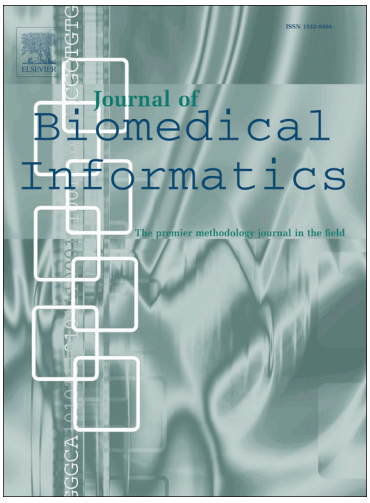
Journal Pre-proofs



Deep Learning to Refine the Identification of High-Quality Clinical Research Articles from the Biomedical Literature: Performance Evaluation

Cynthia Lokker, Elham Bagheri, Wael Abdelkader, Rick Parrish, Muhammad Afzal, Tamara Navarro, Chris Cotoi, Federico Germini, Lori Linkins, R. Brian Haynes, Lingyang Chu, Alfonso Iorio

PII: S1532-0464(23)00105-3
DOI: https://doi.org/10.1016/j.jbi.2023.104384
Reference: YJBIN 104384

To appear in: *Journal of Biomedical Informatics*

Received Date: 15 November 2022
Revised Date: 24 April 2023
Accepted Date: 3 May 2023

Please cite this article as: Lokker, C., Bagheri, E., Abdelkader, W., Parrish, R., Afzal, M., Navarro, T., Cotoi, C., Germini, F., Linkins, L., Brian Haynes, R., Chu, L., Iorio, A., Deep Learning to Refine the Identification of High-Quality Clinical Research Articles from the Biomedical Literature: Performance Evaluation, *Journal of Biomedical Informatics* (2023), doi: https://doi.org/10.1016/j.jbi.2023.104384

Deep Learning to Refine the Identification of High-Quality Clinical Research Articles from the Biomedical Literature: Performance Evaluation

Authors:

Cynthia Lokker[a], PhD MSc, Elham Bagheri[a], PhD; Wael Abdelkader[a], MD MSc; Rick Parrish[a], Muhammad Afzal[b], PhD; Tamara Navarro[a], MLIS, Chris Cotoi[a], BEng, EMBA, Federico Germini[a,c] MD MSc, Lori Linkins[c], MD, MSc, R. Brian Haynes[a,c], MD, PhD; Lingyang Chu[d], Alfonso Iorio[a,c], MD, PhD

Affiliations:

[a]Health Information Research Unit, Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

[b]Department of Computing, Birmingham City University, Birmingham, UK

[c]Department of Medicine, McMaster University, Hamilton, Ontario, Canada

[d]Department of Computing and Software, McMaster University, Hamilton, Ontario, Canada

Corresponding author: Dr. Cynthia Lokker, McMaster University, 1280 Main St W, CRL 137, Hamilton, ON, Canada, L8S 4K1,

Abstract

Background: Identifying practice-ready evidence-based journal articles in medicine is a challenge due to the sheer volume of biomedical research publications. Newer approaches to support evidence discovery apply deep learning techniques to improve the efficiency and accuracy of classifying sound evidence.

Objective: To determine how well deep learning models using variants of Bidirectional Encoder Representations from Transformers (BERT) identify high-quality evidence with high clinical relevance from the biomedical literature for consideration in clinical practice.

Methods: We fine-tuned variations of BERT models ($BERT_{BASE}$, BioBERT, BlueBERT, and PubMedBERT) and compared their performance in classifying articles based on methodological quality criteria. The dataset used for fine-tuning models included titles and abstracts of >160,000 PubMed records from 2012-2020 that were of interest to human health which had been manually labeled based on meeting established critical appraisal criteria for methodological rigor. The data was randomly divided into 80:10:10 sets for training, validating, and testing. In addition to using the full unbalanced set, the training data was randomly undersampled into four balanced datasets to assess performance and select the best performing model. For each of the four sets, one model that maintained sensitivity (recall) at ≥99% was selected and were ensembled. The best performing model was evaluated in a prospective, blinded test and applied to an established reference standard, the Clinical Hedges dataset.

Results: In training, three of the four selected best performing models were trained using $BioBERT_{BASE}$. The ensembled model did not boost performance compared with the best individual model. Hence a solo BioBERT-based model (named DL-PLUS) was selected for further

testing as it was computationally more efficient. The model had high recall (>99%) and 60% to 77% specificity in a prospective evaluation conducted with blinded research associates and saved >60% of the work required to identify high quality articles.

Conclusions: Deep learning using pretrained language models and a large dataset of classified articles produced models with improved specificity while maintaining >99% recall. The resulting DL-PLUS model identifies high-quality, clinically relevant articles from PubMed at the time of publication. The model improves the efficiency of a literature surveillance program, which allows for faster dissemination of appraised research.

Keywords: bioinformatics; machine learning; evidence-based medicine; literature retrieval; medical informatics; Natural Language Processing.

Abbreviations

BERT Bidirectional Encoder Representations from Transformers

HiRU Health Information Research Unit

PLUS McMaster Premium Literature Service

PRLM pre-trained language models

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

1. Introduction

Evidence based medicine integrates clinical expertise; patient circumstances, values, and preferences; and best available evidence from clinical research in the process of health care decision making. It provides a coherent and structured framework for assessing and applying the best evidence to complement or reduce reliance on expert opinion for patient care

decisions [1]. More than 1.5 million new citations were indexed in PubMed in 2021 alone [2]; this volume of publication is a major barrier to retrieving high quality content, especially given that only about 1% of published clinical studies meet methodologic criteria for scientific rigor for use in clinical decision making [3].

Statement of significance

Problem or Issue Access to high-quality, clinically relevant research is hindered by the volume of published research, required critical appraisal skills, and available time.
What is Already Known Various informatics approaches have been applied to retrieve high quality evidence to make it timely and accessible to clinicians. Pretrained language models, such as BERT and its variants, make it easier to tackle natural language processing tasks. Generally, models have been trained to identify high-quality articles focused on treatment studies.
What this Paper Adds Using a large database of clinical articles from 2012 to 2020 that were manually classified for methodological rigor, we fine-tuned BERT-variant deep learning models to identify high-quality, clinically relevant evidence from the biomedical literature at the time of publication for use in a real-time literature surveillance program. We selected and tested a model trained on BioBERT that classifies, by quality, articles across a range of study purpose categories with >98% sensitivity and 73% specificity, that improves the efficiency of the literature surveillance process by >60%.

1.1 Background

Information retrieval to support evidence-based practice relies on retrieving articles from the bibliographic repositories and then screening and appraising the articles manually. It is impossible for readers to keep up with all potentially clinically relevant articles and ensure they are of high quality. Automatic knowledge extraction and mining from the biomedical literature is therefore in high demand [4]. Empirical search filters like PubMed Clinical Queries [5] which are based on text mining and logical combinations of text strings, Medical SubHeadings terms, and database tags, have been developed to optimize sensitivity, specificity, or the best balance between them, for different clinical study purpose categories [6]. Such search filters reduce the manual screening burden and improve the retrieval of relevant clinical studies [7], [8], but they can be limited by their partial reliance on Medical SubHeadings indexing terms, as it can take up to a year for articles to be indexed in MEDLINE [9].

Machine learning is a powerful approach to automate and increase the speed and efficiency of manual processes, and researchers have employed these techniques to find high quality evidence. Earlier studies applied conventional machine learning approaches and feature engineering for this purpose [10]–[12]. Recent advances in deep neural networks have been established as the state-of-the-art models for biomedical text classification. While conventional supervised machine learning models require manual feature engineering, deep learning models take raw text directly as the input and work in an end-to-end manner, i.e., the model learns all the steps between the initial input phase and the final output result. Neural network models, such as convolutional neural networks, have shown superior performance for identifying high quality clinical treatment studies compared with manually created Boolean search filters [13], [14].

Owing to recently developed pre-trained language models (PTLMs), natural language processing (NLP) is gradually shifting to a two-stage pre-training and fine-tuning paradigm that is suitable when supervised data is limited but large-scaled unsupervised data is readily available, which is a typical scenario in the biomedical domain [15]. PTLMs are based on transfer learning in which a machine learning model developed for one task is reused as the starting point for a model on a different but related task. When trained on a large body of text, pretrained models can learn universal language representations which can be beneficial for downstream NLP tasks and avoid the need for training a model anew. Fine-tuning language models is mainstream for PTLM adaption [16] and examples include Universal Language Model Fine-Tuning [17] and Bidirectional Encoder Representations from Transformers (BERT) [18].

