Automating Risk of Bias Inference in Clinical Studies

Abel C. Dias[1], **Viviane P. Moreira**[1], **João Luiz D. Comba**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{abel.correa,viviane,comba}@inf.ufrgs.br

Abstract. *One of the best quality indicators in the clinical domain is the risk of bias (RoB). Bias refers to any systematic error in results that may lead to misinterpretation. In the systematic review process, human reviewers manually assess the RoB. Existing works attempt to automate this process using support vector machines (SVM), convolutional neural networks (CNN), or logistic regression. To the best of our knowledge, no previous work has explored Transformer-based models for the RoB assessment in clinical studies. In this work, we propose a novel model for RoB inference based on the Transformers architecture, called RoBIn (i.e., Risk of Bias Inference). We employ a machine reading comprehension (MRC) approach to extract evidence that is then classified with a RoB label. Furthermore, we use distant supervision to annotate a dataset for MRC and RoB inference. As a final contribution, a large language model (LLM) application was created to receive clinical trials as input and to assess the RoB. The proposed model outperforms state-of-the-art approaches and other LLMs in many settings, with high accuracy (AUC-ROC*$= 0.83$*) for different bias types.*

1. Introduction

The biomedical literature plays a crucial role in advancing healthcare research by providing clinical evidence that informs decision-making and healthcare policy development. One of the biggest challenges in extracting reliable information is **bias**, which can lead to incorrect conclusions and potentially harm patients if treatments are based on flawed data. The **RoB** is a key indicator used to assess the quality of clinical studies, traditionally evaluated by trained experts such as researchers and clinicians. However, the rapid increase in scientific publications makes manual assessment increasingly difficult [Landhuis 2016].

An example of this challenge emerged during the COVID-19 pandemic, which caused a surge in biomedical research. The medRxiv preprint repository, launched in 2019, saw an exponential increase in publications, with 62% of preprints in December 2020 related to COVID-19, and by October 2024, it hosted nearly 60,000 papers, half of which were published. The urgency to disseminate findings quickly increased the risk of errors, fraud, and retractions, particularly due to a lack of peer review and pressure for immediate results. Studies on treatments like Ivermectin, Tocilizumab, and Oseltamivir during the COVID-19 and H1N1 pandemics exemplify cases where clinical trials exhibited a high RoB, sometimes influenced by commercial or political interests [Hasdeu and Tortosa 2021, Brainard 2020].

Given these challenges, relying solely on manual RoB assessment is impractical. *This thesis focuses on developing automated tools to evaluate RoB in clinical studies,*

offering a scalable and efficient solution. We propose leveraging Natural Language Processing (NLP) techniques to analyze biomedical literature and provide evidence-based assessments, helping researchers and clinicians make more informed decisions.

2. Research Problem and Contributions

The RoB in biomedical literature can arise from errors in study design or analysis, affecting the reliability of clinical research. Categorizing RoB types helps systematically assess study quality. Examples include **selection bias**, where patient selection deviates from proper methodology, **performance bias**, when participants are aware of the treatment, and **detection bias**, when outcome assessors know the treatment being administered [Phillips et al. 2022]. Various guidelines and methodologies exist for evaluating RoB, including the Jadad Score, Delphi List, CONSORT, and Cochrane Collaboration.

The Cochrane Collaboration is a widely recognized reference for RoB evaluation, maintaining the Cochrane Database of Systematic Reviews (CDSR), which includes nearly 10,000 reviews as of October 2024. Given the structured RoB evaluations available in the CDSR database, researchers have explored automating RoB assessment using machine learning models like SVM, CNNs, and logistic regression [Millard et al. 2015, Marshall et al. 2016, Zhang et al. 2016, Marshall et al. 2020]. These models attempt to classify RoB types based on previously annotated data. However, a limitation is the lack of a public dataset to train machine learning models effectively.

To address this, our first task was to create a public dataset, referred to **RoBIn dataset**, by processing the RoB sections from CDSR reviews. This dataset captures information such as bias types, judgments, and supporting evidence from clinical studies, and can be used to enable future research to enhance automated RoB assessment. Figure 1 illustrates the steps for creating RoBIn dataset.

The proposal for automated RoB assessment relies on two key NLP tasks: MRC and Sequence Classification. MRC enables machines to extract evidence by reading text and answering related questions, while Sequence Classification assigns labels (*e.g.,* low, high, or unclear RoB) to text sequences. The goal is to extract and classify RoB evidence from clinical trial texts.

We propose Transformer-based models [Vaswani et al. 2017] designed for two tasks (depicted in Figure 2): (1) identifying supporting sentences from clinical texts (MRC task) and (2) classifying the RoB from the extracted evidence (RoB inference task).

