

1 ChatGPT for assessing risk of bias of randomized trials using the RoB

2 2.0 tool: A methods study

3

4 Tyler Pitre

5 *Division of Respiriology, Department of Medicine*

6 *University of Toronto*

7 Tanvir Jassal

8 *Departments of Anesthesia*

9 *McMaster University, Hamilton, ON*

10 Jhalok Ronjan Talukdar

11 *Departments of Anesthesia and Health Research Methods, Evidence, and Impact*

12 *McMaster University, Hamilton, ON*

13 Mahnoor Shahab

14 *Faculty of Health Sciences*

15 *McMaster University, Hamilton, ON*

16 Michael Ling

17 *Departments of Anesthesia*

18 *McMaster University, Hamilton, ON*

19 Dena Zeraatkar*

20 *Departments of Anesthesia and Health Research Methods, Evidence, and Impact*

21 *McMaster University, Hamilton, ON*

22

23 *Corresponding author:

24 **Disclaimers:** None.

25 **Funding:** None.

26 **Data:** Available on OSF (<https://osf.io/aq85p>)

27 **Acknowledgements:** None.

28 **Authors' Contributions:** DZ and TP conceived this study. TJ, JRT, MS, and ML collected data. DZ and TP

29 analyzed the data. DZ and TP wrote the first draft of the manuscript and all authors reviewed and

30 approved the final version.

31 **Word count:** 5,918

32 **Tables:** 3

33 **Figures: 3**

34 **Abstract**

35 **Background:** The assessment of risk of bias is a critical component of systematic review methods.
36 Assessing risk of bias, however, can be time- and resource-intensive. AI-based solutions may increase
37 efficiency and reduce burden.

38 **Objective:** To evaluate the reliability of ChatGPT for performing risk of bias assessments of randomized
39 trials.

40 **Methods:** We sampled recently published Cochrane systematic reviews of medical interventions (up to
41 October 2023) that included randomized controlled trials and assessed risk of bias using the Cochrane-
42 endorsed revised risk of bias tool for randomized trials (RoB 2.0). From each eligible review, we collected
43 data on the risk of bias assessments for the first three reported outcomes. Using ChatGPT-4, we
44 assessed the risk of bias for the same outcomes using three different prompts: a minimal prompt
45 including limited instructions, a maximal prompt with extensive instructions, and an optimized prompt
46 that was designed to yield the best risk of bias judgments. The agreement between ChatGPT's
47 assessments and those of the systematic reviewers was quantified using weighted kappa statistics.

48 **Results:** We included 34 systematic reviews with 157 unique trials. We found the agreement between
49 ChatGPT and systematic review authors for assessment of overall risk of bias to be 0.16 (95% CI: 0.01 to
50 0.3) for the maximal ChatGPT prompt, 0.17 (95% CI: 0.02 to 0.32) for the optimized prompt, and 0.11
51 (95% CI: -0.04 to 0.27) for the minimal prompt. For the optimized prompt, agreement ranged between
52 0.11 (95% CI: -0.11 to 0.33) to 0.29 (95% CI: 0.14 to 0.44) across risk of bias domains, with the lowest
53 agreement for the deviations from the intended intervention domain and the highest agreement for the
54 missing outcome data domain.

55 **Conclusion:** Our results suggest that ChatGPT and systematic reviewers only have “slight” to “fair”
56 agreement in risk of bias judgments for randomized trials. ChatGPT is currently unable to reliably assess
57 risk of bias of randomized trials. We recommend systematic reviewers avoid using ChatGPT to perform
58 risk of bias assessments.

59

60 **Background**

61 The practice of evidence-based medicine demands knowledge of the best available evidence, which
62 most often comes from rigorous systematic reviews and meta-analyses (1). Systematic reviews,
63 however, are time- and resource-intensive. Empirical evidence suggests they typically require upwards
64 of one year to complete and publish and many are outdated at or shortly following publication (2, 3).

65 One particular time- and resource-intensive component of systematic reviews is the assessment of risk
66 of bias of primary studies—defined as the propensity for studies to systematically over- or
67 underestimate treatment effects (4). Risk of bias assessments are burdensome and time-consuming and
68 demand specialized training. Moreover, to reduce the opportunity for errors, guidance for conducting
69 rigorous systematic reviews typically suggests authors assess risk of bias independently and in duplicate,
70 adding to the complexity and workload of the process (4).

71 In 2019, a new risk of bias tool was introduced that built on the successes of the previous Cochrane
72 endorsed risk of bias tool but also incorporated new advancements (5). This tool was called the revised
73 tool for assessing risk of bias of randomized trials (RoB 2.0) and has now become the gold standard (4).
74 The RoB 2.0 tool rates risk of bias as either high, some concerns, or low across five domains:
75 randomization, deviations from intended intervention, missing outcome data, measurement of
76 outcome, and selective reporting. The overall rating of risk of bias is determined by the domain rated at
77 highest risk of bias.

78 While the RoB 2.0 tool builds off a decade's worth of experience with the original risk of bias tool, recent
79 evidence suggests that reviewers find it more complex and time-consuming (6, 7). Innovations to
80 streamline and simplify risk of bias assessments without compromising their rigor will reduce the time
81 and effort required to perform systematic reviews and aid in maintaining their currency.

82 Previous efforts to streamline and automate risk of bias assessments have shown optimistic results (8-
83 13), suggesting that such endeavors may be feasible. For example, RobotReviewer is an automated tool
84 to extract data from and assess the risk of bias of randomized trials (8, 11, 12). The RobotReviewer,
85 however, was trained on the original Cochrane risk of bias tool and only offers judgments on four of the
86 seven domains of the original tool.

87 ChatGPT (OpenAI, San Francisco, California, USA) is a conversational artificial intelligence (AI) large
88 language model with capabilities in natural language processing and realization (14). Unlike specialized
89 tools for risk of bias assessments, ChatGPT is a general purpose tool, has been developed to emulate
90 human language rather than risk of bias assessments, and has been trained on an internet-scale corpus
91 covering many areas of knowledge, rather than a small training set focused on evidence synthesis and
92 evaluation (14).

93 Nevertheless, ChatGPT has been shown to perform remarkable tasks many of which are similar to
94 performing risk of bias assessments, including passing the United States Medical Licensing exams (15),
95 performing accurate diagnoses (16), and offering medical advice comparable to physicians (17). Further,
96 ChatGPT has been able to construct reasonable search strategies for systematic reviews (18) and other
97 tasks for which it was not intentionally designed (19), suggesting that it may also be able to assess risk of
98 bias despite not originally being designed for this task.

99 This study evaluates the performance of ChatGPT, an AI-based language model, for assessing risk of bias
100 of randomized trials using the RoB 2.0 tool. To do this, we sampled Cochrane systematic reviews using
101 the RoB 2.0 tool and had ChatGPT assess the risk of bias of the trials within these reviews. We compared
102 ChatGPT's assessment with those presented in Cochrane reviews. Consistency in assessments of risk of
103 bias between ChatGPT and Cochrane reviewers will suggest that ChatGPT can provide a reliable
104 assessment of the risk of bias of randomized trials. Conversely, discrepancies in risk of bias assessments
105 between ChatGPT and Cochrane reviewers will suggest that ChatGPT is unreliable for assessing risk of
106 bias.

107 **Methods**

108 We registered our protocol on Open Science Framework (<https://osf.io/aq85p>) in September 2023. We
109 report our study according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses
110 (PRISMA) and Guidelines for Reporting Reliability and Agreement Studies (GRRAS) reporting checklists
111 (20, 21).