Pre-trained transformer-based neural language representation models like $BERT_{BASE}$ and its variants [19]–[23] are contextual language models that excel at several natural language understanding tasks as well as text summarization [24], retrieval, question answering, named entity recognition, document classification [25], and biomedical information extraction tasks [26], [27]. $BERT_{BASE}$ was pre-trained using text data from BookCorpus and English Wikipedia [16]. After a BERT model is pre-trained, it can be shared and fine-tuned for different NLP tasks. $BERT_{BASE}$ has been fine-tuned for the biomedical domain using several relevant datasets; the resulting PTLMs include BioBERT, which was pretrained using PubMed abstracts and PubMed

Central full-text articles [22]; BlueBERT, pretrained using PubMed text and clinical notes from MIMIC-III [28]; PubMedBERT, pretrained using domain-specific text data from 14 million PubMed abstracts [29]; and SciBERT which was pretrained on a random sample of 1.14 M scientific papers from Semantic Scholar [21], a corpus of full text of papers that closely resembles MEDLINE articles. BioBERT, BlueBERT, SciBERT and PubMedBERT have all been applied for downstream biomedical tasks including named entity recognition, relation extraction, text classification, and sentence similarity [15].

Databases of clinical articles that are manually tagged for clinical purpose category and methodological rigor have been used for development of search filters [8], [30], [31] and machine learning models [10], [13], [32], [33]. The creation of Clinical Hedges [8], a dataset of almost 50,000 clinical articles relevant to human health and published in 2000 across 170 journals compiled by the Health Information Research Unit (HiRU) at McMaster University, has pioneered work in this area [34], [35]. Hedges is frequently used as a reference standard for high quality reports of clinical studies for new model development and testing.

To support evidence-based practice, HiRU conducts daily surveillance of PubMed through the McMaster Premium Literature Service (PLUS) process (see Figure 1)[36], [37]. Retrieved articles are appraised for methodological rigor and rated for clinical relevance; these are then shared through push (email alerts) and pull (searchable database) mechanisms to practicing clinicians and other knowledge users such as authors of guidelines, reviews, and online textbooks [38].

1.2 Objectives

In this work we apply deep learning to classify articles across a range of clinical categories and methods from PubMed by evidence quality to support daily evidence surveillance [38]. The objective is to investigate how modern-day deep learning models can be used to identify high-quality, clinically relevant evidence from the biomedical literature at the time of publication for use in a real-time literature surveillance program that supports access to the best clinical evidence for practicing clinicians.

2. Methods

2.1 Dataset construction

The datasets for the current project include articles from ~120 clinical journals published between 2012-2020 that were appraised following the methodological and quality criteria used to create the Clinical Hedges dataset [8]. The process of creating the dataset includes daily searches of PubMed for all indexed journal articles in the ~120 journals using a highly sensitive Boolean search filter of methods terms adapted from Clinical Queries to filter in articles related to human health that are potentially ready for clinical practice (Figure 1). In 2019, the search filter reduced the records indexed in the select journal titles in PubMed from 59 052 to 17 349 (29%). The articles retrieved from PubMed are manually classified by expert research associates to an article type (original study, systematic review, or evidence-based guideline) and one or more of the following purpose categories: treatment, primary prevention, diagnosis, harm from clinical interventions, economics, overall prognosis, clinical prediction guide, or quality improvement [37]. The research associates then apply critical appraisal rules (Appendix A) to establish if the study meets rigor for the specified category or not [39]. For example, an article addressing treatment, primary prevention, or quality improvement questions, would be appraised for random allocation of participants to comparison groups, including ≥ 10 participants per group, having primary outcome(s) assessed in ≥ 80% of those randomized, and reporting an outcome measure of known or probable clinical importance (Appendix A). Articles not meeting the criteria are classified as negative and dropped from the remaining process but added to the ML dataset (Figure 1). Articles that meet rigor criteria are classified as positive articles, confirmed by a clinical editor, and rated by ≥4 clinicians from a community of >4000 for clinical relevance and newsworthiness on Likert scales of 1 to 7 [36]; articles with an average score ≥4 are subsequently disseminated to users of the McMaster PLUS system (Figure 1). This process of selecting clinically relevant articles is further described by Haynes et al [40] and, in an earlier study, the inter-rater reliability of the critical appraisal step had kappa >80% for all categories of articles [41]. The dataset, therefore, contains articles from across types and categories that are classified as positive or negative for meeting methodologic criteria for the particular article type/category.

Since the onset of COVID-19, and in addition to searching the 120 journals, all of PubMed has been searched daily using COVID-19 topic-specific search terms with searches not restricted to the core journal titles. By August 12, 2022, 47,716 additional articles related to COVID-19 were retrieved and assessed for rigor for the same article types and categories described above.

2.2 Machine Learning Models

Using the Python programming language, four pre-trained BERT variants were selected to fine-tune and evaluate. The models used were $BERT_{BASE}$ cased ($BERT_C$)[18], BioBERT [22], BlueBERT [28], and PubMedBERT [29]. These models were selected as they were available in Hugging Face and perform well in the Biomedical Language Understanding and Reasoning Benchmark leaderboard [42].

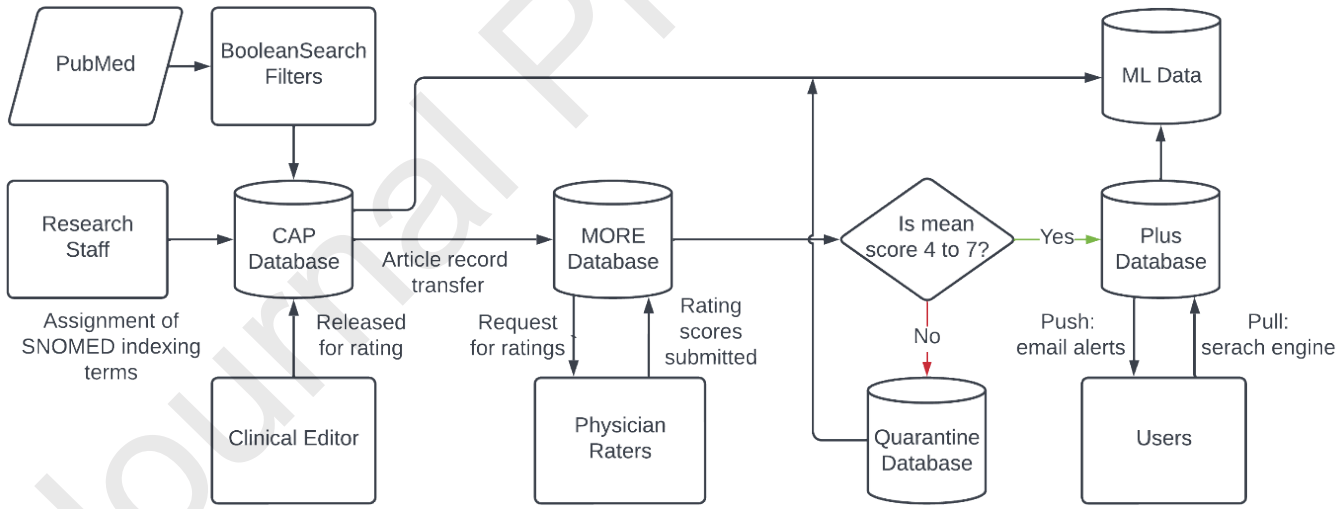


Figure 1. Steps in database creation and article classification during daily literature surveillance for high-quality, clinically relevant articles. Retrievals from PubMed are filtered using empirically validated Boolean searches. CAP = Article Appraisal Process; ML = machine learning; MORE = McMaster Online Rating of Evidence; Plus = Premium Literature Service.

2.3 Experiments

The goal of our experiments was to investigate how well fine-tuned BERT-variant deep learning models could identify, at the time of publication, high-quality, highly relevant clinical evidence across article types and categories with high recall. The process is depicted in Figure 2. We leveraged the high-performance computing resources provided by Compute Canada's Cedar cluster. Our experiments ran on 8 Intel Silver 4216 Cascade Lake CPU cores , an NVIDIA V100 Volta (32G HBM2 memory) GPU, and 40GB RAM.

