7-1-2023

Analysis of predictive performance and reliability of classifiers for quality assessment of medical evidence revealed important variation by medical area

Simon Šuster
School of Computing and Information Systems

Timothy Baldwin
University of Melbourne & Mohamed Bin Zayed University of Artificial Intelligence

Karin Verspoor
University of Melbourne & RMIT University

Recommended Citation

S. Šuster, T. Baldwin, and K. Verspoor, "Analysis of predictive performance and reliability of classifiers for Quality Assessment of medical evidence revealed important variation by Medical Area," *Journal of Clinical Epidemiology,* vol. 159, pp. 58–69, 2023. doi:10.1016/j.jclinepi.2023.04.006

Journal of Clinical Epidemiology 159 (2023) 58−69

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Analysis of predictive performance and reliability of classifiers for quality assessment of medical evidence revealed important variation by medical area

Simon Šuster[a,*], Timothy Baldwin[a,b], Karin Verspoor[c,a]

[a]*School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia*
[b]*Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates*
[c]*School of Computing Technologies, RMIT University, Melbourne, Australia*

Accepted 18 April 2023; Published online 27 April 2023

Abstract

Objectives: A major obstacle in deployment of models for automated quality assessment is their reliability. To analyze their calibration and selective classification performance.

Study Design and Setting: We examine two systems for assessing the quality of medical evidence, EvidenceGRADEr and RobotReviewer, both developed from Cochrane Database of Systematic Reviews (CDSR) to measure strength of bodies of evidence and risk of bias (RoB) of individual studies, respectively. We report their calibration error and Brier scores, present their reliability diagrams, and analyze the risk−coverage trade-off in selective classification.

Results: The models are reasonably well calibrated on most quality criteria (expected calibration error [ECE] 0.04−0.09 for EvidenceGRADEr, 0.03−0.10 for RobotReviewer). However, we discover that both calibration and predictive performance vary significantly by medical area. This has ramifications for the application of such models in practice, as average performance is a poor indicator of group-level performance (e.g., health and safety at work, allergy and intolerance, and public health see much worse performance than cancer, pain, and anesthesia, and Neurology). We explore the reasons behind this disparity.

Conclusion: Practitioners adopting automated quality assessment should expect large fluctuations in system reliability and predictive performance depending on the medical area. Prospective indicators of such behavior should be further researched.

Keywords: Systematic reviews; Automated quality assessment of medical evidence; Risk of bias; Critical appraisal; Calibration; Uncertainty estimation; Reliability; Selective classification; Disparity

1. Introduction

Systematic reviews, which analyze, aggregate and critically appraise all relevant published evidence on a specific question, are a cornerstone of medical evidence−based decision-making [1,2]. Their creation is recognized as an especially arduous process, typically taking 1.3 person years per review [3]. Often, reviews are out of date or simply unavailable for specific clinical questions due to the resources required [4−7]. This has given rise to burgeoning research into automation of different steps of the reviewing process, including article classification, screening for primary studies, data extraction, and quality assessment [7,8], supporting substantial speed-up [9]. In quality assessment specifically, machine learning (ML) and natural language processing (NLP) approaches that (partially) automate risk-of-bias (RoB) estimation of individual studies included in a review have been proposed [10−13].

Funding: The research is funded by the Australian Research Council through an Industrial Transformation Training Center Grant (IC170100030) in collaboration with IBM.

Availability of data and materials: The evaluation data of EvidenceGRADEr and RobotReviewer are available on Zenodo [45,46].

Conflict of interest statement: The authors declare that they have no competing interests.

Authors' contributions: S.Š. collected the data; contributed in conceptualization; conceived and designed the analysis; performed the analysis and the experiments; and wrote the paper. T.B. conceived and designed the analysis; contributed in conceptualization; wrote the paper; and contributed in supervision. K.V. conceived and designed the analysis; contributed in conceptualization; wrote the paper; and contributed in supervision. All authors revised and approved the final manuscript.

* Corresponding author. School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia. Tel.: +61-3-9035-4422.

E-mail address: simon.suster@unimelb.edu.au (S. Šuster).

https://doi.org/10.1016/j.jclinepi.2023.04.006

What is new?

Key findings

- We carry out a reliability analysis on 10 quality criteria from two machine learning-based quality-assessment models and find that on average they are well calibrated.
- Model reliability as well as selective classification and predictive performance vary importantly across medical areas. The distribution of medical evidence and of model outcome probabilities are important factors associated with the varying performance levels.

What this adds to what was known?

- Calibration and the potential of selective prediction for quality assessment systems have not been previously studied.
- Our work is the first to present a detailed analysis of those criteria and shows how these systems can be further calibrated.
- We identify and examine the performance disparity across medical areas, which may pose an obstacle to adoption of such systems in practical settings.

What is the implication and what should change now?

- While systems such as RobotReviewer and EvidenceGRADEr hold promise for supporting semi-automated quality assessment of medical evidence, large fluctuations in model reliability and predictive performance are to be expected by practitioners depending on the medical area.
- Indicators of model reliability and performance during deployment should be further researched.

Other approaches include assessing larger bodies of evidence with a wider set of quality criteria [14].

ML approaches to quality assessment are typically set as text classification tasks, where an abstract or a full article is encoded with a pretrained language model [15,16] and mapped to a quality label (e.g., high and low) using a neural network. Nontextual inputs may be considered as well, such as numerical (e.g., number of participants) and categorical (e.g., type of effect) features in the case of EvidenceGRADEr [14]. Apart from a quality label, RobotReviewer [10] also outputs snippets of text that support its decision and may be useful as a justification when using the classifier in an assistive setting. These quality assessment systems are trained using labeled data obtained from systematic reviews in which the decisions regarding RoB or Grading of Recommendations, Assessment, Development, and Evaluations (GRADE)[1] labels were made by human experts. We describe the systems used in our work in more detail in Sections 2.3.1 and 2.3.2.

While such approaches can speed up quality assessment in semiautomated settings [17], the trustworthiness of systematic reviewing technology in general remains an open question [18,19]. One way to develop trust is by managing uncertainty in the correctness of model predictions and knowing, prior to applying a model in a specific subdomain, what its accuracy will be [20]. Ideally, the confidence of a model in a prediction should be a good indicator of the likelihood for that prediction being correct. For example, if a system says 1,000 of the data points have an event x with probability 0.7, approximately 700 of them should indeed have this event. This would be useful in deployment so that developers can impose a threshold for showing a prediction to the final user. Or a user could decide themselves whether to keep a prediction based on the confidence score [21]. Uncertainty-calibrated classifiers will at least "know what they do not know," making their deployment in practice more reliable, even if their internal mechanisms are opaque or uninterpretable [22], Jiang et al., 2012 [23].

Our work is the first to thoroughly examine the question of the calibration of ML models for quality assessment in medical evidence synthesis. By including in our analysis two systems, RobotReviewer [11] for RoB assessment of individual studies and EvidenceGRADEr for grading the quality of bodies of evidence [14], we wish to lay out a path to the deployment of such tools for practical use.

