(How well) can large language models and AI-based automation tools assist in Risk of Bias Assessment?

Angelika Eisele-Metzger

Institute for Evidence in Medicine, Medical Center & Medical Faculty – University of Freiburg, Germany

Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany

9th March 2025

Conflicts of interest

- Researcher at the Institute for Evidence in Medicine, Medical Center – University of Freiburg, Medical Faculty – University of Freiburg, Germany
- Employee of Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany
- Parts of the work presented here were supported by the Research Commission at the Medical Faculty, University of Freiburg, Germany

- No known conflicts of interest regarding the content of this presentation

Agenda

- Background – risk of bias (RoB) assessment and AI
- Testing the LLM Claude for assessing RCTs with RoB 2
- Comparing our results to those of other studies
- Discussion & Conclusion

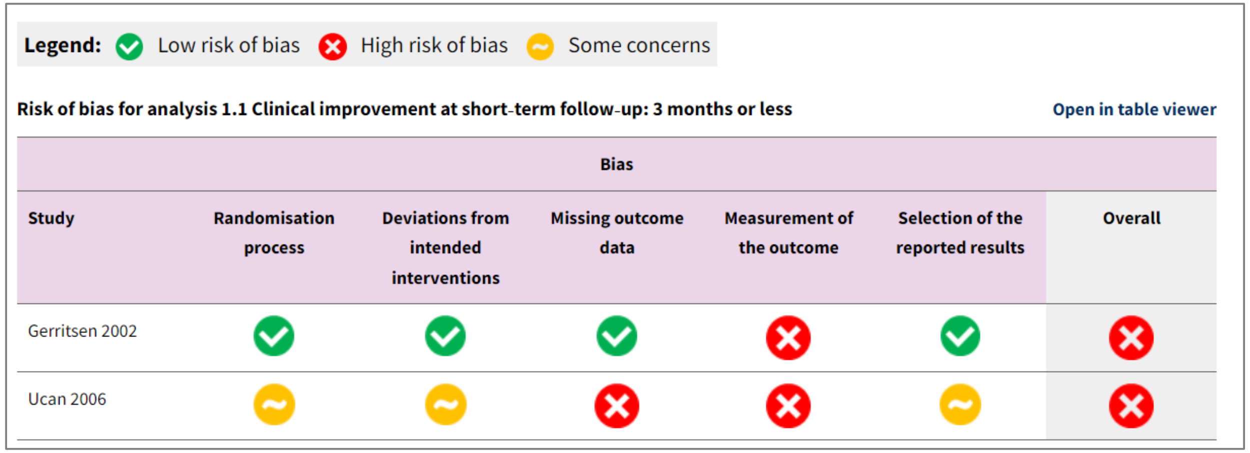
Background

Assess studies (included in a systematic review) for risk of bias

- Randomized controlled trials (RCTs)
  - Cochrane risk of bias tool (revised version; RoB 2)
  - Cochrane risk of bias tool (previous version; „RoB 1“)

https://www.riskofbias.info

- Non-randomized studies of interventions
  - ROBINS-I



Legend: ✓ Low risk of bias ✗ High risk of bias ~ Some concerns

Risk of bias for analysis 1.1 Clinical improvement at short-term follow-up: 3 months or less **Open in table viewer**

	Bias					
Study	**Randomisation process**	**Deviations from intended interventions**	**Missing outcome data**	**Measurement of the outcome**	**Selection of the reported results**	**Overall**
Gerritsen 2002	✓	✓	✓	✗	✓	✗
Ucan 2006	~	~	✗	✗	~	✗

Figure from Lusa et al. 2024, https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD001552.pub3/full

Background

ML <-> LLM

LLMs:

- trained on huge amounts of data
- predict the most likely next token (e.g. text)
- no task specific training necessary
- can be used without programming knowledge

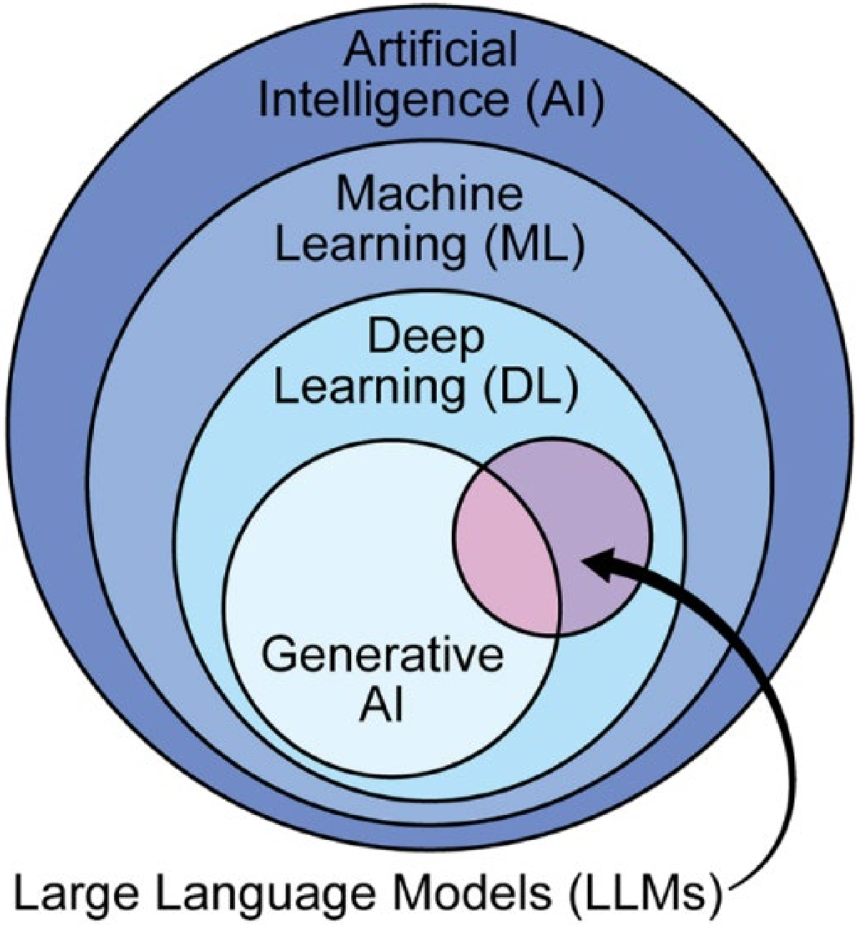
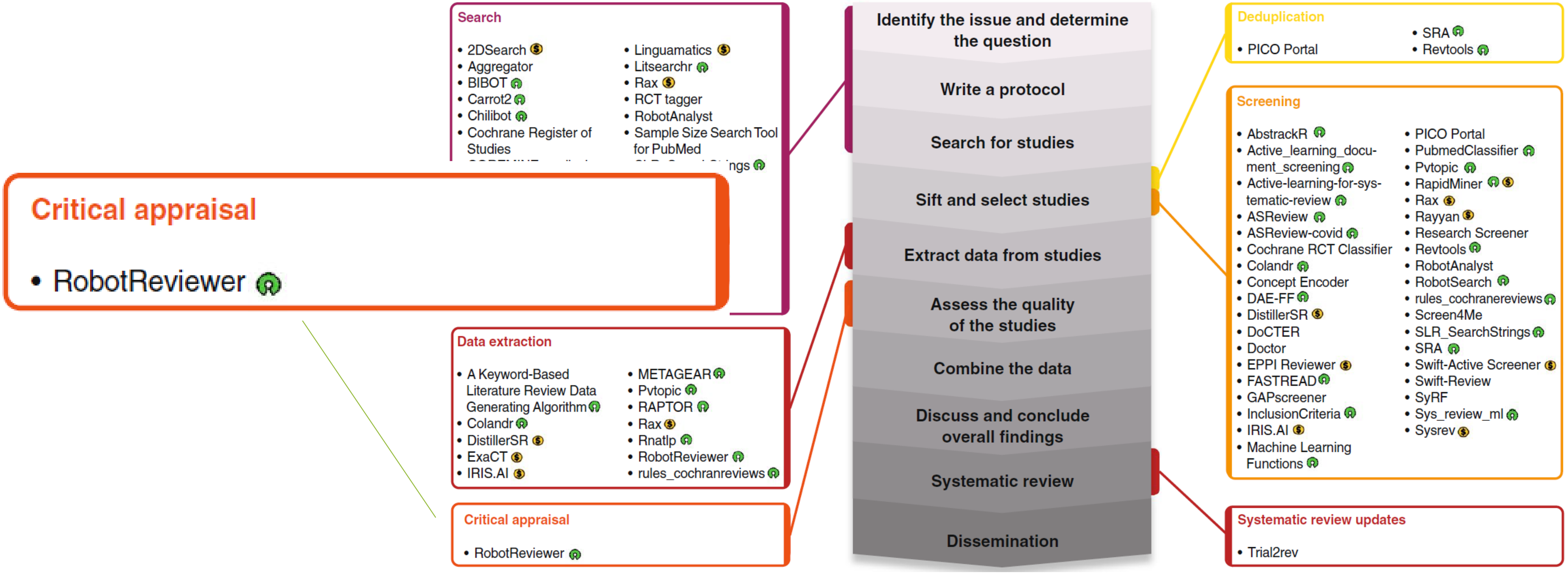


