

Wesentliche neue EbM-relevante Methoden 2024 bis 2026 – Informationsbeschaffung

A decorative horizontal bar spanning the width of the page, composed of 18 small squares in various shades of blue and grey. The word 'ARBEITSPAPIER' is centered in white text on a dark blue rectangular background that covers the middle portion of this bar.

ARBEITSPAPIER

Projekt: GA26-05

Version: 1.0

Stand: 15.09.2026

IQWiG-Berichte – Nr. 2342

DOI: 10.60584/GA26-05

Impressum

Herausgeber

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Thema

Wesentliche neue EbM-relevante Methoden 2024 bis 2026 – Informationsbeschaffung

Auftraggeber

Bearbeitung im Rahmen des Generalauftrags

Datum des Auftrags

21.03.2026

Interne Projektnummer

GA26-05

DOI-URL

<https://doi.org/10.60584/GA26-05>

Anschrift des Herausgebers

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Siegburger Str. 237
50679 Köln

Tel.: +49 221 35685-0

Fax: +49 221 35685-1

E-Mail: info@iqwig.de

Internet: www.iqwig.de

ISSN: 1864-2500

Zitiervorschlag

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Wesentliche neue EbM-relevante Methoden 2024 bis 2026 – Informationsbeschaffung; Arbeitspapier [online]. 2026 [Zugriff: TT.MM.JJJJ]. URL: <https://doi.org/10.60584/GA26-05>.

Schlagwörter

Informationsspeicherung und -Retrieval, Übersichtsliteratur, Evidenzsynthese

Keywords

Information Storage and Retrieval, Review Literature as Topic, Evidence Synthesis

Für die Inhalte des Berichts ist allein das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) verantwortlich.

Mitarbeiterinnen und Mitarbeiter des IQWiG

- Tatjana Hermanns
- Annalena Dunkel
- Elke Hausner
- Simone Heß
- Marco Knelangen
- Ulrike Lampert
- Claudia Lenkewitz
- Anne-Kathrin Petri
- Kristin Rütjes
- Dorothea Sow
- Siw Waffenschmidt

Inhaltsverzeichnis

	Seite
Tabellenverzeichnis	v
Abkürzungsverzeichnis.....	vi
1 Hintergrund.....	1
2 Fragestellung.....	2
3 Projektverlauf.....	3
4 Methoden	4
5 Ergebnisse	6
5.1 Allgemeines.....	6
5.1.1 Reporting Bias (einschließlich Publikationsbias)	7
5.1.2 Expertise	8
5.1.3 Publizieren	8
5.1.3.1 Wissenschaftliches Fehlverhalten	8
5.1.3.2 Paper Mills	9
5.1.3.3 Nicht vertrauenswürdige Studien.....	9
5.2 Automatisieren und Digitalisieren	10
5.2.1 LLM Studienselektion	11
5.2.2 Suche	12
5.2.3 Stopp-Kriterien	13
5.3 Umfassende Informationsbeschaffung.....	13
5.3.1 Bibliografische Datenbanken	13
5.3.1.1 Datenbankauswahl	14
5.3.1.2 Suchfilter	15
5.3.2 Studienregister	16
5.3.3 Weitere Informationsquellen und Suchtechniken	18
6 Literatur	21
Anhang A Übersicht aktueller Suchfilter	27
Anhang B Suchquellen und Suchstrategien	28

Tabellenverzeichnis

	Seite
Tabelle 1: Übersicht über die Kriterien für den Einschluss von Literatur zum Thema Informationsbeschaffung	4
Tabelle 2: Kurzbewertung der Filter.....	27

Abkürzungsverzeichnis

Abkürzung	Bedeutung
AHRQ	Agency for Healthcare and Quality
AI	Artificial Intelligence (künstliche Intelligenz)
ALA	American Library Association
API	Application Programming Interface
CEE	Collaboration for Environmental Evidence
CESAR	Cochrane Evaluation of (Semi-)Automated Review Methods
CSR	Clinical Study Report (klinischer Studienbericht)
CT.gov	ClinicalTrials.gov
EbM	evidenzbasierten Medizin
EGM	Evidence and Gap Map
G-BA	Gemeinsamer Bundesausschuss
GRADE	Grading of Recommendations Assessment, Development and Evaluation (System zur Bewertung der Vertrauenswürdigkeit der Evidenz und der Stärke von Empfehlungen)
HTAi	Health Technology Assessment International
InterTASC	InterTASC Information Specialists Sub-Group
IPD	individuelle Patientendaten
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
IRM	Information Retrieval Meetings
ISSG SFR	InterTASC Information Specialists' Sub-Group Search Filter Resource
JB	Joanna Briggs Institute
LLM	Large Language Model
MeSH	Medical Subject Headings (kontrolliertes medizinisches Schlagwortverzeichnis der U.S. National Library of Medicine)
NICE	National Institute for Health and Care Excellence
NNR	Numbers Needed to Read
OSM	Open-Source Modelle
PRO	Patient-Reported Outcome (patientenberichtetes Endpunkt- bzw. Behandlungsergebnis)
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies 2 (Instrument zur Bewertung des Verzerrungsrisikos diagnostischer Genauigkeitsstudien)
RAISE	Responsible use of AI in evidence SynthEsis

Abkürzung	Bedeutung
ROBINS-1	RoB Assessments for Cochrane 1
RCT	Randomized Controlled Trial (randomisierte kontrollierte Studie)
SAP	statistischer Analyseplan
SBU	Statens beredning för medicinsk och social utvärdering (Schwedische HTA-Agentur)
SR	Systematic Review (systematische Übersicht)
SuRe Info	Summarized Research in Information Retrieval for HTA
WoS	Web of Science

1 Hintergrund

Im Kontext der evidenzbasierten Medizin (EbM) bezeichnet die Informationsbeschaffung den systematischen Prozess der gezielten Suche und Auswahl relevanter wissenschaftlicher Studien und Daten. Ziel ist es, die beste verfügbare Evidenz zuverlässig zu identifizieren, um fundierte gesundheitsbezogene Entscheidungen zu ermöglichen. Dabei sind u. a. die Auswahl geeigneter Suchquellen, strukturierte Suchstrategien und geeignete Suchfilter, sowie die strukturierte Aufbereitung der Suchergebnisse und eine transparente Selektion von Studien von zentraler Bedeutung, um qualitativ hochwertige und aktuelle Informationen sicherzustellen.

Im Bereich der Informationsbeschaffung entwickeln sich Technologien und Methoden kontinuierlich weiter, aktuell insbesondere durch Fortschritte in Künstlicher Intelligenz (AI) und Datenverarbeitung. Die kontinuierliche Beobachtung dieser Entwicklungen ist entscheidend, um die Methoden des Instituts [1] zu verbessern und weiterzuentwickeln. Nur so können aktuelle Anforderungen erfüllt und innovative Lösungsansätze frühzeitig erkannt werden. Vor diesem Hintergrund führt das Ressort Informationsmanagement des Instituts ein kontinuierliches Literaturscreening durch, um sich über neue Methoden und Entwicklungen zum Thema „Informationsbeschaffung“ zu informieren.

2026 werden die wesentlichen Publikationen zu aktuellen methodischen Entwicklungen zum Thema „Informationsbeschaffung“ erstmals in einem Arbeitspapier aufbereitet und veröffentlicht.

2 Fragestellung

Ziel des vorliegenden Projektes ist die jährliche Aufarbeitung der Literatur zu EbM-relevanten Methoden zum Thema „Informationsbeschaffung“.

3 Projektverlauf

Auf Basis einer internen Projektskizze wurde ein Arbeitspapier erstellt. Dieser Bericht wurde an den G-BA übermittelt und 4 Wochen später auf der Website des IQWiG veröffentlicht.

4 Methoden

Die Erfassung von EbM-relevanter Methoden zu dem Thema „Informationsbeschaffung“ erfolgt fortlaufend im Ressort Informationsmanagement des Instituts. Dabei wurden folgende Suchquellen berücksichtigt und in einer Endnote Datenbank aufbereitet (siehe Anhang B für detaillierte Angaben zu den Suchquellen und Suchstrategien):

- Suche in der bibliografischen Datenbank MEDLINE (regelmäßiger AutoAlert)
- Sichtung von relevanten Fachzeitschriften
- Sichtung von relevanten Kongressen
- Sichtung von speziellen Websites und Methodendokumenten ausgewählter Organisationen
- optionale Sichtung von Referenzlisten relevanter Artikel
- Berücksichtigung anderweitig erfasster Literatur

Diese Literatursammlung bildete die Grundlage für die jährliche Aufbereitung.

Bis April 2026 wurden alle neu erfassten Referenzen mit einem Publikationsdatum innerhalb der vorangegangenen 24 Monate thematisch gruppiert, sofern sie seit dem letzten jährlichen Stichtag erfasst und im letzten Literaturscreening noch nicht berücksichtigt worden waren. Die erfasste Literatur umfasste folgende Themenbereiche: Allgemeines (inklusive Reporting Bias, Expertise und Publizieren), Automatisieren und Digitalisieren sowie verschiedene Aspekte der umfassenden Informationsbeschaffung wie bibliografische Datenbanken, Herstellerunterlagen, Studienregister und weitere Suchquellen, ergänzt durch fokussierte Recherche und Studienfilter. Unberücksichtigt blieben Publikationen aus den Vorjahren sowie Kongressbeiträge zu denen lediglich ein Abstract, ein Poster oder eine Präsentation vorlag.

In jedem Themenbereich wurden die Publikationen gesichtet und auf ihre Relevanz bezüglich der Einschlusskriterien (Tabelle 1) geprüft.

Tabelle 1: Übersicht über die Kriterien für den Einschluss von Literatur zum Thema Informationsbeschaffung

Einschlusskriterien	
E1	Die Publikation ist thematisch der Informationsbeschaffung zuzuordnen.
E2	datengestützte Aufarbeitung einer methodischen Fragestellung oder methodisches Dokument jeweils von besonderer Bedeutung für die Arbeit des IQWiG
E3	Publikationsdatum innerhalb der letzten 24 Monate und wurde nicht im Literaturscreening der vergangenen Jahre gesichtet
E4	Vollpublikation verfügbar

Die so ausgewählten Publikationen wurden inhaltlich knapp zusammengefasst. Wenn erforderlich wurde die Relevanz für das Institut eingeordnet bzw. dargestellt und im Arbeitspapier berichtet. Alle Auswahlsschritte und die Zusammenfassungen erfolgten durch 1 Person und wurden durch eine 2. Person überprüft. Diskrepanzen wurden durch Diskussion zwischen beiden aufgelöst.

