Research Synthesis Methods (2025), **0**: 1–14
doi:10.1017/rsm.2025.3

RESEARCH ARTICLE

Machine learning for identifying randomised controlled trials when conducting systematic reviews: Development and evaluation of its impact on practice

Xuan Qin[1,2,3], Minghong Yao[1,2,3], Xiaochao Luo[1,2,3], Jiali Liu[1,2,3], Yu Ma[1,2,3], Yanmei Liu[1,2,3], Hao Li[1,2,3], Ke Deng[1,2,3], Kang Zou[1,2,3], Ling Li[1,2,3] and Xin Sun[1,2,3,4]

[1] Institute of Integrated Traditional Chinese and Chinese Evidence-based Medicine Center and Cochrane China Center and MAGIC China Center, West China Hospital, Sichuan University, Chengdu, China
[2] NMPA Key Laboratory for Real World Data Research and Evaluation in Hainan, Chengdu, China
[3] Sichuan Center of Technology Innovation for Real World Data, Chengdu, China
[4] Sichuan University West China College of Public Health/West China Fourth Hospital, Chengdu, China

Corresponding authors: Xin Sun and Ling Li; Emails: sunxin@wchscu.cn; liling@wchscu.cn

Received: 2 April 2024; **Revised:** 29 November 2024; **Accepted:** 4 December 2024

Keywords: conducting systematic review; ensemble learning; impact on practice; title and abstract screening

Abstract

Machine learning (ML) models have been developed to identify randomised controlled trials (RCTs) to accelerate systematic reviews (SRs). However, their use has been limited due to concerns about their performance and practical benefits. We developed a high-recall ensemble learning model using Cochrane RCT data to enhance the identification of RCTs for rapid title and abstract screening in SRs and evaluated the model externally with our annotated RCT datasets. Additionally, we assessed the practical impact in terms of labour time savings and recall improvement under two scenarios: ML-assisted double screening (where ML and one reviewer screened all citations in parallel) and ML-assisted stepwise screening (where ML flagged all potential RCTs, and at least two reviewers subsequently filtered the flagged citations). Our model achieved twice the precision compared to the existing SVM model while maintaining a recall of 0.99 in both internal and external tests. In a practical evaluation with ML-assisted double screening, our model led to significant labour time savings (average 45.4%) and improved recall (average 0.998 compared to 0.919 for a single reviewer). In ML-assisted stepwise screening, the model performed similarly to standard manual screening but with average labour time savings of 74.4%. In conclusion, compared with existing methods, the proposed model can reduce workload while maintaining comparable recall when identifying RCTs during the title and abstract screening stages, thereby accelerating SRs. We propose practical recommendations to effectively apply ML-assisted manual screening when conducting SRs, depending on reviewer availability (ML-assisted double screening) or time constraints (ML-assisted stepwise screening).

Highlights
What is already known?

- The standard method for retrieving randomised controlled trials (RCTs) from databases during systematic reviews (SRs) often results in only 7% of retrieved citations being real RCTs. Applying machine learning (ML) to identify RCTs is a promising solution.

This article was awarded Open Data badge for transparent practices. See the Data availability statement for details.

https://doi.org/10.1017/rsm.2025.3 Published online by Cambridge University Press

- Existing ML models have seen limited use due to concerns about their performance and practical benefits, which have been questioned by SR practitioners.

What is new?

- We developed a new ML model based on the Cochrane Crowd RCT dataset. This model demonstrated greater workload savings while maintaining a similar recall (0.99) to existing models in both internal and external evaluations.
- When used to assist in manual screening, our model enabled significant labour time savings and improved recall.

Potential impact for readers of research synthesis methods?

- This study may be particularly beneficial for SR practitioners using ML methods to identify RCTs, as it offers a way to accelerate the conduct of SRs.
- Our developed model facilitates workload reduction while minimising the loss of cases, with ML-assisted manual screening offering a promising option for SR practitioners facing limited resources, such as a shortage of reviewers or time constraints.

1. Introduction

Systematic reviews (SRs) focused on randomised controlled trials (RCTs) provide valuable evidence for clinical decision-making.[1–4] Most SRs are assembled from numerous scientific citations that are initially screened by human experts, a process that is both time-consuming and labour-intensive.[5–10] Producing an SR typically takes an average of 67.3 weeks.[11] The standard method for retrieving RCTs from databases often yields results where only 7% of the retrieved citations are real RCTs,[12] making accurate identification of RCTs a crucial task when conducting SRs.[13] Although machine learning (ML)[14,15] models to accelerate the SR workflow by identifying RCTs have been developed, they are not widely used in practice due to concerns about their performance and practical benefits.[15,16]

Existing models[12,13,17–20] have made considerable progress in reducing the workload and minimising missed RCTs, but their performance remains limited; precision is often sacrificed to ensure minimal missing, resulting in limited workload savings.[12,13,17] For example, Thomas et al.[13] reported high recall (0.99) but low precision (0.08) using a Support Vector Machine (SVM) on the Clinical Hedges dataset. Similarly, Marshall et al.[12] obtained comparable results with a convolutional neural network model. More recently, BERT-based models like BioBERT have improved performance, achieving an F1 score of 0.91, which highlights the effectiveness of deep learning in RCT identification.[20] Hybrid approaches that combine crowdsourcing with ML have also been explored to enhance efficiency.[17] Ensemble methods, which combine SVM with PubMed publication-type prediction scores, have achieved a recall of 0.99 and precision of 0.21, demonstrating advances in RCT identification through ML and integrated techniques.[12]