2. Methods

2.1. Calibration evaluation

A model is calibrated if the confidence estimates of its predictions are aligned with the empirical likelihood of the model being correct. The difference between the two is known as calibration error [24]. In our analysis, we report the average over all predictions, known as expected calibration error (ECE), and use reliability diagrams to show how miscalibrated the model predictions are at various steps in the probability range. Finally, we report the Brier score [25–27] which apart from classifier's calibration performance takes into account also its discrimination capabilities (predictive performance). We include further methodological details about these methods in Appendix A.1.

2.2. Selective classification

The tools that we use to study the calibration of classifiers can be further exploited in the framework of selective

[1] Grading of Recommendation, Assessment, Development, and Evaluation framework [32].

classification [28–30]. Here, the model (or alternatively, the user) is granted the ability to decide which predictions should be trusted and kept (possibly for subsequent processing by an expert), and which should be rejected (possibly requiring a complete reassessment). The intuition behind selective prediction is to reduce the error rate (risk) by sacrificing coverage, that is, the proportion of all data points eligible for classification. A necessary condition for successfully applying selectivity is a match between the model's confidence and empirical accuracy, in other words, a well-calibrated model whose confidence scores can be trusted. In real-life applications, a practitioner would prefer the model with better coverage when comparing two models for selective prediction and for some maximum permissible error rate. Conversely, they could fix coverage and select a model that has better discrimination capability.

To show the potential of selective classification in reducing the risk of error in quality assessment tasks, we employ a straightforward approach. We impose a confidence threshold τ on model's predictions and keep those that exceed it while discarding all others. The effect can then be observed in a risk–coverage curve that displays the trade-off between the risk of error and the coverage across the entire spectrum of τ [30,31]. To obtain a single-value indication of the significance of this trade-off, we report the area under the risk–coverage curve (AUC-RCC) [29], where a smaller value indicates a better selective-prediction performance. We include the calculation details in Appendix A.1.

2.3. Models and evaluation data

2.3.1. EvidenceGRADEr

EvidenceGRADEr [14] is a ML system that performs quality assessment in the context of systematic reviews according to GRADE criteria [32]. The system assesses a body of evidence—a set of studies included in a systematic review, grouped by a specific Population, Intervention, Comparison, Outcome question—and outputs predictions for various quality characteristics (or tasks). In this work, we consider all binary quality classification tasks, that is, the presence or absence of a particular downgrading criterion (RoB, imprecision, inconsistency, indirectness, and publication bias) and the overall binary quality grading, in which one class represents low/very low-quality evidence, and another moderate/high-quality evidence.[2] The model output probabilities are taken to represent higher-quality evidence (or the absence of a downgrading criterion) if their value equals or exceeds 0.5, and lower-quality evidence otherwise.

[2] Such binary grading corresponds to how the quality assessment in systematic reviews is used by guideline panels. They typically associate a strong recommendation for an intervention with high/moderate confidence in the effect estimate for critical outcomes and are discouraged from doing so otherwise [47].

2.3.2. Data

In our analysis, we use the dataset created by [14], which was constructed from a 2020 snapshot of the Cochrane Database of Systematic Reviews (CDSR) containing 8,034 reviews.[3] The dataset was developed by extracting and organizing meta data of each review, textual parts of reviews (abstracts and summaries), summaries of findings, and certain characteristics of primary studies. For individual grading criteria, 59% of the total 13,440 data instances (bodies of evidence) are flagged for RoB, 55% for imprecision, 16% for inconsistency, 10% for indirectness, and 5% for publication bias. For two-tier GRADE scoring, low- and very low-quality labels are merged (n = 7,299), as well as moderate- and high-quality labels (n = 6,141), resulting in a roughly balanced set.

2.3.3. RobotReviewer

RobotReviewer is an NLP system for automatically determining the information describing the trial conduct, including the RoB of a study. The system takes as input a full-text article describing the conduct and results of a randomized controlled trial (RCT), and outputs a binary decision about whether the study is at low or high/unclear RoB for four of the Cochrane RoB1 criteria. It also extracts sentences found in the text supporting that decision. The system's output layer predicts the criteria by averaging the output probabilities of two different models, a linear model and a convolutional neural network [33]. As in EvidenceGRADER, the model's output probability represents low RoB for values equaling or exceeding 0.5, and high risk otherwise. We use the openly available implementation of the system, trained on 12,808 trial portable document formats with RoB annotations from CDSR.[4]

2.4. Data

To perform our analysis of RobotReviewer, we first collected all PubMed identifiers of primary studies included in CDSR as of March 2022 (totalling around 64,000) and removed those already used in the development of RobotReviewer, which yielded around 14,000 identifiers. Based on these, we directly crawled for the portable document format documents of open access articles from various publisher websites, as well as downloaded the documents accessible through the PubMed Central Open Access collection.[5] We obtained the ground-truth RoB labels from the judgments made by the review authors. Whenever a medical study was assessed for RoB in multiple reviews, we took the majority vote. Altogether, we were able to obtain 3,197 data instances for random sequence generation (RSG; with a proportion of low-risk evidence of 70%),

[3] https://www.cochranelibrary.com/cdsr/about-cdsr.
[4] https://github.com/ijmarshall/robotreviewer.
[5] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

2,715 for blinding of outcome assessment (BOA; 51% low-risk), 3,256 for allocation concealment (AC; 57% low-risk), and 2,556 for blinding of participants and personnel (BPP; 41% low-risk).

2.5. Prior evaluation

The use of RobotReviewer for RoB assessment in a semiautomated setting has been already extensively evaluated. We can categorize the studies into those looking at: (a) its accuracy in general [20,34–37] or in specific areas, such as nursing [34]; (b) the time used for assessment compared to human evaluators [17,35,36]; and (c) reviewer acceptability [17,36]. We were not able to identify any studies examining the questions of calibration or selective classification. The closest is a brief analysis of reliability of the related Trialstreamer [38]. The system finds and categorizes RCTs, including performing a RoB assessment. However, the tool is optimized for speed of processing, so it processes abstracts rather than full texts, and outputs a single, overall RoB score rather than scores for individual criteria. In this setup, the authors offer a short calibration analysis for the low-risk class only, reporting a Brier score of 0.1. Their calibration plot reveals that predictions closer to 0 (“low risk”) tend to be overconfident and those closer to 0.5 (“undecided”) tend toward underconfidence. They do not address the questions of selective classification and performance disparity.

3. Results and discussion

3.1. Predictive performance

In the case of EvidenceGRADEr, since we use the same evaluation data as in [14], the predictive performance matches that reported in the paper. However, for RobotReviewer, we perform the analysis on a previously unused test set. We therefore compare RobotReviewer’s predictive performance to that reported previously by the authors [39]. Evaluate eight different model variants, including a multitask support-vector machine specifically designed for RoB assessment, and a rationale-augmented convolutional neural network model. We summarize their results as a range of accuracy scores in Table 1. Since we adopt the publicly-available ensemble model as RobotReviewer’s implementation it is fair to assume that its expected performance should lie within the range. Our evaluation run shows that the results are close to, or within the range reported in [39]. While our test data do not represent an out-of-domain set, a large majority of data instances come from articles published after those used in the development of RobotReviewer. This might explain why we observe a small drop in performance.

