Large Language Models in Randomized Controlled Trials Design

Liyuan Jin, Jasmine Chiat Ling Ong, Kabilan Elangovan, Yuhe Ke, Alexandra Pyle, Daniel Shu Wei Ting, Nan Liu

Submitted to: Journal of Medical Internet Research
on: October 12, 2024

Table of Contents

Large Language Models in Randomized Controlled Trials Design

Liyuan Jin[1*] MD; Jasmine Chiat Ling Ong[2*] PharmD; Kabilan Elangovan[3] BE; Yuhe Ke[2] MBBS; Alexandra Pyle[2] PhD; Daniel Shu Wei Ting[4] PhD; Nan Liu[5] PhD

[1]Duke-NUS Medical School Singapor SG
[2]Singapore General Hospital Singapore SG
[3]SingHealth Singapore SG
[4]Singapore National Eye Centre Singapore SG
[5]Duke-NUS Medical School Singapore SG
[*]these authors contributed equally

Corresponding Author:
Nan Liu PhD
Duke-NUS Medical School
8 College Road
Singapore
SG

Abstract

Background: Randomized controlled trials (RCTs) face challenges such as limited generalizability, insufficient recruitment diversity, and high failure rates, often due to restrictive eligibility criteria and inefficient patient selection. Large language models (LLMs) have shown promise in various clinical tasks, but their potential role in RCT design remains underexplored.

Objective: This study investigates the ability of LLMs, specifically GPT-4-Turbo-Preview, to assist in designing RCTs that enhance generalizability, recruitment diversity, and reduce failure rates, while maintaining clinical safety and ethical standards.

Methods: We conducted a non-interventional, observational study analyzing 20 parallel-arm RCTs, comprising 10 completed and 10 ongoing studies published after January 2024 to mitigate pretraining biases. The LLM was tasked with generating RCT designs based on input criteria, including eligibility, recruitment strategies, interventions, and outcomes. The accuracy of LLM-generated designs was quantitatively assessed by comparing them to clinically validated ground truth data from ClinicalTrials.gov. Qualitative assessments were performed using Likert scale ratings (1–3) for domains such as safety, accuracy, objectivity, pragmatism, inclusivity, and diversity.

Results: The LLM achieved an overall accuracy of 72% in replicating RCT designs. Recruitment and intervention designs demonstrated high agreement with the ground truth, achieving 88% and 93% accuracy, respectively. However, LLMs showed lower accuracy in designing eligibility criteria (55%) and outcomes measurement (53%). Qualitative evaluations showed that LLM-generated designs scored above 2 points across all domains, indicating strong clinical alignment. In particular, LLMs enhanced diversity and pragmatism, which are key factors in improving RCT generalizability and addressing failure rates.

Conclusions: LLMs, such as GPT-4-Turbo-Preview, have demonstrated potential in improving RCT design, particularly in recruitment and intervention planning, while enhancing generalizability and addressing diversity. However, expert oversight and regulatory measures are essential to ensure patient safety and ethical standards. The findings support further integration of LLMs into clinical trial design, although continued refinement is necessary to address limitations in eligibility and outcomes measurement.

(JMIR Preprints 12/10/2024:67469)
DOI: https://doi.org/10.2196/preprints.67469

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Large Language Models in Randomized Controlled Trials Design

Liyuan Jin[*], Jasmine Chiat Ling Ong[*], Kabilan Elangovan, Yuhe Ke, Alexandra Pyle, Daniel Shu Wei Ting, Nan Liu[+]

Liyuan Jin[*] MD, Duke-NUS Medical School, Singapore

Jasmine Chiat Ling Ong[*] PharmD, Division of Pharmacy, Singapore General Hospital, Singapore

Kabilan Elangovan BE, AI Office, SingHealth, Singapore

Yuhe Ke MBBS, Singapore General Hospital, Singapore

Alexandra Pyle PhD, Department of Future Health System, Singapore General Hospital, Singapore