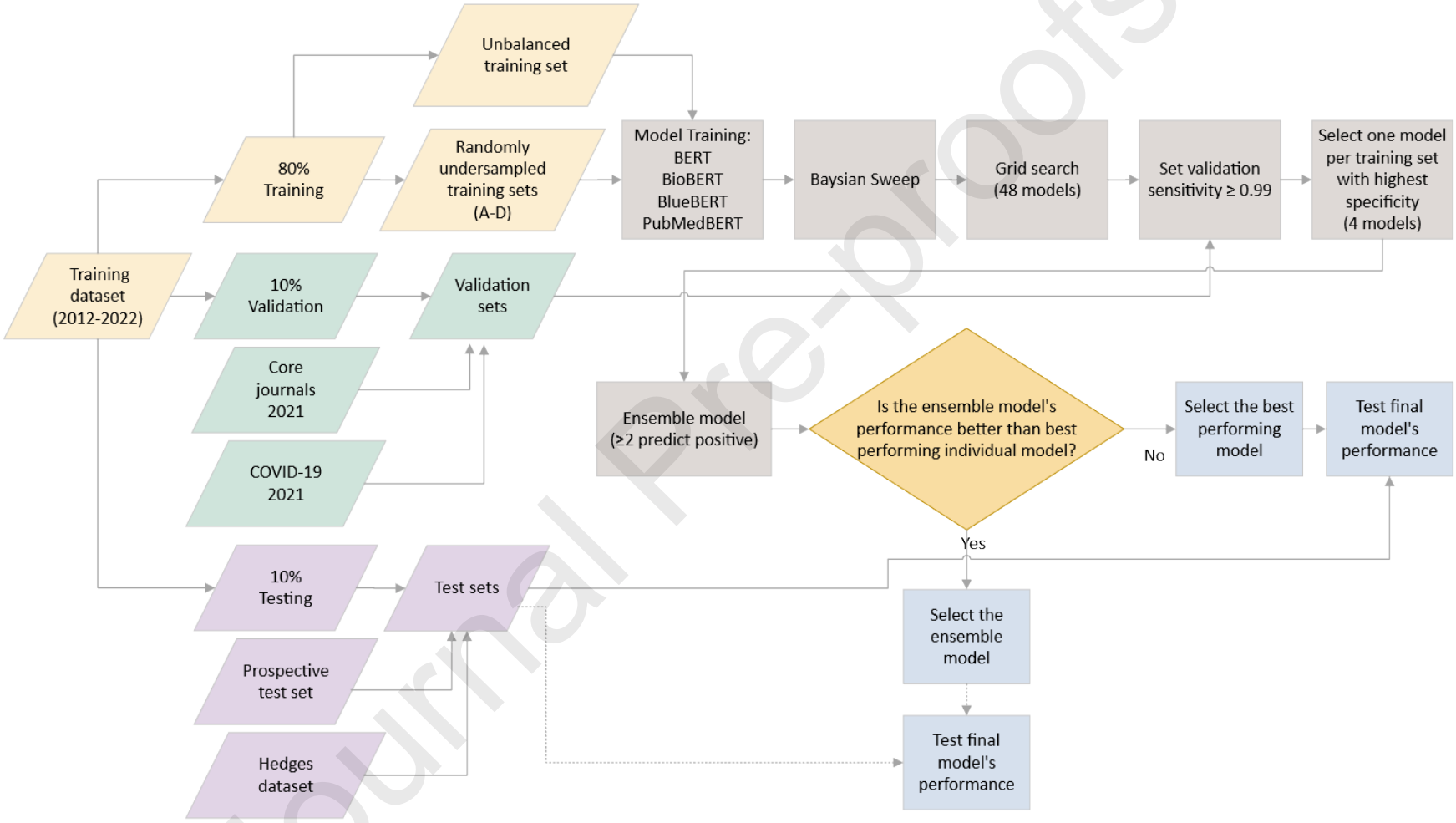


Figure 2. Flow of steps in training, validating, and testing the deep learning models.

2.3.1 Data

For the experiments, training, validation, and test datasets were derived from the ML dataset (Figure 1) that included titles and abstracts of 160,712 articles published in the core journal titles between 2012-2020 identified by PubMed identifier. Across article categories, there were 29,810 positive articles that fulfilled all methodological rigor criteria and 130,902 negative articles that failed to meet ≥1 methodological rigor criteria for the article categories assessed in McMaster PLUS (Table 1). The dataset was randomly split into 80% training, 10% validation, and 10% test subsets. In the training subset, to reduce noise, we removed 7692 articles with <128 words in the text field to remove PubMed entries of corrections, etc., that did not include a full abstract [29]. Additional independent datasets comprising 11,506 articles from the PLUS core journals in 2021, and 19,516 COVID-19-related articles across all journals in PubMed in 2021, were used during validation to support selection of top performing models (Table 1). Test datasets include the hold-out set, a prospective dataset of 11,274 articles, and the Clinical Hedges dataset.

2.3.2. Preparing the training dataset

The dataset is large and unbalanced which may introduce a bias towards the majority class [43]; the training data includes four times as many negative articles as positive. In addition to carrying out the training process using the unbalanced set, we used random undersampling to balance the article positive/negative classes to a 50:50 ratio by creating four smaller training sets. Each smaller balanced training set was created with a random selection of 25% of the negative articles and each of the positive articles (i.e., each set had the same positive articles and a different subset of the negative articles) (Table 1).

Table 1. Datasets used in the experiments.

Dataset	**N**	**Positive articles**	**Negative articles**
Total 2012-2020	160,712	29,810	130,902
-Training 2012-2020 (80%)*	120,877	23,801	97,076
-4 Balanced training subsets (A,B,C,D)†	48,070	23,801	24,269
-Validation 2012-2020 (10%)	16,071	2953	13118

-Test 2012-2020 (10%)	16,072	3000	13072
Validation 2021 Core Journals	11,506	3491	8015
Validation 2021 COVID-19	19,516	827	18,689
Prospective test set 2022	11,274	1376	9898‡
Clinical Hedges test set 2000§	49,024	3036	45,988

*Unbalanced training set; 7692 articles with title + abstract <128 words were removed.

†4 balanced datasets, each containing all positive articles and a random selection of 25% of the negative articles were derived from the training dataset.

‡This value includes 5385 articles that were predicted to be negative but that were not assessed by the research associates.

§Wilczynski et al.[8]

2.3.3. Hyperparameter optimization

Initially, we ran a Bayesian sweep [44] using hyperparameters suggested in the publications of the pre-trained models (Table 2). These ran for 24 hours using 4 epochs, 2 to 4 batch sizes, and learning rates in the range listed in Table 2. From these, we determined that some combinations underperformed and were removed. The second stage included the reduced hyperparameter combinations and a grid search to find the best hyperparameters for each variation of BERT (Figure 2). In this second stage, we trained models using 2 learning rates (3e-5 and 5e-5), 2 epochs (2 and 3), and 3 batch sizes (16, 32, and 64).

Table 2. Bayesian sweep of hyperparameters for each pre-trained model using Weights & Biases platform.

	BERT and BlueBERT	BioBERT	PubMedBERT
Learning rate	2e-5 to 5e-5	1e-5 to 5e-5	1e-5 to 5e-5
Training epochs	1-4	1-4	1-4
Train batch size	16, 32	10, 16, 32, 64	16, 32

2.3.4 Model tuning and selection

Using each of the 4 balanced datasets, we fine-tuned 48 models (4 pre-trained BERT-variant models BY 12 optimized hyperparameter configurations), yielding 192 models; and an additional 12 using BioBERT and the unbalanced dataset. Articles in the validation hold-out and 2021 validation datasets were processed by each model. The area under the curve, sensitivity, and specificity were calculated. Only models achieving ≥99% sensitivity were considered and the model with optimal specificity within the validation datasets (hold-out and articles from 2021, See Figure 2) was selected for each training dataset.

We then ensembled the four top models trained using the balanced datasets by voting using an extra positive vote, i.e., we classified articles as positive if ≥2 predicted the positive class to determine if it boosted performance compared with the solo models. A voting algorithm is one type of ensemble model that aggregates identical or conceptually variant ML classifiers for prediction via voting [45]. It can be viewed as a wrapper for a set of different classifiers that are trained and evaluated in parallel with the purpose of exploiting the peculiarities of each algorithm.

2.3.5 Model testing

The performance of the solo and ensembled models were compared to select a top performing model for testing and prospective implementation. The selected BioBERT model, hereafter termed DL-PLUS, was trained using one of the smaller balanced datasets. It was applied retrospectively to the Clinical Hedges dataset and prospectively tested in real-time and real-world application through our literature surveillance process from March 12th, 2022 to Aug 12, 2022. Following the daily process depicted in Figure 1, articles filtered through Boolean search strategies were downloaded from PubMed to the CAP database; classifications were made using DL-PLUS and all articles predicted to be positive and a random selection of 25% of the articles predicted to be negative were appraised by research associates blinded to the model classification. We calculated sensitivity, specificity, accuracy, precision, the number of articles needed to read (NNR =1/precision) which is a measure of effort required for the literature surveillance program, as well as work saved over sampling @99% recall (WSS@99%) as a measure of efficiency and reduced workload (Table 3).

Table 3. Performance metrics definitions and formula.

Measure	Definition	Formula
Sensitivity (recall)	The proportion of correctly identified positives among the real positive.	TP/TP+FN
Specificity	the proportion of actual negatives, which got predicted as negative (or true negative)	TN/TN+FP
Accuracy	the number of correctly predicted documents out of all classified documents.	TP+TN/TP+FP+FN+TN
Precision	Proportion of correctly identified positives among all classified positives.	TP/TP+FP
AUC	The area under the curve is traced out by graphing the true positive rate against the false-positive rate. The higher the AUC, the better the classifier prediction.	
Number needed to read	The number of articles that need to be read before finding one that is positive (meets criteria)	1/precision
Work saved over sampling at 99% recall [46]	The percentage of all articles that are predicted negative by the algorithm and therefore not reviewed	(TN +FN)/N – (1 – recall) = (TN + FN)/N – 0.01

TP: true positive; TN: true negative; FN: false negative; FP: false positive.

