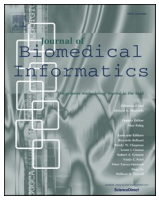
Journal of Biomedical Informatics 166 (2025) 104825

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Benchmarking domain-specific pretrained language models to identify the best model for methodological rigor in clinical studies

Fangwen Zhou [a], Rick Parrish [a], Muhammad Afzal [b], Ashirbani Saha [c], R. Brian Haynes [a], Alfonso Iorio [a,d], Cynthia Lokker [a,*]

[a] Health Information Research Unit, Department of Health Research Methods, Evidence, and Impact, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada
[b] Department of Computing, Faculty of Computing, Engineering and the Built Environment, Birmingham City University, Birmingham, United Kingdom
[c] Department of Oncology, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada
[d] Department of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada

ARTICLE INFO

Keywords:
Deep learning
Encoder-only transformer
Natural language processing
Text classification
Evidence-based medicine
Critical appraisal

ABSTRACT

Objective: Encoder-only transformer-based language models have shown promise in automating critical appraisal of clinical literature. However, a comprehensive evaluation of the models for classifying the methodological rigor of randomized controlled trials is necessary to identify the more robust ones. This study benchmarks several state-of-the-art transformer-based language models using a diverse set of performance metrics.
Methods: Seven transformer-based language models were fine-tuned on the title and abstract of 42,575 articles from 2003 to 2023 in McMaster University's Premium LiteratUre Service database under different configurations. The studies reported in the articles addressed questions related to treatment, prevention, or quality improvement for which randomized controlled trials are the gold standard with defined criteria for rigorous methods. Models were evaluated on the validation set using 12 schemes and metrics, including optimization for cross-entropy loss, Brier score, AUROC, average precision, sensitivity, specificity, and accuracy, among others. Threshold tuning was performed to optimize threshold-dependent metrics. Models that achieved the best performance in one or more schemes on the validation set were further tested in hold-out and external datasets.
Results: A total of 210 models were fine-tuned. Six models achieved top performance in one or more evaluation schemes. Three BioLinkBERT models outperformed others on 8 of the 12 schemes. BioBERT, BiomedBERT, and SciBERT were best on 1, 1 and 2 schemes, respectively. While model performance remained robust on the hold-out test set, it declined in external datasets. Class weight adjustments improved performance in most instances.
Conclusion: BioLinkBERT generally outperformed the other models. Using comprehensive evaluation metrics and threshold tuning optimizes model selection for real-world applications. Future work should assess generalizability to other datasets, explore alternate imbalance strategies, and examine training on full-text articles.

1. Introduction

Randomized controlled trials (RCTs) are the gold standard of primary research evidence for treatment [1]. Nevertheless, RCTs are subject to methodological and resource constraints which can re-introduce bias and undermine the validity of their results [2]. Critical appraisal is a process to evaluate the quality of clinical studies, ensuring that their findings can be interpreted appropriately in clinical practice and knowledge synthesis [3]. Tools for critical appraisal include Cochrane's revised Risk of Bias (ROB) tool [4], Joanna Briggs Institute's RCT checklist [5], and Health Information Research Unit's (HIRU) rigor criteria [6]. These tools are used for assessing limitations and sources of biases in studies, such as the risk of over optimistic results due to a lack of blinding and participant attrition over time.

Due to the increase in the number of published articles [7], manual critical appraisal processes are burdensome for systematic reviewers and knowledge translation professionals. Therefore, exploring automation methods has been a priority [8–10]. Supervised machine learning (ML) has demonstrated a strong potential in medical text classification [11–13]. A relevant example in evaluating RCTs include RobotReviewer

* Corresponding author.
E-mail address: lokkerc@mcmaster.ca (C. Lokker).

https://doi.org/10.1016/j.jbi.2025.104825
Received 1 December 2024; Received in revised form 2 March 2025; Accepted 3 April 2025
Available online 15 April 2025

[14], an automated tool for rating the risk of bias of RCTs. RobotReviewer leverages linear models and convolutional neural networks, and extensive evaluation has identified notable time-savings [15–17]. Other methodologies applying shallow learning (SL) [18–21] and deep learning (DL) models [22,23] for quality appraisal have also been examined.

Unfortunately, the current body of relevant ML literature is constrained by significant limitations. Many articles report the performance of only one model architecture and often do not explore the optimal hyperparameters [14,18,21,22]. Threshold tuning, the process of searching and choosing an optimal probability cutoff for classification, can significantly affect the performance of a binary classification model, especially when trained on an imbalanced dataset [24,25]. Studies mostly rely on discrimination metrics and seldom present calibration plots and metrics (where applicable), which could allow readers to assess the confidence of model predictions in relation to the ground truth [26].

Among the discrimination metrics, there is a noted reliance on a selected few, notably the F1 score [19–23,27] and the area under the receiver operating characteristic curve (AUROC) [18–20,27]. The limitations of these metrics have been extensively documented [28–32]. In short, precision, recall, and F1 scores do not provide insight into a model's ability to distinguish true negatives and are sensitive to class imbalance [32–34]. AUROC, and other threshold-independent metrics such as average precision (AP) and Brier score, represent an averaged performance across a range of decision thresholds and may not necessarily reflect real-world classification performance [29,31]. Depending on the goal(s) of the application of the model, the importance of false negatives and false positives can vary drastically. For example, a researcher looking for confident negative classifications would not benefit from reported AUROC or F1 scores. Thus, a set of different evaluation metrics should be incorporated for understanding a model's performance from various perspectives.

HIRU at McMaster University has been a pioneer in delivering relevant, rigorous evidence to clinicians around the world. In 2000, HIRU curated the Clinical Hedges dataset comprising 49,028 unique records of articles published in 161 journals indexed in MEDLINE. Subsequently, HIRU initiated the Premium LiteratUre Service (PLUS), where daily searches on PubMed retrieve newly published articles that are subsequently classified based on study design, clinical purpose, methodological rigor, and rated for clinical relevance and newsworthiness. Rigorous, relevant articles are sent to clinical practitioners and other partners [36]. Within PLUS, RCTs would be used to compare the effects of treatments, primary prevention strategies, or quality improvement interventions.

Current research at HIRU aims to automate the process of rigor classification by leveraging the PLUS database and domain-specific encoder-only transformer models. Bidirectional Encoder Representations from Transformers (BERT) variants [37–40], and BioELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [41], have demonstrated state-of-the-art performance in multiple clinical text classification tasks [42]. Previously, we fine-tuned 192 models using various hyperparameter combinations on the full PLUS database utilizing a high-sensitivity approach for confident negative classifications [43]. In this experiment, we focus on benchmarking appropriate model rigor classification in RCTs using a wide range of metrics and threshold tuning schemes. Our benchmarking process is uniquely characterized by a systematic approach for selecting models tailored to various preferences in model performance assessment. By leveraging relevant evaluation metrics, we aim to guide decision-making and establish robust standards for model application across diverse scenarios.

Statement of Significance

(*continued on next column*)

(*continued*)

Problem:	The exponential growth of clinical literature has made a manual critical appraisal of RCTs for methodological rigor time-consuming and labor-intensive.
What is already known:	Encoder-only transformer models, such as BioBERT, SciBERT, and BioLinkBERT, have shown promise in automating biomedical text classification, but evaluations often focus on limited metrics and datasets, with minimal exploration of threshold tuning and hyperparameter optimization.
What this paper adds:	This study comprehensively benchmarks 7 encoder-only transformer models for RCT rigor classification, identifies BioLinkBERT as the top performer across most evaluation schemes, and highlights the value of using diverse metrics and threshold tuning for real-world use cases.

2. Methods

2.1. Database description

The database for this project is compiled through PLUS. It includes articles identified in PubMed that were published from 2003 onwards across approximately 120 high-quality clinical journals [44] or related to COVID-19 across all journals (2020–2023), and appraised following the methodological and quality criteria used to create the Clinical Hedges database [35]. Haynes et al. described this process of RCTs selection and appraisal based on the rigor criteria [45]. The classification process had an inter-rater agreement of Cohen's $\kappa > 0.80$ for all steps, including article type, purpose, overall methodological rigor, and newsworthiness [46].

Briefly, the process of creating the PLUS database includes automated daily searches of PubMed using a highly sensitive methods filter adapted from Clinical Queries and COVID-19 topic-specific search terms. The filters are designed to retrieve articles related to human health that are potentially ready for clinical practice. The articles are manually classified by expert research associates into one of four mutually exclusive article types: 1) original study, 2) review, 3) evidence-based guideline, or 4) other. Articles are then tagged with one or more of the following purpose categories: 1) treatment, 2) prevention, 3) quality improvement, 4) diagnosis, 5) prognosis, 6) etiology, 7) economics, 8) clinical prediction guide (subclassified for diagnosis or prognosis), or 9) other.

Methodological rigor is appraised using the full text of the articles. All articles must meet the following criteria: 1) in English, 2) about humans, 3) about topics that are important to the clinical practice of medicine, nursing, rehabilitation, and other health professions, other than descriptive studies of prevalence, and 4) analysis is consistent with the study question. Type and purpose-specific criteria are also applied. For original articles with a purpose of treatment, prevention, or quality improvement to be rated as rigorous, they must be RCTs that have 1) random allocation of participants to study arms, 2) $\geq$ 10 patients/participants per group completing the primary outcome assessment, 3) primary outcome(s) assessed in $\geq$80% of those randomized at the defined follow-up point, 4) at least one clinically important outcome measure, and 5) if reporting subgroup analysis, it is preplanned, with groups analyzed as they were randomized and interaction between two or more subgroups reported. Articles that fail to meet any one of these criteria are deemed to be not rigorous.

2.2. Dataset construction

The datasets used in this project focus on RCTs and include all original studies with the purpose of treatment, prevention, or quality improvement from the database [47]. To tune hyperparameters during fine-tuning and internally validate model performance, articles from 2003 to 2023 inclusive were randomly split into the train (80%), validation (10%), and hold-out test (10%) sets. Articles processed through

PLUS published in 2024, as well as articles of the same categories in the Clinical Hedges database, were used for testing [35]. Thus, in total, three test sets were used, one internal (the hold-out test set) and two external (PLUS articles in 2024 and Clinical Hedges).

2.3. Data preprocessing

The title and abstract of articles were used as inputs. The input sequences were tokenized with a maximum sequence length of 512 tokens. Shorter inputs were padded with semantically meaningless tokens to ensure uniform input sizes across all samples. Tokens exceeding the maximum length were truncated.

