Using artificial intelligence methods for systematic review in health sciences: a systematic review.

BLAIZOT, A., VEETTIL, S.K., SAIDOUNG, P., MORENO-GARCIA, C.F., WIRATUNGA, N., ACEVES-MARTINS, M., LAI, N.M. and CHAIYAKUNAPRUK, N.

2022

Blaizot Aymeric (Orcid ID: 0000-0002-4677-6168)
Aceves-Martins Magaly (Orcid ID: 0000-0002-9441-142X)
Chaiyakunapruk Nathorn (Orcid ID: 0000-0003-4572-8794)

Using artificial intelligence methods for systematic review in health sciences: A systematic review

Aymeric Blaizot[1], Sajesh K. Veettil[1], Pantakarn Saidoung[2], Carlos Francisco Moreno-Garcia[3], Nirmalie Wiratunga[4], Magaly Aceves-Martins[4], Nai Ming Lai[5,6*], Nathorn Chaiyakunapruk[1,7*]

Affiliations:

[1] Department of Pharmacotherapy, College of Pharmacy, University of Utah, Utah

[2] Faculty of Pharmacy, Chiang Mai University, Chiang Mai, Thailand

[3] School of Computing, Robert Gordon University, Aberdeen, Scotland

[4] The Rowett Institute, University of Aberdeen, Aberdeen, Scotland

[5] School of Medicine, Faculty of Health and Medical Sciences, Taylors University, Selangor, Malaysia

[6] School of Pharmacy, Monash University Malaysia, Selangor, Malaysia

[7] IDEAS Center, Veterans Affairs Salt Lake City Healthcare System, Salt Lake City, Utah

*** Corresponding author:**

Nathorn Chaiyakunapruk Pharm D, PhD

Professor, Department of Pharmacotherapy, College of Pharmacy

The University of Utah, 30 2000 E, Salt Lake City, UT 84112, USA

Tel: (801) 585-3092

E-mail: nathorn.chaiyakunapruk@utah.edu

Nai Ming Lai MBBS

Professor, School of Medicine, Faculty of Health and Medical Sciences

Taylor's University, 47100 Subang Jaya, Selangor, Malaysia

E-mail: lainm123@yahoo.co.uk

Abstract

Background: The exponential increase in published articles makes a thorough and expedient review of literature increasingly challenging. This review delineated automated tools and platforms that employ artificial intelligence (AI) approaches and evaluated the reported benefits and challenges in using such methods.

Methods: A search was conducted in 4 databases (Medline, Embase, CDSR, and Epistemonikos) up to April 2021 for systematic reviews and other related reviews implementing AI methods. To be included, the review must use any form of AI method, including machine learning, deep learning, neural network, or any other applications used to enable the full or semi-autonomous performance of one or more stages in the development of evidence synthesis.

Results: Twelve reviews were included, using nine different tools to implement 15 different AI methods. Eleven methods were used in the screening stages of the review (73%). The rest were divided: two in data extraction (13%) and two in risk of bias assessment (13%). The ambiguous benefits of the data extractions, combined with the reported advantages from 10 reviews, indicating that AI platforms have taken hold with varying success in evidence synthesis. However, the results are qualified by the reliance on the self-reporting of the review authors.

Conclusion: Extensive human validation still appears required at this stage in implementing AI methods, though further evaluation is required to define the overall contribution of such platforms in enhancing efficiency and quality in evidence synthesis.

Introduction

Systematic reviews are fundamental to evidence-based decision-making, as they use a comprehensive search and synthesis of the available literature. Such an operation usually requires a team of reviewers to evaluate thousands of articles. With the exponential increase in published articles, more time is needed to review existing literature thoroughly. It has been reported that the average time to complete a systematic review is over 15 months.(1) The current methods of biomedical indexing may have contributed to inefficiency in screening, as the proportion of truly relevant articles may be as low as 1% of the total search yield with a typical search strategy.(1) The long conception-completion interval may render a systematic review outdated by the time they are ready to be submitted and published.(2) Consequently, more expedient methods of screening and data extraction are being developed and employed, and some make use of artificial intelligence (AI) methods. These methods employ various algorithms related to Machine Learning (ML) and Natural Language Processing (NLP) tasks. On the one hand, ML algorithms have the purpose of making automated decisions based on samples of data rather than a fixed mathematical function.(3) On the other hand, NLP refers to the interpretation of human language by a computer, allowing these algorithms to extract the relevant information so that it can be further processed by ML (or other) based algorithms for its interpretation, understanding, answer generation, etc [a]. These methods are used, for instance, to expediting the process of systematic reviews and other evidence synthesis endeavors, such as scoping and rapid reviews that employ similar methodologies. Studies have demonstrated the promise of using AI platforms to reduce the human labor required for an extensive literature review(4–6); however, there is significant doubt on the actual utility of these newly emerging platforms within the community of reviewers.

With automated systems poised to give significant benefits to systematic reviews, we conducted a systematic review to delineate the common automated tools and platforms that employ AI approaches and evaluate the reported benefits and challenges in using such methods.

Methods

Following the referred Preferred Reporting Items for Systematic Reviews and Meta-Analyzes (PRISMA) statement, this review was registered with the International Prospective Register of Systematic Reviews (PROSPERO CRD42021249245).

Literature Search

A search was conducted in Medline, Embase, Cochrane database of systematic reviews and Epistemonikos from database inception to April 2021. The search strategy is provided in Appendix 1.1. After running a scoping search, the names of the most common automation tools were retrieved and included in the additional search (see the list of tools in Appendix 1.2). No language restriction was applied in the search. We also manually searched the cited references of the retrieved articles and reviews. Two reviewers (S.V. and PS.) searched titles or abstracts in Covidence systematic review software (available at www.covidence.org) for eligibility independently. For the full-text screening, two reviewers (AB. and PS.) scrutinized each study for eligibility independently through Excel.