A stacked ensemble approach, which feeds the outputs of several different fine-tuned BERT models into a Light Gradient Boosting Machine (LightGBM)[21] classifier may enhance RCT identification, achieving improved precision and a recall rate as high as 0.99. The transformer architecture[22] of BERT, which underpins models like GPT and ChatGPT, supports advanced text understanding capabilities. Its pre-training[23] on diverse corpora allows for effective fine-tuning and robust cross-language capabilities, which are essential for text classification tasks required in RCT identification.[20] The stacked ensemble method[24] is particularly beneficial as it can detect complex patterns in the data that a single model might overlook, thereby improving overall predictive performance. The combination of LightGBM with BERT's advanced text representation capabilities proves highly effective in text classification tasks, including the detection of fake news.[25]

https://doi.org/10.1017/rsm.2025.3 Published online by Cambridge University Press

Furthermore, the practical benefits of ML in identifying RCTs for title and abstract screening in SRs are often questioned by practitioners due to concerns[26] about the 'black-box' nature of ML. This opacity can obscure the decision-making process, impacting transparency, and reliability, and potentially risking the exclusion of key evidence. SRs typically require at least two reviewers during abstract screening to ensure comprehensiveness[27–29]; however, this process is time-consuming and labour-intensive. ML-assisted screening might address these challenges, potentially replacing the need for a human reviewer.

Few studies have explored the impact of tool-assisted manual screening on practice.[26,30–32] Clark et al.[30] reported that an SR could be completed within 2 weeks using a suite of tools, such as the Word Frequency Analyzer and the Polyglot Search Translator for searching, as well as the SRA Helper and RobotSearch for screening. Their subsequent study[31] demonstrated that teams using these tools saved 70% of the time on six specific tasks compared to those working manually. Perlman-Arrow et al.[32] conducted a real-world study on the impact of ML assistance, reporting a 45.9% reduction in the time spent on title and abstract screening compared to manual methods. These studies highlight the advantages of tool-assisted SRs; however, the presence of various non-ML tools makes it challenging to discern the specific practical impact of ML on the effectiveness of SRs.

This study aims to (1) develop an ensemble learning model to identify RCTs, (2) evaluate its practical impact, and (3) provide recommendations for ML-assisted manual screening to facilitate the rapid conduct of SRs.

2. Methods

We developed and internally evaluated an ensemble learning model using the Cochrane Crowd RCT dataset. The model was then externally evaluated on two RCT datasets, which included annotations with screening time. Following this, we assessed the impact of our model in two practical ML-assisted manual screening scenarios using these newly annotated datasets. The model framework is illustrated in Figure 1.

2.1. *Data sources*

2.1.1. *Development and internal evaluation dataset*

We randomly divided the Cochrane Crowd RCT dataset into three subsets—training, validation, and internal test sets—at a ratio of 3:1:1 for model training, hyperparameter tuning, and internal

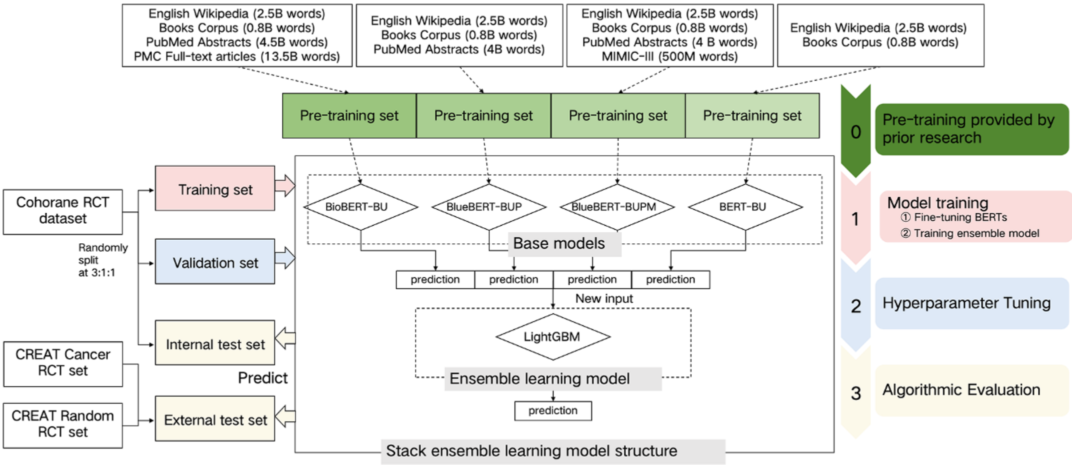

Figure 1. *Model Frame.*

performance evaluation, respectively. The dataset comprised 83,968 citations, including IDs, titles, and abstracts, with 7,356 RCTs (8.8%) and 76,612 not-RCTs (91.2%). The Cochrane Crowd RCT dataset is a robust and diverse resource, sourced from a global community of volunteers who contribute to identifying and classifying RCTs from bibliographic sources such as Embase, PubMed, and ClinicalTrials.gov. The crowd-based annotation process involves volunteers reviewing research study descriptions, with decisions made by at least two reviewers to determine whether the citations describe RCTs or not.[33]

2.1.2. Datasets for external evaluation and assessment of practical impact

To illustrate the practical impact of ML compared with manual screening, we annotated new datasets, as no existing RCT datasets provided screening times and results for each reviewer. We annotated two independent RCT datasets, ensuring no overlap between them, with each citation including screening time and results (see the section "External and practical evaluation datasets" of the Supplementary Material for further details). These external test datasets were distinct from the Cochrane RCT dataset, ensuring an unbiased evaluation of the model's performance on data not included in the original training or validation process. Each dataset comprised 5,000 citations (published between 2013 and 2022 in PubMed) in the form of titles and abstracts. One dataset, randomly selected without specific qualifying topics, was named the CREAT random RCT dataset, while the other, randomly selected from abstracts containing cancer-related topics, was named the CREAT cancer RCT dataset. CREAT represents our research team CREAT (Clinical Research, Evaluation and Translation). Standard manual screening was conducted on these datasets as the gold standard. Paired reviewers independently screened each citation in parallel for titles and abstracts, with a third trained reviewer resolving any disagreements.

