

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 (XX)

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:** Internationally accepted standards for systematic reviews necessitate assessment of the
36 risk of bias of primary studies. Assessing risk of bias, however, can be time- and resource-intensive. AI-
37 based solutions may increase efficiency and reduce burden.

38 **Objective:** To evaluate the reliability of ChatGPT for performing risk of bias assessments of randomized
39 trials using the revised risk of bias tool for randomized trials (RoB 2.0).

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 judgements. The agreement between ChatGPT's
47 assessments and those of Cochrane systematic reviewers was quantified using weighted kappa
48 statistics.

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

56 **Conclusion:** Our results suggest that ChatGPT and systematic reviewers only have “slight” to “fair”
57 agreement in risk of bias judgements for randomized trials. ChatGPT is currently unable to reliably
58 assess risk of bias of randomized trials. We advise against using ChatGPT to perform risk of bias
59 assessments. There may be opportunities to use ChatGPT to streamline other aspects of systematic
60 reviews, such as screening of search records or collection of data.

61

62 **Background**

63 The practice of evidence-based medicine demands knowledge of the best available evidence, which
64 most often comes from rigorous systematic reviews and meta-analyses (1). Systematic reviews,
65 however, are time- and resource-intensive (2-4). Empirical evidence suggests they typically require
66 upwards of one year to complete and publish and many are outdated at or shortly following publication
67 (3, 5).

68 One time- and resource-intensive component of systematic reviews is the assessment of risk of bias of
69 primary studies—defined as the propensity for studies to systematically over- or underestimate
70 treatment effects (6, 7). Risk of bias assessments are burdensome and time-consuming and demand
71 specialized training (6, 7). Moreover, to reduce the opportunity for errors, guidance for conducting
72 rigorous systematic reviews typically suggests authors assess risk of bias independently and in duplicate,
73 adding to the complexity and workload of the process (6).

74 Many tools exist to assess the risk of bias of randomized trials (8, 9), examples of which include the tools
75 from the Joanna Briggs Institute (10), the Jadad Scale (11), and the Critical Appraisal Skills Program
76 (CASP) checklist (12). These tools, however, generally fall short compared to the most commonly used
77 tool, the original Cochrane risk of bias tool for randomized trials, and their application is not
78 recommended (6, 13, 14).

79 In 2019, a new risk of bias tool was introduced that built on the successes of the previous Cochrane
80 endorsed risk of bias tool but also incorporated new advancements (15). This tool was called the revised
81 tool for assessing risk of bias of randomized trials (RoB 2.0) and is now largely considered the gold
82 standard (6).

83 The application of the RoB 2.0 tool, like other risk of bias tools, typically involves reviewers using trial
84 reports and trial registrations or protocols, when available, to make judgements for each risk of bias
85 domain (6, 15). Reviewers who collect data for a systematic review are also typically tasked with
86 assessing the risk of bias of eligible trials (6). The RoB 2.0 tool rates risk of bias as either high, some
87 concerns, or low across five domains: randomization, deviations from intended intervention, missing
88 outcome data, measurement of outcome, and selective reporting. To guide judgements, the RoB 2.0
89 includes signaling questions for each domain. The overall rating of risk of bias is determined by the
90 domain rated at highest risk of bias (6, 15).

91 While the RoB 2.0 tool builds off a decade's worth of experience with the original risk of bias tool, recent
92 evidence suggests that reviewers find it complex and time-consuming—perhaps more complex and
93 time-consuming than previous risk of bias tools (7, 16). Innovations to streamline and simplify risk of
94 bias assessments without compromising their rigor will reduce the time and effort required to perform
95 systematic reviews and aid in maintaining their currency.

96 RobotReviewer is an automated tool to extract data from and assess the risk of bias of randomized trials
97 (17-19). Previous studies on RobotReviewer show optimistic results, with generally moderate to high
98 agreement with systematic reviewers (70% to 90%) (17, 18). The RobotReviewer, however, was trained
99 on the original Cochrane risk of bias tool, rather than the RoB 2.0 tool, and only offers judgements on
100 four of the seven domains of the original tool. To our knowledge, RobotReviewer is the only artificial
101 intelligence (AI) tool for assessing risk of bias in systematic reviews (20).

102 ChatGPT (OpenAI, San Francisco, California, USA) is a conversational AI large language model with
103 capabilities in natural language processing and realization (21). Differing from specialized automated
104 tools for risk of bias assessments, ChatGPT is a general purpose tool, has been developed to emulate
105 human language rather than risk of bias assessments, and has been trained on an internet-scale corpus
106 covering many areas of knowledge, rather than a small training set focused on evidence synthesis and
107 evaluation (21).

108 This study evaluates the performance of ChatGPT, an AI-based language model, for assessing risk of bias
109 of randomized trials using the RoB 2.0 tool. To do this, we sampled Cochrane systematic reviews using
110 the RoB 2.0 tool and used ChatGPT-4—an advanced large language model offered by OpenAI—to assess
111 the risk of bias of the trials within these reviews. We compared ChatGPT’s assessment with those
112 presented in Cochrane reviews. Consistency in assessments of risk of bias between ChatGPT and
113 Cochrane reviewers will suggest that ChatGPT can provide a reliable assessment of the risk of bias of
114 randomized trials. Conversely, discrepancies in risk of bias assessments between ChatGPT and Cochrane
115 reviewers will suggest that ChatGPT is unreliable for assessing risk of bias.

