Received: 2 November 2018 | Revised: 7 August 2019 | Accepted: 6 September 2019

DOI: 10.1002/jrsm.1378

RESEARCH ARTICLE

WILEY

Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources

Michael Gusenbauer[1] | **Neal R. Haddaway**[2,3]

[1]Institute of Innovation Management, Johannes Kepler University Linz, Linz, Austria

[2]Stockholm Environment Institute, Linnégatan 87D, Stockholm, Sweden

[3]Africa Centre for Evidence, University of Johannesburg, Johannesburg, South Africa

Correspondence
Michael Gusenbauer, Institute of Innovation Management, Johannes Kepler University Linz, Linz, Austria.
Email: michael.gusenbauer@jku.at

Rigorous evidence identification is essential for systematic reviews and meta-analyses (evidence syntheses) because the sample selection of relevant studies determines a review's outcome, validity, and explanatory power. Yet, the search systems allowing access to this evidence provide varying levels of precision, recall, and reproducibility and also demand different levels of effort. To date, it remains unclear which search systems are most appropriate for evidence synthesis and why. Advice on which search engines and bibliographic databases to choose for systematic searches is limited and lacking systematic, empirical performance assessments. This study investigates and compares the systematic search qualities of 28 widely used academic search systems, including Google Scholar, PubMed, and Web of Science. A novel, query-based method tests how well users are able to interact and retrieve records with each system. The study is the first to show the extent to which search systems can effectively and efficiently perform (Boolean) searches with regards to precision, recall, and reproducibility. We found substantial differences in the performance of search systems, meaning that their usability in systematic searches varies. Indeed, only half of the search systems analyzed and only a few Open Access databases can be recommended for evidence syntheses without adding substantial caveats. Particularly, our findings demonstrate why Google Scholar is inappropriate as principal search system. We call for database owners to recognize the requirements of evidence synthesis and for academic journals to reassess quality requirements for systematic reviews. Our findings aim to support researchers in conducting better searches for better evidence synthesis.

KEYWORDS

academic search systems, discovery, evaluation, information retrieval, systematic review, systematic search

1 | INTRODUCTION

Research output, as measured by the number of academic publications, continues to grow exponentially,[1,2] placing scientists in danger of becoming decoupled from the discourse with which they are engaged. The growing volume of research makes it ever harder for practitioners and researchers to keep track of past and current findings in a specific discipline and across disciplines. As a result, research agendas neither build on nor advance previous findings but exist in isolation from the greater body of evidence. Scientific discourse and cumulative knowledge are threatened if researchers fail to connect their empirical or theoretical analyses with past knowledge. As a consequence, the relevance and impact of their research is reduced.[3] In academia, it is true that "we are drowning in information, but starving for knowledge"(4, p12).

Evidence syntheses—the collective term for robust summaries of evidence—aim to mitigate the issues of decoupled empirical evidence amidst ever-growing research output. In particular, high-quality evidence synthesis, in the form of systematic reviews, systematic maps, and meta-analyses (which themselves should be based on systematic searches and critical appraisal), aims "to produce an unbiased description of the cumulative state of evidence on a research problem or hypothesis"(5, p32) and syntheses are thus "viewed as the most reliable sources of evidence for practice and policy."[6] In this way, evidence synthesis is capable of highlighting important developments in a particular field of research.

It is the way in which evidence synthesis is undertaken that "may enhance or undermine the trustworthiness of its conclusion or, in common social science parlance, can create threats to the validity of its conclusions."(5, p33) By following strict rules and guidance, a systematic review provides a comprehensive synthesis of a well-defined area of research. The research team conducting the review must be capable of undertaking online searches using a fit-for-purpose set of search systems, which will enable the researchers to search for and identify *all* available relevant research in a procedurally *unbiased* manner.[7] The constructs and phenomena in question need to be well-defined by the review team so that the online search can use these linguistic cues as frames for searching. Documentation of the search process "makes the search replicable and provides a clear starting point for later updates."[8] Building a systematic review team that incorporates diverse expertise in areas such as content, systematic review methods, searching, and quantitative synthesis has been shown to significantly improve the quality of the review work.[9] Building on an understanding of constructs, a systematic search is the gatekeeper that establishes the basis for subsequent synthesis.[10] This process defines the scope of the examination and thus influences the outcome of the analysis. The same analysis employed with different samples might lead to different results. Similarly, as search systems differ in functionality and characteristics, the same query employed with a different search system may result in a different sample.

Today, evidence synthesis can benefit considerably from innovations in information and communication technology. The introduction of improved tools and methods (also called "evidence-synthesis technology") makes it easier to conduct synthesis work. Technologies such as word processing and reference management software, data analysis, and web-based literature searches allow the more efficient and effective identification, analysis, synthesis, and reporting of research. In particular, online literature search tools now cover most disciplines and have made millions of scientific records searchable within seconds; in some cases, free-of-charge. However, the time saved in physically

What is already known

- Evidence identification in systematic reviews and meta-analyses requires the right search strategy using the right search systems.
- Until now, researchers have lacked comprehensive guidance on which search systems are suitable for systematic searches.

What is new?

- This study provides a systematic evaluation and comparison of search and retrieval qualities of 28 widely used academic search systems, including Google Scholar, PubMed, and Web of Science.
- Evaluation profiles for each of those 28 systems allow researchers to assess why and to what degree a particular system is suitable for their search requirements.

Potential impact for RSM readers outside the authors' field

- By making qualities and limitations of search systems transparent, this study creates awareness across disciplines among journals and among database providers to pay particular attention to the search requirements of evidence synthesis.
- We hope our findings assist researchers to perform better searches that require less time, identify more relevant evidence, and adhere to systematic review guidance.

searching for evidence must now be spent in carefully planning searches that make use of complex syntax and search facilities to mine the textual data within titles, abstracts, keywords, and full texts. Accordingly, collaboration with librarians, as information experts in these complex literature search processes, has been shown to benefit study quality.[9,11,12] In evidence syntheses, online databases should "form the backbone of any comprehensive literature search. These sources probably contain the information most closely approximating all research."(5, p112) Indeed, it is now impossible to imagine undertaking academic work without using web-based literature search systems.

Rigorous evidence syntheses, such as systematic reviews, have specific requirements for literature searches.[13] These requirements are stipulated in *conduct* guidance issued by renowned institutions dedicated to warrant and elevate the quality of evidence synthesis in academia. We have based our further analysis on guidance published by three institutions: Cochrane, The Campbell Collaboration, and the Collaboration for Environmental Evidence (CEE). We decided not to include PRISMA and ROSES guidance, as these resources offer guidance on *reporting* rather than *conduct*. Below, we provide an overview of how searches for studies to be included in the evidence base should be performed in systematic reviews.

TABLE 1 Quality requirements of systematic searches derived from evidence synthesis guidelines

Source	Quality Requirements
Cochrane Handbook, 2011	"The key characteristics of a systematic review are: [...] an explicit, reproducible methodology; a systematic search that attempts to identify all studies that would meet the eligibility criteria [...]."[14]
Campbell Methods Guides, 2016	"Systematic reviews of interventions require a thorough, objective and reproducible search of a range of sources to identify as many relevant studies as possible (within resource limits)."[15]
CEE Guidelines and Standards for Environmental Evidence Synthesis, 2018	"To achieve a rigorous evidence synthesis searches should be transparent and reproducible and minimise biases. A key requirement of a review team engaged in evidence synthesis is to try to gather a maximum of the available relevant documented bibliographic evidence in articles and the studies reported therein."[16]

Guidance for systematic reviews (Table 1) refer to three recurring quality requirements that are critical for literature searches: First, the goal must be to identify *all relevant records* (or as many as the resources of the reviewer permit). Second, the search must be *transparent*. Third, the search must be *reproducible*.

Reviewers must take great care when following the steps of the systematic review in order to meet quality requirements. The choice of one or more adequate search systems that allow the user to meet these quality requirements is one major consideration because these systems determine the number of relevant articles that will be identified. If a system has limitations of some sort, even the most skillful searcher might find them difficult or impossible to circumvent. Because the search systems differ in technical characteristics and scope, their suitability for systematic searches and thus for systematic reviews varies; however, reviewers are often unaware of the technical characteristics and limitations of search systems. Reviewers consulting methods-guidance on search systems will find only advice pertaining to certain systems; advice which is often not based on a systematic review of search functionalities. Accordingly, review efforts can still benefit significantly from guidance on which search systems are most suitable for a specific search task.

The first goal of a systematic review is to identify all or as many as possible relevant resources. Hence, the reviewer needs to select a search system that provides the best *coverage* of the chosen search topic. Coverage of a search system is denoted relative to a specific criterion—for example, a specific subject (eg, medicine and physics), resource type (eg, articles and books), time span (eg, retrospective coverage), or geographic location. A search system might provide high coverage of articles in medicine, yet low coverage of articles in physics. Greater overall size of a search system in this sense does not necessarily denote greater coverage on a specific topic. For example, while the multidisciplinary search system JSTOR has more than 12 million records and is considerably larger than IEEE Xplore with four million records, the coverage of IEEE Xplore on the specific topic of engineering records is still broader and thus more appropriate for evidence synthesis in engineering. Accordingly, systematic review guidance advises the use of suitable specialized databases that provide high coverage of a specific topic as well as generic resources that have broad coverage. Reviewers should thus consider their specific review topic when deciding which search systems might prove suitable for a systematic search. To assist with selection, there is considerable research on the coverage of search systems,[17-19] especially with regard to search systems such as Google Scholar which have built up an aura of secrecy around the size of their databases.[20-22]

While coverage is an important criterion based on the specific requirements of the systematic review, it does not indicate how well a reviewer can in fact access these resources on the chosen search system. A search system must allow a reviewer to specify queries that search with high *recall* and *precision*. While recall (or sensitivity) is the percentage of relevant article records that are returned in the result set from all relevant records known to exist, precision (or specificity) is the percentage of records in the result set that are relevant.[23,24] While high recall indicates a search contains many of the relevant items, high precision indicates a search retrieves relatively few irrelevant records.[5] Generally, the more recall is improved, the greater the reduction in precision, and the more precision is improved, the worse the effect on recall.[23] "Although precision and recall are typically at odds, there's one way to overcome the constraints of this trade-off: more features."(23, p80) More features in this context means that recall and precision can both be improved if the search query of a reviewer can be refined so it more accurately includes relevant records, while excluding irrelevant records. The ability of a reviewer to manipulate their search query depends on the capabilities of a search system and thus they significantly influence recall and precision.[25] The capability of search systems to retrieve results in an effective and efficient manner determines its suitability in systematic searches. A suitable search system provides good coverage in the specific area of interest and allows the user to specify a query with high precision and recall. Accordingly, search systems "should be evaluated against the background of what is found for—and what remains hidden from—the users."(26, p1570)

While the goal is to identify all records on a given topic, in practice, this goal has to be pursued within *resource limits*. Reviewers thus must make reasonable judgements on where to best invest time and funds based on a cost/benefit analysis: "Searches for systematic reviews aim to be as extensive as possible in order to ensure that as many as possible of the necessary and relevant studies are included in the review. It is, however, necessary to strike a balance between striving for comprehensiveness and maintaining relevance when developing a search strategy. [...] The decision as to how much to invest in the search process depends on the question a review addresses and the resources that are available."(15, p26) As a consequence, reviewers must select search systems that allow them to make the best use of their resources, that is, to retrieve the most relevant records in exchange for the least amount of time or funds.

The second and third goal of systematic reviews, *reproducibility* (also "replicability," "reliability," and "repeatability") and *transparency*, require an explicit, transparent, and documented search process that allows reviewers to update or replicate a given synthesis search. "The search process needs to be documented in enough detail throughout the process to ensure that it can be reported correctly in the review, to the extent that all the searches of all the databases are reproducible."(15, p41) Conduct and reporting guidance explicitly describe the steps necessary for a reviewer to ensure the rigorous and transparent documentation necessary to foster reproducibility. However, the functionality and capabilities of specific search systems can also influence reproducibility themselves. The reproducibility of a search determines whether that search can be replicated employing the same methods with a given search system. If the same query leads to the same search results, the search is considered reproducible. A lack of reproducibility can indicate a form of sampling bias or so-called search engine bias[27,28]: Repeated tests of the same query can lead to different results.(29, p37) It is important to note that because the size of the database provided by a search system typically increases over time, repeated queries will naturally yield a larger set of results than the initial query, and this has to be taken into account in assessing search system reproducibility and should not be seen as a lack of reproducibility. Researchers must thus pay close attention to reporting the date on which searches took place to make changes in the database of the search system comprehensible.

In summary, it is important to consider how far a search system supports the user in articulating and framing a query in a systematic search context, with special attention to high levels of coverage, recall, precision, and reproducibility.

The objective of the current research is to test and describe the usability and functionality of search systems that are frequently used in evidence syntheses, focusing on limitations that may influence the quality of systematic reviews. While some of these limitations impede the rigor required for systematic reviews as laid out by methodological guidance (necessary condition), others make systematic reviews more challenging or resource intensive (desired condition). Failing at some necessary condition does not mean the search system should be avoided entirely in the systematic review process, but does mean it should perhaps not be used for query-based searching: the fundamental underlying search method for systematic reviews. Nevertheless, such systems might be used in supplementary search methods.