2.2. Model development

We developed an ensemble learning model that integrates multiple base models to identify RCTs in SRs (see the section "Model development" of the Supplementary Material for further details). The ensemble model employed was LightGBM. The base models consisted of four BERT variants, all with the same structure and uncased vocabulary list but pretrained on different datasets, resulting in varying initial parameters. These models included BioBERT-BU,[34] BlueBERT-BUP,[35] BlueBERT-BUPM,[35] and BERT-BU[23] (see the section "Detail of different pretrained datasets of BERT models" in the Supplementary Material). The model development proceeded in three stages. The first stage involved fine-tuning the four BERT models using citation text as features to identify RCTs on the same training set. The second stage entailed training LightGBM with the classification results from the four BERT models as features on the training set. In the final stage, we fine-tuned the hyperparameters of LightGBM, including learning rate, tree complexity, and regularisation, to optimise the area under the curve (AUC) during cross-validation on the training set. After optimisation, we calibrated the cutoff on the validation set, prioritising a high recall (0.99) while maximising precision based on the precision-recall curve, to ensure comprehensive identification of all true RCTs while minimising workload.

2.3. Model evaluation

2.3.1. Algorithmic evaluation

Algorithmic evaluation: prior to cutoff adjustment. We performed an algorithmic evaluation of our model (balance) with a 0.5 cutoff to assess its capability to enhance RCT identification. Before adjusting the cutoff, we evaluated the base BERT models and introduced an ensemble voting method *Vote* on the internal test set for comparison. To identify RCTs, the voting method *Vote* generates a score by averaging the prediction probabilities. For this evaluation, the voting method *Vote* (balance) used a 0.5 cutoff. We used standard manual screening as a benchmark to evaluate our model, applying both ML evaluation metrics (precision, recall, and F1 score) and epidemiological metrics (sensitivity and specificity). Recall and sensitivity are included to reflect the dual terminology of ML and epidemiology,

https://doi.org/10.1017/rsm.2025.3 Published online by Cambridge University Press

despite being equivalent measures. Further details of these metrics can be found in the section "Algorithmic evaluation" of the Supplementary Material.

Algorithmic evaluation: post cutoff adjustment. After hyperparameter tuning, we identified the optimal cutoff value for our model using the validation set and uniformly applied this value to evaluate model performance on both internal and external test sets. This evaluation aimed to demonstrate the model's potential to improve the RCT screening process, focusing on enhancing precision while maintaining a high recall rate of 0.99, which is crucial for capturing the maximum number of relevant RCTs.

To validate the model's robustness across diverse data environments, we assessed its performance on both internal and external test datasets. With the adjusted cutoff value, we benchmarked our model against two established methods: the voting method *Vote* and a high-recall SVM[13] model (details of which are provided in the section on "Model evaluation" of the Supplementary Material). The voting method *Vote* modified the cutoff value on the validation set, prioritising a high recall (0.99) while maximising precision based on the precision-recall curve, to ensure comprehensive identification of all true RCTs while minimising workload. These comparisons provided performance baselines, using both algorithmic and practical evaluation metrics (e.g., workload saving; further details of "Practical evaluation" are shown in the Supplementary Material).

2.3.2 Evaluation of practical impact

We assessed the practical impact of our model on the external test sets across two potential ML-assisted manual screening scenarios: ML-assisted double screening (Figure 2a) and ML-assisted stepwise screening (Figure 2b). Using standard manual screening as a benchmark (Figure 2), we established reference points for evaluating practical impact in terms of accuracy (with a focus on epidemiological metrics, particularly sensitivity) and labour saving (including labour time per scenario and labour time saving, detailed in the section "Labour saving" of the Supplementary Material).

Standard manual screening scenario as the benchmark. In the standard manual screening scenario, the process involves paired reviewers independently screening citations in parallel. When their results are consistent, the RCT identification result is taken from the first reviewer. If their results are inconsistent, a third reviewer resolves the disagreement.

We used standard manual screening as a benchmark to calculate both algorithmic and practical evaluation metrics across two separate external test sets. As illustrated in Figure 2, labour time is measured as the cumulative time taken by the paired reviewers. If the paired reviewers' results are inconsistent, the labour time is increased by the duration required for a third reviewer to mediate and resolve the disagreement.

ML-assisted double screening scenario. In the ML-assisted double-screening scenario, the process is as follows: the ML model and a single reviewer independently screen citations in parallel. When their results are consistent, the RCT identification result is taken from the single reviewer. When their results are inconsistent, a third reviewer resolves the disagreement. As illustrated in Figure 2, the labour time is measured as the cumulative time taken by the single reviewer. If the single reviewer's results do not align with the ML model's, the labour time is increased by the duration required for a third reviewer to mediate and resolve the disagreement.

To enhance the reliability of the efficacy evaluation for the ML-assisted double screening method, we conducted two independent tests on the same external test sets, each with a different single reviewer working in parallel with the ML model. Any disagreements were resolved by a third reviewer. We simulated the performance of the single reviewer by using outcomes from one of the reviewers involved in the standard screening process.

Additionally, we performed two tests of the single reviewer screening scenario on the same sets to compare the effects of ML-assisted double screening. For each test, we evaluated the performance of the single reviewer using outcomes from the single reviewer involved in the ML-assisted double screening. The labour time for the single-reviewer screening scenario was determined by aggregating the time spent by that reviewer.

https://doi.org/10.1017/rsm.2025.3 Published online by Cambridge University Press