Figure from Shahab et al., 2023 https://doi.org/10.1177/17562848241227031

Background

Using ML for RoB assessment



Critical appraisal

- RobotReviewer

Search

- 2DSearch $
- Aggregator
- BIBOT
- Carrot2
- Chilibot
- Cochrane Register of Studies
- Linguamatics $
- Litsearchr
- Rax $
- RCT tagger
- RobotAnalyst
- Sample Size Search Tool for PubMed

Data extraction

- A Keyword-Based Literature Review Data Generating Algorithm
- Colandr
- DistillerSR $
- ExaCT $
- IRIS.AI $
- METAGEAR
- Pvtopic
- RAPTOR
- Rax $
- Rnatlp
- RobotReviewer
- rules_cochranreviews

Critical appraisal

- RobotReviewer

Identify the issue and determine the question

Write a protocol

Search for studies

Sift and select studies

Extract data from studies

Assess the quality of the studies

Combine the data

Discuss and conclude overall findings

Systematic review

Dissemination

Deduplication

- PICO Portal
- SRA
- Revtools

Screening

- AbstrackR
- Active_learning_document_screening
- Active-learning-for-systematic-review
- ASReview
- ASReview-covid
- Cochrane RCT Classifier
- Colandr
- Concept Encoder
- DAE-FF
- DistillerSR $
- DoCTER
- Doctor
- EPPI Reviewer $
- FASTREAD
- GAPScreener
- InclusionCriteria
- IRIS.AI $
- Machine Learning Functions
- PICO Portal
- PubmedClassifier
- Pvtopic
- RapidMiner $
- Rax $
- Rayyan $
- Research Screener
- Revtools
- RobotAnalyst
- RobotSearch
- rules_cochranereviews
- Screen4Me
- SLR_SearchStrings
- SRA
- Swift-Active Screener $
- Swift-Review
- SyRF
- Sys_review_ml
- Sysrev $

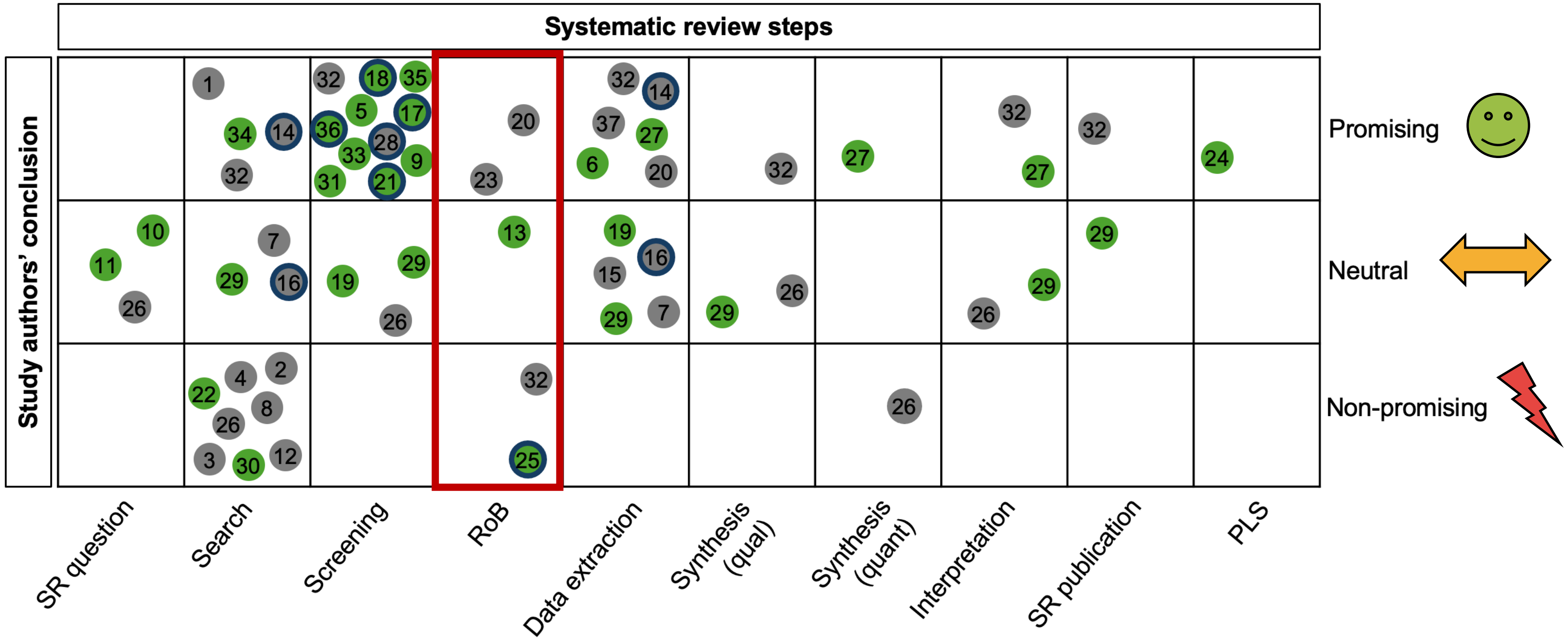
Systematic review updates

- Trial2rev

Cierco Jimenez et al., 2022 https://doi.org/10.1186/s12874-022-01805-4, (search conducted in **April 2021**)

Background

Using LLMs for RoB assessment

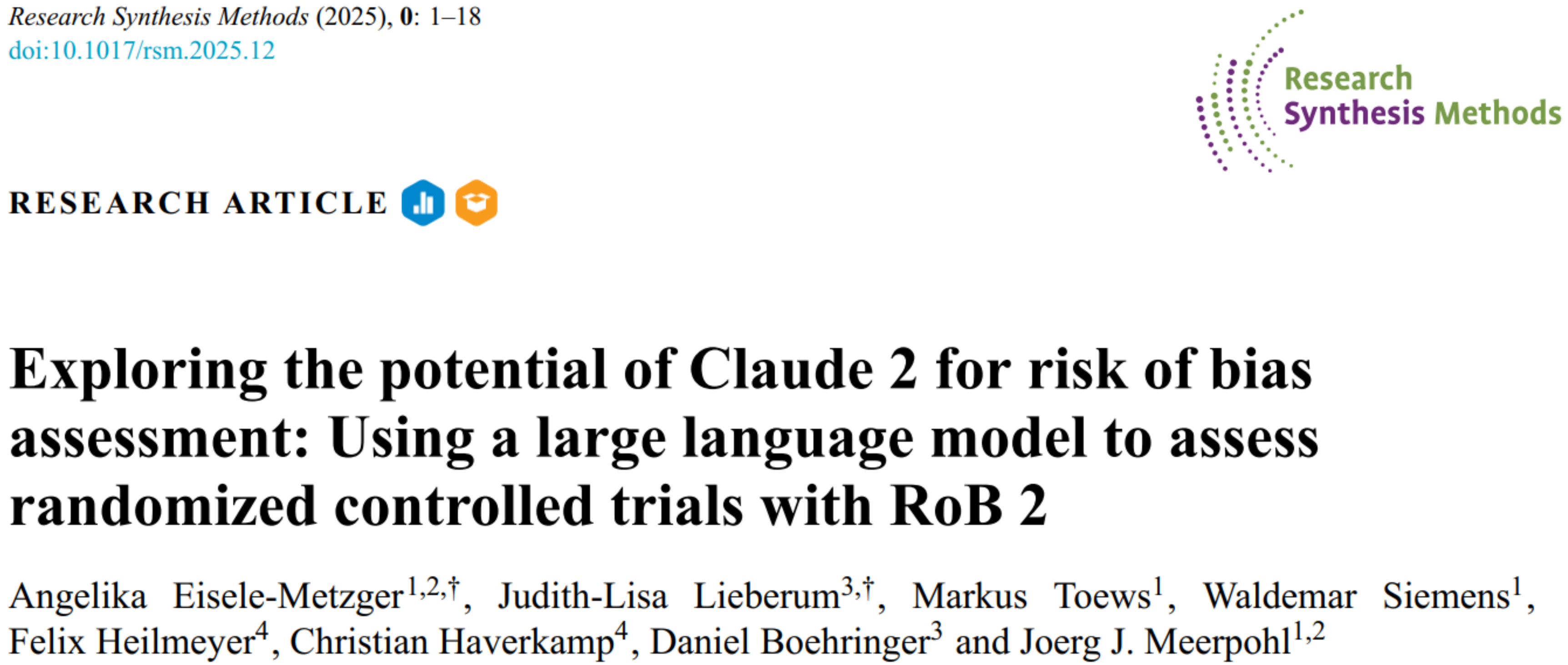


Study authors' conclusion	SR question	Search	Screening	RoB	Data extraction	Synthesis (qual)	Synthesis (quant)	Interpretation	SR publication	PLS
Promising		1, 34, 14, 32	32, 18, 35, 5, 17, 36, 28, 33, 9, 31, 21	20, 23	32, 14, 37, 27, 6, 20	32	27	32, 27	32	24
Neutral	10, 11, 26	7, 29, 16	29, 19, 26	13	19, 16, 15, 7, 29	29, 26		29, 26	29	
Non-promising		2, 4, 22, 8, 26, 3, 30, 12		32, 25			26			