5 Ergebnisse

5.1 Allgemeines

Hausner et al. [2] beschäftigten sich damit, ob die Einbeziehung nicht englischsprachiger Publikationen die Ergebnisse und Schlussfolgerungen von IQWiG Berichten wesentlich verändert. Für die Auswertung wurden 70 IQWiG-Berichte (2007 bis 2018) berücksichtigt, 67 zu nicht medikamentösen Verfahren und 3 Arzneimittel-Berichte. Die Autorinnen und Autoren stellten fest, dass für 38 Berichte (54 %) nicht englischsprachige Publikationen eingeschlossen wurden, sich aber nur in 4 Berichten (6 %) durch die Nichtberücksichtigung dieser Publikationen das Fazit geändert hätte. Dies war dann der Fall, wenn wenig Daten vorlagen bzw. ausschließlich Studien in anderen Sprachen als Englisch eingeschlossen wurden. Die Auswertung bestätigt das Vorgehen des IQWiG, nicht englisch- oder deutschsprachige Literatur nur in Ausnahmefällen, z. B. bei bestimmten Fragestellungen, zu berücksichtigen.

Paynter et al. [3] haben ein White Paper des AHRQ für Tools zur Entwicklung von Suchstrategien veröffentlicht. Hierfür sichteten die Autorinnen und Autoren die Tools der Systematic Review Toolbox (<https://hslib.jabsom.hawaii.edu/systematicreview>) und wählten anhand vorab definierter Kriterien 21 von 102 Tools aus. Jedes der 21 eingeschlossenen Tools wurde einem Anwendungsbereich zugeordnet: Scoping und Testsets identifizieren, Schlüsselbegriffe und Synonyme finden, kontrolliertes Vokabular (z. B. MeSH) identifizieren, Boolesche Suchlogik entwickeln, Suchstrategien evaluieren, Suchstrategien zwischen Datenbanken übersetzen, Graue Literatur durchsuchen, Dubletten erkennen und entfernen. Zusätzlich wurden die wichtigsten Hintergrundinformationen und Charakteristika zu dem jeweiligen Tool beschrieben. Die Autorinnen und Autoren stellten fest, dass kein Tool den gesamten Prozess abdecken kann und dass gerade schwierige Aspekte der Suchstrategieentwicklung, wie etwa die Entwicklung des Suchkonzeptes, immer noch vorrangig durch Informationsspezialistinnen und -spezialisten erfolgen müsse.

Cooper et al. [4] skizzieren einen neuen Ansatz für die Informationsbeschaffung von Evidenzsynthesen. Sie bemerken zu Recht, dass die Suche in bibliografischen Datenbanken immer noch die Hauptquelle der Informationsbeschaffung darstellt und die Einbindung weiterer Suchquellen oft unstrukturiert erfolgt. Sie schlagen vor einen 3-stufigen Prozess für die Informationsbeschaffung von randomisierten kontrollierten Studien (RCTs) zu etablieren:

- 1) Identifikation von Studien über Studienregister
- 2) Identifikation publizierter Dokumente sowie unregistrierter Studien in bibliografischen Datenbanken
- 3) Identifikation unpublizierter Dokumente und ggf. zusätzlicher Studien mittels ergänzender Suchtechniken, beispielsweise durch die Sichtung von Webseiten, Zitationssuchen oder die Auswertung von Studienberichten

Dieser Ansatz ordnet die verschiedenen Suchquellen in logische Arbeitsschritte und stellt zudem die Identifikation der Studie und nicht der Publikation zur Studie in den Mittelpunkt. Das Konzept ist noch nicht ausreichend erprobt, um sichere Aussagen zu den Vorteilen gegenüber dem herkömmlichen Verfahren zu treffen. Die Autorinnen und Autoren planen ihre Evaluationen zu veröffentlichen.

5.1.1 Reporting Bias (einschließlich Publikationsbias)

Für die Erstellung systematischer Übersichten (SRs) verdeutlichen 3 aktuelle Publikationen die Relevanz weiterer Suchquellen wie Studienregister, Studienprotokolle und Studienberichte zur Identifikation und Bewertung von Publikations- und Reporting-Bias.

Taji Heravi et al. [5] untersuchten für 488 RCTs die Übereinstimmung zwischen Studienprotokollen und Ergebnispublikationen von RCTs hinsichtlich patientenberichteter Endpunkte (PROs). Die Autorinnen und Autoren zeigten, dass die Berichterstattung von PROs häufig von den ursprünglichen Studienprotokollen abweicht. Nur 44 % der Ergebnispublikationen berichteten alle PROs entsprechend zugehöriger Protokolle, während 21 % keine der geplanten PROs berichteten und 36 % zusätzliche, weniger oder andere PROs berichteten. Darüber hinaus war die Qualität der PROs-Berichterstattung häufig unzureichend.

Dobrescu et al. [6] analysierten den Einfluss unveröffentlichter Daten auf die Ergebnisse einer Netzwerk-Metaanalyse am Beispiel von akuter Migräne. Hierfür führten sie eine ergänzende Recherche zur Identifikation von Studienregistereinträgen, Kongressabstracts und regulatorischen Dokumenten durch. Die Einbeziehung der dort verfügbaren Informationen führte zwar zu einer Erweiterung des Netzwerks, veränderte jedoch weder die Schlussfolgerungen noch die Sicherheit der Ergebnisse in relevantem Ausmaß. Studienregister erwiesen sich dabei als die wichtigste Quelle zusätzlicher Informationen, während andere Quellen nur einen begrenzten Mehrwert lieferten. Die Autorinnen und Autoren fordern weitere Untersuchungen, um zu klären, wann der zusätzliche Aufwand der Informationsbeschaffung tatsächlich einen relevanten Mehrwert bietet.

Ergänzend untersuchten Hirt et al. [7] anhand einer Metaanalyse zu Interventionen in der Peritonealdialyse den Einfluss von Studienberichten (Clinical Study Reports, CSRs) auf die Ergebnisse der Evidenzsynthese. Hierzu verglichen sie Metaanalysen auf Basis publizierter Daten mit denen, in die zusätzlich CSR-Daten einbezogen wurden. Die Berücksichtigung von CSRs erhöhte die Verfügbarkeit von Daten, insbesondere zu schwerwiegenden unerwünschten Ereignissen, führte jedoch zu keinen wesentlichen Änderungen der Gesamtergebnisse oder Schlussfolgerungen aus den Metaanalysen. Dennoch verbesserten die zusätzlichen CSR-Daten die Präzision der Effektschätzungen und lieferten eine vollständigere und verlässlichere Evidenzbasis, insbesondere für Endpunkte zur Sicherheit wie schwerwiegende unerwünschte Ereignisse. Die Autorinnen und Autoren heben hervor, dass

die strukturierte Aufbereitung von CSRs die Identifikation und Extraktion relevanter Studiendaten erleichtert und empfehlen, diese Dokumente bei Evidenzsynthesen systematisch zu berücksichtigen.

5.1.2 Expertise

Harper et al. [8] vergleichen die Kernkompetenzen der American Library Association (ALA) mit den von der AI Alliance (<https://thealliance.ai/>) definierten AI-Kompetenzen. Dabei zeigt sich eine deutliche Übereinstimmung, wie z. B. „Bewertung von Datenqualität und Datenbeschränkungen“ oder „Datenanalyse und Informationsorganisation“. Die Autorinnen und Autoren argumentieren, dass Bibliotheksmitarbeitende für die Entwicklung, Verwaltung und Bewertung von AI-Systemen zunehmend wichtig werden. Viele der Fähigkeiten, die heute in der AI-Branche benötigt werden – Informationsorganisation, Metadatenmanagement, Datenbewertung, Ethik und Wissensstrukturierung – gehören bereits zum Kernprofil des Bibliotheks- und Informationswesens.

5.1.3 Publizieren

5.1.3.1 Wissenschaftliches Fehlverhalten

Der Artikel von Hegedus et al. [9] befasst sich mit dem Problem sogenannter „hijacked journals“. Dabei werden die Webseiten von Online-Zeitschriften „nachgebaut“, indem Logos und Webseitenstruktur einer bekannten Zeitschrift kopiert werden. Forschende werden dazu verleitet, Manuskripte einzureichen und Publikationsgebühren zu zahlen, ohne dass ein Peer Review stattfindet. Die Autoren zeigen, dass einige dieser gefälschten Zeitschriften durch gezielte Suchmaschinenoptimierung bei Suchanfragen sehr weit oben erscheinen – teilweise sogar vor seriösen wissenschaftlichen Quellen. Sie stellen dadurch eine wachsende Gefahr für Forschung, wissenschaftliche Integrität und die Qualität digitaler Informationssysteme dar.

Spick et al. [10] untersuchten den Missbrauch frei verfügbarer Datenquellen. Über eine scientometrische Analyse (quantitative Auswertung wissenschaftlicher Publikationen und ihrer Metadaten) stellten sie in OpenAlex ein überproportionales Wachstum an Publikationen fest, die 9 bekannte frei verfügbare Datenquellen (z. B. UK Biobank) als Datenbasis nutzten. Hier wird vermutet, dass es sich häufig um nicht vertrauenswürdige Studien bzw. um Missbrauch handelt (z. B. über Paper Mills). Die Autorinnen und Autoren betonen, dass nicht die frei verfügbaren Daten das Problem sind, sondern das ausbeuterische Ausnutzen von Open Science. Die Autorinnen und Autoren sehen darin eine ernstzunehmende Herausforderung für die Qualität und Vertrauenswürdigkeit der biomedizinischen Forschung und schlagen vor, zukünftig den Zugang besser zu kontrollieren.

5.1.3.2 Paper Mills

Scancar et al. [11] untersuchten, wie verbreitet von Paper Mills erzeugte wissenschaftliche Publikationen in der Krebsforschung sind. Laut dem Committee on Publication Ethics [12] sind Paper Mills Organisationen oder Unternehmen, die gegen Bezahlung wissenschaftliche Publikationen erstellen, indem sie den Publikationsprozess systematisch manipulieren. Dabei werden unterschiedliche Praktiken angewendet, beispielsweise die Verwendung erfundener oder manipulierter Daten. Die Autorinnen und Autoren entwickelten und validierten ein Verfahren auf Basis von Machine Learning, um typische sprachliche und strukturelle Merkmale bekannter Paper-Mill-Artikel zu erkennen. Das Modell wurde anhand von 2202 zurückgezogenen Paper-Mill-Artikeln trainiert und anhand unabhängiger Daten validiert. Das Modell erreicht eine Genauigkeit von 91 %. Bei der Anwendung auf die Krebsforschungsliteratur kennzeichnete es 261 245 von 2 647 471 Artikeln (10 %) und zeigte einen starken Anstieg der gekennzeichneten Artikel von 1999 bis 2024. Mehr als 170 000 Artikel, die chinesischen Institutionen zugeordnet waren, wurden markiert, was 36 % der chinesischen Krebsforschungsartikel ausmachte. Die Studie liefert Hinweise darauf, dass ein erheblicher Anteil der Krebsforschungsliteratur Merkmale aufweisen könnte, die auf Paper-Mill-Aktivitäten hindeuten. Sie zeigt zugleich, dass AI-gestützte Verfahren ein nützliches Werkzeug zur Qualitätssicherung wissenschaftlicher Publikationen sein können.

