ASIDE Health Sciences

Review Article

Search Filters to Identify Automation in HEOR: An Umbrella Review of Performance and Overlap

Maria Rizzo[1], Omar Irfan[2,*]

1-Independent Consultant, London, United Kingdom
2-Independent Consultant, Milton, Ontario, Canada

ARTICLE INFO

Article history:
Received 3 Aug. 2025
Received in revised form 28 Nov. 2025
Accepted 7 Dec. 2025
Published 21 Dec. 2025

Keywords:
Automation
Artificial intelligence
Health economics and outcomes research
HEOR
Search filters

ABSTRACT

Background: Search strategies used to identify evidence on automation in Health Economics and Outcomes Research (HEOR) often lack sensitivity and specificity, resulting in information overload or missed studies. This umbrella review evaluated and compared the performance and overlap of search filters commonly used to retrieve automation-related HEOR evidence.
Methods: Systematic literature reviews (SLRs) and search filters focusing on any form of automation in HEOR, were included. Searches (January 01, 2023–July 03, 2024) were conducted in EMBASE, ISSG Resource, and Google Scholar. Subject headings, search terms, and performance metrics were extracted. Reference lists were cross-checked. Screening was performed by one reviewer, with 20% verified by a second reviewer. The PRESS checklist was used to assess search strategy quality. The protocol was registered with Open Science Framework (OSF).
Results: Seven SLRs and one standalone filter, reporting 11 search strategies, met inclusion criteria. HEOR relevance was defined by studies applying search filters in contexts of SLRs, indirect treatment comparisons, and economic modelling. Included SLRs retrieved between 5-273 studies. PubMed was the most frequently searched database. Commonly exploded subject-headings included "artificial intelligence," "deep learning," "machine learning," and "natural language processing," with "artificial intelligence" the most frequent free-text term. Inclusion rates varied: title/abstract (1%–8%), full-text (27%–86%), final inclusion (0.13%–2.31%). Time to include one study ranged from 0.6-8 hours.
Conclusions: Considerable variability in search filter performance was observed, causing lower specificity and inefficient evidence retrieval. Standardized, high-performing search strategies are needed to enhance efficiency and reliability in identifying automation-related HEOR evidence.

1. Introduction

Health Economics and Outcomes Research (HEOR) plays a pivotal role in ensuring patients have access to new technologies that have been evaluated for efficacy, safety, and cost-effectiveness, thereby optimizing resource allocation within the healthcare sector [1]. With the escalating costs of new and innovative medicines and the constraints on public funding, the efficient allocation of healthcare resources is imperative to ensure that cost-effective treatments are prioritized and accessible to those in need. However, the health technology appraisal (HTA) landscape is also becoming more complex, with frequent changes to reimbursement legislations, and with more advanced and complex HEOR analytics (e.g., multi-level network meta regression [ML-NMR], synthetic control arms, quantitative bias analysis), it is not surprising that the consideration on the use of artificial intelligence (AI) in HEOR is being considered, to expedite timelines, to get new technology to patients on time [2]. HTA bodies have begun using AI in reimbursement submissions. Most recently, the National Institute for Health and Care Excellence (NICE) accepts the use of AI, however, its use must be declared, the methods transparently reported, should only be applied if there is a rationale, and its use adds value, and most critically, that the AI only augments human involvement, and does not replace it [3]. Additionally, CADTH outlines the use of AI in systematic literature reviews (SLRs) to automate data extraction and searching [4]. Additionally, the Institut National d'Excellence en Santé et Services Sociaux (INESSS) created a proof-of-concept demonstrating the application of a large language model (LLM) based on GPT-4 to aid title and abstract screening [5].

As AI continues to be integrated into HTA submissions, researchers, decision-makers, and HEOR experts must remain up to date with the latest developments in AI methodologies, applications, and software. The rapid expansion of research on AI and automation in HEOR has created a vast, continually growing body of literature, posing significant challenges for efficiently identifying the most relevant studies. Current search strategies often struggle to balance sensitivity (capturing all relevant studies) and specificity (excluding irrelevant studies), leading to either information overload or missed key evidence [6]. There is therefore an urgent need for robust, validated search filters that can accurately navigate the complex, abundant research landscape in this domain. This umbrella review aims to systematically identify, evaluate, and compare the performance and overlap of search filters used

* Corresponding author: Omar Irfan, Independent Consultant, Milton, Ontario, Canada, Email: omarirfan1@hotmail.com

Published by the American Society for Inclusion, Diversity, and Equity in Healthcare (ASIDE). ISSN (Print) 3067-8730, ISSN (Online) 3067-8749. Hosting by ASIDE Journals.