Lieberum et al., 2025, https://doi.org/10.1016/j.jclinepi.2025.111746 (search conducted in **Februray 2024**)

green: „validation studies“, grey: other designs, blue circle: preprint articles

Testing Claude for assessing RCTs with RoB 2



Research Synthesis Methods (2025), **0**: 1–18
doi:10.1017/rsm.2025.12

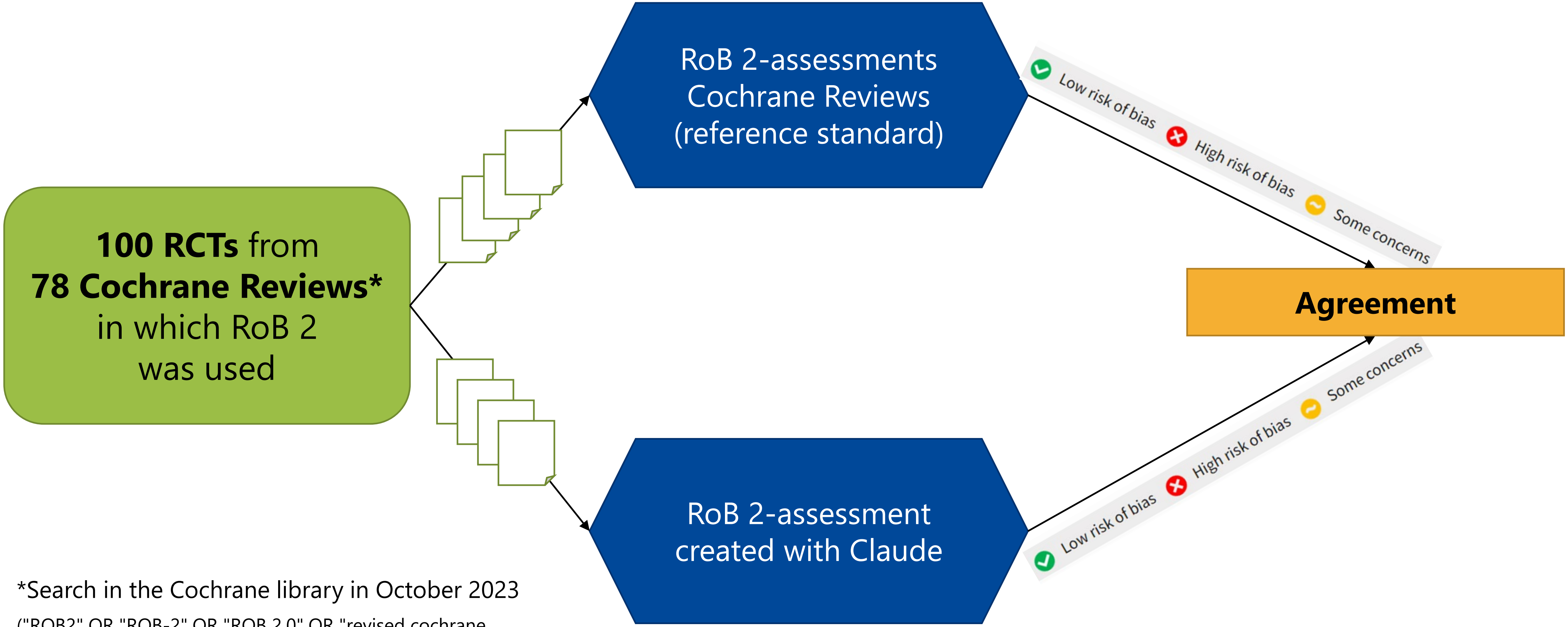
Research Synthesis Methods

RESEARCH ARTICLE

Exploring the potential of Claude 2 for risk of bias assessment: Using a large language model to assess randomized controlled trials with RoB 2

Angelika Eisele-Metzger[1,2,†], Judith-Lisa Lieberum[3,†], Markus Toews[1], Waldemar Siemens[1], Felix Heilmeyer[4], Christian Haverkamp[4], Daniel Boehringer[3] and Joerg J. Meerpohl[1,2]

Testing Claude for assessing RCTs with RoB 2



100 RCTs from
78 Cochrane Reviews*
in which RoB 2
was used

RoB 2-assessments
Cochrane Reviews
(reference standard)

RoB 2-assessment
created with Claude

Agreement

*Search in the Cochrane library in October 2023

("ROB2" OR "ROB-2" OR "ROB 2.0" OR "revised cochrane risk-of-bias" (all text), limit for publication date: January 2019 onwards, filter for review type “intervention”)

UNIVERSITÄTS
KLINIKUM FREIBURG

Testing Claude for assessing RCTs with RoB 2

Prompt

- Pilot phase: Prompt engineering using a sample of 30 RCTs from three Cochrane Reviews (excluded from the main testing)

Some of our prompt principles:

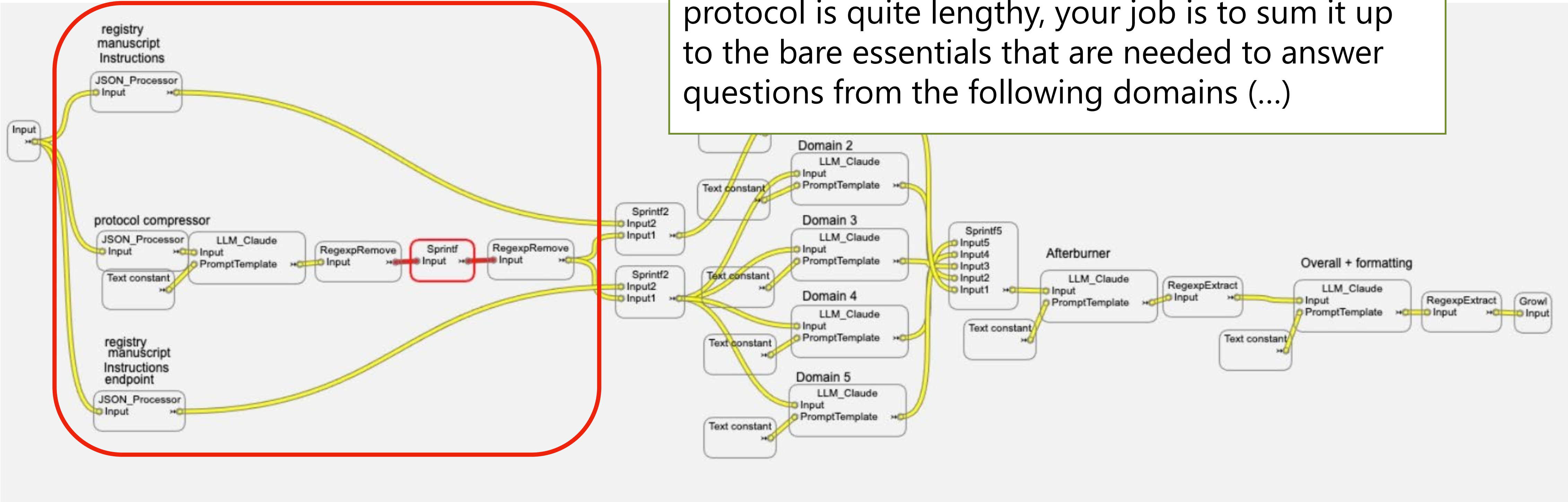
- Separate prompts for each domain (minimize reasoning complexity)
- Specify study outcome for which RoB should be assessed
- Include RoB 2 guidance & provide detailed instruction
- Do not mention the name of the tool (avoid simple recall of results / data contamination)
- Compress protocols & register entries

Testing Claude for assessing RCTs with RoB 2

Program - to automate the process of assembling the single prompts

Extract relevant details from protocol (if any)