Tang et al. [13] untersuchten in welchem Ausmaß zurückgezogene Artikel aus Paper Mills in SRs eingeflossen sind. Obwohl deren Anteil derzeit noch gering ist (0,15 % in einem Sample von 200 000 SRs), stellen sie ein reales und zunehmendes Risiko dar.

5.1.3.3 Nicht vertrauenswürdige Studien

Der Artikel von Papapanou et al. [14] vergleicht 7 Instrumente (REAPPRAISED, TRACT, CPC-TST, RIA, IPD-Integrity, OGEIG, INSPECT-SR) zur Beurteilung der Vertrauenswürdigkeit von RCTs. Diese unterscheiden sich hinsichtlich Umfang, Kriterien und Anwendungsweise. Alle Instrumente berücksichtigen Aspekte wie Ethik, prospektive Studienregistrierung, Baseline-Daten und Plausibilität der Follow-up Daten. Auffällige Muster in Datensätzen können Hinweise auf Fehler oder wissenschaftliches Fehlverhalten liefern. Die Autorinnen und Autoren plädieren deshalb dafür, vor der klassischen Bewertung des Verzerrungspotenzials zunächst zu prüfen, ob eine Studie überhaupt vertrauenswürdig ist. Sie schlagen einen strukturierten Ansatz mit 6 Kernprüfungen vor, um potenziell problematische Studien frühzeitig zu erkennen:

- 1) Prüfung vor der Publikation, wie Studienregistrierung und Ethikvotum,
- 2) Prüfung nach der Publikation, wie Meldung von Auffälligkeiten an Zeitschriften oder Plattformen wie PubPeer sowie Nachverfolgung von Korrekturen und Retractions,

- 3) Prüfung in SRs / Metaanalysen, wie vertrauenswürdiges Screening vor der Datenextraktion sowie Ausschluss von Studien mit schwerwiegenden oder ungeklärten Problemen,
- 4) Prüfung auf Ebene der individuellen Patienten Daten (IPD), wie Analyse der Rohdaten sowie Suche nach ungewöhnlichen Datenmustern oder statistischen Auffälligkeiten,
- 5) Prüfung in Umbrella Reviews, wie Kontrolle, ob die eingeschlossenen SRs bereits Vertrauenswürdigkeitsprüfungen durchgeführt haben, sowie ggf. erneute Bewertung der Primärstudien,
- 6) Prüfung bei Leitlinienentwicklung, wie Bewertung der Vertrauenswürdigkeit der Evidenz vor der Formulierung von Empfehlungen.

Weibel et al. [15] betrachteten anhand des Research Integrity Assessment (RIA)-Tools die Vertrauenswürdigkeit von Studien zu COVID-19 (ausgewertet wurden 23 SRs). Von den geprüften Studien wurden nur 28 % als völlig unbedenklich eingeschätzt, 29 % wurden aufgrund mangelnder Integrität ausgeschlossen. 38 % wurden in eine Wartekategorie einsortiert, da Angaben zum Ethikvotum unklar waren. Die Autorinnen und Autoren kommen zu dem Fazit, dass eine Prüfung der Integrität mithilfe eines Tools notwendig ist, jedoch auch Zeit und Ressourcen erfordert. Allein das Vorliegen einer Retraction als Kriterium für die Einstufung einer Studie als nicht vertrauenswürdig halten sie für nicht ausreichend. Die Anzahl der eingeschlossenen Studien sinkt bei der Verwendung eines solchen Tools deutlich und führt dadurch auch zu einem Verlust an Präzision von Metaanalysen. Die statistische Signifikanz der Ergebnisse hingegen ändert sich nur selten.

Wilkinson et al. [16-18] beschreiben in insgesamt 3 Artikeln das Tool INSPECT-SR bzw. dessen Entwicklung. Das Tool kann zur Identifikation problematischer RCTs im Rahmen der Erstellung von SRs eingesetzt werden. Die Autorinnen und Autoren gehen davon aus, dass etablierte Qualitätsprüfungsmechanismen wie GRADE oder Risk of Bias manipulierte / nicht vertrauenswürdige Studien häufig nicht erkennen. Deshalb haben sie Kriterien entwickelt, die explizit zur Prüfung der Vertrauenswürdigkeit herangezogen werden können. Es handelt sich um insgesamt 21 Kriterien, die 4 verschiedenen Domänen (Überprüfung 1. auf Hinweise nach einer Veröffentlichung, z. B. Retractions, 2. von Durchführung, Governance und Transparenz, 3. von Text und Abbildungen und 4. der Ergebnisse der Studie) zugeordnet werden. Wilkinson et al. betonen, dass Fachwissen und Erfahrung für die Einschätzung der einzelnen Kriterien erforderlich sind.

5.2 Automatisieren und Digitalisieren

Die Automatisierung durch AI-Tools verspricht insbesondere im Bereich der Evidenzsynthesen erhebliche Zeit- und Ressourceneinsparungen. Dennoch gibt es bei Forschenden Bedenken, diese Tools in der Praxis einzusetzen. O'Keefe et al. [19] haben diese Bedenken im Rahmen

von 2 Präsenz- sowie 2 Onlineveranstaltungen und einem Online Horizon Scanning Survey untersucht. Insgesamt konnten 10 übergreifende Themen identifiziert werden. Bedenken hinsichtlich des eigenen Wissens und Fähigkeiten (24 %) wurde am häufigsten genannt; die anderen Themen waren Datenmanagement, Transparenz, Rechtmäßigkeit und Gerechtigkeit, Qualität der AI, Umweltauswirkungen, unabhängiges Denken, Nutzung und Rechenschaftspflicht, Wirtschaftlichkeit und AI-Markt.

In einem Positionspapier von Cochrane, der Campbell Collaboration, Joanna Briggs Institute (JBI) und der Collaboration for Environmental Evidence (CEE) befürworten die Autorinnen und Autoren [20] einen verantwortungsvollen Einsatz von AI bei der Evidenzsynthese. Sie stützen sich hierbei vorrangig auf die RAISE-Guidance (Responsible use of AI in evidence Synthesis) [21]. Zentrale Elemente für einen solchen Einsatz sind aus Sicht der Autorinnen und Autoren unter anderem die vollständige Transparenz und Dokumentation, die menschliche Aufsicht über die AI und dass letztendlich die Verantwortung für alle Inhalte und Entscheidungen in der Evidenzsynthese bei den jeweiligen Autorinnen und Autoren liegt.

5.2.1 LLM Studienselektion

Adam et al. [22] stellen in ihrem 2. Surveillance-Bericht ein Update (letzte Suche am 16.10.2025) hinsichtlich neuer Studien dar, die sich seit der letzten Recherche mit Tools zur (Teil-)Automatisierung von Evidenzsynthesen befassen. Insgesamt konnten 39 neue Studien eingeschlossen werden. Nach aktuellem Stand sind einige AI-Tools für bestimmte Aufgaben und unter menschlicher Aufsicht einsatzbereit. Hierzu zählen die Autorinnen und Autoren:

- 1) die Identifizierung von RCTs,
- 2) das halbautomatisierte Abstract-Screening,
- 3) das automatisierte Volltextscreening,
- 4) Datenextraktion von Studiencharakteristika und Bewertung des Verzerrungspotenzials nach RoB assessments for Cochrane 1 (ROBINS-1), aber nicht nach ROBINS-2.

Laignelot et al. [23] erstellten eine Übersichtsarbeit (Suchdatum 14.01.2025) mit dem Ziel die Leistungsfähigkeit von Large Language Models (LLMs) bei der Automatisierung einiger oder aller Schritte von SRs und Metaanalysen im Bereich der biomedizinischen Forschung zusammenzufassen. 63 Studien wurden eingeschlossen, davon 52 quantitative Studien. Die Studien deckten mit den untersuchten Reviewthemen ein breites Spektrum medizinischer Fachrichtungen ab. Für 57 % (n = 30) der eingeschlossenen Studien war das Verzerrungspotenzial nach adaptierter QUADAS-2-Bewertung unklar. 22 (42 %) der eingeschlossenen quantitativen Studien untersuchten Title-Abstract-Screening und jeweils 4 (8 %) Studien untersuchten Volltextscreening bzw. Suchstrategieerstellung mithilfe von LLMs. Das am häufigsten untersuchte Modell war GPT-4 von OpenAI (n = 28; 54 %). Für das Abstract-

Screening betrug der Median der globalen Sensitivität 92 % (IQR: 69–98 %; lt. Publikation „Positive Percent Agreement“) und der globalen Spezifität 89 % (IQR: 72–95 %; lt. Publikation „Negative Percent Agreement“). Verschiedene Subgruppenanalysen legen nahe, dass es Unterschiede geben könnte, die sich auf die Performance auswirken (u. a. ältere versus neuere LLMs, API versus web-basierter Chat).

Cochrane legt das Protokoll für eine adaptive Plattform-Studie CESAR (Cochrane Evaluation of (Semi-)Automated Review Methods) vor [24]. Darin streben sie an, AI-Tools anhand von etwa 15 Cochrane-Review Aktualisierungen zu testen und mit den Standard-Methoden zu vergleichen. Ziel des Studiendesigns ist es, mehrere AI-Tools parallel unter einem einzigen Protokoll zu bewerten und diese aber auch abubrechen bzw. mit anderen Tools zu ersetzen, wenn bestimmte Mindestwerte in der Performance nicht erreicht werden. Hierfür wird für das Abstract-Screening eine Punktschätzung der Sensitivität von mindestens 80 % und für die obere Grenze des 95 %-Konfidenzintervalls mindestens 95 % angesetzt. Unter anderem sieht das Protokoll vor, die Antwortstabilität der Systeme mittels Jaccard Similarity zu untersuchen. Laut Cochrane wurden als erste AI-Tools NestedKnowledge und LASER AI ausgewählt. Zahlreiche weitere Publikationen befassen sich derzeit zunehmend mit agentischen Workflows im Abstract-Screening, um mit der schwankenden Antwort-Stabilität beim Einsatz eines einzelnen LLMs im Abstract-Screening umzugehen (u. a. [25-27]).