3. Results

3.1 Model performance

Based on performance in the hold-out dataset and 2021 core journal and COVID-19 validation datasets, we selected one model per dataset among the models developed on each set that optimized specificity when sensitivity was set to ≥99%. Three using the balanced datasets were BioBERT-based and one was BlueBERT based (Table 4). They were combined in a voting ensemble with articles classified as positive if ≥2 of the 4 models predicted positive. Performance in the hold-out validation set is presented in Table 4. The model derived from

dataset D which performed better compared to other models and the ensembled model was applied to the 2021 datasets to confirm performance (Table 5). In the hold-out and independent validation datasets from 2021, the ensembled model did not provide a boost in performance. Therefore, the model derived from dataset D (hearafter named DL-PLUS) was selected for prospective evaluation.

Table 4. Performance of the top performing models derived from the unbalanced dataset, each balanced data set, and the ensembled majority vote in the hold-out validation dataset

Parameter (95% CI)	**performance of model from each dataset (95% CI)**					
	Unbalanced model	**Set A model**	**Set B model**	**Set C model**	**Set D model (DL-PLUS)**	**ABCD Ensemble***
Pretrained model	**BioBERT**	**BioBERT**	**BlueBERT**	**BioBERT**	**BioBERT**	
Sensitivity†	99.0% (98.6-99.4)	99.0% (98.7-99.4)				99.1% (98.7-99.4)
Specificity	66.6% (65.8-67.4)	66.2% (65.4-67.0)	59.5% (58.7-60.4)	70.2% (69.5-70.9)	70.2% (69.4-71.0)	69.7% (69.0-70.5)
Accuracy	72.6% (71.9-73.3)	72.2% (71.5-72.9)	66.8% (66.1-67.5)	75.5% (74.8-76.1)	75.5% (74.8-76.2)	75.14% (74.5-75.8)
Precision	40.1 (38.9-41.2)	39.7% (38.6-40.9)	35.5% (34.5-36.6)	42.7% (41.6-43.9)	42.8% (41.6-44.0)	42.4% (41.3-43.6)
AUC	0.97	0.97	0.96	0.97	0.97	NA
NNR	2.50 (2.43-	2.52	2.81 (2.74-	2.34	2.34	2.36 (2.29-

	2.57)	(2.45-2.59)	2.90)	(2.28-2.40)	(2.27-2.40)	2.42)
WSS@99%	54%	53%	48%	56%	56%	56%

AUC = area under the curve; NA = not applicable; NNR = number needed to read (1/precision); WSS@99% = work saved over sampling at 99% recall.

*Articles classified as positive if ≥2 of the 4 models predict positive.

†Sensitivity was set at 99% hence each model that was selected had equivalent values.

Table 5. Performance of DL-PLUS, the ensembled model, and the model trained using the unbalanced dataset in independent validation datasets from 2021 representing core journal articles and COVID-19 articles.

	Performance in 2021 core journal dataset (95% CI)			**Performance in 2021 COVID-19 dataset (95% CI)**		
	DL-PLUS	**Ensemble**	**Unbalanced**	**DL-PLUS**	**Ensemble**	**Unbalanced**
N	11506	11506	11506	19516	19516	19516
Sensitivity	99.5% (99.3-99.8)	99.6% (99.3-99.8)	99.3% (99.0-99.6)	98.7% (97.9-99.4)	98.9% (98.2-99.6)	98.1% (97.1-99.0)
Specificity	60.7% (59.5-61.6)	59.8% (58.8-60.9)	56.0% (54.9-57.1)	77.3% (76.7-77.9)	77.7% (77.1-78.3)	75.5% (74.9-76.1)
Accuracy	72.4% (71.6-73.2)	71.9% (71.1-72.7)	69.1% (68.3-70.0)	78.2% (77.6-78.8)	78.6% (78.1-79.2)	76.5% (75.9-77.1)
Precision	52.4% (51.2-53.6)	51.9% (50.7-53.1)	49.6% (48.4-50.7)	16.1% (15.1-17.1)	16.4% (15.4-17.4)	15.0% (14.1-16.0)
NNR	1.91 (1.87-1.95)	1.93 (1.88-1.97)	2.02 (1.97-2.07)	6.21 (5.84-6.62)	6.10 (5.74-6.51)	6.66 (6.26-7.11)
WSS@99%	42%	41%	38%	73%	73%	71%

NNR = number needed to read (1/precision); WSS@99% = work saved over sampling at 99% recall.

3.2 Independent tests

During the independent test phase, 11,274 articles pre-filtered from PubMed were processed by DL-PLUS. Of those, 4068 were predicted positive and 7206 were predicted negative. All positive and a random 1771 predicted negative were critically appraised by blinded research associates (Figure 3). Performance of the model across all articles, those from the core journal set, and those related specifically to COVID-19 outside of the core journal set are described in Table 6. DL-PLUS performance in 49,024 articles in Clinical Hedges is also in Table 6.

Table 6. Performance of DL-PLUS in a prospective test within an active literature surveillance process between March 12th, 2022 and Aug 12, 2022 and the Hedges dataset.

	Performance (95% CI)			
Parameter (95% CI)	**All 2022 articles (n=11,274)**	**Core journal articles (n=3774)**	**COVID-19 articles (n=7500)**	**Hedges (n=49,024)**
Sensitivity	99.7% (99.4-100)	99.8% (99.5-100)	99.4% (98.5-100)	83.7% (82.4-85.0)
Specificity	72.8% (71.9 - 73.6)	60.2% (58.4 - 62.1)	77.5% (76.5 - 78.5)	80.8% (80.5-81.2)
Accuracy	76.1% (75.3 - 76.8)	71.4% (69.9 - 72.8)	78.4% (77.5 - 79.3)	81.0% (80.7-81.3)
Precision	33.3% (32.3 - 35.7)	50.0% (47.6 - 52.3)	16.7% (14.5 – 14.5)	22.2% (21.7 -23.2)
NNR	3.0 (2.8 - 3.1)	2.0 (1.9 - 2.1)	6.2 (5.6 - 6.9)	4.5 (4.3 to 4.6)
WSS@99%	63%	42%	73%	61%

NNR = number needed to read (1/precision); WSS@99% = work saved over sampling at 99% recall.

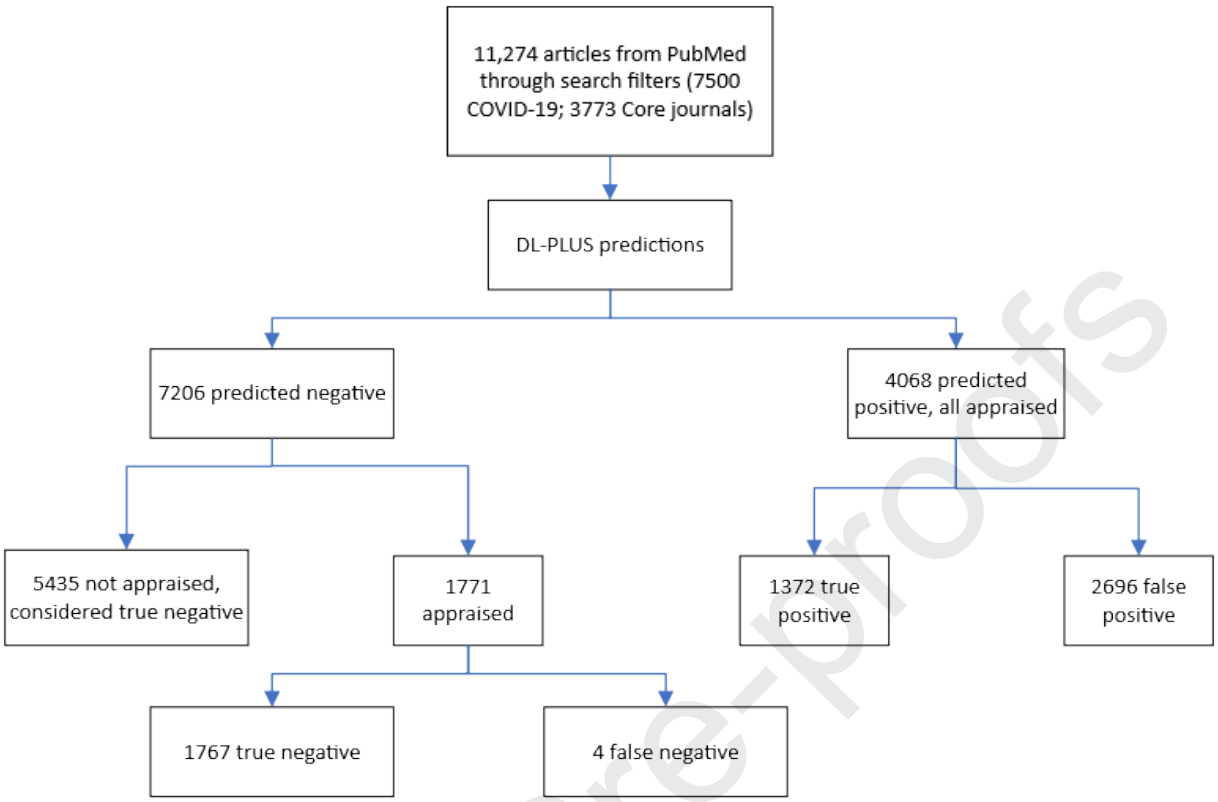


Figure 3. Real-world test of the DL-PLUS model predictions in active surveillance of the literature between March 12th, 2022 to Aug 12, 2022. Research associates blind to model predictions classified and appraised 5839 articles, including all predicted to be positive articles and 25% of those predicted to be negative.