2.4. Model configuration during fine-tuning

A linear learning rate scheduler and the AdamW optimizer [48] with β1 set to 0.9 and β2 set to 0.999 were utilized. Gradient accumulation between batches was employed when memory was insufficient. Mixed precision training was enabled (fp16=True) to optimize computational efficiency. A warmup ratio of 0.1 was used to gradually increase the learning rate at the beginning of the fine-tuning process. Weight decay was applied with a factor of 0.01 to prevent overfitting. All layers of the model, including the embedding, encoder, pooler, and classifier layers, were fine-tuned to minimize cross-entropy loss.

Each model and hyperparameter combination was fine-tuned for up to 10 epochs. An early stopping patience of 3 was selected, in which the fine-tuning process prematurely terminated if the model failed to decrease validation cross-entropy loss for three consecutive epochs. The weights from the epoch with the lowest cross-entropy loss on the validation set were chosen.

2.4.1. Pretrained model selection

Seven domain-specific pretrained models, including BioBERT [37], BioELECTRA [41], BioLinkBERT [38], BiomedBERT (abstracts only) [39], BiomedBERT (abstracts + full text) [39], SciBERT-cased [40], and SciBERT-uncased [40], were fine-tuned. These models were chosen based on their performance in previous literature [37–41] and on the Biomedical Language Understanding and Reasoning Benchmark leaderboard [42].

We focused exclusively on encoder-only architectures as they have demonstrated efficiency and ease of use in text classification tasks, particularly in biomedical domains [42]. Compared with shallow learning architectures, such as Naïve Bayes and support vector machines, transformers preclude the need for meticulous feature engineering [49–51]. Decoder-only and encoder-decoder transformers, including generative large language models, require significant prompt engineering and computational demands, making them less pragmatic for text classification over a large number of articles [52–54]. Furthermore, while these models excel at text generation, they often suffer from inconsistent performance in classification tasks, especially when only the abstract of articles is available as input [55–57].

2.4.2. Hyperparameter optimization

A grid search of 3 learning rates (LR) (1e-5, 3e-5, 5e-5) and 5 batch sizes (BS) (16, 32, 64, 128, 256) was performed. While it is suggested that learning rate selection can be influenced by batch size due to their interplay in gradient updates for Adam-style optimizers, their relationship is often not linear [58,59]. we opted to tune them independently to ensure a comprehensive search for optimal configurations. This decision was made to better capture model behaviors across a wide range of settings and avoid potential biases introduced by predefining relationships between these parameters.

2.5. Addressing class imbalance

To account for class imbalances in the train set, an alternate model with each hyperparameter combination was also fine-tuned with class weights (CW) using the following formula:

$$weight_i = \frac{N}{2n_i}$$

where N is the total number of samples and n_i is the number of samples in class i. This ensures that the minority class would receive a proportionally higher weight to compensate during the training process. This weighting strategy was applied directly to the loss function rather than through data resampling techniques such as undersampling or oversampling.

2.6. Evaluation metrics

The final dense layer was transformed with Softmax to convert logits to probabilities. Model performance was evaluated based on 4 threshold-independent metrics: cross-entropy loss, Brier score, AUROC, AP; as well as 11 threshold-dependent metrics: sensitivity/recall, specificity, accuracy, precision/positive predictive value (PPV), negative predictive value (NPV), number needed to read (NNR), F1 score, F2 score, Matthew's correlation coefficient (MCC), work saved over sampling (WSoS), and Youden's Index. The definitions and interpretations of these metrics are in Appendix **Table A1**.

2.7. Model performance evaluation

2.7.1. Evaluation

Twelve evaluation schemes were chosen to provide a comprehensive assessment of model performance across various real-world use cases (Table 1). The average model performances by pretrained model and class weight adjustments on the validation set based on 12 evaluation schemes were assembled. The model performance across all hyperparameter combinations and seeds was presented using mean and 95% confidence interval (CI). Subsequently, the models that achieved the best performance on one or more schemes were selected for further evaluation on the three test sets. For these models, the 95% CI was estimated using bootstrapping over 1,000 iterations.

2.7.2. Threshold tuning

To dichotomize predicted probabilities into rigor classifications, the maximum and minimum probability thresholds that would achieve 99% sensitivity and specificity were selected for the target sensitivity and specificity schemes. For threshold-dependent metrics, the probability threshold that would achieve the best score, corresponding to the metric, was selected. The threshold was selected using the validation set and was kept consistent during testing.

2.8. Post-hoc sensitivity analysis

The Clinical Hedges test set has a different class balance than the other evaluation subsets. This difference could be due to changes to the article retrieval process and minor amendments to the rigor assessment criteria in the intervening years. Specifically, PLUS filters PubMed content using Clinical Queries Boolean filters, significantly reducing the number of non-rigorous articles, while Clinical Hedges includes cover-to-cover content from 160 journals published in 2000. For rating criteria, articles from the Clinical Hedges were not required to have ≥10 patients per group and a clinically important outcome to be considered rigorous. Overall, the impact of the filtering outweighs the looser rigor criteria, resulting in significantly more rigorous articles by proportion in PLUS. To evaluate model performance independent of class balance differences, we performed undersampling to ensure that the class balance of Clinical Hedges was identical to the train set. Specifically, instances from the non-rigor class in the Clinical Hedges test set were randomly removed.

Table 1
Model evaluation schemes.

Evaluation Scheme	Threshold Tuning	Purpose
Threshold-independent		
Lowest cross-entropy loss	N/A	To improve the overall certainty of classifications by minimizing the difference in bits between predicted probability and true probability (i.e., actual class labels 0/1 in a binary classification) distributions.
Lowest Brier score	N/A	To improve calibration performance by minimizing the mean-squared difference between predicted probabilities and actual class labels (i.e., 0/1 in a binary classification).
Highest AUROC	N/A	To enhance the overall ability to distinguish between positive and negative classes, particularly useful in imbalanced datasets
Highest AP	N/A	To reduce overall false positives and maximize precision, particularly when high recall is critical in imbalanced datasets.
Target sensitivity or specificity		
Highest specificity at 99% sensitivity	Minimum probability that achieves 99% sensitivity	To ensure near-perfect identification of positive instances to improve the performance and confidence in identifying negative instances (minimizing false positives).
Highest sensitivity at 99% specificity	Maximum probability that achieves 99% specificity	To ensure near-perfect identification of negative instances to improve the performance and confidence in identifying positive instances (minimizing false negatives).
Threshold-dependent		
Highest accuracy	Maximizes accuracy	To balance true positives and true negatives, assuming class distribution is balanced. Best when both false positives and false negatives are equally important.
Highest F1	Maximizes F1	To optimize the balance between precision and recall, especially useful when there is a class imbalance, and false positives and false negatives have similar costs. It ignores true negatives.
Highest F2	Maximizes F2	To prioritize recall over precision, making it more suitable when minimizing false negatives is more critical than minimizing false positives. It ignores the true negatives.
Highest MCC	Maximizes MCC	To optimize overall performance by balancing both types of errors (false positives and false negatives), particularly effective in more imbalanced datasets. It incorporates true positives, false positives, true negatives, and false negatives in computation.
Highest WSoS	Maximizes WSoS	To maximize efficiency in screening tasks by minimizing the number of false positives and saving effort in reviewing non-relevant (i.e., non-rigorous in our study) items. Note that WSoS is dataset-specific and equivalent to specificity when normalized [60].
Highest Youden's Index	Maximizes Youden's Index	To achieve an optimal balance between sensitivity and specificity, especially when false negatives and false positives are equally important. Mostly agnostic to class balance like ROC AUC.

AP Average precision; **AUROC** Area under the receiver operating characteristic curve; **MCC** Matthews correlation coefficient; **ROC** Receiver operating characteristic; **WSoS** Work saved over sampling.

2.9. Stochastic stability analysis

To investigate model stochastic stability, all model configurations were trained over two additional random seeds (2, 3). The difference between the best- and worst-performing seeds for all model configurations was presented and analyzed.

2.10. Hardware and software

All fine-tuning was conducted using the resources from the Cedar cluster of the Digital Research Alliance of Canada. Each model was trained using one NVIDIA V100 Volta (32G HBM2 memory), as well as an allocation of 8 cores and 40 GB of memory.

Visual Studio Code and Python 3.11.5 were used for all software development. We utilized the transformers library by Hugging Face [61] to obtain pretrained models, and torch was used for evaluation purposes. Data management and statistical analysis were conducted using Pandas, NumPy, and scikit-learn. Data visualization was done with matplotlib. The full software environment can be found in **Table A2**.

3. Results

3.1. Characteristics of datasets

The five data subsets include a total of 60,802 articles from 2003 to 2024 and from 2000 only for the Clinical Hedges test set (see Table 2); overall, 37,488 (61.66%) of the included articles were appraised as methodologically rigorous. In each data subset, the proportion of articles meeting rigor assessment ranged from 24.1% to 60.2%. The undersampled Clinical Hedges Test set for sensitivity analysis had 2,643 articles.

3.2. Average performance on the validation set

Each of the 7 encoder models was trained with 15 possible hyperparameter combinations and with CW and without CW. In total, 210 models were trained and evaluated. Table 3 details the average performance of the models by pretrained model and CW on the validation set using the default threshold of $\geq$0.50. The average performance of the models after tuning their thresholds can be found in Appendix **Tables A3 to A11**.

Using the default threshold of $\geq$ 0.50, BioLinkBERT demonstrated the best average performance across 9 and 12 of the 15 metrics among models with and without class weight adjustments, respectively. After tuning for thresholds, BioLinkBERT achieved the best performance on all schemes, except for the highest sensitivity at 99% specificity which SciBERT-cased obtained the best performance. In general, models with CW adjustments outperformed those without.

3.3. Best models on the internal sets

Of the 210 fine-tuned models, 6 unique models and hyperparameter combinations achieved the best validation performance across one or more of the 12 evaluation schemes, in which 4 models used CW adjustments (Table 4). The performance of the 6 models across all evaluation metrics in the PLUS-2003–2023 validation set and test set is in **Table A12** and **A13**, respectively. In general, the models had good performance on their corresponding metrics, and the performance on the validation and test sets was similar. Depending on the optimization scheme, performance on other metrics varied.

The ROC curve and precision-recall curve of the model that achieved the best performance on the four threshold-independent metrics

Table 2
The characteristics of included datasets.