Then, the spreadsheets were combined, and the discrepancies were resolved by a third reviewer (S.V.).

Inclusion criteria

We included AI-assisted systematic reviews and similar reviews, such as rapid reviews, umbrella reviews, evidence gap maps, evidence mapping, and scoping reviews on health science involving human subjects. To be included, an AI-assisted review must use any form of AI method, including machine learning, deep learning, neural network, or any other applications that are used to enable the full or semi-autonomous performance of one or more stages in the development of evidence synthesis, published in English. Reviews that used any tools for only data management, such as Covidence were excluded. Abstracts only or conference abstracts were excluded. Narrative reviews, review protocols, and studies that assessed the effectiveness of automation tools in reviews were also excluded.

Data extraction and synthesis

Data extraction was carried out independently by two reviewers (AB. and PS.) and revised by a third reviewer (MA-M), and in case of disagreements, a fourth reviewer was referred to (S.V.). For each eligible review, the following items were extracted: study country, study design, research question, category of health science investigated, type of review, the AI tool employed, stage of the process that involved AI, number of articles that went through AI, description of AI method, description of use, validation of use, reported unplanned human interventions, reported advantages, reported concerns, the types of AI methods used, stages in the development of evidence synthesis where AI methods are employed, and the extent of reliance on AI in evidence synthesis.

We performed descriptive analyzes and presented the findings narratively. This analysis was guided by established questions, as outlined in the PROSPERO protocol. Although there was some overlap in the reported items, specifically between the description of use and description of method, this allowed for more ways to answer the established questions depending on the reporting of the author of the review. Although a quantitative analysis of the registration submission time of AI-assisted systematic reviews compared to manually conducted systematic reviews submitted in the same period was planned, no such comparison could be made, as very few reviews provided clear information on the above.

To grade the methodological quality of included systematic reviews (high, moderate, low, and critically low), we used the revised AMSTAR-2 (A Measurement Tool to Assess Systematic Review, version 2) tool.(7) Although this tool targets systematic reviews, it was applied to the other types of reviews for a consistent evaluation and to give insight into their quality.

Results

Study Selection

From 4579 identified records, 1020 duplicates were removed, and 3514 articles were excluded based on title and abstract screening, leaving 45 articles to be assessed for eligibility. An additional search based on a list of automation tools identified 332 records from which 111 duplicates were removed, and 124 were excluded by title and abstract screening, leaving 97 to be assessed for eligibility. After the deduplication of the 142 records of the two searches, 107 articles were eligible for full-text screening, but 20 were removed because they were abstract only. Of the 87 available full-text studies, 12 were included in this systematic review as 16 were not reviews, 54 did not use AI or machine learning, three were not in English, one was another type of review, and one was duplicate. An overview of the study selection process is shown in **Figure 1.**

Study Characteristics

The 12 included reviews(8–19) implemented 15 different AI methods **(Table 1)**. Among the 12 reviews, there were five (42%) systematic reviews,(8–12) two (17%) systematic reviews with meta-analysis,(13,16) one (8%) review that conducted a quantitative analysis without a qualitative analysis,(18) two (17%) integrative reviews,(14,15) one (8%) rEM (rapid-Evidence-Mapping),(17) and one scoping review (8%).(19) Most of the studies from developed countries with five from the US (42%),(8,10,13,17,18) four from the UK (33%),(8,9,12,19) two from Italy (17%),(8,11) Canada (17%),(11,16) and Brazil (17%),(14,15) one from Australia (8%),(9) Austria (8%),(10) Germany (8%),(10) New Zealand (8%),(16) and China (8%),(16) with five (42%) reviews including multinational collaborations.(8–11,16) All the publications were from 2018 onwards.

Categorically, three of the reviews were etiology and/or risk reviews (25%),(16–18) four effectiveness reviews (33%),(10,11,13,19) two expert opinion/integrative literature reviews (17%),(14,15) two experimental/qualitative reviews (17%),(8,12) and one was a prognostic review (8%).(9) In terms of the health science areas, two reviews investigated mental health (17%),(8,11) one review investigated injury (8%),(9) three reviews investigated cancer (25%),(10,13,18) two reviews investigated medical education (17%),(14,15) three reviews investigated cardiovascular diseases (CVS) (25%),(12,16,19) and one review investigated nutrition-related topics.(17)

The quality of the systematic reviews was evaluated using AMSTAR-2. Of the seven systematic reviews, including those with MA, one was of high quality,(13) four were of moderate quality,(8–11) one was of low quality,(16) and one was critically low quality.(12) The quality of the two integrative reviews was also evaluated using AMSTAR-2 and found to be of critically low quality, which is to be expected as they are not systematic reviews. The other reviews could not be evaluated using AMSTAR-2 as they did not have methodologies with similar quality control as a systematic review (e.g. rapid reviews). More information on the evaluation of the studies is available in Appendix 2.1.

Study Findings

Because of their multi-step nature, multiple AI methods can be implemented within one review. Because of the multi-step nature of systematic reviewing, several AI methodologies could be implemented within one study. Therefore, in this present systematic review, we will consider that only one AI methodology is being implemented on each step of the process. Moreover, some reviews implement the same AI tool in different ways or at different stages of their process. Hence, we consider that each implementation is an AI method of its own. Among the 12 reviews considered in this study, we were able to identify nine different tools, leading to a total of 15 AI methods implemented **(Table 1)**. Each method branches at the tool used or at the stage of the process column **(Table 1)**. The percent of utilization of the nine tools is displayed in **Figure 2**. Five methods employed Rayyan (33%),(11–15) three methods involved Robot Reviewer (20%),(10,19) and one method used EPPI-reviewer (7%),(8) K-means clustering (7%),(16) SWIFT-review (7%),(17) SWIFT-Active Screener (7%),(17) Abstrackr (7%),(9) Wordstat and QDA (Qualitative Data Analysis) Miner (7%),(9) and the natural language processing (7%).(18) It must be noted that Deng *et al.* developed their own NLP tool for the study.