5.2.2 Suche

Featherstone et al. [28] vergleichen anhand von 7 abgeschlossenen Projekten zu unterschiedlichen Gesundheitsthemen verschiedene Tools u. a. Microsoft Copilot zur Suchstrategieerstellung. Sie kommen zu dem Schluss, dass Microsoft Copilot bei der Identifikation von Suchbegriffen unterstützen könnte. Featherstone et al. weisen jedoch auch darauf hin, dass dieses Vorgehen nur von Suchexpertinnen und -experten angewendet werden sollten, um Fehleinschätzungen zu vermeiden und im Vorfeld sorgfältig überlegt werden sollte, für welche Projekte ein tatsächlicher Nutzen des Einsatzes zu erwarten ist.

Wang et al. [29] legen eine Reproduktionsstudie und Erweiterung ihrer Studie von 2023 vor, indem sie noch weitere LLMs untersuchten. Die Sensitivität der LLM-generierten Suchstrategien bleibt auch in der Reproduktionsstudie deutlich hinter manuell erstellten Suchstrategien zurück, die mit 84 % angegeben wird. Die Autoren geben den Unterschied zwischen der durchschnittlichen Sensitivität zwischen manueller Erstellung und LLM-Generierung mit 15 % an. Keines der Modelle erreichte eine Sensitivität von > 70 %, und lag meistens sogar unter 50 %.

Ebenfalls von Wang et al. [30] wurde die AutoBool Studie durchgeführt. Darin präsentieren die Autoren Ergebnisse zu Basis-Modellen, die mittels Fine-Tuning (Training mit zusätzlichen Trainingsdaten) optimiert wurden. Der Trainings-Datensatz basiert auf 65 588 in PubMed

Central als Systematic Review verschlagworteten Volltexten und weiteren bereits publizierten Datensätzen (z. B. CLEF TAR 2017 und 2018). Dieser als AutoBool bezeichnete Trainingsprozess wurde auf die Open-Weights Modelle Qwen3-4B, Qwen3-8B und Llama3.1-8B angewendet, welche sich dadurch auszeichnen, dass sie deutlich kleiner sind als Modelle wie GPT-4o und z. B. auch lokal ausgeführt werden können. Die Studie zeigt, dass kleinere Modelle nach dem Fine-Tuning ähnliche Sensitivitäten erreichen, wie die großen Basis-Modelle (GPT-4o, GPT-O3) und diese teilweise sogar übertreffen. Eine Demoversion ist unter <https://huggingface.co/spaces/wshuai190/AutoBool-Demo> verfügbar (zuletzt geprüft am 25.06.2026).

Das Institute of Health Economics (IHE) [31] stellen ihre „Learnings“ und Überlegungen zu zahlreichen Anwendungsfällen für generative AI-Tools im Bereich der Evidenzsynthesen vor (u. a. Suchstrategieerstellung). Unter anderem untersuchten sie ChatGPTs Deep Research Funktion, Perplexity, Elicit, Covidence und DistillerSR. Für die Suchstrategieerstellung sahen die Autorinnen und Autoren vielversprechende Ergebnisse z. B. von ChatGPT und Perplexity. Sie weisen dennoch darauf hin, dass eine sorgfältige Validierung unerlässlich ist.

5.2.3 Stopp-Kriterien

Nach wie vor ist es eine offene Diskussion, zu welchem Zeitpunkt ein priorisiertes Abstract-Screening vorzeitig beendet werden kann und wurde in 2 Publikationen behandelt: Repke et al. [32] evaluieren verschiedene Stopp-Kriterien. Von 15 statistischen oder regelbasierten Stopp-Kriterien erreichte nur eine Methode zuverlässig die apriori angesetzte Sensitivität bei gleichzeitiger Aufwandsreduktion. Auch diese Methode verfehlte allerdings regelhaft, den optimalen Stopp-Zeitpunkt, sodass mehr gescreent würde als nötig. Boetje und van de Schoot [33] präsentieren die SAFE Methode als pragmatische Heuristik für Stopp-Kriterien im Abstract-Screening. Die Methode besteht aus 4 Phasen (z. B. screenen einer Zufallsstichprobe) und geht davon aus, dass ein Active-Learning-Algorithmus zum Einsatz kommt. Boetje und van de Schoot legen keine Validierung des Ansatzes vor.

5.3 Umfassende Informationsbeschaffung

5.3.1 Bibliografische Datenbanken

Wilson [34] überprüfte beispielhaft Fehler bei der automatisierten Indexierung in PubMed. Hierzu wurden 1705 MEDLINE-Datensätze, die automatisch mit dem MeSH-Begriff „Malus“ indexiert wurden, dem Gattungsnamen für Apfelbäume, mit zugehörigen Titeln, Abstracts und Volltexten inhaltlich abgeglichen. Die Untersuchung zeigte, dass 135 (8 %) Datensätze falsch indexiert waren. Der häufigste Indexierungsfehler entstand bei der Verwendung des Wortes „apple“ in Gleichnissen, Metaphern oder Redewendungen (80 Fälle), gefolgt von Fällen, in denen „apple“ in einem Namen oder Begriff vorkam (50 Fälle). Weitere Indexierungsfehler zeigten sich z. B., wenn der Begriff in Akronymen zu finden war. Die Ergebnisse zeigen, dass

es bei der automatischen Indexierung zu Fehlern kommen kann, insbesondere wenn Datensätze Begriffe mit nicht wörtlicher oder alternativer Bedeutung enthalten.

5.3.1.1 Datenbankauswahl

Frandsen et al. [35] untersuchten die Abdeckung mehrerer bibliografischer Datenbanken. Sie stellten sich die Frage, ob neben PubMed und Embase auch andere Datenbanken für eine umfassende Literaturrecherche für SRs unerlässlich sind. Hierzu wurden 343 Cochrane-Reviews aus den Jahren 2017 bis 2022 der Cochrane-Gruppen: Public Health, Inkontinenz, Hepatologie und Gastroenterologie sowie Schlaganfall analysiert. Die in diesen Reviews eingeschlossenen Publikationen wurden manuell in PubMed und Embase gesucht. Wenn die Publikationen nicht gefunden wurden, erfolgte eine zusätzliche Suche in Datenbanken wie CENTRAL, PsycINFO, CINAHL und ClinicalTrials.gov (CT.gov). Die durchschnittliche Abdeckung der Publikationen in PubMed und Embase über alle 4 Fachgruppen hinweg betrug 72 %, während die Einbeziehung zusätzlicher Datenbanken die Trefferquote auf 94 % erhöht, insbesondere durch Central und CT.gov. Durchschnittlich 6 % der Publikationen konnte in keiner der untersuchten Datenbanken gefunden werden.

Rajit et al. [36] bewerten die Abdeckung von PubMed, Embase, OpenAlex und Semantic Scholar für das Monitoring der Evidenzlage für Leitlinien anhand der eingeschlossenen Publikationen in einer Leitlinie zum polyzystischen Ovarialsyndrom 2023 [37]. OpenAlex enthält die meisten Publikationen (99 %), gefolgt von Semantic Scholar (98 %), Embase (97 %) und PubMed (93 %). Zusätzlich wurde das Risk of Bias für die Publikationen analysiert, die nicht in den Datenbanken enthalten waren. Die Autorinnen und Autoren schlagen OpenAlex als alleinige Quelle für die Evidenzsuche bei der Entwicklung von „Living Guidelines“ vor, da die Datenbank – im Vergleich zu MEDLINE und Embase – eine hohe Abdeckung aufweist und ein geringes Risiko besteht, hochwertige Studien (niedriges Risk of Bias) zu übersehen. Allerdings wurde lediglich die Abdeckung (also „Coverage“) geprüft und nicht, ob die Publikationen auch tatsächlich mit einer realistischen Suche identifiziert werden könnten. Siehe dazu auch die Zusammenfassung zu Stansfield et al. [38] in Abschnitt 5.3.3.

Bethel et al. [39] gehen beispielhaft der Frage nach, welche Suchquellen bzw. Suchtechniken für die Erstellung von Evidence and Gap Maps (EGMs) zum Thema „Peer support interventions“ verwendet werden sollten. Es wurden hierfür 2 „Search summary tables“ mit Datenbanken sowie Suchtechniken und deren Abdeckung für die Suche nach SRs und RCTs vorgestellt. Aus solchen Tabellen geht z. B. hervor, wie viele der eingeschlossenen Publikationen auf die einzelne Suchquelle zurückgehen. Die Untersuchung zeigt u. a., dass es sinnvoll ist, mehrere bibliografische Datenbanken zu durchsuchen, um einen hohen Abdeckungsgrad zu erreichen. MEDLINE und PsycINFO waren hierbei die wichtigsten Datenbanken für die Identifizierung sowohl von SRs als auch von RCTs, ergänzt durch CINAHL für SRs und CENTRAL für RCTs. Nur vereinzelt konnten ergänzende Publikationen mittels

Vorwärtszitationssuche oder Google Scholar identifiziert werden, keine mittels Rückwärtszitationssuche. Die Autorinnen und Autoren schlagen vor, solche „Search summary tables“ zu veröffentlichen, um mehr Transparenz der Suchprozesse zu ermöglichen und die Effizienz von Suchmethoden für EGMs und andere Arten von Evidenzsynthesen zu verbessern.

Horton et al. [40] untersuchten die Abdeckung bibliografischer Datenbanken für die Suche nach qualitativen Studien. Die Abdeckung der Datenbanken MEDLINE, CINAHL, PsycINFO, Scopus und Web of Science (WoS) wurde anhand einer Liste von 191 relevanten Zeitschriftentitel, die von der Canadian Drug Agency auf Grundlage der von ihren Wissenschaftlerinnen und Wissenschaftlern häufig herangezogenen Titel erstellt worden war, bewertet. Nicht alle 191 Zeitschriften waren in den untersuchten Datenbanken indexiert. Die multidisziplinären Zitationsindices wiesen den höchsten Anteil an Zeitschriften auf, wobei WoS bei 91 % und Scopus bei 82 % lagen. MEDLINE (51 %) hatte den dritthöchsten Anteil am Gesamtbestand, gefolgt von den eher fachspezifischen Datenbanken CINAHL (48 %) und PsycINFO (38 %). Die Autorinnen und Autoren kamen zu dem Schluss, dass neben den zentralen Datenbanken (MEDLINE / PubMed, PsycINFO, CINAHL) es sinnvoll sein kann, Scopus oder WoS als ergänzende Quelle für die Suche nach qualitativer Forschung heranzuziehen.

Lagisz et al. [41] beschreiben die Validierung und Entwicklung von Suchstrategien anhand von sogenannten „Benchmark-Publikationen“ und geben hierzu ein ausführliches Tutorial. Das Vorgehen ähnelt dem Einsatz von Testsets bei der Validierung von Suchstrategien am IQWiG [42]. Die Autorinnen und Autoren bemängeln auf der Grundlage von 100 ausgewerteten systematischen Übersichten, dass in der Regel Details über die Entwicklung von Suchstrategien und deren Validierung fehlen und plädieren für eine detailliertere Darstellung zu diesen Aspekten in SRs bzw. zugehörigen Protokollen.