112 This study does not involve human participants and is thus exempt from ethics review.

113 Figure 1 presents an overview of our methods.

114 ***Search strategy and screening***

115 For this study, we intended to include a reasonably representative sample of Cochrane systematic
116 reviews. We did not perform a search of medical research databases. Instead, we used the Cochrane
117 Database of Systematic Reviews (CDSR) that provides a chronological catalogue of published and
118 updated Cochrane systematic reviews to identify eligible reviews.

119 Reviewers worked independently and in duplicate to screen Cochrane reviews for eligibility, starting
120 with the most recently published (August 2023) and working backwards in time. We preferentially
121 included the most recently published Cochrane systematic reviews since these reviews are most likely to
122 have used the most up-to-date version of the RoB 2.0 tool instead of preliminary pilot versions of the
123 tool (5). Reviewers continued screening until we had identified our target sample size of approximately
124 160 trials.

125 ***Eligibility criteria***

126 Our sampling approach was designed to include randomized trials addressing a diverse range of
127 questions (i.e., selected from different systematic reviews) and both dichotomous and continuous
128 outcomes.

129 We included newly published or updated Cochrane systematic reviews addressing the benefits and/or
130 harms of health interventions that included one or more parallel randomized trials and reported
131 consensus-based risk of bias judgments using the Cochrane-endorsed RoB 2.0 tool (5). We define
132 consensus-based as two reviewers agreeing on the final risk of bias judgments. This may involve two
133 reviewers independently assessing risk of bias and resolving conflicts by discussion or a reviewer
134 assessing risk of bias and a second reviewer confirming the first reviewers' judgments.

135 We excluded systematic reviews that were not published by Cochrane, since such reviews may not
136 involve reviewers with sufficient training to appropriately apply the RoB 2.0 tool. We also excluded
137 Cochrane systematic reviews that investigated prognosis or the performance of diagnostic tests and

138 systematic reviews that only include observational studies since these reviews will necessitate the use of
139 other risk of bias tools.

140 Cochrane systematic reviews use summary of findings tables to present their results (4, 22). These tables
141 list outcomes in order of importance, the number of trials and patients that contributed data to the
142 meta-analysis for each outcome, the relative and absolute effect estimates based on meta-analyses, and
143 judgments about the certainty of evidence (4, 22). From each eligible review, we selected the first two
144 listed outcomes (suggesting that they are the most important) that were informed by one or more trials.
145 If either of the first two outcomes were continuous, we then selected the third outcome listed in the
146 summary of findings table. If the two reported outcomes were both dichotomous, we then selected the
147 first listed continuous outcome reported in the summary of findings table. When summary of findings
148 tables reported on the same outcome at different timepoints, we selected entirely unique outcomes.

149 From each review, we included all parallel randomized trials published in English that were included in
150 analyses addressing the outcomes of interest. We excluded crossover and cluster randomized trials
151 since these trial designs require unique considerations in their assessment of risk of bias and different
152 versions of the RoB 2.0 tool.

153 Cochrane reviews often include unpublished trial data. When reviews reported that information for a
154 particular trial was unpublished or was drawn from a combination of unpublished and published data,
155 we excluded those trials since we did not have access to the same unpublished information as the
156 Cochrane reviewers for risk of bias assessments. For feasibility, we also excluded trials for which data
157 was drawn from multiple publications. Including such trials would have necessitated an exhaustive
158 review of all related publications to identify those containing the outcome data and the comprehensive
159 details required for risk of bias assessment.

160 ***ChatGPT prompts***

161 A key component in the use of ChatGPT is the design of the text used to instruct the model (called
162 ‘prompts’) to generate an answer. We anticipated that ChatGPT’s risk of bias judgments may depend on
163 the nature of the prompts that it is provided. To study how different prompts may influence risk of bias
164 judgments, we iteratively designed three different prompts: a minimal prompt including limited
165 instructions for assessing risk of bias, a maximal prompt with extensive instructions, and an optimized
166 prompt that was designed to include sufficient information to yield the best risk of bias judgments.

167 We piloted the prompts using 15 trials drawn from systematic reviews previously performed by our own
168 team and refined the prompts by iterative discussion and input by the co-authors (23-25). All prompts
169 asked ChatGPT to judge risk of bias for all RoB 2.0 domains (bias due to randomization, deviation from
170 intended intervention, missing outcome data, measurement of outcome, and selective reporting) as low
171 risk of bias, some concerns, or high risk of bias—consistent with RoB 2.0 guidance (5). Supplement 1
172 presents these three prompts.

173 The RoB 2.0 tool is accompanied by a document that describes the tool and offers guidance on its
174 implementation. All three prompts included the RoB 2.0 full guidance document ([riskofbias.info](https://www.riskofbias.info)), which
175 were fed to ChatGPT using the AskYourPDF ChatGPT plugin that allows ChatGPT to read and query PDF
176 documents. All prompts also included a PDF copy of the trial publication, a PDF copy of the trial
177 registration or protocol (if one was available), and specified the outcome of interest for which risk of
178 bias assessment was being performed.

179 The RoB 2.0 tool offers two options for assessing the risk of bias due to deviations of the intended
180 intervention: one for the effect of assignment to the intervention and the other for the effect of
181 adhering to the intervention. In Cochrane systematic reviews, the subsection on risk of bias typically
182 reports whether Cochrane reviewers assessed risk of bias for the effect of assignment or adherence to
183 the intervention. Our ChatGPT prompts also specified whether to assess risk of bias for the effect of
184 assignment or adherence to the intervention, depending on the option selected by the Cochrane review
185 authors. For systematic reviews that failed to specify whether they assessed risk of bias for the effect of
186 being assigned to the intervention or adherence to the intervention, we assumed they assessed risk of
187 bias for assignment to the intervention.

188 The ChatGPT prompts do not include any information related to the consensus-based risk of bias
189 judgments presented in the systematic reviews. Hence, ChatGPT is 'blind' to the risk of bias judgments
190 that are presented in the review.

191 ***Data collection***

192 RoB 2.0 guidance demands that reviewers perform risk of bias judgments for each particular result
193 rather than each trial or outcome, since risk of bias may differ across outcomes in a trial or across
194 different ways of statistically summarizing the results for the same outcome (5). We took this approach
195 in this study. For each eligible trial and outcome, we collected information on the consensus-based risk
196 of bias judgments presented in the Cochrane systematic reviews. Subsequently, for each eligible trial,
197 we used the ChatGPT-4 chatbot to assess the risk of bias of the outcomes of interest, using each of the
198 three ChatGPT prompts. ChatGPT-4 is a more advanced iteration of its predecessor ChatGPT-3. Unlike
199 ChatGPT-3, ChatGPT-4 is only available with a paid subscription to OpenAI. We implemented each of the
200 prompts in unique chats.

201 We did not collect data in duplicate because the nature of the data did not require any subjective
202 judgments and we anticipated that the only potential source of error is mistakes in copying and pasting
203 prompts to the ChatGPT interface, which we deemed unlikely.

204 We anticipated that the reliability of ChatGPT may depend on the objectivity of the outcome for which
205 risk of bias is being assessed. We considered outcomes objective if they were based on established
206 laboratory measures or if they were not subject to interpretation by patients or healthcare providers.
207 Conversely, we considered outcomes subjective if they were patient-reported or subject to
208 interpretation by patients or healthcare providers. We classified outcomes as either definitely objective
209 (e.g., mortality), probably objective (e.g., unscheduled physician visits), probably subjective (e.g., serious
210 adverse events), and definitely subjective (e.g., quality of life) to facilitate stratified analyses based on
211 the degree of objectivity of the outcome.