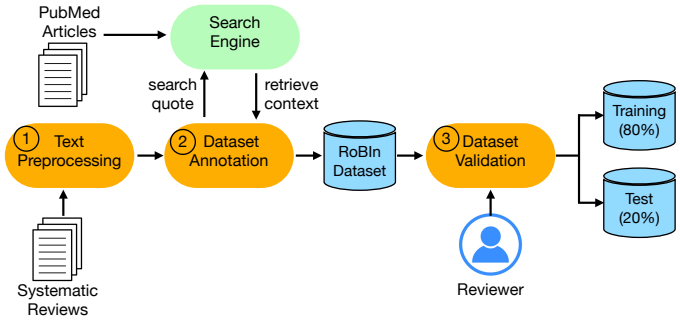


Figure 1. Steps for creating the RoBIn dataset.

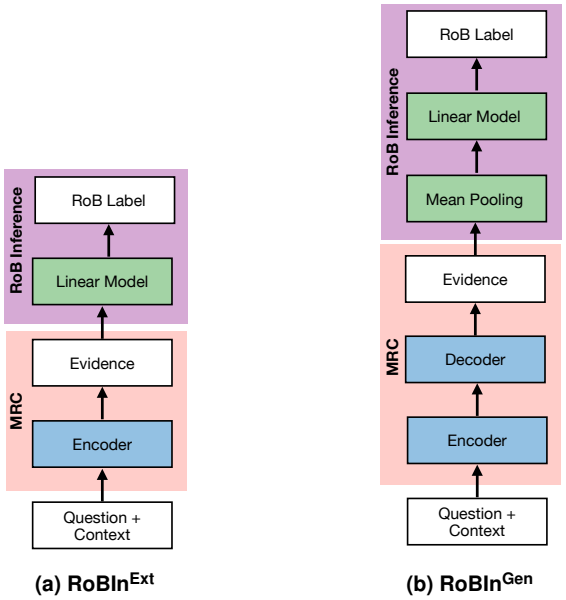


Figure 2. Architecture of the Extractive and Generative RoBIn models

Transformers are well-suited for these tasks due to their ability to capture complex contextual relationships and handle domain-specific biomedical language. We developed two models: **RoBIn^{Ext}**, an extractive model based on the encoder of the transformer architecture [Vaswani et al. 2017] that identifies relevant text spans, and **RoBIn^{Gen}**, a generative model based on the encoder-decoder that produces textual outputs for RoB classification. Both models were evaluated against state-of-the-art approaches.

To make our models accessible, we developed the **RoBIn chatbot**, an application using retrieval-augmented generation (RAG) and RoBIn model for RoB assessment. The chatbot, available as a web application, allows users to upload clinical trial reports for automated RoB evaluation, providing a practical tool for researchers and clinicians.

The main contributions of this thesis can be summarized as follows:

- **RoBIn dataset**[1], a dataset automatically annotated from CDSR reviews that contains information about RoB judgment and supporting evidence;
- **RoBIn extractive and generative transformer-based models**[2], trained for assessing the RoB retrieving supporting evidence; and
- **RoBIn chatbot**[3], a web-based application that relies on an LLM-based agent to evaluate RoB and answer related questions in natural language.

3. Results

RoBIn was evaluated using the RoBIn dataset on two tasks: MRC to identify evidence-supporting sentences and RoB inference to classify the RoB itself. We also tested LLMs for both tasks using GPT-4o-mini 8B, Llama3.1 8B, and Gemma2 9B, with prompts designed for zero-, one-, and few-shot learning. Traditional machine learning baselines such

[1]RoBIn dataset: `https://github.com/phdabel/RoBIn/tree/main/data`
[2]RoBIn: `https://github.com/phdabel/robin`
[3]RoBIn chatbot: `https://github.com/phdabel/robin-chatbot`

Table 1. Results for the MRC Task. Best results in red.

Metric	RoBIn^{Gen}	RoBIn^{Ext}	Llamma 3.1 8B			Gemma2 9B			GPT-4o-Mini 8B		
			0-shot	1-shot	few-shot	0-shot	1-shot	few-shot	0-shot	1-shot	few-shot
F1	93.81	97.10	40.95	50.99	53.25	50.78	58.74	61.90	55.54	52.47	56.71
F_{BERT}	89.52	91.30	58.97	70.30	74.68	64.36	76.80	80.91	72.85	68.97	75.36
EM	82.09	87.76	17.51	24.94	27.29	28.64	33.84	34.99	29.76	29.99	32.37

as SVM and Logistic Regression (LR) were also implemented, as they have been widely used for RoB inference [Millard et al. 2015, Marshall et al. 2016, Marshall et al. 2020, Gonçalves Pereira et al. 2020].