Dataset Name	Publication Year of the Articles	Purpose (Proportion)	No. of Articles	Rigorous Articles (%)	Non-Rigorous Articles (%)
PLUS-2003–2023	2003 to 2023	Train (80%)	42,575	25,561 (60.0)	17,014 (40.0)
	2003 to 2023	Validation (10%)	5,322	3,203 (60.2)	2,119 (39.8)
	2003 to 2023	Hold-out Test (10%)	5,322	3,164 (59.5)	2,158 (40.5)
PLUS-2024	2024	External Test	1,011	575 (56.9)	436 (43.1)
Clinical Hedges	2000	External Test	6,572	1,587 (24.1)	4,985 (75.9)
Clinical Hedges – Undersampled	2000	External Test	2,643	1,587 (60.0)	1,056 (40.0)

(BioLinkBERT [CW: Yes; LR: 1e-05; BS: 32]) are in Fig. 1. The model's performance on threshold-dependent metrics against the probability threshold and the calibration curve can be found in Fig. 2 and Fig. 3, respectively. The ROC curve and precision-recall curve of the 5 other models can be found in **Figs. A1** to **A5**. The threshold-dependent metrics plot can be found in **Figs. A6** to **A10**. The calibration plots can be found in **Figs. A11** to **A15**. The models were generally well-calibrated on the validation and tests.

3.3.1. Threshold-independent metrics

BioLinkBERT (CW: Yes; LR: 1e-05; BS: 32) achieved the best cross entropy loss of 0.294 (0.275, 0.311), Brier score of 0.089 (0.083, 0.094), AUROC of 0.952 (0.945, 0.959), and AP of 0.942 (0.935, 0.949). There was a nonsignificant improvement across all four threshold-independent metrics on the hold-out test set.

3.3.2. 99% Target sensitivity or specificity

BioBERT (CW: No; LR: 1e-5; BS: 16) achieved the best specificity of 0.628 (0.606, 0.650) at 99% sensitivity. SciBERT-cased (CW: Yes; LR: 1e-5; BS: 16) achieved the best sensitivity of 0.335 (0.319, 0.350) at 99% specificity. The specificity and sensitivity decreased by a nonsignificant amount on the hold-out test set.

3.3.3. Threshold-dependent metric

BioLinkBERT (CW: Yes; LR: 1e-5, BS: 256) achieved the best accuracy of 0.882 (0.874, 0.891), F1 of 0.906 (0.899, 0.914), and MCC of 0.753 (0.734, 0.771). BioLinkBERT (CW: No; LR: 1e-5; BS: 32) performed the best on F2 with a value of 0.946 (0.941, 0.95). BiomedBERT (abstracts only) (CW: Yes; LR: 1e-5; BS: 32) achieved the best WSoS of 0.298 (0.286, 0.310), and Youden's index of 0.749 (0.730, 0.769). The models achieved similar performance on all threshold-dependent metrics on the hold-out test set with a maximum absolute difference of 0.011 on Youden's index.

3.4. Performance on the external sets

The 6 models, using the same threshold where applicable, were applied to predict outcomes for the PLUS-2024 and the Clinical Hedges set (Table 4). The performance across all evaluation metrics is in **Table A14** and **A15**. Overall, model performance in the PLUS-2024 set showed a significant decline across all metrics. Performance on the Clinical Hedges set was mixed, in which metrics prioritizing true negatives (specificity, NPV, and WSoS) improved while the others further declined.

The models were, in general, well calibrated on the PLUS-2024 set. However, the models severely overestimated the predicted probability of rigor across the ranges of probability scores on the Clinical Hedges set (Fig. 3, **Figs. A10** to **A15**).

3.4.1. Threshold-independent metrics

Compared with the internal validation set, the performance across all four metrics significantly declined in the PLUS-2024 set and further declined in the Clinical Hedges set. Notably, there was an approximate 0.091 increase in Brier score and a 0.192 decrease in AP on the Clinical Hedges set.

3.4.2. 99% Target sensitivity or specificity

For the PLUS-2024 set, the specificity and sensitivity decreased by 0.119 and 0.225, respectively, resulting in the worst sensitivity at 99% specificity across all four evaluation sets. On the Clinical Hedges set, the models saw a decrease in specificity by 0.204 and an increase in sensitivity by 0.128, resulting in the best sensitivity at 99% specificity of 0.463 (0.438, 0.487).

3.4.3. Threshold-dependent metrics

Significant decreases in performance were evident for the PLUS-2024 set. Notably, MCC and Youden's index decreased by 0.131 and 0.150, respectively. For the Clinical Hedges set, F1, F2, and MCC decreased by over 0.15, while accuracy and Youden's index decreased by 0.12 and 0.082, respectively. The best WSoS of 0.506 (0.488, 0.525) was achieved, a 0.208 increase over the validation set.

3.5. Post-hoc sensitivity analysis

Undersampling of the Clinical Hedges dataset created a subset of 1587 (60%) rigorous articles and 1056 (40%) non-rigorous articles (Table 2). The performance of the 6 models is in Table 4 and **Table A16**. Compared with the original Clinical Hedges dataset, there was significant improvement in cross-entropy loss, Brier score, AP, accuracy, MCC, F1, and F2. As expected, AUROC, sensitivity, specificity, and Youden's Index were similar as they are independent of class prevalence. WSoS decreased as it is dependent on the proportion of negative instances. The performance on the Clinical Hedges – undersampled set was similar to the PLUS-2024 set. The models were better calibrated on the undersampled dataset compared with the original (Fig. 3, **Figs. A10** to **A15**).

3.6. Stochastic stability analysis

The stability of the fine-tuned models was analyzed by additionally fine-tuning each configuration over two additional seeds and assessing the range between the best and worst performing seeds for each model. **Table A17** to **A21** tabulate the mean range and 95% CI for all metrics and the optimal decision thresholds.

For both internal datasets, the models demonstrated strong stability. On the validation and hold-out test sets, the mean cross-entropy loss range was 0.061 and 0.057, respectively. The differences in the optimal decision threshold across different seeds were substantive for most tuning schemes. Notably, for accuracy, F1, MCC, WSoS, and Youden's index, the mean range of optimal thresholds were $\geq$0.239.

For the external datasets, the stability was generally worse compared to the internal datasets. The mean cross-entropy loss range was 0.090, 0.384, and 0.202 for the PLUS 2024, Clinical Hedges, and Clinical Hedges – Undersampled sets, respectively.

4. Discussion

This experiment is a comprehensive evaluation of encoder-only transformers on classifying RCTs per rigor, utilizing a large dataset, a leaderboard top set of models and hyperparameter combinations, threshold tuning, and multiple evaluation schemes. Our previous publication included articles addressing any purpose category and article

Table 3
Average performance of model configurations on the PLUS-2003–2023 validation dataset using a threshold of ≥ 0.50.