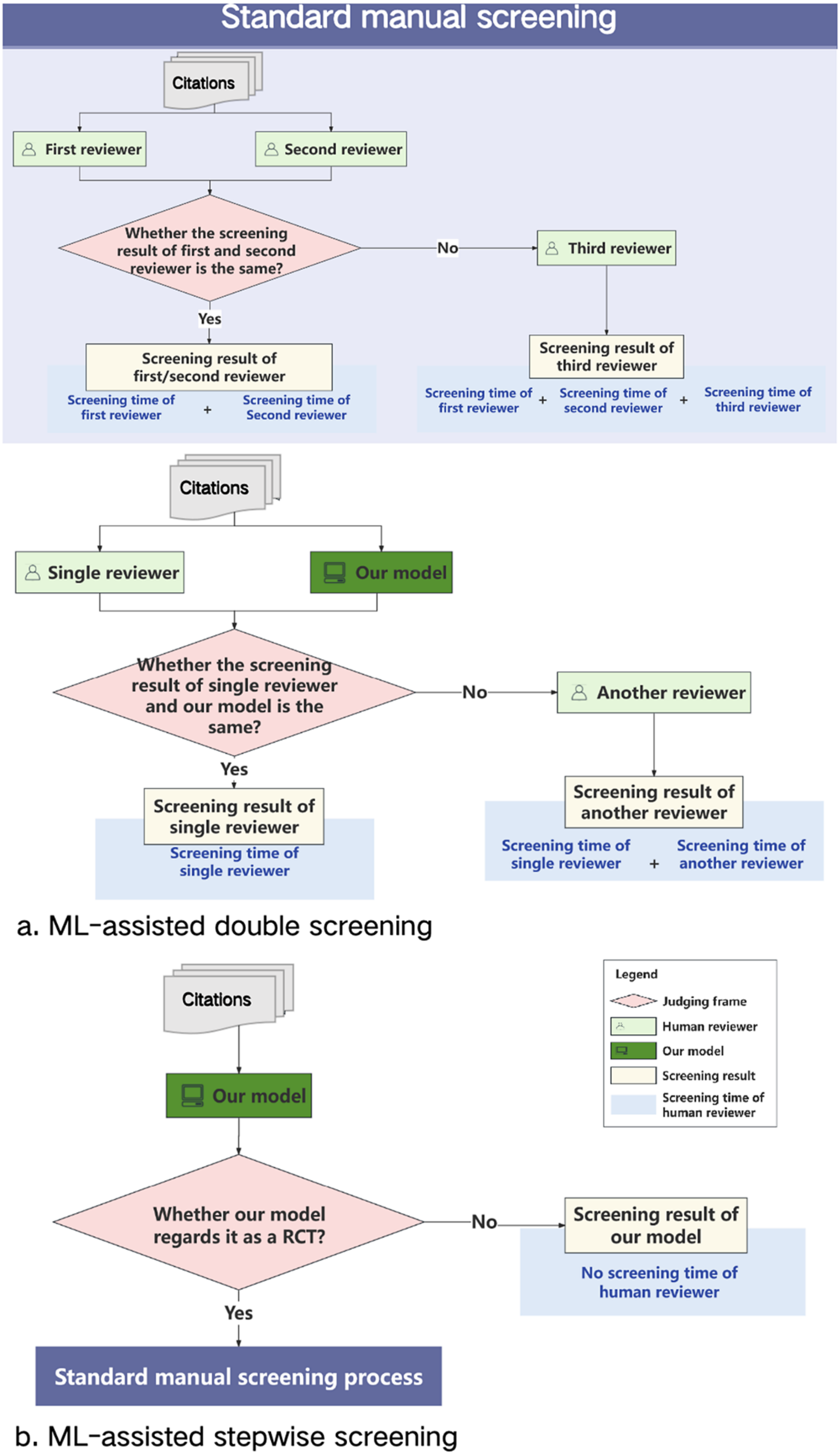
Standard manual screening
Citations
First reviewer
Second reviewer
Whether the screening result of first and second reviewer is the same?
No
Third reviewer
Yes
Screening result of first/second reviewer
Screening time of first reviewer + Screening time of Second reviewer
Screening result of third reviewer
Screening time of first reviewer + Screening time of second reviewer + Screening time of third reviewer
Citations
Single reviewer
Our model
Whether the screening result of single reviewer and our model is the same?
No
Another reviewer
Yes
Screening result of single reviewer
Screening time of single reviewer
Screening result of another reviewer
Screening time of single reviewer + Screening time of another reviewer
a. ML-assisted double screening
Citations
Our model
Whether our model regards it as a RCT?
No
Screening result of our model
No screening time of human reviewer
Yes
Standard manual screening process
Legend
Judging frame
Human reviewer
Our model
Screening result
Screening time of human reviewer
b. ML-assisted stepwise screening

Figure 2. *Overview of workflow and labour time in different screening scenario.*

https://doi.org/10.1017/rsm.2025.3 Published online by Cambridge University Press

ML-assisted stepwise screening scenario. In the ML-assisted stepwise screening scenario, the process involves two steps. In the first step, the ML model screens all citations and flags potential RCTs. In the second step, human reviewers conduct standard manual screening of the flagged citations.

We applied this screening method once to two separate external test sets. The performance of standard manual screening on the flagged citations was assessed using outcomes from the standard screening process. As depicted in Figure 2, the new labour time reflects the time spent on manual screening of citations identified as RCTs by the ML model.

3. Results

3.1. Algorithmic evaluation

3.1.1. Algorithmic evaluation: prior to cutoff adjustment

At this stage, prior to any adjustment of the cutoff, our model demonstrated strong algorithmic performance. It achieved a recall of 0.902, a precision of 0.953, and an F1-score of 0.927 on the internal test dataset from the Cochrane Crowd RCT dataset. Our model outperformed all baseline models (see Table 1), with an F1-score of 0.927 compared to 0.922 for the best baseline models, and a recall of 0.902 compared to 0.889 for the best baseline models.

3.1.2. Algorithmic evaluation: post cutoff adjustment

As shown in Table 2, after adjusting the cutoff, our model maintained a high recall rate of 0.99 while achieving precision scores of 0.353, 0.483, and 0.400 across the Cochrane, CREAT random, and CREAT cancer datasets, respectively. These precision scores significantly surpass those of the SVM and *Vote* methods, effectively reducing false positives. Additionally, our model delivered substantial workload savings, reducing the manual review effort by 75.3%, 73.9%, and 86.4% for the respective datasets. This demonstrates the model's ability to streamline the RCT screening process by enhancing precision and efficiency while preserving a high recall rate of 0.99.