For MRC evaluation, metrics included F1 score, which considers partial matches, Exact Match (EM), which requires a perfect match, and BERTScore (F_{BERT}), which leverages contextual embeddings for comparison. The RoB inference task was assessed using precision, recall, macro F1-score, AUC-ROC, and the precision-recall curve[4].

Table 1 shows the overall MRC results for each model. RoBIn^{Ext} outperforms all models in both F1, F_{BERT}, and EM for all bias types. Some errors in RoBIn^{Ext} refer to when the sentence should end. For example, in some results, the sentence ended early at the first punctuation mark. RoBIn^{Gen} performs slightly lower than RoBIn^{Ext} but still better than the LLMs. Some of the errors made by RoBIn^{Gen} refer to word replacement.

When evaluating LLM predictions, it was observed that some responses were longer and more detailed than the ground truth, which negatively impacted scores due to the nature of the evaluation metrics. In some cases, the models provided the correct evidence along with an additional, unnecessary one, further affecting performance. Among the LLMs, Gemma2 in the few-shot setting achieved the highest overall score, outperforming Llama 3.1 and GPT-4o-mini in different configurations. Llama 3.1 had the lowest performance, while GPT-4o-mini produced competitive results but remained inferior to Gemma2 in key metrics such as EM and F1.

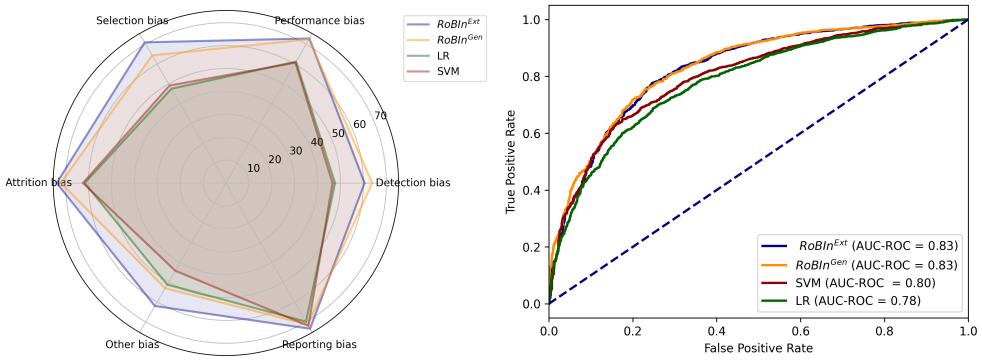
Table 2. Risk of Bias Inference Results: Models are listed in columns, with overall metric results in rows (best results in red).

Metric	RoBIn^{Gen}	RoBIn^{Ext}	SVM	LR	Llamma3.1 8B			Gemma2 9B			GPT-4o-Mini 8B		
					0-shot	1-shot	few-shot	0-shot	1-shot	few-shot	0-shot	1-shot	few-shot
F1	72.75	74.18	71.29	69.11	64.54	68.15	68.12	67.57	67.56	70.03	69.20	69.62	69.60
Prec.	73.21	75.49	70.84	68.96	65.06	67.75	68.43	67.18	68.04	71.16	68.83	69.17	69.14
Rec.	74.51	73.62	73.02	71.26	67.15	69.35	67.87	68.39	67.21	69.35	69.78	70.99	70.72

Table 2 presents the results for the RoB inference task, where scores are macro-averaged. RoBIn^{Ext} achieved the best F1 and precision, while RoBIn^{Gen} had the highest recall. A statistical test showed no significant differences between RoBIn^{Ext} and RoBIn^{Gen}, but both outperformed the baseline models, with Logistic Regression (LR) being the worst overall. Figure 3 shows the performance of the models in terms of F1 score for each bias (i.e., the radar plot in Figure 3a) and the area under the receiver operating characteristic curve (i.e., AUC-ROC in Figure 3b). Among LLMs, Gemma was the top model, which was consistent with previous results.

The RoBIn chatbot (Figure 4) uses ReAct prompting [Yao et al. 2023] to generate reasoning traces, capturing the LLM's decision-making process with internal logic and

[4]Note: Omitted due to lack of space.



(a) Macro F1 **(b) AUC-ROC**

Figure 3. Macro F1 and AUC-ROC of the best models.

external feedback for transparency. The prototype allows users to upload files for RoB evaluation, summarization, and other LLM-driven tasks.

