Automating risk-of-bias assessment with generative AI

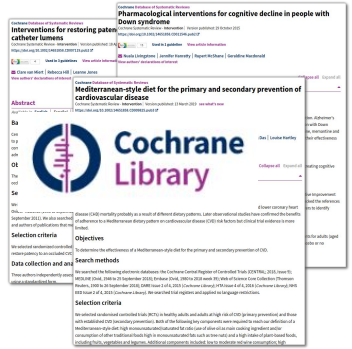
Simon Šuster, Tim Baldwin, and Karin Verspoor

13 September 2024

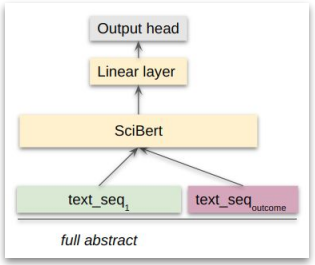
Conflict of interest disclosure

I have no actual or potential conflict of interest in relation to this presentation.

Introduction



a “traditional” automated risk-of-bias (RoB) assessment process…



high

low/unclear

RCTs with RoB annotations

Train a classifier and use pretrained language models to represent text inputs

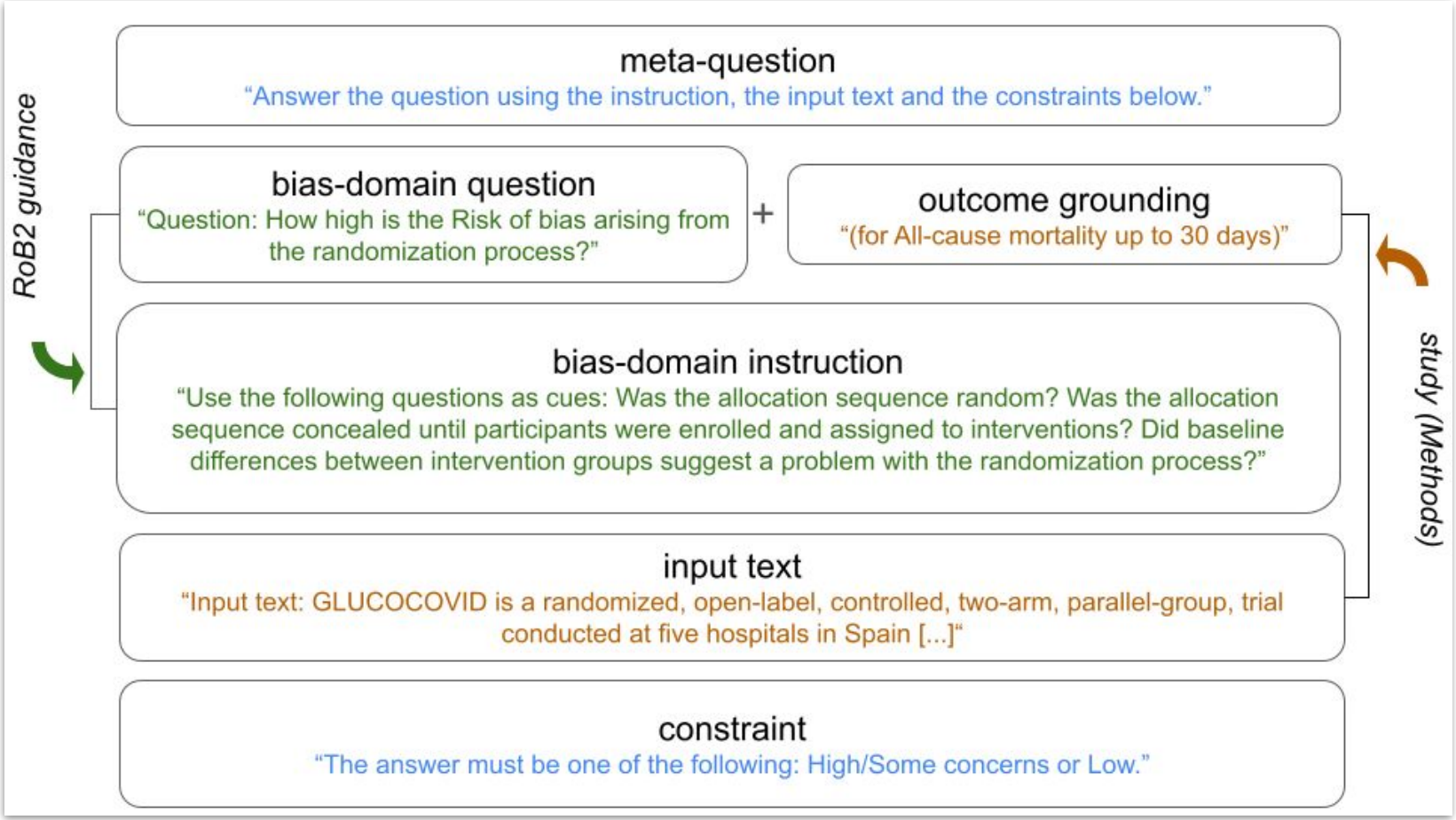
Predict RoB

… requires extensive ground-truth annotations

Can LLMs assess RoB effectively (without specific training)?