Previous studies examining the suitability of search systems for evidence synthesis have focused on a limited number of search systems and/or based their analysis on a review of the search interface, yet without any in-depth examination of core functionalities that allow reliable query-based searching.[19,30-32] Previous studies have also calculated precision and recall of search systems from data reported by specific evidence-synthesis

TABLE 2 Description of tests performed to evaluate academic search systems, based on systematic search requirements regarding coverage, recall, precision, efficiency, and reproducibility

Number	Tested Quality	Test Scope	Test Criterion	Test Procedure	Performance Threshold of Test Criterion	Necessity of Meeting Threshold
1	C, R, P, E	DB	Subject coverage	Review of content description: type of academic disciplines covered	Subject coverage: maximum	Desired
2	C, R, P, E	DB	Size	Review of information provided by the official website of each search system concerning number of records provided on a database. Estimations of sizes based on query method by Gusenbauer, 2019[37].	Size: maximum	Desired
3	C, R, P, E	DB	Record type	Review of types of scholarly resources offered.	Types of records: maximum	Desired
4	C, R, P, E	DB	Retrospective coverage	Limitation of time span to oldest years available.	Time span: maximum	Desired
5	C, R, P, E	DB	Open access content	Review of search system and description of content provided: type of usage rights attributed to resources.	Content: open	Desired
6	R, P, E	Q	Controlled vocabulary?	Review of search options. Is a controlled vocabulary available and what is its coverage and accessibility? 1. Available: yes/no 2. Retrospective coverage: years available 3. Hierarchically structured: yes/no 4. Searchable: yes/no	Available: yes	Desired
7	R, P, E	Q	Field code search: query refinement	Listing of all field codes and limiters visible in the search interface.	≥5 field codes	Necessary
8	R, P, E	Q	Full text search option available?	Review of search options (field codes).	Available: yes	Desired
9	R, P, E	Q	Search string: 1. maximum length of word combination 2. maximum number of characters	1. maximum search string length (trial and error) 2. maximum number of characters (web search)	≥25 terms	Necessary
10	R, P, E	Q	Server response (time and number of records) for max. word combination	Test of the longest search string the system still could handle and review of search results in terms of whether longer search strings produced more results. 1. test via maximum word combination from test 9 2. test of next shorter word combination	Longer search string = more hits Timeout: no	Necessary

(Continues)

TABLE 2 (Continued)

Number	Tested Quality	Test Scope	Test Criterion	Test Procedure	Performance Threshold of Test Criterion	Necessity of Meeting Threshold
11	R, P, E	Q	Search string language support: English, Chinese, Cyrillic	1. English tested with test number 9 2. Chinese tested with 的 (means "bright") 3. Cyrillic tested with и (single letter)	Support: full (Chinese and Cyrillic)	Desired
12	R, P, E	Q	Boolean "OR" functional?	Test of six terms (research, define, paper, Asterix, table, and analysis) through step-wise addition with "OR" operators → does record count increase?	Functional: yes	Necessary
13	R, P, E	Q	Boolean "AND" functional?	Test of six terms (research, define, paper, Asterix, table, and analysis) through step-wise addition with "AND" operators → does record count decrease?	Functional: yes	Necessary
14	R, P, E	Q	Boolean "NOT" functional?	Test of six terms (research, define, paper, Asterix, table, and analysis) through step-wise addition with "NOT" operators → does record count decrease?	Functional: yes	Necessary
15	R, P, E	Q	Comparative test: "AND"/"NOT"comparative test: "OR"/"NOT"	*Comparative test* of AND/NOT-operators: Does "research" minus "research AND define" equal "research NOT define"? *Comparative test* of OR/NOT-operators: Does "define" plus "research NOT define" equal "research OR define"?	Functional: yes	Necessary
16	R, P, E	Q	Literal vs. expanded queries	Comparison of various similar (ill-written) terms: 1. Defin, "Defin" 2. Definx, "Definx" 3. Define, "Define" Do databases interpret queries literally, ie, retrieve only few hits for ill-written terms? Do quotation marks (usually used as limiters to search for verbatim) reduce the number of results, indicating that queries without quotation marks were automatically expanded? Comparison of British and American English: 4. Organise, "Organize" 5. Colour, "Color" Does variation in British/American spelling lead to differences in search results. If not, queries are expanded automatically to include both British and American versions.	Literal queries: yes	Desired

(Continues)

TABLE 2 (Continued)

Number	Tested Quality	Test Scope	Test Criterion	Test Procedure	Performance Threshold of Test Criterion	Necessity of Meeting Threshold
17	R, P, E	Q	Truncation/wildcards available?	Use of most common symbols: 1. "*" for right-hand truncation 2. "?" for single character placeholder 3. "$" for zero- or one-character placeholder	Functional: yes	Desired
18	R, P, E	Q	Exact phrase search functional?	Exact phrase search was tested using "": 1. Organise team, "Organise team" Does the use of quotation marks change search results?	Functional: yes	Necessary
19	R, P, E	Q	Parenthesis functional?	Scope tested with different positions of parentheses: 1. (Research OR define) AND Asterix 2. Research OR (define AND Asterix)	Functional: yes	Necessary
20	R, P, E	F	Filtering: Post-query refinement	Listing of all filters that allow refinement of a query results set visible in the search interface.	Number of filters: maximum	Desired
21	R, P, E	CS	Forward citation search available?	Review of search options.	Available: yes	Desired
22	E	Q	Advanced search string input field	Review of search options.	Available: yes	Desired
23	E	Q, F, CS	Search help?	Review of search options.	Available: yes	Desired
24	R, E	Q, F, CS	Maximum number of accessible hits	Last accessible results page of a large results set.	≥1000 results	Necessary
25	E	Q, F, CS	Bulk download supported?	For how many records can citation information be downloaded at once (involving a single selection and download request) to be exported to some reference management software or other destination?	Supported: maximum	Desired
26	RP	Q	Reproducibility of search results at different times	Repeated query after few seconds.	Search result: same	Necessary
27	RP	Q	Reproducibility of search results at different locations	Repeated query after few seconds with different, foreign IP address, or different institutional access.	Search result: same	Necessary

Abbreviations: C, coverage; CS, citation search; DB, database; E, efficiency; F, filter; P, precision; Q, query; R, recall; RP, reproducibility.

studies.[24,32-34] This analysis focuses instead on evaluation criteria for systematic reviews across disciplines, following universally accepted conduct guidance (ie, Cochrane, Campbell, and CEE). This research thus fills a need for support in the choice of search system, currently lacking in evidence-synthesis methodology, and follows calls for comparative studies on the effectiveness of search systems[25] or the "need to develop 'bias profiles' for search engines."(35, p1193) This study provides a much-needed overview of academic search systems from a user perspective. It compares a large selection of popular academic search systems and examines their unique characteristics to draw conclusions on their suitability for evidence synthesis. Specifically, the study tests whether a system allows the user to precisely specify a search so it retrieves as many relevant results as possible, how efficiently search results can be retrieved, and if the search results could be reproduced with the same methods. Hence, our study contributes to evidence synthesis as "[...], the value of a systematic review depends on what was done, what was found, and the clarity of reporting."(36, p1) Overall, the question framing this research is: How suitable and usable are commonly-used academic search systems for systematic searches in evidence synthesis? Definitions of the terms used throughout this study can be found in Appendix I, the detailed tests in Appendix II (supplementary online material).

2 | METHOD AND ANALYSIS

This study measures the suitability of a number of popular search systems for evidence synthesis using specific criteria. These criteria were assessed based on the 27 tests outlined in Table 2.

2.1 | Selection of search systems

The search systems analyzed in this study represent common resources in highly cited systematic reviews and meta-analyses in recent years. According to a search of Web of Science, for systematic reviews and meta-analyses across all databases available to us,* there were 63 "hot papers" that were "published in the past two years and received enough citations in September/October 2018 to place it in the top 0.1% of papers in its academic field." All search systems and databases that were mentioned in at least two of these 63 studies were included in our analysis. The result was a list of 16 databases and search systems: CINAHL, ClinicalTrials.gov, Cochrane Library, EbscoHost, Embase, ERIC, Google Scholar, LILACS, ProQuest, PsycINFO, PubMed, ScienceDirect, Scopus, SportDiscus, TRID, and Web of Science. While most of the search systems mentioned were databases, some authors mentioned platforms without stating the exact databases searched (eg, Web of Science is a platform, while Web of Science Core Collections is its main database)—a common reporting error of search scope.

In addition, to obtain a broader picture of the qualities of academic search systems, we also included other search systems that are regularly used among academic researchers across disciplines[38]: AMiner, ACM, arXiv, Bielefeld Academic Search Engine (BASE), CiteSeerX, Digital Bibliography & Library Project (DBLP), Directory of Open Access Journals (DOAJ), IEEE Xplore Digital Library, JSTOR, Microsoft Academic, Semantic Scholar, SpringerLink, Wiley Online Library, WorldCat, and WorldWideScience. Thus, we examine the quality of a total of 28 search systems that access 34 databases either via web search engines (eg, Google Scholar or Microsoft Academic), via platforms that allow access to one or more discrete databases (eg, ProQuest or OVID) or other bibliographic databases (eg, Transport Research International Documentation). Below, we present an overview of the 28 search systems; if the database is accessed via a platform, the database's name is given in parentheses as follows:

1. ACM Digital Library	11. Education Resources Information Center	21. Semantic Scholar
2. AMiner	12. Google Scholar	22. SpringerLink
3. arXiv	13. IEEE Xplore Digital Library	23. Transport Research International Documentation
4. Bielefeld Academic Search Engine	14. JSTOR	24. Virtual Health Library *(LILACS)*
5. CiteSeerX	15. Microsoft Academic	25. Web of Science *(Medline, Web of Science Core Collection)*
6. ClinicalTrials.gov	16. OVID *(Embase/Embase Classic, PsycINFO)*	26. Wiley Online Library

(Continues)

*Web of Science Core Collection, BIOSIS Citation Index, BIOSIS Previews, Data Citation Index, Derwent Innovations Index, KCI-Korean Journal Database, MEDLINE, Russian Science Citation Index, SciELO Citation Index, and Zoological Record (accessed on February 11, 2019; search string: ti("systematic review" OR "meta-analysis")

7. Cochrane Library *(CENTRAL)*	17. ProQuest *(ABI/Inform Global, Nursing & Allied Health Database, Public Health Database)*	27. WorldCat–Thesis/Dissertations
8. Digital Bibliography & Library Project	18. PubMed *(Medline)*	28. World WideScience
9. Directory of Open Access Journals	19. ScienceDirect	
10. EbscoHost *(CINAHL Plus, EconLit, ERIC, Medline, SportDiscus)*	20. Scopus	

We selected a large set of popular specialized and multidisciplinary search systems that are relevant not only for disciplines where evidence synthesis is already well-established (eg, medicine, health sciences, or environmental studies) but also other disciplines, such as management, where these methods have been increasingly used just in the last years. To include all search systems or databases in this study would be an impossible task, as hundreds of bibliographic databases and search systems exist across subjects.

Our sample of search systems covers a range of types of technology (eg, platforms and web search engines), target audience (eg, academic discipline and resource restriction), and provided content (eg, traditional academic literature and grey literature). Some platforms examined provide access to multiple databases at once, allowing us to assess the basic qualities and functionalities on a system level as well as a database level. While we find that most of the functionalities are determined by the system itself, other qualities might be closely linked to the underlying database—such as the number and type of field codes or the availability of a controlled vocabulary. We included proprietary, nonproprietary, and Open Access databases, a distinction especially relevant for reviewers who have only limited access to expensive database subscriptions. While the focus was on bibliographic databases, we also included sources that include grey literature (eg, arXiv, Google Scholar, and WorldCat-Thesis/Dissertations). Grey literature refers to any document produced by an organization at any level whose primary purpose is not commercial publishing and includes theses, white papers, organizational reports, and consultancy documents.[39] By searching for grey literature, systematic reviews aim to maximize comprehensiveness and mitigate publication bias.[16] Typically, systematic reviews will conduct dedicated searches for grey literature (for example, searching organizational websites), but the ability to include grey literature in formalised, systematic searches of bibliographic databases can provide benefits, including the ability to assess eligibility concurrently with bibliographic search results, potentially increasing efficiency.

2.2 | Evaluation approach: Different search systems—One overarching method

It is important to note that the search systems we analyze in our sample are diverse in their functionality, syntax, and features. All of these systems have different underlying databases and indexing methods, data presentation, and curation methods. Crawler-based web search engines (eg, Google Scholar), for example, function differently from bibliographic databases which have a curated catalogue of information (eg, Scopus). Some of these search systems are large and multidisciplinary (eg, Scopus), while others have a narrower focus on a single or a few domains of research (eg, PsycINFO) (see Appendix II). We examine these diverse search systems through the lens of the users that access them and test how well the search facility performs to link the query to the underlying database. We do not test the searchers' ability to formulate such strings/queries[40-44] and we do not test the completeness of the underlying dataset provided by the search system.[24,45] This study instead examines the search system as the gatekeeper that mediates between a database of potentially relevant records and a reviewer that wishes to access, retrieve, analyze, and synthesize that information in a systematic, rigorous manner.

Hence, we assessed the functionality of these databases with standard queries from the perspective of the user (see Table 2). In querying diverse databases with a diverse set of inputs, we tested the capacity of the databases to interpret the user's query so that the dataset is retrieved effectively. We examined the results of the search systems both quantitatively (eg, how many hits a query retrieved or how much time the server needed to respond) and qualitatively (eg, the nature of the search options and the search interface). The quantitative methods based on tested methods that use variations of search queries to iteratively determine sizes of different types of search systems.[37] In our analysis, we did not assess the quality of the retrieved records, in terms of their fit with a given search intent for example. Quality

TABLE 3 Review of search methods used in systematic reviews

Number	Study	Search System with Largest Result Set	Use of Field Codes	Length of Longest Search String (Boolean Operators, Field Codes Not Counted)	Use of Boolean Operators (AND, OR, NOT)	Use of ""	Use of ()	Use of Truncation, Wildcards	Total Number of Studies Screened (Incl. Duplicates)	Studies Identified with a Single Search String	Non-query Identification (Total)	Size of Final Sample	Search Precision
1	Aune et al[58]	PubMed, Embase	N/A	68 terms	OR, AND	Yes	Yes	No	49 772	40 744	4 (handsearch, all relevant)	142	0.28%
2	Barnett et al. (2017)[59]	Scopus	N/A	N/A	N/A	N/A	N/A	N/A	19 005	5681	3 (handsearch, all relevant)	100	0.53%
3	Baur et al[60]	PubMed	Yes	57 terms	OR, AND	Yes	Yes	No	1169	1113 (sum of all databases)	56 (handsearch, not all relevant)	76	6.50%
4	Bediou et al[61]	N/A	N/A	11 terms	OR, AND	Yes	Yes	No	958 147	N/A	0	82	0.01%
5	Bethel et al[62]	PubMed	N/A	16 terms	OR, AND	Yes	N/A	No	12	N/A	N/A (post-query filtering)	4	33.33%
6	Bourne et al. (2017)[63]	Embase (via OVID)	Yes	12 terms (sets were later combined)	Or, And,not	Yes	Yes	Yes	3878	2539	N/A (post-query filtering)	288	7.42%
7	Brunoni et al[64]	PsycInfo (via OVID)	Yes	10 terms	OR, AND	Yes	Yes	Yes	1121	N/A	N/A	81	7.23%
8	Carlbring et al[65]	PubMed	N/A	29 terms	OR, AND	Yes	Yes	No	2078	2078	N/A (handsearch, post-query filtering)	20	0.96%
9	Chu et al[66]	Medline, Healthstar (via OVID)	Yes	25 terms (sets were later combined)	OR, AND	Yes	Yes	Yes	1784	N/A	N/A	26	1.46%
10	Cipriani et al[67]	N/A	N/A	10+ terms (combined with undisclosed keyword list)	OR, AND	Yes	Yes	Yes	28 552	24 200 (sum of all databases)	4352 (handsearch, other sources)	421	1.47%
	Recommended threshold?			≥25 terms	OR, AND, NOT	Yes	Yes	Yes			Post-query		

(Continues)

TABLE 3 (Continued)

Number	Study	Search System with Largest Result Set	Use of Field Codes	Length of Longest Search String (Boolean Operators, Field Codes Not Counted)	Use of Boolean Operators (AND, OR, NOT)	Use of ""	Use of ()	Use of Truncation, Wildcards	Total Number of Studies Screened (Incl. Duplicates)	Studies Identified with a Single Search String	Non-query Identification (Total)	Size of Final Sample	Search Precision
			≥5 field codes							≥1000 accessible records	filtering: m aximum Citation search: yes		
	Necessity?		N	N	N	N	N	D		N	D		

Abbreviations: N, necessary; D, desired.

criteria have previously been used to evaluate search systems in terms of recall and precision indicating their suitability of search in general and systematic search in particular.[25,33,46-57] If available, we searched with the advanced search interface of the search system. Tests were performed between February and March 2019.