Pre-trained Model	Cross-entropy Loss	Brier Score	ROC AUC	AP	Sensitivity/ Recall	Specificity	Accuracy	Precision/ PPV	NPV	NNR	F1 Score	F2 Score	MCC	WSoS	Youden's Index
With Class Weights															
BioBERT	0.334 (0.28, 0.387)	0.097 (0.091, 0.102)	0.935 [missing word] (0.932, 0.938)	0.945 (0.942, 0.948)	0.93 (0.923, 0.938)	0.784 (0.772, 0.796)	0.872 (0.87, 0.874)	0.867 (0.861, 0.873)	0.882 (0.872, 0.892)	1.153 (1.146, 1.161)	0.897 (0.896, 0.899)	0.917 (0.912, 0.921)	0.731 (0.727, 0.736)	0.284 (0.282, 0.287)	0.714 (0.708, 0.721)
BioELECTRA	0.338 (0.321, 0.354)	0.099 (0.096, 0.101)	0.934 (0.931, 0.936)	0.943 (0.941, 0.946)	**0.940 (0.931, 0.949)**	0.764 (0.747, 0.781)	0.870 (0.868, 0.872)	0.858 (0.851, 0.866)	**0.895 (0.882, 0.907)**	1.166 (1.155, 1.176)	0.897 (0.896, 0.898)	**0.922 (0.917, 0.927)**	0.728 (0.725, 0.731)	0.280 (0.277, 0.284)	0.704 (0.696, 0.712)
BioLinkBERT	**0.314 (0.306, 0.323)**	**0.093 (0.092, 0.095)**	**0.937 (0.935, 0.939)**	**0.948 (0.945, 0.95)**	0.924 (0.919, 0.93)	0.802 (0.793, 0.811)	**0.876 (0.874, 0.878)**	0.876 (0.871, 0.881)	0.876 (0.868, 0.883)	1.142 (1.136, 1.148)	**0.899 (0.898, 0.901)**	0.914 (0.91, 0.918)	**0.739 (0.734, 0.743)**	**0.289 (0.287, 0.291)**	**0.726 (0.721, 0.732)**
BiomedBERT (abstracts only)	0.328 (0.314, 0.342)	0.097 (0.094, 0.1)	0.934 (0.932, 0.937)	0.945 (0.942, 0.948)	0.917 (0.906, 0.928)	0.803 (0.788, 0.818)	0.872 (0.869, 0.875)	0.876 (0.869, 0.883)	0.867 (0.853, 0.88)	1.142 (1.133, 1.151)	0.896 (0.893, 0.899)	0.909 (0.901, 0.916)	0.731 (0.725, 0.737)	0.287 (0.284, 0.289)	0.72 (0.713, 0.727)
BiomedBERT (abstracts + full text)	0.323 (0.313, 0.334)	0.096 (0.094, 0.098)	0.936 (0.934, 0.938)	0.946 (0.945, 0.948)	0.916 (0.909, 0.922)	**0.809 (0.799, 0.82)**	0.873 (0.872, 0.875)	**0.879 (0.874, 0.884)**	0.865 (0.857, 0.872)	**1.138 (1.131, 1.144)**	0.897 (0.896, 0.898)	0.908 (0.904, 0.912)	0.734 (0.731, 0.738)	**0.289 (0.287, 0.291)**	0.725 (0.72, 0.73)
SciBERT-cased	0.327 (0.318, 0.336)	0.099 (0.097, 0.101)	0.931 (0.929, 0.933)	0.944 (0.942, 0.946)	0.931 (0.926, 0.936)	0.772 (0.762, 0.782)	0.868 (0.866, 0.869)	0.861 (0.856, 0.866)	0.881 (0.874, 0.888)	1.162 (1.155, 1.168)	0.894 (0.893, 0.896)	0.916 (0.913, 0.919)	0.722 (0.718, 0.726)	0.280 (0.277, 0.282)	0.703 (0.697, 0.709)
SciBERT-uncased	0.349 (0.331, 0.367)	0.103 (0.1, 0.106)	0.927 (0.925, 0.93)	0.938 (0.936, 0.941)	0.925 (0.914, 0.936)	0.772 (0.753, 0.79)	0.864 (0.862, 0.866)	0.86 (0.852, 0.869)	0.874 (0.861, 0.887)	1.163 (1.151, 1.174)	0.891 (0.889, 0.893)	0.911 (0.905, 0.918)	0.715 (0.711, 0.719)	0.277 (0.274, 0.281)	0.697 (0.688, 0.706)
Without Class Weights															
BioBERT	0.334 (0.32, 0.349)	0.098 (0.096, 0.1)	0.934 (0.932, 0.936)	0.944 (0.943, 0.945)	0.936 (0.93, 0.942)	0.773 (0.762, 0.784)	0.871 (0.869, 0.873)	0.862 (0.857, 0.867)	0.889 (0.881, 0.897)	1.161 (1.154, 1.168)	0.897 (0.896, 0.899)	0.920 (0.916, 0.924)	0.730 (0.725, 0.734)	0.282 (0.28, 0.285)	0.709 (0.702, 0.715)
BioELECTRA	0.344 (0.332, 0.356)	0.101 (0.099, 0.103)	0.931 (0.929, 0.934)	0.942 (0.938, 0.946)	0.938 (0.932, 0.945)	0.763 (0.752, 0.774)	0.869 (0.867, 0.87)	0.857 (0.852, 0.862)	0.892 (0.883, 0.901)	1.167 (1.16, 1.174)	0.896 (0.895, 0.897)	0.921 (0.917, 0.925)	0.725 (0.721, 0.728)	0.279 (0.277, 0.282)	0.701 (0.696, 0.707)
BioLinkBERT	**0.330 (0.310, 0.349)**	**0.096 (0.093, 0.099)**	**0.937 (0.935, 0.939)**	**0.947 (0.945, 0.95)**	**0.938 (0.929, 0.946)**	0.781 (0.768, 0.794)	**0.875 (0.873, 0.877)**	0.866 (0.86, 0.872)	**0.893 (0.882, 0.904)**	1.154 (1.146, 1.162)	**0.900 (0.899, 0.902)**	**0.922 (0.917, 0.928)**	**0.739 (0.735, 0.743)**	**0.286 (0.284, 0.288)**	**0.719 (0.713, 0.725)**
BiomedBERT (abstracts only)	**0.330 (0.317, 0.344)**	0.097 (0.094, 0.101)	0.934 (0.932, 0.937)	0.945 (0.942, 0.947)	0.931 (0.921, 0.942)	**0.782 (0.766, 0.799)**	0.872 (0.869, 0.875)	0.867 (0.859, 0.874)	0.884 (0.871, 0.898)	1.154 (1.144, 1.164)	0.898 (0.895, 0.9)	0.917 (0.911, 0.924)	0.732 (0.726, 0.738)	0.284 (0.281, 0.288)	0.714 (0.705, 0.722)
BiomedBERT (abstracts + full text)	**0.330 (0.316, 0.345)**	**0.096 (0.094, 0.099)**	0.935 (0.933, 0.938)	0.945 (0.942, 0.948)	0.929 (0.923, 0.934)	0.79 (0.782, 0.799)	0.874 (0.872, 0.875)	**0.870 (0.866, 0.874)**	0.880 (0.873, 0.887)	**1.149 (1.144, 1.155)**	0.898 (0.897, 0.900)	0.916 (0.913, 0.920)	0.735 (0.731, 0.738)	**0.286 (0.284, 0.288)**	0.719 (0.714, 0.724)
SciBERT-cased	0.338 (0.324, 0.353)	0.101 (0.098, 0.104)	0.931 (0.928, 0.934)	0.944 (0.941, 0.946)	0.934 (0.926, 0.941)	0.766 (0.755, 0.777)	0.867 (0.865, 0.869)	0.858 (0.853, 0.863)	0.885 (0.875, 0.895)	1.166 (1.159, 1.172)	0.894 (0.892, 0.896)	0.918 (0.913, 0.922)	0.721 (0.716, 0.726)	0.279 (0.276, 0.281)	0.700 (0.694, 0.706)
SciBERT-uncased	0.350 (0.334, 0.365)	0.103 (0.101, 0.105)	0.927 (0.925, 0.93)	0.939 (0.936, 0.942)	0.936 (0.928, 0.944)	0.754 (0.741, 0.768)	0.864 (0.862, 0.866)	0.852 (0.846, 0.858)	0.887 (0.876, 0.898)	1.173 (1.165, 1.182)	0.892 (0.89, 0.894)	0.918 (0.913, 0.923)	0.714 (0.71, 0.719)	0.275 (0.272, 0.278)	0.690 (0.683, 0.697)

Note: All values are presented as the mean (95% CI) of the 15 hyperparameter combinations. Bolded values indicate best performance.

AP Average precision; **AUROC** Area under the receiver operating characteristic curve; **BERT** Bidirectional Encoder Representations from Transformers; **ELECTRA** Efficiently Learning an Encoder that Classifies Token Replacements Accurately; **ROC** Receiver operating characteristic.

Table 4
Best model performances according to the 12 evaluation schemes.

Scheme	Best Performing Model (CW, LR, BS)	Threshold	Scores Corresponding to the Metric Used in the Scheme				
			PLUS-2003–2023 Validation Set	**PLUS-2003–2023 Hold-out Test Set**	**PLUS-2024**	**Clinical Hedges**	**Clinical Hedges —Undersampled**
Threshold-independent							
Lowest cross-entropy loss	BioLinkBERT (Yes, 1e-05, 32)	–	0.294 (0.275, 0.311)	0.284 (0.266, 0.303)	0.422 (0.379, 0.469)	0.585 (0.561, 0.61)	0.472 (0.437, 0.506)
Lowest Brier score		–	0.089 (0.083, 0.095)	0.087 (0.082, 0.092)	0.134 (0.119, 0.149)	0.180 (0.173, 0.187)	0.147 (0.137, 0.158)
Highest AUROC		–	0.942 (0.935, 0.948)	0.947 (0.941, 0.953)	0.889 (0.869, 0.906)	0.882 (0.871, 0.891)	0.883 (0.869, 0.895)
Highest AP		–	0.952 (0.944, 0.959)	0.959 (0.953, 0.964)	0.899 (0.873, 0.922)	0.760 (0.738, 0.779)	0.919 (0.907, 0.931)
Target sensitivity or specificity							
Highest specificity at 99% sensitivity	BioBERT (No, 1e-05, 16)	0.063	0.628 (0.607, 0.648)	0.644 (0.624, 0.665)	0.509 (0.462, 0.558)	0.440 (0.427, 0.453)	0.438 (0.407, 0.47)
Highest sensitivity at 99% specificity	SciBERT-cased (Yes, 1e-05, 16)	0.983	0.335 (0.319, 0.352)	0.326 (0.311, 0.343)	0.080 (0.058, 0.102)	0.463 (0.438, 0.487)	0.463 (0.439, 0.487)
Threshold-dependent							
Highest accuracy	BioLinkBERT (Yes, 1e-05, 256)	0.425	0.882 (0.874, 0.891)	0.877 (0.867, 0.886)	0.815 (0.791, 0.839)	0.762 (0.752, 0.772)	0.809 (0.794, 0.824)
Highest F1		0.422	0.906 (0.899, 0.914)	0.901 (0.894, 0.909)	0.849 (0.827, 0.869)	0.636 (0.619, 0.652)	0.844 (0.830, 0.858)
Highest MCC		0.422	0.753 (0.734, 0.771)	0.745 (0.726, 0.762)	0.622 (0.573, 0.671)	0.514 (0.493, 0.533)	0.600 (0.568, 0.630)
Highest F2	BioLinkBERT (No, 1e-05, 32)	0.172	0.946 (0.941, 0.95)	0.944 (0.940, 0.949)	0.917 (0.902, 0.931)	0.731 (0.717, 0.743)	0.891 (0.880, 0.902)
Highest WSoS	BiomedBERT (abstracts only) (Yes, 1e-05, 32)	0.707	0.298 (0.286, 0.310)	0.299 (0.288, 0.312)	0.258 (0.232, 0.288)	0.506 (0.488, 0.525)	0.274 (0.258, 0.29)
Highest Youden's Index		0.707	0.749 (0.730, 0.769)	0.738 (0.721, 0.756)	0.599 (0.550, 0.646)	0.667 (0.645, 0.689)	0.685 (0.659, 0.713)

Note: Values are presented as score (95% CI from bootstrapping over 1,000 iterations).
AP Average precision; **AUROC** Area under the receiver operating characteristic curve; **BERT** Bidirectional Encoder Representations from Transformers; **BS** Batch size; **CW** Class weights; **LR** Learning rate; **MCC** Matthews correlation coefficient; **NNR** Number needed to read; **NPV** Negative predictive value; **PPV** Positive predictive value; **ROC** Receiver operating characteristic; **WSoS** Work saved over sampling.

