatically retrieved after search (n = 22709, i/e: 16942/5767), pre-processed, and analysed via topic modelling (625 topics). Titles and abstracts of full-text papers in the 14 relevant topics were screened manually (n = 519, i/e: 119/400). From 119 selected full-text papers, 20 were eligible. Additional 7 papers were identified among the references of included papers.</td>
<td>[Python]</td>
<td>R: na / WLS: 93.2%</td>
</tr>
</tbody>
</table>

\(^a\) Number of records after removing duplicates

\(^b\) WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

\(^c\) the authors reported that all manual tasks were performed by two independent reviewers
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of records&lt;sup&gt;a&lt;/sup&gt; / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaskins, 2021&lt;sup&gt;120&lt;/sup&gt;</td>
<td>To review from professional (healthcare, exercise and fitness) staff perspective the factors affecting the implementation of aerobic exercise after stroke.</td>
<td>11449 / 20</td>
<td>Screening</td>
<td>Screening was completed manually by pairs of reviewers (n = 11449, i/e: 331 / 11118). Rayyan was trained on manual results, and 200 most relevant records were screened manually by a single reviewer (i/e: 162 / 38). Records were re-screened manually (n = 493, i/e: 63/434). 63 full-text papers were assessed for eligibility (i/e: 20/43).</td>
<td>Rayyan</td>
<td>R: na / WLS: na</td>
</tr>
<tr>
<td>Carlson, 2022&lt;sup&gt;132&lt;/sup&gt;</td>
<td>Prepare a systematic evidence map for per- and polyfluoroalkyl substances (PFAS).</td>
<td>52219 / 339</td>
<td>Search Screening Full text selection</td>
<td>(Two reviews’ results are combined: Naftion + 150 PFAS) Database search yielded 52219 records after deduplication (Naftion PubMed)</td>
<td>SWIFT Review, Distiller SR</td>
<td>For SWIFT-Review Active Screener: R: 96% / WLS: 59.0%</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of records after removing duplicates

<sup>b</sup> WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

<sup>c</sup> the authors reported that all manual tasks were performed by two independent reviewers
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of records&lt;sup&gt;a&lt;/sup&gt; / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results, recall / WLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>syntax was</td>
<td></td>
<td>The title/abstract screening was completed in 94 working hours.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>automatically</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>created by</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SWIFT Review</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>„Find Chemical</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Synonyms“)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16378 records</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>were de-</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>prioritised and</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>removed.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Records relevant</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>to human health</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>were selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>by SWIFT-</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Review Evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Streams (n =</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35841, i/e:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15414/20427).</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Relevant PFAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>records were</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>screened by</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>title/abstract</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>via active</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>learning to</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>meet PECO</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>criteria using</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SWIFT-Review</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Active</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Screener, with</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cut-off set at</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.95 recall (n =</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13161, i/e:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1483/11678).</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>In active</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>screening</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5390 records</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>were screened</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>manually.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Additional</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>records from grey</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sources were</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>added, and records</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>were screened</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>using Distiller</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SR via keyword</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of records after removing duplicates

<sup>b</sup> WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

<sup>c</sup> the authors reported that all manual tasks were performed by two independent reviewers
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of records a / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>De Menezes, 2022 133</td>
<td>Geographical distribution of gender-related topics in arboviral vector control literature</td>
<td>7367 / 2812</td>
<td>Data extraction</td>
<td>After manual search, geographic locations were extracted from 2812 records (title, abstract).</td>
<td>[R]</td>
<td>R: na / WLS: na</td>
</tr>
<tr>
<td>Jackson, 2022 136</td>
<td>To evaluate the efficacy of mindfulness-based interventions for smoking cessation among smokers.</td>
<td>2900 / 55</td>
<td>Search</td>
<td>Conventional database search yielded 3557 records. 112 records were added from an automated search in Microsoft Academic using a search strategy from the Human Behaviour Change Database. After deduplication, 2900 records were processed in a manual review.</td>
<td>Microsoft Academic</td>
<td>R: na / WLS: na</td>
</tr>
</tbody>
</table>

a Number of records after removing duplicates

b WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

c the authors reported that all manual tasks were performed by two independent reviewers
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of records&lt;sup&gt;a&lt;/sup&gt; / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>van Lissa, 2022&lt;sup&gt;146&lt;/sup&gt;</td>
<td>A text-mining systematic review of phenomena relevant to adolescent emotion regulation.</td>
<td>6584 / 6305</td>
<td>Screening</td>
<td>A search string was manually constructed from keywords to retrieve relevant seed records (n = 29, retrieved: 25, missed: 4). From 6584 records after deduplication, 559 were screened by Rayyan (i/e: 367/192), followed by screening 541 records in ASReview (i/e: 456/85). Missed records were added, 6305 papers were suitable for text mining (out of scope).</td>
<td>Rayyan, ASReview</td>
<td>R: na / WLS: na</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of records after removing duplicates