212 ***Data synthesis and analysis***

213 Sample size estimation

214 We used the kappaSize package in R (Vienna, Austria, Version 4.1.3) to estimate sample size (26). We
215 aimed to calculate the number of required trials to obtain a sufficiently precise estimate of a value of
216 kappa for which systematic reviewers will feel confident using ChatGPT for risk of bias assessments. We
217 assumed that most reviewers will feel confident using ChatGPT for risk of bias assessments if it yields a
218 kappa of 0.70, indicating substantial agreement, with the lower bound of the confidence interval no less

219 than 0.55. We anticipated the risk of bias distribution to be approximately 30% low, 30% with some
220 concerns, and 40% high.

221 We inflated the estimated sample size by a design effect to account for correlation between the risk of
222 bias of trials from the same review. We assumed an intra-review correlation of 0.05 and an average of
223 10 trials per review, yielding a design effect of 1.45. This resulted in a minimum sample size of 120 trials
224 from 12 reviews. We investigated the sensitivity of our estimated sample size to different assumptions
225 about the anticipated distribution of risk of bias judgments across the three categories and the potential
226 correlation between trials from the same review. To account for other potential scenarios (e.g., kappa =
227 0.6, intrareview correlation of 0.1), we ultimately intended to include approximately 160 trials from 16
228 reviews.

229 Agreement between ChatGPT and consensus-based risk of bias assessments

230 We present the inter-rater agreement, represented by weighted kappa, between each of the three
231 ChatGPT prompts and consensus-based risk of bias judgments from Cochrane authors. Unlike
232 percentage agreement, the weighted kappa accounts for the possibility of agreement due to chance and
233 for the ordinal nature of the response options of the RoB 2.0 tool (low risk of bias, some concerns, high
234 risk of bias) (27).

235 We present separate analyses for each RoB 2.0 domain and for the overall rating of risk of bias. Each
236 analysis only includes one outcome from each included trial. Our primary analysis includes the most
237 important outcome, based on the order in which outcomes were listed in Cochrane systematic review
238 summary of findings tables. We adjusted for clustering of trials within each systematic review by
239 inflating the variance of all estimates by the design effect (28).

240 We interpreted Cohen's kappa statistics using previously established guidelines: values from 0.0 to 0.2
241 indicating slight agreement, 0.21 to 0.40 indicating fair agreement, 0.41 to 0.60 indicating moderate
242 agreement, 0.61 to 0.80 indicating substantial agreement, and 0.81 to 1.0 indicating perfect agreement
243 (29).

244 We hypothesized that ChatGPT may be more reliable to assess risk of bias when there are few subjective
245 judgments. Therefore, we expected better agreement for: (i) trials addressing pharmacologic
246 interventions because trials of pharmacologic interventions are more likely to blind patients and
247 healthcare providers thus simplifying judgments related to deviations from intended intervention and
248 measurement of outcomes; (ii) trials addressing risk of bias of assignment of the intervention because
249 assignment to the intervention does not necessitate making judgments about adherence; (iii) objective
250 outcomes since these outcomes do not need additional judgments about whether failure to blind may
251 have resulted in differential measurement of the outcome, and (iv) dichotomous instead of continuous
252 outcomes since continuous outcomes are more likely to be subjective. To test these hypotheses, we
253 performed secondary analyses stratified by these factors.

254 We also performed a secondary analysis in which we collapsed ratings of "some concerns" and "high risk
255 of bias" into a single category.

256 In our primary analysis, we excluded ratings of uncertain risk of bias from analyses. We had planned to
257 perform additional sensitivity analyses treating these ratings as some concerns or high risk of bias but
258 there were too few uncertain ratings to affect estimates of reliability.

259 We performed all statistical analyses using the psych package in R (Vienna, Austria, Version 4.1.3) (30).

260 Review of ChatGPT justifications for discrepant risk of bias judgments between Cochrane systematic 261 reviewers and ChatGPT

262 Our prompts queried ChatGPT to provide a justification for its ratings of risk of bias. To understand
263 reasons why ChatGPT may produce unreliable risk of bias judgments, we also qualitatively reviewed
264 justifications provided by ChatGPT to support its judgments for potential errors or problems.

265 ***Deviations from protocol***

266 To account for correlation between trials in the same systematic review, we planned to calculate
267 weighted kappa within each review individually and pool the weighted kappa statistics across systematic
268 reviews using random-effects meta-analysis (31). The sampling distribution of kappa, however, is
269 asymmetric. While with a large enough number of observations, the sampling distribution of kappa is
270 approximately normal, we found there to be too few trials within each systematic review to assume
271 normality, precluding our approach to perform meta-analyses. Instead, we adjusted the variance of all
272 estimates for the correlation within each systematic review.

273 **Results**

274 ***Systematic review and trial characteristics***

275 We included 157 trials from 34 systematic reviews. Figure 2 presents the selection of systematic
276 reviews. Supplement 2 presents a list of included reviews and supplement 3 presents a list of excluded
277 reviews.

278 More than half of reviews were published in 2023 and addressed pharmacologic interventions. Reviews
279 most commonly addressed infectious, ophthalmologic, and respiratory conditions. Reviews either rated
280 the risk of bias for assignment to the intervention or did not report whether they assessed the risk of
281 bias of assignment to or adherence to the intervention. More than half of included outcomes were
282 dichotomous and rated as either definitely or probably objective.

283 In our analyses, each trial contributed data only for one outcome. Our primary analysis included data
284 from 157 trials. Of these, 45 (28.7%) were rated at low risk of bias by Cochrane systematic reviewers, 75
285 (47.8%) at some concerns, and 37 (24.6%) at high risk of bias. Fifty-two trials (33.1%) were rated at high
286 risk of bias or some concerns for bias due to randomization, 37 (23.6%) for bias due to deviations from
287 the intended intervention, 23 (14.7%) for missing outcome data, 29 (18.5%) for measurement of the
288 outcome, and 72 (45.9%) for selective reporting.

289 ***Agreement between ChatGPT and consensus-based risk of bias judgments from Cochrane review*** 290 ***authors***

291 In our analyses, each trial contributed data only for one outcome. In our primary analysis, when a trial
292 reported data on more than one outcome of interest, we included data for the outcome reported first in
293 the systematic review.

294 We found overall only slight agreement between ChatGPT risk of bias judgments and consensus-based
295 risk of bias judgments from systematic reviewers. Agreement for overall risk of bias ranged between
296 0.11 (95% CI: -0.04 to 0.27) and 0.17 (95% CI: 0.02 to 0.32) for the minimal and optimized prompts,
297 respectively. Figure 2 presents a flow diagram representing categorical changes in the overall rating of
298 risk of bias between systematic reviewers and the optimized ChatGPT prompt.

299 For the optimized prompt, agreement ranged between 0.11 (95% CI: -0.11 to 0.33) to 0.29 (95% CI: 0.14
300 to 0.44) across risk of bias domains, with the lowest agreement for the deviations from the intended
301 intervention domain and the highest agreement for the missing outcome data domain.