5.3.1.2 Suchfilter

Generell wird vor der Verwendung von Suchfiltern eine Evaluation der Filtererstellung empfohlen [43,44]. Von hochwertigen bzw. vertrauenswürdigen Filtern ist zu fordern, dass die Identifikation des Goldstandards an relevanten Referenzen beschrieben ist und dass die Validierung anhand eines unabhängigen Sets relevanter Referenzen durchgeführt wurde. Gegebenenfalls finden sich hierzu bereits Angaben auf der Website von InterTASC [45]. Für ausgewählte Publikationen wurde eine eigene Kurzbewertung vorgenommen. Dabei dienten die Bewertungskriterien von Bak et al. [46] als Orientierung. Diese Kurzbewertung finden sich in Tabelle 2 in Anhang A.

Muente et al. [47] verglichen validierte geografische Suchfilter. In das SR wurden 6 Suchfilter (Afrika, Spanien, das Vereinigte Königreich, deutschsprachige Länder – eingeschränkt auf den Bereich Krankenpflege, OECD-Länder und USA) einbezogen. Nur 3 davon wurden unter Verwendung objektiver Methoden erstellt und 2 einer internen Validierung unterzogen. Die

gleiche Arbeitsgruppe [48] veröffentlichten außerdem eine Anpassung eines PubMed-Filters für deutsche Studien für die Datenbanken Embase (Ovid) sowie MEDLINE (Ovid) und validierten diese neu entstandenen Filter.

Moberg und Gornitzki [49] von der schwedischen HTA-Agentur SBU bewerteten die Sensitivität und die Numbers Needed to Read (NNR) für den Cochrane-RCT-Classifer im Vergleich sowie in Kombination mit etablierten RCT-Suchfiltern (sensitivity-maximizing Cochrane-RCT-Filter für MEDLINE [Ovid] und Cochrane-Embase-RCT-Filter für Embase.com) im Screeningtool Covidence. Der Goldstandard bestand aus 1103 RCTs aus 47 abgeschlossenen SBU-Berichten von 2019 bis 2024. Darüber hinaus wurden die Filter in 3 Fallstudien validiert, um die Auswirkungen auf die Suchergebnisse und die NNR zu bewerten. Der Cochrane-RCT-Classifer wies mit 99,6 % die höchste Sensitivität auf, gefolgt vom Cochrane-RCT-Filter in MEDLINE (Ovid) mit 98,7 % und dem Cochrane-Embase-RCT-Filter mit 98,5 %. Allerdings wies der Cochrane RCT-Classifer in allen Fallstudien eine höhere NNR auf als die RCT-Filter. Die Kombination der RCT-Filter mit dem Cochrane RCT-Classifer führte zu einer Verringerung der NNR im Vergleich zur alleinigen Verwendung der RCT-Filter, wobei bei der Kombination aus MEDLINE (Ovid) und dem RCT-Classifer eine Sensitivität von 98,5 % und bei der Kombination aus Embase und dem RCT-Classifer eine Sensitivität von 98,3 % erreicht wurde. Außerdem werden wichtige Limitierung erwähnt (z. B. wurden Referenzen ohne Abstract ausgeschlossen), die vor einer praktischen Umsetzung beachtet werden sollten.

5.3.2 Studienregister

Die Veröffentlichung von Studienergebnissen in Studienregistern spielt eine große Rolle für wissenschaftliche Transparenz. Dennoch werden Ergebnisse nicht immer veröffentlicht. Vor diesem Hintergrund diskutieren Chan et al. [50] in ihrer Arbeit, warum eine Veröffentlichung der Ergebnisse wichtig ist, wie der aktuelle Umgang hierzu in insgesamt 17 Studienregistern ist sowie die neue WHO-Richtlinie von 2024 [51] zur Veröffentlichung zusammenfassender Ergebnisse. Diese Richtlinie definiert 8 Mindestkriterien, die für das Reporting der Ergebniszusammenfassungen wichtig sind und in den Registern umgesetzt sein sollten. Um eine möglichst flächendeckende Umsetzung der neuen WHO-Richtlinie zu erzielen, sprechen die Autorinnen und Autoren hierfür 3 zentrale Empfehlungen aus: 1. sollen Prüfer und Sponsoren sicherstellen, dass Ergebnisse innerhalb von 12 Monaten nach Abschluss veröffentlicht werden, 2. sollten Regierungen oder Betreibende der Register ausreichende Ressourcen bereitstellen, um die neue WHO-Richtlinie umzusetzen (z. B. Infrastruktur, neue Datenfelder etc.) und drittens sollten relevante Akteure wie Sponsoren, Behörden oder Gesetzgeber Maßnahmen umsetzen, die die Einhaltung der WHO-Richtlinie vorschreiben.

Mit der gleichen Problematik befassen sich auch Kaul et al. [52] im Rahmen einer 3-geteilten Mixed-Methods Studie. In einer deskriptiven Analyse betrachteten die Autorinnen und Autoren 201 265 RCTs aus 6 Registern. Nur 17 % der Studien berichteten Ergebnisse, wovon

63 % direkt über das Register einsehbar waren und bei 21 % eine Verlinkung auf eine Zeitschriftenpublikation erfolgte. Eine 2-stufige Online-Befragung von Studienregistrierenden sowie Nutzerinnen und Nutzern zeigte, dass zwar 86 % der Registrierenden die Möglichkeit zur Ergebnisberichterstattung im Register kannten, jedoch 61 % die Veröffentlichung in Fachzeitschriften priorisieren und häufig Bedenken hinsichtlich Beeinträchtigungen für Veröffentlichungen in Zeitschriften haben. Bei den Nutzerinnen und Nutzern hatten bereits über 50 % Ergebnisse aus Studienregistern verwendet. Das größte Hindernis für die Nutzung von Ergebnissen stellte das mangelnde Vertrauen in Daten, die keinem Peer Review unterzogen wurden. Zuletzt wurden im Rahmen einer Expertendiskussion Empfehlungen erarbeitet, um den 10 identifizierten Hauptbedenken entgegenzuwirken, darunter die Standardisierung von Registeroberflächen und Prozessen, E-Mail Erinnerungen an die Registrierenden sowie Maßnahmen zur Förderung von Vertrauen in berichtete Ergebnisse in Registern.

Lee et al. [53] untersuchen die Entwicklung bei der jährlichen Anzahl registrierter klinischer Studien in Ostasien (China, Japan, Südkorea und Taiwan) gegenüber den USA im Zeitraum von 2014 bis 2023. Datengrundlage bietet hierbei das WHO ICTRP Search Portal, sowie das taiwanesisches Studienregister (Taiwan Clinical Trial Registry). Mehrfachregistrierung von Studien wurden identifiziert und bereinigt. Die Daten wurden hinsichtlich Studientyp (RCT), Studiengröße, Indikationsgebieten sowie dem Anteil nationaler und internationaler Studien ausgewertet. Während im Jahr 2014 in China 2578 klinische Studien neu registriert wurden, stieg diese Zahl bis zum Jahr 2023 auf mehr als das 6-fache (16 612). In den USA blieb die jährliche Zahl der Registrierungen im gleichen Zeitraum nahezu konstant (2014: 8841; 2023: 9100). Die USA hat aber einen deutlichen Schwerpunkt auf international durchgeführte Studien, während Studien in China überwiegend lokal durchgeführt werden. Die Studie adressiert zudem nur die jährliche Anzahl registrierter Studien, führt aber keine Auswertung zu den Ergebnissen, Vollständigkeit oder Vertrauenswürdigkeit der Studien durch. Nichtsdestotrotz zeigt die Analyse, dass wir in Zukunft mit einer erhöhten Anzahl von Studien aus China rechnen müssen, von denen viele, wie Auswertungen [54,55] zeigen, methodisch fragwürdig sind.

Zhang et al. [56] untersuchten die Angaben zum Data Sharing Plan in Studienregistrierungen und deren Übereinstimmung mit den Angaben in den zugehörigen Publikationen. Hierzu wurden Studien aus 6 hochrangigen Fachzeitschriften analysiert und die zugehörigen Registereinträge (z. B. CT.gov) hinsichtlich der geplanten Datenfreigabe ausgewertet. Von 383 erfassten Studien plante weniger als die Hälfte (45 %) eine Datenweitergabe im Rahmen der Registrierung. 133 (35 %) planten keine Weitergabe von Daten und bei 79 (21 %) war die Entscheidung zur Weitergabe von Daten noch offen bzw. fehlte sie. Dabei zeigten sich häufige Diskrepanzen zu den dazugehörigen Publikationen, was auf eine mangelnde Übereinstimmung der berichteten Data Sharing Plans hinweist. Unabhängig davon planten in der

Registrierung 20 von 383 Studien (5 %) die Bereitstellung des Analysecodes, 101 (26 %) die Freigabe des statistischen Analyseplans (SAP) bzw. Studienprotokolls und 105 (27 %) die Bereitstellung von IPD. Die Studie bewertete jedoch nicht, ob die angekündigten Daten tatsächlich veröffentlicht oder bereitgestellt wurden.

5.3.3 Weitere Informationsquellen und Suchtechniken

Stansfield et al. [38] untersuchen die Eignung von OpenAlex als Suchquelle für SRs in der Gesundheitsforschung anhand eines Beispiels. Hierzu analysieren die Autorinnen und Autoren Abdeckung und Auffindbarkeit der 131 eingeschlossenen Referenzen aus dem ausgewählten SR. Von diesen waren 128 (98 %) in OpenAlex verfügbar. Mit einer booleschen Suche in OpenAlex konnten 126 Referenzen (96 %) identifiziert werden. Über eine anschließende Zitationssuche ausgehend von den 126 identifizierten Artikeln konnte eine der beiden fehlenden Referenzen identifiziert werden. Bei einer Wiederholung der Suche zu einem späteren Zeitpunkt waren 3 Studien aufgrund entfernter Abstracts zu Closed-Access-Artikeln (Publikationen mit Zugriffsbeschränkungen) nicht mehr über eine boolesche Suche auffindbar. Generell zeigt sich eine geringere Verfügbarkeit von Abstracts in OpenAlex im Vergleich zu anderen Datenbanken [57]. Zudem wurden 2 zusätzliche relevante Studien in OpenAlex identifiziert, die in der ursprünglichen Suche für das SR nicht enthalten waren. Herausforderungen zeigten sich jedoch insbesondere bei den eingeschränkten Suchfunktionalitäten, der höheren Gesamttrefferzahl in OpenAlex und dem damit verbundenen Screening-Aufwand sowie der Reproduzierbarkeit aufgrund von dynamischen Inhalten (z. B. verschwinden von Abstracts). Im Ergebnis kommen die Autorinnen und Autoren zu dem Schluss, dass OpenAlex eine vielversprechende Ressource für SRs ist, insbesondere bei begrenztem Zugang zu kommerziellen Datenbanken, allerdings ist der Umfang der Recherche gegen die verfügbaren Ressourcen gut abzuwägen.