Table 1. Accuracy of RobotReviewer on the new test set

Criterion	[39]	Ours
RSG	0.72–0.77	0.72
AC	0.72–0.76	0.67
BPP	0.73–0.76	0.74
BOA	0.63–0.70	0.62

Abbreviations: RSG, random sequence generation; AC, allocation concealment; BPP, blinding of participants and personnel; BOA, blinding of outcome assessment.

3.2. Off-the-shelf reliability

We first examine how well-calibrated the different quality assessment models are. For EvidenceGRADEr tasks, the calibration quality is best on the binary grading task (Fig. 1a), where the model achieves an ECE of 0.05 and the lowest maximum calibration error, 0.13.

The calibration on individual downgrading criteria is variable. The models for imprecision (Fig. 1b) and RoB (Fig. 1c) are reasonably well-calibrated, but they tend to be overly confident in predicting the absence of the downgrading criterion (the positive class),[6] and underconfident when predicting the presence of the criterion (negative class). This is a similar pattern to that observed in Trialstreamer (cf. Section 2.3.2). The results on the remaining tasks (Fig. 1d and f) display the same tendency, but there the predictions are heavily skewed toward the positive class.[7] We report all ECE scores in Tables 2 and 3.

The prediction of RoB criteria by RobotReviewer (Table3) is similarly well-calibrated to the binary GRADE, imprecision and RoB models of EvidenceGRADEr (Table 2). The ECEs are in the range 0.03–0.06, but the models make no high-confidence predictions for the negative class (presence of the risk criterion), that is, there are no predictions around the $[0, 0.2]$ confidence band (Fig. 2). This is unlikely to cause problems when the end user is interested in identifying higher quality evidence (without identified RoB criteria), but can be problematic when trying to flag with high confidence the evidence that is at a higher RoB.

3.3. Selective prediction

Selective classification enables the end users to focus on those predictions that are more likely to be correct, and treat them differently to those that are likely to be incorrect (these could be discarded or flagged for manual

[6] For example, when looking at the imprecision, we see that the predictions binned under 0.69–0.75 correspond to the true positive-class relative frequency of 0.6.

[7] This can be explained by a natural imbalance in the training and testing data distributions. The larger the imbalance, the larger the skew found in our reliability plots. For example, on the publication-bias task (Fig. 1f) where only one out of 20 examples are affected by the bias in the training data, the overwhelming majority of model’s predictions are concentrated at the positive end (absence of the downgrading criterion).

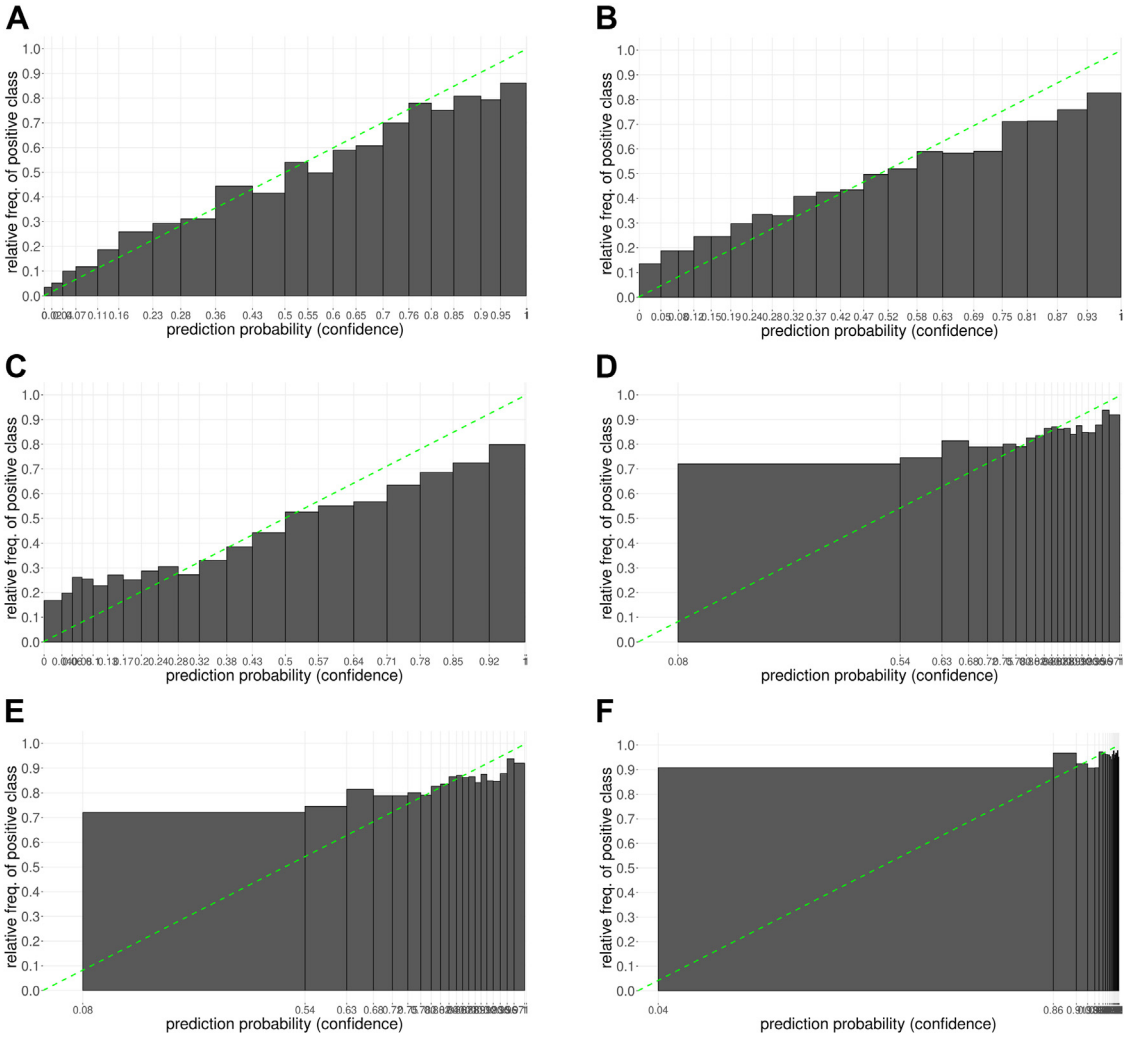

Fig. 1. Reliability diagrams for EvidenceGRADEr. The results are shown for overall binary grading as well as for the five individual downgrading criteria. A, Task: binary grading. B, Task: imprecision. C, Task: risk of bias. D, Task: inconsistency. E, Task: indirectness. F, Task: publication bias.

reassessment). In this way, the user is trading off coverage for accuracy. The exact implementation of such a mechanism is to be determined by the end application. For our purposes, we take a broad view of selective classification and therefore examine how selective prediction varies on the entire spectrum of coverage, that is, without committing to specific trade-offs.