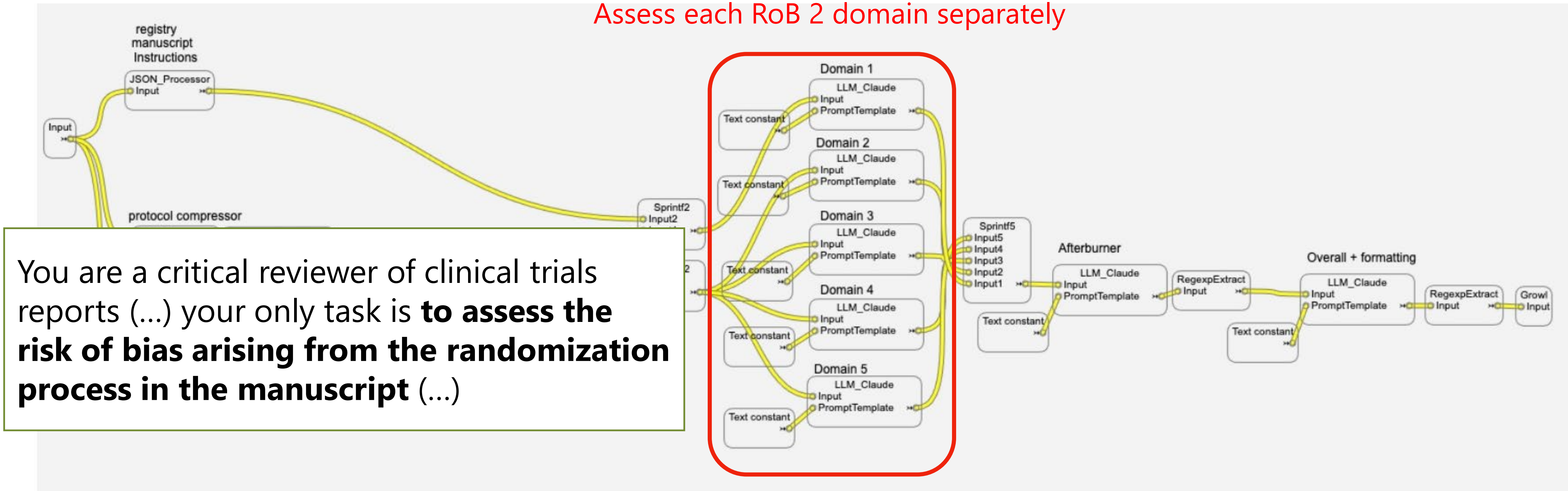
I give you a **clinical trials protocol** (…) As this protocol is quite lengthy, your job is to sum it up to the bare essentials that are needed to answer questions from the following domains (…)



https://github.com/daboe01/LLMPatchbay

Testing Claude for assessing RCTs with RoB 2

Program - to automate the process of assembling the single prompts



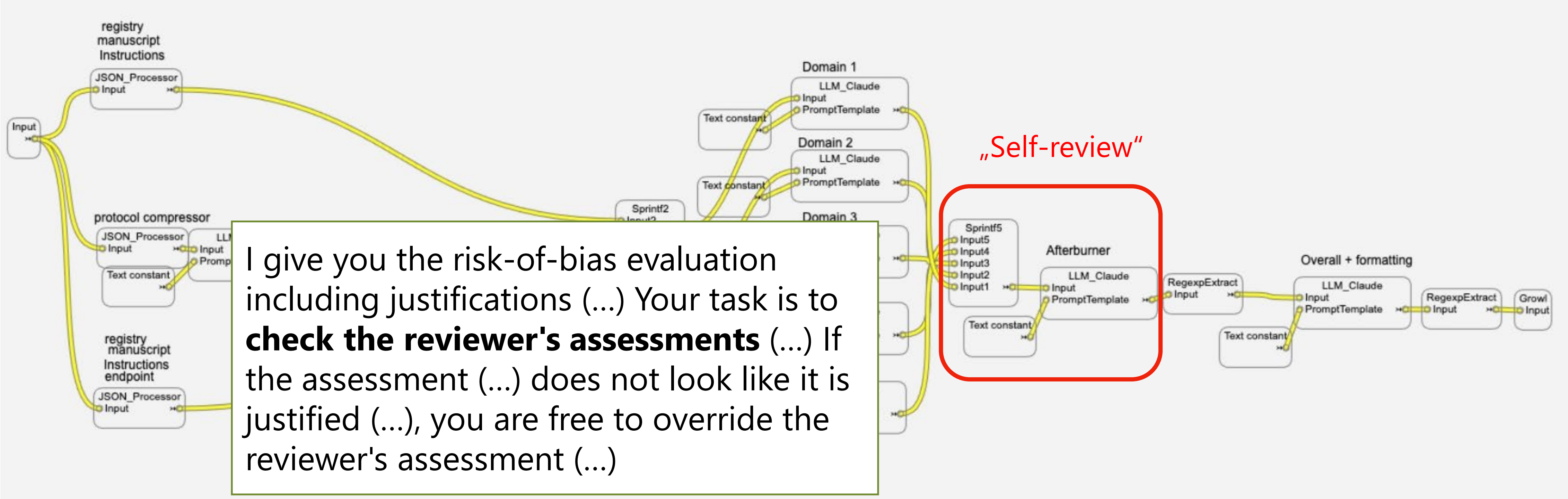
Assess each RoB 2 domain separately

You are a critical reviewer of clinical trials reports (...) your only task is **to assess the risk of bias arising from the randomization process in the manuscript** (...)

https://github.com/daboe01/LLMPatchbay

Testing Claude for assessing RCTs with RoB 2

Program - to automate the process of assembling the single prompts



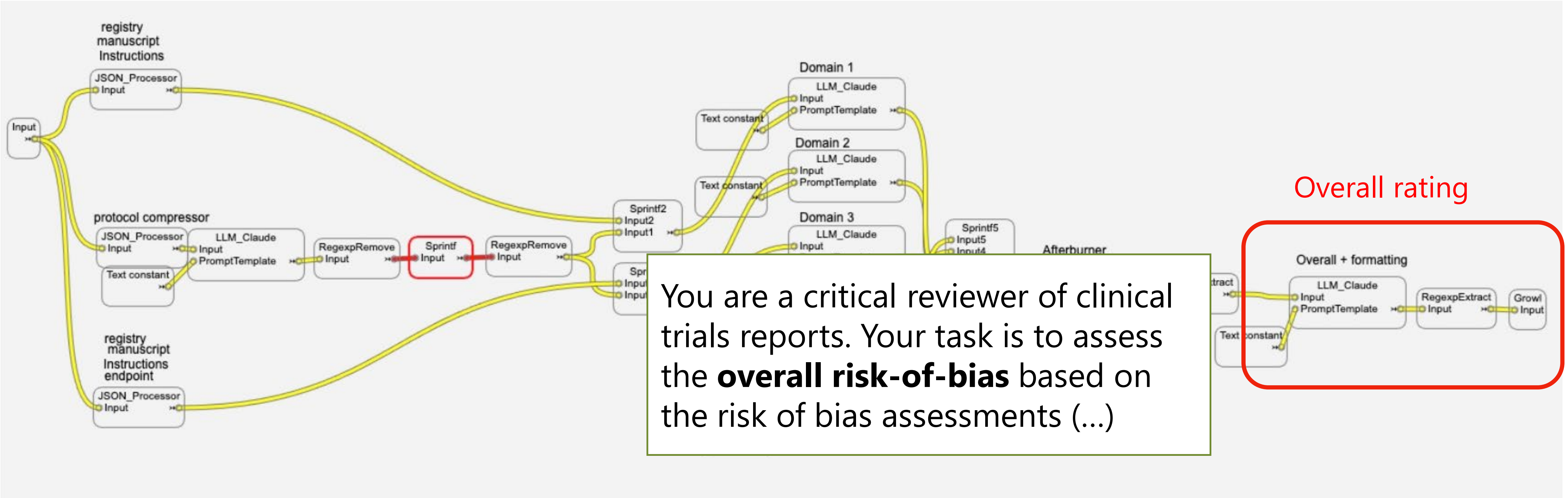
„Self-review“

I give you the risk-of-bias evaluation including justifications (...) Your task is to **check the reviewer's assessments** (...) If the assessment (...) does not look like it is justified (...), you are free to override the reviewer's assessment (...)