3.2. Evaluation of practical impact

The evaluation of the impact of ML in two practical scenarios is presented in Table 3. Figure 3 visually represents the labour time across different screening scenarios for the CREAT random and CREAT cancer RCT datasets.

3.2.1. Standard manual screening scenario as the benchmark

In the standard manual screening scenario, the results indicated that screening 5,000 citations, which included 632 RCTs in the CREAT random RCT dataset and 338 RCTs in the CREAT cancer

Table 1. *Algorithmic evaluation of models on the internal test from Cochrane Crowd RCT dataset.*

	Total		Algorithmic evaluation			
Model	citations	RCTs	Specificity	Recall/sensitivity	Precision	F1 score
BioBERT-BU	16,795	1472	**0.996**	0.875	0.953	0.912
BlueBERT-BUP			0.995	0.874	0.946	0.909
BlueBERT-BUPM			0.993	0.889	0.929	0.909
BERT-BU			0.994	0.889	0.938	0.912
Vote (balance)			**0.996**	0.889	**0.958**	0.922
Our model (balance)			**0.996**	**0.902**	0.953	**0.927**

Note: The "total citations" column represents the total number of citations in the internal test dataset from Cochrane Crowd RCT dataset. The "RCTs" column indicates the number of citations that were identified as RCT in the internal test dataset from Cochrane Crowd RCT dataset. Bolded values are used to highlight key performance metrics that are particularly significant in the context of our study.

https://doi.org/10.1017/rsm.2025.3 Published online by Cambridge University Press

https://doi.org/10.1017/rsm.2025.3 Published online by Cambridge University Press

Table 2. *Single Model Performance in RCT Identifying: Algorithmic evaluation and practical evaluation on the internal and external test datasets.*

				Algorithmic evaluation				Practical evaluation
Dataset	Model	Total citations	RCTs	Specificity	Recall/sensitivity	Precision	F1score	Workload saving (%)
Cochrane data	SVM	16,795	1472	0.567	0.994	0.181	0.306	51.8
	Vote			0.772	**0.995**	0.295	0.455	70.4
	Our model			**0.825**	**0.995**	**0.353**	**0.521**	**75.3**
CREAT random RCT data	SVM	5,000	632	0.520	**0.998**	0.231	0.376	45.5
	Vote			0.675	**0.998**	0.308	0.471	59.0
	Our model			**0.846**	0.997	**0.483**	**0.651**	**73.9**
CREAT cancer RCT data	SVM	5,000	338	0.646	0.976	0.167	0.285	60.4
	Vote			0.717	**0.997**	0.203	0.338	66.8
	Our model			**0.926**	0.994	**0.400**	**0.570**	**86.4**

Note: In Table 2, the bolded values highlight the improvements in sensitivity, specificity, and workload saving achieved by our model over existing methods like the SVM and Vote ensemble.

Table 3. *Comparative analysis of epidemiological metrics and labour time in screening scenarios for practical evaluation.*

		Epidemiological metrics Average (first, second)		Labour saving Average (first, second)	
Model	Dataset	Sensitivity	Specificity	Labour time per scenario(h)	Labour time saving (%)
Single reviewer	CREAT random RCT dataset	0.927 (0.973,0.881)	0.985 (0.980,0.990)	17.62 (17.37,17.87)	52.0 (52.7,51.3)
	CREAT cancer RCT dataset	0.911 (0.944,0.879)	0.988 (0.986,0.991)	17.77 (17.79,17.74)	51.5 (51.4,51.6)
	Average of these two datasets #	0.919	0.987	17.70	51.7
ML-assisted double screening	CREAT random RCT dataset	0.998 (0.997,1.000)	0.986 (0.981,0.991)	20.17 (19.79,20.55)	45.1 (46.1,44.0)
	CREAT cancer RCT dataset	0.998 (0.997,1.000)	0.990 (0.987,0.992)	19.84 (19.74,19.94)	45.8 (46.1,45.6)
	Average of these two datasets #	0.998	0.988	20.00	45.4
ML-assisted stepwise screening	CREAT random RCT dataset*	0.997	1.000	11.20	69.5
	CREAT cancer RCT dataset*	0.994	1.000	7.59	79.3
	Average of these two datasets #	0.996	1.000	9.39	74.4

Note: Average (first, second): Since we conducted two tests of single reviewer screening and ML-assisted double screening, this represents the result from the first test, the result from the second test, and average values of the results from these two tests. '*': Because we only conducted one test of single ML-assisted stepwise screening, there is only one set of results. The symbol '#' represents the mean performance across two datasets.

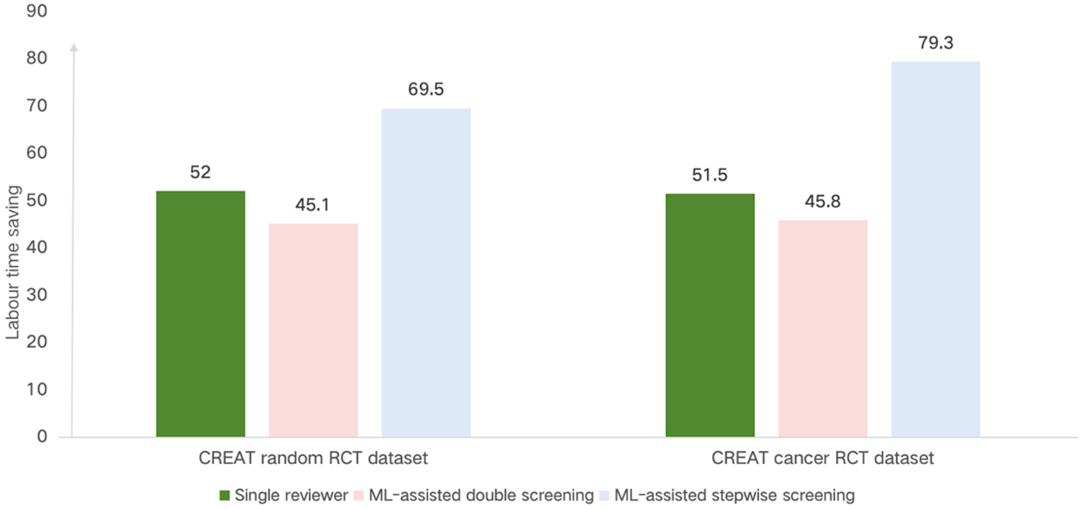

Figure 3. *Comparative analysis of labor time saved (%) compared to standard manual screening across different screening scenarios.*