116 **Methods**

117 We registered our protocol on Open Science Framework (<https://osf.io/aq85p>) in September 2023. We
118 report our study according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses
119 (PRISMA) and Guidelines for Reporting Reliability and Agreement Studies (GRRAS) reporting checklists
120 (22, 23).

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

122 Figure 1 presents an overview of our methods.

123 ***Search strategy and screening***

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

128 Reviewers worked independently and in duplicate to screen Cochrane reviews for eligibility, starting
129 with the most recently published (August 2023) and working backwards in time. We preferentially
130 included the most recently published Cochrane systematic reviews since these reviews are most likely to
131 have used the most up-to-date version of the RoB 2.0 tool instead of preliminary pilot versions of the
132 tool (15). Reviewers continued screening until we had identified our target sample size of approximately
133 160 trials.

134 ***Eligibility criteria***

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

138 We included newly published or updated Cochrane systematic reviews addressing the benefits and/or
139 harms of health interventions that included one or more parallel randomized trials and reported
140 consensus-based risk of bias judgements using the Cochrane-endorsed RoB 2.0 tool (15). We defined

141 consensus-based as two reviewers agreeing on the final risk of bias judgements. This may involve two
142 reviewers independently assessing risk of bias and resolving conflicts by discussion or a reviewer
143 assessing risk of bias and a second reviewer confirming the first reviewers' judgements.

144 We excluded systematic reviews that were not published by Cochrane, since such reviews may not
145 involve reviewers with sufficient training to appropriately apply the RoB 2.0 tool. We also excluded
146 Cochrane systematic reviews that investigated prognosis or the performance of diagnostic tests and
147 systematic reviews that only included observational studies since these reviews will necessitate the use
148 of other risk of bias tools.

149 Cochrane systematic reviews use summary of findings tables to present their results (6, 24). These tables
150 list outcomes in order of importance, the number of trials and patients that contributed data to the
151 meta-analysis for each outcome, the relative and absolute effect estimates based on meta-analyses, and
152 judgements about the certainty of evidence (6, 24). From each eligible review, we selected the first two
153 listed outcomes (suggesting that they are the most important) that were informed by one or more trials.
154 If either of the first two outcomes were continuous, we then selected the third outcome listed in the
155 summary of findings table. If the two reported outcomes were both dichotomous, we then selected the
156 first listed continuous outcome reported in the summary of findings table. When summary of findings
157 tables reported on the same outcome at different timepoints, we selected entirely unique outcomes.

158 From each review, we included all parallel randomized trials published in English that were included in
159 analyses addressing the outcomes of interest. We excluded crossover and cluster randomized trials
160 since these trial designs require unique considerations in their assessment of risk of bias and different
161 versions of the RoB 2.0 tool. Cochrane reviews often include unpublished trial data. When reviews
162 reported that information for a particular trial was unpublished or was drawn from a combination of
163 unpublished and published data, we excluded those trials since we did not have access to the same
164 unpublished information as the Cochrane reviewers for risk of bias assessments.

165 For feasibility, we also excluded trials for which data were drawn from multiple publications. Including
166 such trials would have necessitated an exhaustive review of all related publications to identify those
167 containing the outcome data and the comprehensive details required for risk of bias assessment.

168 ***ChatGPT prompts***

169 A key component in the use of ChatGPT is the design of the text used to instruct the model (called
170 'prompts') to generate an answer. We anticipated that ChatGPT's risk of bias judgements may depend
171 on the nature of the prompts that it is provided. To study how different prompts may influence risk of
172 bias judgements, we iteratively designed three different prompts: a minimal prompt including limited
173 instructions for assessing risk of bias, a maximal prompt with extensive instructions, and an optimized
174 prompt that was designed to include sufficient information to yield the best risk of bias judgements.

175 We piloted the prompts using 15 trials drawn from systematic reviews previously performed by our own
176 team and refined the prompts by iterative discussion and input by the co-authors (25-27). All prompts
177 asked ChatGPT to judge risk of bias for all RoB 2.0 domains (bias due to randomization, deviation from
178 intended intervention, missing outcome data, measurement of outcome, and selective reporting) as low
179 risk of bias, some concerns, or high risk of bias—consistent with RoB 2.0 guidance (15). Supplement 1
180 presents these three prompts.

181 The RoB 2.0 tool is accompanied by a document that describes the tool and offers guidance on its
182 implementation. All three prompts included the RoB 2.0 full guidance document (riskofbias.info), which
183 were fed to ChatGPT using the AskYourPDF ChatGPT plugin that allows ChatGPT to read and query PDF
184 documents. All prompts also included a PDF copy of the trial publication, a PDF copy of the trial
185 registration or protocol (if one was available), and specified the outcome of interest for which risk of
186 bias assessment was being performed. The prompts also specified the order in which the RoB 2.0
187 domains should be assessed and that the responses should include a judgment and rationale.

188 The RoB 2.0 tool offers two options for assessing the risk of bias due to deviations of the intended
189 intervention: one for the effect of assignment to the intervention and the other for the effect of
190 adhering to the intervention. In Cochrane systematic reviews, the subsection on risk of bias typically
191 reports whether Cochrane reviewers assessed risk of bias for the effect of assignment or adherence to
192 the intervention. Our ChatGPT prompts also specified whether to assess risk of bias for the effect of
193 assignment or adherence to the intervention. We specified the same option used by the Cochrane
194 systematic reviews. For systematic reviews that failed to specify whether they assessed risk of bias for
195 the effect of being assigned to the intervention or adherence to the intervention, we assumed they
196 assessed risk of bias for assignment to the intervention.