Daniel Shu Wei Ting PhD, Singapore National Eye Centre, Singapore

Nan Liu[+] PhD, Duke-NUS Medical School, Singapore (Email: liu.nan@duke-nus.edu.sg, Contact Number: + 65 66016503)

*Contributed Equally

+Corresponding Author

Large Language Models in Randomized Controlled Trials Design

Abstract

Background:
Randomized controlled trials (RCTs) face challenges such as limited generalizability, insufficient recruitment diversity, and high failure rates, often due to restrictive eligibility criteria and inefficient patient selection. Large language models (LLMs) have shown promise in various clinical tasks, but their potential role in RCT design remains underexplored.

Objective:
This study investigates the ability of LLMs, specifically GPT-4-Turbo-Preview, to assist in designing RCTs that enhance generalizability, recruitment diversity, and reduce failure rates, while maintaining clinical safety and ethical standards.

Methods:
We conducted a non-interventional, observational study analyzing 20 parallel-arm RCTs, comprising 10 completed and 10 ongoing studies published after January 2024 to mitigate pretraining biases. The LLM was tasked with generating RCT designs based on input criteria, including eligibility, recruitment strategies, interventions, and outcomes. The accuracy of LLM-generated designs was quantitatively assessed by comparing them to clinically validated ground truth data from ClinicalTrials.gov. Qualitative assessments were performed using Likert scale ratings (1–3) for domains such as safety, accuracy, objectivity, pragmatism, inclusivity, and diversity.

Results:
The LLM achieved an overall accuracy of 72% in replicating RCT designs. Recruitment and intervention designs demonstrated high agreement with the ground truth, achieving 88% and 93% accuracy, respectively. However, LLMs showed lower accuracy in designing eligibility criteria (55%) and outcomes measurement (53%). Qualitative evaluations showed that LLM-generated designs scored above 2 points across all domains, indicating strong clinical alignment. In particular, LLMs enhanced diversity and pragmatism, which are key factors in improving RCT generalizability and addressing failure rates.

Conclusions:
LLMs, such as GPT-4-Turbo-Preview, have demonstrated potential in improving RCT design, particularly in recruitment and intervention planning, while enhancing generalizability and addressing diversity. However, expert oversight and regulatory measures are essential to ensure patient safety and ethical standards. The findings support further integration of LLMs into clinical trial design, although continued refinement is necessary to address limitations in eligibility and outcomes measurement.

Keywords:
Large language models; randomized controlled trials; clinical trial design; recruitment diversity; eligibility criteria; clinical research ethics; trial failure reduction.

Introduction

Randomized controlled trials (RCT), serve as the backbone of modern evidence-based clinical practice. RCT provides a carefully controlled environment to investigate cause-effect relationships between intervention and outcomes. Landmark RCTs often inform clinical practice. However, trial designs face criticisms of poor generalizability from fixed eligibility criteria[1], lack of diversification in recruitment[2], and practical implementation concerns[1]. Patients with complex co-morbidities or late-stage disease excluded from phase III trials fail to benefit from breakthrough discoveries in real-world practice. Thus, challenges need to be addressed to maximize the yield of each study. High failure rate of clinical trials is a key stumbling block in drug development pipelines. RCTs failure rate has been reported for various reasons[3-5], including safety and toxicity concerns, poor accrual and recruitment challenges, logistics, and funding. Of which, a key contributory factor to failure of phase III trials is an inefficient patient selection process[6]. Failure of clinical trials bears significant implications for both drug development companies and patients. Clinical research remains the most expensive and time-consuming process of drug development, costing up to a billion dollars of investment and taking more than a decade of work to bring a new drug into market[7]. Reform of clinical research is much needed to accelerate this process.