RCT dataset, took 36.71 h for the random dataset and 36.63 h for the cancer dataset, respectively. Individually, a reviewer took an average of 12.71 s to screen a citation in the random dataset and 12.81 s in the cancer dataset.

3.2.2. ML-assisted double screening scenario

We conducted two tests of ML-assisted double screening on the CREAT cancer RCT dataset and the CREAT random RCT dataset. Our model achieved a higher recall rate than single reviewers, with an average improvement from 0.919 to 0.998. Specifically, the average recall was 0.998 for ML-assisted double screening versus 0.927 for single reviewer screening on the CREAT cancer dataset, and 0.999 versus 0.911 on the CREAT random RCT dataset. However, this improvement in recall resulted in a modest reduction in labour time savings, averaging 45.3% compared to 51.7% for the single reviewer scenario. In detail, the labour time savings were 45.1% versus 52.0% for the CREAT random dataset and 45.8% versus 51.5% for the CREAT cancer dataset. Figure 3 illustrates these findings, with the labour time savings for ML-assisted double screening denoted by a relatively shorter bar, indicating a more efficient labour time requirement compared to the standard screening method.

3.2.3. ML-assisted stepwise screening scenario

In the ML-assisted stepwise screening scenario, our model performed comparably to standard manual screening while achieving substantial labour time savings, averaging 74.4%. Specifically, the savings were 69.5% for the CREAT random RCT dataset and 79.3% for the CREAT cancer dataset. Figure 3 highlights the labour time savings for this scenario, showing ML-assisted stepwise screening with the shortest bar, indicating it as the most time-efficient method among those evaluated.

4. Discussion

4.1. Recommendations for applying ML in identifying RCTs for rapid title and abstract screening in SRs

Based on the varying requirements of different application scenarios, we offer recommendations for 2 ML-assisted manual screening approaches, as outlined in Table 4.

When the number of reviewers available for SRs is limited, ML-assisted double screening is a promising solution. This method reduces labour time by half and outperforms a single reviewer.

https://doi.org/10.1017/rsm.2025.3 Published online by Cambridge University Press

Table 4. *Recommendations for SR reviewers to apply ML-assisted manual screening to identify RCTs for rapid title and abstract screening in SRs.*

Application scenarios	Limited the number of reviewers	Limited time
Application	ML-assisted double screening	ML-assisted stepwise screening
Implementation	Replace one human reviewer with ML in standard manual screening	Step 1: Use the model to screen all the citations Step 2: Conduct standard manual screening on the citations included by ML
Effect	• Reduces labour time by nearly half • Higher recall than a single human reviewer	• Reduces labour time by nearly three-fourths • Performance close to that of standard manual screening
Limitation	When human reviewers and models misclassify simultaneously, real RCTs are lost.	When the model misclassifies RCTs, real RCTs is lost.
Suggestions	Human reviewers should have extensive screening experience.	Randomly sample the citations excluded by the model to check for missing.

Reviewers engaged in double screening with ML should have extensive screening experience to minimise the risk of missing real RCTs. Our model is designed to enhance precision while maintaining a high recall of 0.99, thus ensuring minimal omission of real RCTs and reducing workload. Consequently, the individual consistency between our model and human reviewers in double screening is not very high, but it effectively serves as a useful alert mechanism.

When time constraints are a significant factor in conducting an SR (e.g., for rapid reviews[36]), ML-assisted stepwise screening can reduce labour time by three-fourths while performing comparably to standard manual screening. To prevent missing real RCTs, random sampling of citations excluded by the model should be conducted.

4.2. Main findings and implications

The main findings of this study are as follows. (1) The LightGBM model, which integrates multiple BERT models, demonstrated superior performance in identifying RCTs compared to the existing high-recall SVM model. It reduced the workload significantly while maintaining similar high recall in both internal and external tests. (2) The model showed promising results in 2 ML-assisted manual screening scenarios. In the ML-assisted double screening scenario, it reduced labour time by half and increased the recall compared to a single reviewer. In the ML-assisted stepwise screening scenario, it cut labour time by three-fourths while achieving performance comparable to standard manual screening. (3) Recommendations for ML-assisted manual screening in rapid title and abstract screening for SRs were provided, highlighting the potential benefits of integrating ML techniques in these processes.

4.2.1. Improved workload savings with minimal missing cases in RCTs identification in SRs

Our results demonstrate that the ensemble model, which integrates various BERT models, consistently provides a higher recall than individual BERT models. The different BERT base models, pre-trained on diverse datasets, capture distinct aspects of the underlying patterns and discriminative properties. This allows ensemble learning to extract and interpret semantics from citations more effectively, resulting in more accurate RCT identification.

Previous studies[13,18,3,7] have employed SVMs to reduce workload and minimise missing cases, but these approaches often overlooked semantic information, limiting their effectiveness. By adjusting the cutoff value of our model to achieve a sensitivity/recall of 0.99, we improved performance significantly compared to existing high-recall SVM models. Our modified model achieved a three-fourths reduction in workload while maintaining the same level of sensitivity (0.99) in both internal and external evaluations. The BERT ensemble leverages advanced language understanding capabilities to capture the semantic richness and contextual nuances of medical literature. Its deep learning framework facilitates sophisticated feature extraction and representation, which is crucial for complex classification tasks such as RCT identification. Additionally, the ensemble's combined learning approach addresses individual model weaknesses, enhancing overall accuracy.