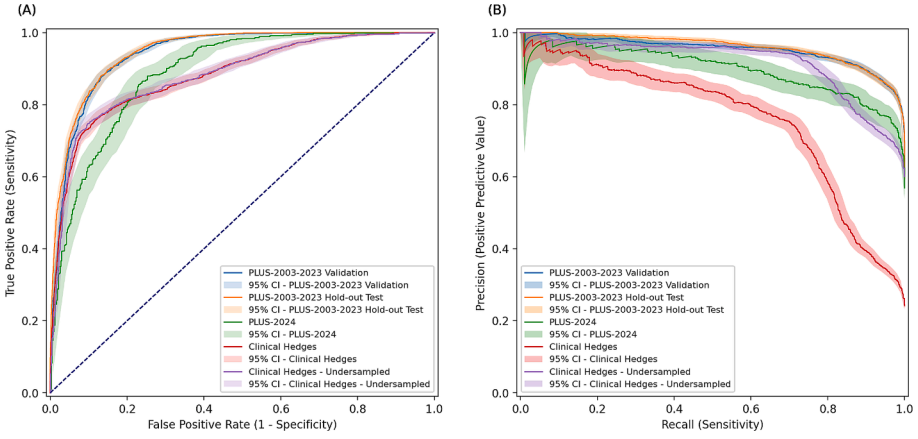


Fig. 1. Receiver operating characteristic curve and precision recall curve for BioLinkBERT (CW: Yes; LR: 1e-05; BS: 32). **A:** Receiver operating characteristic curve; **B:** Precision recall curve.

type, focusing on a high-sensitivity approach with a single best-performing model [43], while this experiment discovered models for a range of use cases where identifying high-quality RCTs is critical. This study serves as a benchmark for classifying methodological rigor in RCTs using transformer-based models and as a framework for selecting models based on the specific needs of real-world applications. By incorporating

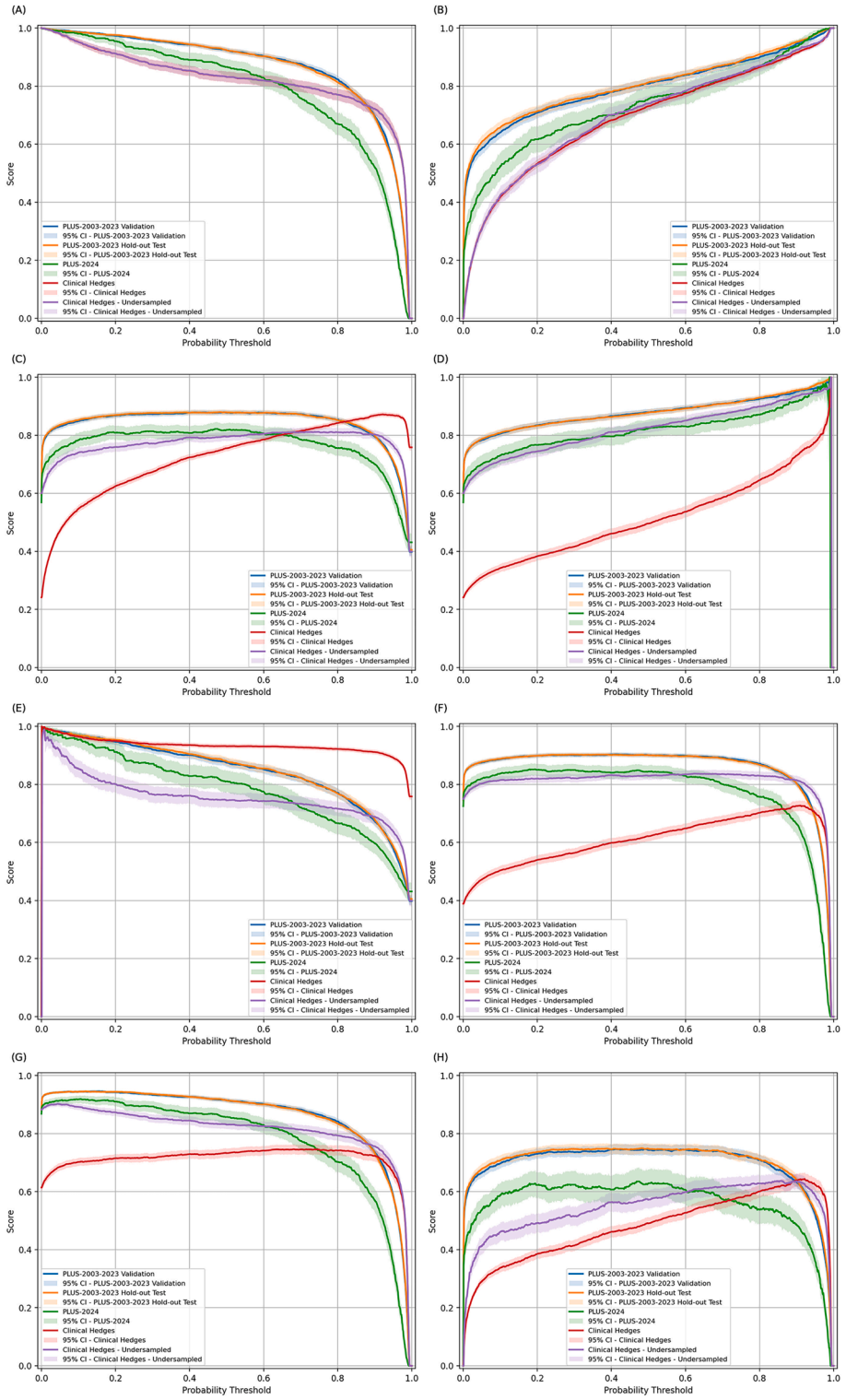


Fig. 2. Threshold-dependent metrics versus probability threshold plot for BioLinkBERT (CW: Yes; LR: 1e-05; BS: 32). **A:** Sensitivity/recall; **B:** Specificity; **C:** Accuracy; **D:** PPV/precision; **E:** NPV; **F:** F1 score; **G:** F2 score; **H:** MCC; **I:** Youden's Index; **J:** WSoS.

diverse evaluation metrics and threshold-tuning approaches, we provide a structured methodology for optimizing classification performance depending on the user priorities. Additionally, we perform stochastic stability analysis and sensitivity analyses to assess model robustness and generalizability over time. These insights address critical gaps in previous research and establish a more robust foundation for automated classification in evidence-based medicine workflows.

4.1. Model performance

On average, all chosen models displayed satisfactory performance on the validation set after fine-tuning. BioLinkBERT outperformed other

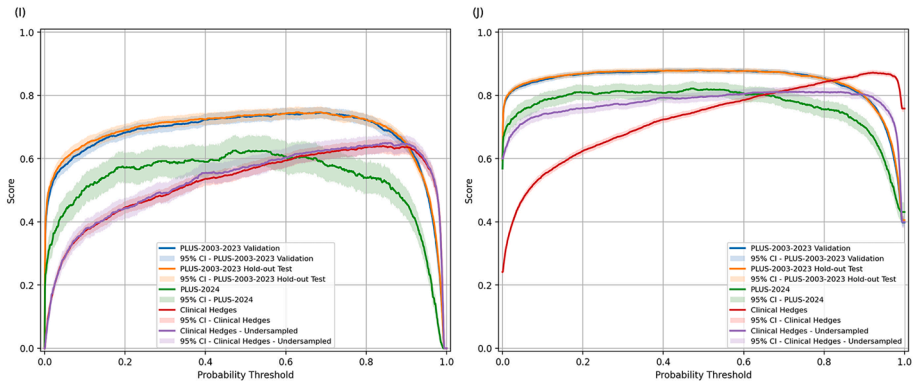


Fig. 2. (*continued*).

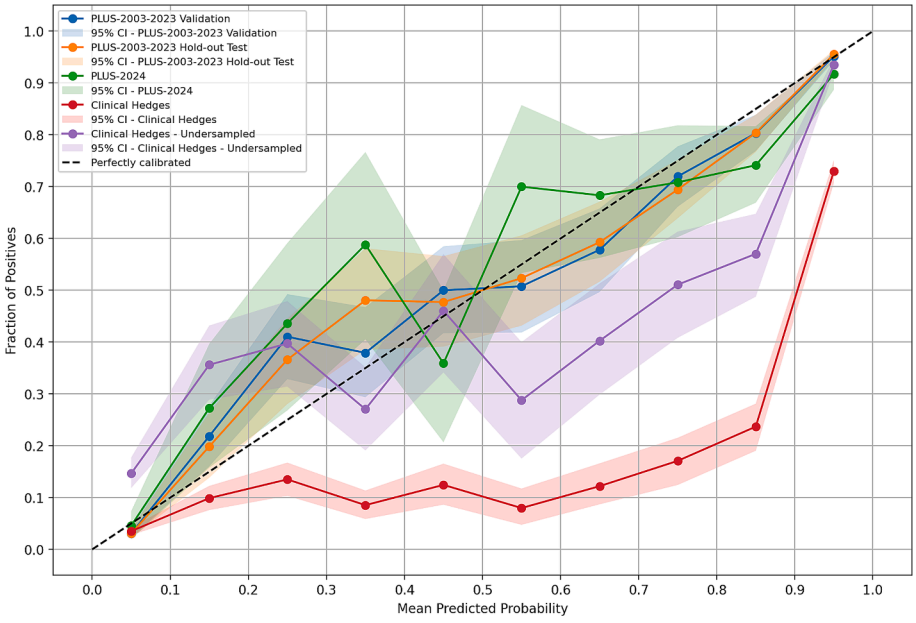


Fig. 3. Calibration plot for BioLinkBERT (CW: Yes; LR: 1e-05; BS: 32).

models in most instances, while SciBERT had a worse performance for this task in general. The inclusion of class weight adjustments marginally improved performance in most instances. Of the 210 models, the 6 best-performing models achieved respectable performance on the validation set in the context of clinical literature classification. Three of the 6 achieved the best performance on multiple evaluation schemes. Regarding global, threshold-independent performances, our top models achieved a Brier score of <0.1, and $\sim$ 95% AUROC and AP. We explored numerous schemes with threshold tuning and obtained the best possible performance. Notably, 64% specificity was achieved when sensitivity was set at 99%, and F1 and F2 scores >0.90 and MCC >0.75 when thresholds were tuned to maximize these metrics.

While we saw no notable degradation on the hold-out test set, we did on the external datasets. Sensitivity analysis revealed that the models performed similarly on the undersampled Clinical Hedges set compared with the PLUS-2024 set. While the models overestimated on the original Clinical Hedges set, undersampling brought calibration in line. These results demonstrate that the models had good generalizability to Clinical Hedges despite slight changes in the rigor evaluation criteria over time. However, the modest decrease in performance in the PLUS-2024 and undersampled Clinical Hedges set could not be explained by class prevalence. We suspect these differences may be attributable to systematic changes in clinical literature over time, including text structure and complexity [62–64]. Experiments incorporating time series data during training or analyzing whether the exclusion of older articles from training would lead to better predictive performance on recent studies may be warranted. Additionally, it may be justified to continuously fine-tune deployed models to mitigate performance degradation over time [64].

4.2. Comparison with existing literature

Our previous experiment [43] utilized the entire PLUS and Clinical Hedges datasets which include a broader array of study types. On the validation sets, the best model from the previous experiment achieved 97% AUROC and 70% specificity at 99% sensitivity, which are 3% and 7% higher, respectively. In general, deep learning models tend to perform better when the dataset is more homogeneous, anticipating that the models in this experiment should be equal to or better than the previous [65]. However, due to the inclusion of other types of articles, the comparison is not straightforward. We suspect that the higher performance in the previous experiment could be from including other types of articles, and that the model may perform worse on RCTs but significantly better on the other types of articles. On the Clinical Hedges dataset, the sensitivity of the model from the previous experiment dropped to 84% using the same threshold, and specificity was 81%. In

contrast, the model and threshold in this experiment were more resilient in maintaining a high sensitivity (96%), albeit achieving lower specificity (44%).