https://github.com/daboe01/LLMPatchbay

Testing Claude for assessing RCTs with RoB 2

Program - to automate the process of assembling the single prompts



Overall rating

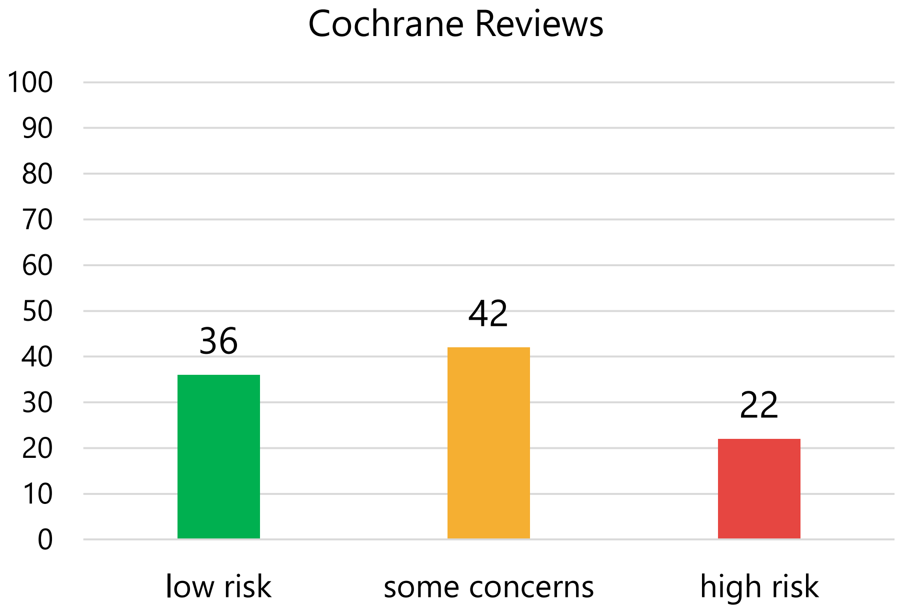
You are a critical reviewer of clinical trials reports. Your task is to assess the **overall risk-of-bias** based on the risk of bias assessments (...)

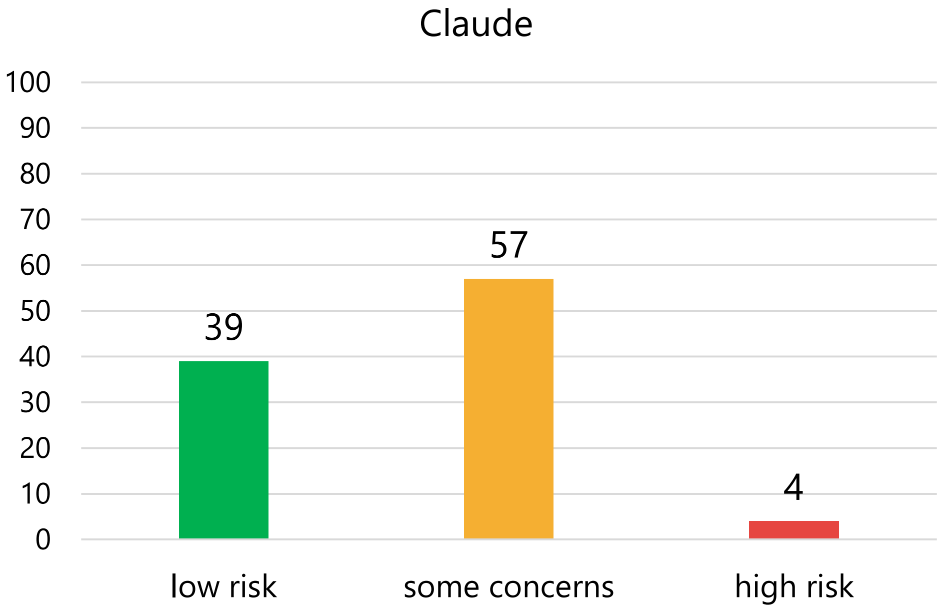
https://github.com/daboe01/LLMPatchbay

Testing Claude for assessing RCTs with RoB 2

Results for domain: „Overall judgement“

	CR: low risk	CR: some concerns	CR: high risk	Total
Claude: low risk	18	17	4	39
Claude: some concerns	18	22	17	57
Claude: high risk	0	3	1	4
Total	36	42	22	100





n= 100 RCTs

Testing Claude for assessing RCTs with RoB 2

Results: Agreement Claude – Cochrane Review authors

n= 100 RCTs

Domain	% agreement (accuracy)
D1 (“randomization”)	65%
D2 (“deviations from interventions”)	63%
D3 (“missing data”)	70%
D4 (“outcome measurement”)	71%
D5 (“selective reporting”)	58%
Overall	41%

Testing Claude for assessing RCTs with RoB 2

Results: Agreement Claude – Cochrane Review authors

n= 100 RCTs

Domain	% agreement (accuracy)	Weighted Cohen's Kappa (95%-CI)
D1 ("randomization")	65%	0.11 (-0.08; 0.29)
D2 ("deviations from interventions")	63%	0.12 (-0.08; 0.32)
D3 ("missing data")	70%	0.31 (0.10; 0.52)
D4 ("outcome measurement")	71%	0.15 (-0.11; 0.41)
D5 ("selective reporting")	58%	0.10 (-0.10; 0.31)
Overall	41%	0.22 (0.06; 0.38)

Rough interpretation - kappa

"slight" (0.00-0.20)
"fair" (0.21-0.40)

Testing Claude for assessing RCTs with RoB 2

Review of discrepancies:

Domain	**Issues with…**
D1 ("randomization")	
D2 ("deviations from interventions")	
D3 ("missing data")	
D4 ("outcome measurement")	
D5 ("selective reporting")	

Testing Claude for assessing RCTs with RoB 2

Review of discrepancies:

Domain	**Issues with…**
D1 (“randomization”)	… detecting absence of concealment of allocation
D2 (“deviations from interventions”)	
D3 (“missing data”)	
D4 (“outcome measurement”)	
D5 (“selective reporting”)	

Testing Claude for assessing RCTs with RoB 2

Review of discrepancies:

Domain	Issues with...
D1 ("randomization")	... detecting absence of concealment of allocation
D2 ("deviations from interventions")	... interpreting the potential impact of lack of blinding (of participants or carers)
D3 ("missing data")	
D4 ("outcome measurement")	
D5 ("selective reporting")	

Testing Claude for assessing RCTs with RoB 2

Review of discrepancies:

Domain	**Issues with…**
D1 ("randomization")	… detecting absence of concealment of allocation
D2 ("deviations from interventions")	… interpreting the potential impact of lack of blinding (of participants or carers)
D3 ("missing data")	… interpreting the potential impact of missing data … detecting available data
D4 ("outcome measurement")	
D5 ("selective reporting")	

Testing Claude for assessing RCTs with RoB 2

Review of discrepancies:

Domain	**Issues with…**
D1 (“randomization”)	… detecting absence of concealment of allocation
D2 (“deviations from interventions”)	… interpreting the potential impact of lack of blinding (of participants or carers)
D3 (“missing data”)	… interpreting the potential impact of missing data … detecting available data
D4 (“outcome measurement”)	… detecting absence of assessor blinding … interpreting the impact of non-blinded assessors
D5 (“selective reporting”)	

Testing Claude for assessing RCTs with RoB 2

	RCT013 - Cochrane Review	RCT013 - Claude
D4	**high risk**	**low risk**
	The ADL Staircase is considered **appropriate**, and there were **no differences in the measurement** or ascertainment between groups. The **assessors were not blinded**, and it was therefore considered **likely** that knowledge of the intervention could influence the outcome, given the likely strong belief in the benefits of the intervention ward.	The outcome measurement method was **appropriate**, did **not differ between groups**, and **assessor knowledge** of interventions was **unlikely** to impact the standardized ADL Staircase ratings.

Testing Claude for assessing RCTs with RoB 2

Review of discrepancies:

Domain	**Issues with…**
D1 (“randomization”)	… detecting absence of concealment of allocation
D2 (“deviations from interventions”)	… interpreting the potential impact of lack of blinding (of participants or carers)
D3 (“missing data”)	… interpreting the potential impact of missing data … detecting available data
D4 (“outcome measurement”)	… detecting absence of assessor blinding … interpreting the impact of non-blinded assessors
D5 (“selective reporting”)	… detecting absence (or presence) of pre-specified protocols/analysis plans

Overall judgement: largely followed the guidance (only 2/100 Claude judgements deviated from the given algorithm)

Comparing our results to those of other studies

→ „Traditional“ ML-approaches (RobotReviewer)