The stages of the reviews that implemented AI methods are shown in **Figure 3**. Out of the 15 methods, 11 were used in the screening stages of the review (73%), which includes eight title and abstract screenings,(8,9,11,13–17) and one post-protocol screening,(12) one abstract classification and filtering and text mining,(18) and one full-text screening.(9) Of the methods employed in other stages, two were used in data extraction (13%)(10,17) and two in risk of bias assessment (13%).(10,19) Abstract classification and filtering, and text mining methods are categorized as screening because they facilitated the triage of articles fulfilling the same purpose as a traditional title and abstract assessment. Although the post-protocol screening took place after the data extraction, it was also categorized as screening since it contributed to the number of included studies for that review.

Three of the 15 AI/ML methods (20%) (EPPI-Reviewer 4,(8) K-means clustering,(16) and one of the Rayyan methods(12) were fully autonomous, meaning that AI operated on without continuous input from reviewers. The rest of the AI/ML operated semi-autonomously (or human in the loop), requiring sustained human input to confirm the articles' relevance.

Method validation was conducted in 11 of 15 methods (73%). For Russel *et al.,* decisions about inclusion made in the AI method were independently reassessed by senior authors.(8) Giummarra *et al.* resolved disagreements about eligibility through discussion and consultation with senior authors for AI-assisted title and abstract screening and full-text screening.(9) Goldkuhle *et al.* had two review authors resolving discrepancies in their data extraction and the AI method through discussion.(10) Pinna *et al.* resolved disagreements emerging from the AI-assisted title and abstract screening through joint discussion with a senior reviewer.(11) Gaskins *et al.* manually re-screened the relevant articles provided by the post-protocol screening.(12) Siqueira *et al.* verified that the studies selected by the researchers in the AI-assisted title and abstract screening were the same.(14) Nascimento *et al.* managed the divergence generated in the AI method by sending discrepancies in inclusion and exclusion to a third individual.(15) Xiong *et al.* and Deng *et al.* conducted a duplicate manual title and abstract screening alongside the AI method.(16,18) In Aali *et al.*, one reviewer double-checked and revised RobotReviewer's risk of bias assessment.(19)

Out of the 15 AI methods, reviewers reported concerns in 5 (33%) of them. Giummarra *et al.* reported a risk of missing a few relevant studies with both the Abstrackr title and abstract screening and the WordStat and QDA full-text screening. However, a separate evaluation of the methods suggests they were not detrimental to the review's integrity.(9) Goldkuhle *et al.* reported issues with the software (Robot Reviewer) recognizing randomized controlled trials (RCTs) where there were none.(10) This concern led to human intervention, as reviewers could not use the extracted data because it erroneously flagged two of them as RCTs. The reviewers had to conduct data extraction manually. Lam *et al.* reported that the software (SWIFT-Review) could not automate all aspects of data extraction.(17) A human intervention was required to manually extract study sample size and review of automated tagging for each category as SWIFT-Review did not perform those tasks effectively. Deng *et al.* reported that the natural language processing missed one paper out of ten critical to the review.(18)

In 10 of the 12 reviews (83%), authors reported advantages in using the AI-assisted method. Giummarra *et al.* reported that the AI methods (Abstrackr, WordStat, and QDA Miner) reduced workload demands.(9) Four of the reviews that used Rayyan in their methods reported it helped expedite title and abstract screening and had a high level of usability.(11,13–15) Siqueira *et al.* also reported that Rayyan made the initial triage process of abstract and title reading faster in the initial triage process.(14) Gaskins *et al.* reported that Rayyan was efficient, accurate, and freely available while increasing recall of relevant studies, thereby strengthening the review.(12) Xiong *et al.* reported that the K-means clustering algorithm facilitated study selection, stating that "the burden of manual screening is reduced from all the articles returned by the initial online strategic search to those in the training set and in the principal cluster(s) identified by supervised machine learning."(16) Lam *et al.* reported substantial time savings relative to similar studies due to the use of SWIFT Active screener and SWIFT-Reviewer. However, the specific time saved was not quantifiable because of confounding variables.(17) Deng *et al.* reported a six-fold decrease in the abstract review workload equivalent to saving 42501 minutes (approximately 30 full days) of human effort with a 93% coverage on the final included papers. The method primarily missed studies due to a lack of abstracts of the included studies.(18) Aali *et al.* reported that the use of RobotReviewer was a strength of the review to save time and other resources while maintaining study quality.(19) The extracted information not displayed in **Table 1** is available in Appendix 3.1.

Discussion

AI methods in healthcare reviews are progressively being incorporated into practice. Given the repetitive nature of screening that is often associated with a large volume of literature and certain mechanistic aspects in data extraction and possibly the risk of bias assessment, AI pattern recognition algorithms are developed to expedite the process. Generally, the researchers provide labeled training data and then apply it to search results (6) depending on the tool. For instance, Rayyan is a free web and mobile app that extracts all the words and word pairs and previously computed MeSH terms. It then employs support-vector machines to classify the extracted terms.(20) Robot Reviewer is a free system that uses several ML methods, including convolutional neural networks and support-vector machines.(21) EPPI-reviewer is a subscription-based web-based software for all reviews, including systematic reviews.(22) SWIFT-review and SWIFT-Active Screener are a freely available interactive workbench that provides numerous tools to assist with problem formulation and literature prioritization.(23) Abstrackr is a freely available application that allows the reviewer to tag the records depending on the ML's relevance. WordStat (Version 7.1.21) and QDA Miner (Version 5.0.21) are text mining software. The K-means algorithm and the natural language processing were developed and implemented by the reviewers themselves. More information on the general characteristics of each of the tools is available in Appendix 4.1.