Citation: Rizzo M, Irfan O. Search Filters to Identify Automation in HEOR: An Umbrella Review of Performance and Overlap. ASIDE Health Sci. 2025;2(2):9-17, doi:10.71079/ASIDE.HS.122125202

Title and Abstract Inclusion Rate:

$$\text{Inclusion Rate}_{\text{TA}} = \frac{\text{Full-texts Retrieved}}{\text{Title and Abstract Screened}}$$

Full-text Inclusion Rate:

$$\text{Inclusion Rate}_{\text{FT}} = \frac{\text{Final Included Studies}}{\text{Full-texts Retrieved}}$$

Overall Inclusion Rate:

$$\text{Inclusion Rate}_{\text{Overall}} = \frac{\text{Final Included Studies}}{\text{Title and Abstract Screened}}$$

Figure 1: Search performance metrics
FT, Full-text; TA, Title and Abstract

to identify evidence on automation in HEOR. To our knowledge, few studies have compared the overlap and performance of filters used to retrieve automation studies in HEOR-relevant contexts, highlighting a critical gap in the literature.

2. Methods

The umbrella review followed the reporting requirements of the Preferred Reporting Items for Systematic reviews and Meta-Analyses 2020 (PRISMA) [7] and the methodology outlined in the Cochrane Handbook for Overviews of Reviews [8]. The protocol for this systematic review is registered in the Open Science Framework (OSF) platform (https://osf.io/dzumv).

Studies were included if they were (a) SLRs, with reproducible search filters, or publications on search filters; (b) SLR or search filter was on any automation; (c) SLRs on automation was tested in HEOR, defined as SLRs, indirect treatment comparisons (ITCs) or health economic modeling; (d) full-text publication; (e) written in English; and (f) published in 2023 or later, to capture studies since the availability of ChatGPT.

2.1. Search Strategy

The search was run from January 01, 2023, to July 03, 2024. The review was limited to studies published from January 2023 onwards to capture the post-ChatGPT era, during which large language models (LLMs) became widely accessible and integrated into academic and clinical discourse. This period also reflects a shift in indexing terminology, focus areas, and applications relevant to current LLM capabilities [9].

Terms on automation and artificial intelligence (AI) were combined with terms for SLR, ITCs, or economic models. The full search strategy is available in **(Supplement Table 1)**. Studies were searched in EMBASE via OVID. EMBASE was selected as the primary database because it offers comprehensive coverage of biomedical, health technology, and pharmacoeconomic literature, and provides more detailed indexing of artificial intelligence and automation-related subject headings than other databases, such as MEDLINE. Its broad inclusion of conference abstracts and global publications also enhances the capture of emerging AI-HEOR evidence, which is often presented in early or technical formats before journal publication.

The reference lists of included studies were checked, and forward citation chasing was conducted in Google Scholar to identify any missing SLRs. The InterTASC Information Specialists' Sub-Group (ISSG) Search Filter Resource was also manually searched to identify additional search filters.

2.2. Selection Process

Studies were screened by one reviewer, with 20% of records checked by a second reviewer to ensure accuracy in study selection. Any discrepancies were resolved through discussion and consensus, with a third reviewer available for arbitration if needed, although this was not required. This process was designed to ensure consistency and mitigate potential selection bias inherent in single-reviewer screening approaches.

2.3. Data Extraction

All studies were extracted in Microsoft Excel® using a piloted data extraction template. A second reviewer validated all data. Details on SLR characteristics (objectives, search sources, inclusion criteria, exclusion criteria); search terms; and search results (number of records screened, number of reports retrieved, and number of studies included) were extracted. The time to include studies during the SLR phases of title/abstract screening and full-text screening was directly extracted from the included reviews, not calculated by the authors.

2.4. Quality Assessment

The Peer Review of Electronic Search Strategies (PRESS) was used to evaluate the electronic search strategies [18].

2.5. Evidence Synthesis

Subject headings and search terms used across search filters were qualitatively summarized. In addition, search performance was assessed at the title and abstract, full-text, and overall levels **(Figure 1)**.

3. Results

A total of 11,319 records were identified from EMBASE via OVID. No duplicates were identified after the records were imported into EndNote X9 software for deduplication. Out of these, 25 reports were retrieved for full-text review. Among these 25 reports, five SLRs were included. A flow of literature and reasons for exclusion are outlined in **(Figure 2)**.

An additional three reports were included: a search filter from the ISSG Search Filter Resource [17] and two SLRs identified through forward citation chasing in Google Scholar [10, 15]. Among the eight reports [10–17], seven were search filters used in an SLR [10–16] and one was a standalone search filter on the ISSG website [17]. One SLR, combined searches from four previously identified reviews on SLR automation [15]. As a result, this umbrella review identified 11 individual search filters.

Among the seven SLRs [10–16], two had objectives related to the automation and processing of biomedical literature, as well as to the preparation of SLRs [11, 14]. Two evaluated the automation of any SLR task [10, 15]; one restricted to the automation of data extraction in preparation of an SLR [13]; another to the automation of data extraction from grey literature and soft data [12]. Lastly, one SLR was restricted to automation used in the screening of cancer topics [16]. The number of title and abstract records screened across the SLRs ranged from 3,947 [16] to 27,542 [13]; between 7 [16] and 721 [14] full-texts assessed for eligibility; and between 5 [16] and 273 [11] reports included. A summary of the seven SLRs is provided in **(Table 1)**.

