Systematic Review with Artificial Intelligence (SRAI): A Few-Shot Learning Framework for Efficient Article Screening in Systematic Reviews

Phongphat Wiwatthanasetthakarn[1], Wanchana Ponthongmak[*1], Panu Looareesuwan[1], Amarit Tansawet[2], Pawin Numthavaj[1], and Ammarin Thakkinstian[1]

[1]Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand
[2]Department of Surgery, Faculty of Medicine Vajira Hospital, Navamindradhiraj University, Bangkok, Thailand
[*]Corresponding author, E-mail: wanchana.pon@mahidol.edu

Abstract

A systematic review (SR) is a comprehensive method to summarize and analyze existing research on a specific topic or question. The selection of studies is labor-intensive and time-consuming, and it is typically conducted by at least two reviewers. Applying artificial intelligence (AI) using a few-shot learning (FSL) framework may be helpful in reducing the workload and improving efficiency and accuracy. This study aimed to develop a model framework for literature screening in SR using the FSL approach. The data sources for conducting this study were drawn from nine SR studies conducted between 2016 and 2022, in which the number of identified studies ranged from 426 to 7341. These SR studies could be classified into four types, including therapeutic, prognostic/risk, genetic association, and economic evaluation studies. This study had two phases. 1) finding the optimal number of positive studies; and 2) finding the cosine similarity threshold for selecting positive studies.

The findings revealed that the overall median (IQR) of optimal *n* positive studies was 9 (8-12), with the corresponding median (IQR) reduced workload of 95.78% (93.49% - 97.98%). However, we initially used 4-6 positive studies to conduct the second study phase. In the second study phase, the mean (range) of the optimal cosine similarity threshold of the therapeutic, prognostic/risk, genetic association, and economic evaluation SRs were 0.5345 (0.4387 - 0.6168), 0.5048 (0.4317 - 0.5778), 0.5903 (0.5458 - 0.6348) and 0.6358 (0.6358-0.6358), respectively. The corresponding reduced workloads were 78.75% (64.81%-96.94%), 67.64% (51.11%- 84.16%), 83.39% (69.11% - 97.67%), and 95.34%.

The proposed framework, SRAI, can potentially reduce the workload for article screening in the SR process by requiring the reviewer to select a few positive studies for initial model training. However, prospective evaluation is necessary to ascertain the performance of the model.

Keywords**:** *Systematic Review, Artificial Intelligence, Machine Learning, Few-Shot Learning, Natural Language Processing*

1. Introduction

A systematic review (SR) is a rigorous and comprehensive method to synthesize existing research findings on a specific topic or question (Ahn, & Kang, 2018; Uman, 2011). SR is commonly used in healthcare and other fields to inform decision-making, policy development, and further research. In addition, the SR also plays a vital role in summarizing existing evidence in order to identify a gap in knowledge for further research (Ahn, & Kang, 2018; Uman, 2011). It involves selecting studies based on eligibility criteria, which is time-consuming and resource-intensive and requires at least two reviewers who independently select studies as the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines (Hoffmann et al., 2021). As the number of PubMed publications in Medline has increased over time (Hoffmann et al., 2021), an artificial intelligence (AI) tool that can reduce workload and time in selecting identified studies is required. Such AI relies on natural language processing (NLP) and machine learning (ML) techniques to perform screening tasks as a secondary reviewer for SR. Several tools for the selection

process of SR have been developed, which could perform the title and abstract screening, including EPPI-Reviewer (EPPI Centre, 2024; Tsou et al., 2020), Abstrackr (Gates et al., 2019, 2020; Rathbone et al., 2015; Wallace et al., 2012), Covidence (Veritas Health Innovation, 2024), DistillerSR (Gates et al., 2019; Hamel et al., 2020; Patel et al., 2021; Wallace et al., 2010), RobotAnalyst (Gates et al., 2019; Przybyła et al., 2018), and Rayyan (Olofsson et al., 2017; Ouzzani et al., 2016; Yu et al., 2022). These tools apply NLP and supervised learning (SL) with an active learning framework (Gates et al., 2019; Wallace et al., 2010) as a kernel mechanism for screening identified studies. Although active learning enables semi-automatic screening tasks, the major drawback is that it requires some or even many initial positive studies to train the model to achieve targeted performance. Moreover, active learning is not generally applicable because new projects sometimes require different iterative annotation processes. Such a limitation has resulted in less applicability for conducting multiple SR projects. In addition, some available tools are commercially constrained, costing from $165 to $635 per year (EPPI Centre, 2024; Veritas Health Innovation, 2024), thus limiting accessibility. These tools have used various feature extractions, as follows: EPPI-Reviewer used trigrams (Forsgren et al., 2023) and term frequency - inverse document frequency (TF-IDF) (EPPI Centre, 2024; Tsou et al., 2020), Abstrackr used n-grams (Przybyła et al., 2018; Wallace et al., 2012), Rayyan used unigrams, bigrams (Ouzzani et al., 2016), Robot Analyst used bag-of-words, TF-IDF, and latent dirichlet allocation (LDA) model (Przybyła et al., 2018). However, TF-IDF, unigrams, bigrams, n-grams, and LDA methodologies exhibit limitations in capturing semantic nuances within the text, particularly in comparing textual similarity. The support vector machines (SVM) classifier is commonly used in most tools, including EPPI-Reviewer (Tsou et al., 2020), Abstrackr (Wallace et al., 2012), Covidence (Veritas Health Innovation, 2024), Rayyan (Ouzzani et al., 2016), RobotAnalyst (Przybyła et al., 2018), and DistillerSR (DistillerSR, 2024). Also, the naive Bayes classifier was applied in DistillerSR (Burgard & Bittermann, 2023; DistillerSR, 2024). Nevertheless, the model's performance mainly depends on feature extraction/vector representation.

The existing knowledge gap in AI for SR revolves around the necessity of annotated data for active learning. To bridge this disparity, we propose an innovative approach, SRAI: an AI using a few-short learning (FSL) framework (Bražinskas et al., 2020; Wang et al., 2020), to circumvent the need for extensive and repetitive annotation. This method empowers the model to learn effectively from a concise set of training samples. Moreover, this method harnesses the cutting-edge capabilities of the sentence-bidirectional encoder representations from transformers (sentence-BERT) pre-trained model (Devlin et al., 2019; Reimers, & Gurevych, 2019). By leveraging SBERT, we enhance the semantic similarity between user queries and pertinent studies, thereby actively optimizing screening outcomes. The SRAI offers potential utility for model training. Theoretically, this method is expected to make ML processes more quickly identify studies most relevant to SR than previous AI approaches.