Khalil et al. [58] zeigen anhand einer Querschnittsanalyse von 100 SRs deutliche Mängel bei der Einbindung von grauer Literatur und versuchen, eine methodische Anleitung für die Integration grauer Literatur in Evidenzsynthesen zu entwickeln. Graue Literatur, die z. B. über eine Suche in Studienregistern, über Behördenwebsites oder eine Handsuche identifiziert wird, gewinnt laut den Autorinnen und Autoren zunehmend an Bedeutung und umfasst Materialien, die außerhalb des traditionellen akademischen oder kommerziellen Verlagswesens entstehen. Im Ergebnis zeigte sich, dass am häufigsten händisch nach grauer Literatur gesucht wird (58 %) und Automatisierungstools zur Unterstützung des Screenings und der Datenextraktion aus grauer Literatur trotz ihres Potenzials zur Effizienzsteigerung nur selten eingesetzt werden (19 %) Automatisierungstools, spezialisierte Datenbanken und konsistente Bewertungsinstrumente werden kaum eingesetzt, sodass die Suche variabel und intransparent bleibt. Die Ergebnisse flossen in die Entwicklung einer Checkliste zur Suche nach grauer Literatur ein.

Rajit et al. [59] untersuchen in einer Simulationsstudie die Leistungsfähigkeit automatisierter Zitationssuchen. Hierzu wurde die automatisierte Zitationssuche anhand von 27 SRs aus den 3 Bereichen Gesundheit, Umweltmanagement und Sozialpolitik getestet, indem ausgehend von einigen Schlüsselpublikationen (aus dem Hintergrund- und Methodenteil des SR) das direkte Zitationsnetzwerk ausgewertet wurde. Als Datenbanken für die Zitationssuche wurden OpenAlex und Semantic Scholar genutzt. Die Ergebnisse der Zitationssuche wurden anhand von Sensitivität, Präzision und F-Scores anschließend mit den ursprünglichen booleschen Suchstrategien verglichen. Es zeigte sich eine deutlich höhere Präzision der automatisierten Zitationssuche, allerdings lag die Sensitivität im Median bei nur 36 %. Die automatisierte Zitationssuche ist dementsprechend derzeit nicht als eigenständige Suchmethode geeignet, kann jedoch ergänzend gute Ergebnisse liefern. Dies bestätigt bekannte Ergebnisse älterer Untersuchungen [60], auch wenn sich durch ausgefeiltere Methoden wie Co-Zitate oder zusätzliche Durchläufe im Zitationsnetzwerk über mehr als eine direkte Zitationsebene hinaus (sogenannte „Hops“) durch das Zitationsnetzwerk eine weitere Leistungsverbesserung von Zitationssuchen zeigt.

Die Arbeit von Edwards et al. [61] befasst sich mit den Herausforderungen bei der Informationsbeschaffung für SRs zu textbasierter Evidenz (z. B. narrative Texte, Meinungsbeiträge, Leitlinien, Richtlinien etc.), welche überwiegend aus grauer Literatur resultiert. Es wird ein 3-stufiger Suchprozess nach der Methodik des JBI dargestellt, der im Wesentlichen anderen systematischen Literaturrecherchen entspricht, jedoch durch die Verwendung anderer Dokumentarten außerhalb bibliografischer Datenbanken erschwert wird. Die Autorinnen und Autoren betonen, dass klare Suchparameter (z. B. Suchbegriffe, Suchquellen und Suchumfang) sowie definierte Ein- und Ausschlusskriterien entscheidend bei der Durchführung sind, da viele Online- und Website-Suchen umfangreiche und nur schwer reproduzierbare Suchergebnisse liefern. Weitere Suchquellen sind beispielsweise ProQuest für Thesisdokumente und Dissertationen, Nachrichtendatenbanken für Medienberichte, Google oder Google Scholar Suchen für z. B. Stellungnahmen oder öffentliche politische Dokumente, diverse Repositorien (z. B. medRxiv) für Preprints sowie das Einholen von Expertenmeinungen. Die Auswahl der Quellen sollte auf das jeweilige Thema abgestimmt sein, sowohl öffentliche als auch unveröffentlichte Dokumente berücksichtigen und gut dokumentiert werden.

Henderson et al. [62] kartieren mit einer SR die Landschaft gesundheitsökonomischer Open-Source Modelle (OSM) mit dem Ziel einer systematischen Erfassung und Analyse von OSMs hinsichtlich Verfügbarkeit, Transparenz und Reproduzierbarkeit in der Gesundheitsökonomie. Hierzu wurden 11 digitale Datenbanken (z. B. GitHub oder Zenodo), 6 spezialisierte OSM-Repositorien sowie die bibliografische Datenbank PubMed nach „Open-Source“ oder „Open-Access“ Modellen durchsucht. Im Ergebnis konnten 182 OSMs identifiziert werden, wovon der größte Anteil (74 %) auf der Plattform GitHub gehostet wurde; nur 19 OSMs wurden über die

Suche in PubMed identifiziert. Dies zeigt die große Schlüsselrolle digitaler Plattformen für dieses Feld, wobei Datenbanken wie GitHub häufig nur eingeschränkte Suchfunktionalitäten und eher schwache Metadaten aufweisen. Die Autorinnen und Autoren empfehlen, dass OSM-Entwickler zukünftig die Begriffe „decision-analytic model“ oder „health economic model“ sowie einen Begriff zur Krankheitsbeschreibung, einen Begriff zur Beschreibung des Modelltyps (z. B. Markov) sowie den Begriff „Open Source“ zur Beschreibung ihrer Modelle auf Titel- / Abstract-Ebene verwenden sollen, um eine bessere Auffindbarkeit in bibliografischen Datenbanken zu gewährleisten. Eine weitere Herausforderung zeigte sich darin, dass ca. 24 % der OSMs keine erkennbare Open-Source-Lizenz auswiesen. Hierzu empfehlen die Autorinnen und Autoren die Nutzung von solchen Lizenzen (Licenses – Open Source Initiative). Künftige Forschung sollte darauf abzielen, die Beschränkungen bei den Suchfunktionen zu beseitigen und das Reporting über Modelle zu vereinheitlichen.

6 Literatur

1. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden; Version 8.0 [online]. 2025 [Zugriff: 06.01.2026]. URL: https://doi.org/10.60584/Allgemeine-Methoden_V8.0.
2. Hausner E, Sturtz S, Molnar S et al. Searching for non-English literature may be unnecessary for German HTA Reports. F1000Res 2024; 13: 1134. <https://doi.org/10.12688/f1000research.151365.2>.
3. Paynter RA, Adam GP, Hedden-Gross A et al. Systematic Review Search Strategy Development Tools: A Practical Guide for Expert Searchers [online]. 2025 [Zugriff: 23.07.2026]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/related_files/search-strategy-white-paper.pdf.
4. Cooper C, Premji Z, Worsley C et al. A New Process Model of Study Identification Specific to the Identification of Randomised Studies for Systematic Reviews of Medical Interventions. Cochrane Evid Synth Methods 2025; 3(3): e70026. <https://doi.org/10.1002/cesm.70026>.
5. Taji Heravi A, Gryaznov D, Busse JW et al. Metaresearch on patient-reported outcomes in trial protocols and results publications suggested large outcome reporting bias. J Clin Epidemiol 2025; 185: 111822. <https://doi.org/10.1016/j.jclinepi.2025.111822>.
6. Dobrescu A, Nussbaumer-Streit B, Wagner G et al. Assessing the Impact of Unpublished Data on Network Meta-Analysis Outcomes in Outpatient Adults with Acute Migraine: A Study within a Review. J Clin Epidemiol 2025: 112014. <https://doi.org/10.1016/j.jclinepi.2025.112014>.
7. Hirt J, Pieper D, Becker M et al. Clinical study report data did not substantially alter point estimates but improved precision in a nephrology systematic review. J Clin Epidemiol 2025: 111890. <https://doi.org/10.1016/j.jclinepi.2025.111890>.
8. Harper C, Groth P. The Librarian Skillset and the Needs of Artificial Intelligence. Cataloging & Classification Quarterly 2025; 63(6-7): 546-565. <https://doi.org/10.1080/01639374.2025.2539787>.
9. Hegedus M, Dadkhah M, David LD. Hijacked medical journals rank first via search engine optimization and threaten academic integrity. Eur J Intern Med 2025; 140: 106450. <https://doi.org/10.1016/j.ejim.2025.106450>.
10. Spick M, Onoja A, Harrison C et al. Quantifying new threats to health and biomedical literature integrity from rapidly scaled publications and problematic research. J Clin Epidemiol 2026: 112203. <https://doi.org/10.1016/j.jclinepi.2026.112203>.

11. Scancar B, Byrne JA, Causeur D et al. Machine learning based screening of potential paper mill publications in cancer research: methodological and cross sectional study. *BMJ* 2026; 392: e087581. <https://doi.org/10.1136/bmj-2025-087581>.
12. Committee on Publication Ethics. Paper mills [online]. [Zugriff: 07.08.2026]. URL: <https://publicationethics.org/topic-discussions/paper-mills>.
13. Tang G, Cai H. Citation Contamination by Paper Mill Articles in Systematic Reviews of the Life Sciences. *JAMA Netw Open* 2025; 8(6): e2515160. <https://doi.org/10.1001/jamanetworkopen.2025.15160>.
14. Papapanou M, Siristatidis C, Paraskevas T et al. Enhancing trustworthiness checks of randomized controlled trials: a pathway to more reliable evidence. *J Clin Epidemiol* 2026; 195: 112259. <https://doi.org/10.1016/j.jclinepi.2026.112259>.
15. Weibel S, Meybohm P, Oeser A et al. Impact of a Research Integrity Assessment (RIA) of Randomized Controlled Trials Included in Interventional COVID-19 Systematic Reviews: A Meta-Epidemiological Study. *Cochrane Evid Synth Methods* 2026; 4(2): e70076. <https://doi.org/10.1002/cesm.70076>.
16. Wilkinson J, Heal C, Antoniou GA et al. Assessing the feasibility and impact of clinical trial trustworthiness checks via an application to Cochrane Reviews: Stage 2 of the INSPECT-SR project. *J Clin Epidemiol* 2025; 184: 111824. <https://doi.org/10.1016/j.jclinepi.2025.111824>.
17. Wilkinson J, Heal C, Antoniou GA et al. Protocol for the development of a tool (INSPECT-SR) to identify problematic randomised controlled trials in systematic reviews of health interventions. *BMJ Open* 2024; 14(3): e084164. <https://doi.org/10.1136/bmjopen-2024-084164>.
18. Wilkinson J, Tovey D. How should we assess trustworthiness of randomized controlled trials? *J Clin Epidemiol* 2025; 180: 111670. <https://doi.org/10.1016/j.jclinepi.2025.111670>.
19. O'Keefe H, Eastaugh C, Yazar F et al. Concerns of AI use in evidence synthesis based practices: collective views from the community. *BMC Med Res Methodol* 2026. <https://doi.org/10.1186/s12874-026-02844-x>.
20. Flemyng E, Noel-Storr A, Macura B et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence 2025. *Environ Evid* 2025; 14(1): 20. <https://doi.org/10.1186/s13750-025-00374-5>.
21. Thomas J, Flemyng E, Noel-Storr A et al. Responsible use of AI in evidence Synthesis (RAISE): recommendations and guidance [online]. 2025 [Zugriff: 16.07.2025]. URL: <https://doi.org/10.17605/OSF.IO/FWAUD>.

