



EVALUATING CHATGPT'S EFFICACY AND EFFICIENCY IN THE RISK OF BIAS ASSESSMENT IN HEALTH ECONOMIC EVALUATIONS: A COMPARATIVE ANALYSIS USING THE DRUMMOND'S CHECKLIST

Ohra S¹, Rajeeswari C¹, Siddiqui MT¹, Gupta J²,
Siddiqui MK²

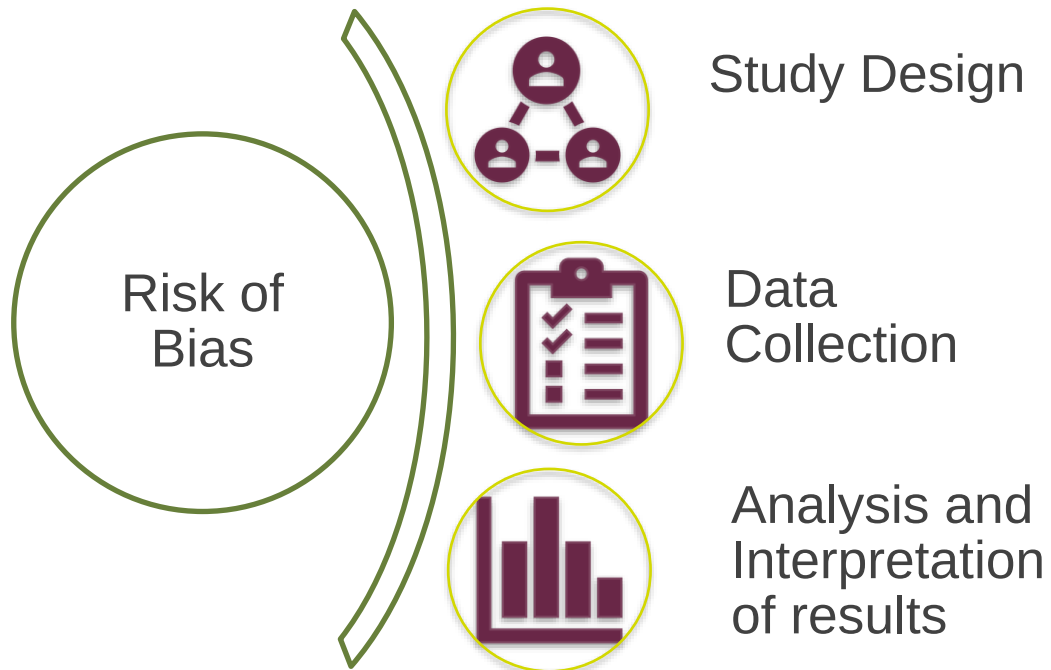
Affiliation:

¹EBM Health, New Delhi, India

²EBM Health, Cleckheaton, West Yorkshire, UK



Drummond's Checklist is a gold standard for assessing RoB in HEE



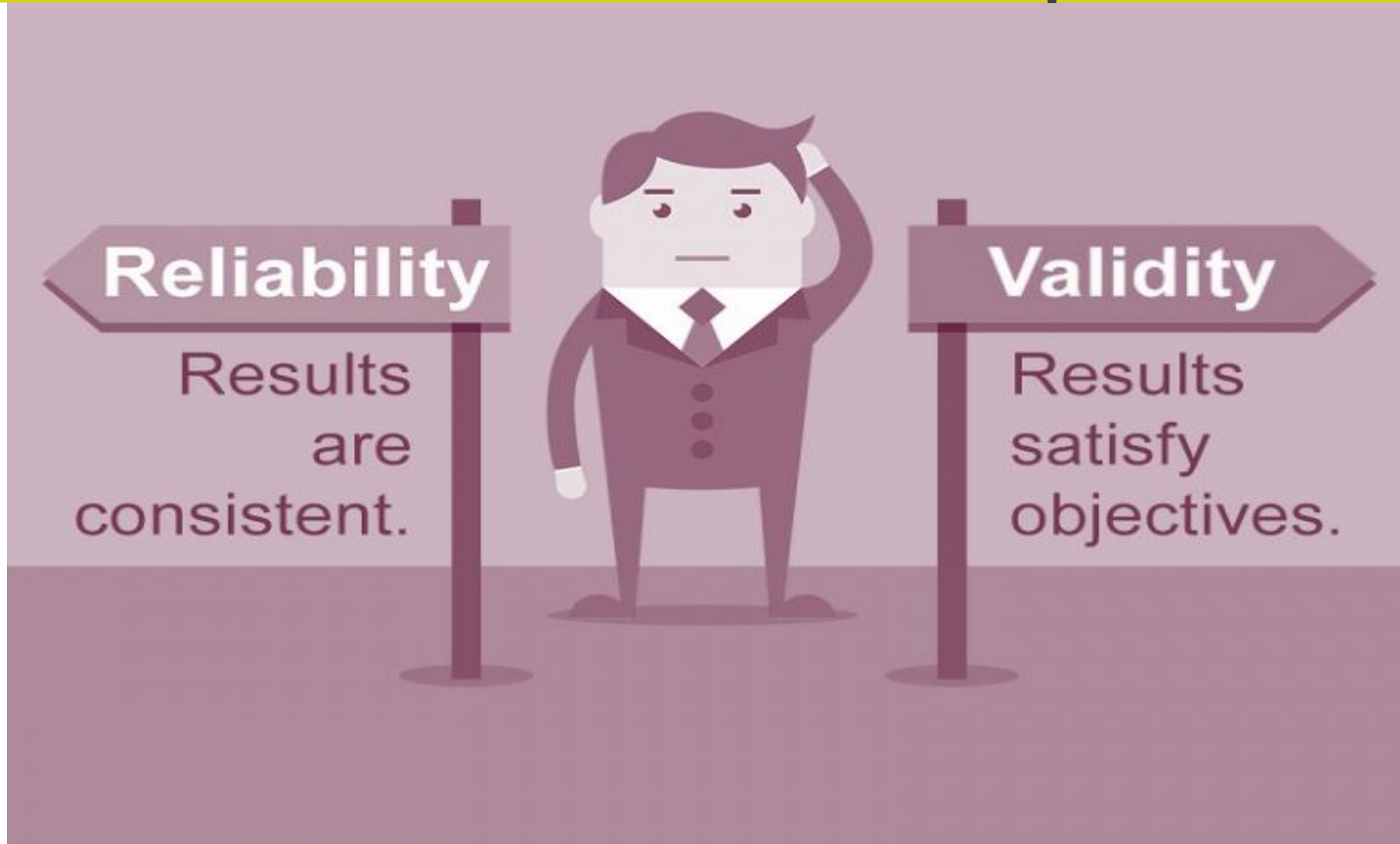
Drummond checklist contains^{1,2}:

- Research question
- Description of the study/intervention
- Effectiveness of the programme
- Identification of cost & consequences for each alternative
- Measurement of cost & consequences
- Valuation of costs and consequences
- Whether discounting was carried out
- Incremental analysis
- Presentation of results with uncertainty and sensitivity analyses
- Discussion of results in the context of policy relevance and existing literature.

(1) Drummond MF, Jefferson TO (1996) Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. British Medical Journal 313(7052):275-83.

(2) Drummond M, Sculpher M, Torrance G, et al. (2005) Methods for the economic evaluation of health care programs. Oxford: Oxford University Press; 2005.

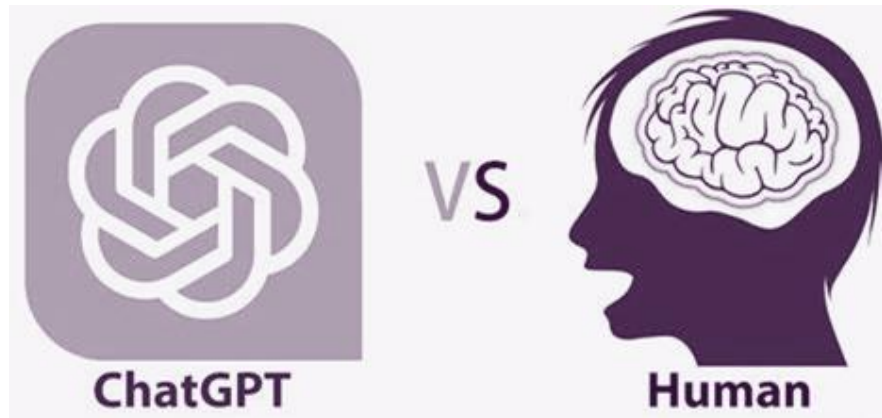
RoB Assessment of HEE studies is a resource and time-intensive process