Nevertheless, review teams with little to no AI or machine learning expertise can freely use available tools such as Abtrackr, Rayyan, and RobotReviewer to reduce workloads. The effectiveness of some of these tools has been evaluated in some studies. Although several benefits have been reported, there are some limitations reported as well. For instance, Gates *et al.* found that the automated text mining program Abstrackr allowed for large workload savings but possibly missed relevant articles.(24) Other studies such as Rathbone *et al.* and Giummarra *et al*. found Abstrackr to reliably reduce workloads with very little risk of omitting records.(5,6) The 12 identified studies have implemented AI methods in the review and evidence synthesis process to varying degrees of success. Certain stages in the review process are easier to expedite with AI methods. Lam *et al.* and Goldkhule *et al.* both attempted data extraction with SWIFT-Review and RobotReviewer, respectively.(10,17) Both methods ran into issues with SWIFT-Review being unable to extract certain items and the results from RobotReviewer's data extraction not being usable. These issues required human interventions to complete the data extraction successfully. Given these problems, the benefits AI-assisted data extraction are not definitive.

By comparison, screening was able to be completed by some AI methods in various ways. Both reviews with AI experience implemented their algorithms and reviews that used freely available tools such as Rayyan successfully screened without unplanned human interventions. The reported concerns associated AI screening demonstrate a relatively small risk of impacting the review quality. While Giummarra *et al.* reported a risk of missing a small number of relevant studies in the AI-assisted screening, they also suggest the methods were not detrimental to the review's integrity.(9) Gaskins *et al.* reported that their implementation Rayyan “enhanced the screening process,” describing it as “user-friendly,” “accurate,” and “efficient.”(12) The authors also advocated for the future use of automated screening in systematic reviews. Deng *et al.* reported that the AI method had missed 10% of included studies, which substantiates the risk.(18) However, the risk of missing included studies can be mitigated with separate manual screening or method validation. As is suggested in systematic reviews,(25) the title and abstract screening, full-text screening, data extraction, and risk of bias assessment are completed in duplicate to reduce bias. Similarly, the bias of an AI screening can be reduced with a duplicate manual screening as was done by Xiong *et al.* and Deng *et al.*(16,18) The benefits of AI screening are substantial, with 10 of the 11 (91%) screening methods reported advantages to use these methodologies. These benefits usually involve time savings and workload reductions. Lam *et al.* reported having spent 70 person-hours conducting the screening and data extraction for the rEM compared to the 480–960 person-hours of a similar study though compounding factors such as the reviewers' experience makes it difficult to quantify the time saved due to SWIFT Active Screener and SWIFT-Reviewer.(17) However, Lam *et al.* demonstrate the applicability of these AI methods to the broader review community. Deng *et al.* report concrete benefits with a six-fold decrease in the workload for the screening process amounting to 708 hours of human efforts saved.(18) Although no quantitative analyzes were done on the mean difference in publication time, the time saves reported by Lam *et al.* and Deng *et al.* might resolve the issue reported by Borah *et al.* of increasingly long systematic review publication time.(1) It must also be noted that Deng *et al.* specifically designed their natural language processing to fit the nature of their study, and the time saved does not account for the time spent developing the AI method.(18) The workload reductions of a freely available application are likely to be less dramatic than those reported by Deng *et al.*, but that comes with the benefit of the reviews not needing AI expertise to execute the method. Four of the studies that used Rayyan(11,13–15) substantiate this benefit by reporting a high level of usability.

Study Limitations

The qualities and impacts of the AI methods on their respective studies were gathered from the judgements of the authors, which were prone to inconsistencies. It was assumed that no AI was used in that review if the review used an AI tool without mentioning its capabilities. This assumption may have led to the exclusion of some AI-assisted reviews. Our initial screening strategy was limited to the title and abstract of the relevant references. It is likely that the review also overlooks those papers that report the use of AI methods in the text but not in the abstract. With restrictions on abstract word counts, it is not surprising that the reviewers omit the use of AI methods, which may be part of a broader issue with reporting, but that is beyond the scope of this study. The title and abstract screening imperfections were demonstrated with a post-hoc full-text screening of the additional studies from the Cochrane database of systematic reviews (n=67). Out of these 67 studies, which were excluded in the title and abstract screening, could be highlighted as relevant for our current work, Crossingham et al.(26), only one implemented AI methods (1.5%). This study does not mention the AI nature of its methodology in its abstract. Adding Crossingham et al. to the twelve included studies would require the full-text screening of all the studies excluded by the title and abstract screening. More information on Crossingham et al. is available in Appendix 5.1. However, to conduct a full-text screening of the 3514 studies excluded from the preliminary search and the 124 studies excluded from the additional search would be unreasonable. Such a task could be completed in a timely manner with AI methods, which further supports the development and publicizing of AI methods. Though quality assessment was conducted on the included systematic reviews, the quality of the reviews was not accounted for in the study findings.

Conclusion

AI platforms have begun to take hold with varying success in evidence synthesis. The benefits of data extraction conducted with AI methods remain unclear. The current AI platforms are still undergoing refinement, as no single platform appeared to be sufficiently accurate and reliable to date. Existing methods still need humans in the loop and human judgment whenever AI platforms are used in evidence synthesis. Evaluation is needed on the relationship between AI methods and publication time and study quality to delineate AI platforms' efficiency in evidence synthesis.

Highlights

- AI methods show promise in reducing human labor for reviews. A few health science reviews have begun implementing AI methods.
- This paper is the first systematic review of published reviews that implemented AI methods, which allowed for a delineation of the characteristics of those methods.
- The results of this study are relevant to all literature searches and evidence syntheses, as they make known the available methods and their qualities.