Del Fiol et al. [22] tested neural networks on treatment articles from the Clinical Hedges dataset and achieved a precision, recall and F1 score of 35%, 97% and 0.51, respectively. Our best-F1 model achieved an F1 of 0.64, albeit with a lower recall (86%). Our best specificity at 99% sensitivity model achieved similar recall and precision. Ambalavanan and Devarakonda [66] used SciBERT as a cascade learner to classify all steps of Clinical Hedges, including article format, interest to human healthcare, article purpose, and methodological rigor. The model achieved an F1 score of 0.778 at 91% sensitivity. Devarakonda's model achieved a precision of 38% at 99% recall when applied to the dataset used by Del Fiol et al. [22]. Aphinyanaphongs et al. [20] examined shallow learning models to classify internal medicine articles. They retrieved MEDLINE articles and considered all articles from the ACP Journal Club as high quality. The polynomial support vector machine achieved an average cross-validation AUROC of 0.97 on treatment articles, which is slightly higher than our best models. However, the authors utilized extensive feature engineering, including the original journal and MeSH terms as inputs plus encoded texts, and optimized specifically for AUROC. These limitations undermine the goal of automation and may lead to over-optimistic AUROC. Nevertheless, this indicates that sophisticated shallow learning methods remain relevant in text classification tasks.

To our knowledge, there are no other experiments that used articles labelled using HIRU's criteria. Two studies used ML to classify RCTs based on the Cochrane ROB [14,18]. Marshall et al. [14] presented evaluations on RobotReviewer and achieved a document-level accuracy of 71%. However, the lack of reporting on class prevalence precludes meaningful comparison. Millard et al. [18] evaluated logistic regression models to classify RCTs based on the random sequence generation, allocation concealment, and blinding domains of ROB. The highest average AUROC across 10-fold cross-validation was 78% on allocation concealment. Wang et al. [23] evaluated deep learning models for risk of bias assessment in preclinical literature. Depending on the domain of the tool, the F1 scores ranged from 0.82 to 0.92 on the held-out test set. However, the difference in class prevalence, where positive instances made up 12% to 78% of the dataset, precludes any meaningful comparison as well. Afzal et al. [67] developed multiple models to identify kidney disease studies cited in Cochrane reviews, which they labelled as being scientifically rigorous. On the test set, a maximum AUROC of 91% was achieved by the multilayer perceptron model.

4.3. Proxying rigor criteria from abstracts and tokenization concerns

One concern of this study is that classification labels are derived from full-text appraisals, yet the models are trained on abstracts due to input token limitations for encoder-only transformers. Rigor criteria for human assessments require details not included in abstracts (e.g., attrition and preplanned subgroup analyses) so the model cannot directly assess these [68,69]. Additionally, criteria-level labels were not available, and the model was fine-tuned using the final rating of rigor. The classifier likely relies on indirect indicators or patterns detected during fine-tuning that would proxy the final rigor evaluation, which may not correspond to the specific criteria applied to full text or external datasets. Considering the strong performance in the internal datasets, it is likely that the abstracts contain sufficient information for effective proxying. However, performance degradation on the two external test sets suggests that generalizability remains a concern.

BERT and ELECTRA models have a maximum token limit of 512 (~400 words). This is unlikely to be an issue as most titles and abstracts fall under this limit [69,70]. Models like Longformer [71] or BigBird [72], with a higher limit of 4096 tokens, utilize a sliding window to reduce the complexity associated with self-attention [73]. A study examining clinical Longformer and BigBird for clinical document classification found that they offered limited benefits for shorter inputs, but may uplift performance for longer sequences [74]. Future studies working with longer inputs could explore these architectures.

Another concern is the limited ability of transformer models to process numerical information due to the tokenization of numbers [75], which could lead to difficulties in evaluating and generalizing to criteria associated with sample sizes. While this is unlikely to be an issue in this study as RCTs with <10 participants per group are rare [76,77], it could be an issue for a different criterion (e.g., attrition) or study design.

4.4. Threshold tuning

Threshold tuning is an important process to improve classification performance [78–80]. Our experiment demonstrated that there is relatedness in the optimal thresholds that would maximize a certain metric. Accuracy, F1 and MCC were maximized using similar thresholds, and WSoS and Youden's Index were maximized using the same threshold. However, depending on the prioritized metric, optimal thresholds can differ significantly from each other and from the typical arbitrary threshold of ≥0.50, especially for imbalanced datasets. For instance, a lower threshold improved metrics that prioritize sensitivity, such as F2. Importantly, threshold tuning, similar to hyperparameter tuning, must be conducted on the validation set and not the training or independent test sets to prevent overfitting [25]. As setting the threshold to prioritize one metric may affect the performance of another, researchers should carefully consider and justify their tuning process in the context of their objectives.

4.5. Evaluation metric selection

As previously mentioned, a limitation of many studies is the reliance on F1 scores and AUROC [18–23,27]. We mitigated this issue by providing 15 evaluation metrics, 12 optimization schemes, and the best-performing models. These schemes cover a range of use cases. For instance, those looking to obtain high-quality articles while minimizing effort should focus on metrics that optimize false positives, such as F1 [28], and setting a high specificity. A researcher who wishes to exclude as many low-quality articles as possible without misclassifying rigorous ones may find WSoS [81] and setting a high sensitivity beneficial [43]. Sensitivity and specificity, unlike PPV, NPV, WSoS, among others, are independent of class prevalence [82]. Therefore, AUROC [83] and Youden's Index [84] provide a better assessment of a model's capabilities independent of class prevalence on the testing data, and metrics that are sensitive to class prevalence should not be used to compare models across datasets unless normalized by class distribution [60]. For ranking tasks, models that have a low cross-entropy loss or Brier score would likely produce the most accurate predictions [85,86]. While in our experiment, the models generally performed well on related metrics, such as F1, accuracy, and MCC, this may not be the case on another dataset. As we demonstrate, a model that performed well on one metric may suffer in another. We argue that it is important to comprehensively evaluate model performance and present multiple models, bearing in mind the most common use cases, to ensure that model selection and deployment align with the articulated goal.

4.6. Effect of class balance during training

The most common methods to improve performance are resampling and using class weight adjustments [87]. We used class weights as the training set was relatively balanced, and there is no consensus on the superiority of any one method. This contrasts with our previous experiment where we undersampled the majority class as only 20% of articles were rigorous [43]. In this study, 7 of the 8 best-performing models utilized class weight adjustments, indicating that even slight imbalances may adversely affect performance despite the validation set having a similar balance as the training set. Interestingly, there was no notable

degradation in the predictive performance on the independent and prospective testing sets and the Clinical Hedges set compared to the hold-out test set in our previous experiment [43]. However, a direct comparison is not feasible due to the differences in the included articles. Comparing resampling versus class weight adjustments and the effect of class balance of the training and evaluation sets on model performance are areas for continued exploration.

4.7. Stochastic stability

The stochastic nature of neural networks, including random weight initialization, may impact model performance after training, and the general consensus for random seeds is for reproducibility. However, how random seeds should be utilized in the training and evaluation process is still debated [88–90]. A seed may consistently lead to better results on a particular dataset due to the prioritization of important features over noise in the training data by chance, yet selecting a specific seed may result in overfitting and poorer performance on an external dataset [90,91].

We analyzed the stochastic stability of all models by fine-tuning over two additional seeds and analyzing the differences between the best- and worst-performing seeds across all models, metrics, and threshold tuning schemes. In general, the models tended to be more stable on datasets that performed well. For datasets with worse performance, including Clinical Hedges, the stability deteriorated as well. The optimal threshold during tuning differed substantially across seeds, indicating that the best threshold for a model trained on one seed may not necessarily perform well on another seed. Based on these results, it was unlikely that averaging performance scores over multiple seeds as opposed to one would have led to significant differences in model selection during hyperparameter tuning. Nevertheless, model performances on external datasets could differ significantly based on random seeds [90–92].

4.8. Limitations

Limitations include that the models are specific to HIRU's rigor evaluation criteria; generalizability to other tools has not yet been examined. Cross-validation and nested cross-validation could not be conducted for hyperparameter selection due to computational and storage limitations. Nevertheless, the large number of included articles should mitigate the risk of sampling bias [93–95]. Some hyperparameters, such as the optimizer, remained fixed due to infeasible computational costs associated with a larger search grid. The choice of certain hyperparameters, especially the optimizer, may alter the effect of others [58,96,97]. Using less demanding hyperparameter tuning methodology, such as Bayesian optimization and Hyperband [98], may be warranted in future research. Lastly, we did not examine other methods to address class imbalance [79].

5. Conclusion

In this study, we evaluated the performance of 7 domain-specific, encoder-only transformer models for the classification of methodological rigor in RCTs. By fine-tuning 210 models with comprehensive hyperparameter tuning and applying a range of evaluation metrics, we demonstrated that BioLinkBERT with specific hyperparameters consistently outperformed other models across most evaluation schemes. Our findings highlight the importance of using a broad set of metrics, including threshold-independent and dependent approaches, to optimize model performance based on specific use cases. Future work should explore model generalization to newer articles, the impact of class balance on training and testing sets, and the performance benefits of training on full text data.

Role of funding sources

FZ was funded by the Mitacs Business Strategy Internship grant (IT42947), Mitacs, Canada, with matching funds from EBSCO Canada. The funders were not involved in any aspect of the project.

CRediT authorship contribution statement

Fangwen Zhou: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Rick Parrish:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Muhammad Afzal:** Writing – review & editing, Validation, Methodology, Conceptualization. **Ashirbani Saha:** Writing – review & editing, Validation, Methodology. **R. Brian Haynes:** Writing – review & editing, Visualization, Conceptualization. **Alfonso Iorio:** Writing – review & editing, Validation, Resources, Investigation, Conceptualization. **Cynthia Lokker:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: McMaster University, a non-profit public academic institution, operates contracts through the Health Information Research Unit under the supervision of AI and RBH. These contracts involve professional and commercial publishers to provide newly published studies and systematic reviews, which are critically appraised for research methodology and assessed for clinical relevance as part of the McMaster Premium LiteratUre Service (McMaster PLUS). CL and RP receive partial compensation through these contracts, and RBH is remunerated for supervisory responsibilities and royalties. AS, FZ, and MA have no affiliations with McMaster PLUS.