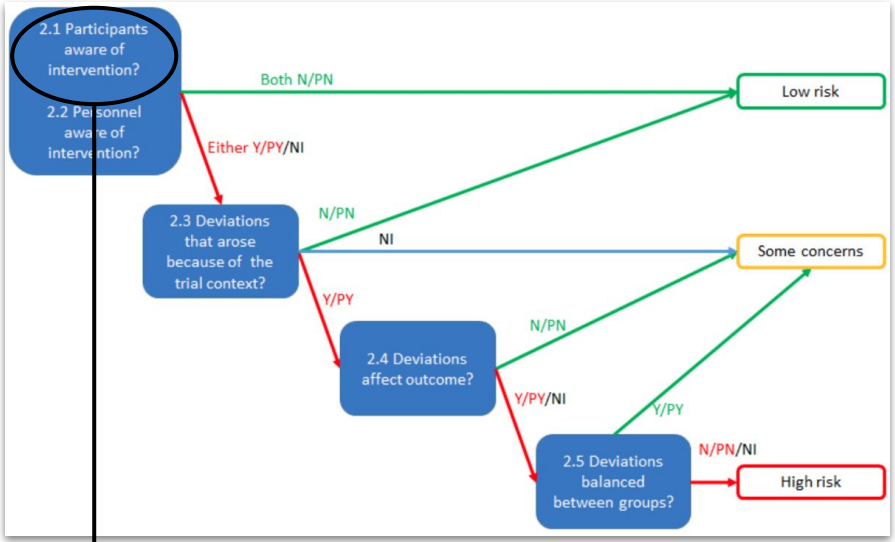
Can we use them to overcome data scarcity for RoB v2?

Prompting an LLM for RoB prediction

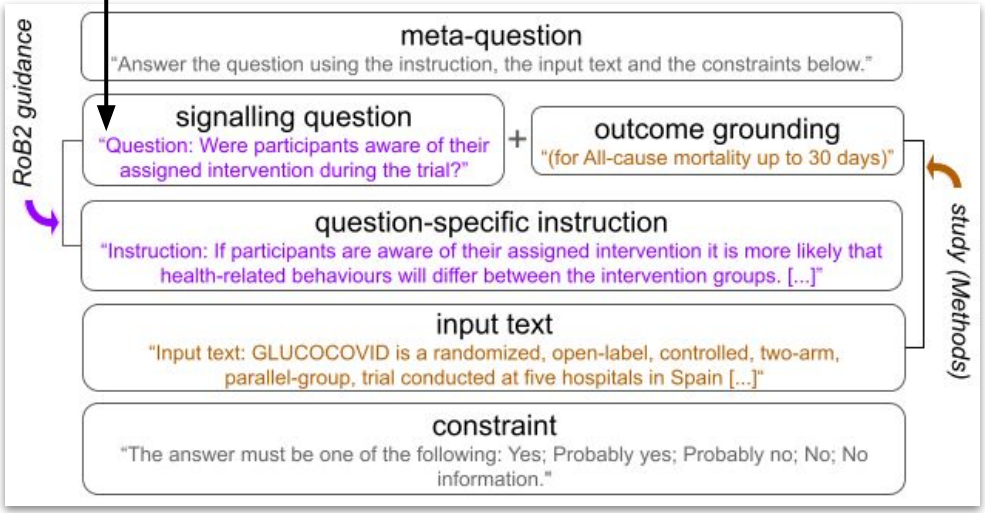


Decomposition

An RoB label is obtained via a decision algorithm proposed by RoB2 authors

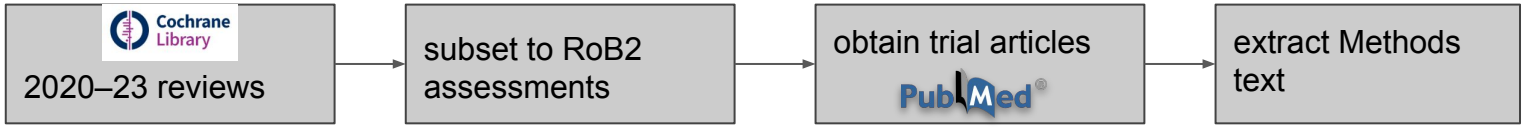


Prompting an LLM to answer a signalling question



LLM experiments

Data



- 6,000 RoB2 decisions, 218 studies
- 71% low risk, 29% high & some concerns
- some signalling question answers (9 studies)

Models

General (ChatGPT, FlanT5XL) and medical (Meditron-70B, Med42-70B) queried on OpenAI / university servers

Fewshot learning

Exemplars: Author justifications with RCT excerpts supporting a decision

Additional:

Simplified prompts: Omit instructions, use short questions

RoB1 instead of RoB2: Known predictive performance

LLM finetuning: Parameter efficient task-specific adaptation

Results

None of the LLMs can make accurate RoB predictions!

Overall F1 score for binary classification is ~.5
Little affected by bias domain, LLM type, and prompting strategy (decomposed or not)
On a par with trivial baselines

Fewshot learning & prompt simplification make little difference

RoB1 performance also low
In the range .3–.6 F1, cf. RobotReviewer at ~.7 F1

But LLM finetuning is promising
Observe improvements of ~.2 F1
Sensitive to training data sampling

Observations and future work

Overconfidence
Models reluctant to output “No information” in signalling question answering

RoB2 guidelines
High annotator agreement reported for authors with content/methodological expertise
Similar results with RoB1 guidelines

Input augmentation
Trial protocols and registry entries could enhance assessment for certain RoB domains

Ground-truth data for signalling questions
Better evaluation, more targeted prompt development for difficult questions

Finetuning regimes
RoB1 data to support RoB2 assessment

More information

Zero- and Few-Shot Prompting of Generative Large Language Models Provides Weak Assessment of Risk of Bias in Clinical Trials. Simon Šuster, Timothy Baldwin, Karin Verspoor. *Research Synthesis Methods*, 2024.

Related work: **ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: A methods study.** Tyler Pitre et al. *medRxiv* (2023), 2023.

See **simonsuster.github.io/evidence_grading/**
for more on evidence assessment