To our knowledge, the FSL has not yet been used as an automated tool for conducting SRs. This study aimed to create a novel automated framework utilizing FSL to enhance the SR screening process, achieving performance as high as traditional approaches with a faster method.

2. Objectives

1) To develop an AI tool by applying an FSL with SBERT embedding for the selection of studies in various types of SRs.

2) To find an optimal number of positive studies that minimize workload while maintaining recall at 100% for training the model.

3) To estimate and calibrate the cosine similarity score threshold between the support (training) and query (test) sets.

3. Materials and Methods

3.1 Data sources

This study used data from nine SRs, which were conducted by researchers of the Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, between 2016 and 2022. The nine SRs are as follows:

1) Mesh position for hernia prophylaxis after midline laparotomy: A systematic review and network meta-analysis of randomized clinical trials (Tansawet et al., 2020), hereafter called SR1.

2) The efficacy of antibiotic treatment versus surgical treatment of uncomplicated acute appendicitis: systematic review and network meta-analysis of a randomized controlled trial (Poprom et al., 2019), hereafter called SR2.

3) Efficacy and safety of urate-lowering agents in asymptomatic hyperuricemia: systematic review and network meta-analysis of randomized controlled trials (Sapankaew et al., 2022), hereafter called SR3.

4) Efficacy and safety of antiviral agents in the prophylaxis and pre-emptive strategies for cytomegalovirus infection in kidney transplantation: a systematic review and network meta-analysis (Ruenroengbun et al., 2021), hereafter called SR4.

5) Association between vitamin D and uric acid in adults: a systematic review and meta-analysis (Isnuwardana et al., 2020), hereafter called SR5.

6) Prognostic model for complications in type 2 diabetes: systematic review and meta-analysis (Saputro et al., 2021), hereafter called SR6.

7) The association between genetic polymorphisms in ABCG2 and SLC2A9 and urate: an updated systematic review and meta-analysis (Lukkunaprasit et al., 2020), hereafter called SR7.

8) AHSG gene polymorphisms, Serum fetuin-A levels and association with type 2 diabetes and cardiovascular diseases: a systematic review and meta-analysis (Bassey et al., 2022), hereafter called SR8.

9) Evaluation of the cost utility of phosphate binders as a treatment option for hyperphosphatemia in chronic kidney disease patients: A systematic review and meta-analysis of the economic evaluation studies (Chaiyakittisopon et al., 2021), hereafter called SR9.

These SRs were classified into four types: therapeutic, prognostic/risk, genetic association, and economic evaluation studies, as seen in Table 1.

The title and abstract of individual SRs were extracted and concatenated to use as input data for the model development process. The total number of texts and other characteristics of 9 SRs are summarized in Table 1.

Table 1 Characteristics of systematic review projects

No.	Project name	Type of study	Number of identified studies	Number of tokens	Number of unique tokens	Median (IQR) (range)	Studies contained > 384 tokens[1] (%)
1	SR1	Therapeutic[2]	3,966	1,120,514	37,776	285 (206.25 – 343.00) (32 – 1,629)	507 (12.78)
2	SR2	Therapeutic	1,702	411,157	18,174	244 (164.00 - 307.00) (37 - 877)	126 (7.40)
3	SR3	Therapeutic	7,341	2,205,482	71,645	297 (214.00 - 364.00) (20 – 1,385)	1,486 (20.24)
4	SR4	Therapeutic	3,144	874,261	30,226	272 (195.00 - 337.00) (33 – 3,509)	408 (12.98)
5	SR5	Prognostic[3]	699	191,189	15,298	274 (189.00 - 340.50) (20 - 821)	97 (13.88)
6	SR6	Prognostic	426	125,362	10,843	292 (232.50 - 339.00) (17 – 1,112)	57 (13.38)
7	SR7	Genetic[4]	1,708	444,383	26,883	259 (204.00 - 312.00) (29 - 635)	81 (4.74)
8	SR8	Genetic	1,053	318,339	18,787	305 (255.00 - 346.00) (70 - 647)	126 (11.97)

No.	Project name	Type of study	Number of identified studies	Number of tokens	Number of unique tokens	Median (IQR) (range)	Studies contained > 384 tokens[1] (%)
9	SR9	Economic[5]	1,653	463,892	30,601	244 (176 - 322) (18 - 2542)	221 (13.37)

[1] 384 is the maximum sequence length of text input for the model, [2] Therapeutic studies, [3] Prognostic/ risk studies, [4] Genetic association studies, [5] Economic evaluation studies

3.2 Finding optimal n positive

The optimal *n* positive was an experiment to find the number of positive studies that yielded the best performance with a restriction of 100% recall. The experiment consisted of three processes, as follows:

3.2.1 Data preparation

If crucial data, i.e., titles and abstracts of positive studies, were missing, they were searched and retrieved from PubMed, Scopus, or other databases. In contrast, those negative studies were excluded from the experiment. The duplicated studies with the same year, author, and title were excluded.

3.2.2 Scenario creation

We created 10 scenarios or fewer for each SR project, depending on the number of positive studies, as shown in the following steps. First, the corpus was split into three sets (i.e., training, validation, and test pools) with a splitting ratio of 50:25:25 of the positive studies, denoted as S_+, S_+^{trn}, S_+^{val}, S_+^{tst} for total, training, validation, and test pools, respectively, as an example of SR1 in Figure 1. Meanwhile, the training and validation pools were concatenated with negative studies ($S\_$), selected randomly without replacement at a ratio of 10 times, which were denoted as S_-^{trn} and S_-^{val}. Additionally, the remaining negative studies belonged to the test set (S_-^{tst}). Second, the next scenarios were created by reducing the number of positive studies by 10% of the previous training (S_+^{trn}) and validation (S_+^{val}) sets. However, the test set generated by the combination of S_+^{tst} and S_-^{tst} of the first scenario remained the same for all scenarios, comparing model performance fairly. A minimal number of four positive studies in the scenario was required to get the positive sample pair in the training and validation sets. As a result, most SR projects with less than 24 positive studies could only generate less than 10 scenarios.