→ Other studies using LLM-approaches

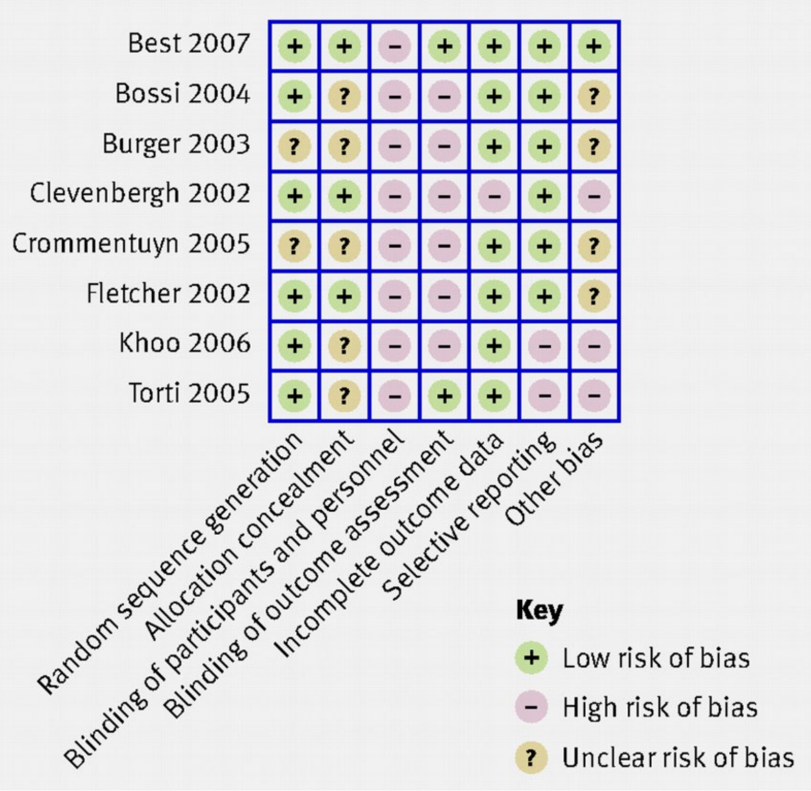
→ Humans

Comparing our results to those of other studies

„Traditional“ ML-approaches:

RobotReviewer versus humans (RoB 1, D1 – D4)

	Tian et al. 2024
RCTs assessed (n)	1955
Reference standard	Cochrane
% agreement (range)	63 – 83%
Cohen‘s Kappa (range)	0.25 – 0.59



	Random sequence generation	Allocation concealment	Blinding of participants and personnel	Blinding of outcome assessment	Incomplete outcome data	Selective reporting	Other bias
Best 2007	+	+	–	+	+	+	+
Bossi 2004	+	?	–	–	+	+	?
Burger 2003	?	?	–	–	+	+	?
Clevenbergh 2002	+	+	–	–	–	+	–
Crommentuyn 2005	?	?	–	–	+	+	?
Fletcher 2002	+	+	–	–	+	+	?
Khoo 2006	+	?	–	–	+	–	–
Torti 2005	+	?	–	+	+	–	–

Key

+ Low risk of bias

– High risk of bias

? Unclear risk of bias

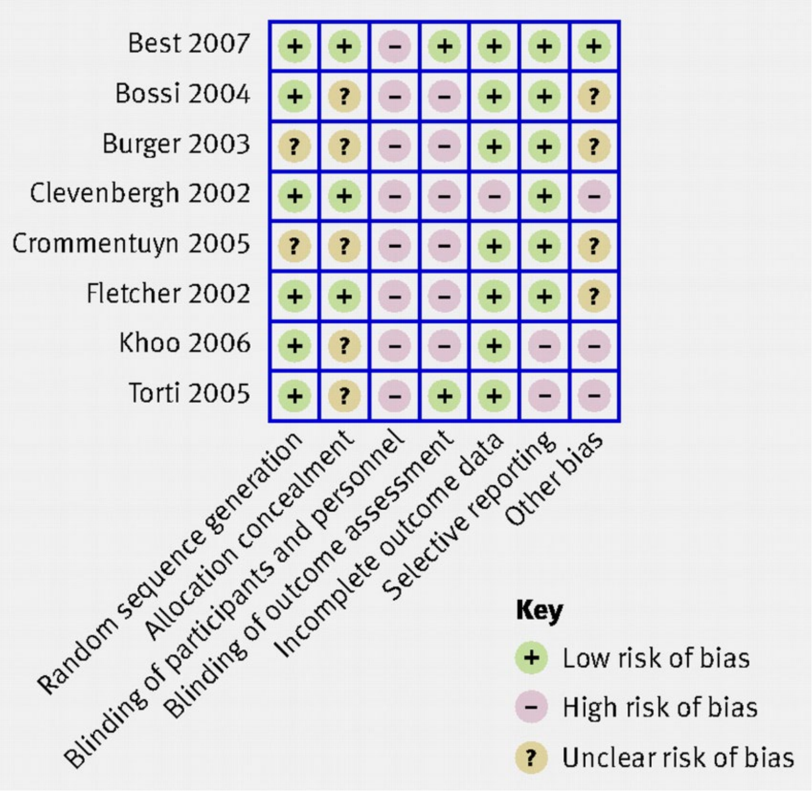
Figure from Higgins et al. 2011, https://doi.org/10.1136/bmj.d5928

Comparing our results to those of other studies

„Traditional“ ML-approaches:

RobotReviewer versus humans (RoB 1, D1 – D4)

	Tian et al. 2024	Hirt et al. 2021
RCTs assessed (n)	1955	190
Reference standard	Cochrane	Cochrane
% agreement (range)	63 – 83%	50 – 87%
Cohen’s Kappa (range)	0.25 – 0.59	0.04 – 0.60



	Random sequence generation	Allocation concealment	Blinding of participants and personnel	Blinding of outcome assessment	Incomplete outcome data	Selective reporting	Other bias
Best 2007	+	+	−	+	+	+	+
Bossi 2004	+	?	−	−	+	+	?
Burger 2003	?	?	−	−	+	+	?
Clevenbergh 2002	+	+	−	−	−	+	−
Crommentuyn 2005	?	?	−	−	+	+	?
Fletcher 2002	+	+	−	−	+	+	?
Khoo 2006	+	?	−	−	+	−	−
Torti 2005	+	?	−	+	+	−	−

Key

+ Low risk of bias

− High risk of bias

? Unclear risk of bias

Figure from Higgins et al. 2011, https://doi.org/10.1136/bmj.d5928

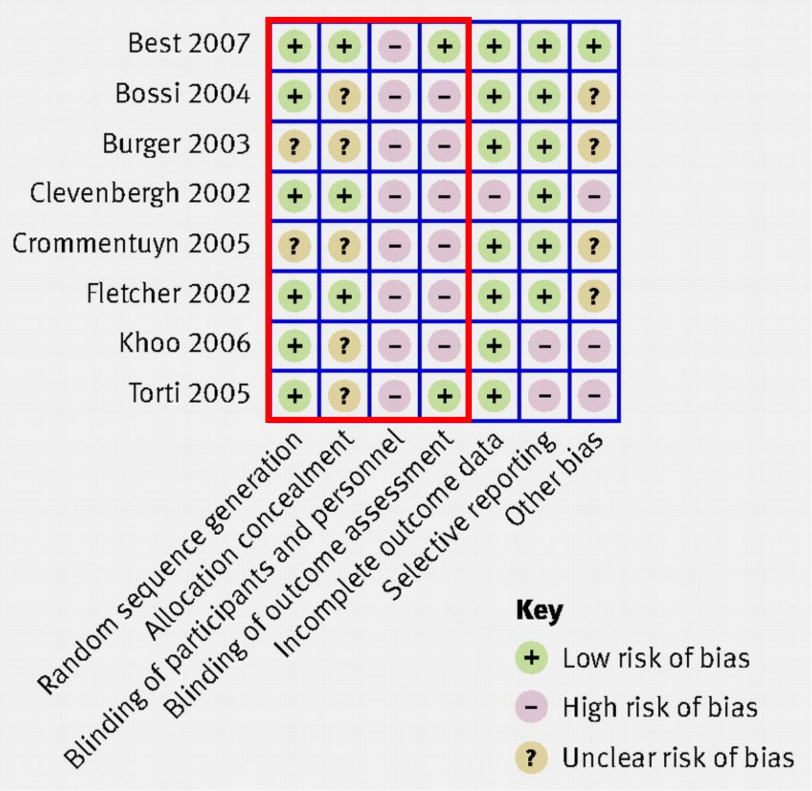
Comparing our results to those of other studies

„Traditional“ ML-approaches:

RobotReviewer versus humans (RoB 1, D1 – D4)

	Tian et al. 2024	Hirt et al. 2021	Armijo-Olivo et al. 2020
RCTs assessed (n)	1955	190	372
Reference standard	Cochrane	Cochrane	Own judgements
% agreement (range)	63 – 83%	50 – 87%	56 – 81%
Cohen's Kappa (range)	0.25 – 0.59	0.04 – 0.60	0.06 – 0.62

D1 – D4:
% agreement: 63 – 71%
Cohen's Kappa: 0.11 – 0.31



	Random sequence generation	Allocation concealment	Blinding of participants and personnel	Blinding of outcome assessment	Incomplete outcome data	Selective reporting	Other bias
Best 2007	+	+	−	+	+	+	+
Bossi 2004	+	?	−	−	+	+	?
Burger 2003	?	?	−	−	+	+	?
Clevenbergh 2002	+	+	−	−	−	+	−
Crommentuyn 2005	?	?	−	−	+	+	?
Fletcher 2002	+	+	−	−	+	+	?
Khoo 2006	+	?	−	−	+	−	−
Torti 2005	+	?	−	+	+	−	−