Figures 3 and 4 show the risk–coverage curves for both systems. On most tasks, coverage can be successfully traded off for a reduced risk of error (the red "all" curves). Specifically, in binary quality scoring using EvidenceGRADEr (Fig. 3a), we see that the risk of error is over 25% at full coverage, but can be reduced to around 10% by keeping 20% of predictions. Comparable reduction in

Table 2. Calibration results for EvidenceGRADEr

Task	F1	P	R	ECE	Brier	AUC-RCC
Binary GRADE	0.74	0.74	0.74	0.05	0.18	0.15
Risk of bias	0.67	0.68	0.67	0.09	0.24	0.24
Imprecision	0.67	0.68	0.67	0.08	0.28	0.23
Indirectness	0.53	0.57	0.53	0.07	0.83	0.06
Inconsistency	0.50	0.57	0.52	0.06	0.69	0.13
Publication bias	0.50	0.52	0.50	0.04	0.93	0.03

Abbreviations: ECE, expected calibration error; AUC-RCC, area under the risk–coverage curve.

Table 3. Calibration results for RobotReviewer

Task	Acc	F1	P	R	ECE	Brier	AUC-RCC
AC	0.67	0.67	0.67	0.68	0.05	0.20	0.24
BPP	0.75	0.74	0.75	0.73	0.06	0.18	0.35
BOA	0.64	0.64	0.64	0.64	0.03	0.22	0.33
RSG	0.72	0.67	0.66	0.67	0.06	0.18	0.14

Abbreviations: ECE, expected calibration error; AUC-RCC, area under the risk–coverage curve; AC, allocation concealment; BPP, blinding of participants and personnel; BOA, blinding of outcome assessment; RSG = random sequence generation.

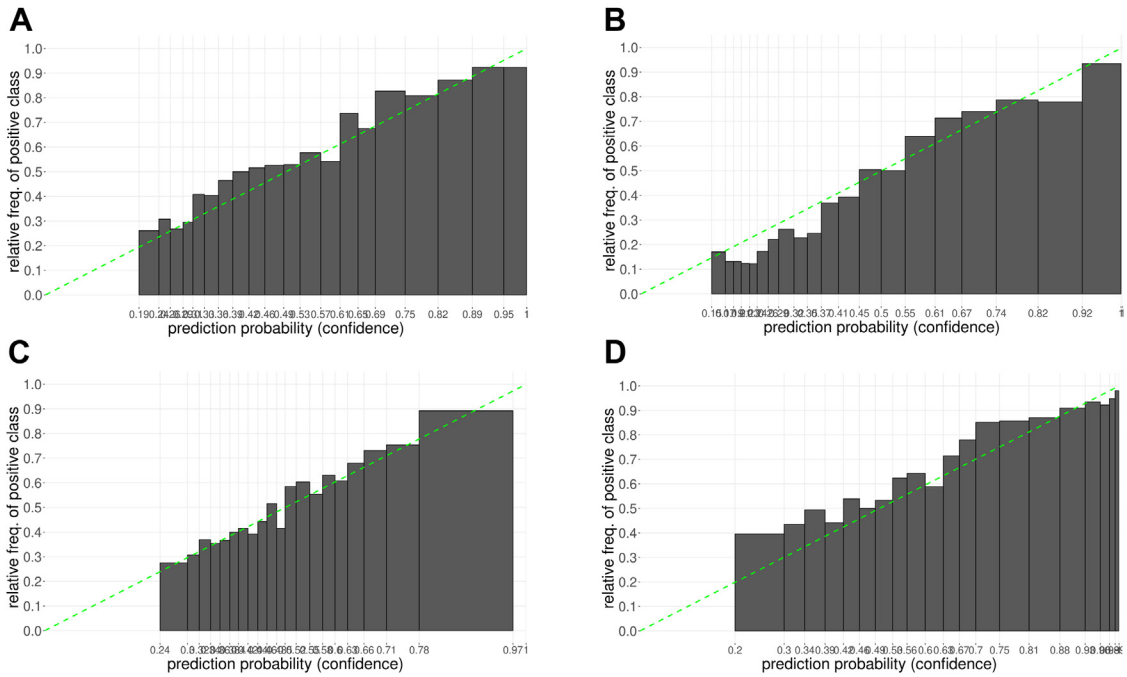

Fig. 2. Reliability diagrams for RobotReviewer. The results are shown for four different tasks (individual risk-of-bias criteria). A, Task: allocation concealment. B, Task: blinding of participants and personnel, C, Task: blinding of outcome assessment. D, Task: random sequence generation.

the risk of error can be observed for imprecision (Fig. 3b) and RoB (Fig. 3c) criteria, as well as for RobotReviewer's BOA (Fig. 4c) and RSG tasks (Fig. 4d). In AC (Fig. 4a) and BPP (Fig. 4b), the risk reduction is even greater, from 60% (full coverage) to 20% (20% coverage), and from 40% to 10%. For certain EvidenceGRADEr tasks like indirectness and publication bias, selective prediction is less effective. This can be explained by the limited data annotated with

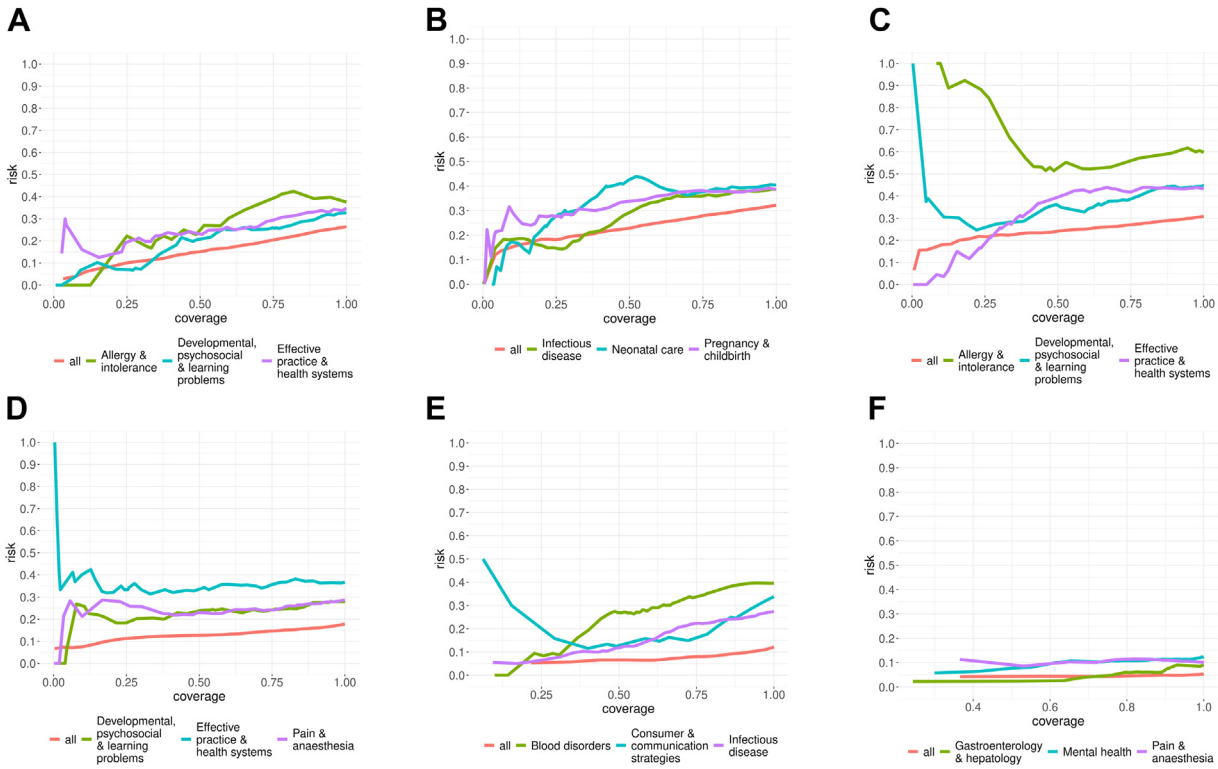

Fig. 3. Risk–coverage curves for EvidenceGRADEr on different tasks. The red ''all'' curve represents the risk–coverage trade-off without selecting a specific medical area. Each plot also includes the curves for the three medical areas (in blue, green, and magenta) with the highest risk of error at full coverage. A, Task: binary grading. B, Task: imprecision. C, Task: risk of bias. D, Task: inconsistency. E, Task: indirectness. F, Task: publication bias. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article).