Table 1: Summary of Included SLRs

Primary author	Year	Objectives	Inclusion criteria	Exclusion criteria	Title/abstrac records	Full-text reports	Included reports
de la Torre-Lopez [10]	2023	To determine which phases of the SLR process have been automated using AI, identify supporting AI techniques, and assess human involvement in AI-based SLR automation.	Automation of one or more SLR phases using an AI-based approach.	Non-English language reports.	9,027	50	34
Santos [11]	2023	To examine AI use in automated or semi-automated analysis of biomedical literature and identify state-of-the-art methods and knowledge gaps.	AI methods and tools applied to medical literature; AI tools used in preparation of SLRs.	AI in diagnosis or treatment; AI applied to non-literature medical documents (e.g., electronic medical records).	12,145	316	273
Schmidt [12]	2023a	To review methods and tools for (semi-)automated data extraction in medical SLRs using a living SLR approach.	Original NLP-based full-text studies describing complete implementation and evaluation cycles; published from 2005; data from RCTs or comparative observational studies.	Image-only processing; protocol-only or synthesis-only studies; electronic health records or genetic data mining.	27,542	278	76
Schmidt [13]	2023b	To provide an overview of automated data extraction methods for health-related research using grey and soft data.	Original extraction tools or methods; non-peer-reviewed healthcare datasets; English-language full texts.	Patient-level electronic health records; genomic or biological data extraction.	8,927	131	84
Doneva [14]	2024	To identify BioNLP tasks addressed using large language models, map LLM architectures, and assess methodological transparency.	LLMs applied to biomedical text; SLRs or meta-analyses using LLMs for automation.	Non-English language reports; clinical questionnaires or surveys; reviews.	13,825	721	197
Toth [15]	2024	To determine which SLR stages were automated, the applied tools and data sources, and the research impact of SLR automation studies.	SLRs using automation tools to assist or replace human judgment tasks; full-text reports.	Non-English language reports.	5,321	411	123
Yao [16]	2024	To evaluate accuracy and workload savings of AI-based screening tools versus human reviewers in cancer-related SLRs.	AI tools for automated screening; cancer topics; reported sensitivity/specificity or workload savings.	Non-English language reports; letters, editorials, commentaries; non-public tools.	3,947	7	5

AI, Artificial Intelligence; LLM, Large Language Models; MA, Meta-analysis; NLP, Natural Language Processing; PDFs, Portable Document Formats; RCT; Randomized Controlled Trial; SLR, Systematic Literature Revi

Across the seven SLRs [10–16], 16 bibliographic databases were searched, with PubMed used most frequently (6 out of 7 SLRs) [11–16]; followed by EMBASE [11, 14, 16], Web of Science [10, 11, 13] and Institute of Electrical and Electronics Engineers (IEEE) [10, 11, 13], each searched by three SLRs. In addition to IEEE, three more bibliographic databases were identified and searched by two SLRs each, specializing in computing and technology; these databases were: the Association for Computing Machinery (ACM) Digital Library [10, 11]. ACL Anthology [12, 13], and dblp [12, 13]. For a full list of bibliographic databases searched by the SLRs, see **(Table 2)**. One of the reports identified was a search filter published by the University of Alberta, and available on the ISSG Search Filter Resource website [17]. The search filter, which is not validated, was translated for EMBASE, MEDLINE, PsycINFO, Cumulative Index to Nursing and Allied Health Literature (CINAHL), Scopus, and Cochrane Library [17].

Of the eight reports summarizing 11 search filters on automation, five used subject headings [11–14, 17]. Four search filters exploded the subject headings ‘artificial intelligence’ [11–13, 17], ‘deep learning’[11–13, 17], ‘machine learning’[11–13, 17], and ‘natural language processing’ [11–14]. Followed by two search filters, each exploding subject headings on ‘data mining’ [12, 13], ‘supervised machine learning’ [13, 17], support vector machine’[13, 14], and ‘unsupervised machine learning’ [13, 17]. Subject headings such as ‘artificial intelligence’, ‘automation’, ‘data mining’, ‘machine learning’ (using both the exact term and the wildcard search ‘machine learn*’), and ‘natural language processing’ were also

Table 2: Electronic Databases Searched for Literature on AI

Database	de la Torre-Lopez	Santos	Doneva	Schmidt(a)	Schmidt(b)	Toth	Yao
PubMed		x	x	x	x	x	x
Embase		x	x				x
IEEE Xplore	x	x	x				
Web of Science	x	x		x			
ACL-Anthology				x	x		
ACM Library	x	x					
Cochrane Library		x					x
dblp				x	x		
Scopus	x				x		
arXiv					x		
CINAHL		x					
Google Scholar		x					
MedRxiv					x		
Prospero						x	
SciELO		x					

ACM, Association for Computing Machinery; ACL, Association for Computational Linguistics; ACM, Association for Computing Machinery; AI, Artificial Intelligence; CINAHL, Cumulative Index to Nursing and Allied Health Literature; dplp, database systems and logic programming; IEEE, Institute of Electrical and Electronics Engineers; SciELO, Scientific Electronic Library Online.

searched as major headings in PubMed and as descriptors in EMBASE. For a full list of subject headings searched, see (**Table 3**).