4.2.2. Impact of our model in two practical scenarios

Our model demonstrated a promising impact on newly annotated RCT datasets in two potential ML-assisted manual screening scenarios: ML-assisted double screening and ML-assisted stepwise screening.

ML-assisted double-screening scenario. In this scenario, our model achieved a higher recall than a single human reviewer while reducing labour time by nearly half. Although single-reviewer screening could accelerate the process, it is not recommended by guidelines due to the increased risk of missing studies and potential bias.[28,29] ML-assisted double screening offers a solution by improving recall while maintaining similar labour time. Discrepancies between the single reviewer's results and those of the ML model are resolved by a third-party reviewer, ensuring no RCTs are overlooked.

For cases missed by ML-assisted double screening, a thorough review of full texts confirmed that none were RCTs. These omissions were due to the information available in titles and abstracts being insufficient to accurately determine the study type, thus highlighting the limitations of relying solely on these sources for classification.

ML-assisted stepwise screening scenario. The ML-assisted stepwise screening scenario performed comparably to standard manual screening, with a labour time saving of three-fourths. This approach significantly reduces the workload by excluding citations through ML screening before manual review. However, there is a risk of missing important information, making it crucial to use a model with high recall to ensure that real RCTs are not overlooked. This scenario is particularly beneficial for rapid reviews when time is constrained.

4.3. Strengths

A key strength of our study is that the developed model outperformed the high-recall SVM model across diverse datasets, achieving greater workload reduction with minimal missing cases. We also annotated two new RCT datasets, demonstrating the practical benefits of ML-assisted manual screening. To our knowledge, no previous studies have directly compared the impact of ML on labour time and recall improvement with human reviewers. Our findings provide reviewers with insights into the actual benefits of ML. Additionally, we offer recommendations to assist reviewers in applying ML effectively, ensuring the timely completion of SRs.

4.4. Limitations

This study has several limitations. Firstly, our model is limited to identifying English RCTs and may not apply to languages with non-Latin alphabets due to the constraints of BERT's tokenizer. Additionally, the model's complexity requires substantial computational resources, resulting in longer training and prediction times compared to a single BERT model or an SVM model. However, a notable advantage is the reduced need for data pre-processing compared to SVM models, which enhances efficiency and reduces computational demands. There is also a risk of performance overestimation due to overlap with BERT's pre-training datasets, particularly with external PubMed datasets from 2012–2023. Finally, it is crucial to validate the applicability of ML-assisted strategies and their recall thresholds with new data to ensure their continued effectiveness, especially when applied to datasets that differ significantly from the Cochrane RCT dataset used in our study.

5. Conclusions

In conclusion, we developed an ensemble learning model that integrates multiple BERT models. This model effectively identifies RCTs for the rapid title and abstract screening in SRs, significantly reducing workload and minimising missed cases. Our findings suggest that ML-assisted double screening is a promising solution when the number of reviewers is limited. Additionally, ML-assisted stepwise screening proves valuable when time constraints are present in producing SRs.

Author contributions. XQ implemented and wrote the source codes, evaluated the model performance, and drafted the manuscript. MY helped conceive and design the study. XL, JL, YM, YL, and HL helped prepare the data. KD and KZ were the project administrators. LL and XS provided helpful feedback and revised the manuscript.

Competing interest statement. The authors declare that no competing interests exist.

Data availability statement. The annotated data and source code are available on GitHub (https://github.com/RuringQinXuan/RCT_recognition_lightgbm.git).

Funding statement. This study was supported by the National Science Fund for Distinguished Young Scholars (Grant No. 82225049), the National Natural Science Foundation of China (Grant No. 81904134), the Sichuan Provincial Central Government Guides Local Science and Technology Development Special Project (Grant No. 2022ZYD0127), and the Fundamental Research Funds for the Central Public Welfare Research Institutes (Grant No. 2020YJSZX-3).

Supplementary material. To view supplementary material for this article, please visit http://doi.org/10.1017/rsm.2025.3.

References

[1] Kolaski K, Logan LR, Ioannidis JPA. Guidance to best tools and practices for systematic reviews. *Syst Rev*. 2023;12(1): 96. https://doi.org/10.1186/s13643-023-02255-9.

[2] Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: study design: randomised controlled trials. *BJOG*. 2018;125(13): 1716. https://doi.org/10.1111/1471-0528.15199.

[3] Chalmers I The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann N Y Acad Sci*. 1993;703: 156–163; discussion 63–5. https://doi.org/10.1111/j.1749-6632.1993.tb26345.x.

[4] Gupta S, Rajiah P, Middlebrooks EH, et al. Systematic review of the literature: best practices. *Acad Radiol*. 2018;25(11): 1481–1490. https://doi.org/10.1016/j.acra.2018.04.025.

[5] Scheidt S, Vavken P, Jacobs C, et al. Systematic reviews and meta-analyses. *Z Orthop Unfall*. 2019;157(4): 392–399. https://doi.org/10.1055/a-0751-3156.

[6] Akobeng AK. Understanding systematic reviews and meta-analysis. *Arch Dis Child.* 2005;90(8): 845–848. https://doi.org/10.1136/adc.2004.058230.

[7] Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med*. 2010;7(9): e1000326. https://doi.org/10.1371/journal.pmed.1000326.

[8] Tarsilla M. Cochrane handbook for systematic reviews of interventions. *Journal of MultiDisciplinary Evaluation*. 2010;6(14): 142–148. https://doi.org/10.56645/jmde.v6i14.284

[9] National Center for Biotechnology Information (NCBI). PubMed 1988. Available at: https://pubmed.ncbi.nlm.nih.gov/ (Accessed: February 18, 2024).