Abbreviations

AI: Artificial intelligence; ML: Machine learning; AI/ML: Artificial intelligence/Machine learning; CVS: Cardiovascular disease; T&A: Title and abstract; NLP: Natural language processing; ROB: Risk of bias;

Authors' contributions

N.C., N.L., and S.V. had the original idea for the study and designed the study. S.V., PS. and AB. conducted the literature search and literature screening. AB., S.V., and MA-M. extracted data. N.C. and N.W. supervised the study. AB., S.V., and N.C. wrote the first draft of the paper. CFM-G., N.L., and MA-M. reviewed the first draft and AB. wrote the final draft. All authors interpreted the data, read the manuscript, and approved the final version. N.C. is guarantor.

Acknowledgements

We thank Mr. Josh Higashi for aiding in the post-hoc full-text screening.

Funding

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

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

DECLARATIONS

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Ethical approval

Not applicable.

References

1. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017 Feb 1;7(2):e012545.

2. Yaffe J, Montgomery P, Hopewell S, Shepard LD. Empty reviews: a description and consideration of Cochrane systematic reviews with no included studies. PloS One. 2012;7(5):e36626.

3. Chollet F. Deep learning with Python. Shelter Island, New York: Manning Publications Co; 2018. 361 p.

4. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. Syst Rev. 2018 Mar 12;7(1):45.

5. Giummarra MJ, Lau G, Gabbe BJ. Evaluation of text mining to reduce screening workload for injury-focused systematic reviews. Inj Prev J Int Soc Child Adolesc Inj Prev. 2020 Feb;26(1):55–60.

6. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev. 2015 Jun 15;4:80.

7. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017 Sep 21;358:j4008.

8. Viner R, Russell S, Saulle R, Croker H, Stansfeld C, Packer J, et al. Impacts of school closures on physical and mental health of children and young people: a systematic review. medRxiv. 2021 Feb 12;2021.02.10.21251526.

9. Giummarra MJ, Lau G, Grant G, Gabbe BJ. A systematic review of the association between fault or blame-related attributions and procedures after transport injury and health and work-related outcomes. Accid Anal Prev. 2020 Feb 1;135:105333.

10. Goldkuhle M, Dimaki M, Gartlehner G, Monsef I, Dahm P, Glossmann J, et al. Nivolumab for adults with Hodgkin's lymphoma (a rapid review using the software RobotReviewer). Cochrane Database Syst Rev. 2018 Jul 12;2018(7):CD012556.

11. Pinna F, Manchia M, Paribello P, Carpiniello B. The Impact of Alexithymia on Treatment Response in Psychiatric Disorders: A Systematic Review. Front Psychiatry [Internet]. 2020 [cited 2021 Jul 18];0. Available from: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.00311/full

12. Gaskins NJ, Bray E, Hill JE, Doherty PJ, Harrison A, Connell LA. Factors influencing implementation of aerobic exercise after stroke: a systematic review. Disabil Rehabil. 2019 Dec 25;0(0):1–15.

13. Rogers CR, Matthews P, Xu L, Boucher K, Riley C, Huntington M, et al. Interventions for increasing colorectal cancer screening uptake among African-American men: A systematic review and meta-analysis. PLOS ONE. 2020 Sep 16;15(9):e0238354.

14. Siqueira TV, Nascimento J da SG, Oliveira JLG de, Regino D da SG, Dalri MCB. The use of serious games as an innovative educational strategy for learning cardiopulmonary resuscitation: an integrative review. Rev Gaucha Enferm. 2020;41:e20190293.

15. Nascimento J da SG, Siqueira TV, Oliveira JLG de, Alves MG, Regino D da SG, Dalri MCB. Development of clinical competence in nursing in simulation: the perspective of Bloom's taxonomy. Rev Bras Enferm [Internet]. 2021 Mar 24 [cited 2021 Jul 18];74. Available from: http://www.scielo.br/j/reben/a/zgmY8gmZF3Q98JrxzLdCLrC/?lang=en

16. Xiong Z, Liu T, Tse G, Gong M, Gladding PA, Smaill BH, et al. A Machine Learning Aided Systematic Review and Meta-Analysis of the Relative Risk of Atrial Fibrillation in Patients With Diabetes Mellitus. Front Physiol [Internet]. 2018 [cited 2021 Jul 18];0. Available from: https://www.frontiersin.org/articles/10.3389/fphys.2018.00835/full

17. Lam J, Howard BE, Thayer K, Shah RR. Low-calorie sweeteners and health outcomes: A demonstration of rapid evidence mapping (rEM). Environ Int. 2019 Feb 1;123:451–8.

18. Deng Z, Yin K, Bao Y, Armengol VD, Wang C, Tiwari A, et al. Validation of a Semiautomated Natural Language Processing-Based Procedure for Meta-Analysis of Cancer Susceptibility Gene Penetrance. JCO Clin Cancer Inform. 2019 Aug;3:1–9.

19. Aali G, Drummond A, das Nair R, Shokraneh F. Post-stroke fatigue: a scoping review. F1000Research. 2020 Aug 25;9:242.

20. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016 Dec 5;5(1):210.

21. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying Randomized Controlled Trials: An evaluation and practitioner's guide. Res Synth Methods. 2018;9(4):602–14.

22. Park SE, Thomas J. Evidence synthesis software. BMJ Evid-Based Med. 2018 Aug;23(4):140–1.

23. Howard BE, Phillips J, Tandon A, Maharana A, Elmore R, Mav D, et al. SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. Environ Int. 2020 May 1;138:105623.

24. Gates A, Gates M, DaRosa D, Elliott SA, Pillay J, Rahman S, et al. Decoding semi-automated title-abstract screening: findings from a convenience sample of reviews. Syst Rev. 2020 Nov 27;9(1):272.

25. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M. Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021) [Internet]. Cochrane; 2021. Available from: www.training.cochrane.org/handbook

26. Crossingham I, Turner S, Ramakrishnan S, Fries A, Gowell M, Yasmin F, et al. Combination fixed-dose beta agonist and steroid inhaler as required for adults or children with mild asthma. Cochrane Database Syst Rev [Internet]. 2021 [cited 2021 Nov 24];(5). Available from: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD013518.pub2/full

Table 1 Characteristics of the included studies

Author, Year, Country and Design	Category and Health Science Area	Tool	Stage of Process that AI involved	ML	NLP	Method of use	Description of Methods	Validation	Advantages	AMSTAR
Russell Viner, 2021(8) UK, Italy, USA SR	Experiential/ Qualitative reviews Mental health	EPPI-Reviewer 4	T&A screening	Yes	No	Alone	Reviewers trained the ML algorithm, and then a classifier model was built to rank subsequent studies.	Yes	NA	Moderate
M.J. Giummarra, 2020(9) Australia, UK SR	Prognostic reviews Injury	Abstrackr	T&A screening	Yes	No	Human in the loop	Abstrackr uses an active learning algorithm from judgements made by the reviewer to generate predictions of relevance.	Yes	Reduction in workload	Moderate
		Wordstat and QDA Miner	Full-text screening	Yes	No	Human in the loop	Text mining to identify studies that included fault-related terms in the methods and results.	Yes	Reduction in workload	
Goldkuhle M, 2018(10) Germany, Austria, US Rapid review	Effectiveness review Cancer	RobotReviewer	Data extraction	Yes	Yes	Human in the loop	RobotReviewer applied ML to extract data so it could be compared to manual data extraction.	Yes	NA	Moderate
			ROB assessment	Yes	Yes	Human in the loop	RobotReviewer was made to assess the risk of	No	NA	

							bias with ML, and a review author would have compared these results with the results from the manual assessment.			
Pinna, 2021(11) Italy, Canada SR	Effectiveness Review Mental health	Rayyan	T&A screening	Yes	Yes	Human in the loop	The Rayyan Web app applied a ML algorithm to expedite the screening of titles and abstracts of all identified studies for possible inclusion.	Yes	Accelerate T&A screening and high level of usability	Moderate
Gaskins, 2020(12) UK SR	Experiential /Qualitative reviews CVS	Rayyan	Post-protocol screening	Yes	Yes	Alone	Reviewers used Rayyan autonomously to enhance the screening process after data analysis was conducted. The relevant studies indicated by Rayyan were re-screened.	Yes	Efficient, accurate, and freely available, increasing recall of relevant studies, thereby strengthening a review	Critically Low
Riley, 2020(13) US MA	Effectiveness Review Cancer	Rayyan	T&A screening	Yes	Yes	Human in the loop	The Rayyan Web app applied a ML algorithm to expedite the screening of titles and abstracts of all identified studies for possible inclusion.	No	Accelerate T&A screening and high level of usability	High
Siqueira, 2020(14)	Expert opinion	Rayyan	T&A screening	Yes	Yes	Human in the loop	Reviewers used Rayyan to make the initial triage	Yes	Accelerate T&A screening	Critically Low

Brazil Integrative Review	/Integrative literature review Medical education						process of abstract and title reading faster by identifying studies for possible inclusion applying a ML algorithm.		and high level of usability	
Nascimento, 2021(15) Brazil Integrative Review	Expert opinion /Integrative literature review Medical education	Rayyan	T&A screening	Yes	Yes	Human in the loop	Reviewers used Rayyan to make the initial triage process of abstract and title reading faster by identifying studies for possible inclusion applying a ML algorithm.	Yes	Accelerate T&A screening and high level of usability	Critically Low
Xiong, 2021(16) New Zealand, China, Canada MA	Etiology and /Risk reviews CVS	K-means clustering algorithm	T&A screening	Yes	No	Alone	The reviewers used the K-means clustering algorithm to provide an alternative to manual T&A screening by clustering articles. The algorithm was trained on a set of relevant studies.	Yes	Accelerate T&A screening, Reduction in workload; accurate	Low
Lam, 2019(17) US rEM(rapid Evidence Mapping)	Etiology and /Risk reviews Nutrition	SWIFT-Active Screener	T&A screening	Yes	No	Human in the loop	SWIFT-Active Screener used active learning to prioritize relevant references and estimated the number of remaining relevant articles.	No	Substantial time savings relative to a similar study	NA

		SWIFT-Review	Data extraction	Yes	No	Human in the loop	SWIFT-Review assisted with searching and pattern visualization using machine learning methods.	No	Substantial time savings relative to a similar study	
Deng, 2019(18) US MA	Etiology and /Risk reviews Cancer	Semi-automated natural language processing	Abstract classification and filtering and text mining	Yes	Yes	Human in the loop	Reviewers implemented a semi-automated natural language processing (NLP) to classify abstracts. Text mining was then employed.	Yes	A six-fold decrease in the abstract review workload is equivalent to saving 42,501 minutes (approximately 30 full days) of human effort. 93% of coverage on the review process	NA
Aali, 2020(19) UK Scoping Review	Effectiveness Review CVS	RobotReviewer	ROB assessment	Yes	Yes	Human in the loop	RobotReviewer applied its ML methods to evaluate the risk of bias for certain bias categories automatically	Yes	Strength in terms of saving time and other resources while maintaining study quality	NA

