Original Paper

The Phases of Living Evidence Synthesis Using AI: Living Evidence Synthesis (Version 1)

Xuping Song[1,2,3,4], PhD; Zhenjie Lian[1,2], MSc; Rui Wang[2,5], MSc; Ruixin Li[1,2], MSc; Zhenzhen Yang[1,2], MSc; Xufei Luo[4,6], PhD; Lei Feng[1], BSc; Zhiming Ma[1], BSc; Zhen Pu[1], BSc; Qi Wang[7], PhD; Long Ge[1,2,3,4], PhD; Caihong Li[8], PhD; Yaolong Chen[1,3,4,6], PhD; Kehu Yang[1,2,3,4], MSc; John Lavis[9], PhD

[1]School of Public Health, Lanzhou University, Lanzhou, Gansu, China

[2]The Centre of Evidence-based Social Science, Lanzhou University, Lanzhou, Gansu, China

[3]Key Laboratory of Evidence Based Medicine & Knowledge Translation of Gansu Province, Lanzhou, Gansu, China

[4]WHO Collaborating Centre for Guideline Implementation and Knowledge Translation, Lanzhou, Gansu, China

[5]Dingxi Center for Disease Control and Prevention, Dingxi, Gansu, China

[6]Evidence-Based Medicine Center, School of Basic Medicine, Lanzhou University, Lanzhou, Gansu, China

[7]School of Public Health, University of Hong Kong, Hong Kong, China

[8]School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China

[9]Department of Health Research Methods, Evidence, and Impact, McMaster Health Forum, McMaster University, Hamilton, ON, Canada

Corresponding Author:

Kehu Yang, MSc
School of Public Health, Lanzhou University
No. 222 South Tianshui Road, Lanzhou
Lanzhou, Gansu 753000
China
Phone: +86 13893117077
Fax: +86 13893117077
Email: yangkh-ebm@lzu.edu.cn

Abstract

Background: Living evidence (LE) synthesis refers to the method of continuously updating systematic evidence reviews to incorporate new evidence. It has emerged to address the limitations of the traditional systematic review process, particularly the absence of or delays in publication updates. The emergence of COVID-19 accelerated the progress in the field of LE synthesis, and currently, the applications of artificial intelligence (AI) in LE synthesis are expanding rapidly. However, in which phases of LE synthesis should AI be used remains an unanswered question.

Objective: This study aims to (1) document the phases of LE synthesis where AI is used and (2) investigate whether AI improves the efficiency, accuracy, or utility of LE synthesis.

Methods: We searched Web of Science, PubMed, the Cochrane Library, Epistemonikos, the Campbell Library, IEEE Xplore, medRxiv, COVID-19 Evidence Network to support Decision-making, and McMaster Health Forum. We used Covidence to facilitate the monthly screening and extraction processes to maintain the LE synthesis process. Studies that used or developed AI or semiautomated tools in the phases of LE synthesis were included.

Results: A total of 24 studies were included, including 17 on LE syntheses, with 4 involving tool development, and 7 on living meta-analyses, with 3 involving tool development. First, a total of 34 AI or semiautomated tools were involved, comprising 12 AI tools and 22 semiautomated tools. The most frequently used AI or semiautomated tools were machine learning classifiers (n=5) and the Living Interactive Evidence synthesis platform (n=3). Second, 20 AI or semiautomated tools were used for the data extraction or collection and risk of bias assessment phase, and only 1 AI tool was used for the publication update phase. Third, 3 studies demonstrated the improvement in efficiency achieved based on time, workload, and conflict rate metrics. Nine studies applied AI or semiautomated tools in LE synthesis, obtaining a mean recall rate of 96.24%, and 6 studies achieved a mean F_1-score of 92.17%. Additionally, 8 studies reported precision values ranging from 0.2% to 100%.

Conclusions: AI and semiautomated tools primarily facilitate data extraction or collection and risk of bias assessment. The use of AI or semiautomated tools in LE synthesis improves efficiency, leading to high accuracy, recall, and F_1-scores, while precision varies across tools.

Trial Registration: OSF Registries 87tp4; https://osf.io/4fvdq/overview

Introduction

Evidence synthesis refers to an approach where data across studies are identified and combined to gain a clearer understanding of a body of research [1]. There is typically a significant gap between the time when a search is performed and the time when the results are published, often exceeding a year [2]. Furthermore, only a limited number of reviews are updated once they have been published [3]. This process can result in missing evidence, potentially affecting the accuracy of the findings. The approach of living evidence (LE) synthesis has been developed to address this challenge.

The method of constantly updating a systematic synthesis of evidence to incorporate newly available evidence is known as LE [4]. Elliott et al [5] developed the basis of the LE model in 2014, which effectively incorporates and summarizes new evidence. The LE synthesis process includes 4 phases: database searching and eligibility assessment, data extraction or collection and risk of bias assessment, synthesis and analysis, and publication update [6]. It has also been adapted in areas such as network meta-analysis and guidelines. The onset of COVID-19 increased the incentive to use LE [7]. Unlike traditional evidence synthesis, which requires the redeployment of significant resources for updates, the maintenance of an LE synthesis can require more modest resources [8]. However, LE synthesis that focuses on evolving topics may have a reduced reliability compared to traditional evidence synthesis. The incorporation of artificial intelligence (AI) techniques has the potential to enhance the reliability of LE synthesis by, for example, leveraging advanced algorithms to continuously assess and filter the most relevant and high-quality evidence [9].

The field of AI, which encompasses machine learning, deep learning, natural language processing, data mining, image recognition, and computer vision, to name a few, has the potential to enhance the efficiency of LE synthesis [10, 11]. In 2013, Adams et al [11] indicated that leveraging AI to automate the LE synthesis procedures could simplify the regular updating and maintenance of evidence. The development of AI systems, particularly AI based on large language models (LLMs), such as the generative pretrained transformer, has significantly advanced natural generative language systems [12]. Various AI-driven tools have been developed for different phases of LE synthesis, such as crowdsourcing and task-sharing platforms like HDAS [13]. However, the performance of the AI techniques and the phases of LE synthesis where AI is used remain unclear.

Overall, the objectives of this review are (1) to conduct a review analyzing the phases of LE synthesis that use AI and (2) to explore whether AI can improve the efficiency, accuracy, or utility of LE synthesis.

Methods

This is the first version of an LE synthesis. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement for living systematic reviews (PRISMA-LSR; Checklist 1) was used as a guide for reporting this LE synthesis [14]. The review has been registered in the Open Science Forum [15].

Search Strategy

We systematically searched the Web of Science, PubMed, the Cochrane Library, Epistemonikos, the Campbell Library, IEEE Xplore, medRxiv, COVID-19 Evidence Network to support Decision-making, and McMaster Health Forum for publications up to April 2, 2025. The details of the search strategy used can be found in Table S1 in Multimedia Appendix 1. We subscribed to the Web of Science, PubMed, the Cochrane Library, the Campbell Library, and IEEE Xplore for monthly dynamic updates and used Covidence to facilitate the screening and extraction processes for maintaining an LE synthesis. We plan to conduct living updates for a 12-month period (from April 2025 to April 2026). The final update is scheduled for April 2, 2026, after which we will assess whether to retire the living mode based on the following established triggers: (1) evidence on "the AI application in LE synthesis" has reached conclusiveness, (2) the topic no longer holds decision-making value for the field, (3) no new eligible studies emerge during the 12-month update period, or (4) subsequent resource or funding support is unavailable [16,17].

Inclusion and Exclusion Criteria

First, the LE synthesis includes living systematic review, living meta-analysis, living network meta-analysis, living guideline, living scoping review, living overview, living umbrella review, and living mapping. In this review, the types of included studies were classified into 2 categories based primarily on whether a meta-analysis had been performed. These categories include the LE synthesis (without a meta-analysis) and living meta-analysis (with a meta-analysis conducted).

Second, the criteria for inclusion in this review are studies that use AI or semiautomated tools in the following phases of LE synthesis: (1) database searching and eligibility

assessment, (2) data extraction or collection and risk of bias assessment, (3) synthesis and analysis, or (4) publication update [6]. The LE syntheses from any field were included. In addition, studies that developed AI or semiautomated tools for LE synthesis were also included. Textbox 1 provides further details.

Textbox 1. Inclusion and exclusion criteria for the study.

Inclusion criteria
- The studies using artificial intelligence (AI) or semiautomated tools in the following phases of living evidence (LE) synthesis: (1) database searching and eligibility assessment, (2) data extraction or collection and risk of bias assessment, (3) synthesis and analysis, or (4) publication update. A study can be any type of LE synthesis in any field, including but not limited to all scientific journals in the social sciences.
- Studies that developed AI or semiautomated tools for LE synthesis.