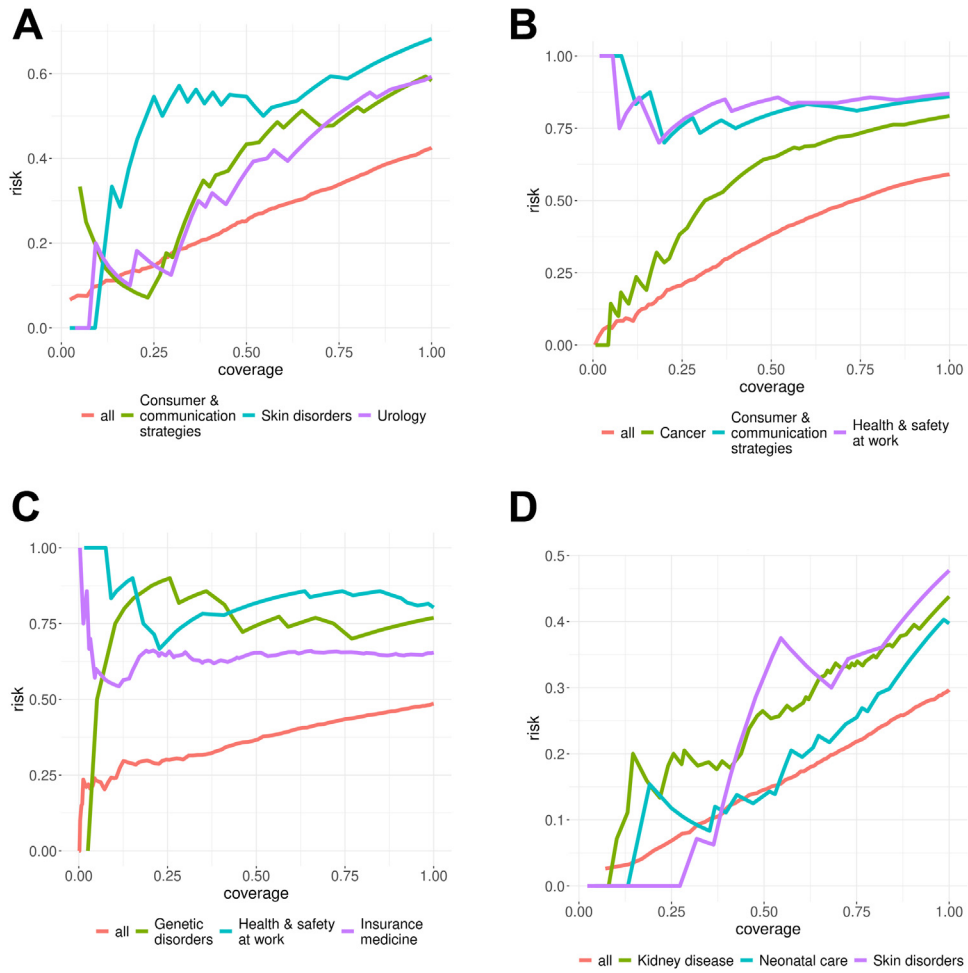

Fig. 4. Risk–coverage curves for RobotReviewer on different tasks. The red ''all'' curve represents the risk–coverage trade-off without selecting a specific medical area. Each plot also includes the curves for the three medical areas (in blue, green, and magenta) with the highest risk of error at full coverage. A, Task: allocation concealment. B, Task: blinding of participants and personnel, C, Task: blinding of outcome assessment. D, Task: random sequence generation. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article).

these two criteria and the resulting weak performance of the classifier.[8]

Overall, we find the decrease in risk to be largely linear, apart from a sharp reduction seen in certain tasks when abstaining from all but the most confident predictions (ie, closer to 0% coverage). Although these trade-offs look very promising for employing a selective approach in practice, we find that they may give a distorted picture of what takes place in certain ''subregions'' of medical evidence, namely on the instances grouped by individual medical areas.[9] In the risk–coverage plots, we illustrate this by including the expected trade-offs on three medical areas on which the risk of error is the greatest at full coverage. This paints a more realistic picture that illustrates the disparity across those areas. Although one may expect to attain higher-quality predictions by keeping more confident ones (reducing coverage), this is often not the case, and the risk of error can even increase dramatically on some medical areas. This is a reason for concern as it implies that selective prediction would work differently well depending on the medical area to which it is being applied. In the following, we discuss the disparity issue in more depth, and analyze it also with regard to predictive performance and reliability.

3.4. Disparity

We use the term disparity to mean that outcomes are group-dependent, which, if unobserved, may potentially be harmful in clinical decision making. In this work, we define the group as a collection of data instances belonging to a specific medical area. While groups could be defined in a number of ways, for example, using demographic or specific population criteria, we focus here only on medical areas, and leave others for future work. We would like to

[8] Although the risk of error is low across the coverage range, this is due to highly unbalanced predictions that are heavily biased toward the absence of the criteria.

[9] We define the medical areas to be the topics assigned by Cochrane to each review, cf. https://www.cochranelibrary.com/cdsr/reviews/topics.

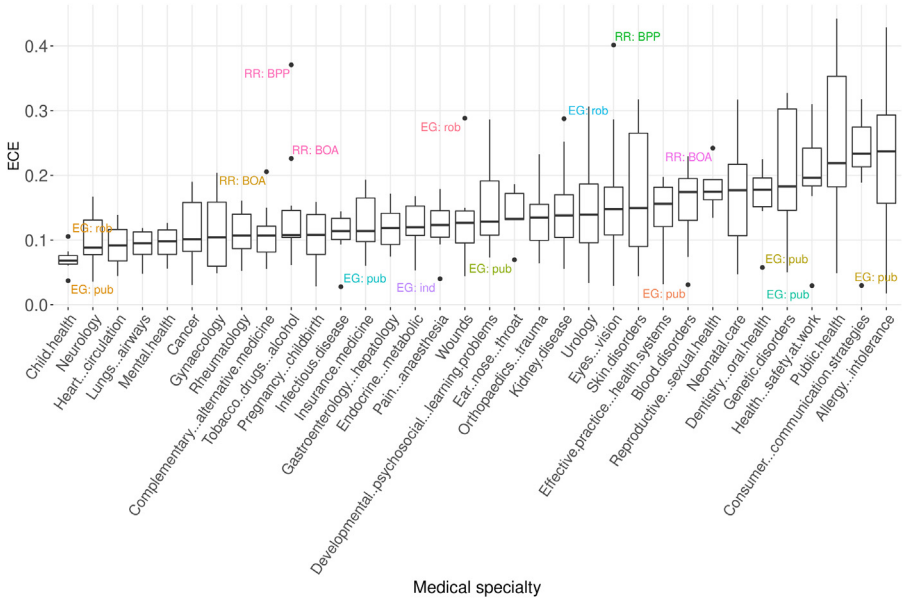

Fig. 5. Disparity of classifier reliability as measured by expected calibration error across different medical areas. Each boxplot represents all EvidenceGRADEr and RobotReviewer tasks (classifiers). ECE, expected calibration error; BPP, blinding of participants and personnel; BOA, blinding of outcome assessment; EG, EvidenceGRADEr; RR, RobotReviewer.