302 We hypothesized that ChatGPT may be more reliable to assess risk of bias when there are few subjective
303 judgments: trials addressing pharmacologic interventions, reviews that assessed risk of bias of
304 assignment rather than adherence to the intervention, objective outcomes, and dichotomous outcomes.
305 To test these hypotheses, we performed secondary analyses stratified by these factors. We did not find
306 evidence that ChatGPT had importantly different reliability in these stratified analyses (Supplements 4 to
307 10). ChatGPT showed “slight” to “fair” agreement for these subgroups.

308 Likewise, we performed a secondary analysis in which we collapsed ratings of “some concerns” and
309 “high risk of bias” into a single category. This secondary analysis also showed “slight” to “fair”
310 agreement (Supplement 11).

311 ***Discrepant risk of bias judgments between Cochrane systematic reviewers and ChatGPT***

312 For all risk of bias judgments, our prompts queried ChatGPT to provide a justification for its rating of risk
313 of bias. To understand reasons why ChatGPT may produce unreliable risk of bias judgments, we also
314 qualitatively reviewed justifications provided by the optimized ChatGPT prompt to support its judgments
315 for potential errors or problems. An analysis of the justifications provided by ChatGPT suggests four
316 major types of problems.

317 First, it appears that ChatGPT could not distinguish between characteristics of trials that are at low risk
318 of bias and characteristics at high risk of bias. For example, one trial reported randomization by an
319 “interactive web-response system”, which suggests central randomization and allocation concealment
320 (32). The ChatGPT optimized prompt rates the trial at some concerns for randomization because the
321 trial report “does not explicitly mention whether the allocation sequence was concealed”. One trial
322 reported using “system-generated random numbers” to randomize participants (33). The ChatGPT
323 prompt rated risk of bias due to randomization at low risk of bias with the justification that “an open list
324 of random numbers for concealment” indicates “proper randomization process”—an incorrect
325 statement since an open list allows those recruiting participants in a trial to predict the arm to which
326 subsequent participants will be randomized.

327 Second, ChatGPT was unable to make reasonable assumptions about risk of bias. Cochrane systematic
328 reviewers rated a trial investigating the effects of convalescent plasma on all-cause mortality in COVID-
329 19 patients at low risk of bias for missing outcome data (34). ChatGPT rated the trial at some concerns
330 because it “does not provide explicit details about the availability of outcome data for all participants or
331 if there was significant dropout of participants”. The trial however reports that no patients were lost to
332 follow-up and it is reasonable to assume that all-cause mortality would be one of the outcomes for
333 which there would be no missing outcome data without loss to follow-up since it does not involve active
334 measurement or monitoring by investigators.

335 Third, ChatGPT made errors that suggested that it was unfamiliar with recommended processes for risk
336 of bias assessments. For example, for the domain bias due to deviations from the intended intervention
337 an open-label trial of aspirin for COVID-19 was judged at high risk of bias by systematic reviewers and
338 low risk of bias by ChatGPT because the outcome ‘all-cause mortality’ is objective (35). While the
339 outcome is objective, the domain of bias due to deviations from intended intervention is meant to solely

340 assess risk of bias due to imbalances in cointerventions or differences in how the intervention is
341 implemented rather than objectivity of the outcome, which is assessed by the bias due to measurement
342 of the outcome domain. Similarly, in making judgments about risk of bias due to selective reporting,
343 ChatGPT often considered discrepancies between all outcomes and results between the trial publication
344 or the registration or protocol instead of the results for which risk of bias was being assessed. Although
345 bias due to randomization should be consistent across outcomes from the same trial, we identified
346 instances in which ChatGPT rated the domain inconsistently across outcomes from the same trial.

347 Finally, ChatGPT made random errors in assessing risk of bias. For example, in another trial described as
348 double-blind and rated at low risk of bias due to deviations from intended intervention by systematic
349 reviewers, ChatGPT rated risk of bias as high because the trial “does not provide information on whether
350 participants and personnel were aware of the intervention” (36).

351 **Discussion**

352 ***Main findings***

353 We performed a study evaluating ChatGPT for assessing the risk of bias of randomized trials using the
354 Cochrane-endorsed RoB 2.0 tool (5). To do this, we sampled Cochrane systematic reviews that reported
355 RoB 2.0 judgments for randomized trials, assessed the risk of bias of trials using ChatGPT via three
356 variations of prompts, and compared the degree of agreement between RoB 2.0 judgments presented in
357 systematic reviews and those by ChatGPT.

358 We found only slight to fair agreement between ChatGPT risk of bias judgments and those presented in
359 systematic reviews. Our results suggest that ChatGPT, at least as it stands today, is suboptimal for
360 facilitating risk of bias assessments. We found similar results when we restricted our analysis to
361 subgroups for which we hypothesized that ChatGPT may be more reliable, including trials addressing
362 pharmacologic interventions, reviews assessing the risk of bias associated with assignment to the
363 intervention, objective outcomes, and dichotomous outcomes.

364 We also reviewed cases in which ChatGPT's risk of bias judgments differed from those of Cochrane
365 systematic reviewers with the goal of identifying ways in which we can refine future prompts. Our
366 findings indicate that ChatGPT might make more accurate risk of bias judgments if informed about both
367 low and high risk of bias methodological traits. For example, one trial reported randomization by an
368 “interactive web-response system”, which suggests central randomization and allocation concealment
369 (32). ChatGPT, however, rated the trial at some concerns for randomization because the trial report
370 “does not explicitly mention whether the allocation sequence was concealed”. Training ChatGPT to
371 recognize features of trials at low versus high risk of bias may improve the reliability of its risk of bias
372 assessments.

373 Though our results appear discouraging, they must also be contextualized considering general poor
374 agreement between even experienced reviewers in implementing the RoB 2.0 tool. For example, a
375 previous investigation of the reliability of RoB 2.0 using experienced systematic reviewers reported
376 inter-rater reliability ranging between 0.04 to 0.45, indicating only slight to fair agreement (7). The
377 original Cochrane risk of bias tool also demonstrated poor inter-rater reliability for select domains (37).

378 Our results may also be explained by ChatGPT's limited memory, which may not be sufficient to fully
379 process RoB 2.0's extensive and lengthy guidance (38, 39). An improvement in ChatGPT's performance

380 in risk of bias assessment might be achieved by enhancing its memory capabilities, by utilizing other
381 plans from OpenAI that offer expanded memory options such as ChatGPT Enterprise, or by fine-tuning
382 ChatGPT's base model—a process that involves additional training of the model.

383 Finally, while we evaluated the degree of agreement between risk of bias judgments reported in
384 systematic reviews and those made by ChatGPT, we did not consider the impact of these discrepancies.
385 For example, discrepancies in risk of bias judgments may not necessarily lead to an overall change in the
386 rating of the certainty (quality) of evidence and the material conclusions of systematic reviews.

387 ***Strengths and limitation***

388 The primary strength of our study is its generalizability to diverse research questions, reviews, and
389 research teams. Risk of bias judgments are subjective and different research groups and teams may
390 have different understandings and thresholds for expressing concerns about risk of bias. Similarly,
391 assessing risk of bias involves unique considerations related to the research question being investigated.
392 As our sample included systematic reviews from multiple diverse research teams, ChatGPT's reliability is
393 not confined to the specific nuances of a single group's approach to risk of bias assessments or to a
394 single topic.