Across the 11 search filters, 196 variations of free text terms were used to capture studies on automation. Less than a third of these free-text terms were used by more than one search, with only twelve used in three or more search filters. "Artificial intelligence" was the most frequently utilized in all search filters, followed by "machine learning" and "natural language processing", each present in six search filters (see (**Figure 3**). For the full list of search terms used, please see (**Supplement Table 2**).

When grouping subject headings and free text terms into common concepts, for example, the following search terms were grouped under machine learning (general terms): "machine learning", "ml", [machine NEXT/1 (intelligence OR learning)], exp machine learning/, machine learning/de); "machine learning (general terms)" and "automation" were the most common constructs searched across the search filters (8 search filters each). This was followed by "mining" and "natural language processing" (7 search filters each), and "text grouping", "neural network", "identification, interpretation and pattern recognition", "support vector machine", and "artificial intelligence" (6 search filters each). "Deep learning" and "machine learning approaches" were searched in 5 search filters, while "semi automation" and "decision support systems" were searched in 4. The remaining concepts were searched in 3 ("AI models and application", "extraction", and "text analysis"), and 2 search filters ("knowledge systems"). The cluster with the most variation in terms used was "AI models and applications", followed by "machine learning approaches". All concepts identified are shown in (**Figure 4**), with a list of terms for each concept in (**Supplement Table 3**). Similar search filter characteristics were identified by the PRESS checklist (**Supplement Table 4**): only 5 of the included reports used subject headings, and the majority of the search filters reported a narrow list of keywords. Only five of the 11 search filters included truncation. No incorrect use of system syntax and spelling errors were identified. A mapping framework (**Supplement Table 5**) was developed to link each included automation review or search filter to HEOR-relevant evidence-generation tasks. Reviews were categorized by their core automation focus (e.g., AI/NLP, ML, decision systems) and cross-referenced with HEOR functions, including SLR automation for economic model inputs, ITCs, and HTA submissions. Each review/filter was coded for "HEOR applicability (Y/N)" with justification provided under the column "Basis."

Among the seven SLRs [10–16], one of which searched using 4 previously identified search filters, and combined using "or" [15]; none had a title and abstract inclusion rate (defined as the proportion of studies selected for full-text retrieval out of the total number of title and abstract records screened) above 8%, with two searches below 1% [10, 16] (see (**Table 4**)). The full-text inclusion rate (defined as the proportion of studies included after full-text review compared to the total number of full-text articles retrieved for evaluation) varied from 27% [13, 14] to 86% [11]. The overall inclusion rate (defined as the proportion of studies included after the full review compared to the total number of title and abstract records initially screened) ranged from 0.13% [16] to 2.31% [15]. When evaluating the number of hours needed to include one study the SLR by Santos [11], followed by Toth [15] had the best performance (0.6 hours and 0.8 hours, respectively), followed by Doneva (1.1 hours) [14], Schmidt(b) (1.2 hours) [12], De la Torre-Lopes (2.8 hours), Schmidt(a) (4.0 hours) [13], with Yao, which was only interested in AI used in HEOR among oncology studies, requiring 8 hours of screening to identify one included study [16] (see (**Table 4**)). Of note: performance metrics could not be calculated for the standalone search filter identified from the ISSG Search Filter resource, as it did not present a flow of literature

Table 3: Subject Headings Used to Identify Literature on AI

Search Filter	Corresponding Citations
artificial intelligence/	Campbell [17], Schmidt(a) [12], Schmidt(b) [13]
artificial intelligence/de	Doneva [14]
artificial intelligence[MH]	Schmidt(a) [12]
artificial intelligence[MeSH Terms]	Santos [11]
artificial neural network/exp	Doneva [14]
automated pattern recognition/	Doneva [14]
automation/	Schmidt(b) [13]
automation[MH]	Schmidt(a) [12]
automated[MH]	Schmidt(a) [12]
automating[MH]	Schmidt(a) [12]
data mining/	Schmidt(a) [12], Schmidt(b) [13]
data mining[MH]	Schmidt(a) [12]
decision trees/	Schmidt(b) [13]
deep learning/	Campbell [17], Schmidt(a) [12], Schmidt(b) [13]
deep learning [MeSH Terms]	Santos [11]
knowledge bases/	Schmidt(b) [13]
learning algorithm*[MH]	Schmidt(a) [12]
literature mining[MH]	Schmidt(a) [12]
machine learn*[MH]	Schmidt(a) [12]
machine learning/	Campbell [17], Schmidt(a) [12], Schmidt(b) [13]
machine learning/de	Doneva [14]
machine learning [MeSH Terms]	Santos [11]
natural language processing/	Schmidt(a) [12], Schmidt(b) [13], Doneva [14]
natural Language processing[MH]	Schmidt(a) [12]
natural Language processing [MeSH Terms]	Santos [11]
neural networks (computer)/	Schmidt(b) [13]
neural networks/	Schmidt(a) [12]
neural network [MeSH Terms]	Santos [11]
predictive modelling[MH]	Schmidt(a) [12]
semi-automated[MH]	Schmidt(a) [12]
semi-automating[MH]	Schmidt(a) [12]
semi-automation[MH]	Schmidt(a) [12]
supervised machine learning/	Schmidt(a) [12], Schmidt(b) [13]
support vector machine/	Schmidt(b) [13], Doneva [14]
text mining[MH]	Schmidt(a) [12]
unsupervised machine learning/	Schmidt(a) [12], Schmidt(b) [13]