investigate the group-specific effects relevant in the assessment of the strength of evidence. We first extend our analysis of disparity to other categories beyond selective classification. We see in Figure 5 that the choice of a medical area to a large degree affects the *reliability* of the classifier as measured by ECE. When assessing the evidence within child health, neurology, and heart and circulation, the calibration error is very low (median ECE $<$ 0.1), while for other areas, such as allergy and intolerance, consumer and communication strategies, and bublic health, it is much higher (median ECE $>$ 0.2). While we have looked at this by aggregating over different models, there is noticeable variation depending on the model choice, as can be seen through the elongated boxplots for some areas and the presence of outliers, such as RobotReviewer's BPP task when assessing evidence for eyes and vision, or EvidenceGRADEr's RoB assessment for Wounds. In the same spirit, we plot F1 as a measure of *predictive performance* (Fig. 6). There is a 0.22 difference in the median F1 between the area with the most accurate assessment (reproductive and sexual health, 0.71 F1) and the least accurate one (wounds, 0.49 F1). The areas that we have found above to be the most or least calibrated, now also tend to have higher or lower F1, providing some indication that there is an interplay between reliability and predictive performance.[10]

3.4.1. Explanatory factors

We create a series of statistical models to find whether there is an association between certain data characteristics and different performance measures. We treat the proportion of higher-quality evidence (''proportion''), the number of data instances (''quantity''), as well as their interaction, as independent variables (fixed effects), and observe their effect on a performance measure (dependent variable). We include the measures for selective prediction (AUC-RCC), calibration error (ECE) and predictive performance (F1). Our observations are additionally influenced by two other factors, namely the chosen task–model combination and the medical area. To capture the random effects of these factors, we fit a linear mixed effect model [40]. We create a separate model for each dependent variable and assess the effect of proportion and quantity for significance.[11] When significant, the null hypothesis of no effect of either proportion or quantity on a performance measure is rejected, meaning there is a data-characteristic effect.

The results are shown in Table 4. First, we find that the disparity in *selective-prediction performance* is strongly associated with the proportion of higher-quality evidence available in a medical area. The relationship is negative, so the higher the proportion, the lower our estimate of the selective prediction trade-off (AUC-RCC). From the point of view of selective classification, lower AUC-RCC is better, meaning that those medical areas that display higher proportions of high-quality evidence tend to fare better in

[10] A Spearman correlation test confirms that per-area ECE and F1 scores are mostly negatively associated, but that the strength of this association varies per task–model combination. Coefficients: risk of bias (EG): −0.7, AC (RR): −0.5, publication bias (EG): −0.4, binary GRADE (EG): −0.4, imprecision (EG): −0.4, indirectness (EG): −0.4, RSG (RR): −0.3, BPP (RR): −0.2, BOA (RR): −0.2, inconsistency (EG): 0.3.

[11] We use the significance level of $\alpha = 0.05$ and correct for multiple comparisons using the Bonferroni method (for three models, $\alpha = 0.017$).

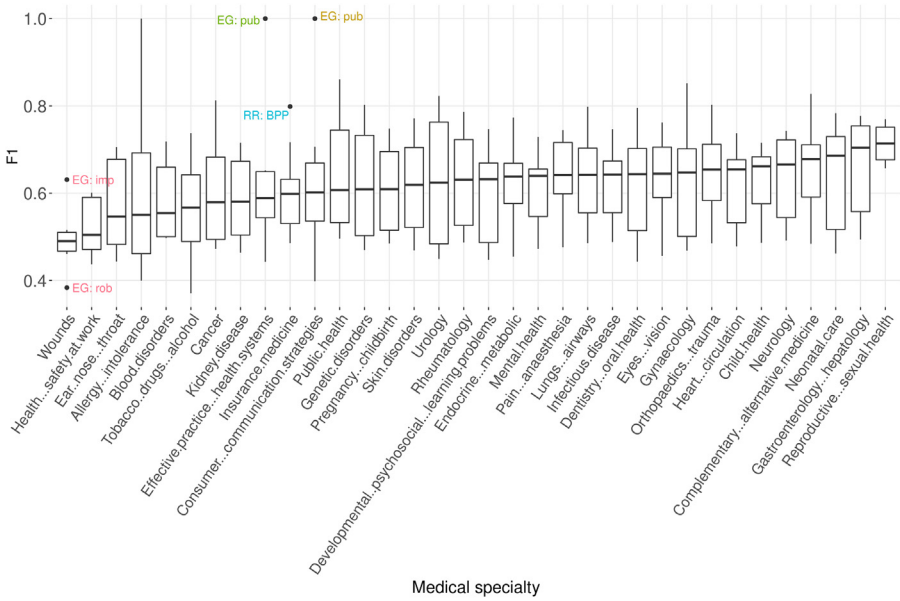

Fig. 6. Disparity of predictive performance as measured by F1 across different medical areas. Each boxplot represents all EvidenceGRADEr and RobotReviewer tasks (classifiers). BPP, blinding of participants and personnel; EG, EvidenceGRADEr; RR, RobotReviewer.

trading off coverage for risk. The role of quantity is less clear as the effect is on the brink of significance due to Bonferroni correction. Secondly, more data together with more higher-quality evidence are predictive of a decrease in calibration error (i.e., improved reliability). Thirdly, the unequal distribution of *predictive performance* (F1 across medical areas in Fig. 6) is explained well by both quantity and proportion. The direction here is positive, so more data and more high-quality evidence mean higher F1.

In our statistical model above, we have assumed that the quality labels are given, enabling us to calculate the proportion of higher-quality evidence per medical area. However, in practical applications on previously unseen data, such ground-truth labels will not be available, but users would still like to be aware of the risk of reduced calibration or selective-prediction performance. We therefore examine using model's predictions alone in quantifying such risk. We employ *variance* to obtain an estimate of the dispersion of output probabilities, and use it as a predictor of the three performance measures above. As before, we fit a mixed effect model with the same two random factors and treat variance as a fixed effect. Our results (Table 5) indeed show that variance strongly predicts the calibration and selective-prediction performance (but not the predictive performance as measured by F1). The direction of the found effect is negative, which means higher variance is associated with reduced ECE and AUC-RCC scores (desirable).[12] This means that an indication of low variance could be helpful in practical applications as a warning sign that the model's outputs are poorly calibrated and that confidence-based selection may not work as expected, requiring additional oversight from the user.