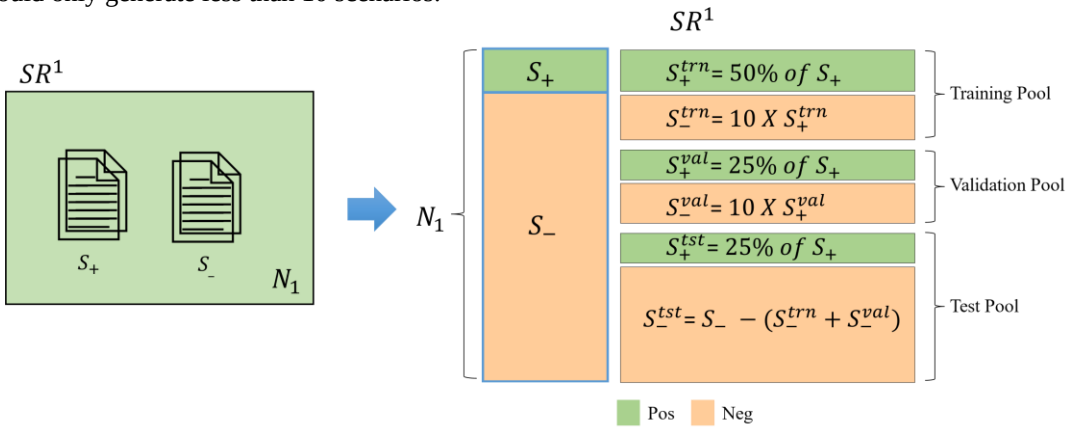

Figure 1 Corpus splitting into training, validation, and test pools for SR1

1) Pairing data for the development dataset

Two datasets were generated during the model development phase, including training and validation datasets. Each dataset contained several pairs of studies (referred to as paired samples), which were retrieved from each data set (i.e., training and validation sets) in the scenario creation process. Each sample was generated by pairing two studies, which were annotated with a new label paired as "1" for a pair of positive-

positive studies and "0" for a pair of positive-negative studies, whereas a pair of negative-negative studies was not used.

2) Pairing data for a test dataset

The test dataset was created by pairing between the query set and support set, which were the studies in the test pool and the positive studies from the training pool, respectively. The cosine similarity score for each paired sample was calculated for model inference. After that, the cosine similarity of the paired query set was averaged. Given Q_1 denoted the first study of the query set, was paired with all positive studies of the support set, which were denoted by S_1, S_2, S_3, …, S_m; where *m* was the number of positive studies of the support set. The cosine similarity score of each pair (paired sample) was calculated as Sim_1^1, Sim_2^1, Sim_3^1, …, Sim_m^1. Finally, the average cosine similarity score of $Q_1$was calculated to represent the cosine similarity score of Q_1. The Q_1 study was classified as positive (selected) if the score was more than or equal to the threshold; otherwise, the study was classified as negative (not selected). The process of classifying Q_1 was applied to the rest of the studies in the query set Q_2, Q_3, Q_4, …, Q_n; where *n* was the total number of studies in the query set. Using all models trained for each scenario, the optimal *n* positive was determined by identifying the minimum number of positive studies while maximizing workload reduction. This process is indicated by the Kneedle algorithm (Satopaa et al., 2011).

3.3 Finding the cosine similarity threshold

After selecting the optimal scenario from Section 3.2, the model from such a scenario was applied to find the cosine similarity threshold of each SR project using the new test dataset (generated by the new test pool) while maintaining the recall at 100%. The sample preparation method for this phase was similar to Section 3.2.2. The development datasets (training and validation datasets) were drawn from the optimal scenario for each SR project. In contrast, the test dataset was regenerated from the new test pool, which was altered to utilize all available studies (all studies that were not included in the training and validation pools of the optimal scenario).

3.4 Model development framework

This study applied the model training framework using FSL, which helped the model learn the underlying representation with only a few samples and their similarity score rather than applying a fully supervised fine-tuning approach. An FSL refers to the capability of a model to learn new tasks or concepts from a small number of examples. Unlike conventional learning paradigms, which necessitate a substantial amount of labeled data for effective training, few-shot learning models excel in scenarios where only a handful of labeled instances are available per class.

The SBERT was used, with pre-trained weight and architecture acquired from "all-mpnet-base-v2", the best pre-trained model among available models in Hugging Face (Hugging Face, 2024). SBERT can capture the semantics at the sentence level by comparing two sentence embeddings with cosine-similarity. The title and abstract were extracted and concatenated for each SR project for fine-tuning. By fine-tuning the SBERT model with the paired data from the training and validation sets, the embedding layer should have a good document representation that captures the candidate studies with high similarity to the selected studies. During fine-tuning, the two studies, A and B, were passed through the BERT and pooling layer to reduce the dimension of the vectors, and the output is two vectors (i.e., u and v). The cosine similarity was applied to measure the similarity between the vectors, as illustrated in Figure 2. The SBERT hyperparameters were configured as follows: batch size = 8, epochs = 1, optimizer params (lr) = $2e^{-05}$, max seq length = 384, word embedding dimension = 768. After fine-tuning the model, such a model was used for feature extraction for the query set (test set) and support set (train set).

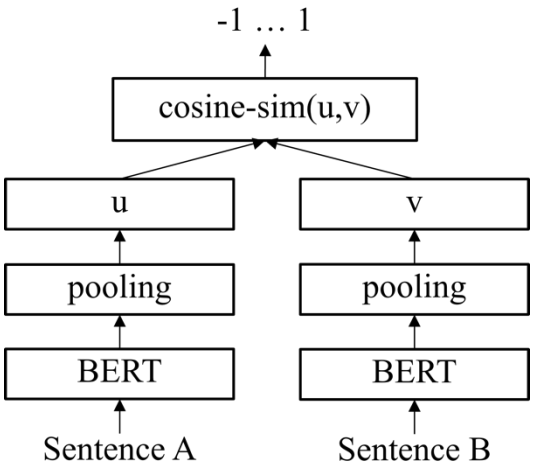

Figure 2 Model architecture of SBERT (Reimers & Gurevych, 2019)

3.5 Model evaluation

We evaluated the model by 1) examining the number of studies that needed to be reviewed during the screening process and 2) the percentage of reduced workload. These metrics were evaluated based on a false negative of zero. The evaluation was done on the test set. The following formula calculated both of the metrics:

$$Workload\ (\%) = 100 * \frac{(TP + FP)}{(TP\ +\ FP\ +\ TN\ +\ FN)} \quad (1)$$

$$Reduced\ workload\ (\%) = 100 - Workload\ (\%) \quad (2)$$

$$Number\ needed\ to\ read = Total\ studies * Workload\ (\%) \quad (3)$$

Where TP is true positive, FP is false positive, TN is true negative, FN is false negative, $(TP + FP)$ is the number of studies that were predicted as positive, and $(TP\ +\ FP\ +\ TN\ +\ FN)$ is the total individual studies.

A higher percentage of reduced workload is better in fewer studies that are required to be reviewed. In contrast, the lower percentage of reduced workload results in many studies to be reviewed. However, all experiments must be based on a recall that remains at 100%.