[10] McKibbon KA, Wilczynski NL, Haynes RB. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. *Health Info Libr J.* 2009;26(3): 187–202. https://doi.org/10.1111/j.1471-1842.2008.00827.x.

[11] Borah R, Brown AW, Capers PL, et al. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*. 2017;7(2): e012545. https://doi.org/10.1136/bmjopen-2016-012545.

[12] Marshall IJ, Noel-Storr A, Kuiper J, et al. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. *Res Synth Methods*. 2018;9(4): 602–614. https://doi.org/10.1002/jrsm.1287.

https://doi.org/10.1017/rsm.2025.3 Published online by Cambridge University Press

[13] Thomas J, McDonald S, Noel-Storr A, et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *J Clin Epidemiol*. 2021;133: 140–151. https://doi.org/10.1016/j.jclinepi.2020.11.003.

[14] Tsafnat G, Dunn A, Glasziou P, et al. The automation of systematic reviews. *BMJ*. 2013;346: f139. https://doi.org/10.1136/bmj.f139.

[15] van Altena AJ, Spijker R, Olabarriaga SD. Usage of automation tools in systematic reviews. *Res Synth Methods.* 2019;10(1): 72–82. https://doi.org/10.1002/jrsm.1335.

[16] O'Connor AM, Tsafnat G, Gilbert SB, et al. Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collabouration for the Automation of Systematic Reviews (ICASR). *Syst Rev*. 2018;7(1): 3. https://doi.org/10.1186/s13643-017-0667-4.

[17] Wallace BC, Noel-Storr A, Marshall IJ, et al. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *J AmMed Inform Assoc*. 2017;24(6): 1165–1168. https://doi.org/10.1093/jamia/ocx053.

[18] Dunn AG, Arachi D, Bourgeois FT. Identifying clinical study types from PubMed metadata: the active (machine) learning approach. *Stud Health Technol Inform*. 2015;216: 867–871.

[19] Cohen AM, Smalheiser NR, McDonagh MS, et al. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. *J AmMed Inform Assoc*. 2015;22(3): 707–717. https://doi.org/10.1093/jamia/ocu025.

[20] Kim J, Kim J, Lee A, et al. Bat4RCT: a suite of benchmark data and baseline methods for text classification of randomized controlled trials. *PLoS One*. 2023;18(3): e0283342. https://doi.org/10.1371/journal.pone.0283342.

[21] LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Neural Information Processing Systems*. 2017, 30. https://dl.acm.org/doi/pdf/10.5555/3294996.3295074

[22] Attention is All you Need. *Neural Information Processing Systems*, 2017. https://doi.org/10.48550/arXiv.1706.03762

[23] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT, 2019*. https://doi.org/10.18653/v1/N19-1423.

[24] Wolpert DHJNN. Stacked generalization. *Neural Networks*. 1992;5(2): 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.

[25] Essa E, Omar K, Alqahtani A. Fake news detection based on a hybrid BERT and LightGBM models. *Complex Intell Syst*. 2023;9(6): 6581–6592. https://doi.org/10.1007/s40747-023-01098-0.

[26] Blaizot A, Veettil SK, Saidoung P, et al. Using artificial intelligence methods for systematic review in health sciences: a systematic review. *Res Synth Methods*. 2022;13(3): 353–362. https://doi.org/10.1002/jrsm.1553.

[27] Waffenschmidt S, Knelangen M, Sieben W, et al. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Med Res Methodol*. 2019;19(1): 132. https://doi.org/10.1186/s12874-019-0782-0.

[28] Gartlehner G, Affengruber L, Titscher V, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *J Clin Epidemiol*. 2020;121: 20–28. https://doi.org/10.1016/j.jclinepi.2020.01.005.

[29] Higgins JP, Green SJW-B. *Cochrane Handbook for Systematic Reviews of Interventions*. *Cochrane Collaboration and John Wiley & Sons, Ltd.: Chichester, UK*. 2008. https://doi.org/10.1002/9780470712184.ch1

[30] Clark J, Glasziou P, Del Mar C, et al. A full systematic review was completed in 2 weeks using automation tools: a case study. *J Clin Epidemiol*. 2020;121: 81–90. https://doi.org/10.1016/j.jclinepi.2020.01.008.

[31] Clark J, McFarlane C, Cleo G, et al. The impact of systematic review automation tools on methodological quality and time taken to complete systematic review tasks: case study. *JMIR Med Educ*. 2021;7(2): e24418. https://doi.org/10.2196/24418.

[32] Perlman-Arrow S, Loo N, Bobrovitz N, et al. A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. *Res Synth Methods*. 2023;14(4): 608–621. https://doi.org/10.1002/jrsm.1636.

[33] Noel-Storr A, Dooley G, Elliott J, et al. An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. *J Clin Epidemiol.* 2021;133: 130–139. https://doi.org/10.1016/j.jclinepi.2021.01.006.

[34] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36(4): 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

[35] Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets, 2019.

[36] Ganann R, Ciliska D, Thomas H. Expediting systematic reviews: methods and implications of rapid reviews. *Implement Sci*. 2010;5: 56. https://doi.org/10.1186/1748-5908-5-56.

[37] Al-Jaishi AA, Taljaard M, Al-Jaishi MD, et al. Machine learning algorithms to identify cluster randomized trials from MEDLINE and EMBASE. *Syst Rev*. 2022;11(1): 229. https://doi.org/10.1186/s13643-022-02082-4.

Cite this article: Qin X, Yao M, Luo X, Liu J, Ma Y, Liu Y, Li H, Deng K, Zou K, Li L, Sun X. Machine learning for identifying randomised controlled trials when conducting systematic reviews: Development and evaluation of its impact on practice. *Research Synthesis Methods*. 2025: 1–14. https://doi.org/10.1017/rsm.2025.3