Acknowledgements

We gratefully acknowledge the computational support provided by the Digital Research Alliance of Canada.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2025.104825.

References

[1] P.B. Burns, R.J. Rohrich, K.C. Chung, The levels of evidence and their role in evidence-based medicine, Plast. Reconstr. Surg. 128 (1) (2011 Jul) 305–310.
[2] E. Hariton, J.J. Locascio, Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials, BJOG 125 (13) (2018 Dec) 1716.
[3] A. Al-Jundi, S. Sakka, Critical appraisal of clinical research, J. Clin. Diagn. Res. 11 (5) (2017 May). JE01–5.
[4] J.A.C. Sterne, J. Savović, M.J. Page, R.G. Elbers, N.S. Blencowe, I. Boutron, et al., RoB 2: a revised tool for assessing risk of bias in randomised trials, BMJ 28 (366) (2019 Aug) l4898.
[5] T.H. Barker, J.C. Stone, K. Sears, M. Klugar, C. Tufanaru, J. Leonardi-Bee, et al., The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials, JBI Evid. Synth. 21 (3) (2023 Mar 1) 494–506.
[6] [cited 2024 Aug 19]. Available from: https://hiruweb.mcmaster.ca/hkr/what-we-do/methodologic-criteria/.
[7] Number of clinical trials by year, country, WHO region and income group (1999-2022) [Internet]. [cited 2024 Sep 16]. Available from: https://www.who.int/observatories/global-observatory-on-health-research-and-development/monitoring/number-of-clinical-trials-by-year-country-who-region-and-income-group.
[8] K. Kolaski, L.R. Logan, J.P.A. Ioannidis, Guidance to best tools and practices for systematic reviews, Jun;11(6) Available from: JBJS Rev. [Internet] (2023) https://doi.org/10.2106/jbjs.rvw.23.00077.
[9] B. Tóth, L. Berek, L. Gulácsi, M. Péntek, Z. Zrubka, Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in PubMed, Syst. Rev. [Internet] 13 (1) (2024), https://doi.org/10.1186/s13643-024-02592-3. Jul 8, Available from:.

[10] K. Kolaski, L.R. Logan, J.P.A. Ioannidis, Guidance to best tools and practices for systematic reviews, JBJS Rev. [Internet] 11 (6) (2023), https://doi.org/10.2106/JBJS.RVW.23.00077. Jun 1, Available from:.
[11] I. Spasic, G. Nenadic, Clinical text data in machine learning: Systematic review, JMIR Med. Inform. 8 (3) (2020) e17984. Mar 31.
[12] [cited 2024 Sep 16]. Available from: https://ebooks.iospress.nl/publication/46339.
[13] S.K. Prabhakar, D.O. Won, Medical text classification using hybrid deep learning models with multihead attention, Comput. Intell. Neurosci. [Internet]. 1 (2021), https://doi.org/10.1155/2021/9425655, 2021 Jan, Available from:.
[14] I.J. Marshall, J. Kuiper, B.C. Wallace, RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials, J. Am. Med. Inform. Assoc. 23 (1) (2016 Jan) 193–201.
[15] J. Clark, P. Glasziou, C. Del Mar, A. Bannach-Brown, P. Stehlik, A.M. Scott, A full systematic review was completed in 2 weeks using automation tools: a case study, J. Clin. Epidemiol. 121 (2020 May) 81–90.
[16] J. Clark, C. McFarlane, G. Cleo, C. Ishikawa Ramos, S. Marshall, The impact of systematic review automation tools on methodological quality and time taken to complete systematic review tasks: Case study, May 31, JMIR Med. Educ. 7 (2) (2021) e24418.
[17] F. Soboczenski, T.A. Trikalinos, J. Kuiper, R.G. Bias, B.C. Wallace, I.J. Marshall, Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study, BMC Med. Inform. Decis. Mak. [Internet]. 19 (1) (2019), https://doi.org/10.1186/s12911-019-0814-z. Available from:.
[18] L.A.C. Millard, P.A. Flach, J.P.T. Higgins, Machine learning to assist risk-of-bias assessments in systematic reviews, Int. J. Epidemiol. 2016-2;45(1):266–77.
[19] H. Kilicoglu, D. Demner-Fushman, T.C. Rindflesch, N.L. Wilczynski, R.B. Haynes, Towards automatic recognition of scientifically rigorous clinical research evidence, J. Am. Med. Inform. Assoc. 16 (1) (2009 Jan) 25–31.
[20] Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, C.F. Aliferis, Text categorization models for high-quality article retrieval in internal medicine, J. Am. Med. Inform. Assoc. 12 (2) (2005 Mar) 207–216.
[21] I.J. Marshall, J. Kuiper, B.C. Wallace, Automating risk of bias assessment for clinical trials, IEEE J. Biomed. Health Inform. 19 (4) (2015 Jul) 1406–1412.
[22] G. Del Fiol, M. Michelson, A. Iorio, C. Cotoi, R.B. Haynes, A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: comparative analytic study, J. Med. Internet Res. 20 (6) (2018 Jun 25) e10281.
[23] Q. Wang, J. Liao, M. Lapata, M. Macleod, Risk of bias assessment in preclinical literature using natural language processing, Res. Synth. Methods 13 (3) (2022 May) 368–380.
[24] Q. Zou, S. Xie, Z. Lin, M. Wu, Y. Ju, Finding the best classification threshold in imbalanced classification, Big Data Res. 5 (2016 Sep) 2–8.
[25] P. von Stackelberg, R. Goedhart, Ş.İ. Birbil, R.J.M.M. Does, Comparison of threshold tuning methods for predictive monitoring, Qual. Reliab. Eng. Int. [Internet] (2023), https://doi.org/10.1002/qre.3436. Sep 3; Available from:.
[26] Y. Huang, W. Li, F. Macheret, R.A. Gabriel, L. Ohno-Machado, A tutorial on calibration measurements and calibration models for clinical prediction models, J. Am. Med. Inform. Assoc. 27 (4) (2020 Apr 1) 621–633.
[27] Y. Aphinyanaphongs, Text categorization models for high-quality article retrieval in internal medicine, J. Am. Med. Inform. Assoc. 12 (2) (2004 Nov 23) 207–216.
[28] S.A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M.A. Riegler, P. Halvorsen, et al., On evaluation metrics for medical applications of artificial intelligence, Sci. Rep. 12 (1) (2022 Apr 8) 5979.
[29] D. Chicco, G. Jurman, The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification, BioData Min. [Internet] 16 (1) (2023), https://doi.org/10.1186/s13040-023-00322-4. Feb 17, Available from:.
[30] B.J. Erickson, F. Kitamura, Magician's corner: 9. Performance metrics for machine learning models, Radiol. Artif. Intell. 3 (3) (2021) e200126.
[31] M. Assel, D.D. Sjoberg, A.J. Vickers, The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models, Diagn. Progn. Res. 1 (1) (2017 Dec 2) 19.
[32] D.M.W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation [Internet], arXiv [cs.LG], 2020, Available from: http://arxiv.org/abs/2010.16061.
[33] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (1) (2020 Jan 2) 6.
[34] D. Powers, Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation, J. Mach. Learn. Technol. 2 (1) (2011) 37–63.
[35] N.L. Wilczynski, D. Morgan, R.B. Haynes, Hedges Team, An overview of the design and methods for retrieving high-quality studies for clinical care, BMC Med. Inform. Decis. Mak. (2005). Jun 21;5(1):20.
[36] [cited 2025 Feb 14]. Available from: https://hiruweb.mcmaster.ca/hkr/what-we-do/plus-directory/.
[37] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020 Feb 15) 1234–1240.
[38] M. Yasunaga, J. Leskovec, P. Liang, LinkBERT: Pretraining language models with document links [Internet], arXiv [cs.CL]. 2022 [cited 2024 Aug 13], Available from: http://arxiv.org/abs/2203.15827.
[39] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, et al., Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthcare 3 (1) (2022 Jan 31) 1–23.
[40] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text [Internet], arXiv [cs.CL]. 2019 [cited 2024 Aug 13], Available from: http://arxiv.org/abs/1903.10676.
[41] K.R. Kanakarajan, B. Kundumani, M. Sankarasubbu, BioELECTRA: Pretrained Biomedical text Encoder using Discriminators, in: D. Demner-Fushman, K.B. Cohen, S. Ananiadou, J. Tsujii, (Eds.), Proceedings of the 20th Workshop on Biomedical Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 143–54.
[42] BLURB leaderboard [Internet]. [cited 2024 Aug 13], Available from: https://microsoft.github.io/BLURB/leaderboard.html.
[43] C. Lokker, E. Bagheri, W. Abdelkader, R. Parrish, M. Afzal, T. Navarro, et al., Deep learning to refine the identification of high-quality clinical research articles from the biomedical literature: Performance evaluation, J. Biomed. Inform. 142 (104384) (2023 Jun) 104384.
[44] Journals Reviewed [Internet]. [cited 2025 Feb 14]. Available from: https://hiru.mcmaster.ca/hiru/journalslist.asp.
[45] R.B. Haynes, J. Holland, C. Cotoi, R.J. McKinlay, N.L. Wilczynski, L.A. Walters, 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. Inform. Assoc. 13 (6) (2006 Nov 1) 593–600.
[46] N.L. Wilczynski, C.J. Walker, K.A. McKibbon, R.B. Haynes, Assessment of methodologic search filters in MEDLINE, Proc. Annu. Symp. Comput. Appl. Med. Care. 601–5 (1993).
[47] MCMASTER+ [Internet]. [cited 2024 Aug 13]. Available from: https://plus.mcmaster.ca/McMasterPLUSDB/.
[48] I. Loshchilov, F. Hutter, Decoupled weight decay regularization [Internet], arXiv [cs.LG], 2017, Available from: http://arxiv.org/abs/1711.05101.
[49] X.C. Yin, C. Yang, W.Y. Pei, H.W. Hao, Shallow classification or deep learning: An experimental study, in: In: 2014 22nd International Conference on Pattern Recognition. IEEE, 2014, pp. 1904–1909.
[50] M. Oleynik, A. Kugic, Z. Kasáč, M. Kreuzthaler, Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification, J. Am. Med. Inform. Assoc. 26 (11) (2019 Nov 1) 1247–1254.
[51] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, et al., A survey on text classification: From shallow to deep learning [Internet], arXiv [cs.CL], 2020, Available from: http://arxiv.org/abs/2008.00364.
[52] B. Clavié, A. Ciceu, F. Naylor, G. Soulié, T. Brightwell, Large language models in the workplace: A case study on prompt engineering for job type classification, in: Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2023, p. 3–17. (Lecture notes in computer science).
[53] H. Naveed, A.U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, et al., A comprehensive overview of large Language Models [Internet], arXiv [cs.CL], 2023, Available from: http://arxiv.org/abs/2307.06435.
[54] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, et al., A survey of large language models [Internet], arXiv [cs.CL], 2023, Available from: http://arxiv.org/abs/2303.18223.
[55] A. Edwards, J. Camacho-Collados, Language models for text classification: Is in-context learning enough? [Internet], arXiv [cs.CL], 2024, Available from: http://arxiv.org/abs/2403.17661.
[56] A. Kostina, M.D. Dikaiakos, D. Stefanidis, G. Pallis, Large Language Models for text classification: Case study and comprehensive review [Internet], arXiv [cs.CL], 2025 [cited 2025 Feb 13], Available from: http://arxiv.org/abs/2501.08457.
[57] Y. Guo, A. Ovadje, M.A. Al-Garadi, A. Sarker, Evaluating large language models for health-related text classification tasks with public social media data, J. Am. Med. Inform. Assoc. 31 (10) (2024 Oct 1) 2181–2189.
[58] S. Li, P. Zhao, H. Zhang, X. Sun, H. Wu, D. Jiao, et al., Surge phenomenon in optimal learning rate and batch size scaling [Internet], arXiv [cs.LG], 2024, Available from: http://arxiv.org/abs/2405.14578.
[59] D. Granziol, S. Zohren, S. Roberts, Learning rates as a function of batch size: A random matrix theory approach to neural network training [Internet], arXiv [stat.ML], 2020, Available from: http://arxiv.org/abs/2006.09092.
[60] W. Kusa, A. Lipani, P. Knoth, A. Hanbury, An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews, Intelligent Systems with Applications. 18 (200193) (2023 May) 200193.
[61] The AI community building the future [Internet]. [cited 2024 Nov 28]. Available from: https://huggingface.co/.
[62] N. Markey, B. Howitt, I. El-Mansouri, C. Schwartzenberg, O. Kotova, C. Meier, Clinical trials are becoming more complex: a machine learning analysis of data from over 16,000 trials, Sci. Rep. 14 (1) (2024 Feb 12) 3514.
[63] N. Heßler, M. Rottmann, A. Ziegler, Empirical analysis of the text structure of original research articles in medical journals, PLoS One 15 (10) (2020 Oct 8) e0240288.
[64] A.J.A. Ty, Z. Fang, R.A. Gonzalez, P.J. Rozdeba, H.D.I. Abarbanel, Machine learning of time series using time-delay embedding and precision annealing [Internet], Available from: arXiv [cs.LG]. (2019) http://arxiv.org/abs/1902.05062.
[65] J.E. Lo, E.Y.C. Kang, Y.N. Chen, Y.T. Hsieh, N.K. Wang, T.C. Chen, et al., Data homogeneity effect in deep learning-based prediction of type 1 diabetic retinopathy, J. Diabetes Res. 28 (2021) (2021 Dec) 2751695.
[66] A.K. Ambalavanan, M.V. Devarakonda, Using the contextual language model BERT for multi-criteria classification of scientific articles, J. Biomed. Inform. 112 (103578) (2020 Dec) 103578.
[67] M. Afzal, B.J. Park, M. Hussain, S. Lee, Deep learning based biomedical literature classification using criteria of scientific rigor, Electronics (Basel) 9 (8) (2020 Aug 5) 1253.