Exclusion criteria
- Studies that did not document the use of AI or semiautomated tools in LE synthesis.
- Protocol, commentaries, editorials, letters to the editor, and updating studies.

We excluded studies that did not document the use of AI or semiautomated tools in LE synthesis. In addition, protocols, commentaries, editorials, letters to the editor, and updating studies were also excluded, as shown in Textbox 1.

Third, AI tools are characterized by autonomous learning and end-to-end decision-making. They enable the independent execution of data collection, feature extraction, model training, and inference and generate output results without any human intervention. However, semiautomated tools incorporate human review or decision support at critical stages, using a "machine assistance and human oversight"

collaborative paradigm [18,19]. Textbox 2 shows the types of AI or semiautomated tools, where AI or semiautomated tools were categorized by the application phases. First, the first segment of the AI or semiautomated tools for each phase is sourced from Bendersky et al [13]. Second, the subsequent segment is derived from the work of Khalil et al [20]. Third, for the final segment, AI or semiautomated tools were identified and summarized from relevant studies using a manual search. The AI techniques based on LLMs, such as the generative pretrained transformer, were also included.

Textbox 2. Artificial intelligence (AI) or semiautomated tools used in the 4 phases of living evidence (LE) synthesis.

Phase 1. Database searching and eligibility assessment
- Segment 1.1: Automatic, continuous database search with push notification, database aggregators (such as HDAS, Epistemonikos), notification from clinical trial registries, randomized clinical trial classifier, text mining technologies, and automatic retrieval of full-text papers
- Segment 1.2: RCT tagger, LitSuggest, Evidence mapping tool, SRA-Polyglot Search Translator, QuickClinical, HDAS, ROBOTsearch, SRA-word, frequency analyzer, The Search Refiner, Sherlock, SRA De-duplicate, Distiller, R package-rev tools, Rayyan, EPPI-reviewer, Abstrackr, SRA helper, LibSVM classifier, Bibot, Active Screener, RobotAnalyst, Swift-Review, Evidence Pipeline, JBI Sumari, EndNote, SARA, eSuRFr, ParsCit, and Citation searcher
- Segment 1.3: Natural language processing–assisted abstract screening tool, automatic text classifiers supported by deep learning–based language models, machine learning classifiers, Cochrane Crowd, Living Interactive Evidence (LIvE) synthesis platform, Cochrane RCT classifier, OpenAlex, Risklick AI, Bayesian classifier, Generative Pretrained Transformer models, and RobotReviewer LIVE

Phase 2. Data extraction or collection and risk of bias assessment
- Segment 2.1: Machine learning information-extraction systems, automated structured data extraction tools for PDFs, machine learning–assisted RoB tool, data repositories, and linked data
- Segment 2.2: RobotReviewer, DistelleR, JBI Sumari, in-house data extraction tool written in R, statistical package R, ExaCT, Revman, Raptor, ContentMine, Graph2Data, and Evidence mapping tool
- Segment 2.3: BioMart, MetaInsight COVID-19, LIvE synthesis platform, Open Science Framework (OSF), PsychOpen CAMA, and Generative Pretrained Transformer models

Phase 3. Synthesis and analysis
- Segment 3.1: Structured data extraction tools, which automatically provide data in a suitable format for statistical analysis; continuous analysis updating based on availability of structured extracted data; and statistical surveillance of key analysis results, with threshold set for potential conclusion change
- Segment 3.2: MetaPreg, MetaXL, NetMetaXL, Meta-analyst, Webplotdigitizer, Evidence mapping tool, PRISMA flow diagram generator, Evidence mapping tool, R package-rev tools
- Segment 3.3: Risklick AI, Web Source Processing Pipeline, LIvE synthesis platform, and generative pretrained transformer models

Phase 4. Publication update

- Segment 4.1: Templated reporting of some report items, automatic text generation tools for synthesis and writing, automatization in the identification of changes between LSR versions for peer review, and editorial process (such as Archie)
- Segment 4.2: Trial2rev, RevManHAL, DistelleR, SRA replicant writer, SRA-RevMan Replicant, and JBI Sumari
- Segment 4.3: Generative pretrained transformer models

Study Screening and Data Collection

Two reviewers independently screened the titles and abstracts of all selected studies, followed by a full-text review. Any disagreements regarding selection were resolved by a third researcher. Data were extracted using a predesigned Microsoft Excel sheet. Two reviewers independently extracted data from all included studies, including information such as title, first author, journal, year of publication, LE synthesis type, types of tool or technology, types of AI or semiautomated tools, phases of LE synthesis, outcomes, and so forth. Any disagreements were resolved by a third researcher. During data extraction, representative outcomes (such as means or ranges) were prioritized for synthesis, with the range of values considered subsequently when outcomes were similarly representative.

Methodological Quality Assessment

Given the lack of a standardized tool for assessing the methodological quality of AI-related studies, the 24 studies were categorized into 3 types by methodological characteristics and primary objective (diagnostic test, tool development, or—when neither applied—a general synthesis) and assessed for methodological quality using the modified version of the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) tool, Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Textual Evidence: Narrative, and AMSTAR 2 tool. First, 10 studies were assessed with the modified version of the QUADAS-2 tool: these studies specifically assessed the application of AI in the database searching and eligibility assessment phase, which aligns with a diagnostic test accuracy (DTA) framework. We adopted the modified version of the QUADAS-2 proposed by Rashid et al [21-23]. As QUADAS-2 is designed for DTA research contexts, this framework was only applicable to those studies where one of the objectives included the application of AI

in the database searching and eligibility assessment phase [21, 24,25]. The core elements of QUADAS-2 were revised to adapt it to AI-related research scenarios, as follows: "patient" was replaced with "study," "index test" with "AI," "reference standard" with "comparator," and "case-control design" with "DTA framework." We also constructed a 2×2 table, categorizing studies into "included" or "excluded" based on both "AI screening results" and "reference/original systematic review (SR) screening results," with counts denoted as a, b, c, and d, respectively. The details of the modified QUADAS-2 are provided in Table S2 in Multimedia Appendix 1. Second, 5 studies, which specifically developed AI or semiautomated tools for LE synthesis without DTA-related accuracy evaluation and were not designed as LE synthesis themselves, were assessed using the JBI Critical Appraisal Checklist for Textual Evidence: Narrative [26]. Third, 9 studies, which were designed as LE syntheses without DTA-related accuracy evaluation and not primarily focused on AI or semiautomated tool development (or tool development was only an auxiliary means), were assessed using the AMSTAR 2 tool [27, 28]. The details are shown in Tables S3 and S4 in Multimedia Appendix 1. All of the included studies were evaluated independently by 2 reviewers (RL and ZY), and disagreement was resolved by a third reviewer (ZL). The LE synthesis did not involve a statistical combination of results (meta-analysis), as its aims were to document the phases of LE synthesis where AI is used and to investigate whether AI improves the efficiency, accuracy, or utility of LE synthesis. Therefore, several systematic review procedures—including sensitivity analyses, reporting bias assessment, certainty assessment, and investigations of heterogeneity—were not used.

Data Analysis

This review conducted 3 complementary analyses, as shown in Figure 1.

Figure 1. Road map for the use of artificial intelligence (AI): applications and extractable clinical outcomes across 4 phases of living evidence synthesis. LE: living evidence.
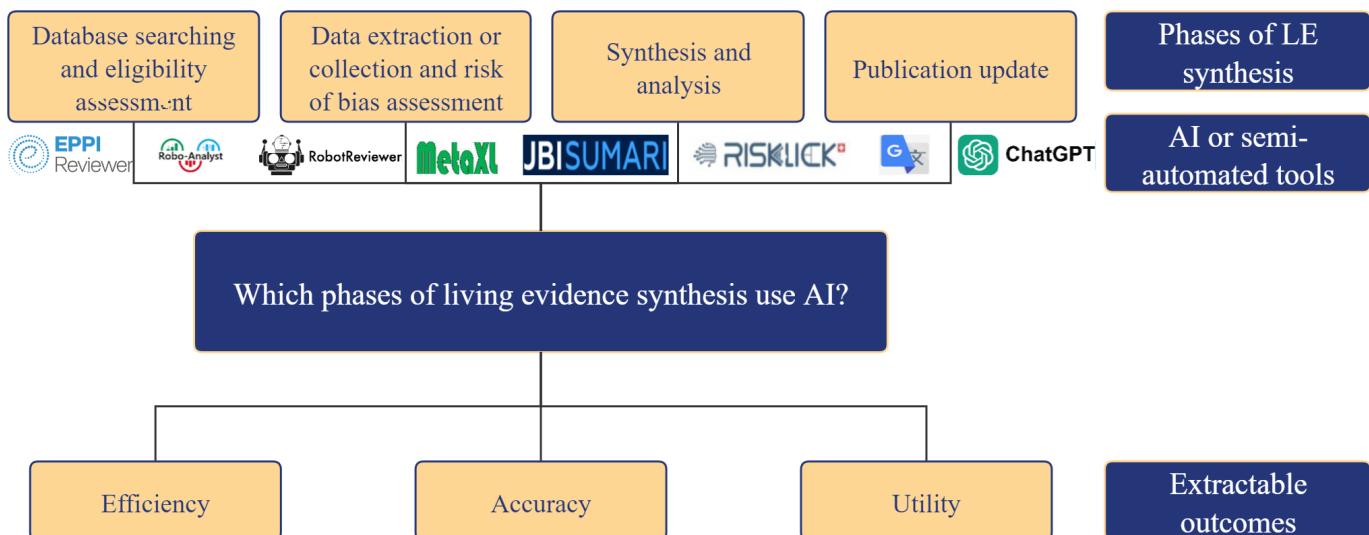
Analysis 1: Phases of LE Synthesis Utilizing AI or Semiautomated Tools