4. Results and Discussion

4.1 Results

4.1.1 Finding optimal n positive

The overall median (IQR) of optimal *n* positive studies for our experiment was 9 (8-12) with a range of 4-12, while the corresponding median (IQR) reduced workload was 95.78% (93.49%-97.98%) with a range of 80.87%-99.37%. However, this study could further classify SR projects into 4 sub-categories. First, in the therapeutic study group (SR1-SR4), the performance of the model yielded a median (IQR) of optimal *n* positive studies of 10 (7-12) with a range of 4-12, and the corresponding reduced workload was 96.10% (95.21%-96.81%) with a range of 93.49%-97.98%. Second, the prognostic/risk study group (SR5-SR6) provided the *n* positive studies of 9 and 12, with reduced workloads of 93.56% and 88.12%, respectively. Third, the genetic association study (SR7-SR8) provided the optimal *n* positive studies of 8 and 4 with the corresponding reduced workloads of 80.87% and 99.37%, respectively. Fourth, only one economic evaluation

study showed an optimal *n* positive studies of 9 and a reduced workload of 98.21%. These results are summarized in Table 2.

Table 2 Describe percent reduced workload and optimal *n* positive by systematic review projects

Type of study	Project	N identified studies / total *n* positive	% Reduced workload (*n* positive)		
			Minimum	Maximum	Optimal
Therapeutic study	SR1	3,966/20	84.61 (4)	97.82 (14)	96.42 (8)
	SR2	1,702/9	95.48 (5)	97.07 (6)	95.78 (4)
	SR3	7,341/19	90.72 (5)	98.55 (14)	97.98 (12)
	SR4	3,144/21	69.69 (9)	93.92 (14)	93.49 (12)
Prognostic/ risk study	SR5	699/32	87.13 (4)	93.79 (15)	93.56 (9)
	SR6	426/21	59.00 (5)	89.66 (15)	88.12 (12)
Genetic association study	SR7	1,708/48	56.55 (26)	80.87 (8)	80.87 (8)
	SR8	1,053/11	98.22 (8)	99.37 (4)	99.37 (4)
Economic evaluation study	SR9	1,653/24	96.63 (12)	98.21 (9)	98.21 (9)

4.1.2 Finding the cosine similarity threshold

Figures 3 to 5 illustrate the optimal threshold of SR1-SR9, given that a recall of 100% is maintained in each SR. The relationships between cosine similarity versus recall and reduced workload were plotted on the X and Y axes, and the optimal threshold and the corresponding reduced workload were assessed. We can see that the optimal cosine similarity threshold varies according to the types of SR, as different types of SRs might have different thresholds. Only the economic evaluation study did not have minimum and maximum thresholds because there was only one study in the project. The mean (min-max) optimal cosine similarity of the therapeutic study, prognostic/risk study, genetic association study, and economic evaluation study had thresholds of 0.5345 (0.4387 - 0.6168), 0.5048 (0.4317 - 0.5778), 0.5903 (0.5458 - 0.6348) and 0.6358, respectively. The mean (min-max) of reduced workload for these corresponding SRs were 78.75% (64.81% - 96.94%), 67.64% (51.11% - 84.16%), 83.39% (69.11% - 97.67%), and 95.34%. The summary of findings by SR types is illustrated in Table 3.