197 The ChatGPT prompts do not include any information related to the consensus-based risk of bias
198 judgements presented in the systematic reviews. Hence, ChatGPT is 'blind' to the Cochrane systematic
199 reviewers' risk of bias judgements.

200 ***Data collection***

201 RoB 2.0 guidance demands that reviewers perform risk of bias judgements for each particular result
202 rather than each trial or outcome, since risk of bias may differ across outcomes in a trial or across
203 different ways of statistically summarizing the results for the same outcome (15). We took this approach
204 in this study. For each eligible trial and outcome, we collected information on the consensus-based risk
205 of bias judgements presented in the Cochrane systematic reviews. Subsequently, for each eligible trial,
206 we used the ChatGPT-4 chatbot to assess the risk of bias of the outcomes of interest, using each of the
207 three ChatGPT prompts. ChatGPT-4 is a more advanced iteration of its predecessor ChatGPT-3. Unlike
208 ChatGPT-3, ChatGPT-4 is only available with a paid subscription to OpenAI. We implemented each of the
209 prompts in unique chats.

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

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

221 ***Data synthesis and analysis***

222 Sample size estimation

223 We used the kappaSize package in R (Vienna, Austria, Version 4.1.3) to estimate sample size (28). We
224 aimed to calculate the number of required trials to obtain a sufficiently precise estimate of a value of
225 kappa for which systematic reviewers will feel confident using ChatGPT for risk of bias assessments. We
226 assumed that most reviewers would feel confident using ChatGPT for risk of bias assessments if it yields
227 a kappa of 0.70, indicating substantial agreement, with the lower bound of the confidence interval no
228 less than 0.55. We anticipated the risk of bias distribution to be approximately 30% low, 30% with some
229 concerns, and 40% high.

230 We inflated the estimated sample size by a design effect to account for correlation between the risk of
231 bias of trials from the same review. We assumed an intra-review correlation of 0.05 and an average of
232 10 trials per review, yielding a design effect of 1.45. This resulted in a minimum sample size of 120 trials
233 from 12 reviews. We investigated the sensitivity of our estimated sample size to different assumptions
234 about the anticipated distribution of risk of bias judgements across the three categories and the
235 potential correlation between trials from the same review. To account for other potential scenarios
236 (e.g., kappa = 0.6, intrareview correlation of 0.1), we ultimately intended to include approximately 160
237 trials from 16 reviews.

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

239 We present the inter-rater agreement, represented by weighted kappa, between each of the three
240 ChatGPT prompts and consensus-based risk of bias judgements from Cochrane authors. Unlike
241 percentage agreement, the weighted kappa accounts for the possibility of agreement due to chance and
242 for the ordinal nature of the response options of the RoB 2.0 tool (low risk of bias, some concerns, high
243 risk of bias) (29).

244 We present separate analyses for each RoB 2.0 domain and for the overall rating of risk of bias. Each
245 analysis only includes one outcome from each included trial. Our primary analysis includes the most
246 important outcome, based on the order in which outcomes were listed in Cochrane systematic review
247 summary of findings tables. We adjusted for clustering of trials within each systematic review by
248 inflating the variance of all estimates by the design effect (30).

249 We interpreted Cohen's kappa statistics using previously established guidelines: values from 0.0 to 0.2
250 indicating slight agreement, 0.21 to 0.40 indicating fair agreement, 0.41 to 0.60 indicating moderate
251 agreement, 0.61 to 0.80 indicating substantial agreement, and 0.81 to 1.0 indicating perfect agreement
252 (31).

253 We hypothesized that ChatGPT may be more reliable to assess risk of bias when there are few subjective
254 judgements. Therefore, we expected better agreement for: (i) trials addressing pharmacologic
255 interventions because trials of pharmacologic interventions are more likely to blind patients and
256 healthcare providers thus simplifying judgements related to deviations from intended intervention and
257 measurement of outcomes; (ii) trials addressing risk of bias of assignment of the intervention because
258 assignment to the intervention does not necessitate making judgements about adherence; (iii) objective
259 outcomes since these outcomes do not need additional judgements about whether failure to blind may
260 have resulted in differential measurement of the outcome, and (iv) dichotomous instead of continuous

261 outcomes since continuous outcomes are more likely to be subjective. To test these hypotheses, we
262 performed secondary analyses stratified by these factors.

263 We also performed a secondary analysis in which we collapsed ratings of “some concerns” and “high risk
264 of bias” into a single category.

265 We performed all statistical analyses using the psych package in R (Vienna, Austria, Version 4.1.3) (32).

266 Review of ChatGPT justifications for discrepant risk of bias judgements between Cochrane systematic
267 reviewers and ChatGPT

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

271 **Results**

272 ***Systematic review and trial characteristics***

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

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

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

287 ***Agreement between ChatGPT and consensus-based risk of bias judgements from Cochrane review*** 288 ***authors***

289 In our analyses, each trial contributed data only for one outcome. When a trial reported data on more
290 than one outcome of interest, we included data for the outcome reported first in the systematic review.