Research Objectives

1

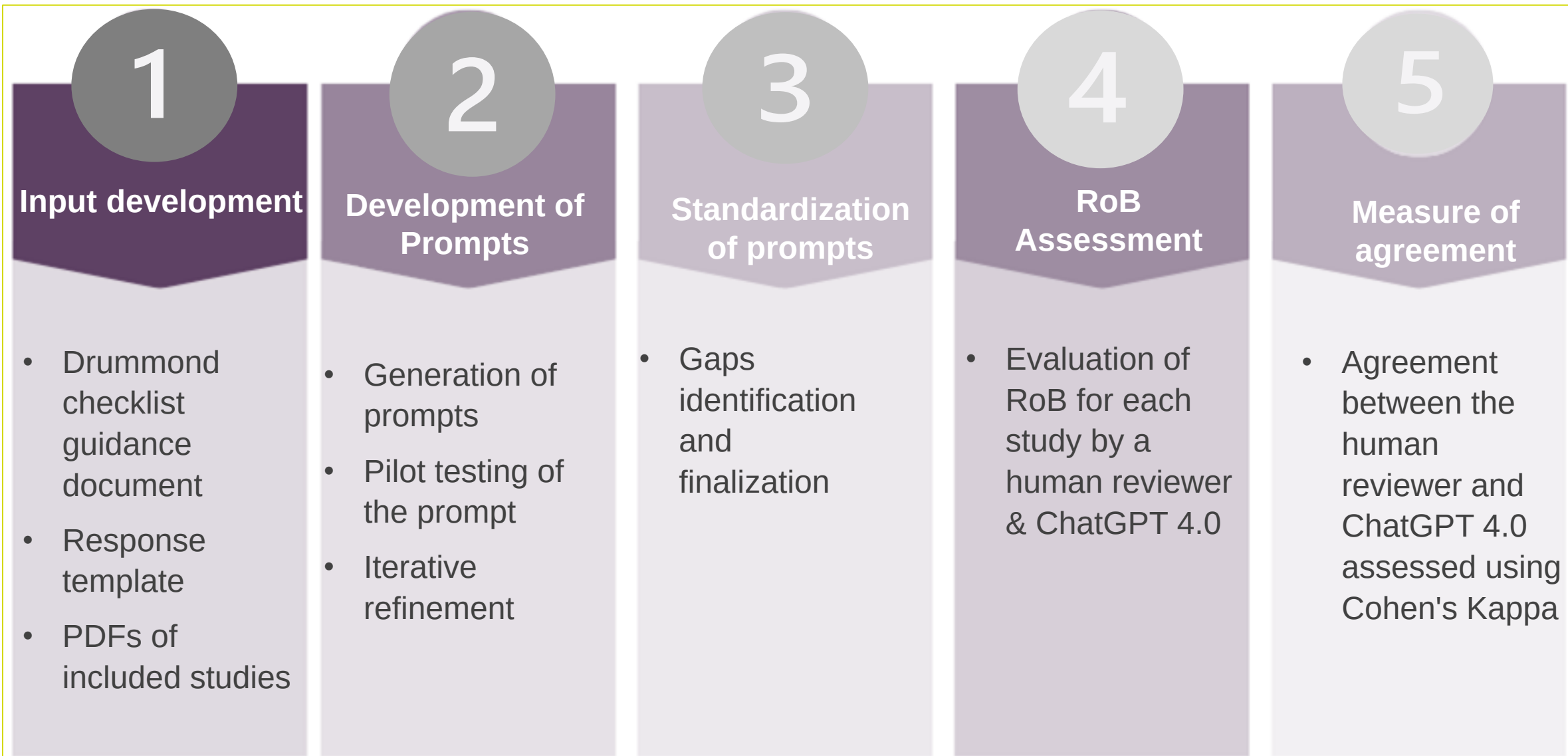
To evaluate the ChatGPTv4.0 **Efficacy and Efficiency** in the RoB Assessment using Drummond's checklist in economic evaluations