Figure 1. PRISMA flowchart

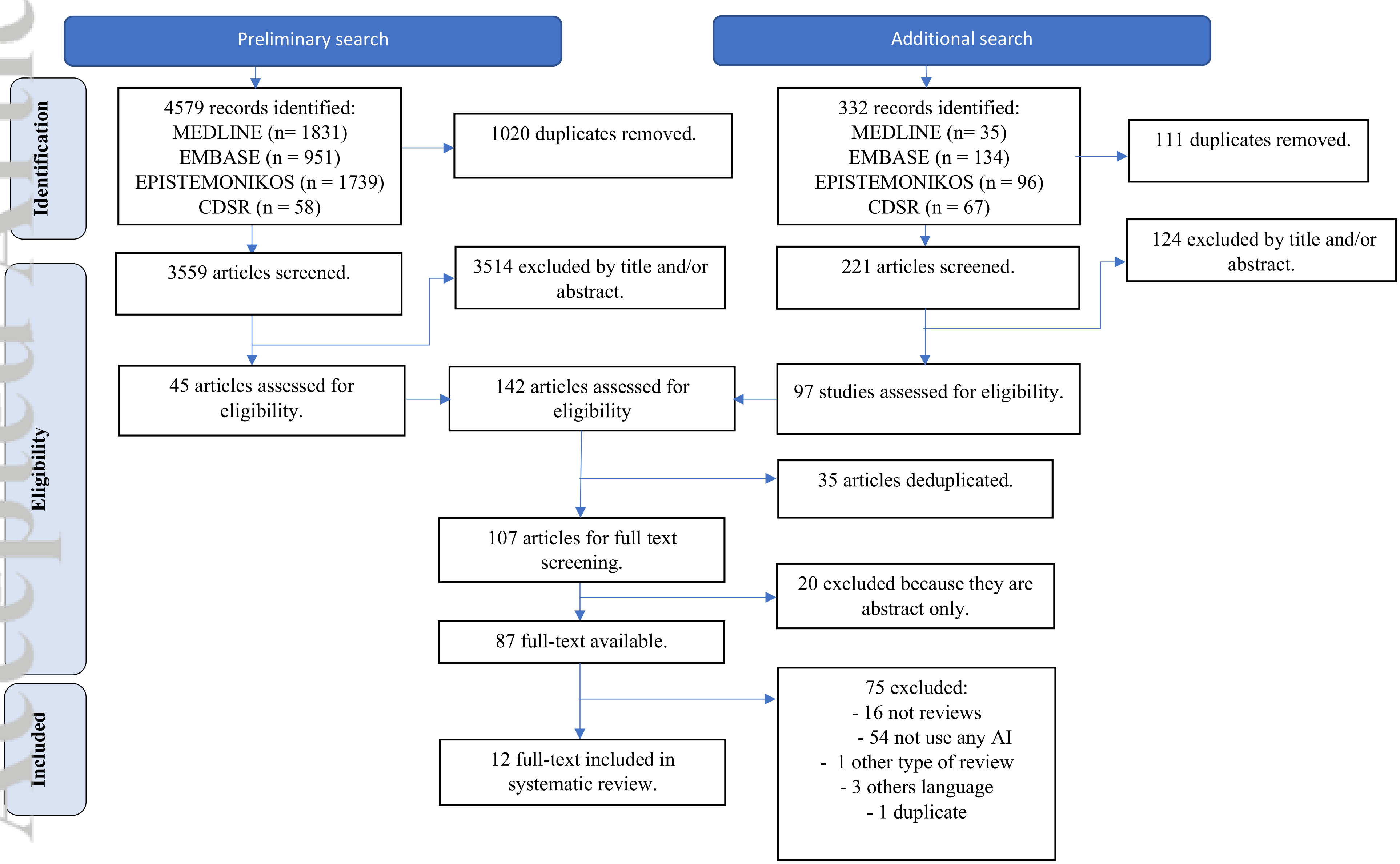


Figure 2: AI methods that employed each tool

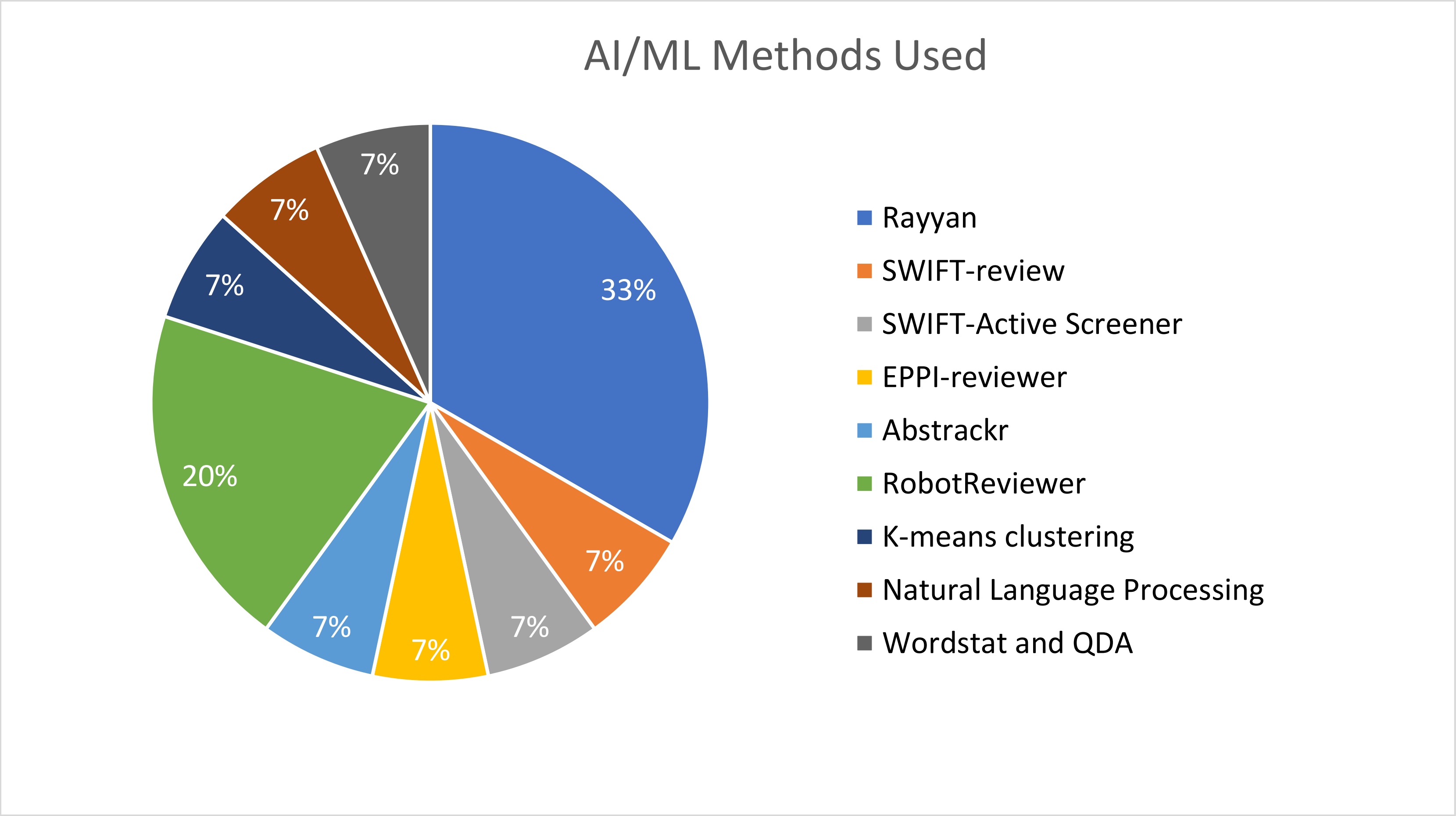