291 We found overall only slight agreement between ChatGPT risk of bias judgements and consensus-based
292 risk of bias judgements from systematic reviewers. Agreement for overall risk of bias ranged between
293 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,
294 respectively. Figure 2 presents a flow diagram representing categorical changes in the overall rating of
295 risk of bias between systematic reviewers and the optimized ChatGPT prompt.

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

299 We did not find evidence that ChatGPT had importantly different reliability in stratified analyses based
300 on whether trials addressed pharmacologic or non-pharmacologic interventions, objective or subjective
301 outcomes, dichotomous or continuous outcomes or whether reviews specified assessing the risk of bias
302 of assignment to the intervention (Supplements 4 to 10). ChatGPT showed “slight” to “fair” agreement
303 for these subgroups.

304 Likewise, our secondary analysis that collapsed ratings of “some concerns” and “high risk of bias” into a
305 single category also showed “slight” to “fair” agreement (Supplement 11).

306 Supplement 12 presents qualitative observations about discrepant risk of bias judgements between
307 ChatGPT and Cochrane systematic reviewers.

308 **Discussion**

309 ***Main findings***

310 We performed a study evaluating ChatGPT for assessing the risk of bias of randomized trials using the
311 Cochrane-endorsed RoB 2.0 tool (15). To do this, we sampled Cochrane systematic reviews that
312 reported RoB 2.0 judgements for randomized trials, assessed the risk of bias of trials using ChatGPT via
313 three variations of prompts, and compared the degree of agreement between RoB 2.0 judgements
314 presented in systematic reviews and those by ChatGPT.

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

321 We also reviewed cases in which ChatGPT's risk of bias judgements differed from those of Cochrane
322 systematic reviewers with the goal of identifying ways in which we can refine future prompts. Our
323 findings indicate that ChatGPT might make more accurate risk of bias judgements if informed about
324 both low and high risk of bias methodological traits. For example, one trial reported randomization by
325 an “interactive web-response system”, which suggests central randomization and allocation
326 concealment (33). ChatGPT, however, rated the trial at some concerns for randomization because the
327 trial report “does not explicitly mention whether the allocation sequence was concealed”. Training
328 ChatGPT to recognize features of trials at low versus high risk of bias may improve the reliability of its
329 risk of bias assessments.

330 Though our results appear discouraging, they must also be contextualized considering general poor
331 agreement between even experienced reviewers in implementing the RoB 2.0 tool. For example, a
332 previous investigation of the reliability of RoB 2.0 using experienced systematic reviewers reported
333 inter-rater reliability ranging between 0.04 to 0.45, indicating only slight to fair agreement (16). The
334 original Cochrane risk of bias tool also demonstrated poor inter-rater reliability for select domains (34).

335 Our results may also be explained by ChatGPT’s limited memory, which may not be sufficient to fully
336 process RoB 2.0’s extensive and lengthy guidance (35, 36). An improvement in ChatGPT's performance
337 in risk of bias assessment might be achieved by enhancing its memory capabilities, by utilizing other

338 plans from OpenAI that offer expanded memory options such as ChatGPT Enterprise, or by fine-tuning
339 ChatGPT's base model—a process that involves additional training of the model.

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

344 ***Strengths and limitation***

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

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

359 Evidence suggests that risk of bias assessments in Cochrane reviews, despite their rigor, are sometimes
360 unreliable and inconsistent with established guidance (16). Hence, differences in risk of bias judgements
361 between ChatGPT and Cochrane systematic reviewers may also represent errors on part of reviewers.
362 Previous studies suggest that agreement between reviewers in assessing risk of bias may be very poor
363 (37, 38). To minimize the potential for this error, we limited our sample to Cochrane systematic reviews,
364 which are known for their methodological rigor (39, 40).

365 The performance of ChatGPT is also not static. The infrastructure, interfaces, and applications built
366 around ChatGPT are continuously updated (35, 41, 42). Our experiment was performed over a two-week
367 time period between September and October 2023. It is possible that the performance that we
368 observed may not be replicable in the future—though it is more likely that the capabilities of ChatGPT
369 will improve rather than deteriorate. Even with identical prompts, ChatGPT might provide slightly
370 different answers due to the inherent stochasticity in its response generation (41).

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

375 Our prompts queried ChatGPT to provide a justification for its ratings of risk of bias. To understand
376 reasons why ChatGPT may produce unreliable risk of bias judgements, we also reviewed justifications
377 provided by ChatGPT to support its judgements for potential errors or problems. While we performed a

378 general review of justifications for which ChatGPT and Cochrane reviewers made discrepant risk of bias
379 judgements, we did not perform a formal qualitative analysis of the justifications.

380 While we did not record the exact duration our team spent using ChatGPT, we estimate that each trial
381 took no longer than 15 minutes—less time than on average required for a reviewer to conduct an
382 individual risk of bias assessment and consensus meeting according to empirical evidence (7, 16).

383 Finally, our systematic review includes minor deviations from the protocol. To account for correlation
384 between trials in the same systematic review, we planned to calculate weighted kappa within each
385 review individually and pool the weighted kappa statistics across systematic reviews using random-
386 effects meta-analysis (43). The sampling distribution of kappa, however, is asymmetric. While with a
387 large enough number of observations, the sampling distribution of kappa is approximately normal, we
388 found there to be too few trials within each systematic review to assume normality, precluding our
389 approach to perform meta-analyses. Instead, we adjusted the variance of all estimates for the
390 correlation within each systematic review. Likewise, in our primary analysis, we excluded ratings of
391 uncertain risk of bias from analyses. We had planned to perform additional sensitivity analyses treating
392 these ratings as some concerns or high risk of bias but there were too few uncertain ratings to affect
393 estimates of reliability.