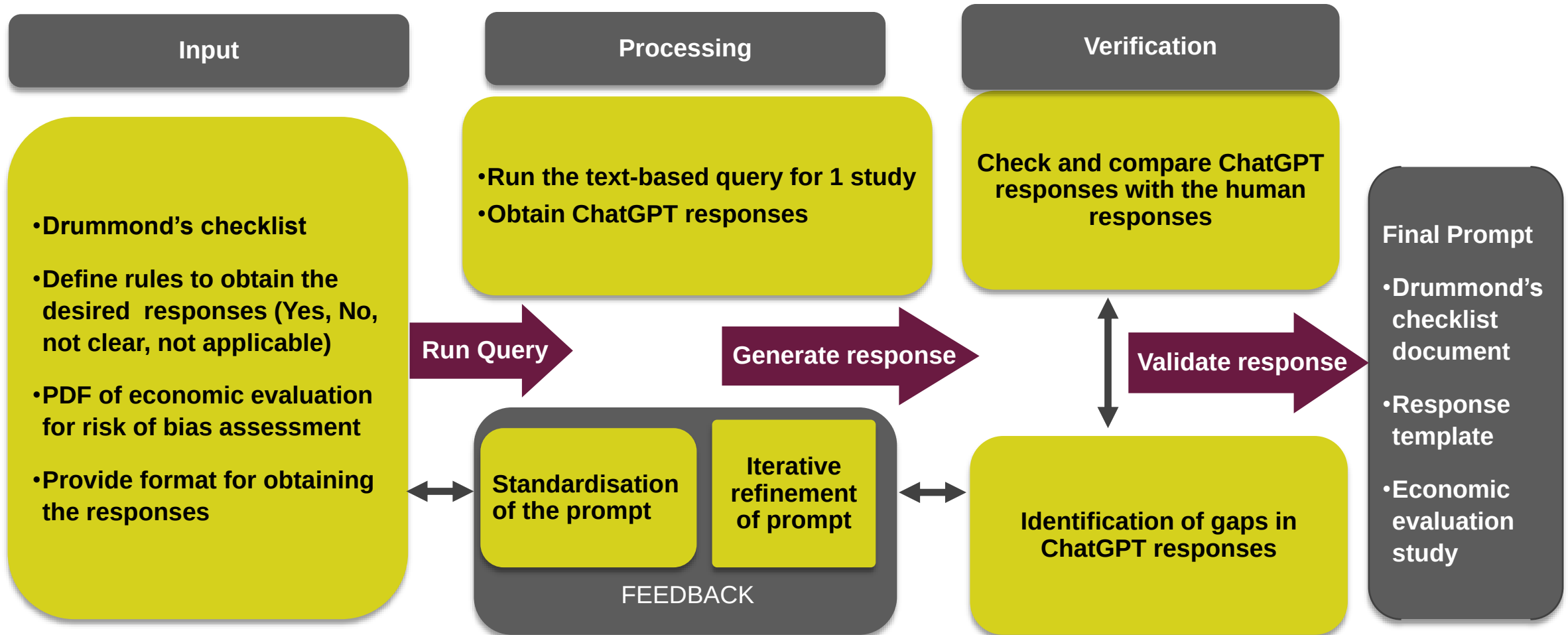
2

To assess the **level of agreement** between ChatGPTv4.0 and a trained human reviewer in assessing the RoB using Drummond's checklist in economic evaluations

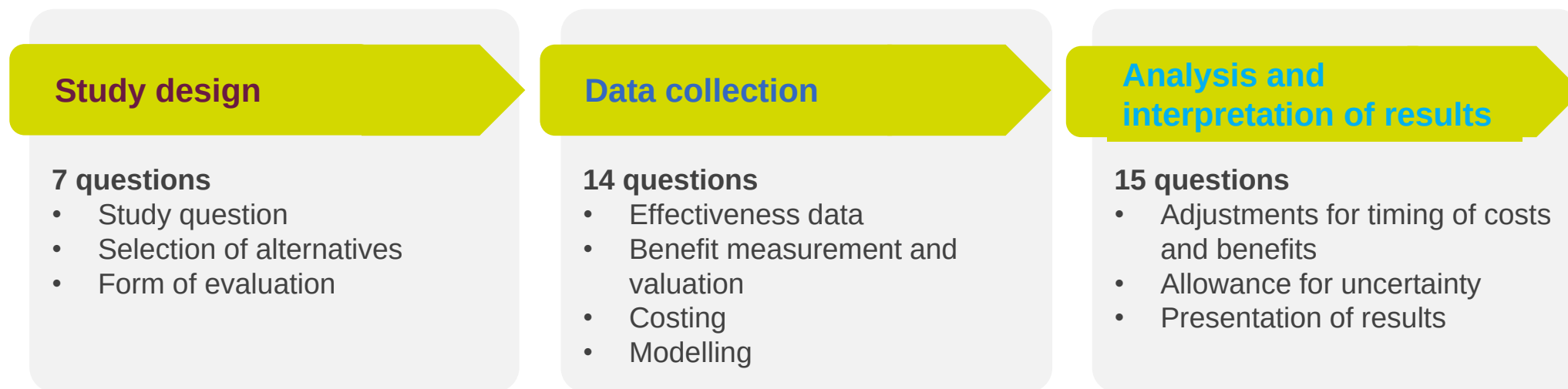
Methodology : Approach for Streamlined RoB Assessment



Prompt development & Standardization



RESULTS ARE DIVIDED ACROSS THREE DOMAINS



- RoB assessment was recorded as **Yes, No, Not clear, and not applicable or not appropriate (NA)** for each question across three domains



How was the study reported?

How was the study performed and reported, is it justified?