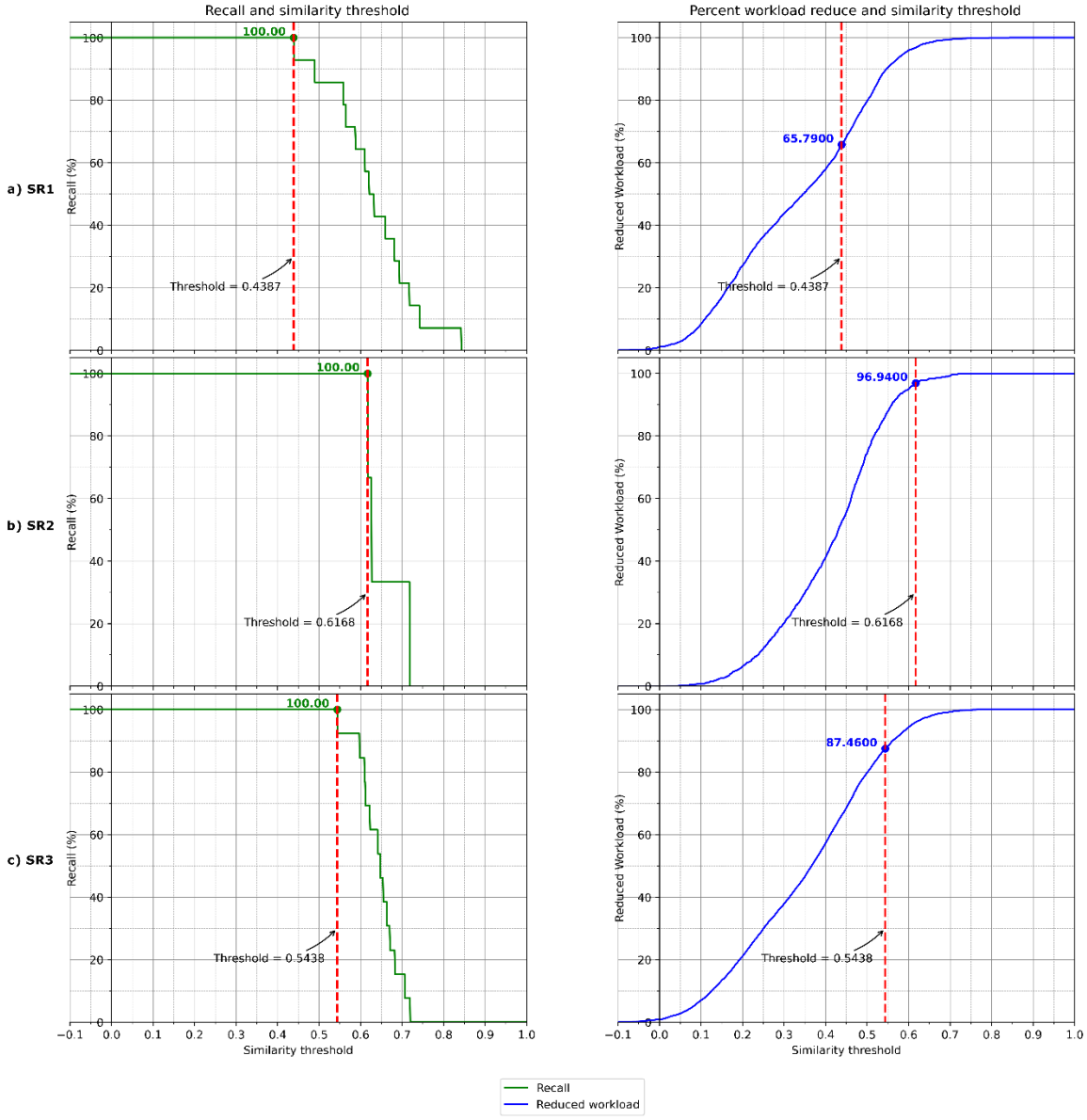

Figure 3 Optimal cosine similarity threshold of SR1, SR2, and SR3

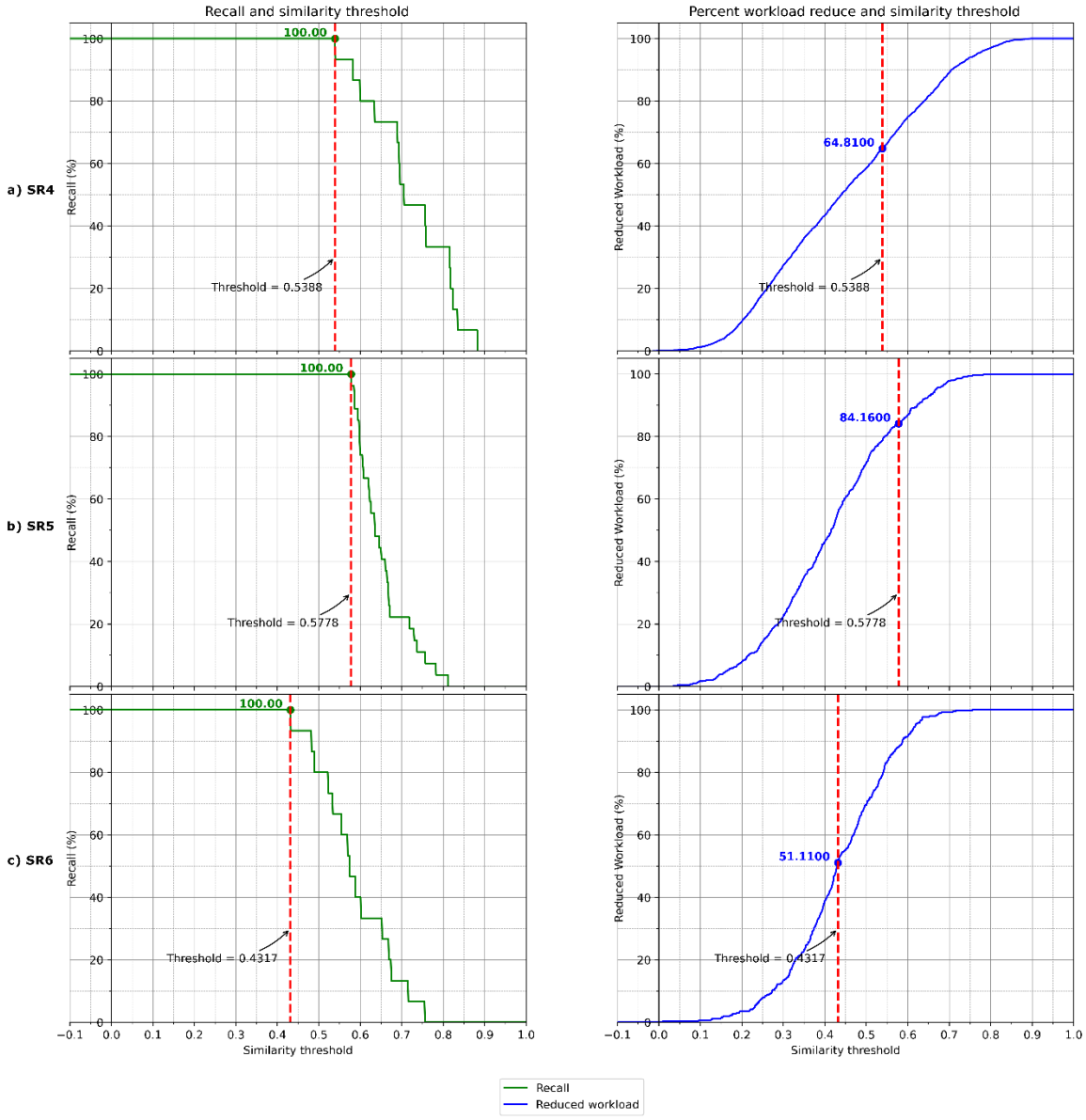

Figure 4 Optimal cosine similarity threshold of SR4, SR5, and SR6

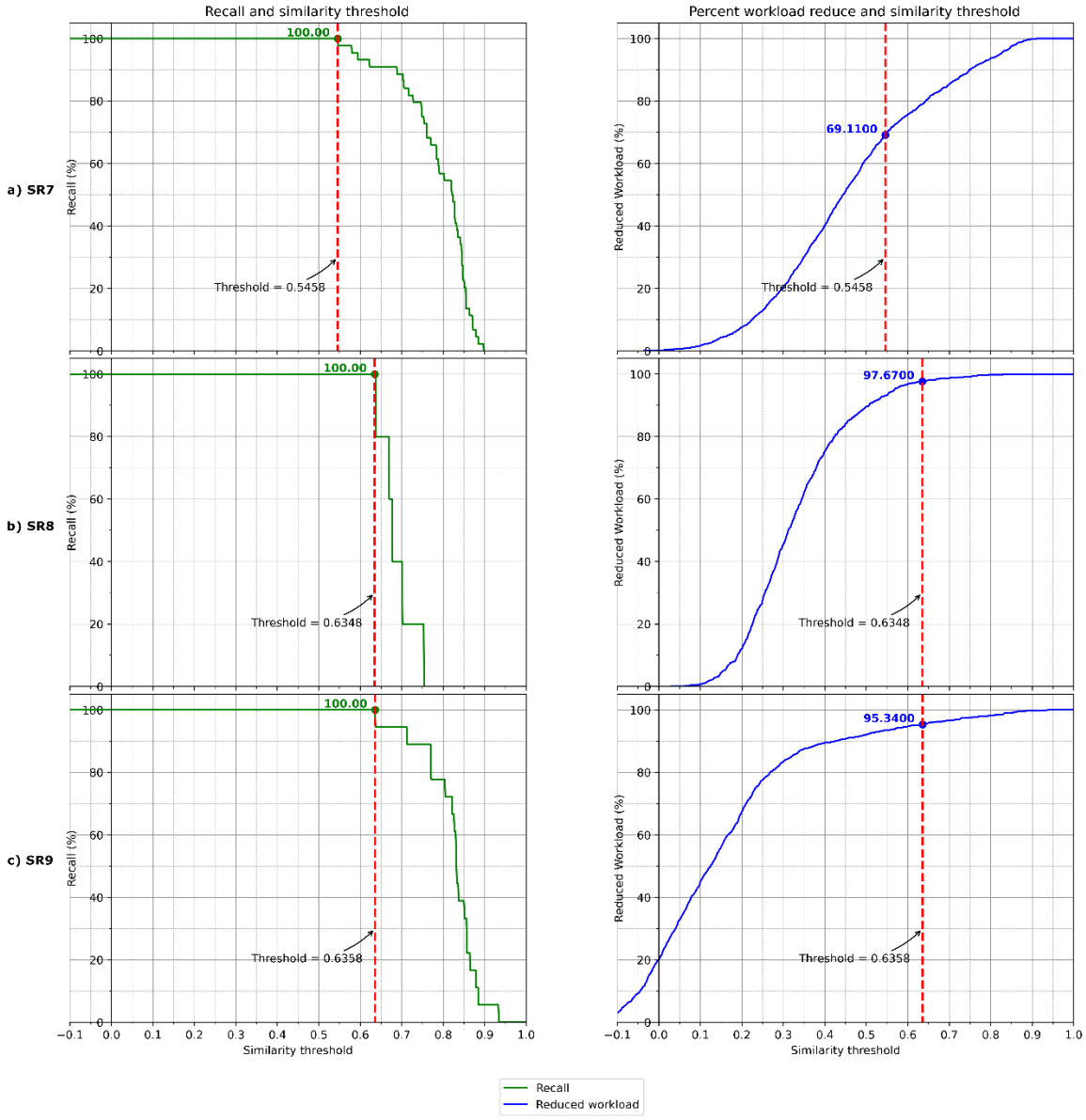

Figure 5 Optimal cosine similarity thresholds of SR7, SR8, and SR9

Table 3 Reduced workload and similarity score thresholds categorized by type of systematic review

Type of study	Threshold	
	Similarity score Mean (min-max)	% Reduced workload Mean (min-max)
Therapeutic study	0.5345 (0.4387 - 0.6168)	78.75 (64.81 - 96.94)
Prognostic/risk study	0.5048 (0.4317 - 0.5778)	67.64 (51.11 - 84.16)
Genetic association study	0.5903 (0.5458 - 0.6348)	83.39 (69.11 - 97.67)
Economic evaluation study	0.6358 -	95.34 -

4.2 Discussion

The study could identify the optimal *n* positive in the training and validation pool used for training the model. The findings showed that the overall median (IQR) of optimal *n* positive studies was 9 (8-12) with a range of 4-12, while the corresponding median (IQR) reduced workload at 100% recall was 95.78% (93.49%-97.98%) with a range of 80.87%-99.37%. As a result, the numbers needed to be read ranged from 7 to 251, given the total number of identified studies of 426 to 7,341.

The minimum number of *n* positive studies for model development was four, of which two were for the training pool and another two for the validation pool, as they could be paired into a single positive paired sample (positive-positive pairing) for each training and validation dataset. However, the maximum number of *n* positive studies during the model development process could be more than four, but manually reviewing to get positive studies was time-consuming and must be considered and justified. In practice, reviewers typically perform an initial search to identify potentially positive studies before proceeding with a comprehensive review. If several positive studies (approximately 4 - 6) are identified during this preliminary phase, the SRAI could significantly reduce the effort for subsequent reviewing. SRAI is less time-consuming than the AI tools reviewed in this study, which were commonly used in SR screening.

As reviewed in this study, the current tools applied NLP and SL with active learning to train the model (Gates et al., 2019; Wallace et al., 2010). Active learning is a framework that allows the model to be initially trained with a small sample size and then re-trained iteratively with increasing numbers of samples. It does not guarantee when the model will converge to the expected performance.

These tools used frequency-based feature extractions (EPPI Centre, 2024; Forsgren et al., 2023; Ouzzani et al., 2016; Przybyła et al., 2018, 2018; Tsou et al., 2020; Wallace et al., 2012). This method cannot capture the text's semantics; it only counts the frequency of tokens without considering context. In contrast, the proposed method in this study utilized SBERT for text vectorization, which improved finding relevant studies by capturing the sentences' dependency and meaning better than the word frequency-based techniques.