Large language models (LLMs) have recently emerged as an efficient tool in various clinical tasks[8] with comparable clinical alignment to human experts[9]. As a result, LLMs tools are expected to assist clinical practice ranging from basic healthcare related administrative work [10], educational chatbot for medical knowledge[11,12], to advance clinical notes generation[13-15], complex clinical cases diagnosis[16], and patient triaging[17]. Recently, there is increasing interest in LLM applications in clinical trials[18-21]. Generative AI introduced new paradigms in drug development, from the design and validation of novel pharmaceutical compounds to eligibility screening of patients for clinical trials[18-20]. These approaches show promise in streamlining clinical research but fail to address problems related to trial design and generalizability of RCTs including eligibility criteria, diversification and practicability. RCTs provide the highest level of scientific evidence of therapeutic interventions and, their design requires in-depth clinical understanding and rigorous scientific methodologies[22-24]. In this study, we explore and validate the use of LLMs as a pilot application for efficient and clinically aligned RCT design, to help improve study generalizability and reduce failure rate.

Methods

We performed an observational, non-interventional study using GPT-4-Turbo-Preview as state-of-the-art LLM.

Validation and Testing Datasets

We randomly selected 20 parallel-arm RCTs (Phase III or IV): 10 completed RCTs, with results published in leading clinical journals (JAMA, Nature Medicine, NEJM, and The Lancet); and 10 ongoing RCTs registered on ClinicalTrials.gov. To mitigate the risks of LLM's pretraining utilization on such studies, we used studies published or newly registered after January 2024 (after GPT-4-Turbo-Preview pretraining date of December 2023).
Details of the dataset are presented in **eTable 1 (Supplementary text).**

Reference standard and LLM Prompt

We extracted the respective study designs from ClinicalTrials.gov (information cross-checked against publication if available), to serve as our ground truth. We provided the LLM with the following inputs: Official Titles, Brief Summaries, Study Type, Study Phase, Study Design, Conditions and Intervention/Treatment. We then prompted the LLM for the following outputs:

Eligibility Criteria (Inclusion and Exclusion Criteria), Recruitment (Sex/Gender and Age), Arm/Intervention (Active and Control Arms), and Outcomes Measurement (Measurement design and Measurement time frame).

Large Language Model

In this current study, we selected GPT-4-Turbo-Preview. We chose a Temperature of 0.2 to balance replicability and clinical rigor. Detailed prompts and example output are presented in **eFigure 1** and **eFigure 2 (supplementary text**), respectively.

Quantitative Evaluation

We quantitatively evaluated the accuracy (degree of agreement) of the LLM's outputs by comparing them with the clinically defined ground truth. We first collect ground truth for published studies from publication (cross-examined with corresponding study from ClinicalTrials.gov), and recent registered ongoing trials from ClinicalTrials.gov. For outputs with numerical or categorical answers, such as gender or age in recruitment and measurement time frame in outcome measures, we define correct answers as completely matching numerical values in ground truth. For outputs with clinical answers, such as eligibility criteria, active and control arm in intervention and measurement design in outcome measures, we defined correct answers if clinically align with ground truth. Specifically, for eligibility criteria designs, the accuracy of was determined by numbers of matched LLM designs divided by total number of eligibility criteria LLM has listed.

We created a qualitative assessment metric to evaluate both LLM and ground truth designs. This metric comprised of safety, clinical accuracy, objectivity (bias), pragmatic (adapted from PRECIS-2 guidance)[25], inclusivity and diversity (adapted from United States Food and Drug Administration (FDA) draft guidance to clinical trial design)[2] measured on a three-point Likert Scale (1 is the worst, 3 is the best). For selected ongoing RCT studies, we performed a blinded qualitative evaluation without knowledge of ground truth designs to provide a more objective analysis.

Statistical analysis

We employed average, non-weighted NLP based objective scoring, including BLEU, ROGUE-L and METEOR, for LLM outputs. Details are represented in **eTable 2 (Supplementary text)**.

Ethics statement and informed consent

As current study is retrospective in nature, and no real patient was involved in current research, regulatory approval and informed consent is not applicable. Human clinical experts received no compensation for rating.