[68] M. Villa, M. Le Pera, T. Cassina, M. Bottega, Reporting quality of abstracts from randomised controlled trials published in leading critical care nursing journals: a methodological quality review, BMJ Open 13 (3) (2023 Mar 15) e070639.
[69] F. Hua, T. Walsh, A.M. Glenny, H. Worthington, Structure formats of randomised controlled trial abstracts: a cross-sectional analysis of their current usage and association with methodology reporting, BMC Med. Res. Methodol. 18 (1) (2018 Jan 10) 6.
[70] J. Hartley, Current findings from research on structured abstracts: an update, J. Med. Libr Assoc. 102 (3) (2014 Jul) 146–148.
[71] I. Beltagy, M.E. Peters, C.A. Longformer, The Long-Document Transformer [Internet], Available from: arXiv [cs.CL] (2020) http://arxiv.org/abs/2004.05150.
[72] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, et al., Big bird: Transformers for longer sequences [Internet], Available from: arXiv [cs.LG] (2020) http://arxiv.org/abs/2007.14062.
[73] A. Hassani, S. Walton, J. Li, S. Li, H. Shi, Neighborhood Attention Transformer [Internet], arXiv [cs.CV], 2022, Available from: http://arxiv.org/abs/2204.07143.
[74] Y. Li, R.M. Wehbe, F.S. Ahmad, H. Wang, Y. Luo, Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences [Internet], arXiv [cs.CL], 2022 [cited 2025 Feb 14], Available from: https://www.alphaxiv.org/abs/2201.11838v3.
[75] E. Wallace, Y. Wang, S. Li, S. Singh, M. Gardner, Do NLP models know numbers? Probing numeracy in embeddings, in: K. Inui, J. Jiang, V. Ng, X. Wan, (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. p. 5307–15.
[76] J.C. Rothwell, S.A. Julious, C.L. Cooper, A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal, Trials 19 (1) (2018 Oct 10) 544.
[77] S.J. Walters, R.M. Jacques, I.B. Dos Anjos Henriques-Cadby, J. Candlish, N. Totton, M.T.S. Xian, Sample size estimation for randomised controlled trials with repeated assessment of patient-reported outcomes: what correlation between baseline and follow-up outcomes should we assume? Trials 20 (1) (2019 Sep 13) 566.
[78] J.L. Leevy, J.M. Johnson, J. Hancock, T.M. Khoshgoftaar, Threshold optimization and random undersampling for imbalanced credit card data, J. Big Data [Internet] (2023), https://doi.org/10.1186/s40537-023-00738-z. May 6;10(1). Available from:.
[79] M. Abdelhamid, A. Desai, Balancing the scales: A comprehensive study on tackling class imbalance in binary classification [Internet], arXiv [cs.LG], 2024 [cited 2024 Oct 9], Available from: http://arxiv.org/abs/2409.19751.
[80] Z.C. Lipton, C. Elkan, B. Narayanaswamy, Thresholding classifiers to maximize F1 score [Internet], arXiv [stat.ML], 2014 [cited 2024 Oct 9], Available from: http://arxiv.org/abs/1402.1892.
[81] A.M. Cohen, W.R. Hersh, K. Peterson, P.Y. Yen, Reducing workload in systematic review preparation using automated citation classification, J. Am. Med. Inform. Assoc. 13 (2) (2006 Mar 1) 206–219.
[82] O. Rainio, J. Teuho, R. Klén, Evaluation metrics and statistical tests for machine learning, Sci. Rep. 14 (1) (2024 Mar 13) 6086.
[83] K. Hajian-Tilaki, Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, Caspian J. Intern. Med. 4 (2) (2013) 627–635. Spring.
[84] M.D. Ruopp, N.J. Perkins, B.W. Whitcomb, E.F. Schisterman, Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection, Biom. J. 50 (3) (2008 Jun) 419–430.
[85] T.J.T. Heiser, M.L. Allikivi, M. Kull, Shift happens: Adjusting classifiers [Internet], Available from: arXiv [cs.LG] (2021) http://arxiv.org/abs/2111.02529.
[86] A.F. Machado, A. Charpentier, E. Flachaire, E. Gallic, F. Hu, Probabilistic scores of classifiers, calibration is not enough [internet], Available from: Arxiv [cs.LG] (2024) http://arxiv.org/abs/2408.03421.
[87] E. Loffredo, M. Pastore, S. Cocco, R. Monasson, Restoring balance: principled under/oversampling of data for optimal classification [Internet], Available from: arXiv [cond-Mat.dis-Nn] (2024) http://arxiv.org/abs/2405.09535.
[88] S. Bethard, We need to talk about random seeds [Internet], arXiv [cs.CL], 2022, Available from: http://arxiv.org/abs/2210.13393.
[89] P. Madhyastha, R. Jain, On model stability as a function of random seed [Internet], Available from: arXiv [cs.LG] (2019) http://arxiv.org/abs/1909.10447.
[90] D. Picard, Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision [Internet], Available from: arXiv [cs.CV]. (2021) http://arxiv.org/abs/2109.08203.
[91] J. Åkesson, J. Töger, E. Heiberg, Random effects during training: Implications for deep learning-based medical image segmentation, Comput. Biol. Med. 180 (108944) (2024 Sep) 108944.
[92] L. Schader, W. Song, R. Kempker, D. Benkeser, Don't let your analysis go to seed: On the impact of random seed on machine learning-based causal inference, Epidemiology 35 (6) (2024 Nov 1) 764–778.
[93] T.J. Bradshaw, Z. Huemann, J. Hu, A. Rahmim, A guide to cross-validation for artificial intelligence in medical imaging, Radiol. Artif. Intell. 5 (4) (2023 Jul) e220232.
[94] D. Wilimitis, C.G. Walsh, Practical considerations and applied examples of cross-validation for model development and evaluation in health care: Tutorial, JMIR AI. 18 (2) (2023 Dec) e49023.
[95] S. Raschka, Model evaluation, model selection, and algorithm selection in machine learning [Internet], Available from: arXiv [cs.LG] (2018) http://arxiv.org/abs/1811.12808.
[96] S. Zheng, H. Lin, S. Zha, M. Li, Accelerated large batch optimization of BERT pretraining in 54 minutes [internet], Available from: arXiv [cs.LG] (2020) http://arxiv.org/abs/2006.13484.
[97] N. Gkouti, P. Malakasiotis, S. Toumpis, I. Androutsopoulos, Should I try multiple optimizers when fine-tuning pre-trained transformers for NLP tasks? Should I tune their hyperparameters? [internet], Available from: Arxiv [cs.CL] (2024) http://arxiv.org/abs/2402.06948.
[98] S. Falkner, A. Klein, F.B.O.H.B. Hutter, Robust and efficient hyperparameter optimization at scale [Internet], Available from: arXiv [cs.LG] (2018) http://arxiv.org/abs/1807.01774.