We analyzed the prevalence and distribution of AI or semiautomated tools across 4 phases of LE synthesis. Phase 1 is database searching and eligibility assessment. This process includes going through the databases, retrieving the results, importing them into the citation management software, removing any duplicate results, and assessing their eligibility individually. Phase 2 is data extraction or collection and risk of bias assessment; once the eligibility of studies has been verified and they have been included in the review process, it becomes crucial to systematically extract and collect information about their main characteristics and results. Additionally, it is very important to assess the risk of bias associated with the conduct and methodology used in the studies. In phase 3—synthesis and analysis—the data that have been assessed to conform to the criteria are integrated, and the data are analyzed. In phase 4—publication update—after going through the aforementioned phases 1-3, sections of a review are generated based on their results, and conclusions are updated.

Analysis 2: AI or Semiautomated Tools Used in LE Synthesis

First, the types of AI or semiautomated tools applied in each LE synthesis phase were investigated. Second, the frequency of AI or semiautomated tools applied in the LE synthesis was analyzed.

Analysis 3: Primary Outcomes Investigating AI or Semiautomated Tools in LE Synthesis

The impact of applied AI or semiautomated tools in LE synthesis was analyzed across 3 outcomes [29]. First, efficiency, defined as the relationship between the time required to complete a workload and the workload itself, was evaluated to determine whether either the duration or workload was reduced with the use of AI or semiautomated tools. This outcome may be described as time reduction, workload reduction, and conflict rates with and without the tool.

Second, accuracy is used to assess performance with and without AI or semiautomated tools. It may be described as accuracy, recall, precision, F_1-score, area under the receiver operating characteristic curve, number needed to read, and study relevance. In addition, we calculated the overall mean recall and mean F_1-score using the following formula:

$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

where M_i is the representative value for study i, defined as the reported single value, if provided, or the midpoint of the reported range [L, U], calculated as (L+U)/2, if a range was provided. N is the number of studies reporting that metric [30,31].

Third, utility is used to assess whether user decisions align with those of AI or semiautomated tools, including user consistency, user satisfaction, perceived ease of use, and study quality.

Results

Search Results

Out of 9180 studies, 24 studies applied AI or semiautomated tools in LE synthesis, including 17 LE syntheses (4 developing tools) and 7 living meta-analyses (3 developing tools), as shown in Figure 2 [29,32-54]. In addition, 8 studies exclusively applied AI tools in LE synthesis, 11 studies exclusively applied semiautomated tools, and 5 studies utilized both AI and semiautomated tools. The basic characteristics of the included studies are shown in Table S5 in Multimedia Appendix 1. The details of the studies excluded at the full-text eligibility stage with reasons are shown in Table S6 in Multimedia Appendix 1 [5,9,55-75].

Figure 2. Database search flow diagram. LE: living evidence.
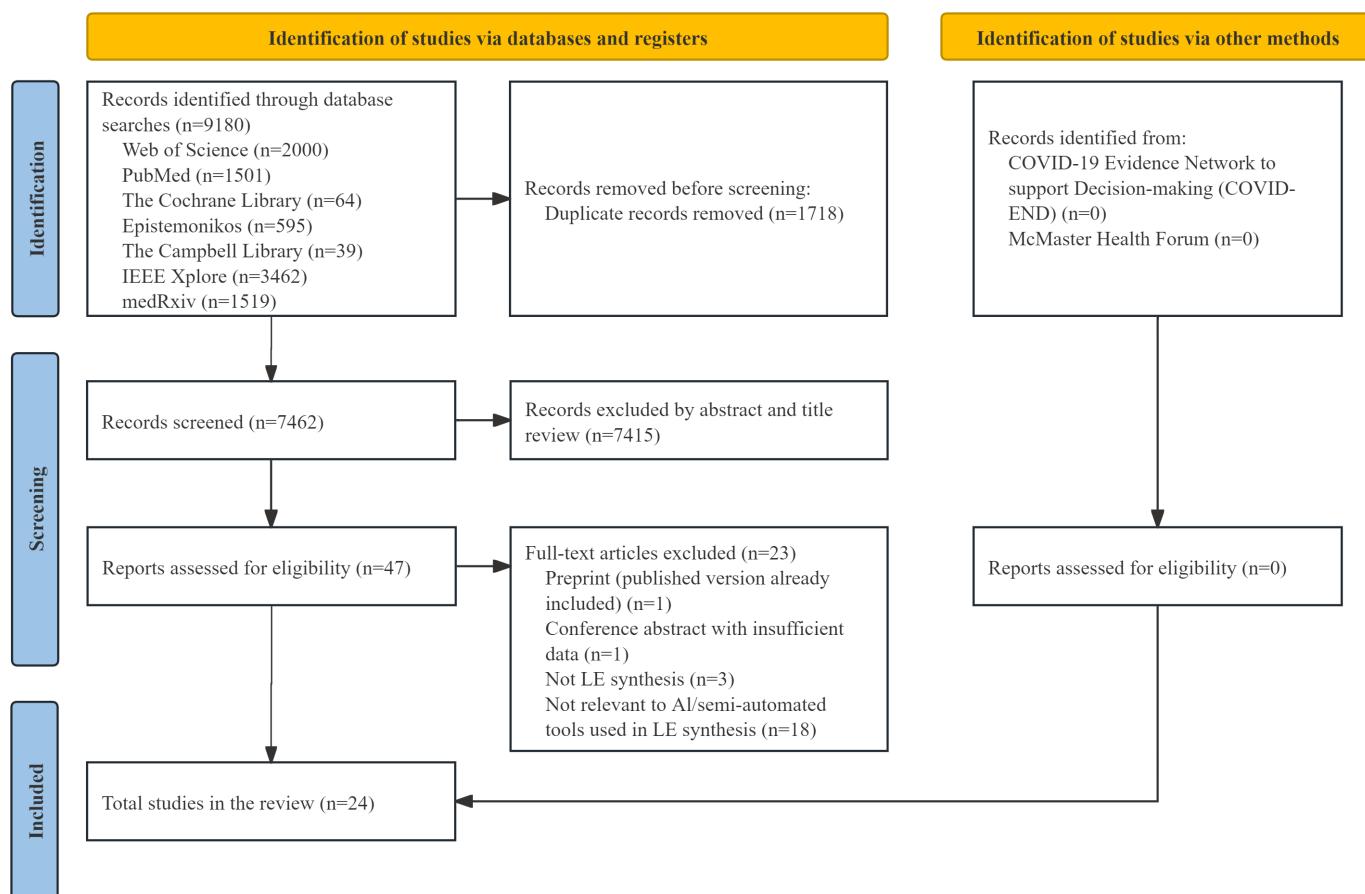
Methodological Quality of Included Studies

We conducted a methodological quality assessment of 10 studies using a revised QUADAS-2 tool within the DTA framework [29,32,35,36,42-44,51,52,54]. All studies were assessed as low-risk in the "Study selection," "Index test (AI)," and "Reference (comparator)" domains. While none of the studies specified the time interval between the task execution of AI and comparator-based analysis, all were determined as low-risk in the "Flow and timing" domain. Additionally, we did not identify any applicability concerns, as all studies were classified as low-risk in the "Applicability" domain (Table 1). Five studies were subjected to methodological quality assessment using the JBI Critical Appraisal Checklist for Textual Evidence: Narrative [41,46, 48-50]. Four studies obtained a score of 5/6, with a narrative appraisal of "Exclude" owing to failure to meet the narrative classification criterion [41,46,48,49]. One study achieved a full score of 6/6 and was thus appraised as "Include" (Table S7 in Multimedia Appendix 1) [50]. In addition, we conducted a methodological quality assessment of 9 studies using AMSTAR 2 [33,34,37-40,45,47,53]. The methodological quality scores of the included studies ranged from 11 to 15. Overall, the methodological quality of eight studies [34, 37-40,45,47,53] was rated as moderate, while only 1 study [33] was rated as low in methodological quality. The most common limitation was that the authors failed to provide a list of excluded studies (Table S8 in Multimedia Appendix 1).