Results

Our results showing LLM demonstrated 72% accuracy in overall RCT designs (**Figure 1**). Specifically, it showed high agreement in Recruitment and Arm/Intervention, with accuracy of 88% and 93%, respectively. However, it demonstrated discrepancies in designing Eligibility Criteria and Outcomes Measurement, with accuracy of 55% and 53%, respectively. We observed marginal difference in accuracy between LLM outputs and published RCTs and ongoing RCTs except improvement in exclusion criteria designs on latest RCTs. We employed statistical analysis using natural language processing (NLP) based methods, including BLEU[26], ROGUE-L[27] and METEOR[28], for corresponding LLM outputs, presented in **Supplementary Method eTable 2**. Qualitatively, LLM designs produced comparable clinical alignment in RCT design compared to ground truth, with Likert scales scoring above 2 points across all domains (**Figure 2**).

Our findings suggest that LLM, represented by GPT-4-Turbo-Preview in this study, can replicate RCT designs with reasonable clinical alignment. LLM was able to match RCTs with over 80% accuracy in designing Recruitment requirements and Active/Control Intervention. When assessed qualitatively, we observed marginal difference in overall clinical accuracy of LLM design compared with ground truth, highlighting multiple accepted clinical decisions related to RCT design. Upon qualitative analysis, LLM RCT designs closely aligned documented consensus in safe, accurate, and objective domains, while showing enhanced diversity and pragmatism. Notably, diversity and pragmatism are key determinants of LLM generalizability and reasons for RCT failure. Additionally, LLM could avoid critical safety and ethical issues identified in the ground truth from the analysis of the selected registered ongoing RCTs.

Discussion

Principal Results

RCTs serve key roles in clinical practice, and inclusivity has been heavily emphasized by FDA[29] to ensure consistently high-quality design that is scientifically justifiable. Current results highlight the potential role in LLM for such an important design principle. Unique attributes of LLM architecture bring distinct advantages over conventional deep learning and NLP in text-based comprehension capabilities. General-purpose LLMs like GPT-4 can perform tasks with little or no task-specific fine-tuning. Emergent properties set them apart from conventional machine learning or deep learning models, simulating clinical reasoning and inferential skills across diverse disciplines[30,31]. The large knowledge corpus in pre-training dataset of LLMs enabled stochastic responses to tasks that are non-deterministic in nature, such as in clinical trial design. We infer that LLM was capable of recommending most commonly used comparator arms for trials of similar nature and discipline; logical deduction of active intervention dosage regimen based on pre-clinical or phase I/II published studies captured in its knowledge corpus. Recommended exclusion criteria and outcome measurement time frames differed to a greater extent between LLM-designed trials and actual published design. These design elements often vary widely across different studies and intervention tested in real-world. Qualitatively, the overall safety and clinical accuracy of these reported differences was not compromised significantly. Coupled with further tailored RCT designs through prompting with LLMs regarding various patient and condition-related concerns, as well as financial and pragmatic challenges, the current pilot LLM-based RCT framework is expected to improve generalizability, enhance patient recruitment, and reduce RCT failure rates.

Limitations

Our study suffers the following limitations. First, the generalizability of our findings is constrained by the specific LLM architecture used, GPT-4-Turbo-Preview, which may not reflect the performance of other LLMs or future versions. Our analysis was limited to text-based outputs, which do not capture the full complexity of clinical trial design, such as availability of funding, ease of patient recruitment and ethical considerations. The study also relied on a relatively small sample of RCT designs, which may not provide a comprehensive view of the LLM's capabilities across diverse medical specialties. Finally, alternative trial designs such as open-label, cross-over or pragmatic trials were not considered in this study.

Comparison with Prior Work

Existing clinical trial related LLM studies, presented in **Table 1**, have only focused on preliminary text classification task and are mostly limited to last generation LLM, such as BERT[32]. With rapid advancement in LLM development and taking advantage of LLM's accessibility and efficiency as demonstrated in current study, it holds great promise as an assistive tool for RCT design. In our quantitative analysis, LLMs could recommend study designs using gold standard control groups and appropriate active group intervention.