4. Discussion

4.1 Findings

Our goal was to develop models, using modern deep learning techniques and PTLMs developed for the domain, to retrieve high quality studies across a range of clinical study questions, including diagnosis, prognosis, clinical prediction guides, etc. at the time of publication by using minimal features (e.g., title and abstract only). Deep learning approaches have been used to classify high quality clinical studies [13], [34], [47] , but only one has used a pretrained language model for training [34], and the focus was on identifying studies on

treatments only. Our model is the first we are aware of that identifies articles across the purpose categories trained using PTLMs.

A primary goal for literature surveillance is to quickly identify research that could impact care decisions. Maximizing specificity while maintaining high sensitivity to reduce the burden of manual critical appraisal stages while ensuring that all relevant articles are assessed improves the efficiency of the process. In the prospective evaluation of the model performance, specificity was 73% overall and the NNR, calculated as 1/precision, was 3.0 (95% CI, 2.8 - 3.1) meaning research associates would need to read 3 articles before they found one that met methodological criteria. This compares to NNRs of 4.6 (CI, 4.5 to 4.8) for 2019 when only Boolean search filters were in place, and 3.7 (CI, 3.5 to 3.8) using an earlier LightGBM model we developed in 2021 (Lokker, unpublished). WWS@ 99% also reflects the reduction in labour required for reviewing articles [46], with a 42% reduction for core journal articles and 73% for COVID-19 articles and 63% overall.

DL-PLUS has been fully implemented into the McMaster PLUS pipeline as a processing step between the Boolean search filters and entry into the CAP database (Figure 1) since Aug 13, 2022. By Mar 31, 2023, the model predicted that 6063 of 15,862 articles retrieved from PubMed were positive, a 62% reduction in articles requiring appraisal by research associates, which aligns with WSS@99% results from the prospective evaluation. This increase in efficiency, has saved an estimated >800 hours of staff time (assuming 5 minutes per article for assessment). Notably, for articles relating to COVID-19 retrieved from all journals indexed in PubMed, not just the PLUS core journals, the model has greater specificity, and higher NNR. The PubMed search strategy for these articles is more specific as it contains content words to exclude off target articles. The vast volume and variable quality of COVID-19 articles across all journals explains the lower precision, the greater NNR, and the higher WWS@99% of 73%. All retrieved articles, regardless of meeting quality criteria, are added to COVID-19 Evidence Alerts, a searchable web platform with optional free registration for email alerts to support access to the appraised research [48].

Given the longevity of the McMaster PLUS surveillance program, we were able to leverage a large, consistently produced dataset that is collated using accepted standards for quality assessment. The Clinical Hedges dataset is considered an established standard for search strategies and machine learning model development for retrieving high quality studies. It was produced in 2000 and is comprised of all items indexed in PubMed from 160 journals, including letters and editorials, etc. [8]. The other datasets used in this study include only articles that have been prefiltered using Boolean search strategies that purposely exclude some indexed items such as letters. The lower performance of DL-PLUS in the Hedges dataset could be explained by the noise added by these entries.

In this study, training was done with the unbalanced training set and smaller datasets to address the imbalance in the positive and negative classes and included evaluating the effect of ensembling models trained using the smaller sets. However, ensembling when tested in independent sets of data did not boost performance over the solo models as anticipated [45]. This lack of gain in performance could be due to each training dataset having the same positive class articles. Given the need for greater processing power for ongoing classification of articles using an ensemble model with no gain in performance, we focused further efforts on the solo DL-PLUS model.

4.2 Comparison with prior work

Machine learning applied to finding high quality clinical research is an active area of discovery, particularly recent training of deep learning models. Del Fiol et al. [13] used a noisy dataset of >400,000 PubMed articles to train a convolutional neural network to identify high quality articles on treatment. They compared the performance of their model to Clinical Queries search filters in the Hedges dataset and achieved 97% sensitivity, 35% precision and an F-measure of 0.51. Our model achieved similar performance in identifying articles across all purpose categories in our prospective test—not only treatment—a category of studies that have a relatively standard and consistent language in reporting. Ambalavanan and Devarakonda [34] used more elements of the Clinical Hedges dataset to develop ensemble models, cascade ensemble, and a single integrated model based on SciBERT. The elements included article format (original or review), of interest to human healthcare (yes/no), article purpose category,

and rigor (yes/no). They concluded that at a fixed recall, an individual task learner model outperformed the others, while suggesting that their cascade ensemble which had a higher F-measure (0.75) was more suitable for interactive searching. In a subset of treatment articles, at a fixed recall of 98.5%, their individual task learner model had 38.1% precision. Afzal, et al.[47] used articles identified from Cochrane reviews as methodologically rigorous and developed a deep learning model based on multi-layer perceptron and compared it with a few conventional machine learning models including Support Vector Machine and Gradient Boosted Tree. They achieved a higher performance using deep learning, with accuracy of 97.3%, 95.1% recall, and 86.2% precision.

Transformer-based deep learning, particularly BioBERT, is showing promise in addressing the challenge of identifying high-quality evidence in the vast pool of clinical articles being published. So far, much of the work is on studies of treatment, which support the majority of clinical questions posed by clinicians [49].

4.3 Limitations and future work

We demonstrated that fine-tuning PTLMs on our dataset is effective for identifying high quality evidence. We anticipate that further improvements in performance could be achieved by addressing the class imbalance and lack of articles for some of the categories, such as prognosis or diagnosis. In this work, we used an under-sampling technique to balance the dataset for training and ensembling to address class imbalance with an expectation of boosting performance. Oversampling has yet to be explored to balance the dataset in upward direction by enriching the rare class. Plans include exploring state-of-the-art data augmentation approaches to address the imbalance and to assess impact on performance. Additionally, the model is not designed for identifying articles within particular categories but rather serves to retrieve all high-quality studies across categories while limiting off-target articles. We have not assessed how the model would perform for uses other than a broad literature surveillance program or in PubMed without the Boolean search filters. We plan to train PTLMs for category specific models, especially for smaller categories such as prognosis and diagnosis, using the

data in our ML dataset. Thus far, we have focused on binary classification by quality criteria and have not leveraged the other features added by research associates.

5. Conclusions

We trained DL-PLUS using state-of-the-art PTLMs to identify high-quality, clinically relevant articles from PubMed at the time of publication using minimal features. The model maintains high recall and improves upon specificity compared with other approaches. It has been implemented into a real-time literature surveillance program, reducing the burden of manual critical appraisal by >60%; a significant savings of research staff time and improvement in efficiency to support quick dissemination.

Competing statement

McMaster University, a not-for-profit institution, has contracts, managed by the Health Information Research Unit, supervised by AI, RBH, and LL, with several professional and commercial publishers, to supply newly published studies and systematic reviews that are critically appraised for research methods and assessed for clinical relevance through the McMaster Premium Literature Service (McMaster PLUS). TN, RP, CC, and CL are partly paid through these contracts and RBH receives remuneration for supervisory time and royalties. EB, MA, WA, GF, and LC are not affiliated with McMaster PLUS.

Author contributions

Cynthia Lokker: Conceptualization, Methodology, Writing - Original Draft, Supervision **Elham Bagheri:** Validation, Writing - Review & Editing, Visualization **Rick Parrish:** Data curation, Software, Investigation, Formal analysis, Writing - Review & Editing **Muhammad Afzal:** Conceptualization, Methodology, Writing - Review & Editing **Wael Abdelkader:** Conceptualization, Writing - Review & Editing **Tamara Navarro:** Conceptualization, Writing - Review & Editing **Chris Cotoi:** Conceptualization, Project administration **Federico Germini:** Conceptualization, Writing - Review & Editing **Lori Linkins:** Conceptualization **R. Brian Haynes:** Conceptualization, Methodology, Writing - Review & Editing **Lingyang Chu:** Methodology,

Writing - Review & Editing **Alfonso Iorio:** Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision.

Role of funding source: This study received no external funding.

Data sharing: The PLUS database is not publicly available, but the authors would be glad to consider request for its use on a case-by-case basis. Please contact the corresponding author.

References

[1] S. Kamath and G. Guyatt, “Importance of evidence-based medicine on research and practice.,” *Indian J. Anaesth.*, vol. 60, no. 9, pp. 622–625, Sep. 2016, doi: 10.4103/0019-5049.190615.

[2] “MEDLINE PubMed Production Statistics,” 2021. https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html (accessed Aug. 06, 2021).

[3] R. B. Haynes, “Where’s the meat in clinical journals?,” *ACP J. Club*, vol. 119, no. 3, p. A22, Nov. 1993, doi: 10.7326/ACPJC-1993-119-3-A22.