2.3 | Necessity to meet requirements

Our evaluation of search systems involves applying 27 unique criteria each of which tests the performance of a specific quality. In our evaluation of the search systems, we differentiate between capabilities that are *necessary* or merely *desired* for a systematic review. In order to meet the requirements of the guidance of Cochrane, The Campbell Collaboration, and CEE for systematic reviews, a necessary criterion needs to be fulfilled by a search system, irrespective of the context of the study. A desired criterion is necessary to be met only for systematic reviews with specific requirements or foci. Further, a desired criterion that is not met, can, if the reviewer is aware, be circumvented with extra effort of using suboptimal search methods. Reviewers should decide whether the fulfilment of a desired criterion is important for their specific search. Necessary criteria, however, should always be met by search systems. Each criterion was classified as either desired or necessary according to evidence synthesis guidance (see Table 2).

In order to support our decision to determine meaningful performance thresholds, we reviewed the search methods of 10 random articles from the sample of 63 - articles (see Table 3). We reasoned search systems should at least come close to enabling the searches described in these studies. We extracted information concerning the search methods these studies used to obtain their search results. The single thresholds are explained in detail in the description of each single criterion used in our test.

2.4 | Requirements translated to evaluation criteria of search systems

The requirements for systematic reviews are largely agreed upon in evidence-synthesis guidance: (a) identify a maximum number of relevant records for a specific topic, (b) within the resource limits, and (c) use transparent and reproducible search methods. These requirements focus on the search process of the reviewer and the type of methods that are employed; yet, evidence-synthesis guidance provides no clear technical requirements for search systems. Hence, for our analysis, we

translated these requirements to technical criteria that search systems needed in order to meet the requirements of evidence synthesis. A search system thus needs to be (a) *effective* in finding most of the relevant results while filtering out the irrelevant, (b) *efficient* allowing the reviewer fast identification and retrieval of records, and (c) must allow the *reproduction* of search results with the same methods:

First, effective search depends in the reviewer's choice of a suitable search system offering the best coverage of records searched for. Coverage can be determined in multiple ways: for example, concerning time frame (retrospective coverage), academic subject (discipline) or usage rights (Open Access). Additionally, *effectiveness* is determined by the search system's capability to translate the search frame determined by the reviewer to enable precise searching with a high level of recall. To provide a thorough search for relevant records, the reviewer can combine different methods of (a) queries with keywords and a controlled vocabulary, (b) post-query filtering, and (c) handsearching of relevant journals, issues, or reference lists (citation search). While all of these methods provide value for a rigorous search that aims to identify all or at least most relevant records, it is particularly *queries* using keywords or a controlled vocabulary that are able to search the corners of a database that would be inaccessible with citation searching, handsearching, or post-query filtering alone. The Cochrane Review of Stacey et al[68] is an example of a study that uses an elaborate query-based search strategy relying on "AND," "NOT," "OR" operators, database-specific field codes and controlled vocabularies. Searching five databases from different providers, they retrieved a total of 46 054 hits from which they included 105 studies in their final meta-analysis. A perfect query would, in theory, make citation searching and handsearching obsolete, yet perfect citation searching and handsearching could not do the same to make the use of queries obsolete. This logic is supported by our review (see Table 3) where most identified results were derived from query-based searches—some studies based their search strategies on queries alone. Accordingly, using *queries* for systematic search is necessary for systematic reviews as evidenced in both research practice and in evidence-synthesis guidance. The other search methods are supplementary techniques desired to improve the search result of the query. Second, the efficiency with which a reviewer can retrieve relevant results is largely determined by recall and precision. Therefore, the choice of a suitable search system with suitable coverage and capabilities of searching with functional search strings, filters, and citation search impacts tremendously on effectiveness, that is, precision and recall. Nevertheless, other functionalities associated with downloading search results or user-friendly data-input also influence search efficiency (along with the subsequent stage in a systematic review, eligibility assessment). Third, *reproducibility* can be determined how well the system is capable to retrieve same results again with same search methods.

2.5 | Test procedures and performance requirements

We reviewed and tested the 28 search systems with 27 criteria determining each search system's (a) coverage and (b) capability to perform systematic searches via queries, filters, and handsearching so that a reviewer can obtain *reproducible* results, *efficiently*, and with high recall and precision.

2.5.1 | Coverage

Generally, it is assumed that more coverage is better than less coverage, as without a comprehensive database, searches would identify few relevant records. This means higher coverage typically increases the recall of a query. Nevertheless, it is important to note that while recall increases, precision simultaneously decreases, disproving the statement that more coverage is *always* better. What records are considered relevant depends on the specific requirements of the reviewer and thus cannot be generalised. For example, a search system with a smaller size, covering only a single discipline, might bring more relevant search results than a large search system covering multiple disciplines. Accordingly, all performance requirements on coverage were framed as *desired* criteria as reviewers must decide for themselves what search system best fits their unique study requirements.

Criterion 1: "Subject coverage" assesses the type of academic disciplines that are predominantly covered by a search system. This criterion determines whether a search system specializes in single disciplines or is multidisciplinary. While a greater coverage of disciplines might generally be regarded as beneficial, the greater breadth of records available might harm search precision. When working with such multidisciplinary search systems, the reviewer needs to be more specific about search context to receive the same precision than when using a specialized search system.

Criterion 2: "Size" informs about the absolute number of records available on a database that is made available through a search system. Searching larger databases, all things being equal, results in higher search recall. We assessed sizes by reviewing the official information

provided by the search systems' websites. If this information was up-to-date, we reported the official number; if it was either outdated or unavailable, we used the method suggested by Gusenbauer[37] to assess search system size.

Criterion 3: "Record type" informs about the types of records offered by a search system. Here we relied on the information provided by the search systems. Naturally, each search system had its own definition of how to categorize and classify records, which made direct comparison of record types difficult. Nevertheless, the availability of more—as opposed to fewer—document types meant that reviewers could search with increased precision, if this field code was available to specify a search.

Criterion 4: "Retrospective coverage" informs about what year the oldest records on a database are from. When information on retrospective coverage was provided by the search system, we included this information, if not, we manually searched for the oldest records on the database and reported this year. In doing so, we took care not to consider incorrectly dated records in our assessment of retrospective coverage.

Criterion 5: "Open Access" was assessed in reviewing the usage rights of the records offered. If a search system offered mostly proprietary content with only a marginal focus on Open Access resources, then it was considered "proprietary." If the search system was nonprofit and/or provided strong emphasis to support Open Access content—but was also linked to proprietary resources—it was considered "mixed." If the search system offered only Open Access content, it was considered "open."

2.5.2 | Search

All of our query tests (criteria 6-17) have in common that the results determine whether a given search system allows the user to specify a systematic search query that is targeted at high precision or recall. All these query-features are helpful to compile a comprehensive search string to specify exactly what lies inside and outside the search scope of the evidence synthesis.

Criterion 6: "Controlled Vocabulary" was assessed by reviewing the search options provided by the search system for a given database. Where databases are accessed via broader platforms (eg, ProQuest), the options available often differed across databases, such that a single platform may have different search functionalities for its databases. The controlled vocabulary is more useful for some disciplines, while it is less useful for others. For example, "databases in the social sciences tend not to be as thoroughly indexed as those in medicine and may use methodological indexing inconsistently, if at all."(15, p33) Hence, guidance by Campbell Collaboration, for example, advises for cautious use of controlled vocabularies: *"When searching for studies for a systematic review, [...] the extent to which subject terms are applied to references should be viewed with caution. Authors may not describe their methods or objectives well and indexers are not always experts in the subject areas or methodological aspects of the articles that they are indexing."*(15, p28) Further, because retrospective coverage of the controlled vocabulary may be limited, reviewers should take care if relying on this method, especially for searches over longer time periods including earlier studies. It is difficult to quantify the quality of such controlled vocabulary as their features are diverse. Accordingly, we provide additional information on coverage, the option of a searchable index, and the availability of a hierarchical structure. We leave it to reviewers to decide whether such information is helpful for their specific search tasks. In summary, we regard the availability of a controlled vocabulary a desired condition as a thorough query-based search can compensate for some of the advantages of a controlled vocabulary.

Criterion 7: "Field Code Search: Query Refinement" is important for systematic searches to provide the user with options to search with high precision and recall. We reviewed the search options to assess which field codes were provided by search systems allowing reviewers to detail which parts of documents should be searched. The availability of field codes is a necessary criterion, since the user must be able to specify exactly where the requested information is located within the records. We defined *five* field codes as the lower threshold. With the exception of the option of a full text search (criterion 8), we do not test for the availability of single field codes, as we assume that in general, the more field codes are available the better the chances of a reviewer being able to search with high recall and precision.

Criterion 8: "Full Text Search". In some cases, reviewers need to search the full texts to identify specific study types. As full text search is however not necessary in every systematic search, we considered this criterion desirable.

Criterion 9: "Maximum Search String Length" was an essential determinant of how long and thus how specific the search string can be. We determined the maximum length of search strings that still retrieved results but did not result in timeouts or other system failures through trial and error of search strings varying in length. For this purpose, we used the 2008 Oxford Word List[69] and interlinked strings of the top 1000, top 500, top 100, top 50, top 25, top 10, or fewer most utilized English words with Boolean "OR"-operators. It happened that some Boolean queries were aborted due to timeouts or search string length limitations. Here, we iteratively searched

for the largest query the search system could still handle. We considered a Boolean search needs to support at least *25 terms*, so the user is able to specify the minimum necessary outlines of a search, i.e. to adapt the search string to the linguistic particularities of the scientific (sub-) topics of interest using keywords. Our quick review of systematic searches (Table 3) revealed that while appropriate search strings can be significantly longer, 25 terms should allow reviewers to specify their search scope to a reasonable extent. Search string length is critical for searching with high recall and precision and thus a necessary criterion. If reviewers needed a larger search string than supported by the search system, it might be possible to circumvent this limitation by splitting search strings. This practice can, however, be extremely laborious and is prone to error.

Criterion 10: "Server Response (Time and Number of Records)" was a test to determine the server's response time for the longest search string still supported by the system and was conducted together with the test for criterion 9. Additionally, we tried the next shorter search string combination to determine whether longer search strings actually resulted in longer loading times. Further, in order to pass this test, the system needed to produce more results for broader searches than for narrower ones. As the technical performance of the search system and the correct interpretation of long search strings are critical for information retrieval, we deemed this test necessary.

Criterion 11: "Search String Language Support" tested the search systems' capacity to interpret English, Chinese, and Cyrillic characters using frequently used terms and characters to determine the response of the search system. If the system retrieved results, it was deemed to support such characters, if the result was an error message or zero results, we deemed the system was not working. This was considered a desired criterion, as reviewers typically search with English characters only and because a number of databases index non-English records using translated English language titles and abstracts, thus being identifiable also with English language.

Criteria 12 to 15: "Boolean Functionality" were some of the most important tests in our study and tested the search systems' capability to effectively interpret the most common Boolean operators OR, AND, and NOT.[55] The system must retrieve results as anticipated by the Boolean logic. Boolean operators (AND, OR, NOT) are an integral part of systematic searches,[26] allowing the user to precisely specify the scope of the query as no other technique could. While AND and OR are used to link single concepts to a common search string, NOT is mostly used for disambiguation. Because evidence synthesis has very specific information needs determined by choices of constructs, methods, and research questions, it needs complex search strings to determine queries that link concepts of interest. Boolean operators have been shown to be essential for sampling in evidence synthesis[70] as they "allow the searcher to use set theory to help define the items that will be retrieved by a search."(5, p103) They provide "a great range of strategies are available to increase 'recall' and 'precision.'"(26, p1570) We used a combination of a total of six terms—research, define, paper, Asterix, table, and analysis—to see whether the result set increased, decreased, or remained constant after adding one more term. For strings with OR operators the results set should increase or at least remain constant with each additional term, with AND operators, it should decrease or remain constant, and with NOT, it should decrease or remain constant. We deliberately chose words from the research context and one—the word "Asterix"—that is very unlikely to appear frequently in scholarly articles to test the systems' responses. Adding the nonscholarly term Asterix to a Boolean OR string would not add many additional hits on a scholarly database yet would reduce the set of an AND string to almost zero or decrease it only slightly in a NOT combination. We also tested how many results the term Asterix would retrieve when searched as a single term to cross-check these results against the changes in the different Boolean strings. Further, we used alternative notations for Boolean operators if they were explicitly stated in the help or FAQ files of the individual search systems. We confirmed that "how websites are represented and the precise commands used to do the Boolean syntax search will differ somewhat for each search engine."(5, p103) It was not uncommon, for example, that "AND NOT" would be used instead of NOT or blank would be interpreted as AND. If we could not find any information on whether and how Boolean operators were supported by the search system, we utilized the most frequently used syntax - OR, AND, and NOT - to test these systems. In addition, to test the Boolean operators individually, we added two comparative tests (criterion 15) and evaluated whether queries with different Boolean operators retrieved a valid number of hits. Specifically, we tested first whether the number of hits for "research" minus the number of hits for "research AND define" equalled the number of hits for "research NOT define." Second, we tested whether the number of hits for "research OR define" minus the number of hits for "define" equalled the number of hits for research NOT define. If both tests were passed, we considered Boolean operators functional. Accordingly, the functioning of all three Boolean operators were considered necessary for systematic reviews.

Criterion 16: "Literal vs Expanded Queries" determined whether a search system automatically expands

queries impacting on precision and recall. We tested different correct and incorrect versions of the word define to check whether the search system would use autocorrect or would expand the query to different word forms. Additionally, we compared the number of results for terms with different spellings for British and American English. If the number of records for different spellings was the same, we assumed automatic query expansion. As knowledgeable reviewers are able to circumvent automatic query expansion via the use of quotation marks, this criterion was considered desired.