MH, Medical Subject Headings; MeSH, Medical Subject Headings.

as part of an SLR. These time metrics were directly extracted from the included reviews, not calculated by the authors.

For this umbrella review, a comprehensive search was conducted in EMBASE (via Ovid), the InterTASC Information Specialists' Sub-Group (ISSG) Search Filter Resource, and Google Scholar. The search covered studies published between January 1, 2023, and July 3, 2024, to capture the post-ChatGPT era when large language models (LLMs) became widely adopted. Search terms combined concepts for automation and artificial intelligence (AI) with terms representing systematic evidence synthesis (systematic review, indirect treatment comparison, economic model). In contrast, the included SLRs and search filters used a variety of databases and approaches. Across the eight included publications, 16 databases were searched, with PubMed being the most frequently used (6/7 SLRs), followed by EMBASE (3/7). This distinction between our

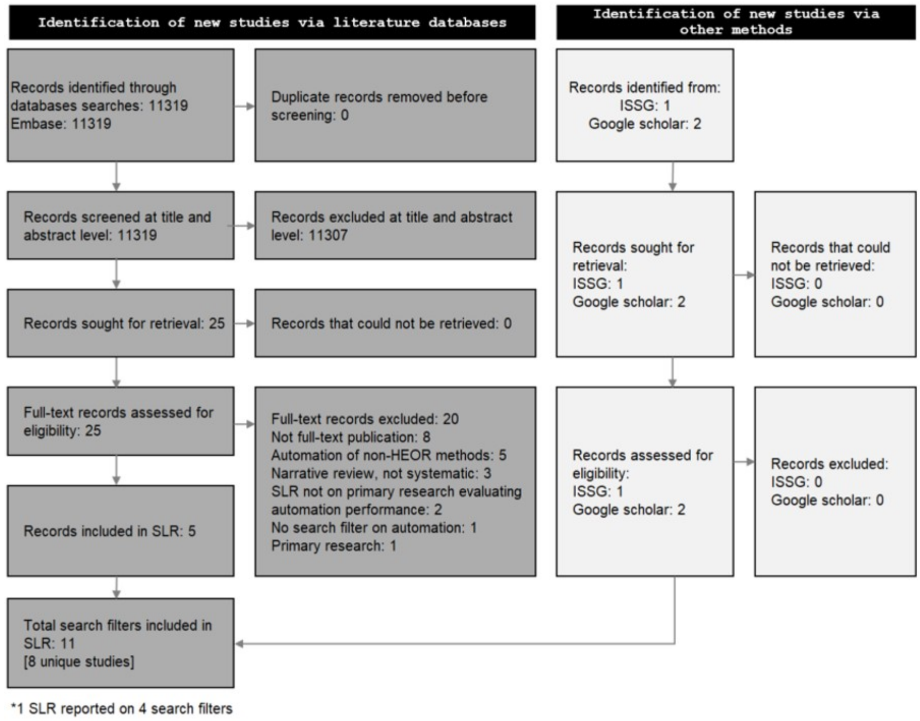

Figure 2: PRISMA flow diagram of the Umbrella review study selection

HEOR, Health Economics Outcomes Research; ISSG, InterTASC Information Specialists' Sub-Group; SLR, Systematic Literature Review

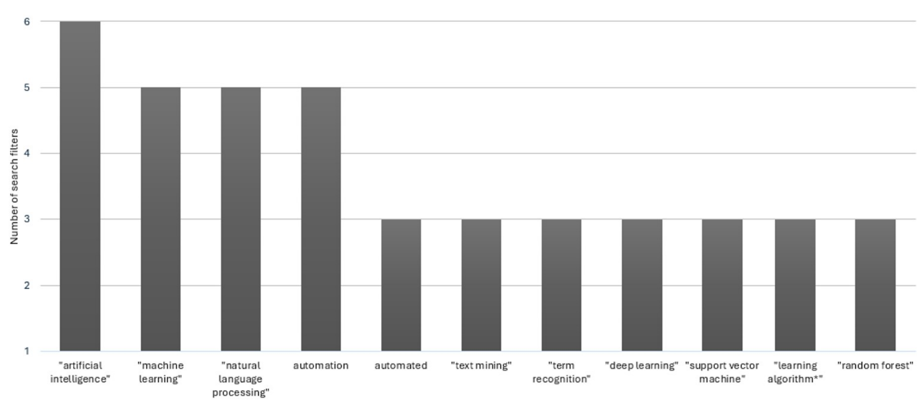

Figure 3: Free text terms used in three or more search filters.