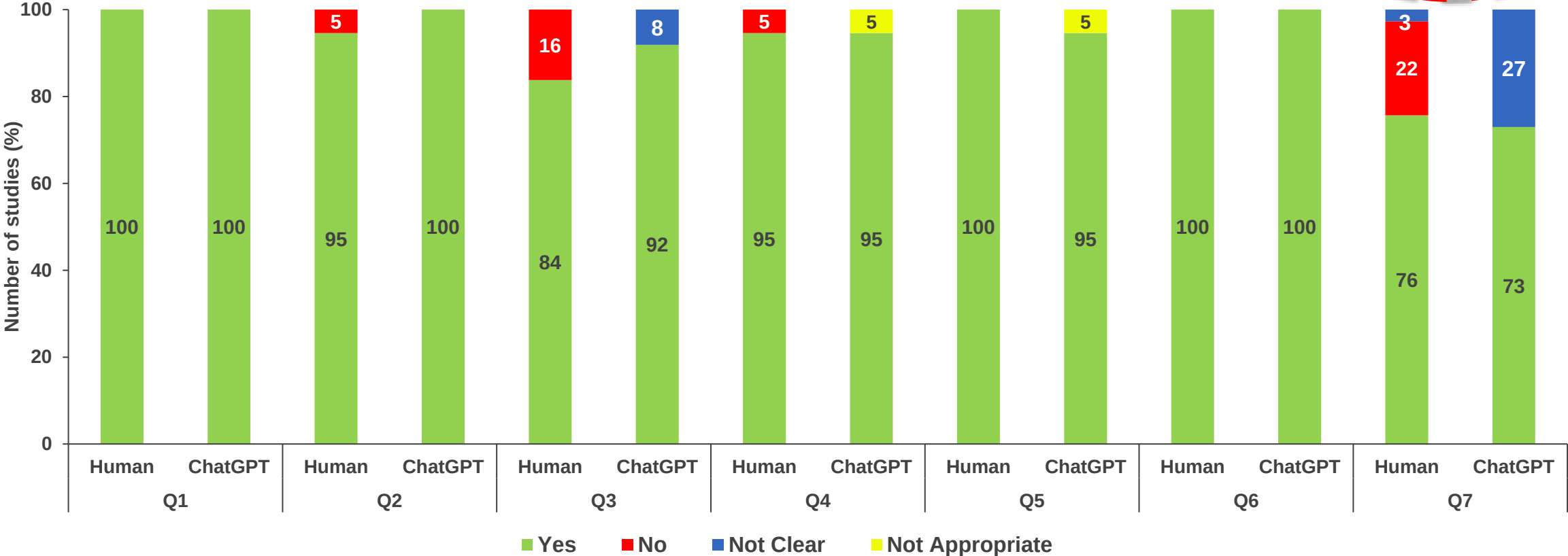
- Results are presented as bar charts providing a comparison of responses as % of studies reporting Yes/No/Not clear/NA from ChtGPT4.0 and a human reviewer
- Cohen kappa agreement statistics is reported to provide a level of agreement between ChatGPT4.0 and the human reviewer

RoB Assessment by ChatGPT vs Human: Study design

Domain: Study Design (Q1 – Q7)

Level of Agreement

Q1	Q2	Q3	Q4	Q5	Q6	Q7
100%	94.6%	89.2%	97.3%	94.6%	100%	70.3%

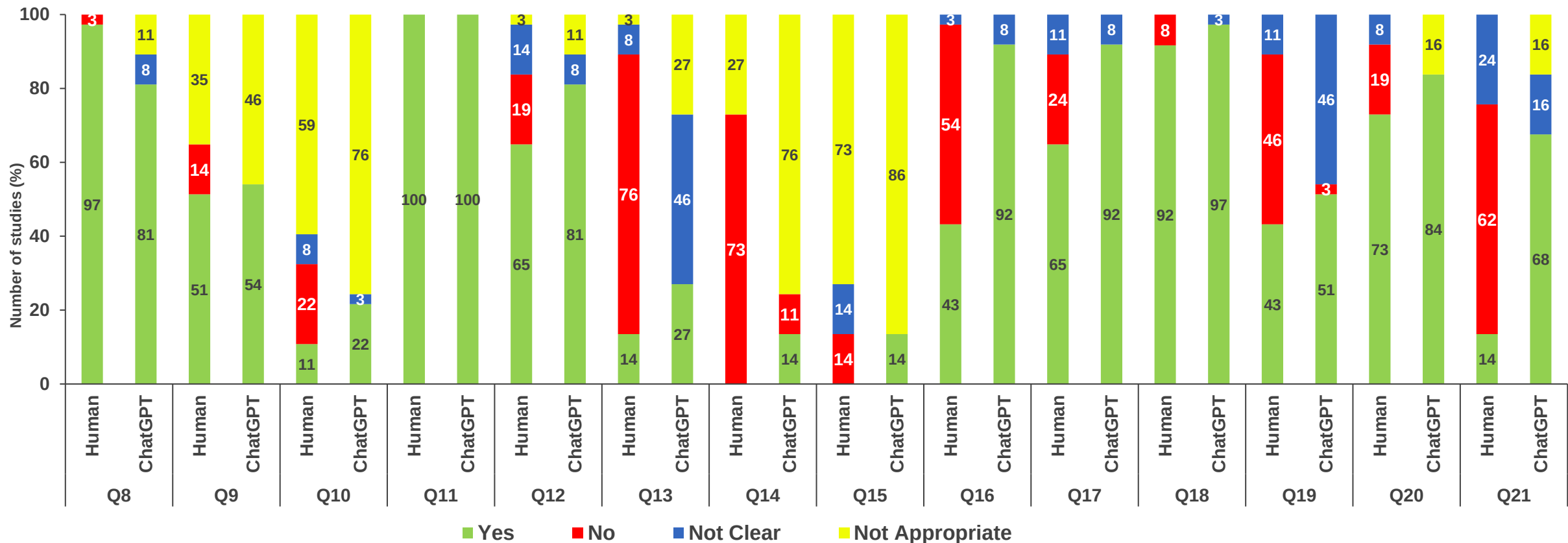


Q7: The choice of form of economic evaluation justified in relation to the questions addressed