<sup>b</sup> WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

<sup>c</sup> the authors reported that all manual tasks were performed by two independent reviewers
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Aim</th>
<th>N of records a / included full-text</th>
<th>Automation stage</th>
<th>Automation process</th>
<th>Automation Tool</th>
<th>Automation results (recall / WLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viner, 2022</td>
<td>To review the association of school closures with mental health, health behaviours, and well-being in children and adolescents during COVID lockdown.</td>
<td>16817 / 36</td>
<td>Screening</td>
<td>From 16817 records the authors screened 1500 to train a ML classifier to rank records by relevance. Records with relevance score above threshold were screened by two independent authors (title/abstract). A single reviewer also screened records with lower relevance (title only). Altogether 151 records were reviewed in full text.</td>
<td>EPPI-Reviewer 4</td>
<td>R: na / WLS: na</td>
</tr>
</tbody>
</table>

a Number of records after removing duplicates

b WLS: workload saved on screening (assuming that manual tasks were performed by a single reviewer unless tasks performed by two independent reviewers are explicitly reported in the manuscript)

c the authors reported that all manual tasks were performed by two independent reviewers

**Discussion**

We provided a comprehensive overview of SLR automation studies across all stages of the SLR workflow, featuring a detailed catalogue of 123 articles indexed in PubMed and published until November 2022. The number of papers and available tools have shown rapid growth over time. Automation tools were developed for all stages of the SLR workflow, with majority of research (72%) focusing on the record screening phase. Most included articles (88%) were SSAMs with only 12% ASLRs, suggesting that the uptake of SLR automation tools in real practice is still in its infancy. The use of automated search, screening, full text selection and data extraction was demonstrated in published ASLRs, even in combination 132.

It has been demonstrated that an integrated automation workflow over multiple SLR stages can lead to savings in reviewer effort and expedite the SLR process 18,119,130. While some integrated SLR automation toolkits are
available \cite{18,119,130}, most available tools can automate only a single SLR stage, with potentially limited impact on the entire review process. Even when employing automation on multiple SLR stages, the time savings compared to the total review process duration remained modest \cite{119}. It is difficult to predict, what are the effects of SLR automation on the entire review. The performance of automation tools varies largely across review topics \cite{32,40,59,65,76}. Achievable time savings depend on various factors, including the extent to which automation replaces human reviewers \cite{121}, the impact of automating one SLR stage on the workload of subsequent review tasks \cite{109}, the baseline speed of the manual reviewer team \cite{99,119}, the complexity of the research question \cite{97,148}, the learning curve of reviewers \cite{108}, and the overall size of the review (i.e., the number of records and eligible articles). Moreover, the diverse automation strategies, datasets and performance metrics complicate the assessment of the utility of available tools. Altogether, standardized reporting practices and evaluation metrics would be helpful to keep track of the progress in SLR automation. The frequently incomplete reporting of automation performance in ASLRs also calls for better reporting standards.

Workload savings via automated record screening may come at the cost of imperfect sensitivity, which has been shown to impact the results of meta-analyses \cite{107}. The consequences of reduced sensitivity may vary between SLRs and should be carefully considered on a case-by-case basis. However, automation can increase the sensitivity of SLRs, when applied in addition to manual screening. In some ASLRs extending manual work with automated record screening increased the sensitivity of SLRs with minimal extra effort \cite{95,120}.

The citation analysis provided insights into the most impactful research articles concerning SLR automation. While the introduction of an off-the self SLR management tool was the most cited paper in this review \cite{69}, some highly cited papers indicated considerable interest about open-source tools \cite{41,100}, multiple stages of automation including screening \cite{41,100}, text mining \cite{65}, data extraction \cite{39} and risk of bias assessment \cite{67}. Solutions enabling extreme performance, such as completing a SLR in 2 weeks \cite{18} or the screening of 1 million records \cite{53} were also frequently cited.

Compared to existing reviews in SLR automation, our review has unique features. Although the SLR automation toolbox (http://systematicreviewtools.com), an online inventory of SLR automation tools provides a comprehensive collection of available solutions \cite{154}, our review also covered methods in development and published SLRs using automation techniques. By combining the search syntaxes of four published reviews in the field, the coverage of our study was broader than reviews focusing on specific aspects of SLR automation, including a review of text-mining for study identification (N = 44) \cite{22}, data extraction (N = 26) \cite{28}, retrieval of high-quality clinical studies (N = 10) \cite{30}, SLR software packages including those with automation features \cite{155,156}, reviews using AI-based automation (N = 12) \cite{157}, a living review of automated data extraction tools (N = 53) \cite{158}, or the syntheses of workload reduction via automated screening (N = 21 and N = 86) \cite{27,159}. Some reviews aimed for full coverage of SLR automation. Van Dinter et al. \cite{29} identified 41 studies, while a recent scoping review on the use of AI in biomedical literature analyses covered 273 research articles, although with broader focus including the assembly of evidence (N = 127), literature mining (N = 112) and quality analysis (N = 34) \cite{160}.