394 ***Relation to previous findings***

395 Attempts to reduce the time, resources, and expertise needed to perform systematic reviews are not
396 new. For example, RobotReviewer is an automated tool to extract data from and assess the risk of bias
397 of randomized trials (17). The RobotReviewer, however, was trained on the original Cochrane risk of bias
398 tool and only offers judgements on four of the seven domains of the original tool. Since then, Cochrane
399 has adopted a revised risk of bias assessment tool that requires more nuanced judgements and is more
400 resource- and time-intensive (7). Given the performance of ChatGPT, however, adapting RobotReviewer
401 to provide risk of bias assessments using the RoB 2.0 tool may be more promising.

402 ***Implications***

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

407 Our study also has implications for future research. Since the completion of this study, OpenAI has
408 released the option to create custom GPTs (42). Custom GPTs offer users the option to customize their
409 ChatGPT using additional instructions. While our prompts in their current form could not be used to
410 reliably assess risk of bias, other prompts or custom GPTs may be able to provide more reliable
411 assessments.

412 Further, more granular prompts may also lead to more reliable judgements. For example, for each
413 domain, RoB 2.0 contains a series of signaling questions designed to help reviewers think systematically
414 about the different aspects of trial conduct that might lead to bias. These signaling questions are
415 answered with "Yes," "Probably yes," "Probably no," "No," or "No information." Based on the answers to
416 these questions, a judgment is made about the risk of bias for that domain as "Low," "Some concerns,"
417 or "High." Instead of asking ChatGPT to assess the risk of bias of each domain, ChatGPT may be
418 prompted to go through the RoB 2.0 signalling questions. Future research may address the usefulness of

419 having systematic reviewers reconcile their risk of bias assessments with ChatGPT or the role of ChatGPT
420 in training systematic reviewers.

421 There are also opportunities to use ChatGPT to streamline other aspects of systematic reviews. Early
422 studies suggest that ChatGPT can be used to devise search strategies (44). ChatGPT may also assist with
423 screening search records, extracting data from eligible studies, or performing evaluations of the
424 certainty of evidence. Though, at this time, based on the results of the current study, we are not
425 optimistic about ChatGPT's ability to reliably extract data or evaluate the certainty of evidence.
426 Screening studies is less subjective and perhaps better suited to ChatGPT's abilities.

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

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

437 **Conclusion**

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

446 **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.

447

Table 2: Degree of Agreement

		Consensus based risk of bias judgements 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%)

448

449

450

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

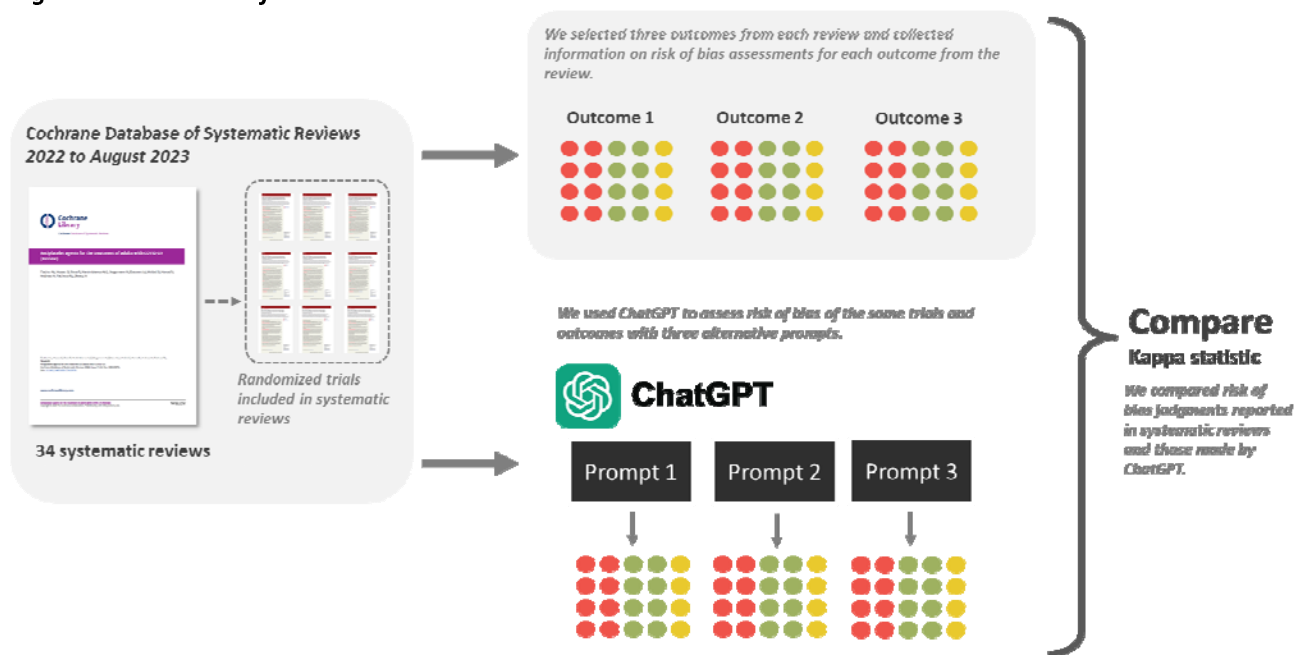
	Optimized prompt	Minimal prompt	Maximal prompt
	Weighted kappa (95% CI)		
Overall risk of bias rating	0.17 (0.02, 0.32)	0.11 (-0.04, 0.27)	0.16 (0.01, 0.3)
Risk of bias due to randomization	0.24 (0.02, 0.47)	0.09 (-0.16, 0.33)	0.09 (-0.15, 0.34)
Risk of bias due to deviations from the intended intervention	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)