RoB Assessment by ChatGPT vs Human: Data collection (1)

Domain: Data Collection (Q8 – Q21)

Level of agreement	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21
	82.9%	52.7%	65.8%	100%	74.8%	55.9%	44.6%	83.8%	64.9%	77.0%	93.1%	51.4%	81.1%	54.1%



Q9: Details of the design and results of effectiveness study are given (if based on a single study).
Q10: Were details of the methods of synthesis or meta-analysis of estimates given (if based on an overview of a number of effectiveness studies)?
Q13: Were the details of the subjects from whom valuations were obtained given?

Q14: Productivity changes (if included) are reported separately.
Q16: Were quantities of resources reported separately from their unit cost?
Q19: Details of currency of price adjustments for inflation or currency conversion are given.
Q21: The choice of model used and the key parameters on which it is based are justified.



RoB Assessment by ChatGPT vs Human: Data collection (2)

- Effectiveness data

9. Were details of the design and results of effectiveness study given (if based on a single study)?

52.7%

10. Were details of the methods of synthesis or meta-analysis of estimates given (if based on a synthesis of a number of effectiveness studies)?

65.8%

- Benefit measurement and valuation

13. Were the details of the subjects from whom valuations were obtained given?

55.9%

14. Were productivity changes (if included) reported separately?

44.6%

- Cost

16. Were quantities of resource use reported separately from their unit costs?

64.9%

19. Were details of price adjustments for inflation, or currency conversion, given?

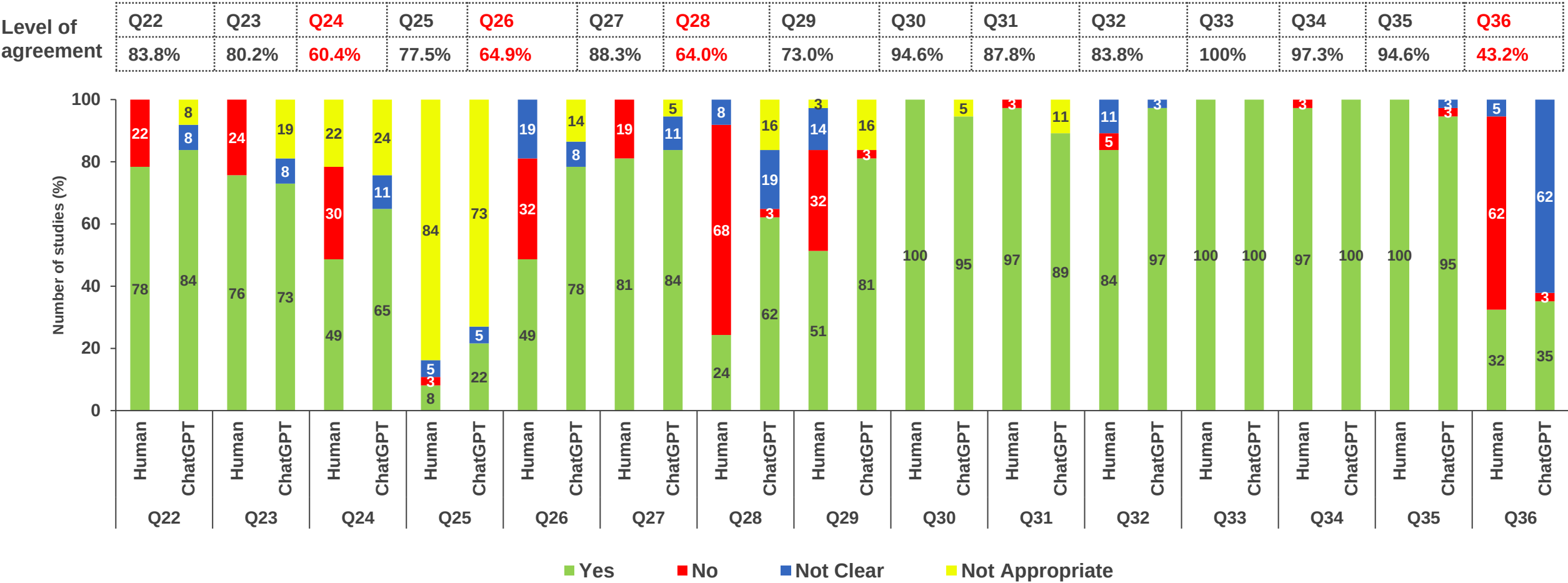
51.4%

21. Was there a justification for the choice of model used and the key parameters on which it was based?

54.1%

RoB Assessment by ChatGPT vs Human: Analysis and results (1)

Domain: Analysis and interpretation of results (Q22 – Q36)



Q24: The choice of discount rate(s) are justified
Q26: Details of statistical tests and confidence intervals are given for stochastic data.