395 Our study was limited to parallel randomized trials published in English. We excluded crossover and
396 cluster randomized trials since these trial designs require unique considerations in their assessment of
397 risk of bias and different versions of the RoB 2.0 tool. Thus, the results of our study may lack
398 generalizability beyond English language parallel randomized trials, though these are the most common
399 studies typically included in systematic reviews. Further, it is unlikely for ChatGPT to be able to perform
400 remarkably differently for other types of trials, since assessing the risk of bias of these trials necessitates
401 the same considerations as parallel randomized trials in addition to several additional unique
402 considerations.

403 Evidence suggests that risk of bias assessments in Cochrane reviews, despite their rigor, are sometimes
404 unreliable and inconsistent with established guidance (7). Hence, differences between risk of bias
405 judgments between ChatGPT and Cochrane systematic reviewers may also represent errors on part of
406 reviewers. Previous studies suggest that agreement between reviewers in assessing risk of bias may be
407 very poor (40, 41). To minimize the potential for this error, we limited our sample to Cochrane
408 systematic reviews, which are known for their methodological rigor (42, 43).

409 The performance of ChatGPT is also not static. The infrastructure, interfaces, and applications built
410 around ChatGPT are continuously updated. Our experiment was performed over a two-week time
411 period between September and October 2023. It is possible that the performance that we observed may
412 not be replicable in the future—though it is more likely that the capabilities of ChatGPT will improve
413 rather than deteriorate. Even with identical prompts, ChatGPT might provide slightly different answers
414 due to the inherent stochasticity in its response generation.

415 The reliability of ChatGPT risk of bias assessments is likely to depend on the nature of the prompts. We
416 tested three different prompts. Our results suggest that the performance of the three prompts is
417 comparable. It is possible that reviewers may be able to produce more reliable risk of bias assessments
418 using alternative prompts.

419 Our prompts queried ChatGPT to provide a justification for its ratings of risk of bias. To understand
420 reasons why ChatGPT may produce unreliable risk of bias judgments, we also reviewed justifications
421 provided by ChatGPT to support its judgments for potential errors or problems. While we performed a
422 general review of justifications for which ChatGPT and Cochrane reviewers made discrepant risk of bias
423 judgments, we did not perform a formal qualitative analysis of the justifications.

424 While we did not record the exact duration our team spent using ChatGPT, we estimate that each trial
425 took no longer than 15 minutes—less time than on average required for a reviewer to conduct an
426 individual risk of bias assessment and consensus meeting according to empirical evidence (6, 7).

427 ***Relation to previous findings***

428 Attempts to reduce the time, resources, and expertise needed to perform systematic reviews are not
429 new. For example, RobotReviewer is an automated tool to extract data from and assess the risk of bias
430 of randomized trials (8). The RobotReviewer, however, was trained on the original Cochrane risk of bias
431 tool and only offers judgments on four of the seven domains of the original tool. Since then, Cochrane
432 has adopted a revised risk of bias assessment tool that requires more nuanced judgments and is more
433 resource and time intensive (6). Given the performance of ChatGPT, however, adapting RobotReviewer
434 to provide risk of bias assessments using the RoB 2.0 tool may be more promising.

435 ***Implications***

436 The practice of evidence-based medicine demands knowledge of the best available evidence, which
437 most often comes from rigorous systematic reviews and meta-analyses (1). Systematic reviews are
438 resource and time intensive. For example, empirical evidence suggests that systematic reviews typically
439 require upwards of one year to complete and publish (2). Tools that efficiently and reliably conduct risk
440 of bias assessments can conserve time and resources, free reviewers to concentrate on other critical
441 tasks, and potentially enhance the accuracy of risk of bias judgments.

442 Our results suggest that ChatGPT, in its current form, is not able to reliably assess the risk of bias of
443 randomized trials. Since assessment of the risk of bias of observational studies or diagnostic studies is
444 even more complicated, it is reasonable to expect that ChatGPT might encounter even more challenges
445 with these other types of study designs.

446 Our study also has implications for future research. While our prompts in their current form could not
447 be used to reliably assess risk of bias, other prompts may be able to provide more reliable assessments.
448 For example, for each domain, RoB 2.0 contains a series of signaling questions designed to help
449 reviewers think systematically about the different aspects of trial conduct that might lead to bias. These
450 signaling questions are answered with "Yes," "Probably yes," "Probably no," "No," or "No information."
451 Based on the answers to these questions, a judgment is made about the risk of bias for that domain as
452 "Low," "Some concerns," or "High." Instead of asking ChatGPT to assess the risk of bias of each domain,
453 ChatGPT may be prompted to go through the RoB 2.0 signalling questions. Future research may address
454 the usefulness of having systematic reviewers reconcile their risk of bias assessments with ChatGPT or
455 the role of ChatGPT in training systematic reviewers.

456 There are also opportunities to use ChatGPT to streamline other aspects of systematic reviews. Early
457 studies suggest that ChatGPT can be used to devise search strategies (18). ChatGPT may also assist with
458 screening search records, extracting data from eligible studies, or performing evaluations of the
459 certainty of evidence. Though, at this time, based on the results of the current study, we are not

460 optimistic about ChatGPT's ability to reliably extract data or evaluate the certainty of evidence.
461 Screening studies is less subjective and perhaps better suited to ChatGPT's abilities.

462 If ChatGPT's performance improves or if other tools emerge that can reliably perform various systematic
463 review tasks, systematic review authors will need to consider whether the time and resource savings
464 afforded by these tools are worth potential suboptimal performance. While these tools may not always
465 perform perfectly, they may still be useful in situations in which systematic reviews need to be
466 performed quickly or with limited resources. Similarly, systematic review authors will also need to
467 consider the acceptability of such tools by evidence users. For example, evidence users may be skeptical
468 of systematic reviews that use AI tools.

469 There are ethical implications around the adoption of large language models, artificial intelligence, and
470 ChatGPT, in health research (44). Perhaps the most immediate ethical implication is the replacement of
471 systematic reviewers. Because evidence syntheses are used to make decisions about large numbers of
472 patients, incorrectly replacing human reviewers with an underperforming tool may have serious
473 negative health consequences. We caution against the adoption of ChatGPT for assessments of risk of
474 bias, particularly the replacement of reviewers. Our results suggest that ChatGPT performs poorly in
475 assessing risk of bias.

476 The integration of artificial intelligence and large language models in systematic reviews can also affect
477 trust in health research. We anticipate that due to limited experience, evidence users will be more
478 cautious about the application of studies that use such tools (45, 46).

479 There are also ethical issues in outsourcing important research functions to software developed and
480 operated by commercial entities located in foreign jurisdictions that may not be incentivized to ensure
481 that health decisions are free of conflicts of interest. Undue influence or attacks on artificial intelligence
482 systems by corporations, interest groups, and even hostile governments represent new threats against
483 which research should be protected (47, 48). Further, there is limited details on how ChatGPT works
484 internally, including model architecture and the training data. The benefits, risks, and costs of
485 outsourcing risk of bias assessments to software operated by commercial entities should be evaluated,
486 from the perspective of research resiliency and scientific accountability.

487 **Conclusion**

488 We performed a study evaluating the usefulness of ChatGPT for assessing the risk of bias of parallel
489 randomized trials using the Cochrane-endorsed RoB 2.0 tool. We found only slight to fair agreement
490 between ChatGPT risk of bias judgments and risk of bias judgments presented in systematic reviews.
491 Our results suggest that ChatGPT, at least as it stands today, is suboptimal for performing risk of bias
492 assessments. The practice of evidence-based medicine demands knowledge of the best available
493 evidence, which most often comes from rigorous systematic reviews. Systematic reviews, though, are
494 time and resource intensive. Tools to assist with systematic reviews, be it with risk of bias assessments
495 or other tasks, are critically needed.