451

452

453 **Figures**

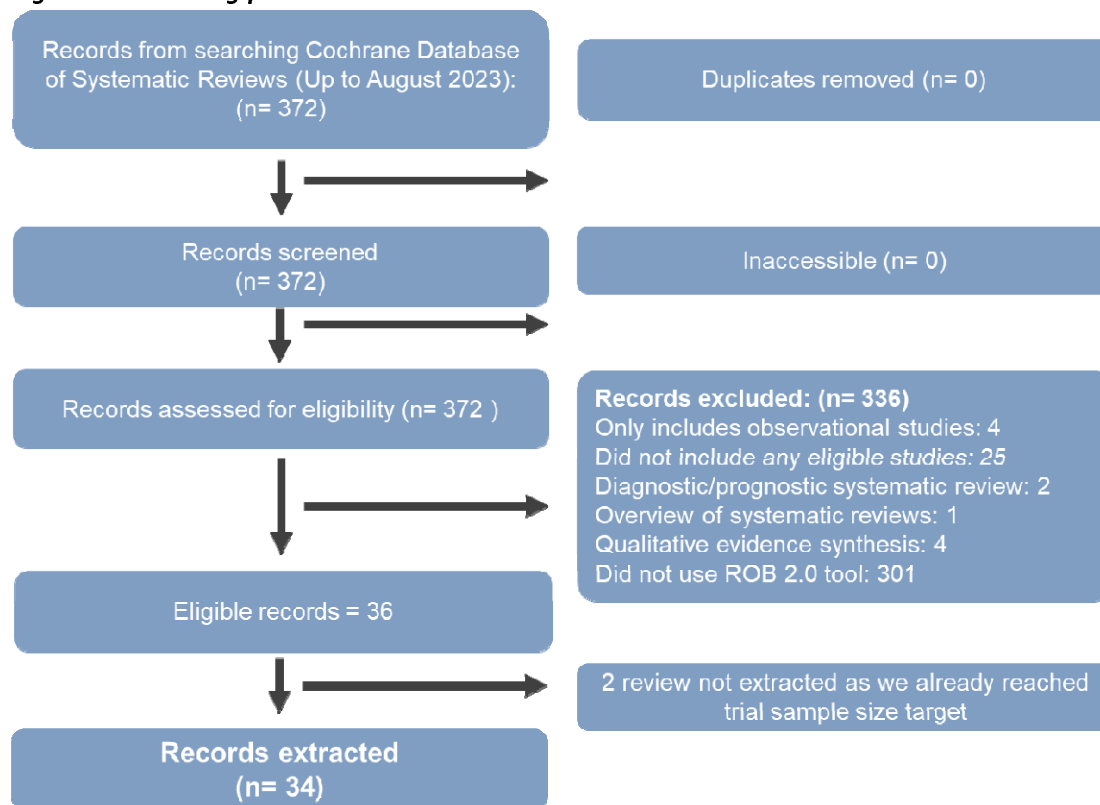
454 **Figure 1: Overview of methods**



455

456

457 **Figure 2: Screening process**

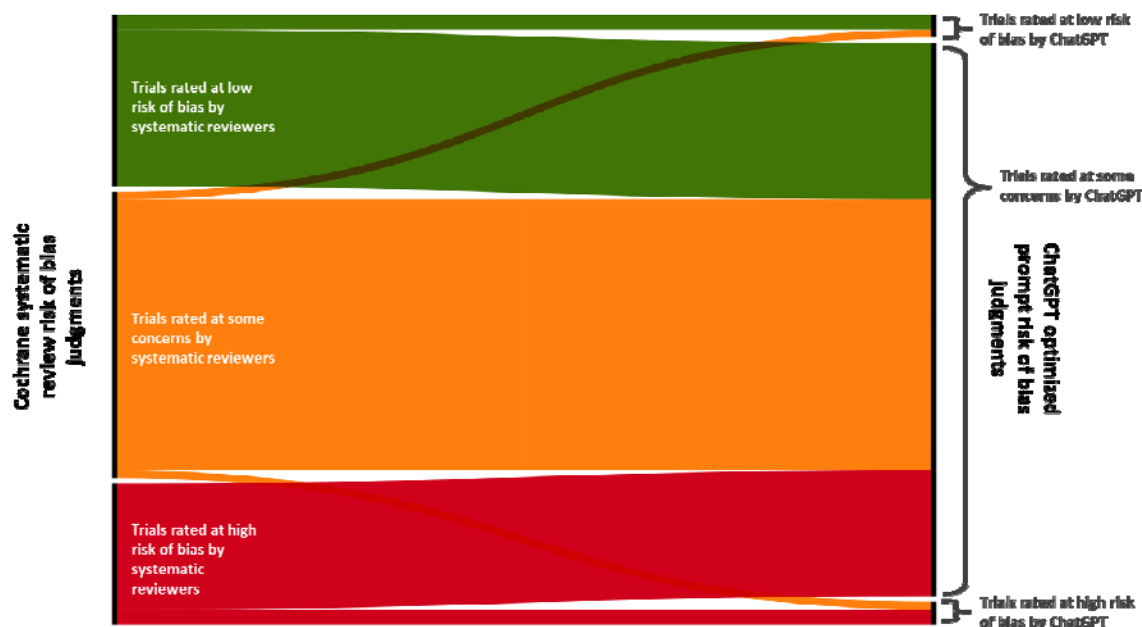


458

459

460 **Figure 3: Flow diagram representing changes in risk of bias judgements**

461



462

463 The bars on the left represent ratings of low risk of bias (represented in green), some concerns
464 (represented in orange), and high risk of bias (represented in red) by Cochrane systematic reviewers.
465 The bars on the right represent ratings of low risk of bias, some concerns, and high risk of bias by
466 ChatGPT. The graph represents differences in ratings of overall risk of bias between Cochrane systematic
467 reviewers and ChatGPT.

468 Cochrane systematic reviewers rated comparable proportions of trials at low risk of bias, some
469 concerns, and high risk of bias. Conversely, ChatGPT rated few trials at low and high risk of bias and
470 most trials as having some concerns.

471 References

- 472 1. Guyatt, G. H., Rennie, D., Meade, M. O., & Cook, D. J. (2015). *Users' guides to the medical*
473 *literature*: essentials of evidence-based clinical practice (Third edition.). McGraw-Hill Medical.
- 474 2. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for
475 greater involvement of machine learning to assess the promise of clinical trials. *Contemp Clin Trials*
476 *Commun.* 2019;16:100443.
- 477 3. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct
478 systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open.*
479 2017;7(2):e012545.
- 480 4. Nussbaumer-Streit B, Ellen M, Klerings I, Sfetcu R, Riva N, Mahmić-Kaknjo M, et al. Resource use
481 during systematic review production varies widely: a scoping review. *J Clin Epidemiol.* 2021;139:287-96.
- 482 5. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic
483 reviews go out of date? A survival analysis. *Ann Intern Med.* 2007;147(4):224-33.
- 484 6. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane*
485 *Handbook for Systematic Reviews of Interventions* version 6.3 (updated February 2022). Cochrane,
486 2022. Available from www.training.cochrane.org/handbook.
- 487 7. Crocker TF, Lam N, Jordão M, Brundle C, Prescott M, Forster A, et al. Risk-of-bias assessment