The SVM classifier is commonly used in the tools (DistillerSR, 2024; Ouzzani et al., 2016; Przybyła et al., 2018; Tsou et al., 2020; Veritas Health Innovation, 2024; Wallace et al., 2012). The classifier is popular because it is easy to implement, allows different kernels to be utilized, and requires less computation power; however, annotating more samples is still challenging, while the model's performance mainly depends on

feature extraction/vector representation. Hence, we applied the FSL approach to mitigate the workload and time associated with searching the seed studies (positive and negative) and annotating for model training.

A study (Tsou et al., 2020) found that the EPPI-Reviewer could reduce the workload given a 100% recall of 8.68%-60.11%. Another study (Gates, Johnson, & Hartling, 2018) found that Abstrackr could reduce the workload by 9.50%-88.40% while achieving 79%-96% of recall. A study (Valizadeh et al., 2022) found that Rayyan could reduce the workload by 20.00% while achieving 87%-98% recall. Furthermore, a study (Hamel et al., 2020) found that DistillerSR could reduce the workload by 30%- 72.50% while achieving a 95% recall. In addition, a study (Reddy et al., 2020) found that RobotAnalyst could reduce the workload and achieve a 100% recall of 30.69%. The proposed method in this study applied to 4 SR types and could reduce the workload by 67.68%-95.34% while achieving 100% recall. When compared at a recall of less than 100%, the proposed method could reduce workload more than Abstrackr in both the minimum and maximum values of reduced workload and minimize workload more than Rayyan and DistillerSR when considering the maximum value.

The strengths of this study included the following: The experiment was performed using data covering 4 SR types, including therapeutic, prognostic/risk, genetic association, and economic evaluation SRs. The proposed approach applied the FSL framework to support the title and abstract screening processes, which could reduce the workload of selecting studies. This tool required a reviewer to select a small number of positive and negative studies for model training. Subsequently, the reviewer can proceed with a review solely based on the model's suggested studies, thereby cutting down on the time required to review all identified studies. In addition, the number of suggested studies can be calibrated through the percent workload per the reviewer's discretion. This could be implemented into the software that allows the reviewer to adjust the calibration to maximize the chance of relevant studies being reviewed.