umbrella review's search process and that of the included reviews is summarized in Table X ("Databases Searched: This Review vs. Included SLRs"). Among the eight included publications (seven SLRs and one standalone filter), three reviews (38%) explicitly evaluated automation methods within HEOR-relevant contexts, such as SLRs for cost-effectiveness models or indirect treatment comparisons.

4. Discussion

There is very little overlap or consensus among existing search filters for identifying papers on AI in HEOR. Of the 11 search filters identified, only five used subject headings, and among these, no subject heading was searched across all five. For example, only four out of 31 different subject headings ('artificial intelligence', 'deep learning', 'machine learning', and 'natural language processing') were searched across four of the five filters. There were also differences in how subject headings were searched: as major headings in PubMed and as descriptors in EMBASE. The PRESS checklist further confirmed the deficits in the search filters, identifying fewer reports using subject headings and the majority of the searches reporting a narrow list of keywords. Quantitative overlap analysis (e.g., Jaccard similarity) was not conducted because the terminology across filters was highly heterogeneous and inconsistently formatted, making reliable computation of similarity indices infeasible without extensive normalization.

A lack of consistency in search filters was identified among free-text terms, with less than a third used by more than one search, and only 12 by more than three searches. With the abundance of

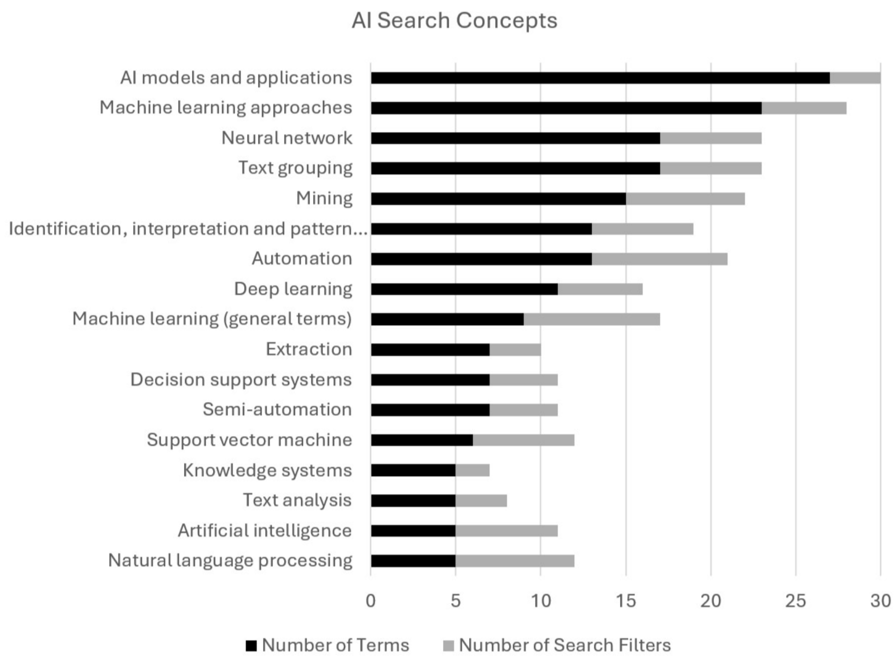

Figure 4: AI Constructs Searched in the search filters.

Table 4: Inclusion rates at different stages of SLR in the included studies

Primary Author	Title and Abstract (%)	Full Text (%)	Overall (%)
de la Torre-Lopez [10]	0.55%	68.00%	0.38%
Doneva [14]	5.22%	27.32%	1.42%
Santos [11]	2.60%	86.39%	2.25%
Schmidt(a) [12]	1.01%	27.34%	0.28%
Schmidt(b) [13]	1.47%	64.12%	0.94%
Toth [15]	7.72%	29.93%	2.31%
Yao [16]	0.18%	71.43%	0.13%

SLR, Systematic Literature Review

Table 5: Number of hours required at different stages of SLR in the included studies

Primary Author	Hours for Title Abstract Screening	Hours for Full Text Screening	Overall Screening Hours	Ratio of Hours to Each 1 Included Study
de la Torre-Lopez [10]	90	5	95	2.8
Doneva [14]	138	72	210	1.1
Santos [11]	121	32	153	0.6
Schmidt(a) [12]	275	28	303	4.0
Schmidt(b) [13]	89	13	102	1.2
Toth [15]	53	41	94	0.8
Yao [16]	39	1	40	8.0

SLR, Systematic Literature Review

AI research in the HEOR field and no uniform, validated search filter available, it is not surprising that the SLRs ranged from just under 4,000 to over 27,000 records. This equated to a wide range of title and abstract screening time, from 0.6 hours to 8 hours, to identify one relevant study. This wide variation underscores the potential importance of a validated, optimized filter that can supplement rapid access to high-quality evidence supporting timely health technology assessment and reimbursement decisions.