496 **Tables**

Table 1: Characteristics of included systematic reviews

Publication year	
2022	12 (35.9%)
2023	22 (64.7%)
Type of intervention	
Pharmacologic	18 (52.9%)
Surgical	6 (17.6%)
Rehabilitation	1 (2.9%)
Lifestyle	4 (11.8%)
Other	5 (14.7%)
Type of condition	
Infectious diseases	9 (26.5%)
Ophthalmologic	7 (20.6%)
Respiratory	4 (11.8%)
Cardiac	2 (5.9%)
Psychiatric	2 (5.9%)
Gastrointestinal	2 (5.9%)
Injury and poisoning	1 (2.9%)
Pediatrics	1 (2.9%)
Cancer	1 (2.9%)
Endocrine	1 (2.9%)
Neurologic	1 (2.9%)
Other	3 (8.8%)
Type of risk of bias assessment	
Assignment to the intervention	24 (%)
Adherence to the intervention	0 (0%)
Not reported	10 (%)
Type of outcome*	
Dichotomous	179 (65.3%)
Continuous	95 (34.7%)
Subjectivity of outcomes*	
Definitely objective	108 (39.4%)
Probably objective	54 (19.7%)
Probably subjective	64 (23.4%)
Definitely subjective	48 (17.5%)
Number of trials included per systematic review	3 [2 to 7]
median [IQR]	

*For each review, we included data on more than one outcome.

497

Table 2: Degree of Agreement

		Consensus based risk of bias judgments reported in systematic reviews		
		Low risk of bias	Some concerns	High risk of bias
Optimized ChatGPT prompt	Low risk of bias	4 (2.55%)	2 (1.27%)	0 (0%)
	Some concerns	41 (26.11%)	71 (45.22%)	33 (21.02%)
	High risk of bias	0 (0%)	2 (1.27%)	4 (2.55%)
Minimal ChatGPT prompt	Low risk of bias	3 (1.91%)	5 (3.18%)	1 (0.64%)
	Some concerns	42 (26.75%)	66 (42.04%)	32 (20.38%)
	High risk of bias	0 (0%)	3 (1.91%)	4 (2.55%)
Maximal ChatGPT prompt	Low risk of bias	1 (0.64%)	2 (1.27%)	0 (0%)
	Some concerns	44 (28.03%)	72 (45.86%)	31 (19.75%)
	High risk of bias	0 (0%)	1 (0.64%)	6 (3.82%)

498

499

500

Table 3: Weighted kappa values representing the degree of agreement between ChatGPT prompts and systematic review risk of bias judgments

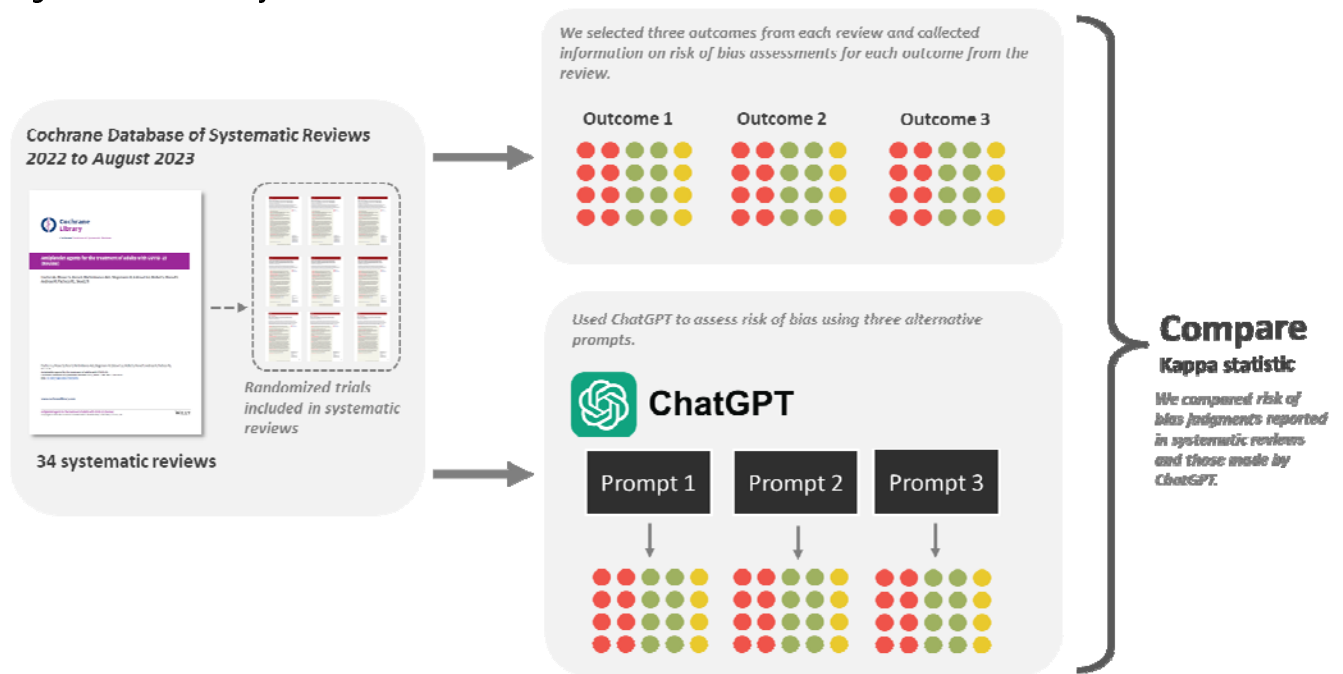
	Optimized prompt	Minimal prompt	Maximal prompt
	Weighted kappa (95% CI)		
Overall risk of bias rating	0.17 (0.02, 0.32)	0.11 (-0.04, 0.27)	0.16 (0.01, 0.3)
Risk of bias due to randomization	0.24 (0.02, 0.47)	0.09 (-0.16, 0.33)	0.09 (-0.15, 0.34)
Risk of bias due to deviations from the intended intervention	0.11 (-0.11, 0.33)	0.12 (-0.12, 0.37)	0.12 (-0.13, 0.36)
Risk of bias due to missing outcome data	0.29 (0.14, 0.44)	0.23 (0.02, 0.45)	0.16 (-0.05, 0.36)
Risk of bias due to measurement of the outcome	0.14 (-0.13, 0.41)	0.04 (-0.18, 0.25)	0.05 (-0.18, 0.28)
Risk of bias due to selective reporting	0.17 (-0.03, 0.37)	0.29 (0.08, 0.49)	0.21 (0.04, 0.37)