22. Adam GP, Davies MC, George J et al. Machine Learning Tools To (Semi-)Automate Evidence Synthesis: A Rapid Review and Evidence Map Surveillance Report [online]. 2026 [Zugriff: 27.02.2026]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/related_files/machine-learning-tools-white-paper-2026-update.pdf.
23. Laignelot F, Martin GL, Ossman M et al. Large language models show promising performance for some systematic review tasks but call for cautious implementation : a systematic review. J Clin Epidemiol 2026: 112221. <https://doi.org/10.1016/j.jclinepi.2026.112221>.
24. Gartlehner G, Banda S, Callaghan M et al. Cochrane Evaluation of (Semi-) Automated Review Methods (CESAR): Protocol for an adaptive platform study within reviews. J Clin Epidemiol 2026: 112390. <https://doi.org/10.1016/j.jclinepi.2026.112390>.
25. Janoudi G, Uzun M, Disher T et al. Validating Loon Lens 1.0 for Autonomous Abstract Screening and Confidence-Guided Human-in-the-Loop Workflows in Systematic Reviews. Value Health 2025; 28(11): 1630-1636. <https://doi.org/10.1016/j.jval.2025.09.008>.
26. Akinseloyin O, Jiang X, Palade V. Large language model-based multiagent collaboration for abstract screening toward automated systematic reviews. Biol Methods Protoc 2026; 11(1): bpag006. <https://doi.org/10.1093/biomethods/bpag006>.
27. Hu B, Tomini E, Corrin T et al. Enhancing Evidence Synthesis Efficiency: Leveraging Large Language Models and Agentic Workflows for Optimized Literature Screening. Cochrane Evid Synth Methods 2025; 3(6): e70042. <https://doi.org/10.1002/cesm.70042>.
28. Featherstone R, Walter M, MacDougall D et al. A Comparative Analysis of Artificial Intelligence Search Tools for Evidence Synthesis [online]. 2025 [Zugriff: 22.07.2026]. URL: <http://dx.doi.org/10.22541/au.174897559.99564896/v1>.
29. Wang S, Scells H, Koopman B et al. Reassessing Large Language Model Boolean Query Generation for Systematic Reviews [online]. 2025 [Zugriff: 23.05.2025]. URL: <https://arxiv.org/abs/2505.07155v1>.
30. Wang S, Scells H, Koopman B et al. AutoBool: An Reinforcement-Learning trained LLM for Effective Automated Boolean Query Generation for Systematic Reviews [online]. 2025 [Zugriff: 20.02.2026]. URL: <https://doi.org/10.48550/arXiv.2602.00005>.
31. Institute of Health Economics. Generative Artificial Intelligence Use at the Institute of Health Economics; Learnings and Considerations for Health Evidence Reviews [online]. 2025 [Zugriff: 23.07.2026]. URL: <https://ihe.ca/wp-content/uploads/2025/10/AI-Methods-FINAL.pdf>.

32. Repke T, Tinsdeall F, Danilenko D et al. Don't Stop Me Now, ;Cause I'm Having a Good Time Screening: Evaluation of Stopping Methods for Safe Use of Priority Screening in Systematic Reviews. *Cochrane Evid Synth Methods* 2026; 4(1): e70068. <https://doi.org/10.1002/cesm.70068>.
33. Boetje J, van de Schoot R. The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Syst Rev* 2024; 13(1): 81. <https://doi.org/10.1186/s13643-024-02502-7>.
34. Wilson P. Sometimes the apple does fall far from the tree: a case study on automatic indexing precision errors in PubMed. *J Med Libr Assoc* 2025; 113(4): 318-326. <https://doi.org/10.5195/jmla.2025.2110>.
35. Frandsen TF, Moos C, Marino C et al. Supplementary databases increased literature search coverage beyond PubMed and Embase. *J Clin Epidemiol* 2025; 181: 111704. <https://doi.org/10.1016/j.jclinepi.2025.111704>.
36. Rajit D, McDonald S, Tay CT et al. Assessing the Coverage of PubMed, Embase, OpenAlex and Semantic Scholar for Automated Single Database Searches in Living Guideline Evidence Surveillance: A Case Study of the International PCOS Guidelines 2023. *J Clin Epidemiol* 2025; 111789. <https://doi.org/10.1016/j.jclinepi.2025.111789>.
37. Teede HJ, Tay CT, Laven JJE et al. Recommendations From the 2023 International Evidence-based Guideline for the Assessment and Management of Polycystic Ovary Syndrome. *J Clin Endocrinol Metab* 2023; 108(10): 2447-2469. <https://doi.org/10.1210/clinem/dgad463>.
38. Stansfield C, Dehdarirad H, Thomas J et al. Analyzing the Utility of Openalex to Identify Studies for Systematic Reviews: Methods and a Case Study. *Cochrane Evid Synth Methods* 2025; 3(4): e70038. <https://doi.org/10.1002/cesm.70038>.
39. Bethel AC, Shaw N, Abbot R et al. Insights from search summary tables for evidence and gap maps: a case study on peer support interventions. *J Med Libr Assoc* 2025; 113(2): 177-183. <https://doi.org/10.5195/jmla.2025.1831>.
40. Horton J, Kaunelis D, Rabb D et al. What's beyond the core? Database coverage in qualitative information retrieval. *J Med Libr Assoc* 2025; 113(1): 49-57. <https://doi.org/10.5195/jmla.2025.1591>.
41. Lagisz M, Yang Y, Young S et al. A practical guide to evaluating sensitivity of literature search strings for systematic reviews using relative recall. *Res Synth Methods* 2025; 16(1): 1-14. <https://doi.org/10.1017/rsm.2024.6>.
42. Hausner E, Waffenschmidt S, Kaiser T et al. Routine development of objectively derived search strategies. *Syst Rev* 2012; 1: 19. <https://doi.org/10.1186/2046-4053-1-19>.

43. Glanville J, Bayliss S, Booth A et al. So many filters, so little time: the development of a search filter appraisal checklist. *J Med Libr Assoc* 2008; 96(4): 356-361.
<https://doi.org/10.3163/1536-5050.96.4.011>.
44. Jenkins M. Evaluation of methodological search filters--a review. *Health Info Libr J* 2004; 21(3): 148-163. <https://doi.org/10.1111/j.1471-1842.2004.00511.x>.
45. InterTASC Information Specialists' Sub-Group. The ISSG Search Filter Resource [online]. 2025 [Zugriff: 16.10.2025]. URL: <https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home>.
46. Bak G, Mierzwinski-Urban M, Fitzsimmons H et al. A pragmatic critical appraisal instrument for search filters: introducing the CADTH CAI. *Health Info Libr J* 2009; 26(3): 211-219. <https://doi.org/10.1111/j.1471-1842.2008.00830.x>.
47. Muentz C, Pachanov A, Hirt J et al. Validated geographic search filters for bibliographic databases: a scoping review. *JBIM Evid Synth* 2025. <https://doi.org/10.11124/JBIES-24-00395>.
48. Pachanov A, Muentz C, Hirt J et al. Translation and validation of a geographic search filter to identify studies about Germany in Embase (Ovid) and MEDLINE(R) ALL (Ovid). *Res Synth Methods* 2025; 16(4): 688-700. <https://doi.org/10.1017/rsm.2025.10016>.
49. Moberg K, Gornitzki C. Combining search filters for randomized controlled trials with the Cochrane RCT Classifier in Covidence: a methodological validation study. *Res Synth Methods* 2025; 16(6): 953-960. <https://doi.org/10.1017/rsm.2025.10023>.
50. Chan AW, Karam G, Pyment J et al. Reporting summary results in clinical trial registries: updated guidance from WHO. *Lancet Glob Health* 2025; 13(4): e759-e768.
[https://doi.org/10.1016/S2214-109X\(24\)00514-X](https://doi.org/10.1016/S2214-109X(24)00514-X).
51. World Health Organization. Guidance for best practices for clinical trials [online]. 2024 [Zugriff: 24.07.2026]. URL: <https://iris.who.int/server/api/core/bitstreams/dbe6d97c-b659-4e7b-8070-b4a3c667d6e5/content>.
52. Kaul T, Damen JA, Seidler AL et al. Improving the reporting and use of trial results in clinical trials registries: global practices, barriers, and recommendations. *J Clin Epidemiol* 2025; 186: 111901. <https://doi.org/10.1016/j.jclinepi.2025.111901>.
53. Lee EH, Lee S, Shin JI et al. Registered clinical trial trends evolved differently in East Asia vs the United States during 2014-2023. *J Clin Epidemiol* 2025; 183: 111791.
<https://doi.org/10.1016/j.jclinepi.2025.111791>.
54. Fan J, Liu X, Li Y et al. Quality problems of clinical trials in China: evidence from quality related studies. *Trials* 2022; 23(1): 343. <https://doi.org/10.1186/s13063-022-06281-1>.