Conclusions

This study highlights the potential of LLMs to enhance RCT design, achieving substantial accuracy with key improvements in diversity and pragmatism. Such advancements could significantly improve the efficiency and effectiveness of clinical trials, driving faster development of therapeutic interventions. While LLM show huge promise, expert oversight remains crucial for ensuring safety and ethics. Future efforts should aim to better integrate LLMs within clinical research frameworks and develop adaptive regulatory measures

Acknowledgements

No funding was required in this manuscript.

Conflicts of Interest

All authors declare no relevant conflicts of interest.

References

1. Nichol A, Bailey M, Cooper D, behalf of the POLAR O. Challenging issues in randomised controlled trials. *Injury*. 2010;41:S20-S23.

2. Gray DM, Nolan TS, Gregory J, Joseph JJ. Diversity in clinical trials: an opportunity and imperative for community engagement. *The Lancet Gastroenterology & Hepatology*. 2021;6(8):605-607.

3. Stensland KD, DePorto K, Ryan J, et al. Estimating the rate and reasons of clinical trial failure in urologic oncology. Elsevier; 2021:154-160.

4. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics*. 2019;20(2):273-286.

5. Pretorius S, Grignolo A. Phase III trial failures: Costly, but preventable. 2016;

6. Artificial Intelligence for Clinical Trial Design: Trends in Pharmacological Sciences. 2024;doi:doi:10.1016/j.tips.2019.05.005

7. Hutson M. How AI is being used to accelerate clinical trials. Nature Index. *Nature*. 2024-03-13 2024;627(8003)doi:doi:10.1038/d41586-024-00753-x

8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nature medicine*. 2023;29(8):1930-1940.

9. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180.

10. Karakas C, Brock D, Lakhotia A. Leveraging ChatGPT in the Pediatric Neurology Clinic: Practical Considerations for Use to Improve Efficiency and Outcomes. *Pediatric Neurology*. 2023;148:157-163.

11. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: Performance of ChatGPT on a PES medical examination. *Cardiology Journal*. 2023;

12. Klang E, Portugez S, Gross R, et al. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. *BMC Medical Education*. 2023;23

13. Waisberg E, Ong J, Masalkhi M, et al. GPT-4 and ophthalmology operative notes. *Annals of Biomedical Engineering*. 2023:1-3.

14. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT-4 on impressions generation in radiology reports. *Radiology*. 2023;307(5):e231259.

15. Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. *Cureus*. 2023;15(4)

16. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA*. 2023;

17. Waisberg E, Ong J, Zaman N, et al. GPT-4 for triaging ophthalmic symptoms. *Eye*. 2023:1-2.

18. Ghim J-L, Ahn S. Transforming clinical trials: the emerging roles of large language models. *Translational and Clinical Pharmacology*. 2023;31(3):131.

19. Wong C, Zhang S, Gu Y, et al. Scaling clinical trial matching using large language models: A case study in oncology. PMLR; 2023:846-862.

20. Jin Q, Wang Z, Floudas CS, Sun J, Lu Z. Matching patients to clinical trials with large language models. *ArXiv*. 2023;

21. Tayebi Arasteh S, Han T, Lotfinia M, et al. Large language models streamline automated machine learning for clinical studies. *Nature Communications*. 2024;15(1):1603.

22. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Bmj*. 2010;340

23. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology and pharmacotherapeutics*. 2010;1(2):100-107.
24. Chan A-W, Tetzlaff JM, Gøtzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *Bmj*. 2013;346
25. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *bmj*. 2015;350
26. Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. 2002:311-318.
27. Lin C-Y. Rouge: A package for automatic evaluation of summaries. 2004:74-81.
28. Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. 2005:65-72.
29. Food, Administration D. Evaluating inclusion and exclusion criteria in clinical trials. 2020.
30. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. Jul 12 2023;doi:10.1038/s41586-023-06291-2
31. Wei J, Tay Y, Bommasani R, et al. Emergent Abilities of Large Language Models. 2022/06/15 2022;
32. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*. 2018;

Figures and Tables

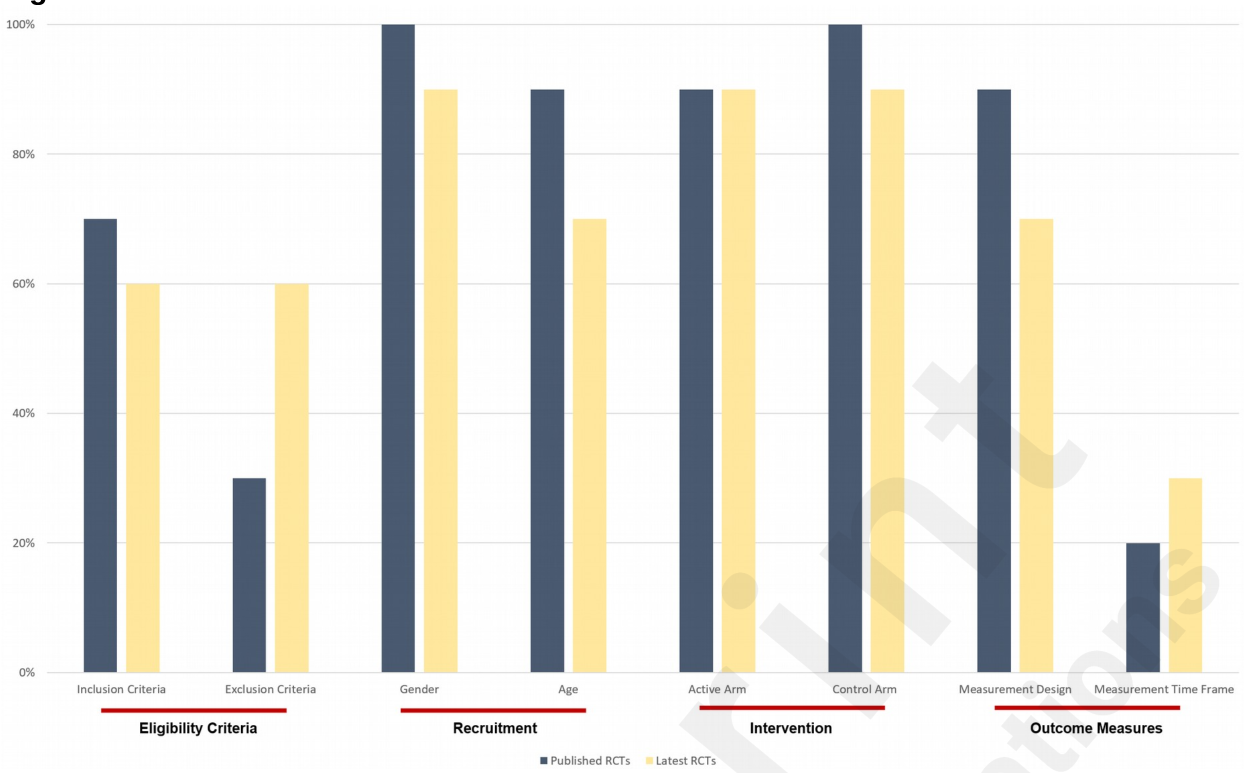


Figure 1. LLM outputs coherence (matching with ground truth) on 20 testing RCT studies (10 published RCTs and 10 ongoing registered RCTs).