To the best of the author's knowledge, this umbrella review appears to be the first systematic review to explore the identification and performance of existing AI search filters in HEOR. One of the identified SLRs combined the search strategies of previous reviews to provide a comprehensive overview of automation studies in the SLR process, but did not evaluate the search filters themselves. Our research identified seven additional search filters that were not included in the previous publication. Our research provides a focused synthesis of the overlap among existing search filters for retrieving AI studies. It explores performance using metrics such as the title-abstract inclusion rate, full-text inclusion rate, and overall inclusion rate.

With key decision-making bodies such as CADTH, INESSS, and NICE exploring the use of AI in HTA submissions, as a community, we must stay ahead of new research to adopt the most reliable and proven automation methods in HEOR. To do this, we must develop a validated search filter that balances the sensitivity and specificity of AI records, specifically for HEOR, to make this manageable. We observed from our search that we had to screen 11,319 records, to full-text review 25 reports, and to include 5 reports from the electronic database search. Most research was excluded at the title and abstract stages because it focused on the direct application of AI in medical and healthcare settings, not HEOR. By creating a tested and validated search strategy, it is more viable that a living SLR on this topic can be maintained, with insights shared with HEOR experts and decision-makers.

Although this was a systematic, umbrella review of search filters used in AI research for HEOR, a validated search filter was not used, as none had yet been developed. In addition, only a single screening was conducted with quality checks, which may indicate some human error in missing studies; however, this was mitigated by reference checking and forward citation chasing. Only EMBASE was searched, while the field of AI may require automation-specific databases such as the ACM Digital Library and IEEE; this needs to be thoroughly investigated. A systematic search restricted to the EMBASE database represents a methodological limitation, especially in the context of AI and LLMs, where relevant work often appears in interdisciplinary and technical venues. Rapidly evolving AI terminology can shorten the longevity of search filters, as keywords and concepts may quickly become outdated or replaced. This can reduce search precision and recall over time, requiring regular updates or adaptive filtering to maintain relevance and effectiveness. As included SLRs were not quality-assessed using a validated tool, such as the AMSTAR2 tool, this may affect the strength of the conclusions. Lastly, restricting studies published in 2023 or later might introduce bias and limit the generalizability of the review's findings. Broader multilingual and multi-database searches could yield a more accurate picture of true variability across automation-related reviews.

To ensure technical accuracy and reproducibility, all search filters extracted from the included reviews were verified against the native database syntax for both PubMed (MEDLINE) and EMBASE (via Ovid). Verification focused on identifying and correcting invalid or outdated field tags, truncation errors, Boolean logic inconsistencies, and incorrect use of subject headings. For PubMed, MeSH terms were verified against the current 2024 MeSH database. Only exact headings (e.g., "Artificial Intelligence"[Mesh]) were retained; wildcards and free-text modifiers were removed from MeSH field tags. Free-text terms were confirmed to use the [tiab] or [tw] fields as appropriate. For EMBASE, Emtree terms were verified using the 2024 Emtree browser. Exploded subject headings (e.g., 'artificial intelligence'/exp) were checked for availability and correct hierarchy placement. Incorrect truncation (e.g., "machine learn"/de*) and misplaced slashes were corrected according to Ovid syntax.

Future studies must focus on using these initial findings to create a validated search filter on AI in HEOR. Filters will need to explore the right AI search terms, but the most appropriate combination of terms to identify research in HEOR specifically. Suppose a search can be highly specific while still sensitive enough to ensure relevant studies are not missed. In that case, we can start standardizing how the industry searches for research in a rapidly evolving field. However, any search filter for AI must acknowledge that periodic review will be required as new AI methods emerge that may require additional search concepts, and as older methods are retired.

This umbrella review identified significant variability in the search filters used for identifying literature on automation in HEOR, highlighting the lack of a standardized approach. The low inclusion rates across the SLRs indicate challenges in optimizing search filter performance. The variation in search strategies and the high yield of records on AI in HEOR call for more research to develop, validate, and standardize search filters in this field to optimize efficiency and keep up to date with the ongoing stream of new evidence.

5. Conclusion

This umbrella review identified significant variability in the search filters used for identifying literature on automation in HEOR, highlighting the lack of a standardized approach. The low inclusion rates across the SLRs indicate challenges in optimizing search filter performance. The variation in search strategies and the high yield of records on AI in HEOR call for the need for more research to create, validate, and standardize search filters in this field to optimize efficiency, and keep up to date, with ongoing stream of new evidence.

Conflicts of Interest

The authors declare no competing interests that could have influenced the objectivity or outcome of this research.