Criterion 17: "Truncation/Wildcards" determined whether different frequently used truncation or wildcard symbols were functional. For this criterion, we tested whether terms with truncation or wildcards resulted in more search results than terms without the use of truncation and wildcards. If words with truncation and wildcards produced more hits, they were assumed to function. Similar to criterion 16, the knowledgeable reviewer might circumvent the absence of functional truncation or wildcards by incorporating diverse word forms manually into a search string. Hence, this criterion was considered desired.

Criterion 18: "Exact Phrase Search" determined whether the use of quotation marks—symbols typically used to deem an expression should be searched literally—would result in fewer results than for terms lacking them. This is an important feature that allows reviewers to specify exact meanings. In reviewing systematic searches (Table 3), we found that exact phrase searching was used by all systematic reviews. Hence, we deem this criterion necessary.

Criterion 19: "Parentheses" determined whether the parentheses functioned in compiling search strings. To create comprehensive search strings with high recall and precision, it is vital to rely on the use of parentheses as these symbols allow a user to group individual concepts and to link them logically. The quick review of systematic searches (Table 3) showed that all systematic reviews used parentheses in their searches. Accordingly, we consider functional parentheses a necessary criterion.

Criterion 20: "Filtering: Post-Query Refinement" determined a search system's capacity for post-query refinement through a so-called faceted search. We listed the different filters available to users to refine their search results sets to increase the precision of their search after a query was computed. The more powerful the post-query filter options are, the greater the potential precision of a given query. If search queries work flawlessly and offer options to comprehensively determine search scope, post-query filtering should not be necessary. Hence, we rated post-query refinement of search results as a desired criterion as it is helpful to further specify search scope.

Criterion 21: "Forward Citation Search" determined a search system's capacity for forward citation search, that is, the system listing records that cite a specific records of interest. The logic is that records that cite a relevant record will be relevant themselves. Through association, the set of relevant search results can be increased beyond what could have been found through query alone. We reviewed whether forward citation information was offered by a search system, yet did not check the quality of the forward citation search, as this depends on the capacity of the citation index. As the forward citation search is considered a supplementary search method to search queries, we consider it a desired criterion. We did not check search systems' capacity for handsearching in terms of backward citation search, or the search of specific issues or journals.

Criterion 22: "Advanced Search String Input Field" was assessed through a review of the search interface. An advanced search input field would allow users to more easily compose advanced search strings. As this limitation can be circumvented with the basic search interface offered by the search system, this criterion is only desired.

Criterion 23: "Search Help" was assessed through reviewing the search interface to determine whether the search system provided some documented form of search help to assist users in formulating their search strategies. Search help would primarily mean guidance on which search operators or field codes are available and how users can use the search interface effectively. We consider this criterion to be desired for systematic search.

Criterion 24: "Maximum Number of Accessible Hits" was assessed by determining the maximum number of hits made accessible by the search system with a single search. This test required navigating to the last search results page of a query with millions of results to determine how many results were accessible to the reviewer. While hit counts may sometimes run into the millions, it is only a relatively small fraction of this theoretical results set that is actually retrievable in practice. Hence, if a reviewer is interested in the full results set and the hit count goes beyond a set threshold, it is impossible or at least cumbersome to retrieve the full set. One workaround might be to fine-slice the query into smaller results sets that lie below this maximum that can then be handled sequentially. Nevertheless, while this procedure requires significant effort, it is far from certain if it is supported by the specific search system to compile such precise search strings with Boolean operators. We introduced the number of accessible hits as a necessary

criterion. If the criterion was above 1000 records, the performance of the search system was considered sufficient for systematic reviews, yet still not ideal. While searches of with more than 1000 results are common as indicated by our review of systematic searches (Table 3), we opted for a rather conservative threshold as reviewers might mitigate this limitation by dividing search strings to retrieve multiple result sets comprising fewer than 1000 records.

Criterion 25: "Bulk Download Supported" was assessed through reviewing the search interface and attempting to download large quantities of results at once. A major time constraint in retrieving search results for subsequent synthesis is search systems' requiring reviewers to download search results in small batches instead of offering full download capabilities. While search system providers want to protect their data from theft and therefore do not provide bulk downloads, this constitutes a tremendous time constraint for reviewers in evidence synthesis and limits their resources for other review activities. We consider this criterion to be desired for systematic search.

Criterion 26: "Reproducibility of Search Results at Different Times" tested whether these search queries show signs of bias.[5, p112,54, p141] These tests show whether queries could be repeated so that identical queries retrieve identical results sets, a criterion also described as "the extent to which the search engine returns, from our query, similar results under consistent conditions."[20, p945] Reproducibility is a quality central to systematic reviews that qualifies a search system capable of retrieving results independent of time and place. Accordingly, it is necessary for any search system to offer reproducible results across time in order to meet the quality guidance for systematic searches.

Criterion 27: "Reproducibility of Search Results at Different Locations" assessed whether changes in the place from which searches were performed influenced the search results. The place was changed in two forms: we used VPN services to simulate foreign IP addresses and repeated the same query and we logged onto the search system and repeated queries with different institutional access schemes. If these variations produced changes in the search results that could not be explained by the periodic database growth of search systems or by minor discrepancies in the database based on institutional subscription (eg, with Web of Science Core Collections made up of multiple indices that can differ across institutions), the services had to be considered biased. Similar to criterion 26, it is necessary for any search system to offer reproducible results across place in order to meet the quality guidance for systematic searches. A systematic search was considered reproducible when it passed both tests for criteria 26 and 27.

2.6 | Principal vs supplementary resources

In our evaluation, a search system could be either rated suitable as a *principal* or *supplementary* resource. A principal resource needed to meet all *necessary* quality requirements or was otherwise considered supplementary. Supplementary resources could be used *in addition to* a principal resource for its specific qualities that could retrieve additional records and to further improve the evidence base. Hence, if a system failed a test for one search method, it might be still considered a good choice for some other search type—for example, while Google Scholar is considered unsuitable for primary review searches, it is considered a suitable supplementary source of evidence (including on grey literature).[57] The distinction between principal and supplementary resources for systematic reviews was also used in previous assessments of search system qualities.[33,57]

Desired quality requirements were not taken into account in this rating of principal and supplementary resources. Instead, these criteria were included in our tests to inform reviewers about functionalities of search systems that were useful or important but still not entirely necessary to meet quality requirements of systematic reviews. The more reviewers are aware of the specific functionalities of a search system, the better they can optimize their search strategy. Accordingly, the tests performed in this study evaluated single search functionalities (see Table 2) so that reviewers received a granular view of how search systems perform for each test. This way, reviewers can quickly consult individual quality criteria for detailed evaluations and reflect on whether a search system of interest offers a service suitable for their needs.

3 | RESULTS

Our analysis assessed the suitability of 28 search systems for systematic reviews with 27 test criteria. Each of these criteria assessed a single functionality of the search system. Jointly, these tests showed to what degree a search system was capable of searching effectively and efficiently: qualities necessary for evidence synthesis in the form of systematic reviews. The systematic assessment with these performance tests showed substantial differences in functionality among search systems (see Table 4). While some search systems could be recommended almost without limitation, others failed important tests limiting their suitability for systematic reviews. In other words, not all search systems allow reviewers to perform queries, apply filters, or undertake

citation searching with the high standards required in systematic reviews. Our work makes it possible to classify search systems transparently and objectively according to their suitability for systematic evidence synthesis. We describe the results of these tests that answer questions of interest to a reviewer engaging in systematic search as follows:

1. What is the coverage of the search system, to ensure I access a database suitable for my review?
2. How effectively can I articulate my search via queries, filters, or citation searches so I can retrieve results with high recall and precision?
3. Can I reproduce my search, so that repeated queries will retrieve the same results?
4. How efficiently can I search the system, so I can perform the review within my resource limits?

3.1 | Coverage

Our coverage tests assessed five desired criteria a reviewer must consider when choosing a suitable search system. We found that there are significant differences in the databases across all tested criteria. Of the 34 databases offered by the 28 search systems, we found that 16 had a multidisciplinary focus while the remainder were specialized, that is, with a focus on medicine, health sciences, sports, computer science, education, economics, electrical engineering and electronics, psychology, business and management, biomedicine, and transportation studies. The sizes of databases indexed ranged from more than 300 000 to almost 400 million records. Similarly, retrospective coverage ranged between 1550 and 1999, while it is important to note that the number of publications dating back as far as 1550 was small. Further, the number of available record types varied between two and 81 different records. Our sample examined five Open Access resources ("open"), six search systems that focused on Open Access literature, while also providing proprietary resources ("mixed"), and 17 search systems that almost exclusively focused on proprietary content ("proprietary").

3.2 | Search queries

We tested the most frequent Boolean operators OR, AND, and NOT via incrementally extended strings adding up to a maximum of six terms and also tested whether Boolean search worked if used with parentheses. Additionally, we assessed exhaustive OR-combinations of different lengths to verify if longer strings also produced more hits. If Boolean operators would not work for short ones, we were not surprised if they also did not work for long ones. Overall, the tests revealed that 17 of the 28 search systems support Boolean operators flawlessly. The remaining search systems retrieved implausible results for search strings consisting of one or more types of Boolean operators. Particularly, AMiner, DBLP, Google Scholar, Microsoft Academic, and WorldWideScience failed all or all but one of the Boolean tests we performed. Further, we found that ERIC seemed to support Boolean searches with our tests of up to six keywords yet failed with longer Boolean searches. The maximum length of the search queries handled without timeouts varies considerably among search systems from only some seven terms (JSTOR) to more than 1000 (EbscoHost, OVID, PubMed, Scopus, Virtual Health Library, Web of Science, and WorldWideScience). For some search systems, we identified restrictions concerning maximum search string length measured in characters, as for example, Google Scholar only allows searches of up to 256 characters. However, for most search strings, the length is determined by the load it puts on the server and thus fails to deliver search results if the request results in a server timeout. The server load seems not only to be determined by search string length but is especially influenced by search scope determined by field codes (eg, title, abstract, or full text search) and the size of the underlying database (searching one or multiple databases via a platform provider).

Of the 28 search systems, 16 seemed to use a form of automatic query expansion to interpret what the user meant instead of processing search strings verbatim. Whenever the default setting is such that queries are expanded automatically, the user can mitigate the effect by adding explicit limiters (in most cases "") to search for keywords literally. On the contrary, explicit limiters ("") were not working or not working correctly in seven search systems. In the cases of AMiner, CiteSeerX, and WorldWideScience, this was especially problematic as these systems expand queries automatically and do not seem to support explicit limiters, forcing reviewers to search via automatically expanded queries. However, for most (21) of the tested search systems, the limiters functioned correctly. We found that while all systems supported the English language, some failed to support Chinese or Cyrillic characters. This may be due to the fact that most texts indexed on these databases were written in English and hence there is little necessity to support additional non-Latin characters. Our tests showed that no search system provided all forms of wildcards and truncation we tested for. Reviewers must exercise caution and test whether the specific truncation or wildcards they use work appropriately.

TABLE 4 Assessment of 28 academic search systems on their suitability for evidence synthesis

Name of Search System	Database(s) Searched; Search Settings	1) Subject	2) Size	3)Record Type (Selectable Separately)	4) Retrospective Coverage (Oldest Entries)	5) Open Access Content?	6) Controlled Vocabulary?	7) Field codes/ Limiters?	8) Full Text Search Option?	9) Search String Length	10) Server Resp. Time/ Records: Max. Word Comb.
		D	D	D	D	D	D	N ≥ 5	D	N ≥ 25	N
ACM Digital Library	Full index: Full-text collection	Computer science	2,000,000+	4	1947	Proprietary	√	17	√	100	√
AMiner	Full index	Multidisciplinary with a focus on computer science	233,127,915	2	Unknown	Mixed	X	3	X	10	X
arXiv	Full index; settings: All fields	Multidisciplinary	1,518,677	2	1991	Open	√	17	√	100	√
Bielefeld Academic Search Engine (BASE)	Full index	Multidisciplinary	144,252,584	10	Unknown	Mixed	√	12	√	100 ≤ 1024 characters	√
CiteSeerX	Full index	Multidisciplinary with focus a on computer and information science	8,401,126	3	Unknown	Open	X	9	X	500	X
ClinicalTrials.gov	Full index	Medicine	301,373	4	1999	Open	√	23	X	37	√
Cochrane Library	Cochrane Central Register of Controlled Trials (CENTRAL)	Medicine	1,317,434	5	1908	Proprietary	√	15	√	100	√
Digital Bibliography & Library Project (DBLP)	Full index	Computer science	4,539,285	9	1936	Open	X	6	X	10	X
Directory of Open Access Journals (DOAJ)	Full index	Multidisciplinary	3,902,698	2	1874	Open	√	13	X	100	√

TABLE 4 (Continued)

Name of Search System	Database(s) Searched; Search Settings	1) Subject	2) Size	3)Record Type (Selectable Separately)	4) Retrospective Coverage (Oldest Entries)	5) Open Access Content?	6) Controlled Vocabulary?	7) Field codes/ Limiters?	8) Full Text Search Option?	9) Search String Length	10) Server Resp. Time/ Records: Max. Word Comb.
		D	D	D	D	D	D	N ≥ 5	D	N ≥ 25	N
EbscoHost	Selection: ERIC; Medline; EconLit	ERIC: Education studies; Medline: Health studies; EconLit: Economics	ERIC: 1,730,508 Medline: 29,456,831 EconLit: 1,661,780	12	ERIC: 1907 Medline: 1946 EconLit: 1886	Proprietary	√	11	X	50	√
EbscoHost	Selection: CINAHL Plus	Health studies	6,304,949	7	1937	Proprietary	√	48	X	500	√
EbscoHost	Selection: SPORTDiscus	Sports studies	2,449,690	6	1800	Proprietary	√	27	X	1,000	√
Education Resources Information Center (ERIC)	Full index	Education studies	1,600,000+	24	1907	Proprietary	√	19	√	100	X
Google Scholar	Full index	Multidisciplinary	389,000,000 +	3	1700	Mixed	X	5	√	25 ≤ 256 characters	X
IEEE Xplore	Full index	Computer science, electrical engineering, electronics	4,831,568	7	1872	Proprietary	√	29	√	10 ≤ 15 search terms (can be longer than stated)	√
JSTOR	Full index	Multidisciplinary	12,000,000+	4	1857	Proprietary	√	12	X	7 (via max. search boxes)	√
Microsoft Academic	Full index	Multidisciplinary	213,850,455	6	Unknown	Mixed	X	0	Unknown	10	X
OVID	Selection: Embase, Embase Classic	Health studies	30,000,000+	9	1947	Proprietary	√	123	X	500	√
OVID			4,000,000+	8	1806	Proprietary	√	97	X	1,000	√