Key

+	Low risk of bias
−	High risk of bias
?	Unclear risk of bias

Figure from Higgins et al. 2011, https://doi.org/10.1136/bmj.d5928

Comparing our results to those of other studies

Other studies using LLM-approaches

	Pitre et al. 2023
RoB tool used	RoB 2
LLM(s) used	GPT-4
Studies assessed (n)	157
Reference standard	Cochrane
% agreement (range)	11 – 29%
Cohen's Kappa (range)	0.11 – 0.29*

"We recommend systematic reviewers avoid using ChatGPT to perform risk of bias assessments."
(Pitre et al. 2023)

Comparing our results to those of other studies

Other studies using LLM-approaches

	Pitre et al. 2023	**Hasan et al. 2024**
RoB tool used	RoB 2	ROBINS-I
LLM(s) used	GPT-4	GPT-4
Studies assessed (n)	157	307
Reference standard	Cochrane	Cochrane
% agreement (range)	11 – 29%	31 – 71%
Cohen's Kappa (range)	0.11 – 0.29*	0.02 – 0.28*

"Considering the agreement level with a human reviewer in the case study, pairing AI with an independent human reviewer remains required at present."
(Hasan et al. 2024)

Comparing our results to those of other studies

Other studies using LLM-approaches

	Pitre et al. 2023	Hasan et al. 2024	Šuster et al. 2024
RoB tool used	RoB 2	ROBINS-I	RoB 2
LLM(s) used	GPT-4	GPT-4	FlanT5XL; GPT-3.5-turbo; Meditron-70B, Med42
Studies assessed (n)	157	307	218
Reference standard	Cochrane	Cochrane	Cochrane
% agreement (range)	11 – 29%	31 – 71%	F1 (range) 0.00 – 0.20
Cohen's Kappa (range)	0.11 – 0.29*	0.02 – 0.28*	

"Using LLMs as an assisting technology for assessing RoB 2 thus currently seems beyond their reach."
(Šuster et al. 2024)

Comparing our results to those of other studies

"…demonstrated substantial accuracy and consistency in evaluating RCTs, suggesting their potential as supportive tools in systematic review processes."
(Lai et al. 2024)

Other studies using LLM-approaches

	Pitre et al. 2023	Hasan et al. 2024	Šuster et al. 2024	Lai et al. 2024	Lai et al. 2025
RoB tool used	RoB 2	ROBINS-I	RoB 2	Modified RoB 1 tool	Modified RoB 1 tool
LLM(s) used	GPT-4	GPT-4	FlanT5XL; GPT-3.5-turbo; Meditron-70B, Med42	GPT; Claude (versions?)	Claude 3.5-sonnet; Moonshot-v1-128k (Kimi-Chat)
Studies assessed (n)	157	307	218	30	107
Reference standard	Cochrane	Cochrane	Cochrane	Own judgements	Own judgements
% agreement (range)	11 – 29%	31 – 71%	F1 (range) 0.00 – 0.20	57 – 98%	88 – 100%
Cohen's Kappa (range)	0.11 – 0.29*	0.02 – 0.28*		0.54 – 0.96	0.42 – 1.00

Comparing our results to those of other studies

Humans – RoB 2

	Minozzi et al. 2020
RCTs assessed (n)	70
% agreement (range)	-
Fleiss' Kappa (range)	0.04 – 0.45

Comparing our results to those of other studies

Humans – RoB 2

	Minozzi et al. 2020	Minozzi et al. 2022 (before calibration)	Minozzi et al. 2022 (after calibration)
RCTs assessed (n)	70	5	11
% agreement (range)	-	-	-
Fleiss' Kappa (range)	0.04 – 0.45	-0.24 – 0.30	-0.01 – 0.93

Minozzi et al. 2022

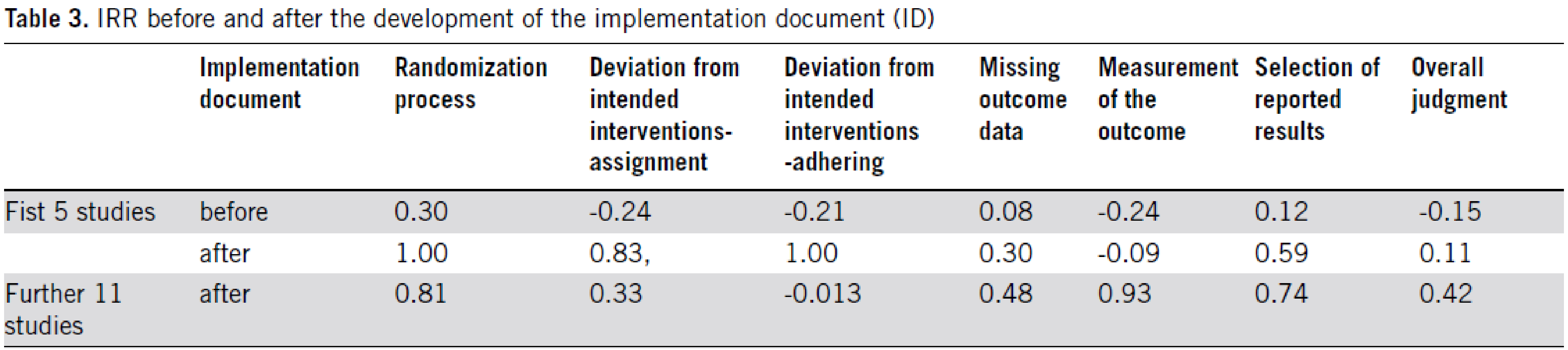
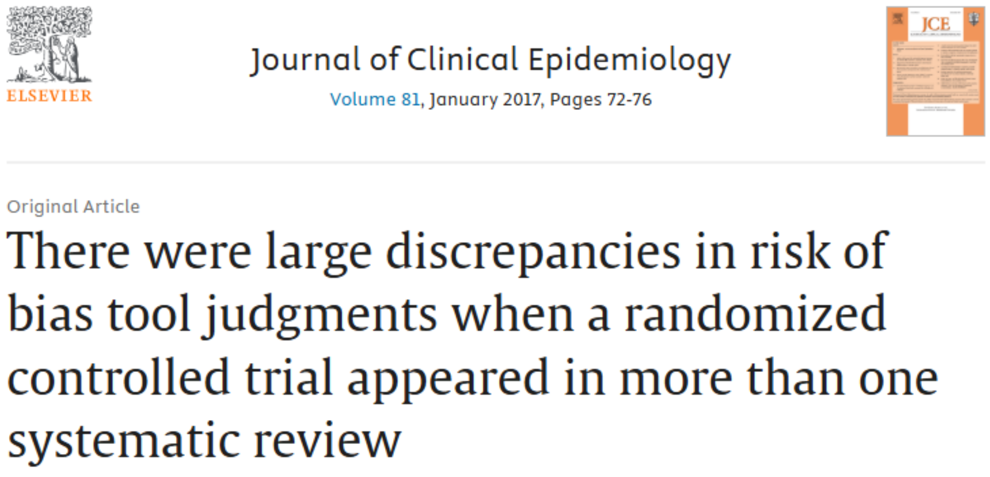


Table 3. IRR before and after the development of the implementation document (ID)

	Implementation document	Randomization process	Deviation from intended interventions-assignment	Deviation from intended interventions -adhering	Missing outcome data	Measurement of the outcome	Selection of reported results	Overall judgment
Fist 5 studies	before	0.30	-0.24	-0.21	0.08	-0.24	0.12	-0.15
	after	1.00	0.83,	1.00	0.30	-0.09	0.59	0.11
Further 11 studies	after	0.81	0.33	-0.013	0.48	0.93	0.74	0.42

Comparing our results to those of other studies

Humans – RoB 1



ELSEVIER

Journal of Clinical Epidemiology

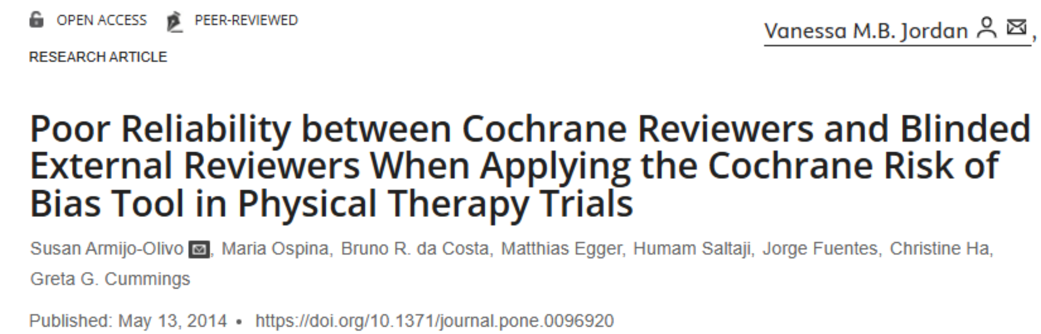
Volume 81, January 2017, Pages 72-76

Original Article

There were large discrepancies in risk of bias tool judgments when a randomized controlled trial appeared in more than one systematic review