However, some limitations were unavoidable. A few SR types contained fewer SR projects; thus, their results may not be repeatable. The model was not prospectively evaluated; without comprehensive validation of diverse types of SR studies, the effectiveness of the framework is uncertain. Implementing and fine-tuning the SRAI framework requires specialized expertise in ML, NLP, and SR methodologies. This is a barrier to adoption for research teams lacking such expertise. While the framework is tailored to different types of SRs, generalizing its performance across diverse research domains and topics can be challenging. Certain domains or topics require additional customization or fine-tuning for optimal performance. There is a risk that researchers over-rely on AI tools like SRAI without critically evaluating their outputs. Unquestioningly trusting automated screening processes leads to the oversight of relevant studies or biases in the review process. AI models are susceptible to algorithmic biases present in the training data, which result in biased decision-making and recommendations. Mitigating these biases requires careful preprocessing of data and ongoing monitoring and validation of the model's outputs. Integrating AI tools like SRAI into existing SR workflows and methodologies poses challenges in terms of compatibility, usability, and acceptance by research communities. Effective integration requires careful consideration of users needs and feedback. While AI tools can automate certain tasks in SR, human oversight and collaboration remain essential for ensuring the accuracy, relevance, and validity of the review process. Balancing automation with human judgment is crucial for maintaining quality standards.

This study's model and proposed framework were developed and evaluated using data from existing completed SR projects. Applying the proposed framework to other SR projects requires a prospective evaluation of the new SR projects to assess if the proposed framework is valid and robust. A web-based application system should be constructed and set up to achieve this goal. User-friendly interfaces and security should be considered when implementing the system. In addition, the performance of this tool should also be prospectively compared with previous tools that were accessible. Finally, the proposed model framework requires a computer with a high GPU specification to train the model.

5. Conclusion

This study uses the FSL framework and SBERT to develop the SRAI tool. After completing data preprocessing for each SR, we individually split the data into training, validation, and a test pool for SR. The

scenarios for each SR were created to find the optimal scenario, and the samples were paired for the training, validation, and test dataset to train the model. We evaluated the model's performance by applying cosine similarity and calculating the percentage of reduced workload by fixing the test dataset for all scenarios. The process maintains the recall of 100%. Finally, we could identify the optimal *n* positive studies. We then used the model of the optimal scenario for each SR to find the cosine similarity threshold. The results revealed that the optimal cosine similarity threshold varies according to the types of SR, as different types of SRs might have different thresholds. The SRAI tool offers a promising solution to the labor-intensive and time-consuming process of selecting studies for SR. Automating parts of this process can significantly reduce the workload and time required. However, this tool is required to evaluate and validate model performance prospectively with other SR projects and various types of studies.

6. Acknowledgements

This manuscript is a part of Phongphat Wiwatthanasetthakarn's thesis within the international M.Sc. program in Data Science for Healthcare, Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University. We wish to convey our appreciation to the department's researchers for providing valuable data to conduct the research, including Dr. Amarit Tansawet, Dr. Napaphat Poprom, Dr. Tunlanut Sapankaew, Dr. Narisa Ruenruengbun, Dr. Ronny Isnuwardana, Dr. Sigit Ari Saputro, Dr. Thitiya Lukkunaprasit, Dr. Philip Bassey, Dr. Kamolpat Chaiyakittiso, and other researchers who have generously supported and distributed the systematic review's dataset for this research.

7. References

Ahn, E., & Kang, H. (2018). Introduction to systematic review and meta-analysis. *Korean Journal of Anesthesiology*, *71*(2), 103–112. https://doi.org/10.4097/kjae.2018.71.2.103

Bassey, P. E., Numthavaj, P., Rattanasiri, S., Sritara, P., McEvoy, M., Ongphiphadhanakul, B., & Thakkinstian, A. (2022). Causal association pathways between fetuin-A and kidney function: A mediation analysis. *The Journal of International Medical Research*, *50*(4), 3000605221082874. https://doi.org/10.1177/03000605221082874

Bražinskas, A., Lapata, M., & Titov, I. (2020). Few-Shot Learning for Opinion Summarization. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4119–4135). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.337

Burgard, T., & Bittermann, A. (2023). Reducing literature screening workload with machine learning: A systematic review of tools and their performance. *Zeitschrift Für Psychologie*, *231*(1), 3–15. https://doi.org/10.1027/2151-2604/a000509

Chaiyakittisopon, K., Pattanaprateep, O., Ruenroengbun, N., Sapankaew, T., Ingsathit, A., Mckay, G. J., Attia, J., & Thakkinstian, A. (2021). Evaluation of the cost-utility of phosphate binders as a treatment option for hyperphosphatemia in chronic kidney disease patients: A systematic review and meta-analysis of the economic evaluations. *The European Journal of Health Economics: HEPAC: Health Economics in Prevention and Care*, *22*(4), 571–584. https://doi.org/10.1007/s10198-021-01275-3

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

DistillerSR. (2024). *Systematic Review and Literature Review Software by DistillerSR*. DistillerSR. Retrieved March 8, 2024, from https://www.distillersr.com/

EPPI Centre. (2024). *EPPI Centre Home*. Retrieved March 8, 2024, from https://eppi.ioe.ac.uk/cms/

Forsgren, E., Wallström, S., Feldthusen, C., Zechner, N., Sawatzky, R., & Öhlén, J. (2023). The use of text-mining software to facilitate screening of literature on centredness in health care. *Systematic Reviews*, *12*(1), 73. https://doi.org/10.1186/s13643-023-02242-0

Gates, A., Gates, M., Sebastianski, M., Guitard, S., Elliott, S. A., & Hartling, L. (2020). The semi-automation of title and abstract screening: A retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and rapid reviews. *BMC Medical Research Methodology*, *20*(1), 139. https://doi.org/10.1186/s12874-020-01031-w

Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. *Systematic Reviews*, *8*(1), 278. https://doi.org/10.1186/s13643-019-1222-2

Gates, A., Johnson, C., & Hartling, L. (2018). Technology-assisted title and abstract screening for systematic reviews: A retrospective evaluation of the Abstrackr machine learning tool. *Systematic Reviews*, *7*(1), 45. https://doi.org/10.1186/s13643-018-0707-8

Hamel, C., Kelly, S. E., Thavorn, K., Rice, D. B., Wells, G. A., & Hutton, B. (2020). An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening – impact on reviewer-relevant outcomes. *BMC Medical Research Methodology*, *20*(1), 256. https://doi.org/10.1186/s12874-020-01129-1