[4] S. Zhao, C. Su, Z. Lu, and F. Wang, “Recent advances in biomedical literature mining,” *Briefings in Bioinformatics*, vol. 22, no. 3. Oxford University Press, May 01, 2021, doi: 10.1093/bib/bbaa057.

[5] N. L. Wilczynski, K. A. McKibbon, S. D. Walter, A. X. Garg, and R. B. Haynes, “MEDLINE clinical queries are robust when searching in recent publishing years,” *J. Am. Med. Informatics Assoc.*, vol. 20, no. 2, pp. 363–368, 2013, doi: 10.1136/AMIAJNL-2012-001075.

[6] “McMaster Health Knowledge Refinery - Hedges Project.” https://hiruweb.mcmaster.ca/hkr/hedges/ (accessed Aug. 06, 2021).

[7] L. M. Bachmann, R. Coray, P. Estermann, and G. Ter Rift, “Identifying diagnostic studies in MEDLINE: Reducing the number needed to read,” *J. Am. Med. Informatics Assoc.*, vol. 9, no. 6, pp. 653–658, Nov. 2002, doi: 10.1197/jamia.M1124.

[8] N. L. Wilczynski, D. Morgan, and R. B. Haynes, “An overview of the design and methods for retrieving high-quality studies for clinical care,” *BMC Med. Inform. Decis. Mak.*, vol. 5, no. 1, pp. 1–8, Jun. 2005, doi: 10.1186/1472-6947-5-20/FIGURES/1.

[9] A. N. Irwin and D. Rackham, “Comparison of the time-to-indexing in PubMed between biomedical journals according to impact factor, discipline, and focus,” *Res. Soc. Adm. Pharm.*, vol. 13, no. 2, pp. 389–393, Mar. 2017, doi: 10.1016/J.SAPHARM.2016.04.006.

[10] H. Kilicoglu, D. Demner-Fushman, T. C. Rindflesch, N. L. Wilczynski, and R. B. Haynes, “Towards automatic recognition of scientifically rigorous clinical research evidence.,” *J. Am. Med. Inform. Assoc.*, vol. 16, no. 1, pp. 25–31, Jan. 2009, doi: 10.1197/jamia.M2996.

[11] Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, and C. F. Aliferis, “Text categorization models for high-quality article retrieval in internal medicine,” *J. Am. Med.*

Informatics Assoc., vol. 12, no. 2, pp. 207–216, 2005, doi: 10.1197/jamia.M1641.

[12] E. V. Bernstam, J. R. Herskovic, Y. Aphinyanaphongs, C. F. Aliferis, M. G. Sriram, and W. R. Hersh, "Using citation data to improve retrieval from MEDLINE," *J. Am. Med. Informatics Assoc.*, vol. 13, no. 1, pp. 96–105, Jan. 2006, doi: 10.1197/jamia.M1909.

[13] G. Del Fiol *et al.*, "A Deep Learning Method to Automatically Identify Reports of Scientifically Rigorous Clinical Research from the Biomedical Literature: Comparative Analytic Study.," *J. Med. Internet Res.*, vol. 20, no. 6, p. e10281, Jun. 2018, doi: 10.2196/10281.

[14] I. J. Marshall, A. Noel-Storr, J. Kuiper, J. Thomas, and B. C. Wallace, "Machine learning for identifying Randomized Controlled Trials: An evaluation and practitioner's guide," *Res. Synth. Methods*, vol. 9, no. 4, pp. 602–614, Dec. 2018, doi: 10.1002/JRSM.1287.

[15] B. Wang, Q. Xie, J. Pei, Z. Li, P. Tiwari, and J. Fu, "Pre-trained Language Models in Biomedical Domain: A Systematic Survey," vol. 1, p. 46, 2021, doi: 10.1145/nnnnnnn.nnnnnnn.

[16] X. P. Qiu, T. X. Sun, Y. G. Xu, Y. F. Shao, N. Dai, and X. J. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10. Springer Verlag, pp. 1872–1897, Oct. 01, 2020, doi: 10.1007/s11431-020-1647-3.

[17] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," Jan. 2018, [Online]. Available: http://arxiv.org/abs/1801.06146.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Feb. 24, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805.

[19] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692.

[20] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.11942.

[21] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1903.10676.

[22] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/BIOINFORMATICS/BTZ682.

[23] T. Huang and J. Zhang, "BoostingBERT:Integrating Multi-Class Boosting into BERT for NLP Tasks," 2020.

[24] Q. Xie, J. A. Bishop, P. Tiwari, and S. Ananiadou, "Pre-trained language models with domain knowledge for biomedical extractive summarization," *Knowledge-Based Syst.*, vol. 252, p. 109460, Sep. 2022, Accessed: Oct. 31, 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705122007328.

[25] A. Adhikari, A. Ram, R. Tang, and J. Lin, "DocBERT: BERT for Document Classification," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.08398.

[26] Y. Si, J. Wang, H. Xu, and K. Roberts, “Enhancing clinical concept extraction with contextual embeddings,” *J. Am. Med. Informatics Assoc.*, vol. 26, no. 11, pp. 1297–1304, Jul. 2019, doi: 10.1093/jamia/ocz096.

[27] H. Guan and M. Devarakonda, “Leveraging Contextual Information in Extracting Long Distance Relations from Clinical Notes.,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2019, pp. 1051–1060, 2019, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/32308902.

[28] Y. Peng, Q. Chen, and Z. Lu, “An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining,” May 2020, [Online]. Available: http://arxiv.org/abs/2005.02799.

[29] Y. U. Gu *et al.*, “Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing,” *ACM Trans. Comput. Healthc.*, vol. 3, no. 1, pp. 1–23, Oct. 2021, doi: 10.1145/3458754.

[30] N. L. Wilczynski *et al.*, “Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: An analytic survey,” *BMC Med.*, vol. 2, Jun. 2004, doi: 10.1186/1741-7015-2-23.

[31] N. L. Wilczynski, K. A. McKibbon, and R. B. Haynes, “Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature.,” *Stud. Health Technol. Inform.*, vol. 84, no. Pt 1, pp. 390–3, 2001, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11604770.

[32] Y. Aphinyanaphongs and C. Aliferis, “Prospective validation of text categorization filters for identifying high-quality, content-specific articles in MEDLINE.,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, pp. 6–10, Jan. 2006, Accessed: Nov. 21, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/17238292/?tool=EBI.

[33] M. Afzal, M. Hussain, R. B. Haynes, and S. Lee, “Context-aware grading of quality evidences for evidence-based decision-making,” *Health Informatics J.*, vol. 25, no. 2, pp. 429–445, Jun. 2019, doi: 10.1177/1460458217719560.

[34] A. K. Ambalavanan and M. V. Devarakonda, “Using the contextual language model BERT for multi-criteria classification of scientific articles,” *J. Biomed. Inform.*, vol. 112, p. 103578, Dec. 2020, doi: 10.1016/j.jbi.2020.103578.

[35] G. J. Geersing, W. Bouwmeester, P. Zuithoff, R. Spijker, M. Leeflang, and K. Moons, “Search filters for finding prognostic and diagnostic prediction studies in medline to enhance systematic reviews,” *PLoS ONE*, vol. 7, no. 2. Feb. 29, 2012, doi: 10.1371/journal.pone.0032844.

[36] J. Holland and R. B. Haynes, “McMaster Premium Literature Service (PLUS): An Evidence-based MedicineInformation Service Delivered on the Web,” *AMIA Annu. Symp. Proc.*, vol. 2005, p. 340, 2005, Accessed: Dec. 15, 2021. [Online]. Available: /pmc/articles/PMC1560593/.

[37] “McMaster Health Knowledge Refinery -Our Process,” 2023. https://hiruweb.mcmaster.ca/hkr/what-we-do/.

[38] “McMaster Health Knowledge Refinery - McMaster PLUS Projects,” 2023. https://hiruweb.mcmaster.ca/hkr/what-we-do/plus-projects/.

[39] “McMaster Health Knowledge Refinery - Methodologic Criteria.” https://hiruweb.mcmaster.ca/hkr/what-we-do/methodologic-criteria/ (accessed Aug. 06, 2021).

[40] R. B. Haynes *et al.*, "McMaster PLUS: A Cluster Randomized Clinical Trial of an Intervention to Accelerate Clinical Use of Evidence-based Information from Digital Libraries," *J. Am. Med. Informatics Assoc.*, vol. 13, no. 6, pp. 593–600, Nov. 2006, doi: 10.1197/jamia.M2158.

[41] N. L. Wilczynski, C. J. Walker, K. A. McKibbon, and R. B. Haynes, "Assessment of methodologic search filters in MEDLINE.," *Proceedings. Symp. Comput. Appl. Med. Care*, pp. 601–5, 1993, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8130545.

[42] "BLURB Leaderboard." https://microsoft.github.io/BLURB/leaderboard.html (accessed Nov. 14, 2022).