Funding Source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgment

None

Institutional Review Board (IRB)

None

Large Language Model

None

Authors Contribution

MR conceptualised the study. MR and OI conducted the literature search, study screening, selection, and data extraction, and drafted the manuscript. MR drafted the initial manuscript, and OI reviewed and revised the manuscript. All authors critically reviewed the manuscript for important intellectual content and approved the final manuscript as submitted. MR is the guarantor.

Data Availability

No new datasets were generated or analyzed in this umbrella review. All data supporting the findings are derived from previously published sources and are presented within the article and its supplementary materials.

References

1. van Nooten F, Holmstrom S, Green J, Wiklund I, Odeyemi IA, Wilcox TK. Health economics and outcomes research within drug development: challenges and opportunities for reimbursement and market access within biopharma research. Drug Discov Today. 2012;17(11-12):615-22. [PMID: 22366662, https://doi.org/10.1016/j.drudis.2012.01.021].

2. Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial Intelligence to Automate Health Economic Modelling: A Case Study to Evaluate the Potential Application of Large Language Models. Pharmacoecon Open. 2024;8(2):191-203. [PMID: 38340276, PMCID: PMC10884386, https://doi.org/10.1007/s41669-024-00477-8].
3. NICE. Use of AI in evidence generation: NICE position statement; 2024. Available from: https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation-nice-position-statement. Accessed on: 09 Nov 2025.
4. Mason J, Morrison A, Visintini S. An overview of clinical applications of artificial intelligence. Ottawa: CADTH. 2018.
5. Health NIoEi, Services S. Development and evaluation of a software tool based on GPT-4 to assist in sorting documents in the context of literature reviews: Literature, a proof of concept; 2024. Available from: https://www.inesss.qc.ca/fileadmin/doc/INESSS/Demarche/INESSS_Outil_tri_GPT_4_VF.pdf. Accessed on: 09 Nov 2025.
6. Escobar Liquitay CM, Garegnani L, Garrote V, Sola I, Franco JV. Search strategies (filters) to identify systematic reviews in MEDLINE and Embase. Cochrane Database Syst Rev. 2023;9(9):MR000054. [PMID: 37681507, PMCID: PMC10485899, https://doi.org/10.1002/14651858.MR000054.pub2].
7. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. [PMID: 33782057, PMCID: PMC8005924, https://doi.org/10.1136/bmj.n71].
8. Pollock M, Fernandes RM, Becker LA, Pieper D, Hartling L. Chapter V: overviews of reviews. Cochrane handbook for systematic reviews of interventions version. 2020;6.
9. OpenAI. GPT-4 Technical Report 2023; 2023. Available online: https://openai.com/research/gpt-4. Accessed on: 09 Nov 2025.
10. de la Torre-López J, Ramírez A, Romero JR. Artificial intelligence to automate the systematic review of scientific literature. Computing. 2023;105(10):2171-94. [https://doi.org/10.1007/s00607-023-01181-x].
11. Santos AOD, da Silva ES, Couto LM, Reis GVL, Belo VS. The use of artificial intelligence for automating or semi-automating biomedical literature analyses: A scoping review. J Biomed Inform. 2023;142:104389. [PMID: 37187321, https://doi.org/10.1016/j.jbi.2023.104389].
12. Schmidt L, Mohamed S, Meader N, Bacardit J, Craig D. Automated data analysis of unstructured grey literature in health research: A mapping review. Res Synth Methods. 2024;15(2):178-97. [PMID: 38115736, https://doi.org/10.1002/jrsm.1692].
13. Schmidt L, Finnerty Mutlu AN, Elmore R, Olorisade BK, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: Update of a living systematic review. F1000Res. 2021;10:401. [PMID: 34408850, PMCID: PMC8361807, https://doi.org/10.12688/f1000research.51117.3].
14. Doneva SE, Qin S, Sick B, Ellendorff T, Goldman JP, Schneider G, et al. Large language models to process, analyze, and synthesize biomedical texts: a scoping review. Discover Artificial Intelligence. 2024;4(1). [https://doi.org/10.1007/s44163-024-00197-2].
15. Toth B, Berek L, Gulacsi L, Pentek M, Zrubka Z. Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in PubMed. Syst Rev. 2024;13(1):174. [PMID: 38978132, PMCID: PMC11229257, https://doi.org/10.1186/s13643-024-02592-3].
16. Yao X, Kumar MV, Su E, Flores Miranda A, Saha A, Sussman J. Evaluating the efficacy of artificial intelligence tools for the automation of systematic reviews in cancer research: A systematic review. Cancer Epidemiol. 2024;88:102511. [PMID: 38071872, https://doi.org/10.1016/j.canep.2023.102511].
17. J G, C L, P M, S R, I B, Woods L e. ISSG Search Filter Resource; 2006. York (UK): The InterTASC Information Specialists' Sub-Group; updated 17 September 2025. Available from: https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home. Accessed on: 09 Nov 2025.
18. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. J Clin Epidemiol. 2016;75:40-6. [PMID: 27005575, https://doi.org/10.1016/j.jclinepi.2016.01.021].