(Continues)

TABLE 4 (Continued)

Name of Search System	Database(s) Searched; Search Settings	1) Subject	2) Size	3)Record Type (Selectable Separately)	4) Retrospective Coverage (Oldest Entries)	5) Open Access Content?	6) Controlled Vocabulary?	7) Field codes/ Limiters?	8) Full Text Search Option?	9) Search String Length	10) Server Resp. Time/ Records: Max. Word Comb.
		D	D	D	D	D	D	N ≥ 5	D	N ≥ 25	N
	Selection: PsycINFO	Psychological, social, behavioral, health sciences									
ProQuest	Selection: ABI/INFORM Global	Business, management	24,190,026	13	1855	Proprietary	√	29	√	50	√
ProQuest	Selection: Nursing & Allied Health Database; Public Health Database	Nursing & Allied Health Database: Health studies; Public Health Database: Health studies	12,954,446	14	Nursing & Allied Health Database: 1857; Public Health Database: 1934	Proprietary	√	34	√	25	√
PubMed	Full index: Medline (and others)	Biomedicine; health studies	29,000,000+	3	1790	Mixed	√	29	X	1,000	√
ScienceDirect	Full index	Multidisciplinary	15,000,000+	24	1823	Proprietary	X	14	X	25 ≤ 8 Boolean connectors allowed (actually more possible)	√
Scopus	Full index	Multidisciplinary	70,000,000+	13	1861	Proprietary	√	79	X	1,000	√
Semantic Scholar	Full index	Multidisciplinary with a focus on computer science and medicine	72,366,665	10	1931 (or earlier)	Proprietary	X	0	Unknown	500	X
Springer Link	Full index	Multidisciplinary	12,731,539	10	1815 (single documents)	Proprietary	√	4	X	50	√

TABLE 4 (Continued)

Name of Search System	Database(s) Searched; Search Settings	1) Subject	2) Size	3)Record Type (Selectable Separately)	4) Retrospective Coverage (Oldest Entries)	5) Open Access Content?	6) Controlled Vocabulary?	7) Field codes/ Limiters?	8) Full Text Search Option?	9) Search String Length	10) Server Resp. Time/ Records: Max. Word Comb.
		D	D	D	D	D	D	N ≥ 5	D	N ≥ 25	N
Transport Research International Documentation (TRID)	Full index	Transportation studies	1,200,000+	5	1900	Proprietary	√	14	X	100	√
Virtual Health Library	Full index	Health studies	865,836	10	1902 (single studies); 1966 full	Proprietary	√	13	X	1,000	√
Web of Science	Selection: Web of Science Core Collection[a]	Multidisciplinary (depends on selected databases)	73,000,000+	21	1900 (depends on underlying subscription)	Proprietary	√	18	X	1,000	√
Web of Science	Selection: Medline	Health studies	29,303,305	81	1950	Proprietary	√	24	X	1,000	√
Wiley Online Library	Full index	Multidisciplinary	8,000,000+	3	1798	Proprietary	√	7	X	100	√
WorldCat	Selection: Thesis/ dissertation	Multidisciplinary	8,000,000+	17	About 1550 for earliest theses	Proprietary	√	13	X	25	√
WorldWideScience	Full index	Multidisciplinary	323,000,000	17	1869 (single studies)	Mixed	√	6	√	1,000	X

TABLE 4 Assessment of 28 academic search systems on their suitability for evidence synthesis

Name of Search System	Database(s) Searched; Search Settings	11) Language	12) Boolean Functional? OR	13) Boolean Functional? AND	14) Boolean Functional? NOT	15) Comparative Test	16) Query Interpretation/ Query Expansion	17) Truncation/ Wildcards Available?	18) Exact Phrases Functional?	19) Parenthesis Functional?	20) Post-query Results Refinement
		D	N	N	N	N	D	D	N	N	D
ACM Digital Library	Full index: Full-text collection	√ (E, Ch, Cy)	√	√	√	√	√	X	√	√	9
AMiner	Full index	√ (E, Ch, Cy)	X	X	X	X	X	X	X	X	0
arXiv	Full index; settings: All fields	√ (E), X (Ch, Cy)	√	√	√	X	√	X	√	X	0
Bielefeld Academic Search Engine (BASE)	Full index	√ (E, Ch, Cy)	√	√	√	√	√	X	√	√	9
CiteSeerX	Full index	√ (E, Ch, Cy)	X	√	√	X	X	X	X	X	0
ClinicalTrials.gov	Full index	√ (E, Cy) X (Ch)	√	√	√	√	X	X	√	√	12
Cochrane Library	Cochrane Central Register of Controlled Trials (CENTRAL)	√ (E, Ch, Cy)	√	√	√	√	X	X	√	√	3
Digital Bibliography & Library Project (DBLP)	Full index	√ (E, Cy) X (Ch)	X	X	X	X	√	X	X	X	4
Directory of Open Access Journals (DOAJ)	Full index	√ (E, Ch, Cy)	√	√	√	√	√	X	X	X	6
EbscoHost	Selection: ERIC; Medline; EconLit	√ (E, Ch, Cy)	√	√	√	√	X	X	√	√	13
EbscoHost	Selection: CINAHL Plus	√ (E, Ch, Cy)	√	√	√	√	X	X	√	√	12

TABLE 4 (Continued)

Name of Search System	Database(s) Searched; Search Settings	11) Language	12) Boolean Functional? OR	13) Boolean Functional? AND	14) Boolean Functional? NOT	15) Comparative Test	16) Query Interpretation/ Query Expansion	17) Truncation/ Wildcards Available?	18) Exact Phrases Functional?	19) Parenthesis Functional?	20) Post-query Results Refinement
		D	N	N	N	N	D	D	N	N	D
EbscoHost	Selection: SPORTDiscus	√ (E, Ch, Cy)	√	√	√	√	X	X	√	√	12
Education Resources Information Center (ERIC)	Full index	√ (E) X (Ch, Cy)	√	√	√	√	√	X	X	√	11
Google Scholar	Full index	√ (E, Ch, Cy)	X	X	√	X	X	X	√	X	2
IEEE Xplore	Full index	√ (E) X (Ch, Cy)	√	√	√	√	X	X	√	√	12
JSTOR	Full index	√ (E, Ch, Cy)	√	√	√	√	X	X	√	√	4
Microsoft Academic	Full index	√ (E, Ch, Cy)	X	X	X	X	√	X	X	X	7
OVID	Selection: Embase, Embase Classic	√ (E) X (Ch, Cy)	√	√	√	√	√	X	√	√	5
OVID	Selection: PsycINFO	√ (E) X (Ch, Cy)	√	√	√	√	√	X	√	√	6
ProQuest	Selection: ABI/INFORM Global	√ (E, Ch, Cy)	√	√	√ (with adapted search string)	√	X	X	√	√	10
ProQuest	Selection: Nursing & Allied Health Database; Public Health Database	√ (E, Ch, Cy)	√	√	√ (with adapted search string)	√	X	X	√	√	13
PubMed	Full index: Medline (and others)	√ (E, Cy) X (Ch)	√	√	√	√	X	X	√	√	10

TABLE 4 (Continued)

Name of Search System	Database(s) Searched; Search Settings	11) Language	12) Boolean Functional? OR	13) Boolean Functional? AND	14) Boolean Functional? NOT	15) Comparative Test	16) Query Interpretation/ Query Expansion	17) Truncation/ Wildcards Available?	18) Exact Phrases Functional?	19) Parenthesis Functional?	20) Post-query Results Refinement
		D	N	N	N	N	D	D	N	N	D
ScienceDirect	Full index	√ (E, Ch, Cy)	√	√	√	√	X	X	√	√	4
Scopus	Full index	√ (E, Ch, Cy)	√	√	√	√	X	X	√	√	11
Semantic Scholar	Full index	√ (E, Ch, Cy)	X	√	√	X	√	X	√	X	4
Springer Link	Full index	√ (E, Ch, Cy)	√	√	√	√	√	X	√	√	6
Transport Research International Documentation (TRID)	Full index	√ (E, Ch, Cy)	√	√	√	√	√	X	√	√	14
Virtual Health Library	Full index	√ (E, Ch, Cy)	√	√	√	√	√	X	√	√	13
Web of Science	Selection: Web of Science Core Collection[a]	√ (E) X (Ch, Cy)	√	√	√	√	X	X	√	√	18
Web of Science	Selection: Medline	√ (E, Ch, Cy)	√	√	√	√	X	X	√	√	13
Wiley Online Library	Full index	√ (E, Ch) X (Cy)	√	√	√	√	X	X	√	√	6
WorldCat	Selection: Thesis/ dissertation	√ (E, Ch, Cy)	√	√	√	X	X	X	√	√	7
WorldWideScience	Full index	√ (E, Ch, Cy)	X	X	X	X	X	X	X	X	12

TABLE 4 Assessment of 28 academic search systems on their suitability for evidence synthesis

Name of Search System	Database(s) Searched; Search Settings	21) Citation Search (Forward)	22) Advanced Search String Field?	23) Search Help?	24) No. of Accessible Hits	25) Bulk Download?	26) Repeatable? Time	27) Location-Independent? IP	Assessment
		D	D	D	N ≥ 1,000	D	N	N	
ACM Digital Library	Full index: Full-text collection	√	√	X	Full	2,000	√	√	PRINCIPAL
AMiner	Full index	√	√	X	1,000	X	√	√	SUPPLEMENTARY
arXiv	Full index; settings: All fields	√	√	√	10,000	X	√	√	SUPPLEMENTARY
Bielefeld Academic Search Engine (BASE)	Full index	X	√	√	1,000	100	√	√	PRINCIPAL
CiteSeerX	Full index	√	√	√	500	X	√	√	SUPPLEMENTARY
ClinicalTrials.gov	Full index	X	√	√	Full	Full	√	√	PRINCIPAL
Cochrane Library	Cochrane Central Register of Controlled Trials (CENTRAL)	X	√	√	Full	Full	√	√	PRINCIPAL
Digital Bibliography & Library Project (DBLP)	Full index	X	X	√	Unknown (most likely full)	X	√	√	SUPPLEMENTARY
Directory of Open Access Journals (DOAJ)	Full index	X	√	X	10,010	X	√	√	SUPPLEMENTARY
EbscoHost	Selection: ERIC; Medline; EconLit	X	√	√	25,000	25,000	√	√ (depends on database access)	PRINCIPAL
EbscoHost	Selection: CINAHL Plus	X	√	√	25,000	25,000	√	√ (depends on database access)	PRINCIPAL
EbscoHost	Selection: SPORTDiscus	X	√	√	25,000	25,000	√	√ (depends on database access)	PRINCIPAL

TABLE 4 (Continued)

Name of Search System	Database(s) Searched; Search Settings	21) Citation Search (Forward)	22) Advanced Search String Field?	23) Search Help?	24) No. of Accessible Hits	25) Bulk Download?	26) Repeatable? Time	27) Location-Independent? IP	Assessment
		D	D	D	N ≥ 1,000	D	N	N	
Education Resources Information Center (ERIC)	Full index	X	X	√	Full	200	√	√	
Google Scholar	Full index	√	√	√	1,000	X	X	X	**SUPPLEMENTARY**
IEEE Xplore	Full index	X	√	√	2,000	2,000	√	√	**SUPPLEMENTARY**
JSTOR	Full index	X	√	√	1,000	X	√	√	**SUPPLEMENTARY**
Microsoft Academic	Full index	√	X	X	5,000	X	√	√	**SUPPLEMENTARY**
OVID	Selection: Embase, Embase Classic	√	√	√	Full	1,000	√	√ (depends on database access)	**PRINCIPAL**
OVID	Selection: PsycINFO	√	√	√	Full	1,000	√	√ (depends on database access)	**PRINCIPAL**
ProQuest	Selection: ABI/INFORM Global	√	√	√	10,000	100	√	√ (depends on database access)	**PRINCIPAL**
ProQuest	Selection: Nursing & Allied Health Database; Public Health Database	√	√	√	10,000	100	√	√ (depends on database access)	**PRINCIPAL**
PubMed	Full index: Medline (and others)	√	√	√	Full	Full	√	√	**PRINCIPAL**
ScienceDirect	Full index	√	√	√	6,000	100	√	√	**PRINCIPAL**
Scopus	Full index	√	√	√	2,000	20,000	√	√	**PRINCIPAL**
Semantic Scholar	Full index	√	√	√	10,000	X	√	√	**SUPPLEMENTARY**
Springer Link	Full index	X	√	√	19,980	1,000	√	√	**SUPPLEMENTARY**
Transport Research International Documentation (TRID)	Full index	X	X	√	15,000	X	√	√	**PRINCIPAL**
Virtual Health Library	Full index	X	√	√	Full	Full	√	√	**PRINCIPAL**
Web of Science	Selection: Web of Science Core Collection[a]	√	√	√	100,000	5,000	√	√ (depends on database access)	**PRINCIPAL**

TABLE 4 (Continued)

Name of Search System	Database(s) Searched; Search Settings	21) Citation Search (Forward)	22) Advanced Search String Field?	23) Search Help?	24) No. of Accessible Hits	25) Bulk Download?	26) Repeatable?	27) Location-Independent?	Assessment
		D	D	D	N ≥ 1,000	D	N	N	
Web of Science	Selection: Medline	√	√	√	100,000	5,000	√	√ (depends on database access)	PRINCIPAL
Wiley Online Library	Full index	√	√	√	2,000	X	√	√	PRINCIPAL
WorldCat	Selection: Thesis/ dissertation	X	√	√	5,000	X	√	√	SUPPLEMENTARY
WorldWideScience	Full index	X	√	√	Limited (only top results are shown)	20	X	X	SUPPLEMENTARY

Note. √, passed test; X, failed test;
Abbreviations: D, desired; N, necessary.
[a]Science Citation Index Expanded (1900-present); Social Sciences Citation Index (1900-present); Arts & Humanities Citation Index (1975-present); Conference Proceedings Citation Index- Science (1990-present); Conference Proceedings Citation Index- Social Science & Humanities (1990-present); Book Citation Index– Science (2010-present); Book Citation Index– Social Sciences & Humanities (2010-present); Emerging Sources Citation Index (2015-present); Current Chemical Reactions (2010-present); (Includes Institut National de la Propriete Industrielle structure data back to 1840); Index Chemicus (2010-present).