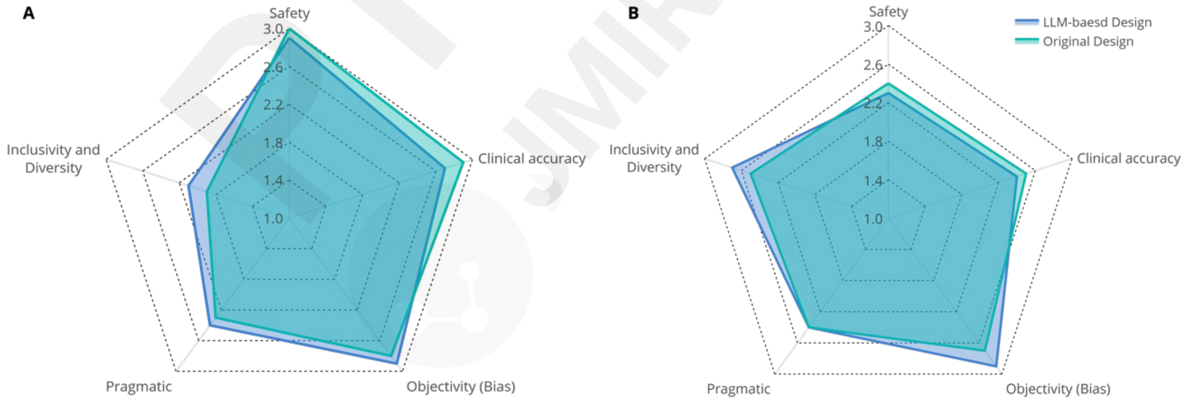


Figure 2. A: Qualitative metric for 10 published RCTs. B. Qualitative metric for 10 ongoing registered RCTs.

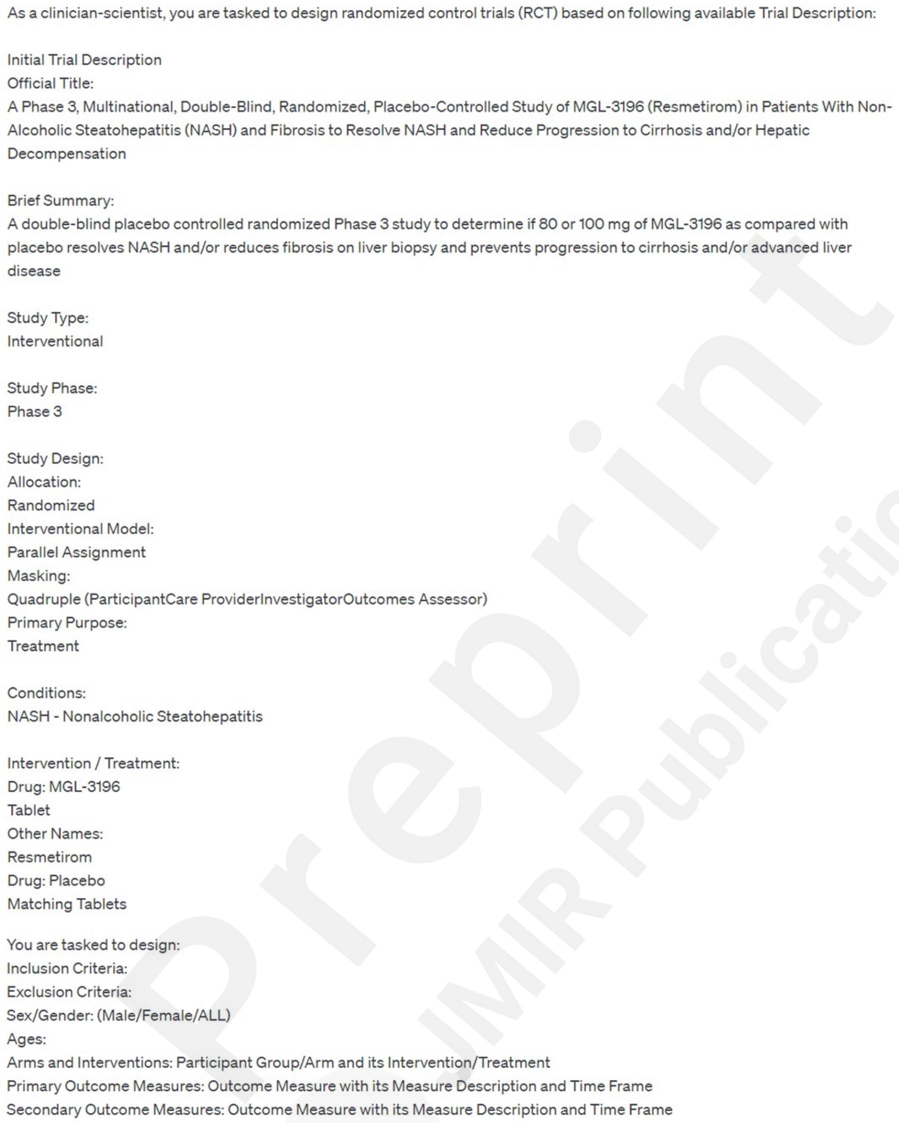
Studies	LLM Application	LLM Base Model	Training or Prompt Technique	Source of Training Dataset	Testing Dataset Sample Size	Evaluation Metrics Used	Model Performance
A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora[33]	Eligibility Screening	BERT	Pre-training	Interview data	470/ 230/ 1000	F1	0.72/ 0.84/ 0.62
AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models[34]	Eligibility Screening	GPT 4	Zero-shot	clinical trial text	180 Trials	F1	0.90
Text Classification of Cancer Clinical Trial Eligibility Criteria[35]	Eligibility Screening	BERT	Zero-shot	Registry	764 Trials	ACC	0.27-0.95
ChatGPT for Sample-Size Calculation in Sports Medicine and Exercise Sciences: A Cautionary Note[36]	Sample Size Calculation	GPT 4	Few-shots	Registry	4 Trials	ACC	0.75
Medical text classification based on the discriminative pre-training model and prompt-tuning[37]	Assist Trial Outcome Measurement	BERT	Pre-training	Interview data	5127 Outcome entities	ACC	0.86
Predicting Publication of Clinical Trials Using Structured and Unstructured Data: Model Development and	Trial Outcome Prediction	BERT	Zero-shot	Registry	76,950 Trials	F1	0.70