55. Teixeira da Silva JA, Nazarovets S, Daly T et al. The Chinese Early Warning Journal List: Strengths, weaknesses and solutions in the light of China's global scientific rise. *Journal of Academic Librarianship* 2024; 50(4): 102898. <https://doi.org/10.1016/j.acalib.2024.102898>.
56. Zhang J, Bierer BE, Van Spall HGC et al. Practice of data sharing plans in clinical trial registrations and concordance between registered and published data sharing plans: a cross-sectional study. *BMC Med* 2025; 23(1): 510. <https://doi.org/10.1186/s12916-025-04328-z>.
57. Culbert JH, Hobert A, Jahn N et al. Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics* 2025; 130(4): 2475-2492. <https://doi.org/10.1007/s11192-025-05293-3>.
58. Khalil H, Johns-Hayden A, Kynoch K. Guidance to including gray literature in systematic reviews-recommendations from an epidemiological study. *J Clin Epidemiol* 2026; 194: 112222. <https://doi.org/10.1016/j.jclinepi.2026.112222>.
59. Rajit D, Du L, Teede H et al. Automated citation searching in systematic review production: A simulation study. *Res Synth Methods* 2025; 16(1): 211-227. <https://doi.org/10.1017/rsm.2024.15>.
60. Robinson KA, Whitlock EP, O'Neil ME et al. Integration of Existing Systematic Reviews [online]. 2014 [Zugriff: 23.007.2026]. URL: https://www.ncbi.nlm.nih.gov/books/NBK216379/pdf/Bookshelf_NBK216379.pdf.
61. Edwards D, Cooper A, McArthur A et al. Textual evidence systematic reviews series paper 2: challenges and strategies in developing a search strategy for systematic reviews of textual evidence. *JBIM Evid Synth* 2025; 23(4): 566-575. <https://doi.org/10.11124/JBIES-24-00292>.
62. Henderson RH, Sampson C, Pouwels X et al. Mapping the Landscape of Open Source Health Economic Models: A Systematic Database Review and Analysis: An ISPOR Special Interest Group Report. *Value Health* 2025; 28(6): 813-820. <https://doi.org/10.1016/j.jval.2025.01.019>.
63. Fontanive VN, Muller LP, Riva F et al. Development and validation of high-sensitivity and high-specificity pubmed search filters for systematic and non-systematic reviews. *BMC Med Res Methodol* 2025; 25(1): 216. <https://doi.org/10.1186/s12874-025-02666-3>.
64. Kysh L, Baker Z, Bogucka R et al. Developing and validating infant hedges for PubMed and Ovid MEDLINE: a Medical Library Association Pediatric Librarians Caucus initiative. *J Med Libr Assoc* 2025; 113(4): 281-289. <https://doi.org/10.5195/jmla.2025.2034>.

Anhang A Übersicht aktueller Suchfilter

Tabelle 2: Kurzbewertung der Filter

	Fontanive et al. [63]		Kysh et al. [64]		
Thema	Suchfilter für (systematische) Übersichten in PubMed		Kinderfilter für Säuglinge (von der Geburt bis zum 23. Lebensmonat)		
Suchoberfläche	PubMed		PubMed, MEDLINE (Ovid)		
Filterdesign	reproduzierbar		unklar		
Goldstandard	2246 Artikel aus 15 zahnmedizinischen Zeitschriften, davon 150 SRs und 132 non-SRs		291 (Sichtung von 5 Recherchen; insgesamt 1000 Referenzen)		
Validierung (unabhängiges Set)	nein (für neu entwickelte Filter)		nein		
Ziel / Ausprägung	Entwicklung von Filtern mit hoher Sensitivität und hoher Spezifität für SR und „any type of Reviews“; außerdem Validierung von Avau-Filter und PubMed SR-Filter		Filterentwicklung für SR, Scoping-Reviews oder für spezifische Suchanfragen „am Krankenbett“		
Gütemaße	High sensitivity filter (any type of review)			PubMed	MEDLINE (Ovid) (Appendix A)
	High specificity filter (systematic reviews)	Se: 98,0 %, Sp: 88,9 %	Hedge 1	Se: 83,8 %, Sp: 88,2 %	Se: 83,6 %, Sp: 88,1 %
	PubMed-Systematic Review-Filter	Se: 96,7 %, Sp: 99,1 %	Hedge 2	Se: 83,2 %, Sp: 89,3 %	Se: 82,9 %, Sp: 89,4 %
	Avau Filter (systematic reviews)	Se: 91,3 %, Sp: 99,7 %	Hedge 3	Se: 83,2 %, Sp: 89,4 %	Se: 82,9 %, Sp: 89,3 %
		Se: 96,0 %, Sp: 98,4 %	Hedge 4	Se: 83,5 %, Sp: 89,7 %	Se: 82,9 %, Sp: 90,8 %
			Hedge 5	Se: 70,1 %, Sp: 96,2 %	Se: 69,6 %, Sp: 96,2 %
Fazit	erste SR-Filter Gütemaße nach Einführung des neuen PubMed Publication Type / MeSH Term „systematic reviews“ bzw. „systematic reviews as topic“ in 2019; außerdem bislang erste Validierung des PubMed SR-Filter eingeschränkt empfehlenswert, da nur für den Bereich Zahnmedizin geprüft; Gütemaße sollten auch an anderen Themen geprüft werden; trotzdem mögliche Alternative für bislang genutzte Filter wegen Aktualität; außerdem erstmals Gütemaße für „non SRs“ in PubMed		eingeschränkt empfehlenswert wegen insgesamt niedriger Sensitivität		
Se: Sensitivität; Sp: Spezifität; SR: systematische Übersicht					

Anhang B Suchquellen und Suchstrategien

Suchstrategien für MEDLINE (Ovid)

Zum Thema AI

- Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present

#	Searches
1	acceler*.ti.
2	(develop* or (proof adj2 concept)).ti.
3	enhanc*.ti.
4	(implementation or approach* or method*).ti.
5	techn*.ti.
6	(tool* or application*).ti.
7	(evaluat* or assess* or accuracy or sensitivit* or specificit* or precision or recall or "ROC" or "CAP curve").ti.
8	or/1-7
9	((clinical or controlled) adj (trial* or stud*)) or "RCT").ti.
10	Systematic Reviews as Topic/
11	((systematic or scoping or literature or evidence or clinical or diagnostic test accuracy or mapping) adj review*).mp.
12	(evidence adj (synthesis or summar*)).mp.
13	or/9-12
14	*Artificial Intelligence/ or Artificial Intelligence/td [Trends]
15	ai.ti.
16	algorithm*.ti.
17	artificial intelligence.ti.
18	automat*.ti.
19	chatbot*.ti.
20	(gpt* or chatgpt*).ti.
21	(generate or generativ*).ti.
22	(retrieval augmented generat* or "RAG").ti,ab.
23	(large language model* or natural language processing or NLP).ti.
24	((machine or deep) adj learning).ti,ab.
25	((text adj mining) or textmining or (text* adj analysis)).ti,ab.
26	prompt*.ti,ab.
27	named entity recognition.ti.
28	retriev*.ti.
29	(transformer-based or ((pre-trained or pretrained) adj transformer*)).ti.
30	classifier*.ti.

#	Searches
31	or/14-30
32	R Package*.ti.
33	or/31-32
34	(((critical or evidence) adj1 appraisal) or ((risk or assess*) adj2 bias)).ti.
35	(evidence adj (synthesis or review* or summar*)).ti.
36	((data or information or relation or study or PICO*) adj1 extrac*).ti.
37	((article* or trial or stud* or record* or paper* or abstract* or title* or literature or fulltext* or full-text* or PDF or report*) and screen*).ti.
38	study selection.ti.
39	(information adj (retrieval or seeking)).ti.
40	((biomedical or academic or medical) adj literature).ti.
41	(literature adj (search or screening*)).ti.
42	search engine*.ti.
43	((scholarly or literature or science or PubMed) adj search).ti.
44	(search adj (strateg* or tool or engine*)).ti.
45	systematic search*.ti.
46	*"Information Storage and Retrieval"/
47	*Search Engine/ or Search Engine/td or Search Engine/td, mt
48	or/34-47
49	("British Medical Journal" or nature or "Nature Medicine").jn.
50	13 and 33
51	8 and 33 and 48
52	13 and 33 and 48
53	33 and 49
54	14 and 49
55	51 or 52 or 53 or 54

Zum Thema Studienregister und unpublizierte Daten

- Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present

#	Searches
1	clinicaltrials*.ti.
2	unpublished.ti.
3	(handsearch* or hand search* or reference list*).ti.
4	((full or abstract or bias or selective) adj3 (publication or reporting or dissemination)).ti.
5	or/1-4
6	Registries/

#	Searches
7	*Databases, Factual/
8	(clinicaltrials* or trial* registr* or ICTRP or European Medicines Agency).ti.
9	6 or 7 or 8
10	Clinical Trials as Topic/
11	*Randomized Controlled Trials as Topic/
12	9 and (10 or 11)
13	5 or 12

Zum Thema Recherche allgemein und Studienfilter

- Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present

#	Searches
1	(Semi-automated or data mining or text mining or (Capture adj3 recapture) or machine learning).ab,ti.
2	(search adj3 (strateg* or term* or filter*)).ab,ti.
3	(query or queries).ab,ti.
4	search*.ti.
5	or/1-4
6	exp "Information Storage and Retrieval"/
7	"Medical Subject Headings"/
8	"Abstracting and Indexing as Topic"/
9	Documentation/
10	reporting.ab,ti.
11	(bibliographic databas* or Pubmed).ab,ti.
12	(MESH or controlled vocabulary or indexing).ab,ti.
13	or/6-12
14	"Review Literature as Topic"/
15	exp "Evidence-Based Practice"/
16	"Technology Assessment, Biomedical"/
17	or/14-16
18	and/5,13,17
19	exp *"Information Storage and Retrieval"/
20	4 and (16 or 19)
21	(Medline or PubMed).ti.
22	(Clinical Queries or Haynes or hedge or search).ti,ab.
23	and/21-22
24	or/18,20,23

Zum Thema Datenbanken

- Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present

#	Searches
1	exp Databases, Bibliographic/
2	(medline* or pubmed* or embase* or cochrane* or cinahl* or psycinfo* or amed* or google* or pedro*).ti.
3	1 or 2
4	exp *"Information Storage and Retrieval"/
5	(systematic and search*).ab,ti.
6	or/4-5
7	review*.ab,ti.
8	Review Literature as Topic/
9	Meta-Analysis as Topic/
10	Evidence-Based Medicine/
11	*Randomized Controlled Trials as Topic/
12	or/7-11
13	and/3,6,12

Gesichtete Fachzeitschriften

- British Medical Journal
- Cochrane Evidence Synthesis and Methods
- Health Information & Libraries Journal
- Journal of Clinical Epidemiology
- Research Synthesis Methods

Gesichtete Kongresse

- Cochrane Colloquia
- Health Technology Assessment International (HTAi): Annual Meetings
- Information Retrieval Meetings (IRM)
- CORE Information Retrieval Forum
- National Institute for Health and Care Excellence (NICE) Joint Information Days

Sichtung Webseiten ausgewählter Organisationen

- Agency for Healthcare and Quality (AHRQ)

- Cochrane Collaboration
- Summarized Research in Information Retrieval for HTA (SuRe Info)
- InterTASC Information Specialists' Sub-Group Search Filter Resource (ISSG SFR)