501

502

503 **Figures**

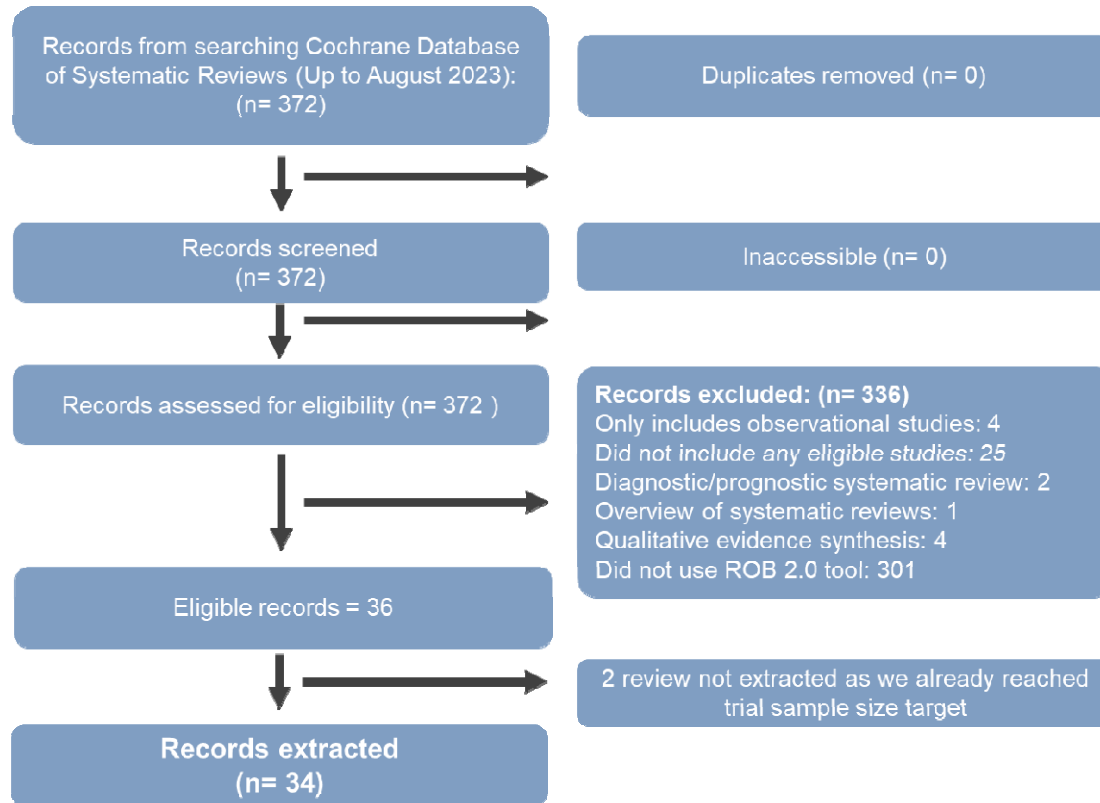
504 **Figure 1: Overview of methods**



505

506

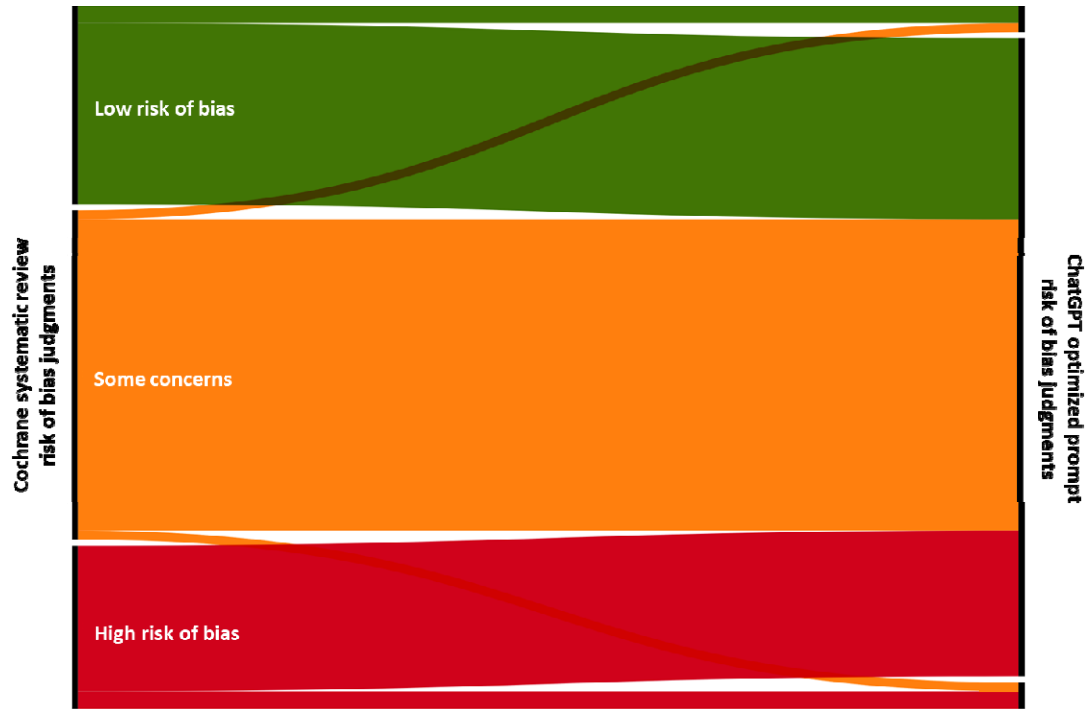
507 **Figure 2: Screening process**



508

509

510 **Figure 3: Flow diagram representing changes in risk of bias judgments**



511

512 The bars on the left represent ratings of low risk of bias (represented in green), some concerns
513 (represented in orange), and high risk of bias (represented in red) by Cochrane systematic reviewers,
514 respectively. The bars on the right represent ratings of low risk of bias, some concerns, and high risk of
515 bias by ChatGPT. The graph represents differences in overall risk of bias ratings between Cochrane
516 systematic reviewers and ChatGPT.

517 References

- 518 1. Guyatt, G. H., Rennie, D., Meade, M. O., & Cook, D. J. (2015). *Users' guides to the medical*
519 *literature*: essentials of evidence-based clinical practice (Third edition.). McGraw-Hill Medical.
- 520 2. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct
521 systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*.
522 2017;7(2):e012545.
- 523 3. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic
524 reviews go out of date? A survival analysis. *Ann Intern Med*. 2007;147(4):224-33.
- 525 4. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane*
526 *Handbook for Systematic Reviews of Interventions* version 6.3 (updated February 2022). Cochrane,
527 2022. Available from www.training.cochrane.org/handbook.
- 528 5. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for
529 assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898.
- 530 6. Crocker TF, Lam N, Jordão M, Brundle C, Prescott M, Forster A, et al. Risk-of-bias assessment
531 using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-
532 intensive: observations from a systematic review. *J Clin Epidemiol*. 2023;161:39-45.
- 533 7. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias
534 tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *J Clin*
535 *Epidemiol*. 2020;126:37-44.
- 536 8. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically
537 assessing bias in clinical trials. *Journal of the American Medical Informatics Association*. 2016;23(1):193-
538 201.
- 539 9. Armijo-Olivo S, Craig R, Campbell S. Comparing machine and human reviewers to evaluate the
540 risk of bias in randomized controlled trials. *Res Synth Methods*. 2020;11(3):484-93.
- 541 10. Hirt J, Meichlinger J, Schumacher P, Mueller G. Agreement in Risk of Bias Assessment Between
542 RobotReviewer and Human Reviewers: An Evaluation Study on Randomised Controlled Trials in Nursing-
543 Related Cochrane Reviews. *J Nurs Scholarsh*. 2021;53(2):246-54.
- 544 11. Arno A, Thomas J, Wallace B, Marshall IJ, McKenzie JE, Elliott JH. Accuracy and Efficiency of
545 Machine Learning-Assisted Risk-of-Bias Assessments in "Real-World" Systematic Reviews : A
546 Noninferiority Randomized Controlled Trial. *Ann Intern Med*. 2022;175(7):1001-9.
- 547 12. Jardim PSJ, Rose CJ, Ames HM, Echavez JFM, Van de Velde S, Muller AE. Automating risk of bias
548 assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a
549 machine learning system. *BMC Med Res Methodol*. 2022;22(1):167.
- 550 13. Soboczenski F, Trikalinos TA, Kuiper J, Bias RG, Wallace BC, Marshall IJ. Machine learning to help
551 researchers evaluate biases in clinical trials: a prospective, randomized user study. *BMC Med Inform*
552 *Decis Mak*. 2019;19(1):96.
- 553 14. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N*
554 *Engl J Med*. 2023;388(13):1233-9.
- 555 15. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of
556 ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit*
557 *Health*. 2023;2(2):e0000198.
- 558 16. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in
559 rheumatology: comparison of physician and ChatGPT-4. *Rheumatol Int*. 2023.
- 560 17. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial
561 Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA*
562 *Internal Medicine*. 2023;183(6):589-96.