Table 1. Summary of modified Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) assessments for studies using artificial intelligence (AI) or semiautomated tools in the database searching and eligibility phase of the living evidence (LE) synthesis process.

Author, year	Risk of bias				Applicability concern		
	Study selection	Index test (AI)	Reference (comparator)	Flow and timing	Study selection	Index test (AI)	Reference (comparator)
Knafou et al [32] (2023)	Low	Low	Low	Low	Low	Low	Low
Perlman-Arrow et al [29] (2023)	Low	Low	Low	Low	Low	Low	Low
Chou et al [35] (2020)	Low	Low	Low	Low	Low	Low	Low

Author, year	Risk of bias			Applicability concern			
	Study selection	Index test (AI)	Reference (comparator)	Flow and timing	Study selection	Index test (AI)	Reference (comparator)
Kamso et al [36] (2023)	Low	Low	Low	Low	Low	Low	Low
Marshall et al [42] (2023)	Low	Low	Low	Low	Low	Low	Low
Haas et al [43] (2021)	Low	Low	Low	Low	Low	Low	Low
Vaghela et al [44] (2021)	Low	Low	Low	Low	Low	Low	Low
Shemilt et al [51] (2024)	Low	Low	Low	Low	Low	Low	Low
Le-Khac et al [52] (2024)	Low	Low	Low	Low	Low	Low	Low
Hair et al [54] (2024)	Low	Low	Low	Low	Low	Low	Low

Types and Frequency of AI or Semiautomated Tools in LE Synthesis

A total of 34 AI or semiautomated tools were involved, including 12 (35.3%) AI tools and 22 (64.7%) semiautomated tools, as shown in Multimedia Appendix 2. The most frequently used AI or semiautomated tools were machine learning classifiers (n=5), followed by the Living Interactive Evidence (LIvE) synthesis platform (n=3), AD-SOLES (n=2), Covidence (n=2), and MAGICapp (n=2).

Phases of AI or Semiautomated Tools Application in LE Synthesis

There were 18 AI or semiautomated tools for database searching and eligibility assessment, 20 for data extraction or collection and risk of bias assessment, and 10 for synthesis and analysis. However, only 1 AI tool was used for publication updates. Out of all the tools, RobotReviewer LIVE can be used for all phases of LE synthesis, as shown in Textbox 3.

Textbox 3. Types of artificial intelligence (AI) or semiautomated tools applications in the 4 phases of living evidence (LE) synthesis.

> Phase 1. Database searching and eligibility assessment
> - LIvE platform, automatic text classifiers, machine learning ensemble classifier, Natural language processing–assisted abstract screening tool, machine learning classifiers, machine learning, PICO annotators, STAR tool, AD-SOLES, Covidence, rcrossref, openalexR, RISmed, RobotReviewer LIVE, Risklick AI, metaCOVID application, supervised text classification models, and text mining techniques
>
> Phase 2. Data extraction or collection and risk of bias assessment
> - LIvE platform, web-based interactive app, open-source living systematic review application, Covidence, AD-SOLES, Google Refine tool, script, REDASA, RobotReviewer LIVE, Risklick AI, Metainsight COVID-19, metaCOVID application, information extraction techniques, EndNote, semiautomated model, supervised text classification models, text mining techniques, GPT-4-turbo, Claude-3-Opus, and EPPI-Reviewer
>
> Phase 3. Synthesis and analysis
> - LIvE platform, MAGICapp, Trial sequential analysis (TSA) software, AD-SOLES, ODDPub, RobotReviewer LIVE, script, Metainsight COVID-19, metaCOVID application, and Dynameta
>
> Phase 4. Publication update
> - RobotReviewer LIVE

Impact of AI or Semiautomated Tools on LE Synthesis

Overview

A total of 10 (41.7%) studies reported on the impact of AI or semiautomated tools on LE synthesis in terms of efficiency, accuracy, or utility in the database searching and eligibility phase or the data extraction or collection and risk of bias assessment phase. Table 2 provides a description of the outcome metrics in the included studies.

Table 2. Summary of the indicator terms for outcome metrics in the included studies.

Metrics	Explanation
Efficiency	

Metrics	Explanation
Time	AI[a] or semiautomated tools were used to save time. Only 2 (8.3%) studies reported on time saving [29,35]. Specifically, Perlman-Arrow et al [29] reported a 45.9% reduction in screening time per abstract in the database searching and eligibility phase. Chou et al [35] estimated the time saving ranged from 2.0 to 13.2 hours in the database searching and eligibility phase.
Workload	Two (8.3%) studies reported on workload metrics related to the use of AI or semiautomated tools [29,42]. Perlman-Arrow et al [29] reported that the semiautomated tool completed 68% of the workload in the database searching and eligibility phase. Marshall et al [42] found that manual screening had an efficiency rate of 23% in obtaining 31 abstracts, whereas AI achieved a rate of 55%, demonstrating an efficiency improvement of approximately 140% in the database searching and eligibility phase.
Conflict rates with and without the tool	The efficiency of abstract screening decreases as the number of conflicting votes increases [29]. Perlman-Arrow et al [29] reported a reduction in conflict rates from 8.32% to 3.64% with the use of semiautomated tool in the database searching and eligibility phase.
Accuracy[b]	
Precision	Precision refers to the ratio of accurately categorized documents among all the documents that the model assigns to a particular class [32]. Eight (33.3%) studies reported on precision [29,32,35,42,43,51,53,54]. 1. Khan et al [53] reported a precision rate of even 100% using AI in the data extraction or collection and risk of bias assessment phase. 2. Perlman-Arrow et al [29] and Haas et al [43] reported precision rates of 92.10% and 96.07%, respectively, using AI or semiautomated tools in the database searching and eligibility phase. 3. Hair et al [54] reported that the average precision rate using AI is about 84.5% in the database searching and eligibility phase. 4. Shemilt et al [51] reported a precision rate of 50%-86% using AI in the database searching and eligibility phase. 5. Marshall et al [42] reported a precision rate of 55% using AI in the database searching and eligibility phase. 6. Knafou et al [32] reported a precision rate of only 29.69% using AI in the database searching and eligibility phase. 7. However, Chou et al [35] reported a precision rate of only 0.2%-8% using AI in the database searching and eligibility phase.
Recall[c]	Recall (also known as sensitivity) refers to the fraction of positive documents that have been accurately identified among all documents for the specified class [32]. Nine (37.5%) studies reported on recall [29,32,35,36,42,43,51,53,54]. All studies reported recall rates in excess of 87%. The average value was about 96.24%. 1. Perlman-Arrow et al [29], Chou et al [35], and Marshall et al [42] found recall rates of even 100% using AI or semiautomated tools in the database searching and eligibility phase. 2. Knafou et al [32], Haas et al [43], and Kamso et al [36] reported a recall rate of 89%, 99.25% and 99.3%, respectively, using AI in the database searching and eligibility phase. 3. Shemilt et al [51] reported a recall rate of 94%-99% using AI in the database searching and eligibility phase. 4. Khan et al [53] reported a recall rate of 92%-96% using AI in the data extraction or collection and risk of bias assessment phase. 5. Hair et al [54] reported that the average sensitivity rate using AI is about 95.1% in the database searching and eligibility phase.
F_1-score[c]	F_1-score refers to the balanced harmonic average between the model precision and recall [32]. Six (25%) studies reported on F_1-score [29,32,43,52-54]. All studies reported F_1-score between 80.47% and 99% after using AI. The average value was about 92.17%. 1. Knafou et al [32], Perlman-Arrow et al [29], and Haas et al [43] reported an F_1-score of 89.2%, 92.6%, and 97.59%, respectively, using AI or semiautomated tools in the database searching and eligibility phase. 2. Le-Khac et al [52] reported an F_1-score of 87% using AI in the data extraction or collection and risk of bias assessment phase. 3. Khan et al [53] reported F_1-scores between 96% and 98% after using AI in the data extraction or collection and risk of bias assessment phase. 4. Hair et al [54] reported that the average F_1-score using AI is about 89.6% in the database searching and eligibility phase.
Area under the receiver operating characteristic curve (AUC-ROC)	AUC-ROC calculates the area under the curve between the true positive rate and the false positive rate [32]. Knafou et al [32] reported higher AUC-ROC performance using AI in the database searching and eligibility phase and had an AUC-ROC performance of 94.25%-94.77%.