Automation or semi-automation of record screening was the most active area of research covered by several systematic reviews. A review of 44 studies reported WSS@95 values between 30%-70\% \cite{22}. A meta-analysis of
15 studies reported WSS at maximal recall levels in a range of -0.3–89.7%. Mean recall was 92.8% (95%CI 87.8%-95.8%) in this sample. A recent meta-analysis of 21 studies reported mean WSS@95 of 55% (95%CI 51%-58%) 27. Similar to our findings, the authors commented on diverse reporting practices, and the scarcity of direct comparative studies on automation tools 22,27. While considerable workload savings are achievable, consistent performance at high recall levels is still elusive, leaving human screening indispensable 159.

The low overlap of between the search results of previous SLRs on SLR automation underscore the challenges associated with identifying relevant research in this field. These challenges arise due to the blurred boundaries between SLR automation and more general approaches in medical information management. For example, the seminal article by Aphinyanaphongs from 2005 161, which is considered by many authors as the inaugural paper for automated record screening, was excluded during our record screening due to the lack of specific reference to systematic reviews. Conversely, we excluded many papers on methods with potential applicability for systematic reviews, but without testing their performance in a systematic review context. Standardised terminology, performance criteria, evaluation methods, and reporting of SLR automation research papers would help the scientific community to keep track of the developments and make informed decisions about the adoption of SLR automation tools. At the meeting point of medicine and computer science, the consolidation of terminology, definitions, and reporting standards seems to be a general challenge including digital health 162 or medical AI research 163.

The breadth and depth of our review, the coverage of both methodological development and the application of automation methods, and unique elements, such as citation analysis are strengths of our review. However, our research has limitations. The search was restricted to PubMed, the main resource for biomedical literature. However, relevant papers indexed elsewhere may have been missed. Also, although uncertain items were discussed, some records may have been lost in the screening by single reviewers. Furthermore, some decisions about the eligibility of certain papers were challenging, and relied on personal judgements, despite the predefined inclusion and exclusion criteria. The same applies to our judgements during data extraction, when characterising the sometimes abundant and complex methodological details of studies. However, the accidentally omitted records or methodological details would not alter the overall findings of our review. While our review focused on SLRs of biomedical literature, we assume that findings about the applied technologies and focus of research may be generalised to automated SLRs in scientific fields outside medicine.

Conclusions

While record screening is the most active area of research, automation tools are being developed for all stages of the SLR workflow (i.e., search, record screening, full-text selection, data extraction, risk of bias assessment, evidence synthesis, assessment of evidence quality, and reporting) and have been shown to save reviewer effort or expedite the SLR process. However, the real world adoption of SLR automation techniques is still limited. The performance (i.e., sensitivity and specificity) of automation techniques varies largely between SLRs, and it is difficult to predict their ultimate benefit in real world applications. Most tools are available for the automation of a single SLR stage, while the potential time savings compared to the entire review process are modest even if multiple stages or the SLR workflow are automated. Standardised terminology, reporting practices and evaluation metrics would enhance the real-life adoption of SLR automation practices. Given the increasing demand for evidence syntheses in medical research and medical decision-making, it is important
that more researchers become familiar with the use of SLR automation techniques, and experience accumulates over a greater evidence base. Until the benefits and risks of SLR automation are better understood, automation tools could be used more often in parallel with manual reviews. Complementing manual reviews with automation techniques could facilitate the developments in the field, with potentially increasing the sensitivity or quality of published SLRs with acceptable extra reviewer effort.

Declarations

Conflict of interest

The authors report no conflict of interest.

Acknowledgement

We thank for the usage of ELKH Cloud (https://science-cloud.hu/) that significantly helped us achieve the results published in this paper. Laura Gulyás-Oldal and Gábor Kertész from Óbuda University provided support in early phases of this review.

This project has been supported by the National Research, Development, and Innovation Fund of Hungary, financed under the TKP2021-NKTA-36 funding scheme (Project: Development and evaluation of innovative and digital health technologies; Subproject: Evaluation of digital medical devices: efficacy, safety and social utility).

Funding

This project has been supported by the National Research, Development, and Innovation Fund of Hungary, financed under the TKP2021-NKTA-36 funding scheme (Project: Development and evaluation of innovative and digital health technologies; Subproject: Evaluation of digital medical devices: efficacy, safety and social utility).

References


**Figures**
Figure 1

PRISMA Flowchart
Figure 2

Distribution of articles by publication year
Figure 3

Number of articles by the automated stage of the systematic literature review (SLR) process

*Articles with automation of multiple SLR stages were counted at each stage.

Supplementary Files

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

- ESM.docx