[43] C. Lanera, P. Berchialla, A. Sharma, C. Minto, D. Gregori, and I. Baldi, "Screening PubMed abstracts: Is class imbalance always a challenge to machine learning?," *Syst. Rev.*, vol. 8, no. 1, pp. 1–9, Dec. 2019, doi: 10.1186/S13643-019-1245-8/TABLES/2.

[44] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization," Mar. 2016, [Online]. Available: http://arxiv.org/abs/1603.06560.

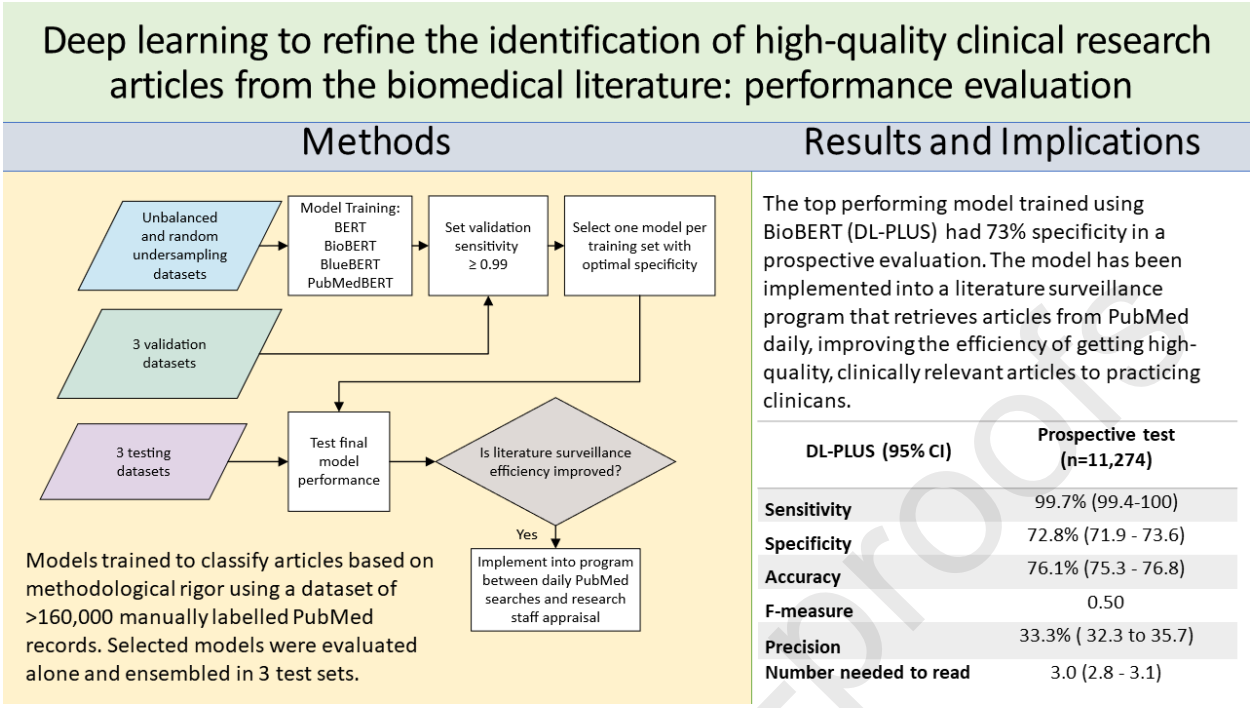
[45] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, p. 105151, Oct. 2022, doi: 10.1016/j.engappai.2022.105151.

[46] A. M. Cohen, W. R. Hersh, K. Peterson, P.-Y. Yen, and A. M. Cohen, "Reducing Workload in Systematic Review Preparation Using Automated Citation Classification," *J Am Med Inf. Assoc*, vol. 13, pp. 206–219, 2006, doi: 10.1197/jamia.M1929.

[47] M. Afzal, B. J. Park, M. Hussain, and S. Lee, "Deep learning based biomedical literature classification using criteria of scientific rigor," *Electron.*, vol. 9, no. 8, pp. 1–12, Aug. 2020, doi: 10.3390/ELECTRONICS9081253.

[48] H. I. R. U. McMaster, "COVID-19 Evidence Alerts from McMaster PLUS | Home," 2022. https://plus.mcmaster.ca/Covid-19/ (accessed Jun. 29, 2022).

[49] G. Del Fiol, T. E. Workman, and P. N. Gorman, "Clinical Questions Raised by Clinicians at the Point of Care: A Systematic Review," *JAMA Intern. Med.*, vol. 174, no. 5, pp. 710–718, May 2014, doi: 10.1001/JAMAINTERNMED.2014.368.



Appendix A

Critical Appraisal Process Inclusion Criteria

from https://hiruweb.mcmaster.ca/hkr/what-we-do/methodologic-criteria/

Basic criteria for original studies, systematic reviews, pooled original studies, and evidence-based guidelines:

- in English
- about humans
- about topics that are important to the clinical practice of medicine, nursing, rehabilitation, and other health professions, other than descriptive studies of prevalence
- analysis of each article consistent with the study question.

Studies of **prevention or treatment** must meet these additional criteria:

- random allocation of participants to comparison groups
- ≥ 10 patients per group (assessed for outcome)
- primary outcome(s) assessed in ≥ 80% of those randomized

- outcome measure of known or probable clinical importance
- subgroup analyses must be preplanned, with groups analyzed as they were randomized; analyses must test for interaction between ≥ 2 subgroups.

Studies of **diagnosis** must meet these additional criteria:

- inclusion of a spectrum of participants, all suspected of having the disease, with some, but not all, found to have the disease of interest after diagnostic testing
- inclusion of ≥ 100 participants, with ≥ 50 participants with the disease and ≥ 50 participants without the disease
- objective diagnostic ("gold") standard (e.g., laboratory test not requiring interpretation) OR current clinical standard for diagnosis (e.g., a venogram for deep venous thrombosis), preferably with documentation of reproducible criteria for subjectively interpreted diagnostic standard (i.e., report of statistically significant measure of agreement beyond chance among observers)
- each participant must receive both the new test and some form of the diagnostic standard
- interpretation of diagnostic standard without knowledge of test result
- interpretation of test without knowledge of diagnostic standard result
- diagnostic test characteristics reported.

Diagnostic tests may also be tested in randomized trials, in which case the criteria for prevention or treatment apply.

Studies of **prognosis** must meet these additional criteria:

- inception cohort of patients at a similar and early point in the course of a disease or condition, all initially free of the outcome of interest
- prospective standardized data collection
- ≥ 80% follow-up until the occurrence of a major study endpoint or to the end of the study.

Studies of **clinical prediction guides** must meet these additional criteria:

- purpose is to validate or compare a rule/index/scale/model that combines ≥ 2 factors into some type of score/ranking that assigns individual patients to different levels of risk for a specific outcome (diagnosis, prognosis, treatment responsiveness) based on the presence/absence of these factors
- data for the prediction guide must be available before data on the outcome that it is predicting
- the guide must be generated in one or more sets of real (not hypothetical) patients (derivation or development cohort)
- the guide must be validated in another set of real (not hypothetical) patients (validation cohort); internal bootstrapping is not acceptable as validation
    - studies validating a previously derived clinical prediction guide should explicitly state that the derivation was done in a separate patient cohort
    - prediction guides developed using individual patient data from > 1 study do not require separate validation
- study must provide information on how to apply the prediction guide in individual patients or cite a reference to this information.

Studies of **etiology of harm from medical interventions** must meet these additional criteria:

- explicit purpose is to assess adverse effects of an intervention
- prospective standardized data collection with clearly identified comparison groups for those at risk for the outcome of interest
- groups are matched or analyses adjusted to create comparable groups (e.g., quasi-randomized controlled trial, nonrandomized controlled trial, cohort study with case-by-case matching or statistical adjustment to create comparable groups, nested case–control study)
- blinding (masking) of observers of outcomes to exposures (criterion assumed to be met if outcome is objective, e.g., all-cause mortality or objective test)
- if harm reported, relative risk (RR) or hazard ratio (HR) or equivalent ≥ 2.0, with a lower 95% CI that excludes 1.5

- if no harm reported, upper 95% CI of RR or HR or equivalent excludes 1.5.

Randomized controlled trials assessing adverse effects are evaluated using criteria for studies of prevention or treatment.

Studies of **quality improvement or continuing education** must meet these additional criteria:

- random allocation of participants or units to comparison groups
- ≥ 10 patients per group (assessed for outcome)
- ≥ 1 specified outcome assessed in ≥ 80% of those randomized at ≥ 1 follow-up point
- outcome measure of known or probable clinical or educational importance
- subgroup analyses must be preplanned, with groups analyzed as they were randomized analysis must test for interaction between ≥ 2 subgroups.

Studies of the **economics** of health care programs or interventions must meet these additional criteria:

- alternate diagnostic or therapeutic services or quality improvement activities must be compared on the basis of both the outcomes produced (effectiveness) and resources consumed (costs) in real patients
- evidence of both effectiveness and costs reported in a single randomized controlled trial that passes criteria for prevention or treatment
- results must be presented in terms of the incremental or additional costs and outcomes of one intervention over another.