We found that all but four search systems (24) included some form of advanced search field where users could structure their systematic searches. While some systems possessed multiple forms with varying degrees of complexity according to the needs and search literacy of the users (eg, ProQuest and Web of Science), others only provided rudimentary interfaces, leaving the user little freedom to specify requests (eg, AMiner and Microsoft Academic). All systems except Microsoft Academic and Semantic Scholar had some form of field codes that allowed the user to specify which parts of the structured information available in the databases to access. Both these search systems relied on semantic searching, where, in contrast to a traditional literal search, the focus is on interpreting what users might have *meant*, instead of retrieving results according to exactly what the search terms specified. Following a semantic search request, the search results can then be customized with a limited number of post-query filters. In contrast, some traditional search systems, like PubMed, and IEEE Xplore offer close to 30 different options to filter their highly structured databases, which is especially convenient for the structured queries necessary in the fields of medicine and engineering. For platforms, the number and type of field codes depends on the underlying database. All but four search systems (24) offered some form of search help to assist users in conducting their search.

While most (21 of 28) search systems offer some kind of controlled vocabulary, the quality differs significantly. In medicine, reviewers rely on frequently updated and rigorously categorized Medical Subject Headings (MeSH), while in other disciplines, specialist databases offer simpler thesauri. The availability of controlled vocabulary depends on the underlying database and its data structure. We found that in 79% of the cases, the controlled vocabulary was presented in a hierarchical form with multiple levels, and in 61% of the cases, the controlled vocabulary was searchable. Full text search functionality was available in only nine search systems, while 17 search systems did not provide that functionality, and for the semantic search engines Microsoft Academic and Semantic Scholar, it was unclear what parts of the records are indexed and searched.

While AMiner, arXiv, and CiteSeerX do not provide any post-query refinements, other systems such as EbscoHost, ProQuest, and Web of Science provide up to 18 different options for filtering content. However, these refinement options depend significantly on the reviewers' selection of databases searched. As the databases hosted on these systems differ in their structured information, the options for filtering them differ

accordingly. Forward citation searching was available on more than half (15) of all search systems.

3.3 | Search results

We found that the maximum number of retrievable records varies greatly among search systems. While some allow access to all records, the functionality of CiteSeerX and WorldWideScience is severely restricted as their threshold lies below 1000 records. Most notably, ACM Digital Library, ClinicalTrials.gov, Cochrane Library, ERIC, OVID, PubMed, and Virtual Health Library allow full access to all datasets returned from a single search. For DBLP, we could not determine the scope of retrievable records since its results set expands dynamically rather than using numeration or pagination. Similarly, bulk download options differed significantly across search systems. Some allow the download of the entire search results set in one go, while almost half of the examined systems provided no support for for exporting multiple records. Most positively, the medical databases of Virtual Health Library, ClinicalTrials.gov, and Cochrane Library supported efficient data retrieval by allowing full download options. While many databases offered an application programming interface (API) to access their database, these options are only accessible to reviewers with programming skills.

3.4 | Search reproducibility

In our sample of 28 academic search systems, all but two—Google Scholar and WorldWideScience—were reproducible in terms of reporting identical results for repeated identical queries. While WorldWideScience failed to deliver replicable results at all times, Google Scholar failed to deliver them only during certain periods: sometimes, search results were replicable with two consecutive queries; then with a third query or with queries after some queries in between, they were no longer replicable and the results set differed in a way not explainable by *natural* database growth. Natural growth means that the dataset indexed on a database increases with the identification and curation of new records and thus that results sets retrieved tend to increase with repeated queries as the underlying database has expanded in the meantime. All the other 26 search systems appeared to provide reproducible results.

Further, in our analysis of search results retrieved via a changed retrieval location, we found differences for varying institutional subscriptions, yet not for differences in IP addresses. Certain subscription-based platforms—for example, EbscoHost, OVID, ProQuest, and Web of Science—delivered notably different results depending on the institution through which we accessed the underlying databases. For these systems, the number of records depended on the subscriptions to different databases or indexes subscribed to by the organization. In most cases, differences depended on the databases available, yet there were also differences within the same databases. We found the coverage of the same database was different as the number of years accessible varies from package to package. These differences can be visible in the description of the database, as with ProQuest offering its popular ABI/Inform package in different versions containing substantially different results sets. Nevertheless, these differences are sometimes not so obvious for users, requiring closer examination. Web of Science's Core Collection also varies significantly in scope containing different indices depending in the subscription. These single indices, again, vary in scope for the subscribing libraries. For the same index, one institution might have subscribed to a retrospective coverage since 1996, another since 2010. The variations highlight that reviewers should be familiar with their institution's subscription and that they need to document this in detail in their review reports.

4 | DISCUSSION

Overall, we found that only 14 of the 28 academic search systems examined are well-suited to evidence synthesis in the form of systematic reviews in that they met all necessary performance requirements (Table 4). These 14 can be used as principal search systems: ACM Digital Library, BASE, ClinicalTrials.gov, Cochrane Library, EbscoHost (tested for ERIC, Medline, EconLit, CINHAL Plus, SportsDiscus), OVID (tested for Embase, Embase Classic, PsychINFO), ProQuest (tested for Nursing & Allied Health Database, Public Health Database), PubMed, ScienceDirect, Scopus, TRID, Virtual Health Library, Web of Science (tested for Web of Science Core Collection, Medline), and Wiley Online Library. In contrast, the remaining 14 were unsuitable for use as the principal search system for systematic reviews due to failing to meet one or more necessary criteria. For these 14 search systems, our tests uncovered severe performance limitations with regard to formulating queries, the correct interpretation of queries by the system, data retrieval capabilities, and the reproducibility of searches. These systems should only be considered supplementary to the principal systems, especially for nonquery-based search methods where they might still provide great benefit. Desired

performance criteria inform about other-nonessential functionalities relevant for systematic search, where reviewers need to assess individually how important these criteria are for their specific search. We next present the results of our analysis. The criteria we base our assessment on can be found in Table 2, and the detailed outcomes of the tests conducted can be found in Appendix II (supplementary online material).

4.1 | Necessary criteria

In most systematic reviews, Boolean queries retrieve the largest portion of relevant records because they allow the user to search large databases with the highest recall. Accordingly, the query-based search form is the backbone of systematic reviews. It is striking that half the search systems we examined have at least some issues with Boolean queries. This is particularly unfortunate because Boolean searching is effective, especially for systematic search strategies: "medical research indicates that expert searching based on Boolean systems is still the most effective method [of searching]."[26, p1570]

Our tests revealed that the help files of numerous search systems promise a Boolean search functionality that our tests could not verify. These findings were especially alarming because users of such systems rely on functionalities that they assume to work properly, but that may not be the case. In a review of search FAQs, we found that arXiv, CiteSeerX, DBLP, ERIC, Semantic Scholar, WorldCat, and WorldWideScience promote support for Boolean searching, yet our tests identified issues with searches using Boolean operators. Semantic Scholar, for example, writes in its FAQ "you may search for papers on Semantic Scholar using AND/OR query terms," yet it failed most query-based tests (criteria 10, 12, 15, and 19). For these services, the negative performance results might hint at glitches that the system administrators should examine. The other systems that failed query tests—AMiner, DOAJ, Google Scholar, and Microsoft Academic—do not state support for Boolean search functionality.

Given the findings in this study, we advocate reassessing the advice given in evidence-synthesis guidance for systematic reviews. For example, the searching guidance of The Campbell Collaboration states: "Given an Internet search engine (Google, Google Scholar, Bing, etc.) [...], many of these search strategies may also be applied. For example, Phrase searching, Boolean Operators and Limiting features are typically all offered. Using the search engine's Advanced search screen can provide an easy way of accessing these features."[15, p34] This passage might incorrectly advise users to pursue full Boolean search strategies with search systems such as Google Scholar that do not offer such functionality. Further, our results contradict systematic review guidance that assumes that "all the search engines in some way [would] permit the use of Boolean syntax operators to expand or restrict the search."[5, p103] The results of our study show that when it comes to search functionalities that are necessary for systematic reviews, a reviewer must look closely at which search systems are in fact suitable and why. If reviewers are comfortable with search systems failing specific *desired* criteria, while query capabilities are sufficient, they should not be discouraged from using it as their principal search system.

4.2 | Desired criteria

Search systems failing one or more necessary test criteria are *always* in conflict with the fundamental quality requirements of systematic reviews, especially for searches with search strings. Accordingly, we advise reviewers to use these systems solely for supplementary search methods as they might still be valuable in improving search outcome. Such supplementary methods include handsearching of backward and forward citations, specific issues or journals, or the use of filters to limit search results. We explicitly tested for handsearching in the form of forward citation searching, as this information needs to be provided by a citation index that contains information on which records have cited a specific record. This citation index is, however, not available through every search system. While our methods did not allow the testing of the comprehensiveness of citation indexes, larger, multidisciplinary search systems seem to provide more complete citation information than smaller specialized search systems. Comparisons of citation indexes generally rate Google Scholar as the most comprehensive.[71-73] These comparisons support the idea that larger search systems tend to have greater citation coverage. While our results show that 15 of the 28 systems examined have cross-citation information, because of the limited coverage of many of these systems', their citation information might be limited as well. Hence, if the reviewer's goal is to reach beyond the limitations of a specialized search system, it might prove beneficial to use citation information from a large, multidisciplinary search system to broaden the search scope.

Desired performance criteria need to be evaluated relative to the *specific* systematic search requirements of the reviewer. Our analysis made some important evaluation criteria transparent, so it is possible for reviewers to reflect on how these criteria could facilitate or limit their

systematic searches. For reviewers, it is important to choose a search system that is suitable for a given research domain, a certain retrospective focus, and that covers the specific record type of interest. Coverage of a search system and/or its underlying database(s) might be important for evaluating a search system's potential recall. Coverage is, however, only beneficial when the necessary retrieval capabilities are offered as well. Otherwise, searching large, multidisciplinary databases might involve low search precision, making systematic search inefficient and laborious. Alternatively, reviewers might want to test systems offering the option to download resources in bulk. Compared with systems without this feature, bulk download allows efficient data handling in combination with reference management software and data analysis tools.

Our analysis showed that of the five Open Access search systems examined that catalogue Open Access records only (arXiv, CiteSeerX, ClinicalTrials.gov, DBLP and DOAJ), only ClinicalTrials.gov passed all tests relating to necessary criteria. Among the six systems offering mixed access, that is, using a dataset offering both proprietary and Open Access content, BASE and PubMed were found to be suitable for use as principal search systems. Accordingly, for reviewers having no access to proprietary databases, our findings mean they only have a limited selection of Open Access database alternatives. Reviewers interested in medical evidence synthesis could access all three—BASE, PubMed, and ClinicalTrials.gov—but should be aware that BASE also indexes PubMed. Reviewers from other disciplines who want synthesise Open Access content systematically, however, are limited to the multidisciplinary system BASE, which provides full texts for 60% of its close to 150 million records under Open Access licence. Other open, or partially open, search systems that fail to meet the criteria for query-based search might still be useful for supplementary search methods.

4.3 | Differences in search systems

The tests applied in this study not only compared individual search systems but also underlying databases accessed through different platforms. For example, ERIC's database is accessible via its dedicated search system, but also through EbscoHost. Similarly, we accessed Medline through EbscoHost, PubMed, and Web of Science (and indirectly through BASE and other systems that use the Medline index). The analysis detailed above detected some performance differences. While the underlying database seemed largely identical, determined by its size, the functionalities of the search system through which it was accessed varied. ERIC (the search system), for example, failed some necessary tests, whereas when accessed through EbscoHost, the search functionalities in searching the ERIC database were superior. We also identified differences in the case of Medline. While PubMed allows bulk download of the full dataset, EbscoHost allows 25 000 and Web of Science 5000. Hence, we conclude that in these cases the search capabilities depended on the system through which it was accessed and less on the underlying database.

Additionally, the platforms in our sample—EbscoHost, OVID, ProQuest, and Web of Science—all underwent multiple tests where individual platforms were tested with varying databases. These repeated tests were aimed to provide another perspective on performance determinants of platforms and should show whether changing underlying databases influenced necessary and desired performance criteria. As the underlying databases changed, so accordingly did the results for tests of these databases, such as scope, available record types, controlled vocabulary, retrospective coverage, field codes, or filters. Further, we found that the maximum length of the search string a platform could handle without timeout differed significantly depending on the size and number of underlying databases. The scope of the search determined by field codes also seems to influence server load and thus maximum search string length. This means a full text search puts a heavier load on the system than, for example, a search of titles or abstracts. Hence, it is not the number of characters (alone) that determines the longest still computable search string. Reviewers may thus need to balance search string length with database selection and field code selection in the case of more exhaustive searches. Another option might be to split the string into pieces and search systems sequentially, although doing so would extend the workload associated with documentation and deduplication.

The quality of systematic searches not only depends on the queries or filters specified for searching a database, but also depends on the database itself. Database providers/platforms, such as EbscoHost, OVID, ProQuest, and Web of Science, provide access to multiple databases simultaneously. Hence, for these platforms, reviewers need to report on the exact databases they have searched. Nevertheless, inexperienced researchers frequently wrongly assume they are searching a single, distinct database, while in truth, that search system aggregates multiple databases. The consequence is that these authors report using a search system but omit to record using its underlying databases. With this limited information, the search process is insufficiently documented and replication is impossible. Hence, for systematic searches of platforms such as EbscoHost, OVID, ProQuest, and Web of

Science, we remind authors that they must report the underlying databases and the indices they contain.[74] Further, it is necessary to bear in mind that databases update frequently—sometimes multiple times a day or even on an hourly basis—thus the underlying dataset changes accordingly. While most of the time the dataset will increase through the addition of records, databases can also shrink through the deletion of duplicates or the occurrence of errors that affect the dataset provided. It is therefore essential to report—in addition to the exact database accessed—the time the dataset was accessed too. One option to facilitate these reporting practices might be to include such requirements in the "guide for authors" section of journals and to advise the use of reporting guidance such as PRISMA[36] or ROSES.[74]

4.4 | Emergence of semantic search systems

There has recently been an upsurge in using semantic search engines over traditional ones, as is evident in the birth of Semantic Scholar (2015), the relaunch of Microsoft Academic (2017), and the expected launch of *Meta*, a project of the Zuckerberg foundation. These semantic search engines tend to be designed to reward exploratory rather than systematic search behavior. These tendencies add to the notion that "the problem is [...] that an ideological tendency to make things 'user friendly' (and the market bigger) tends to hurt the development of systems aimed at increasing the selection power of users and search experts."(26, p1570) Our findings indicating that these systems are inadequate to be used as principal systems in systematic searches support this notion. The criticism of user-friendliness at any cost is especially directed at Google Scholar, which is more concerned with "*tuning*" its first results page(75, p15) than with overall precision. This makes Google Scholar highly precise for exploratory searches conducted by a user interested in only a few relevant results on the first search engine results page.[76,77] Nevertheless, overall, Google Scholar's search precision has been found to be significantly lower than 1% for systematic searches.[25] This is not surprising, since our findings show that Google Scholar does not support many of the features required for systematic searches. Our findings support the criticism of Bramer et al,[33] Bramer et al,[34] and Boeker et al[25] and indicate that Google Scholar's coverage and recall is an inadequate reason to use it as principal search system in systematic searches.[53] If a system such as Google Scholar fails to deliver retrieval capabilities that allow a reviewer to search systematically with high levels of recall, precision, transparency, and reproducibility, its coverage is irrelevant for query-based search. Google Scholar's extraordinary coverage acting as a multidisciplinary compendium of scientific world knowledge should not blind users to the fact that users' ability to access this compendium is severely limited, especially in terms of a systematic search.