Metrics	Explanation
Number needed to read (NNR)	NNR refers to the total number of literature considered within the search divided by the number of literature included from the search [35]. Only 2 (8.3%) studies reported on NNR [29,35]. Perlman-Arrow et al [29] reported an NNR between 1.086 and 1.125 after using a semiautomated tool in the database searching and eligibility phase. Chou et al [35] reported an NNR between 15 and 100 after using AI in the database searching and eligibility phase.
Article relevance	Vaghela et al [44] reported on studies included after searching using AI, and 50.49% were considered relevant to the query in the database searching and eligibility phase.
Utility	
User satisfaction	Perlman-Arrow et al [29] reported that the average satisfaction of users with the tool reached 4.2/5 in the database searching and eligibility phase.
Consistency	Kamso et al [36] reported that consistency in the use of AI between 2 reviewers was assessed using percentage agreement and Kappa scores, revealing a range of percentage agreement from 79.0% to 96.0%, and a variation in Kappa scores from moderate (0.40) to substantial (0.63) in the database searching and eligibility phase.
Article quality	Vaghela et al [44] reported that 64.53% of the included studies possess reliable quality in the database searching and eligibility phase.

[a]AI: artificial intelligence.

[b]Kamso et al [36] achieved an accuracy ranging from 75.9% to 96.9% in research classification using AI in the database searching and eligibility phase. Khan et al [53] reported that the collaborative large language models' accuracy, based on concordant responses in the prompt set, reached 99% in the data extraction or collection and risk of bias assessment phase.

[c]The overall mean recall (96.24%) and F_1-score (92.17%) are the simple averages of study-level values from Table S5 in Multimedia Appendix 1. For studies reporting a range, the midpoint was used as the study-level value.

Efficiency Enhancements Through AI or Semiautomated Tools in LE Synthesis

Three studies showed improved efficiency in the database searching and eligibility phase in terms of 3 indicator terms. A total of 2 (8.3%) studies [29,35] reported on time saving with AI or semiautomated tools, 2 (8.3%) studies [29,42] reported on workload metrics related to the use of AI or semiautomated tools, and 1 study [29] reported a reduction in conflict rates with the use of semiautomated tool, which consequently increases the efficiency.

Accuracy Improvements With AI or Semiautomated Tools in LE Synthesis

A total of 9 and 6 studies that applied AI or semiautomated tools in LE synthesis reported a mean recall rate and a mean F_1-score of 96.24% and 92.17%, respectively. While Khan et al [53] reported a precision rate of even 100% achieved using AI in the data extraction or collection and risk of bias assessment phase. However, in 7 studies, the reported precision rates varied significantly, ranging from 0.2% to 96.07% in the database searching and eligibility phase.

Utility of AI or Semiautomated Tools in LE Synthesis

Three studies reported on the utility of AI or semiautomated tools in the database searching and eligibility phase of LE synthesis, including user satisfaction, consistency, and study quality. Consistency in the use of AI between 2 reviewers was assessed using percentage agreement and Kappa scores [36].

Discussion

Principal Findings

AI or semiautomated tools are actively used to facilitate the process of LE synthesis. We conducted this review to identify the phases of LE synthesis that use AI and explore whether AI can improve the efficiency, accuracy, or utility of LE synthesis.

AI or semiautomated tools have been increasingly used in LE synthesis, particularly in living systematic review. This review discovered that AI or semiautomated tools are most commonly used for data extraction or collection and risk of bias assessment. However, only a few studies have addressed the use of AI or semiautomated systems for publication updates, highlighting the need for further development in this phase.

Diverse types of AI or semiautomated tools were identified in this study. These include the LIvE synthesis platform, AD-SOLES, metaCOVID application, and RobotReviewer LIVE, which are utilized in multiple phases of LE synthesis, indicating their versatility and potential for wider adoption [37,39,40,42,47,54]. The most frequently used AI or semiautomated tools were machine learning classifiers, the LIvE synthesis platform, Covidence, AD-SOLES, and MAGICapp. Furthermore, the rapid rise of AI tools involving LLM types, such as GPT-4-turbo and Claude-3-Opus, has led to their use in LE synthesis. These tools can be suitable for application in multiple or even all phases of LE synthesis, especially in the publication update phase. The application of LLMs to further enhance the efficiency, accuracy, and utility of LE synthesis remains a key focus for researchers and practitioners.

Governments worldwide, particularly those in leading AI nations such as China, the United States, Germany, the United Kingdom, France, and Canada, are especially emphasizing the transformative impact of AI on research and decision-making processes [76,77]. Funding from various sources, including the Economic and Social Research Council, reflects a strong financial commitment to advancing AI technologies in evidence synthesis. Furthermore, a growing number of AI guidance and organizations are emerging to embrace the opportunity that AI has taken in producing LE synthesis. For example, Responsible AI in Evidence SynthEsis has provided recommendations for the main roles of responsible AI in the evidence synthesis ecosystem that are involved in responsible AI use [78]. Furthermore, organizations such as ALIVE aim to improve societal outcomes by producing and utilizing timely, trustworthy, and affordable evidence.

Challenges remain in the application of AI in LE synthesis. Machine learning classifiers suffer from low precision and varying efficiency across different topics [35]. As an example, RobotReviewer LIVE faces challenges in performance variability for complex reviews, limited study types, and data source constraints [42]. Therefore, further research aimed at enhancing the adaptability and stability of AI across various research areas is urgently needed. In addition, ethical issues, data protection measures, and transparency in AI-driven LE synthesis are also key challenges that need to be addressed [79]. At the ethical level, AI is prone to selection bias due to the skewness of its training data, which impairs the inclusivity of evidence, and the mechanism of responsibility attribution remains unclear [80]. Data protection is another area that faces challenges, as research data required for AI training often contain sensitive information, and existing anonymization technologies cannot fully avoid the risk of privacy breaches [81]. Cost considerations in the implementation of AI tools, including initial investment, ongoing operational costs, training expenses, and requirements for hardware and software resources also constitute a significant issue [82].

Policymaking involves judgment, making it more of an art than a science, whereas science is primarily driven by evidence and shapes evidence-informed policymaking [83]. Study has indicated that relying solely on systematic reviews for policymaking is far from sufficient; instead, policymakers need to obtain a more diverse range of synthesized evidence to underpin decision-making [84]. The LE synthesis, especially by incorporating AI into evidence production, can deliver updated evidence to facilitate evidence-informed policymaking. AI could revolutionize policymaking by facilitating ongoing assessments, ensuring that the policies remain aligned with the latest evidence and evolve in response to new information as it emerges [2,5,85]. Furthermore, AI enables policymakers to continuously monitor and assess policies throughout their lifecycle, which allows adaptation to shifting circumstances and evolving societal needs in real time [86]. Furthermore, the advancement of AI capabilities, particularly through LLMs, adds a deeper analytical layer; LLMs can provide nuanced insights and help predict future research directions relevant to policymaking

[87]. The application of AI in LE synthesis could transform policy decision-making, advancing policy formulation for policymakers.

Recent advances in AI provide researchers with new transformative capabilities [79]. Van Dijk et al [88] indicated that AI tools are a promising innovation in the current practice of systematic evaluation, and researchers have reported positive experiences with these tools. The use of AI enhances efficiency by significantly reducing researchers' time and workload [2,89]. Manion et al [90] indicated that natural language processing could enhance accuracy and reduce errors through a "human-in-the-loop" approach. The application of AI in LE synthesis has considerably benefited researchers, significantly enhancing their research capabilities.

This LE synthesis will retain its living mode beyond the present publication, consistent with the methodology. This decision is based on two key considerations: (1) the predefined retirement triggers have not been triggered and (2) the Safe and Responsible Use of AI Working Group (Working Group 3) and the Methods & Process Innovation Working Group (Working Group 4) of the Evidence Synthesis Infrastructure Collaborative will benefit from the continuous updates from this LE synthesis to support their future research initiatives [91-94].

Future Research Directions

In the above discussion, we have suggested the advancement of future work across multiple dimensions. From a technical point of view, efforts are needed to address limitations of existing AI tools, such as inadequate precision and poor adaptability, while deepening research into the LLM applications in the publication update phase of LE synthesis. In the realm of ethics and data governance, it is essential to establish responsibility attribution mechanisms and cross-regulatory data governance frameworks, as well as enhance evidence inclusivity and mitigate privacy risks through algorithmic optimization. Methodologically, we recommend the establishment of a standardized evaluation system for AI applications and refining research design and quality assessment protocols to strengthen the evidence base.