Systematic review articles must meet these additional criteria:

- explicit statement of the clinical topic
- identifiable description of the methods, including the databases searched and inclusion and exclusion criteria for selecting articles for detailed review; reviews of treatment, primary prevention, quality improvement, or economics must search for RCTs; reviews of prognosis must have "inception cohort" as an inclusion criterion
- > 1 major database searched

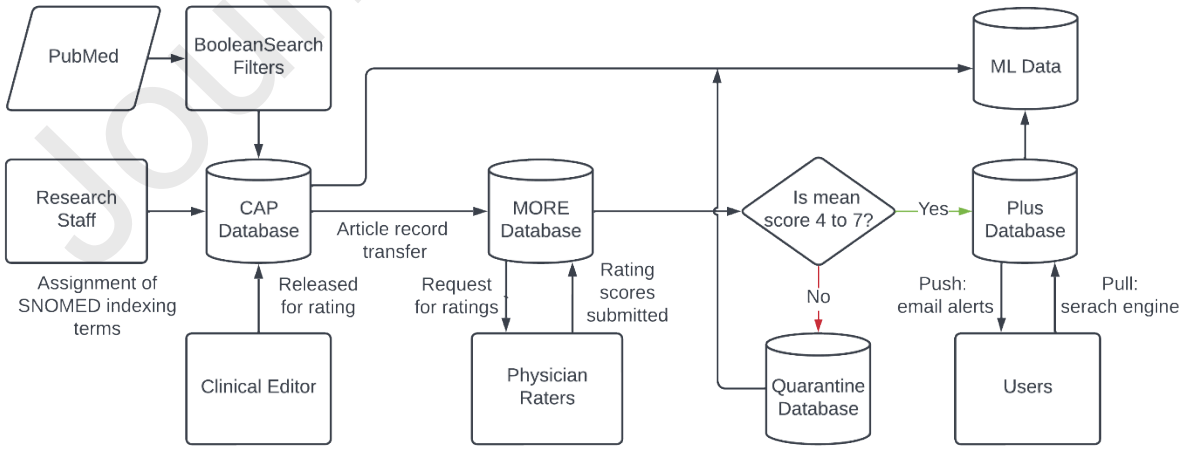
- number of articles retrieved/reviewed, and the number passed/included must be reported.

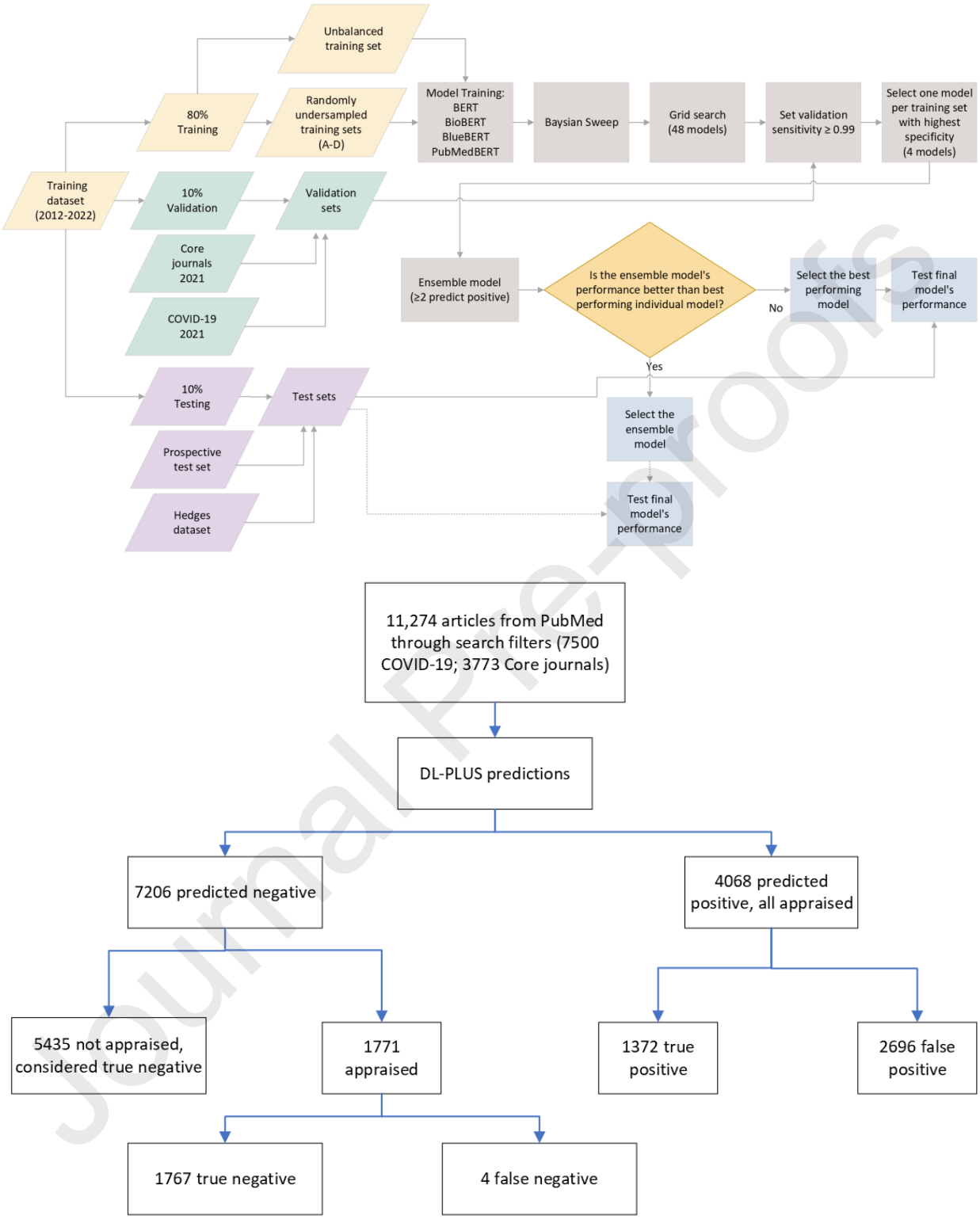
Pooled original studies must meet these additional criteria:

- analysis in which patient-level data are pooled from ≥ 2 studies/cohorts/sources to assess a question related to one of the study categories but article DOES NOT meet criteria for a systematic review

Evidence-based guidelines must meet these additional criteria:

- the Guideline must be based on a published systematic review that passes our current criteria for a Review
- methods and findings of the systematic review may be reported within the Guideline document or in a separate document that accompanies the Guideline or is cited in the Guideline and is accessible
- evidence underpinning the recommendations must be reported (e.g., citations of studies, estimates of effect, etc.)
- the strength of the evidence (such as GRADE) for the recommendations must be reported





Author contributions

Cynthia Lokker: Conceptualization, Methodology, Writing - Original Draft, Supervision **Elham Bagheri:** Validation, Writing - Review & Editing, Visualization **Rick Parrish:** Data curation,

Software, Investigation, Formal analysis, Writing - Review & Editing **Muhammad Afzal:** Conceptualization, Methodology, Writing - Review & Editing **Wael Abdelkader:** Conceptualization, Writing - Review & Editing **Tamara Navarro:** Conceptualization, Writing - Review & Editing **Chris Cotoi:** Conceptualization, Project administration **Federico Germini:** Conceptualization, Writing - Review & Editing **Lori Linkins:** Conceptualization **R. Brian Haynes:** Conceptualization, Methodology, Writing - Review & Editing **Lingyang Chu:** Methodology, Writing - Review & Editing **Alfonso Iorio:** Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision.

Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

McMaster University, a not-for-profit institution, has contracts, managed by the Health Information Research Unit, supervised by Alfonso Iorio, R Brian Haynes, and Lori-Ann Linkins, with several professional and commercial publishers, to supply newly published studies and systematic reviews that are critically appraised for research methods and assessed for clinical relevance through the McMaster Premium Literature Service (McMaster PLUS). Tamara Navarro, Rick Parrish, Chris Cotoi, and Cynthia Lokker are partly paid through these contracts and R Brian Haynes receives remuneration for supervisory time and royalties.

Problem or Issue Access to high-quality, clinically relevant research is hindered by the volume of published research, required critical appraisal skills, and available time.
What is Already Known Various informatics approaches have been applied to retrieve high quality evidence to make it timely and accessible to clinicians. Pretrained language models, such as BERT and its variants make it easier to tackle natural language processing tasks using deep learning techniques. Generally, models have been trained to identify high-quality articles focused on treatment studies.
What this Paper Adds Using a large database of clinical articles from 2012 to 2020 that were manually classified for methodological rigor, we fine-tuned BERT-variant deep learning

models to identify high-quality, clinically relevant evidence from the biomedical literature at the time of publication for use in a real-time literature surveillance program. We selected and tested a model trained on BioBERT that classifies, by quality, articles across a range of study purpose categories with >99% sensitivity and 73% specificity, that improves the efficiency of the literature surveillance process by >60%.