While popular search systems such as Google Scholar or Microsoft Academic being inadequate for query-based search is already unfortunate on its own, the situation is made worse by users seemingly being unaware of these shortcomings. Users are perhaps guided by convenience rather than strategic considerations when choosing their search system. In fact, it was due to its great ease of use and performance in informational and exploratory searches[78] that Google Scholar emerged as the number one go-to academic search engine for most academic users.[79-82] Students in particular utilize Google as a main source in information seeking.[52,83,84] The requirements for evidence synthesis are not always obvious to reviewers as they are used to navigational or exploratory searching that comes intuitively.[78,85-87] Students especially seem to have tremendous difficulty in mastering online literature searches.[83,88-91] Other search systems, such as bibliographic databases or platforms, are less popular due to the elevated skills they require and, in some cases, more difficult access due to paywall restrictions. We advocate educated use of these systems, so users have the right tool for the right purpose fully aware of its strengths and weaknesses.

4.5 | Limitations

While we took the greatest care to include a large evidence-based selection of meaningful methods to test the capacity of search systems, there may be other tests unknown to us that could be performed. The tests conducted here do, however, rigorously and transparently assess whether and to what extent search systems succeed in such tests compared to other systems. Generally, we did not directly test a search system's *level of* precision or recall, but rather *the capacity of allowing the user to specify queries* with high levels of precision and recall. From a methodological standpoint, this study is particularly influenced by the research of Gusenbauer[37] and Boeker et al(25, p11) that sought to, "to compare the effectiveness of Google Scholar and other retrieval tools" Our approach provides great practical benefit, as it illuminates some of the most critical strengths and weaknesses of search systems that are often only communicated without comparative evidence of actual performance criteria. This study contributes such tangible criteria. Nevertheless, from a theoretical standpoint, our study can only

provide evidence that search systems behave incorrectly in failing to comply with certain test criteria. It is impossible to be absolutely certain that a system that has proved successful in our specific tests would not fail in slightly different tests or under different circumstances. For instance, a system providing a functional 1000-term OR-string could theoretically fail one consisting of 1001 terms or let us say 521 terms. Similarly, it is impossible to rule out that search systems have temporal performance variations that cannot be captured with cross-sectional analysis. While we also included a test for temporal variation through reproducibility tests where we determined search engine bias, we have to assume that the systems are otherwise stable in scenarios that lie beyond our tested scope.

One possible limitation might be that whenever a certain threshold is defined, someone asks why it was not some other threshold. In this study, we tried to alleviate this criticism by basing our thresholds on the quality guidance issues by Cochrane, The Campbell Collaboration, and the CEE. Further, if we needed to decide on specific numeric thresholds such as the minimum length of search strings or the minimum number of field codes, we based our decision on a review of best practices from previously published and highly cited systematic reviews.

As most of the search systems update not only their database, but also their search functionalities, the performance results tested in this study might change over time. However, during the period covered by writing-up the results of this study, those results remained relatively stable, which most likely reflects the fact that our performance tests evaluated fundamental functionalities of search systems that are rarely updated. Nevertheless, to be sure to have an accurate picture of search functionalities, reviewers can easily replicate our tests and evaluate them immediately before they access their search system of interest.

5 | CONCLUSION

Selection of suitable search systems is essential for the outcome of evidence-synthesis research. Reviewers must consider the different functionalities offered, or not offered, when interacting with a given search system. In particular, they must be cognisant of the trade-off between search precision and recall. Searching search systems with the greatest effectiveness and efficiency is a skill that is necessary yet generally undervalued in education and research practice. Reviewers should always consult information specialists or librarians and enlist their support in designing systematic review search strategies.[9,13] Only if reviewers are aware of a search system's functionalities, they can take advantage of all methods and functionalities and design good search strategies.

Yet, so often convenience guides the method of search system choice. Unfortunately, awareness of the differences between search systems is not yet sufficiently developed in the area of scientific education. Indeed, librarians affirm the lack of search skills prevalent especially among students[89,92] and so-called digital natives.[88,93] We hope our study helps to create awareness of the importance of search literacy. This study shows the limitations of such convenience. This research encourages responsible and knowledgeable researchers to be aware of search system qualities so they can then use the appropriate tool for the task at hand. Just as artisans have particular tools for particular tasks, we should understand our digital tools are not one-size-fits-all solutions. Crawler-based search engines like Google Scholar or Microsoft Academic function differently to database providers such as ProQuest or EbscoHost, or journal platforms such as SpringerLink or Wiley. The overview provided here should make it easier for scholars to choose the most adequate search system according to their unique information requirements.

The many limitations we identified affecting most of the search systems in our study clearly call for researchers - especially those who engage in systematic searches - to ensure that they possess considerable knowledge of the search systems they intend to use; however, those qualities are not always evident. Without this knowledge, search results might be misinterpreted in a way that impinges on research validity. A high number of hits resulting from an extensive Boolean search string might, for example, be seen as indication of a high number of relevant records—yet in truth, due to faulty interpretation of the search system, might reflect the malfunction of limiting AND or NOT operators.

It has often been unclear exactly why certain search systems perform better or worse than others. Performance issues, especially those concerning the correct interpretation of Boolean search strings by search systems, may have remained undetected so far. We aimed to make these performance differences explicit. Since we used the same metrics for all systems, our assessment makes a large set of systems comparable. If impediments of search systems are made transparent, experienced reviewers could perhaps circumvent these limitations by using search systems differently. However, researchers lacking such knowledge run the risk of expecting too much of search systems (even when searched in a systematic way) and drawing erroneous conclusions based on biased sets of search

results. The establishment of the 27 testing criteria here may help to create awareness among reviewers of where they need to look when selecting and using search systems.

If the results of scientific research are to be cumulative, researchers in general, and especially those aiming to conduct evidence syntheses, should know how to effectively and efficiently gather scientific knowledge. Our evaluation offers reviewers a means of transparent evaluation. While some researchers highlight the benefits of easy-to-use academic search engines like Google Scholar[53] that allow non-experts to make use of scholarly resources,[94] our work highlights the specific pitfalls of those systems. In contrast, we demonstrate that using search systems correctly is not always as straightforward as slick user interfaces might suggest. Further, our detailed assessment based on 27 transparent criteria is also especially helpful for experienced reviewers of all disciplines when they decide on which criteria their search system of choice must meet. The distinction between necessary and desired criteria should create awareness of why certain search systems are suitable and others unsuitable, and that simple distinction could be helpful, especially for non-expert reviewers or those who are not information specialists.

Our analysis reveals that few Open Access search systems can be recommended as a principal resource for systematic searches. It seems there is currently almost no getting around proprietary search systems if one attempts a rigorous systematic review. This finding is extremely unfortunate, as Open Access databases advocate barrier-free access to information, yet for systematic reviews, they most often do not provided the necessary functionalities to be used as principal search systems. For researchers from resource-constrained contexts, we could perhaps recommend the multidisciplinary BASE, as it is a comprehensive resource with a large share of Open Access content and it also met necessary testing criteria in our analysis.

We advocate that search system operators—Open Access or not—review the capabilities and improve performance criteria where necessary. Ideally, search system providers would use our insights to further develop their systems according to the high standards of evidence-synthesis guidance. It becomes evident that these providers need to balance the pros and cons of "exact-match systems" and "best-match systems" or find ways of alleviating the effects of trade-offs between both concepts.[26] The metrics used in this research might prove helpful in defining some of the specifications an improved system should possess—something especially relevant for those systems in which we uncovered severe performance limitations. Such scientific performance requirements might become increasingly relevant with the current research trend of replicating existing studies, and with the continued increase in the number of published systematic reviews.

The criteria established in this study are relatively straightforward as they are defined from the viewpoint of the reviewer. Therefore, it is easily possible for reviewers to update these assessments frequently by identifying changes in the qualities of the single search systems or adding previously unexamined search systems to the comparison set. Until now, studies have largely examined the suitability of search systems for certain scholarly tasks for individual systems or by comparing a few systems.[33,50,95] Our methods allowed a comprehensive review of many different search systems. As a result we found significant performance differences among the search engines examined, confirming that no single search system is perfect. Their efficient use thus demands searchers are well-trained and can weigh up a system's strengths and weaknesses and make informed decisions on where and how to search. Only then can reviewers evaluate search systems and match our tested search system suggestions to their subjective information requirements.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the Appendix (see Supporting Information).

ACKNOWLEDGEMENTS

We like to thank the two anonymous reviewers, associate editor, and editor Prof Gerta Rücker of Research Synthesis Methods for their valuable comments that helped improving this paper. All remaining errors and omissions are ours.

CONFLICT OF INTEREST

The author reported no conflict of interest.

ORCID

Michael Gusenbauer https://orcid.org/0000-0001-7768-2351

Neal R. Haddaway https://orcid.org/0000-0003-3902-2234

REFERENCES

1. Price DJS. *Little Science, Big Science*. New York: Columbia Univ. Press; 1963 Columbia paperback.
2. Larsen PO, von Ins M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation

Index. *Scientometrics.* 2010;84(3):575-603. https://doi.org/10.1007/s11192-010-0202-z
3. Eden D. From the editors: replication, meta-analysis, scientific progress, and AMJ's publication policy. *AMJ.* 2002;45(5):841-846. https://doi.org/10.5465/AMJ.2002.7718946
4. Naisbitt J, Aburdene P. *Megatrends 2000: Ten New Directions for the 1990's*. 1st ed. New York: Morrow; 1990.
5. Cooper HM. *Research Synthesis and Meta-analysis: A Step-by-Step Approach*. Applied social research methods series. Fifth ed. 2 Los Angeles: SAGE; 2017.
6. Littell JH. *Conceptual and Practical Classification of Research Reviews and Other Evidence Synthesis Products*; 2018.
7. Kostoff RN, Shlesinger MF. CAB: citation-assisted background. *Scientometrics.* 2005;62(2):199-212. https://doi.org/10.1007/s11192-005-0014-8
8. Littell JH, Corcoran J, Pillai V. *Systematic Reviews and Meta-Analysis:* Oxford University Press, USA; 2008. https://books.google.at/books?id=UpsRDAAAQBAJ.
9. Rethlefsen ML, Farrell AM, Osterhaus Trzasko LC, Brigham TJ. Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. *J Clin Epidemiol.* 2015;68(6):617-626. https://doi.org/10.1016/j.jclinepi.2014.11.025
10. Bandara W, Furtmueller E, Gorba, Gorbacheva E, Miskon S, Beekhuyzen J. Achieving rigor in literature reviews: insights from qualitative data analysis and tool-support. *Communications of the Association for Information Systems.* 2015;37(8):154-204. http://aisel.aisnet.org/cais/vol37/iss1/8
11. Meert D, Torabi N, Costella J. Impact of librarians on reporting of the literature searching component of pediatric systematic reviews. *J Med Libr Assoc.* 2016;104(4):267-277. https://doi.org/10.3163/1536-5050.104.4.004
12. Koffel JB. Use of recommended search strategies in systematic reviews and the impact of librarian involvement: a cross-sectional survey of recent authors. *Plos One.* 2015;10(5):1-13. https://doi.org/10.1371/journal.pone.0125931
13. Livoreil B, Glanville J, Haddaway NR, et al. Systematic searching for environmental evidence using multiple tools and sources. *Environ Evid.* 2017;6:1-14. https://doi.org/10.1186/s13750-017-0099-6
14. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions*; 2011; Version 5.1.0.
15. Kugley S, Wade A, Thomas J, et al. *Searching for studies: GUidelines on information retrieval for Campbell Systematic Reviews*; 2016; 1.
16. Pullin A, Frampton G, Livoreil B, Petrokofsky G. *Guidelines and Standards for Evidence Synthesis in Environmental Management. Version 5.0*; 2018.
17. Hug SE, Braendle MP. The coverage of Microsoft Academic: analyzing the publication output of a university. *Scientometrics.* 2017;113(3):1551-1571. https://doi.org/10.1007/s11192-017-2535-3
18. Khabsa M, Giles CL. The number of scholarly documents on the public web. *Plos One.* 2014;9(5):1-6. https://doi.org/10.1371/journal.pone.0093949
19. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB journal.* 2008;22(2):338-342. https://doi.org/10.1096/fj.07-9492LSF
20. Orduña-Malea E, Ayllón JM, Martín-Martín A, Delgado L-CE. Methods for estimating the size of Google Scholar. *Scientometrics.* 2015;104(3):931-949. https://doi.org/10.1007/s11192-015-1614-6
21. Harzing A-W. A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics.* 2014;98(1):565-575. https://doi.org/10.1007/s11192-013-0975-y
22. Meier JJ, Conkling TW. Google Scholar's coverage of the engineering literature: an empirical study. *The Journal of Academic Librarianship.* 2008;34(3):196-201. https://doi.org/10.1016/j.acalib.2008.03.002
23. Turnbull D, Berryman J. *Relevant search: with applications for Solr and Elasticsearch*. Shelter Island New York: Manning Publications Co; 2016.
24. Levay P, Ainsworth N, Kettle R, Morgan A. Identifying evidence for public health guidance: a comparison of citation searching with Web of Science and Google Scholar. *Res Synth Methods.* 2016;7(1):34-45. https://doi.org/10.1002/jrsm.1158
25. Boeker M, Vach W, Motschall E. Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. *BMC Med Res Methodol.* 2013;13:1-12. https://doi.org/10.1186/1471-2288-13-131
26. Hjørland B. Classical databases and knowledge organization: a case for Boolean retrieval and human decision-making during searches. *J Assn Inf Sci Tec.* 2015;66(8):1559-1575. https://doi.org/10.1002/asi.23250
27. Weber K. Search engine bias. In: Lewandowski D, ed. *Handbuch Internet-Suchmaschinen 2*. AKA Verlag Heidelberg; 2011:265-285.
28. Vaughan L, Thelwall M. Search engine coverage bias: Evidence and possible causes. *Information Processing & Management.* 2004;40(4):693-707. https://doi.org/10.1016/S0306-4573(03)00063-3
29. Carmines EG, Zeller RA. *Reliability and Validity Assessment*. Quantitative applications in the social sciences. Beverly Hills, London: Sage Publications; 1979 no.07–017.
30. Jacsó P. Google Scholar: the pros and the cons. *Online Information Review.* 2005;29(2):208-214. https://doi.org/10.1108/14684520510598066
31. Jacsó P. Google Scholar revisited. *Online Information Review.* 2008;32(1):102-114. https://doi.org/10.1108/14684520810866010
32. Bethel A, Rogers M. A checklist to assess database-hosting platforms for designing and running searches for systematic reviews. *Health Info Libr J.* 2014;31(1):43-53. https://doi.org/10.1111/hir.12054
33. Bramer WM, Giustini D, Kramer B, Anderson P. The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews. *Syst Rev.* 2013;2:1-9. https://doi.org/10.1186/2046-4053-2-115
34. Bramer WM, Giustini D, Kramer BMR. Comparing the coverage, recall, and precision of searches for 120 systematic reviews in Embase, MEDLINE, and Google Scholar: a prospective study. *Syst Rev.* 2016;5:1-9. https://doi.org/10.1186/s13643-016-0215-7
35. Mowshowitz A, Kawaguchi A. Measuring search engine bias. *Information Processing & Management.* 2005;41(5):1193-1205. https://doi.org/10.1016/j.ipm.2004.05.005

36. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 2009;6(7):1-6. https://doi.org/10.1371/journal.pmed.1000097
37. Gusenbauer M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics.* 2019;118(1):177-214. https://doi.org/10.1007/s11192-018-2958-5
38. Ortega JL. *Academic Search Engines: A quantitative outlook*. Chandos information professional series. Oxford, UK: Chandos Publishing/Elsevier; 2014.
39. Schöpfel J, Farace DJ. Grey literature. In: Bates MJ, Maack MN, eds. *Encyclopedia of library and information sciences*. 3rd ed. / edited by Boca Raton, Fla: CRC London: Taylor & Francis; 2010:2029–2039.Marcia J. Bates and Mary Niles Maack
40. Sampson M, McGowan J. Inquisitio validus Index Medicus: a simple method of validating MEDLINE systematic review searches. *Res Synth Methods.* 2011;2(2):103-109. https://doi.org/10.1002/jrsm.40
41. Rogers M, Bethel A, Abbott R. Locating qualitative studies in dementia on MEDLINE, EMBASE, CINAHL, and PsycINFO: a comparison of search strategies. *Res Synth Methods.* 2017;9(2):579-586. https://doi.org/10.1002/jrsm.1280
42. Rader T, Mann M, Stansfield C, Cooper C, Sampson M. Methods for documenting systematic review searches: a discussion of common issues. *Res Synth Methods.* 2014;5(2):98-115. https://doi.org/10.1002/jrsm.1097
43. O'Mara-Eves A, Brunton G, McDaid D, Kavanagh J, Oliver S, Thomas J. Techniques for identifying cross-disciplinary and 'hard-to-detect' evidence for systematic review. *Res Synth Methods.* 2014;5(1):50-59. https://doi.org/10.1002/jrsm.1094
44. Atkinson KM, Koenka AC, Sanchez CE, Moshontz H, Cooper H. Reporting standards for literature searches and report inclusion criteria: making research syntheses more transparent and easy to replicate. *Res Synth Methods.* 2015;6(1):87-95. https://doi.org/10.1002/jrsm.1127
45. Mahood Q, van Eerd D, Irvin E. Searching for grey literature for systematic reviews: challenges and benefits. *Res Synth Methods.* 2014;5(3):221-234. https://doi.org/10.1002/jrsm.1106
46. Bar-Ilan J. On the overlap, the precision and estimated recall of search engines. A case study of the query "Erdos". *Scientometrics.* 1998;42(2):207-228. https://doi.org/10.1007/BF02458356
47. Kumar BTS, Prakash JN. Precision and relative recall of search engines: a comparative study of Google and Yahoo. *Singapore Journal of Library and Information Management*. 2009;38:124-137.
48. Shafi SM, Rather R. Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. *Webology.* 2005;2(2):1-7.
49. Usmani TA, Pant D, Bhatt AK. A comparative study of Google and Bing search engines in context of precision and relative recall parameter. *International Journal on Computer Science and Engineering (IJCSE).* 2012;4(1):21-34.
50. Giustini D, Boulos MNK. Google Scholar is not enough to be used alone for systematic reviews. *Online J Public Health Inform.* 2013;5(2):1-10. https://doi.org/10.5210/ojphi.v5i2.4623
51. Bramer WM. Variation in number of hits for complex searches in Google Scholar. *Journal of the Medical Library Association.* 2016;104(2):143-145. https://doi.org/10.3163/1536-5050.104.2.009
52. Brophy J, Bawden D. Is Google enough? Comparison of an internet search engine with academic library resources. *Aslib Proceedings.* 2005;57(6):498-512. https://doi.org/10.1108/00012530510634235
53. Gehanno J-F, Rollin L, Darmoni S. Is the coverage of Google Scholar enough to be used alone for systematic reviews. *BMC Medical Informatics and Decision Making.* 2013;13(7):1-5.
54. Sturm B, Sunyaev A. If you want your research done right, do you have to do it all yourself? Developing design principles for systematic literature search systems. In: Maedche A, Vom Brocke J, Hevner A, eds. *Designing the Digital Transformation*; 2017:138–146.
55. Chu H, Rosenthal M. Search engines for the World Wide Web: a comparative study and evaluation methodology. *J. Am. Soc. Inf. Sci.* 1996;33:127-135.
56. Biolcati-Rinaldi F, Molteni F, Salini S. Assessing the reliability and validity of Google Scholar indicators. The case of social sciences in Italy. In: Bonaccorsi A, ed. *The Evaluation of Research in Social Sciences and Humanities: Lessons from the Italian Experience*. Cham: Springer International Publishing; 2018:295-319.
57. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *Plos One.* 2015;10(9):1-17. https://doi.org/10.1371/journal.pone.0138237
58. Aune D, Giovannucci E, Boffetta P, et al. Fruit and vegetable intake and the risk of cardiovascular disease, total cancer and all-cause mortality—a systematic review and dose–response meta-analysis of prospective studies. *International Journal of Epidemiology.* 2017;46(3):1029-1056. https://doi.org/10.1093/ije/dyw319
59. Barnett DW, Barnett A, Nathan A, van Cauwenberg J, Cerin E. Built environmental correlates of older adults' total physical activity and walking: a systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity.* 2017;14(1):1-24. https://doi.org/10.1186/s12966-017-0558-z
60. Baur D, Gladstone BP, Burkert F, et al. Effect of antibiotic stewardship on the incidence of infection and colonisation with antibiotic-resistant bacteria and *Clostridium difficile* infection: a systematic review and meta-analysis. *Lancet Infectious Diseases.* 2017;17(9):990-1001. https://doi.org/10.1016/S1473-3099(17)30325-0
61. Bediou B, Adams DM, Mayer RE, Tipton E, Green CS, Bavelier D. Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin.* 2018;144(1):77-110. https://doi.org/10.1037/bul0000130
62. Bethel MA, Patel RA, Merrill P, et al. Cardiovascular outcomes with glucagon-like peptide-1 receptor agonists in patients with type 2 diabetes: a meta-analysis. *Lancet Diabetes & Endocrinology.* 2018;6(2):105-113. https://doi.org/10.1016/S2213-8587(17)30412-6
63. Bourne RRA, Flaxman SR, Braithwaite T, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a

systematic review and meta-analysis. *Lancet Global Health.* 2017;5(9):E888-E897. https://doi.org/10.1016/S2214-109X(17)30293-0

64. Brunoni AR, Chaimani A, Moffa AH, et al. Repetitive transcranial magnetic stimulation for the acute treatment of major depressive episodes a systematic review with network meta-analysis. *Jama Psychiatry.* 2017;74(2):143-152. https://doi.org/10.1001/jamapsychiatry.2016.3644
65. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlof E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cognitive Behaviour Therapy.* 2018;47(1):1-18. https://doi.org/10.1080/16506073.2017.1401115
66. Chu DK, Kim LH-Y, Young PJ, et al. Mortality and morbidity in acutely ill adults treated with liberal versus conservative oxygen therapy (IOTA): a systematic review and meta-analysis. *Lancet.* 2018;391(10131):1693-1705. https://doi.org/10.1016/S0140-6736(18)30479-3
67. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet.* 2018;391(10128):1357-1366. https://doi.org/10.1016/S0140-6736(17)32802-7
68. Stacey D, Légaré F, Lewis K, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev.* 2017;4:1-343. https://doi.org/10.1002/14651858.CD001431.pub5
69. *Oxford Wordlist*. Oxford University Press; 2008.
70. Fieschi M, Coiera E, Li YCJ. *Medinfo*: IOS Press; 2004. https://books.google.at/books?id=bS2xdt7iufgC.
71. Meho LI, Yang K. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *J. Am. Soc. Inf. Sci.* 2007;58(13):2105-2125. https://doi.org/10.1002/asi.20677
72. Martín-Martín A, Orduna-Malea E, Thelwall M, López-Cózar ED. Google Scholar, Web of Science, and Scopus: a systematic comparison of citations in 252 subject categories. *Journal of Informetrics.* 2018;12(4):1160-1177. https://doi.org/10.31235/osf.io/42nkm
73. Bakkalbasi N, Bauer K, Glover J, Wang L. Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomed Digit Libr.* 2006;3(7):1-8. https://doi.org/10.1186/1742-5581-3-7
74. Haddaway NR, Macura B, Whaley P, Pullin AS. ROSES RepOrting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environ Evid.* 2018;7(7):1-8. https://doi.org/10.1186/s13750-018-0121-7
75. White RW, Roth RA. *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool; 2009.
76. Jansen BJ, Spink A. In: Langendoerfer P, Droegehorn O, eds. IC'2003*An Analysis of Web Documents Retrieved and Viewed*; 2003:65-69.
77. Jansen BJ, Spink A. How are we searching the World Wide Web?: a comparison of nine search engine transaction logs. *Information Processing & Management.* 2006;42(1):248-263. https://doi.org/10.1016/j.ipm.2004.10.007
78. Athukorala K, Głowacka D, Jacucci G, Oulasvirta A, Vreeken J. Is exploratory search different?: a comparison of information search behavior for exploratory and lookup tasks. *J Assn Inf Sci Tec.* 2016;67(11):2635-2651. https://doi.org/10.1002/asi.23617
79. Hemminger BM, Lu D, Vaughan KTL, Adams SJ. Information seeking behavior of academic scientists. *J. Am. Soc. Inf. Sci.* 2007;58(14):2205-2225. https://doi.org/10.1002/asi.20686
80. Athukorala K, Hoggan E, Lehtiö A, Ruotsalo T, Jacucci G. Information-seeking behaviors of computer scientists: challenges for electronic literature search tools. *Proc. Am. Soc. Info. Sci. Tech.* 2013;50(1):1-11. https://doi.org/10.1002/meet.14505001041
81. Nicholas D, Boukacem-Zeghmouri C, Rodríguez-Bravo B, et al. Where and how early career researchers find scholarly information. *Learned Publishing.* 2017;30(1):19-29. https://doi.org/10.1002/leap.1087
82. Niu X, Hemminger BM. A study of factors that affect the information-seeking behavior of academic scientists. *J. Am. Soc. Inf. Sci.* 2012;63(2):336-353. https://doi.org/10.1002/asi.21669
83. Sapa R, Krakowska M, Janiak M. Information seeking behaviour of mathematicians: scientists and students. *Information Research: An International Electronic Journal*. 2014;19(4):1-11.
84. Fast KV, Campbell DG. "I still like Google": University student perceptions of searching OPACs and the web. *Proceedings of the American Society for Information Science and Technology.* 2004;41(1):138-146. https://doi.org/10.1002/meet.1450410116
85. Kuiper E, Volman M, Terwel J. Students' use of Web literacy skills and strategies: searching, reading and evaluating Web information. *Information Research.* 2008;13(3):1-18.
86. Ingwersen P. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation.* 1996;52(1):3-50. https://doi.org/10.1108/eb026960
87. Wellings S, Casselden B. An exploration into the information-seeking behaviours of engineers and scientists. *Journal of Librarianship and Information Science.* 2017;9(2):1-12. https://doi.org/10.1177/0961000617742466
88. Rowlands I, Nicholas D, Williams P, et al. The Google generation: the information behaviour of the researcher of the future. *AP.* 2008;60(4):290-310. https://doi.org/10.1108/00012530810887953
89. Kingsley K, Galbraith GM, Herring M, Stowers E, Stewart T, Kingsley KV. Why not just Google it? An assessment of information literacy skills in a biomedical science curriculum. *BMC Med Educ.* 2011;11(17):1-8. https://doi.org/10.1186/1472-6920-11-17
90. Kurbanoğlu S, Boustany J, Špiranec S, Grassian E, Mizrachi D, Roy L, eds. *Information literacy: moving toward sustainability: Third European conference, ECIL 2015, Tallinn, Estonia, October 19–22, 2015: revised selected papers.* Cham, Heidelberg, New York: Springer; 2015. Communications in computer and information science; 552.
91. Kurbanoğlu S, Boustany J, Špiranec S, et al. (Eds). *Search Engine Literacy: Information Literacy in the*. Workplace: Springer International Publishing; 2018.
92. Brindesi H, Monopoli M, Kapidakis S. Information seeking and searching habits of Greek physicists and astronomers: a case study of undergraduate students. *Procedia - Social and*

Behavioral Sciences. 2013;73:785-793. https://doi.org/10.1016/j.sbspro.2013.02.119

93. Kirschner PA, de Bruyckere P. The myths of the digital native and the multitasker. *Teaching and Teacher Education*. 2017;67:135-142. https://doi.org/10.1016/j.tate.2017.06.001
94. Georgas H. Google vs. the library (part II): student search patterns and behaviors when using Google and a federated search tool. *Portal: Libraries and the Academy*. 2014;14(4):503-532.
95. Halevi G, Moed H, Bar-Ilan J. Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation: review of the literature. *Journal of Informetrics*. 2017;11(3):823-834. https://doi.org/10.1016/j.joi.2017.06.005

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Res Syn Meth*. 2020;11:181–217. https://doi.org/10.1002/jrsm.1378