488 using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-
489 intensive: observations from a systematic review. *J Clin Epidemiol.* 2023;161:39-45.
- 490 8. Farrah K, Young K, Tunis MC, Zhao L. Risk of bias tools in systematic reviews of health
491 interventions: an analysis of PROSPERO-registered protocols. *Syst Rev.* 2019;8(1):280.
- 492 9. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of
493 randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical*
494 *Trials.* 1995;16(1):62-73.
- 495 10. Barker TH, Stone JC, Sears K, Klugar M, Tufanaru C, Leonardi-Bee J, et al. The revised JBI critical
496 appraisal tool for the assessment of risk of bias for randomized controlled trials. *JBI Evid Synth.*
497 2023;21(3):494-506.
- 498 11. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. Assessing
499 the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials.*
500 1996;17(1):1-12.
- 501 12. Critical Appraisal Skills Program (CASP). *Critical Appraisal Checklists* [Available from:
502 <https://casp-uk.net/casp-tools-checklists/>].
- 503 13. Clark HD, Wells GA, Huët C, McAlister FA, Salmi LR, Fergusson D, Laupacis A. Assessing the
504 quality of randomized trials: reliability of the Jadad scale. *Control Clin Trials.* 1999;20(5):448-52.
- 505 14. Berger VW. Is the Jadad score the proper evaluation of trials? *J Rheumatol.* 2006;33(8):1710-1;
506 author reply 1-2.
- 507 15. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for
508 assessing risk of bias in randomised trials. *BMJ.* 2019;366:l4898.
- 509 16. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias
510 tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *J Clin*
511 *Epidemiol.* 2020;126:37-44.
- 512 17. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically
513 assessing bias in clinical trials. *Journal of the American Medical Informatics Association.* 2016;23(1):193-
514 201.
- 515 18. Arno A, Thomas J, Wallace B, Marshall IJ, McKenzie JE, Elliott JH. Accuracy and Efficiency of
516 Machine Learning-Assisted Risk-of-Bias Assessments in "Real-World" Systematic Reviews : A
517 Noninferiority Randomized Controlled Trial. *Ann Intern Med.* 2022;175(7):1001-9.