Finally, we turn our discussion of explanatory factors to a specific medical area, namely public health. It is characterized by a low proportion of high-quality evidence (33% of data instances have high or moderate GRADE scores; 36% have low RoB for RobotReviewer), high ECE (Fig. 5) and an F1 score that sits in the lower half of the medical areas (Fig. 6). Public health reviews are known to be dominated by lower levels of evidence, which in turn receive lower quality and recommendation grades [41]. This has to do with RCTs being too restrictive for some public health interventions, or infeasible due to cost,

Table 4. Statistical models for three different performance measures (dependent variables)

#	Dep. var.	Indep. var.	β	SE	t	P
		Prop	−1.12	0.07	−15.2	<0.001
1	AUC-RCC	Quant	−0.02	0.01	−2.4	0.016
		Prop:quant	0.04	0.01	2.8	0.006
		Prop	−0.33	0.07	−5.1	<0.001
2	ECE	Quant	−0.09	0.01	−10.7	<0.001
		Prop:quant	0.06	0.01	4.9	<0.001
		Prop	0.50	0.12	4.0	<0.001
3	F1	Quant	0.06	0.02	4.1	<0.001
		Prop:quant	−0.11	0.02	−4.9	<0.001

Abbreviations: β, fixed effect of the data characteristic on the outcome; SE, standard error; *t*, *t*-statistic; *P*, *P*-value for the model; AUC-RCC, area under the risk–coverage curve; Prop, proportion of higher quality evidence; Quant, quantity of evidence; Prop:quant, their interaction; ECE, expected calibration error.

[12] In our case, we can interpret the variance as a "tendency to certainty." It represents the dispersion from the absolute uncertainty (0.5), either toward the negative (0) or the positive class (1).

Table 5. Effect of variance of the output probabilities (independent variable) on different performance measures (dependent variables)

#	Dep. var.	Indep. var.	β	SE	t	P
1	AUC-RCC	variance	−1.58	0.15	−10.7	<0.001
2	ECE	variance	−0.90	0.13	−7.2	<0.001
3	F1	variance	−0.42	0.23	−1.8	0.07

Abbreviations: β, fixed effect of the independent variable; SE, standard error; *t*, *t*-statistic; *P*, *P*-value for the model; AUC-RCC, area under the risk–coverage curve; ECE, expected calibration error.

difficulties in controlling randomization, or other practical reasons [42]. Public health—and possibly other areas with similar considerations—would require a specially-adapted evidence grading methodology. An extension of National Institute for Health and Care Excellence guidelines for grading public health evidence has been proposed, which could be considered.[13]

3.5. A note on Cochrane RoB2

While our use of RobotReviewer is based on RoB1 annotations, newer guidance with several revisions to the original tool now exists (RoB2) [43]. Ideally, our findings should be revalidated once a larger amount of data is annotated with RoB2, which will allow retraining RobotReviewer.[14] We expect the main findings around disparity and its role in selective prediction to hold regardless of which tool version underlies RoB assessment: the amount of evidence per clinical question as well as the proportion of higher-quality evidence are likely to differ between medical areas independently of the choice of the RoB tool. In addition, we agree with [36] who anticipate that a portion of future reviews may still reasonably be expected to proceed with RoB1. For example, review updates may opt to use the same RoB tool as the original review, or large reviews with time constraints may prefer the earlier tool.[15]

4. Conclusion

How systematic reviewing technologies work in practice and when (not) to use them is often not clear to practitioners [8]. In this work, we hoped to bring some clarity to these questions by focusing on automated quality and bias assessment within systematic reviewing. We analyzed two systems for automatically grading the quality of evidence according to their reliability and selective classification performance. We show that these tools are already reasonably well calibrated on most tasks, which is important for practitioners seeking to integrate such systems in systematic reviewing workflows. Additionally, we identify and examine the issue of disparity for various measures of system performance, which becomes apparent when looking at the results by specific medical area. As medical evidence in different areas differs in quantity and quality, practitioners should be cognizant of the fact that this affects the reliability and the performance of quality assessment classifiers. Such effects remain under-explored for systematic reviewing technology more generally. Metrics defined over model's output probabilities, as we show, could be used by the practitioners as a defensive step. We leave the exploration of how to precisely implement such mechanisms for future work.

We believe that the current evidence appraisal tools cannot be applied fully independently and require oversight by domain experts. What form such supervision should take is open to debate. One possibility is that systems' most confident decisions are kept (following a selective classification approach) with little or no work to human reviewers, while the remaining decisions would be mere suggestions and would still require substantial human labor for revision. The system could also be considered as an additional annotator that works alongside humans. This scenario could lead to more robust consistency checking, and a model's prediction could sometimes even encourage human reviewers to reconsider their decisions. In the light of large discrepancies found in performance across different medical areas, the prospects for implementation of quality assessment in practice look better for certain medical areas, such as "neonatal care," "gastroenterology," and "reproductive health." To narrow the gap to the areas seeing lower performance, it would be worthwhile in the future to explore different data augmentation and balancing strategies since we have already found a link between the amount of medical evidence (or the prevalence of higher quality evidence) and different measures of performance. Additionally, the use of fairness-promoting techniques may help to equalize the performance across different areas of application [44].

Acknowledgments

We would like to thank Byron Wallace for assisting us with the evaluation of RobotReviewer, and the Cochrane organization for granting us the database of systematic reviews that formed the basis of our analysis.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2023.04.006.

Selective classification

Here, we provide further details about calculating the risk–coverage curves and their corresponding AUC-RCC.

[13] https://www.nccmt.ca/knowledge-repositories/search/28.

[14] The switch to RoB2 is less likely to impact our observations involving EvidenceGRADEr, since four (out of five) criteria involve aspects of quality other than risk of bias, and the RoB criterion in GRADE represents only an overall RoB assessment.

[15] Time use may increase as the tool becomes more refined [48,49].

In our case, we are equally interested in the classification of negative (lower quality of evidence/higher RoB) and positive (higher quality of evidence/lower RoB) examples. A model's prediction probability after the application of the sigmoid function represents the positive class whenever it is greater or equal to 0.5 and the negative class otherwise. The most confident predictions are those that lie at the extremities of the probability interval, that is, closer to 0 for negative- and closer to one for positive-class predictions, with the value of 0.5 representing the utmost uncertainty. To obtain the risk−coverage curves, we change τ_{pos} in increments of $1/50$ in the positive direction, starting at 0.5, and let $\tau_{neg} = 1 - \tau_{pos}$. Concretely, when, for example, $\tau_{pos} = 0.6$, all predictions greater or equal to τ_{pos} are kept, as well as all predictions below τ_{neg} (0.4). When $\tau_{pos} = 0.5$, τ_{neg} is also 0.5, which is equivalent to performing no selection on the model's outputs, leaving the coverage uncompromised. For other thresholding values, a portion of data is rejected to the advantage of risk reduction, namely all predictions $\tau_{neg} \geq \widehat{q}_i < \tau_{pos}$. In effect, the selectivity increases for both positive and negative classes in equal steps.