Strengths and Limitations

The strengths of this review include the following: (1) it systematically analyzes the types of AI and semiautomated tools used across the 4 phases of LE synthesis and (2) it provides insights into the opportunities and challenges of using AI or semiautomated tools in LE synthesis regarding efficiency, accuracy, and utility. However, this review still has a few limitations. First, study screening was based on whether the studies reported on the tools used in LE synthesis. Second, studies that did not document the use of AI or semiautomated tools in LE synthesis were excluded from this review, which may introduce bias. Third, the focus of our search strategy on "living evidence" terminology may have excluded studies describing AI tools for review updates that used different terminology.

Conclusion

Researchers are actively utilizing various AI and semiautomated tools in LE synthesis, primarily for data extraction or collection and risk of bias assessment, while their application in updating publications remains limited. The use of AI or semiautomated tools in LE synthesis improves efficiency in the database searching and eligibility phase and accuracy in the database searching and eligibility phase, as well as in the data extraction or collection and risk of bias assessment phase. The AI or semiautomated tools demonstrate high accuracy, recall, and F_1-scores, while precision varies across tools. AI or semiautomated tools also demonstrate good performance in terms of utility in the database searching and eligibility phase.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

XS designed the study, analyzed the data, and drafted the manuscript. ZL designed the study, analyzed the data, drafted the manuscript, and evaluated the quality of included studies. RW, QW, and XL developed the research design. RL and ZY drafted the manuscript and evaluated the quality of the included studies. LF, ZM, and ZP were in charge of data curation. CL, LG, YC, KY, and JL critically reviewed and revised the manuscript. All authors critically revised the study for important intellectual content and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search strategies, included and excluded study information, and methodological quality assessment methods and results.
[DOCX File (Microsoft Word File), 54 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Frequency of artificial intelligence (AI) or semiautomated tools use.
[PNG File (Portable Network Graphics File), 139 KB-Multimedia Appendix 2]

Checklist 1

PRISMA-LSR checklist.
[DOCX File (Microsoft Word File), 30 KB-Checklist 1]

References

1. Elliott J, Lawrence R, Minx JC, et al. Decision makers need constantly updated evidence synthesis. Nature. Dec 2021;600(7889):383-385. [doi: 10.1038/d41586-021-03690-1] [Medline: 34912079]
2. Sampson M, Shojania KG, Garritty C, Horsley T, Ocampo M, Moher D. Systematic reviews can be produced and published faster. J Clin Epidemiol. Jun 2008;61(6):531-536. [doi: 10.1016/j.jclinepi.2008.02.004] [Medline: 18471656]
3. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med. Aug 21, 2007;147(4):224-233. [doi: 10.7326/0003-4819-147-4-200708210-00179] [Medline: 17638714]
4. Turner T, Lavis JN, Grimshaw JM, Green S, Elliott J. Living evidence and adaptive policy: perfect partners? Health Res Policy Syst. Dec 18, 2023;21(1):135. [doi: 10.1186/s12961-023-01085-4] [Medline: 38111030]
5. Elliott JH, Turner T, Clavisi O, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med. Feb 2014;11(2):e1001603. [doi: 10.1371/journal.pmed.1001603] [Medline: 24558353]
6. Thomas J, Noel-Storr A, Marshall I, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol. Nov 2017;91:31-37. [doi: 10.1016/j.jclinepi.2017.08.011] [Medline: 28912003]
7. Tendal B, Vogel JP, McDonald S, et al. Weekly updates of national living evidence-based guidelines: methods for the Australian living guidelines for care of people with COVID-19. J Clin Epidemiol. Mar 2021;131:11-21. [doi: 10.1016/j.jclinepi.2020.11.005] [Medline: 33188858]
8. Elliott JH, Synnot A, Turner T, et al. Living systematic review: 1. Introduction-the why, what, when, and how. J Clin Epidemiol. Nov 2017;91:23-30. [doi: 10.1016/j.jclinepi.2017.08.010] [Medline: 28912002]
9. Schmidt L, Sinyor M, Webb RT, et al. A narrative review of recent tools and innovations toward automating living systematic reviews and evidence syntheses. Z Evid Fortbild Qual Gesundhwes. Sep 2023;181:65-75. [doi: 10.1016/j.zefq.2023.06.007] [Medline: 37596160]

10. Yang Y, Qin J, Lei J, Liu Y. Research status and challenges on the sustainable development of artificial intelligence courses from a global perspective. Sustainability. 2023;15(12):9335. [doi: 10.3390/su15129335]

11. Adams CE, Polzmacher S, Wolff A. Systematic reviews: work that needs to be done and not to be done. J Evid Based Med. Nov 2013;6(4):232-235. [doi: 10.1111/jebm.12072] [Medline: 24325416]

12. Suárez A, Jiménez J, Llorente de Pedro M, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J. Dec 2024;24:46-52. [doi: 10.1016/j.csbj.2023.11.058] [Medline: 38162955]

13. Bendersky J, Auladell-Rispau A, Urrútia G, Rojas-Reyes MX. Methods for developing and reporting living evidence synthesis. J Clin Epidemiol. Dec 2022;152:89-100. [doi: 10.1016/j.jclinepi.2022.09.020] [Medline: 36220626]

14. Akl EA, Khabsa J, Iannizzi C, et al. Extension of the PRISMA 2020 statement for living systematic reviews (PRISMA-LSR): checklist and explanation. BMJ. Nov 19, 2024;387:e079183. [doi: 10.1136/bmj-2024-079183] [Medline: 39562017]

15. Which phases of living evidence synthesis use artificial intelligence (AI)? an living evidence synthesis. OSF. URL: https://doi.org/10.17605/OSF.IO/4FVDQ [Accessed 2026-01-16]

16. Murad MH, Wang Z, Chu H, et al. Proposed triggers for retiring a living systematic review. BMJ Evid Based Med. Oct 2023;28(5):348-352. [doi: 10.1136/bmjebm-2022-112100] [Medline: 36889900]

17. Cochrane Living Systematic Reviews Network. Guidance for the production and publication of Cochrane living systematic reviews: Cochrane Reviews in living mode. Cochrane; 2019. URL: https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201912_LSR_Revised_Guidance.pdf [Accessed 2026-01-13]

18. Kozhakhmetova A, Mamyrbayev A, Zhidebekkyzy A, Bilan S. Assessing the impact of artificial intelligence on project efficiency enhancement. Knowl Perform Manag. 2024;8(2):109-126. [doi: 10.21511/kpm.08(2).2024.09]

19. Røhl UBU. Automated, administrative decision-making and good administration: friends, foes or complete strangers [Dissertation]. Aalborg Universitetsforlag; 2022. URL: https://vbn.aau.dk/ws/files/549540893/PHD_UBUR.pdf [Accessed 2026-01-19]

20. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol. Apr 2022;144:22-42. [doi: 10.1016/j.jclinepi.2021.12.005] [Medline: 34896236]

21. Rashid M, Yi CS, Lawin S, et al. MT14 role of generative artificial intelligence in assisting systematic review process in health research: a systematic review. Value Health. Jul 2025;28(6):S268. [doi: 10.1016/j.jval.2025.04.1124] [Medline: 40848037]

22. Validity assessment tools for evidence synthesis: your one-stop-shop. Latitudes Network. URL: https://www.latitudes-network.org [Accessed 2026-01-13]

23. Equator Network - Enhancing the QUAlity and Transparency Of health Research. URL: http://equator-network.org [Accessed 2025-08-06]

24. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. Oct 18, 2011;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]

25. Santini A, Man A, Voidăzan S. Accuracy of diagnostic tests. J Crit Care Med (Targu Mures). Jul 2021;7(3):241-248. [doi: 10.2478/jccm-2021-0022] [Medline: 34722928]

26. McArthur A, Klugarova J, Yan H, Florescu S. Chapter 4: systematic reviews of text and opinion. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis JBI. JBI; 2020. [doi: 10.46658/JBIRM-17-04]

27. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. Sep 21, 2017;358:j4008. [doi: 10.1136/bmj.j4008] [Medline: 28935701]

28. Welsh EJ, Normansell RA, Cates CJ. Assessing the methodological quality of systematic reviews. NPJ Prim Care Respir Med. Mar 19, 2015;25:15019. [doi: 10.1038/npjpcrm.2015.19] [Medline: 25789911]

29. Perlman-Arrow S, Loo N, Bobrovitz N, Yan T, Arora RK. A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Res Synth Methods. Jul 2023;14(4):608-621. [doi: 10.1002/jrsm.1636] [Medline: 37230483]

30. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions Version 6.5. Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current [Accessed 2026-01-13]

31. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. Jan 16, 2020;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]