- 563 18. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good boolean query for systematic
564 review literature search? arXiv preprint arXiv:230203495. 2023.
- 565 19. Musser G. How AI Knows Things No One Told It. Scientific American. 2023.
- 566 20. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for
567 Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-
568 106.
- 569 21. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020
570 statement: an updated guideline for reporting systematic reviews. Bmj. 2021;372:n71.
- 571 22. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-
572 GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.
- 573 23. Pitre T, Mah J, Roberts S, Desai K, Gu Y, Ryan C, et al. Comparative Efficacy and Safety of
574 Wakefulness-Promoting Agents for Excessive Daytime Sleepiness in Patients With Obstructive Sleep
575 Apnea : A Systematic Review and Network Meta-analysis. Ann Intern Med. 2023;176(5):676-84.
- 576 24. Pitre T, Jassal T, Angjeli A, Jarabana V, Nannapaneni S, Umair A, et al. A comparison of the
577 effectiveness of biologic therapies for asthma: A systematic review and network meta-analysis. Ann
578 Allergy Asthma Immunol. 2023;130(5):595-606.
- 579 25. Pitre T, Van Alstine R, Chick G, Leung G, Mikhail D, Cusano E, et al. Antiviral drug treatment for
580 nonsevere COVID-19: a systematic review and network meta-analysis. Cmaj. 2022;194(28):E969-e80.
- 581 26. Rotondi MA, Donner A. A confidence interval approach to sample size estimation for
582 interobserver agreement studies with multiple raters and outcomes. J Clin Epidemiol. 2012;65(7):778-
583 84.
- 584 27. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or
585 partial credit. Psychol Bull. 1968;70(4):213-20.
- 586 28. Rao JN, Scott AJ. A simple method for the analysis of clustered binary data. Biometrics.
587 1992;48(2):577-85.
- 588 29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics.
589 1977;33(1):159-74.
- 590 30. William Revelle (2023). psych: Procedures for Psychological, Psychometric, and Personality
591 Research. Northwestern University, Evanston, Illinois. R package version 2.3.9, [https://CRAN.R-](https://CRAN.R-project.org/package=psych)
592 [project.org/package=psych](https://CRAN.R-project.org/package=psych).
- 593 31. Sun S. Meta-analysis of Cohen's kappa. Health Services and Outcomes Research Methodology.
594 2011;11(3):145-63.
- 595 32. Ely EW, Ramanan AV, Kartman CE, de Bono S, Liao R, Piruzeli MLB, et al. Efficacy and safety of
596 baricitinib plus standard of care for the treatment of critically ill hospitalised adults with COVID-19 on
597 invasive mechanical ventilation or extracorporeal membrane oxygenation: an exploratory, randomised,
598 placebo-controlled trial. Lancet Respir Med. 2022;10(4):327-36.
- 599 33. Kannan NB, Kohli P, Parida H, Adenuga OO, Ramasamy K. Comparative study of inverted internal
600 limiting membrane (ILM) flap and ILM peeling technique in large macular holes: a randomized-control
601 trial. BMC Ophthalmology. 2018;18(1):177.
- 602 34. Leo S, Beatriz A, Bruna RF, Murillo MC, Rafael RGM, Edison LD, et al. Convalescent plasma for
603 COVID-19 in hospitalised patients: an open-label, randomised clinical trial. European Respiratory
604 Journal. 2022;59(2):2101471.
- 605 35. Aspirin in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled,
606 open-label, platform trial. Lancet. 2022;399(10320):143-51.
- 607 36. Lewis RT. Oral versus systemic antibiotic prophylaxis in elective colon surgery: a randomized
608 study and meta-analysis send a message from the 1990s. Can J Surg. 2002;45(3):173-80.
- 609 37. Lisa H, Maria O, Yuanyuan L, Donna MD, Nicola H, Jennifer Krebs S, et al. Risk of bias versus
610 quality assessment of randomised controlled trials: cross sectional study. BMJ. 2009;339:b4012.

- 611 38. Wu T, He S, Liu J, Sun S, Liu K, Han QL, et al. A Brief Overview of ChatGPT: The History, Status
612 Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*. 2023;10(5):1122-36.
- 613 39. Shahriar S, Hayawi K. Let's have a chat! A Conversation with ChatGPT: Technology, Applications,
614 and Limitations. *arXiv preprint arXiv:230213817*. 2023.
- 615 40. Bertizzolo L, Bossuyt P, Atal I, Ravaud P, Dechartres A. Disagreements in risk of bias assessment
616 for randomised controlled trials included in more than one Cochrane systematic reviews: a research on
617 research study using cross-sectional design. *BMJ Open*. 2019;9(4):e028382.
- 618 41. Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Testing the risk of
619 bias tool showed low reliability between individual reviewers and across consensus assessments of
620 reviewer pairs. *J Clin Epidemiol*. 2013;66(9):973-81.
- 621 42. Windsor B, Popovich I, Jordan V, Showell M, Shea B, Farquhar C. Methodological quality of
622 systematic reviews in subfertility: a comparison of Cochrane and non-Cochrane systematic reviews in
623 assisted reproductive technologies. *Human Reproduction*. 2012;27(12):3460-6.
- 624 43. Petticrew M, Wilson P, Wright K, Song F. Quality of Cochrane reviews. *Quality of Cochrane*
625 *reviews is better than that of non-Cochrane reviews*. *Bmj*. 2002;324(7336):545.
- 626 44. Keskinbora KH. Medical ethics considerations on artificial intelligence. *J Clin Neurosci*.
627 2019;64:277-82.
- 628 45. Temsah MH, Aljamaan F, Malki KH, Alhasan K, Altamimi I, Aljarbou R, et al. ChatGPT and the
629 Future of Digital Health: A Study on Healthcare Workers' Perceptions and Expectations. *Healthcare*
630 (Basel). 2023;11(13).
- 631 46. Noura A, Khalid A, Rupesh R, Khalid AM, Fadi A, Ibraheem T, et al. Exploring Perceptions and
632 Experiences of ChatGPT in Medical Education: A Qualitative Study Among Medical College Faculty and
633 Students in Saudi Arabia. *medRxiv*. 2023:2023.07.13.23292624.
- 634 47. Qiu S, Liu Q, Zhou S, Wu C. Review of Artificial Intelligence Adversarial Attack and Defense
635 Technologies. *Applied Sciences [Internet]*. 2019; 9(5).
- 636 48. Carlini N, Jagielski M, Choquette-Choo CA, Paleka D, Pearce W, Anderson H, et al. Poisoning
637 web-scale training datasets is practical. *arXiv preprint arXiv:230210149*. 2023.

638