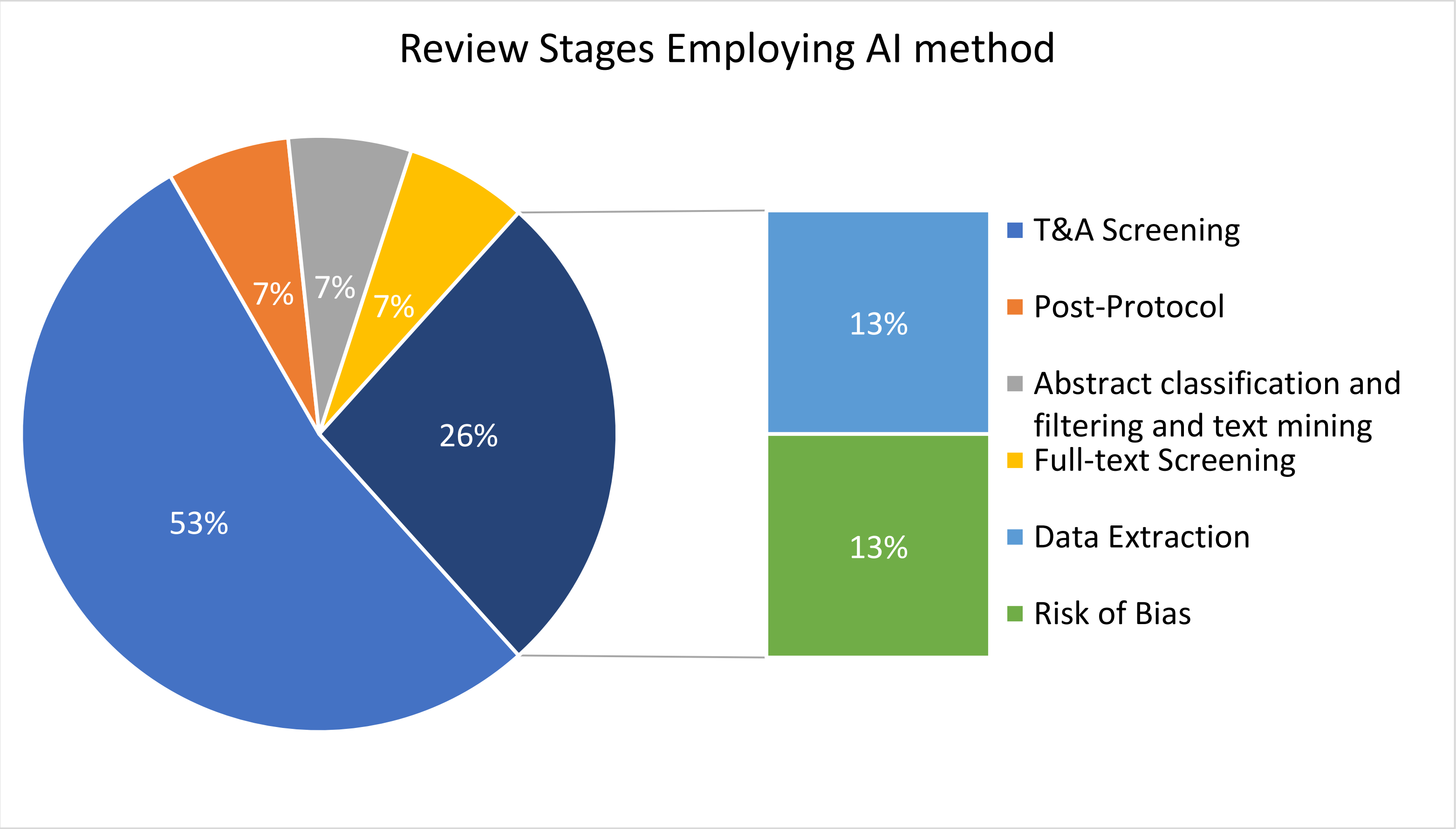


Figure 3: Distribution of AI/ML methods employed in each stage.



The bar displays the methods not involved in screening.