32. Knafou J, Haas Q, Borissov N, et al. Ensemble of deep learning language models to support the creation of living systematic reviews for the COVID-19 literature. Syst Rev. Jun 5, 2023;12(1):94. [doi: 10.1186/s13643-023-02247-9] [Medline: 37277872]

33. Matl S, Brosig R, Baust M, Navab N, Demirci S. Vascular image registration techniques: a living review. Med Image Anal. Jan 2017;35:1-17. [doi: 10.1016/j.media.2016.05.005] [Medline: 27294558]

34. Schmidt L, Finnerty Mutlu AN, Elmore R, Olorisade BK, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: update of a living systematic review. F1000Res. 2021;10:401. [doi: 10.12688/f1000research.51117.3] [Medline: 34408850]

35. Chou R, Dana T, Shetty KD. Testing a Machine Learning Tool for Facilitating Living Systematic Reviews of Chronic Pain Treatments. Agency for Healthcare Research and Quality; 2020. [Medline: 33175480]

36. Kamso MM, Pardo JP, Whittle SL, et al. Crowd-sourcing and automation facilitated the identification and classification of randomized controlled trials in a living review. J Clin Epidemiol. Dec 2023;164:1-8. [doi: 10.1016/j.jclinepi.2023.10.007] [Medline: 37865299]

37. Riaz IB, Sipra Q, Naqvi SAA, et al. Quantifying absolute benefit for adjuvant treatment options in renal cell carcinoma: a living interactive systematic review and network meta-analysis. Crit Rev Oncol Hematol. Jul 2022;175:103706. [doi: 10.1016/j.critrevonc.2022.103706] [Medline: 35537621]

38. Butler AR, Hartmann-Boyce J, Livingstone-Banks J, Turner T, Lindson N. Optimizing process and methods for a living systematic review: 30 search updates and three review updates later. J Clin Epidemiol. Feb 2024;166:111231. [doi: 10.1016/j.jclinepi.2023.111231] [Medline: 38043829]

39. Riaz IB, He H, Ryu AJ, et al. A living, interactive systematic review and network meta-analysis of first-line treatment of metastatic renal cell carcinoma. Eur Urol. Dec 2021;80(6):712-723. [doi: 10.1016/j.eururo.2021.03.016] [Medline: 33824031]

40. Riaz IB, Siddiqi R, Islam M, et al. Adjuvant tyrosine kinase inhibitors in renal cell carcinoma: a concluded living systematic review and meta-analysis. JCO Clin Cancer Inform. May 2021;5:588-599. [doi: 10.1200/CCI.21.00035] [Medline: 34043431]

41. Hair K, Wilson E, Wong C, Tsang A, Macleod M, Bannach-Brown A. Systematic online living evidence summaries: emerging tools to accelerate evidence synthesis. Clin Sci (Lond). May 31, 2023;137(10):773-784. [doi: 10.1042/CS20220494] [Medline: 37219941]

42. Marshall IJ, Trikalinos TA, Soboczenski F, et al. In a pilot study, automated real-time systematic review updates were feasible, accurate, and work-saving. J Clin Epidemiol. Jan 2023;153:26-33. [doi: 10.1016/j.jclinepi.2022.08.013] [Medline: 36150548]

43. Haas Q, Alvarez DV, Borissov N, et al. Utilizing artificial intelligence to manage COVID-19 scientific evidence torrent with Risklick AI: a critical tool for pharmacology and therapy development. Pharmacology. 2021;106(5-6):244-253. [doi: 10.1159/000515908] [Medline: 33910199]

44. Vaghela U, Rabinowicz S, Bratsos P, et al. Using a secure, continually updating, web source processing pipeline to support the real-time data synthesis and analysis of scientific literature: development and validation study. J Med Internet Res. May 6, 2021;23(5):e25714. [doi: 10.2196/25714] [Medline: 33835932]

45. Karakülah G, Suner A, Adlassnig KP, Samwald M. A data-driven living review for pharmacogenomic decision support in cancer treatment. Stud Health Technol Inform. 2012;180:688-692. [doi: 10.3233/978-1-61499-101-4-688] [Medline: 22874279]

46. Xin Y, Nevill CR, Nevill J, et al. Feasibility study for interactive reporting of network meta-analysis: experiences from the development of the Metainsight COVID-19 app for stakeholder exploration, re-analysis and sensitivity analysis from living systematic reviews. BMC Med Res Methodol. Jan 22, 2022;22(1):26. [doi: 10.1186/s12874-022-01507-x] [Medline: 35065603]

47. Evrenoglou T, Boutron I, Seitidis G, Ghosn L, Chaimani A. metaCOVID: a web-application for living meta-analyses of COVID-19 trials. Res Synth Methods. May 2023;14(3):479-488. [doi: 10.1002/jrsm.1627] [Medline: 36772980]

48. Kaiser K, Miksch S. Versioning computer-interpretable guidelines: semi-automatic modeling of "living guidelines" using an information extraction method. Artif Intell Med. May 2009;46(1):55-66. [doi: 10.1016/j.artmed.2008.08.009] [Medline: 18950994]

49. Skinner G, Cooke R, Keum J, et al. Dynameta: a dynamic platform for ecological meta-analyses in R Shiny. SoftwareX. Jul 2023;23:101439. [doi: 10.1016/j.softx.2023.101439]

50. McDonald S, Hill K, Li HZ, Turner T. Evidence surveillance for a living clinical guideline: case study of the Australian stroke guidelines. Health Info Libr J. Nov 9, 2023. [doi: 10.1111/hir.12515] [Medline: 37942888]

51. Shemilt I, Arno A, Thomas J, et al. Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research. Wellcome Open Res. 2024;6:210. [doi: 10.12688/wellcomeopenres.17141.2] [Medline: 38686019]

52. Le-Khac UN, Bolton M, Boxall NJ, Wallace SMN, George Y. Living review framework for better policy design and management of hazardous waste in Australia. Sci Total Environ. May 10, 2024;924:171556. [doi: 10.1016/j.scitotenv.2024.171556] [Medline: 38458450]

53. Khan MA, Ayub U, Naqvi SAA, et al. Collaborative large language models for automated data extraction in living systematic reviews. J Am Med Inform Assoc. Apr 1, 2025;32(4):638-647. [doi: 10.1093/jamia/ocae325] [Medline: 39836495]

54. Hair K, Wilson E, Maksym O, Macleod MR, Sena ES. A systematic online living evidence summary of experimental Alzheimer's disease research. J Neurosci Methods. Sep 2024;409:110209. [doi: 10.1016/j.jneumeth.2024.110209] [Medline: 38964475]

55. Grbin L, Nichols P, Russell F, Fuller-Tyszkiewicz M, Olsson CA. The development of a living knowledge system and implications for future systematic searching. J Aust Libr Inf Assoc. Jul 3, 2022;71(3):275-292. [doi: 10.1080/24750158.2022.2087954]

56. Hearnden J, Dudoit K, Kim E, Tremblay G, Forsythe A. PMU118 use of computer-assisted methods to realize the concept of a living systematic review via an online platform. Value Health. Nov 2019;22:S729. [doi: 10.1016/j.jval.2019.09.1736]

57. Evrenoglou T, Boutron I, Chaimani A. metaCOVID: an R-Shiny application for living meta-analyses of COVID-19 trials. medRxiv. Preprint posted online on Sep 10, 2021. [doi: 10.1101/2021.09.07.21263207]

58. Stoll A, Wilms L, Ziegele M. Developing an incivility dictionary for German online discussions—a semi-automated approach combining human and artificial knowledge. Commun Methods Meas. Apr 3, 2023;17(2):131-149. [doi: 10.1080/19312458.2023.2166028]

59. Meza N, Pérez-Bracchiglione J, Pérez I, et al. Angiotensin-converting-enzyme inhibitors and angiotensin II receptor blockers for COVID-19: a living systematic review of randomized clinical trials. Medwave. Mar 3, 2021;21(2):e8105. [doi: 10.5867/medwave.2021.02.8105] [Medline: 33830976]

60. Verdejo C, Vergara-Merino L, Meza N, et al. Macrolides for the treatment of COVID-19: a living, systematic review. Medwave. Dec 14, 2020;20(11):e8074. [doi: 10.5867/medwave.2020.11.8073] [Medline: 33361755]

61. Baladia E, Pizarro AB, Ortiz-Muñoz L, Rada G. Vitamin C for COVID-19: a living systematic review. Medwave. Jul 28, 2020;20(6):e7978. [doi: 10.5867/medwave.2020.06.7978] [Medline: 32759894]

62. Rada G, Corbalán J, Rojas P, COVID-19 L·OVE Working Group. Cell-based therapies for COVID-19: a living, systematic review. Medwave. Dec 17, 2020;20(11):e8079. [doi: 10.5867/medwave.2020.11.8078] [Medline: 33382060]