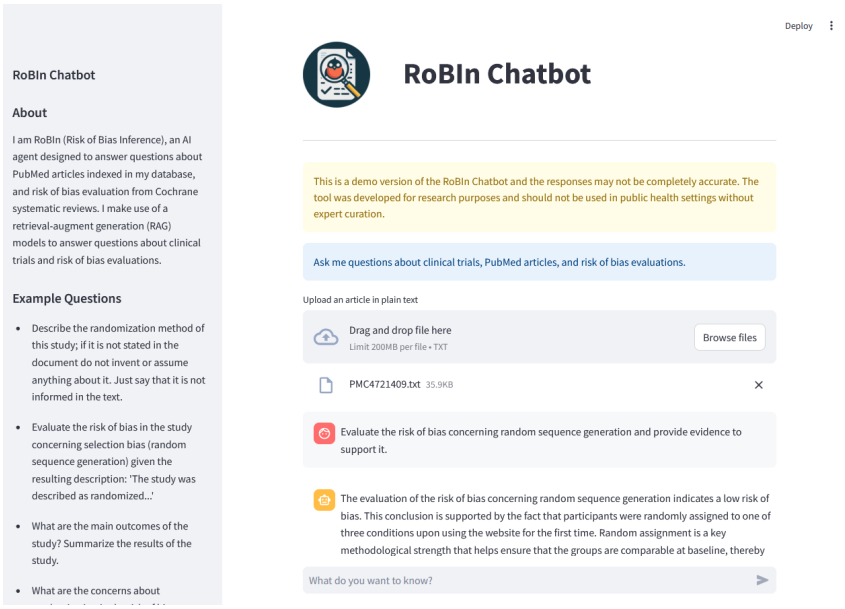


Figure 4. RoBIn chatbot

4. Discussion

This work addressed the assessment of RoB in clinical trials, focusing on retrieving supporting evidence for RoB judgments. This multidisciplinary challenge required adapting biomedical language models to specialized tasks in biomedicine, preclinical research, and risk assessment. Three key contributions were introduced: **RoBIn dataset**, an automatically built dataset for MRC and RoB inference; the **RoBIn$^{\text{Gen}}$** and **RoBIn$^{\text{Ext}}$** models for retrieving evidence from clinical trial publications and performing RoB classification; and the **RoBIn chatbot**, an LLM-powered application using RoBIn$^{\text{Ext}}$ to assess RoB in user-submitted clinical trials.

The RoBIn models outperformed state-of-the-art approaches and LLMs in various scenarios, identifying supporting evidence for RoB assessment. Both models demonstrated strong performance in MRC tasks, with RoBIn$^{\text{Ext}}$ being more efficient. Based on

extracted/generated evidence, RoBIn performs binary classification to determine if a trial has low RoB or high/unclear RoB, achieving an $AUC - ROC = 0.83$.

However, the RoBIn dataset has limitations, such as a small sample size due to automated data acquisition, class imbalance favoring low RoB trials, making high/unclear RoB predictions more challenging, and limited human intervention in annotation, which can introduce noise in the data. Despite these constraints, RoBIn demonstrates the potential of NLP in transforming systematic research methodologies, bridging computer science and healthcare. Automated RoB assessment holds entrepreneurial value, with applications in accelerating research, reducing costs, and enhancing transparency in clinical literature evaluation.

References

Brainard, J. (2020). Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? Science.

Gonçalves Pereira, R., Zanon Castro, G., Azevedo, P., Tôrres, L., Zuppo, I., Rocha, T., and Afonso Guerra, A. (2020). Mcrb: A multiclassifier tool for risk of bias assessment in a systematic review to produce health evidence to decision making. In *IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6.

Hasdeu, S. and Tortosa, F. (2021). Riesgo de sesgo de publicación en intervenciones terapéuticas para la COVID-19. *Pan American Journal of Public Health*, 45.

Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, 535:457–458.

Marshall, I. J., Kuiper, J., and Wallace, B. C. (2016). Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association (JAMIA)*, 23:193–201.

Marshall, I. J., Nye, B., Kuiper, J., Noel-Storr, A., Marshall, R., Maclean, R., Soboczenski, F., Nenkova, A., Thomas, J., and Wallace, B. C. (2020). Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association (JAMIA)*, 27:1903–1912.

Millard, L. A., Flach, P. A., and Higgins, J. P. (2015). Machine learning to assist risk-of-bias assessments in systematic reviews. *International Journal of Epidemiology*, 45(1):266–277.

Phillips, M. R., Kaiser, P., Thabane, L., Bhandari, M., Chaudhary, V., Wykoff, C. C., Sivaprasad, S., Sarraf, D., Bakri, S. J., Garg, S. J., Singh, R. P., Holz, F. G., and Wong, T. Y. (2022). Risk of bias: why measure it, and how?

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., and et. al. (2017). Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010. Curran Associates Inc.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023). React: Synergizing reasoning and acting in language models.

Zhang, Y., Marshall, I., and Wallace, B. C. (2016). Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 795–804. Association for Computational Linguistics.