Vanessa M.B. Jordan, Sarah F. Lensen, Cyn

https://doi.org/10.1016/j.jclinepi.2016.08.012



PLOS One

OPEN ACCESS PEER-REVIEWED

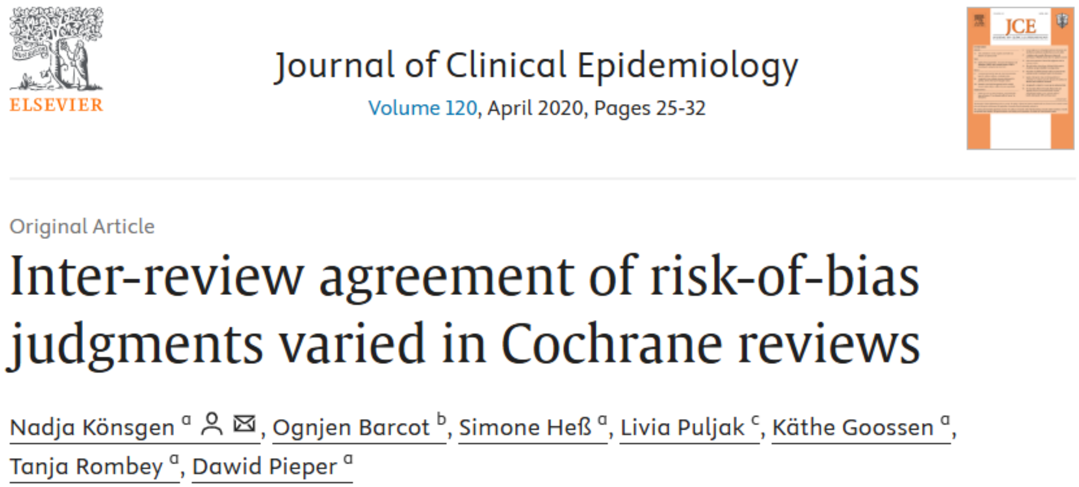
RESEARCH ARTICLE

Poor Reliability between Cochrane Reviewers and Blinded External Reviewers When Applying the Cochrane Risk of Bias Tool in Physical Therapy Trials

Susan Armijo-Olivo, Maria Ospina, Bruno R. da Costa, Matthias Egger, Humam Saltaji, Jorge Fuentes, Christine Ha, Greta G. Cummings

Published: May 13, 2014 • https://doi.org/10.1371/journal.pone.0096920

https://doi.org/10.1371/journal.pone.0096920



ELSEVIER

Journal of Clinical Epidemiology

Volume 120, April 2020, Pages 25-32

Original Article

Inter-review agreement of risk-of-bias judgments varied in Cochrane reviews

Nadja Könsgen [a], Ognjen Barcot [b], Simone Heß [a], Livia Puljak [c], Käthe Goossen [a], Tanja Rombey [a], Dawid Pieper [a]

https://doi.org/10.1016/j.jclinepi.2019.12.016

Discussion & Conclusion

Next steps / Open questions

- Use expert reference standards for testing? (could also introduce bias?)
- Other forms of support than creating full RoB judgements?
- Use only RoB domains that are most promising?
- Strive for high methodological quality
- Release of the new RoB tool ROBUST RCT (Wang et al. 2025)

-> Currently, using LLMs for RoB assessment is not recommended

Graphic created using DALL-E

Acknowledgements

Institute for Evidence in Medicine & Cochrane Germany	**Eye Center,** Medical Center & Medical Faculty – University of Freiburg, Germany	**Institute for Digitalization in Medicine,** Medical Center & Medical Faculty – University of Freiburg, Germany
Markus Töws	Dr Judith-Lisa Lieberum	Felix Heilmeyer
Dr Waldemar Siemens	Prof Daniel Böhringer	Dr Christian Haverkamp
Prof Jörg Meerpohl		

Contact:

angelika.eisele-metzger@uniklinik-freiburg.de

References

- Armijo-Olivo S, Craig R, Campbell S. Comparing machine and human reviewers to evaluate the risk of bias in randomized controlled trials. Res Synth Methods. 2020;11(3):484-493. doi: 10.1002/jrsm.1398.
- Armijo-Olivo S, Ospina M, da Costa BR, Egger M, Saltaji H, et al. Poor Reliability between Cochrane Reviewers and Blinded External Reviewers When Applying the Cochrane Risk of Bias Tool in Physical Therapy Trials. PLOS ONE. 2014;9(5): e96920. https://doi.org/10.1371/journal.pone.0096920
- Cierco Jimenez R, Lee T, Rosillo N, Cordova R, Cree IA, Gonzalez A, Indave Ruiz BI. Machine learning computational tools to assist the performance of systematic reviews: A mapping review. BMC Medical Research Methodology. 2022;22(1):322.
- Chicco D, Warrens MJ, Jurman G. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. IEEE Access. 2021;9:78368-81.
- Eisele-Metzger A, Lieberum J-L, Toews M, Siemens W, Heilmeyer F, Haverkamp C, et al. Exploring the potential of Claude 2 for risk of bias assessment: Using a large language model to assess randomized controlled trials with RoB 2. Research Synthesis Methods. 2025:1-18.
- Hasan B, Saadi S, Rajjoub NS, Hegazi M, Al-Kordi M, Fleti F, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. BMJ Evidence-Based Medicine. 2024:bmjebm-2023-112597.
- Higgins J P T, Altman D G, Gotzsche P C, Jüni P, Moher D, Oxman A D et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials BMJ 2011; 343 :d5928 doi:10.1136/bmj.d5928
- Hirt J, Meichlinger J, Schumacher P, Mueller G. Agreement in Risk of Bias Assessment Between RobotReviewer and Human Reviewers: An Evaluation Study on Randomised Controlled Trials in Nursing-Related Cochrane Reviews. J Nurs Scholarsh. 2021;53(2):246-254. doi: 10.1111/jnu.12628.
- Jordan VMB, Lensen SF, Farquhar CM. There were large discrepancies in risk of bias tool judgments when a randomized controlled trial appeared in more than one systematic review. Journal of Clinical Epidemiology. 2017;81: 72-76. https://doi.org/10.1016/j.jclinepi.2016.08.012.
- Könsgen N, Barcot O, Heß S, Puljak L, Goossen K, Rombey T, Pieper D. Inter-review agreement of risk-of-bias judgments varied in Cochrane reviews. Journal of Clinical Epidemiology, 2020;120:25-32. https://doi.org/10.1016/j.jclinepi.2019.12.016.
- Lai H, Ge L, Sun M, Pan B, Huang J, Hou L, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA Network Open. 2024;7(5):e2412687-e.
- Lai H, Liu J, Bai C, Liu H, Pan B, Luo X, et al. Language models for data extraction and risk of bias assessment in complementary medicine. npj Digital Medicine. 2025;8(1):74.
- Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33(1):159-74.
- Lieberum JL, Töws M, Metzendorf MI, Heilmeyer F, Siemens W, Haverkamp C, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review. J Clin Epidemiol. 2025;181:111746.

References

- Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. Journal of Clinical Epidemiology. 2020;126:37-44.
- Minozzi S, Dwan K, Borrelli F, Filippini G. Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. Journal of Clinical Epidemiology. 2022;141:99-105.
- Pitre T, Jassal T, Talukdar JR, Shahab M, Ling M, Zeraatkar D. ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: A methods study. medRxiv. 2024:2023.11.19.23298727.
- Šuster S, Baldwin T, Verspoor K. Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. Research Synthesis Methods. 2024.
- Tian Y, Yang X, Doi SA, Furuya-Kanamori L, Lin L, Kwong JSW, Xu C. Towards the automatic risk of bias assessment on randomized controlled trials: A comparison of RobotReviewer and humans. Res Synth Methods. 2024 Nov;15(6):1111-1119. doi: 10.1002/jrsm.1761.
- Wang Y, Keitz S, Briel M, Glasziou P, Brignardello-Petersen R, Siemieniuk R A C et al. Development of ROBUST-RCT: Risk Of Bias instrument for Use in SysTematic reviews-for Randomised Controlled Trials BMJ 2025; 388 :e081199 doi:10.1136/bmj-2024-081199