Q28: The choice of variables for sensitivity analysis is justified
Q36: Generalisability issues addressed



RoB Assessment by ChatGPT vs Human: Analysis and results (2)

- Adjustments for timing of costs and benefits

24. Was the choice of discount rate(s) justified?

60.4%

- Allowance for uncertainty

26. Were the details of statistical test(s) and confidence intervals given for stochastic data?

64.9%

28. Was the choice of variables for sensitivity analysis justified?

64.0%

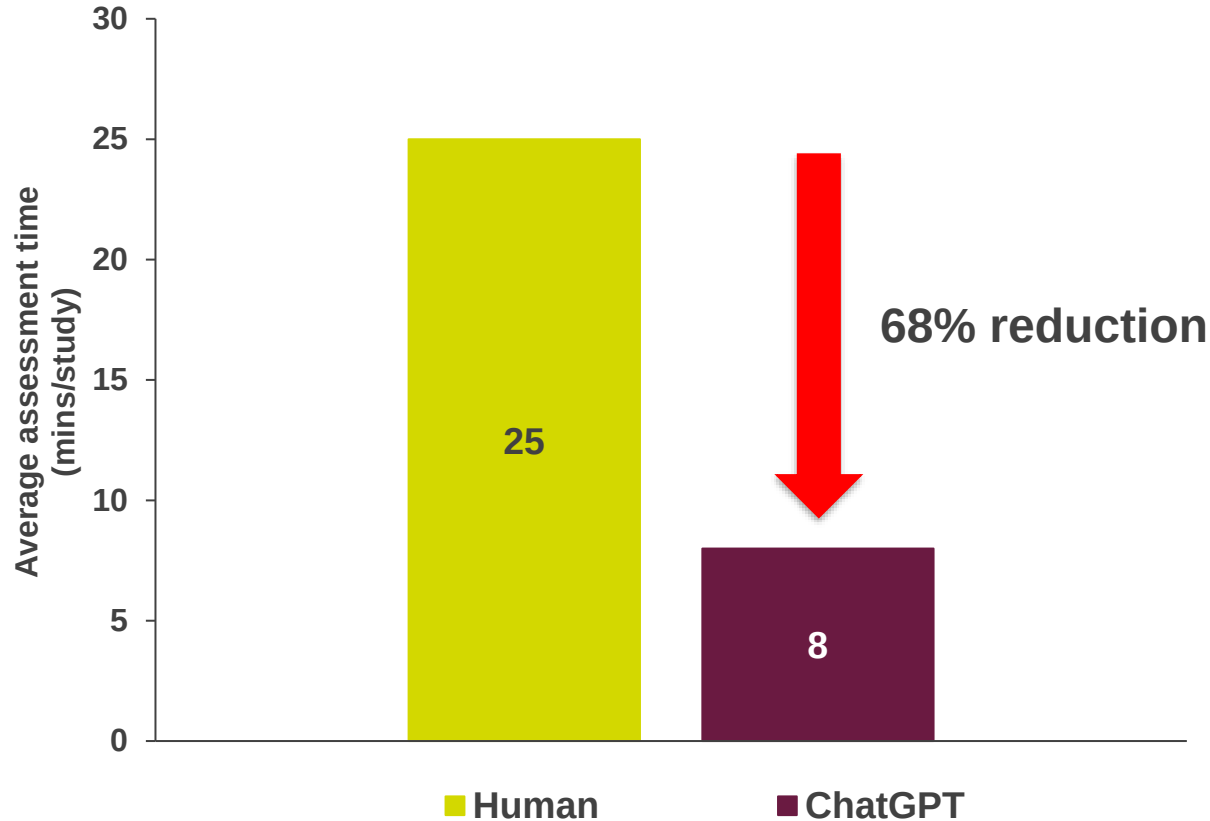
- Presentation of results

36. Were generalisability issues addressed?

43.2%

RoB Assessment by ChatGPT vs Human: Efficiency

Assessment time per study



Recall effect

- Recall effect of ChatGPT was also observed in generating RoB responses
- No temporal effect was identified in the responses generated after 2 months of the index date