- 518 19. Jardim PSJ, Rose CJ, Ames HM, Echavez JFM, Van de Velde S, Muller AE. Automating risk of bias
519 assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a
520 machine learning system. *BMC Med Res Methodol*. 2022;22(1):167.
- 521 20. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping
522 review. *Journal of Clinical Epidemiology*. 2022;144:22-42.
- 523 21. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N*
524 *Engl J Med*. 2023;388(13):1233-9.
- 525 22. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for
526 Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64(1):96-
527 106.
- 528 23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020
529 statement: an updated guideline for reporting systematic reviews. *Bmj*. 2021;372:n71.
- 530 24. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-
531 GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol*. 2011;64(4):383-94.
- 532 25. Pitre T, Mah J, Roberts S, Desai K, Gu Y, Ryan C, et al. Comparative Efficacy and Safety of
533 Wakefulness-Promoting Agents for Excessive Daytime Sleepiness in Patients With Obstructive Sleep
534 Apnea : A Systematic Review and Network Meta-analysis. *Ann Intern Med*. 2023;176(5):676-84.
- 535 26. Pitre T, Jassal T, Angjeli A, Jarabana V, Nannapaneni S, Umair A, et al. A comparison of the
536 effectiveness of biologic therapies for asthma: A systematic review and network meta-analysis. *Ann*
537 *Allergy Asthma Immunol*. 2023;130(5):595-606.
- 538 27. Pitre T, Van Alstine R, Chick G, Leung G, Mikhail D, Cusano E, et al. Antiviral drug treatment for
539 nonsevere COVID-19: a systematic review and network meta-analysis. *Cmaj*. 2022;194(28):E969-e80.
- 540 28. Rotondi MA, Donner A. A confidence interval approach to sample size estimation for
541 interobserver agreement studies with multiple raters and outcomes. *J Clin Epidemiol*. 2012;65(7):778-
542 84.
- 543 29. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or
544 partial credit. *Psychol Bull*. 1968;70(4):213-20.
- 545 30. Rao JN, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics*.
546 1992;48(2):577-85.
- 547 31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*.
548 1977;33(1):159-74.
- 549 32. William Revelle (2023). *psych: Procedures for Psychological, Psychometric, and Personality*
550 *Research*. Northwestern University, Evanston, Illinois. R package version 2.3.9, [https://CRAN.R-](https://CRAN.R-project.org/package=psych)
551 [project.org/package=psych](https://CRAN.R-project.org/package=psych).
- 552 33. Ely EW, Ramanan AV, Kartman CE, de Bono S, Liao R, Piruzeli MLB, et al. Efficacy and safety of
553 baricitinib plus standard of care for the treatment of critically ill hospitalised adults with COVID-19 on
554 invasive mechanical ventilation or extracorporeal membrane oxygenation: an exploratory, randomised,
555 placebo-controlled trial. *Lancet Respir Med*. 2022;10(4):327-36.
- 556 34. Lisa H, Maria O, Yuanyuan L, Donna MD, Nicola H, Jennifer Krebs S, Terry PK. Risk of bias versus
557 quality assessment of randomised controlled trials: cross sectional study. *BMJ*. 2009;339:b4012.
- 558 35. Wu T, He S, Liu J, Sun S, Liu K, Han QL, Tang Y. A Brief Overview of ChatGPT: The History, Status
559 Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*. 2023;10(5):1122-36.
- 560 36. Shahriar S, Hayawi K. Let's have a chat! A Conversation with ChatGPT: Technology, Applications,
561 and Limitations. *arXiv preprint arXiv:230213817*. 2023.
- 562 37. Bertizzolo L, Bossuyt P, Atal I, Ravaud P, Dechartres A. Disagreements in risk of bias assessment
563 for randomised controlled trials included in more than one Cochrane systematic reviews: a research on
564 research study using cross-sectional design. *BMJ Open*. 2019;9(4):e028382.

- 565 38. Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Testing the risk of
566 bias tool showed low reliability between individual reviewers and across consensus assessments of
567 reviewer pairs. *J Clin Epidemiol*. 2013;66(9):973-81.
- 568 39. Windsor B, Popovich I, Jordan V, Showell M, Shea B, Farquhar C. Methodological quality of
569 systematic reviews in subfertility: a comparison of Cochrane and non-Cochrane systematic reviews in
570 assisted reproductive technologies. *Human Reproduction*. 2012;27(12):3460-6.
- 571 40. Petticrew M, Wilson P, Wright K, Song F. Quality of Cochrane reviews. *Quality of Cochrane*
572 *reviews is better than that of non-Cochrane reviews*. *Bmj*. 2002;324(7336):545.
- 573 41. Abdullah M, Madain A, Jararweh Y, editors. *ChatGPT: Fundamentals, Applications and Social*
574 *Impacts*. 2022 Ninth International Conference on Social Networks Analysis, Management and Security
575 (SNAMS); 2022 29 Nov.-1 Dec. 2022.
- 576 42. OpenAI. ChatGPT — Release Notes: The latest update for ChatGPT 2024 [Available from:
577 <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>].
- 578 43. Sun S. Meta-analysis of Cohen's kappa. *Health Services and Outcomes Research Methodology*.
579 2011;11(3):145-63.
- 580 44. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good boolean query for systematic
581 review literature search? *arXiv preprint arXiv:230203495*. 2023.
- 582 45. Temsah MH, Aljamaan F, Malki KH, Alhasan K, Altamimi I, Aljarbou R, et al. ChatGPT and the
583 Future of Digital Health: A Study on Healthcare Workers' Perceptions and Expectations. *Healthcare*
584 (Basel). 2023;11(13).
- 585 46. Noura A, Khalid A, Rupesh R, Khalid AM, Fadi A, Ibraheem T, et al. Exploring Perceptions and
586 Experiences of ChatGPT in Medical Education: A Qualitative Study Among Medical College Faculty and
587 Students in Saudi Arabia. *medRxiv*. 2023:2023.07.13.23292624.

588