63. Gates M, Elliott SA, Gates A, et al. LOCATE: a prospective evaluation of the value of leveraging ongoing citation acquisition techniques for living evidence syntheses. Syst Rev. Apr 19, 2021;10(1):116. [doi: 10.1186/s13643-021-01665-x] [Medline: 33875014]

64. Piechotta V, Iannizzi C, Chai KL, et al. Convalescent plasma or hyperimmune immunoglobulin for people with COVID-19: a living systematic review. Cochrane Database Syst Rev. May 20, 2021;5(5):CD013600. [doi: 10.1002/14651858.CD013600.pub4] [Medline: 34013969]

65. Verdugo-Paiva F, Acuña MP, Solá I, Rada G, COVID-19 L·OVE Working Group. Remdesivir for the treatment of COVID-19: a living systematic review. Medwave. Dec 9, 2020;20(11):e8080. [doi: 10.5867/medwave.2020.11.8080] [Medline: 33361753]

66. Shackelford GE, Martin PA, Hood ASC, Christie AP, Kulinskaya E, Sutherland WJ. Dynamic meta-analysis: a method of using global evidence for local decision making. BMC Biol. Feb 17, 2021;19(1):33. [doi: 10.1186/s12915-021-00974-w] [Medline: 33596922]

67. Verdugo-Paiva F, Izcovich A, Ragusa M, Rada G. Lopinavir-ritonavir for COVID-19: a living systematic review. Medwave. Jul 15, 2020;20(6):e7967. [doi: 10.5867/medwave.2020.06.7966] [Medline: 32678815]

68. Iannizzi C, Chai KL, Piechotta V, et al. Convalescent plasma for people with COVID-19: a living systematic review. Cochrane Database Syst Rev. Feb 1, 2023;2(2):CD013600. [doi: 10.1002/14651858.CD013600.pub5] [Medline: 36734509]

69. Sommer I, Ledinger D, Thaler K, et al. Outpatient treatment of confirmed COVID-19: a living, rapid evidence review for the American College of Physicians (Version 2). Ann Intern Med. Oct 2023;176(10):1377-1385. [doi: 10.7326/M23-1626] [Medline: 37722115]

70. Verdugo-Paiva F, Vergara C, Ávila C, et al. COVID-19 living overview of evidence repository is highly comprehensive and can be used as a single source for COVID-19 studies. J Clin Epidemiol. Sep 2022;149:195-202. [doi: 10.1016/j.jclinepi.2022.05.001] [Medline: 35597369]

71. Paul D, Chakdar D, Saha S, Mathew J. Online research topic modeling and recommendation utilizing multiview autoencoder-based approach. IEEE Trans Comput Soc Syst. 2024;11(1):1013-1022. [doi: 10.1109/TCSS.2023.3253502]

72. Vergara-Merino L, Verdejo C, Carrasco C, Vargas-Peirano M. Living systematic review: new inputs and challenges. Medwave. Dec 23, 2020;20(11):e8092. [doi: 10.5867/medwave.2020.11.8092] [Medline: 33382391]

73. Elbers S, Wittink H, Kaiser U, et al. Living systematic reviews in rehabilitation science can improve evidence-based healthcare. Syst Rev. Dec 7, 2021;10(1):309. [doi: 10.1186/s13643-021-01857-5] [Medline: 34876231]

74. Elvidge J, Hopkin G, Narayanan N, Nicholls D, Dawoud D. Diagnostics and treatments of COVID-19: two-year update to a living systematic review of economic evaluations. Front Pharmacol. 2023;14:1291164. [doi: 10.3389/fphar.2023. 1291164] [Medline: 38035028]

75. Winters M, Lyng KD, Holden S, et al. Infographic. Comparative effectiveness of treatments for patellofemoral pain: a living systematic review with network meta-analysis. Br J Sports Med. Nov 2021;55(22):1311-1312. [doi: 10.1136/ bjsports-2021-104360] [Medline: 34244170]

76. Katanandov SL, Kovalev AA. Technological development of modern states: artificial intelligence in public administration. State and Municipal Management Scholar Notes. Mar 2023;1(1):174-182. [doi: 10.22394/2079-1690- 2023-1-1-174-182]

77. Medaglia R, Gil-Garcia JR, Pardo TA. Artificial intelligence in government: taking stock and moving forward. Soc Sci Comput Rev. 2021;41(1):123-140. [doi: 10.1177/08944393211034087]

78. Thomas J, Flemyng E, Noel-Storr A, et al. Responsible AI in Evidence SynthEsis (RAISE): guidance and recommendations. Open Science Framework; 2024. URL: https://osf.io/fwaud/files/cn7x4 [Accessed 2026-01-13]

79. Filetti S, Fenza G, Gallo A. Research design and writing of scholarly articles: new artificial intelligence tools available for researchers. Endocrine. Sep 2024;85(3):1104-1116. [doi: 10.1007/s12020-024-03977-z] [Medline: 39085566]

80. Ramnani S. Exploring ethical considerations of artificial intelligence in educational settings: an examination of bias, privacy, and accountability. International Journal of Novel Research and Development. 2024;9(2):b173-b191. [doi: 10. 1729/Journal.37869]

81. Li Y, Shao S, He Y, et al. Rethinking data protection in the (generative) artificial intelligence era. arXiv. Preprint posted online on Jul 3, 2025. [doi: 10.48550/arXiv.2507.03034]

82. Umeh II, Umeh KC. A comparative analysis of AI system development tools for improved outcomes. International Journal of Sustainability Management and Information Technologies. 2025;11:1-20. [doi: 10.11648/j.ijsmit.20251101. 11]

83. Ramírez G. Improving the health of populations—evidence for policy and practice action. J Evid Based Med. Nov 2009;2(4):216-219. [doi: 10.1111/j.1756-5391.2009.01044.x] [Medline: 21349019]

84. Manson H. Systematic reviews are not enough: policymakers need a greater variety of synthesized evidence. J Clin Epidemiol. May 2016;73:11-14. [doi: 10.1016/j.jclinepi.2015.08.032] [Medline: 26912122]

85. Berger-Tal O, Wong BBM, Adams CA, et al. Leveraging AI to improve evidence synthesis in conservation. Trends Ecol Evol. Jun 2024;39(6):548-557. [doi: 10.1016/j.tree.2024.04.007] [Medline: 38796352]

86. Jacob S. Artificial intelligence and the future of evaluation: from augmented to automated evaluation. Digit Gov Res Pract. Mar 31, 2025;6(1):1-10. [doi: 10.1145/3696009]

87. Head CB, Jasper P, McConnachie M, Raftree L, Higdon G. Large language model applications for evaluation: opportunities and ethical implications. New Dir Eval. Jun 2023;2023(178-179):33-46. [doi: 10.1002/ev.20556]

88. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open. Jul 7, 2023;13(7):e072254. [doi: 10.1136/bmjopen- 2023-072254] [Medline: 37419641]

89. Thomas IN, Roche P, Grêt-Regamey A. Harnessing artificial intelligence for efficient systematic reviews: a case study in ecosystem condition indicators. Ecol Inform. Nov 2024;83:102819. [doi: 10.1016/j.ecoinf.2024.102819]

90. Manion FJ, Du J, Wang D, et al. Accelerating evidence synthesis in observational studies: development of a living natural language processing-assisted intelligent systematic literature review system. JMIR Med Inform. Oct 23, 2024;12:e54653. [doi: 10.2196/54653] [Medline: 39441204]

91. Glanville J. The role of AI tools in developing search strategies and identifying evidence for systematic reviews. Evidence Synthesis Ireland; 2025. URL: https://evidencesynthesisireland.ie/wp-content/uploads/2025/05/ESI-2025-AI- tools-Julie-Glanville.pdf [Accessed 2026-01-13]

92. New Joint AI Methods Group: guiding responsible use of AI in evidence synthesis. Joanna Briggs Institute (JBI). URL: https://jbi.global/news/article/new-joint-ai-methods-group [Accessed 2026-01-13]

93. Roadmap. Evidence Synthesis Infrastructure Collaborative. 2024. URL: https://evidencesynthesis.atlassian.net/wiki/ spaces/ESE/pages/344817670/English [Accessed 2026-01-13]

94. Scotcher S. Evidence Synthesis Infrastructure Collaborative. European Evaluation Society. 2025. URL: https:// europeanevaluation.org/events/evidence-synthesis-infrastructure-collaborative [Accessed 2026-01-13]

Abbreviations

AI: artificial intelligence

DTA: diagnostic test accuracy
JBI: Joanna Briggs Institute
LE: living evidence
LIvE: Living Interactive Evidence
LLM: large language model
PRISMA-LSR : Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement for living systematic reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies version 2