Overall Summary

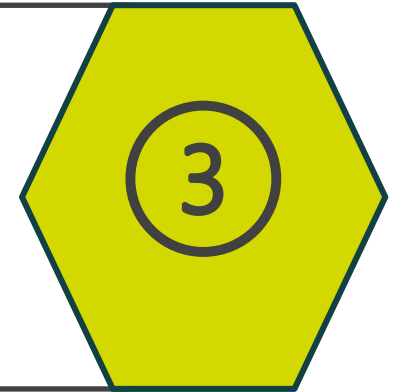
- ChatGPT performed well for '**Study design**' and '**Analysis and interpretation of results**' domain, while a moderate level of agreement was observed between ChatGPT and the Human reviewer for '**Data collection**' domain
- ChatGPT4.0 performed poorly in subjective questions when compared to a trained human reviewer thus a careful review by human experts remains indispensable
- ChatGPT4.0 showed a good recall effect and was able to reproduce the answer after 2 months time gap

Limitations and Future Directions

The **limited sample size** represents a significant constraint, potentially **impacting the robustness** of the findings.



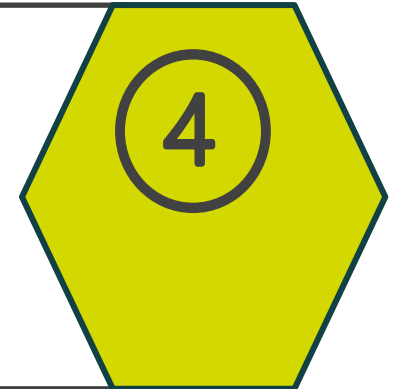
Exploring ChatGPT's use across diverse disease areas could extend its utility beyond this study, enhancing risk-of-bias assessment **in various research domains.**



The calculation of **Cohen's Kappa statistic** is **robust** to variations in data distribution, displaying a limited sensitivity to such fluctuations



To enhance the reliability & robustness of insights derived from the study, **conduct analyses with larger sample sizes** that align with anticipated kappa values.



Collaboration between AI and human reviewers offers a promising avenue for future research

THANK YOU

References

1. Edmunds K., Ling R., Shakeshaft, A. *et al.* (2018) Systematic review of economic evaluations of interventions for high-risk young people. BMC Health Services Research. 18, 660.
<https://doi.org/10.1186/s12913-018-3450-x>
2. Drummond MF., and Jefferson TO., (1996) Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. British Medical Journal. 313(7052):275-83.
3. Drummond M., Sculpher M., Torrance G., *et al.* (2005) Methods for the economic evaluation of health care programs. Oxford: Oxford University Press.
4. McHugh ML., (2012). Interrater reliability: the kappa statistic. Biochemia medica. 22(3): 276–282

Appendix

Drummond's checklist

36 Questions

Domain	Study question
Study Design	1. Was the research question stated?
	2. Was the economic importance of the research question stated?
	3. Was/were the viewpoint(s) of the analysis clearly stated and justified?
	4. Was a rationale reported for the choice of the alternative programmes or interventions compared?
	5. Were the alternatives being compared clearly described?
	6. Was the form of economic evaluation stated?
	7. Was the choice of form of economic evaluation justified in relation to the questions addressed?
Data collection	8. Was/were the source(s) of effectiveness estimates used stated?
	9. Were details of the design and results of the effectiveness study given (if based on a single study)?
	10. Were details of the methods of synthesis or meta-analysis of estimates given (if based on an overview of a number of effectiveness studies)?
	11. Were the primary outcome measure(s) for the economic evaluation clearly stated?
	12. Were the methods used to value health states and other benefits stated?
	13. Were the details of the subjects from whom valuations were obtained given?
	14. Were productivity changes (if included) reported separately?
	15. Was the relevance of productivity changes to the study question discussed?
	16. Were quantities of resources reported separately from their unit cost?
	17. Were the methods for the estimation of quantities and unit costs described?
	18. Were currency and price data recorded?
	19. Were details of price adjustments for inflation or currency conversion given?
	20. Were details of any model used given?
	21. Was there a justification for the choice of model used and the key parameters on which it was based?
Analysis and interpretation of results	22. Was the time horizon of cost and benefits stated?
	23. Was the discount rate stated?
	24. Was the choice of rate justified?
	25. Was an explanation given if cost or benefits were not discounted?
	26. Were the details of statistical test(s) and confidence intervals given for stochastic data?
	27. Was the approach to sensitivity analysis described?
	28. Was the choice of variables for sensitivity analysis justified?
	29. Were the ranges over which the parameters were varied stated?
	30. Were relevant alternatives compared? (That is, were appropriate comparisons made when conducting the incremental analysis?)
	31. Was an incremental analysis reported?
	32. Were major outcomes presented in a disaggregated as well as aggregated form?
	33. Was the answer to the study question given?
	34. Did conclusions follow from the data reported?
	35. Were conclusions accompanied by the appropriate caveats?
	36. Were generalizability issues addressed?