Hoffmann, F., Allers, K., Rombey, T., Helbach, J., Hoffmann, A., Mathes, T., & Pieper, D. (2021). Nearly 80 systematic reviews were published each day: Observational study on trends in epidemiology and reporting over the years 2000-2019. *Journal of Clinical Epidemiology*, *138*, 1–11. https://doi.org/10.1016/j.jclinepi.2021.05.022

Hugging Face. (2024). *Hugging Face – The AI community building the future.* Retrieved February 28, 2024, from https://huggingface.co/

Isnuwardana, R., Bijukchhe, S., Thadanipon, K., Ingsathit, A., & Thakkinstian, A. (2020). Association between vitamin D and uric acid in adults: a systematic review and meta-analysis. *Hormone and Metabolic Research*, *52*(10), 732-741. https://doi.org/10.1055/a-1240-5850

Lukkunaprasit, T., Rattanasiri, S., Turongkaravee, S., Suvannang, N., Ingsathit, A., Attia, J., & Thakkinstian, A. (2020). The association between genetic polymorphisms in ABCG2 and SLC2A9 and urate: An updated systematic review and meta-analysis. *BMC Medical Genetics*, *21*(1), 210. https://doi.org/10.1186/s12881-020-01147-2

Olofsson, H., Brolund, A., Hellberg, C., Silverstein, R., Stenström, K., Österberg, M., & Dagerhamn, J. (2017). Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. *Research Synthesis Methods*, *8*(3), 275–280. https://doi.org/10.1002/jrsm.1237

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Patel, A., Hallock, C., Fusco, N., Cadarette, S. M., & Mody, L. (2021). PNS101 Evaluation of Artificial Intelligence within DistillerSR Software As a Second Reviewer for a Systematic Literature Review. *Value in Health*, *24*, S191. https://doi.org/10.1016/j.jval.2021.04.955

Poprom, N., Numthavaj, P., Wilasrusmee, C., Rattanasiri, S., Attia, J., McEvoy, M., & Thakkinstian, A. (2019). The efficacy of antibiotic treatment versus surgical treatment of uncomplicated acute

appendicitis: Systematic review and network meta-analysis of randomized controlled trial. *American Journal of Surgery*, *218*(1), 192–200. https://doi.org/10.1016/j.amjsurg.2018.10.009

Przybyła, P., Brockmeier, A. J., Kontonatsios, G., Le Pogam, M.-A., McNaught, J., von Elm, E., Nolan, K., & Ananiadou, S. (2018). Prioritising references for systematic reviews with RobotAnalyst: A user study. *Research Synthesis Methods*, *9*(3), 470–488. https://doi.org/10.1002/jrsm.1311

Rathbone, J., Hoffmann, T., & Glasziou, P. (2015). Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic Reviews*, *4*, 80. https://doi.org/10.1186/s13643-015-0067-6

Reddy, S. M., Patel, S., Weyrich, M., Fenton, J., & Viswanathan, M. (2020). Comparison of a traditional systematic review approach with review-of-reviews and semi-automation as strategies to update the evidence. *Systematic Reviews*, *9*(1), 243. https://doi.org/10.1186/s13643-020-01450-2

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (pp. 3982–3992. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Ruenroengbun, N., Numthavaj, P., Sapankaew, T., Chaiyakittisopon, K., Ingsathit, A., Mckay, G. J., Attia, J., & Thakkinstian, A. (2021). Efficacy and safety of conventional antiviral agents in preventive strategies for cytomegalovirus infection after kidney transplantation: A systematic review and network meta-analysis. *Transplant International: Official Journal of the European Society for Organ Transplantation*, *34*(12), 2720–2734. https://doi.org/10.1111/tri.14122

Sapankaew, T., Thadanipon, K., Ruenroengbun, N., Chaiyakittisopon, K., Ingsathit, A., Numthavaj, P., Chaiyakunapruk, N., McKay, G., Attia, J., & Thakkinstian, A. (2022). Efficacy and safety of urate-lowering agents in asymptomatic hyperuricemia: Systematic review and network meta-analysis of randomized controlled trials. *BMC Nephrology*, *23*(1), 223. https://doi.org/10.1186/s12882-022-02850-3

Saputro, S. A., Pattanaprateep, O., Pattanateepapon, A., Karmacharya, S., & Thakkinstian, A. (2021). Prognostic models of diabetic microvascular complications: A systematic review and meta-analysis. *Systematic Reviews*, *10*(1), 288. https://doi.org/10.1186/s13643-021-01841-z

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. *2011 31st International Conference on Distributed Computing Systems Workshops, pp.166–171, Minneapolis, MN, USA*. https://doi.org/10.1109/ICDCSW.2011.20

Tansawet, A., Numthavaj, P., Techapongsatorn, S., Wilasrusmee, C., Attia, J., & Thakkinstian, A. (2020). Mesh position for hernia prophylaxis after midline laparotomy: A systematic review and network meta-analysis of randomized clinical trials. *International Journal of Surgery (London, England)*, *83*, 144–151. https://doi.org/10.1016/j.ijsu.2020.08.059

Tsou, A. Y., Treadwell, J. R., Erinoff, E., & Schoelles, K. (2020). Machine learning for screening prioritization in systematic reviews: Comparative performance of Abstrackr and EPPI-Reviewer. *Systematic Reviews*, *9*(1), 73. https://doi.org/10.1186/s13643-020-01324-7

Uman, L. S. (2011). Systematic Reviews and Meta-Analyses. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *20*(1), 57–59.

Valizadeh, A., Moassefi, M., Nakhostin-Ansari, A., Hosseini Asl, S. H., Saghab Torbati, M., Aghajani, R., Maleki Ghorbani, Z., & Faghani, S. (2022). Abstract screening using the automated tool Rayyan:

Results of effectiveness in three diagnostic test accuracy systematic reviews. *BMC Medical Research Methodology*, *22*(1), 160. https://doi.org/10.1186/s12874-022-01631-8

Veritas Health Innovation. (2024). *Covidence-Better systematic review management*. Covidence. Retrieved March 8, 2024, from https://www.covidence.org/

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 819–824. https://doi.org/10.1145/2110363.2110464

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, *11*(1), 55. https://doi.org/10.1186/1471-2105-11-55

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, *53*(3), 63:1-63:34. https://doi.org/10.1145/3386252

Yu, F., Liu, C., & Sharmin, S. (2022). Performance, Usability, and User Experience of Rayyan for Systematic Reviews. *Proceedings of the Association for Information Science and Technology*, *59*(1), 843–844. https://doi.org/10.1002/pra2.745