EvidenceGRADEr

We use the data splits created by [14] who followed a 10-fold cross validation during development; we therefore include in our analysis the model predictions made on all 10 test subsets.

Disparity: explanatory factors

We use the lme4 package [51] in R [52] together with the lmertest package [53] to provide *P* values for mixed effects models.

References

[1] Higgins JPT, James T, Chandler J, Miranda C, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. Chichester, UK: John Wiley & Sons; 2019.

[2] Sackett DL, Rosenberg WMC, Muir Gray JA, Brian Haynes R, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71.

[3] Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ open 2017; 7(2):e012545.

[4] Chalmers I. The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. Ann New York Acad Sci 1993;703(1):156−65.

[5] Higgins JPT, Altman DG, Gøtzsche PC, Peter J, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

[6] Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? a survival analysis. Ann Intern Med 2007;147:224−33.

[7] Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev 2014;3(1):1−15.

[8] Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev 2019;8(1):1−10.

[9] Clark J, Paul G, Mar CD, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol 2020;121:81−90.

[10] Marshall IJ, Kuiper J, Wallace BC. Automating risk of bias assessment for clinical trials. IEEE J Biomed Health Inform 2015;19(4): 1406−12.

[11] Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. J Am Med Inform Assoc 2015;23:193−201.

[12] Millard LAC, Flach PA, Higgins JPT. Machine learning to assist risk-of-bias assessments in systematic reviews. Int J Epidemiol 2015;45: 266−77.

[13] Sarker A, Mollá D, Paris C. Automatic evidence quality prediction to support evidence-based decision making. Artif Intelligence Med 2015;64:89−103.

[14] Šuster S, Baldwin T, Jey Han L, Yepes AJ, Martinez Iraola D, Otmakhova Y, et al. Automating quality assessment of medical evidence in systematic reviews: model development and validation study. J Med Internet Res 2023;25(e35568).

[15] Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. 3615−3620, Hong Kong, China. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics Available at: https://www.aclweb.org/anthology/D19-1371.

[16] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1. Minneapolis, Minnesota: Association for Computational Linguistics; 2019:4171−86.

[17] Soboczenski F, Trikalinos TA, Kuiper J, Bias RG, Wallace BC, Marshall IJ. Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. BMC Med Inform Decis Making 2019;19:96.

[18] Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as ai say: susceptibility in deployment of clinical decision-aids. NPJ digital Med 2021;4(1):1−8.

[19] O'Connor AM, Guy T, James T, Paul G, Gilbert SB, Hutton B. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? Syst Rev 2019;8(1): 1−8.

[20] Gates A, Vandermeer B, Hartling L. Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the robotreviewer machine learning tool. J Clin Epidemiol 2018;96:54−62.

[21] Gaertig C, Simmons JP. Do people inherently dislike uncertain advice? Psychol Sci 2018;29(4):504−20.

[22] Desai S, Durrett G. Calibration of pre-trained transformers. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 295−302. Association for Computational Linguistics Available at: https://aclanthology.org/2020.emnlp-main.21.

[23] Jiang X. Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. J Am Med Inform Assoc 2012;19:263−74.

[24] Guo C, Pleiss G, Sun Y, Weinberger KQ, On calibration of modern neural networks. In International Conference on Machine Learning, pages 1321−1330. Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia, PMLR 70; 2017.

[25] Brier GW. Verification of forecasts expressed in terms of probability. Monthly weather Rev 1950;78(1):1−3.

[26] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 2007;102:359−78.

[27] Kumar A, Liang PS, Ma T. Verified uncertainty calibration. Adv Neural Inf Process Syst 2019;32.
[28] Chow CK. An optimum character recognition system using decision functions. IRE Trans Electron Comput 1957;EC-6(4):247−54.
[29] El-Yaniv R. On the foundations of noise-free selective classification. J Machine Learn Res 2010;11(5):1605−41.
[30] Geifman Y, El-Yaniv R. Selective classification for deep neural networks. Adv Neural Inf Process Syst 2017;4885−94.
[31] Ding Y, Liu J, Xiong J, Shi Y. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops:4−5.
[32] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924−6.
[33] Zhang GP. Neural networks for classification: a survey. IEEE Trans Syst Man, Cybernetics, C (Applications Reviews) 2000;30(4):451−62.
[34] Armijo-Olivo S, Craig R, Campbell S. Comparing machine and human reviewers to evaluate the risk of bias in randomized controlled trials. Res Synth Methods 2020;11(3):484−93.
[35] Arno A, James T, Wallace B, Marshall IJ, McKenzie JE, Elliott JH. Accuracy and efficiency of machine learning−assisted risk-of-bias assessments in 'real-world' systematic reviews. Ann Intern Med 2022;175(7):1001−9.
[36] Jardim PSJ, Rose CJ, Ames HM, Echavez JFM, van de Velde S, Muller AE. Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. BMC Med Res Methodol 2022;22:1−12.
[37] Vinkers CH, Lamberink HJ, Tijdink JK, Heus P, Bouter L, Paul G, et al. The methodological quality of 176,620 randomized controlled trials published between 1966 and 2018 reveals a positive trend but also an urgent need for improvement. PLoS Biol 2021;19(4):e3001162.
[38] Marshall IJ, Benjamin N, Kuiper J, Noel-Storr A, Marshall R, Maclean R, et al. Trialstreamer: a living, automatically updated database of clinical trial reports. J Am Med Inform Assoc 2020;27:1903−12.
[39] Zhang Y, Marshall I, Wallace BC. Rationale-augmented convolutional neural networks for text classification. Proc Conf Empirical Methods Nat Lang Process Conf Empirical Methods Nat Lang Process 2016. p.795−804.
[40] Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 2008;59(4):390−412.
[41] Weightman AL, Ellis S, Cullum A, Sander L, Turley RL. Grading evidence and recommendations for public health interventions: developing and piloting a framework. London, UK: Health Development Agency; 2005.
[42] Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. Am J Public Health 2004;94:400−5.
[43] Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. Rob 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898.
[44] Han X, Shen A, Cohn T, Baldwin T, Frermann L. Systematic evaluation of predictive fairness. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 1 2022: Association for Computational Linguistics. Available at: https://aclanthology.org/2022.aacl-main.6.
[45] Šuster S. Robotreviewer evaluation data (new test set). 2022. Available at http://dx.doi.org/10.5281/zenodo.6908146. Accessed June 2, 2023.
[46] Šuster S, Baldwin T, Jey Han L, Yepes AJ, Martinez Iraola D, Otmakhova Y, et al. EvidenceGRADEr dataset. 2021. Available at http://dx.doi.org/10.5281/zenodo.5653587. Accessed June 2, 2023.
[47] Schunemann H. GRADE handbook for grading quality of evidence and strength of recommendation. 2008. Available at: http://www.cc-ims.net/gradepro.
[48] Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol 2013;66:973−81.
[49] Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol 2020;126:37−44.
[50] Nixon J, Dusenberry MW, Zhang L, Jerfel G, Tran D. Measuring calibration in deep learning. CVPR Workshops; 2019.
[51] Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. arXiv 2014.
[52] R Core Team. RA language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
[53] Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Softw 2017;82:1−26.