Validation Study[38]							

Table 1. Existing LLM applications in clinical trials related studies. We used the following search strategy We used the following literature search strategy: (("clinical trials as topic"[MeSH Terms] OR "randomized controlled trials as topic"[MeSH Terms] OR "clinical trial"[Title/Abstract]) AND ("artificial intelligence"[MeSH Terms] OR "generative ai"[Title/Abstract] OR "language model"[Title/Abstract])) AND (2022:2024[pdat]). We restricted search to articles published in PubMed between 1st January 2022 and 1st April 2024. We screened a total of 575 articles from PubMed and included a final of 6 publications. We included peer-reviewed articles investigating the performance of generative AI models applied in the conduct of clinical trials or randomized controlled trials. We excluded review papers and studies that did not report any model performance. Legend: BERT: Bidirectional Encoder Representations from Transformers, GPT: Generative Pre-trained Transformer, AUC: area under the receiver operating characteristic (ROC) curve, ACC: Accuracy

Supplementary Figure

eFigure 1: Prompt for LLM based RCT design (Temperature set as 0.2).



As a clinician-scientist, you are tasked to design randomized control trials (RCT) based on following available Trial Description:

Initial Trial Description
Official Title:
A Phase 3, Multinational, Double-Blind, Randomized, Placebo-Controlled Study of MGL-3196 (Resmetirom) in Patients With Non-Alcoholic Steatohepatitis (NASH) and Fibrosis to Resolve NASH and Reduce Progression to Cirrhosis and/or Hepatic Decompensation

Brief Summary:
A double-blind placebo controlled randomized Phase 3 study to determine if 80 or 100 mg of MGL-3196 as compared with placebo resolves NASH and/or reduces fibrosis on liver biopsy and prevents progression to cirrhosis and/or advanced liver disease

Study Type:
Interventional

Study Phase:
Phase 3

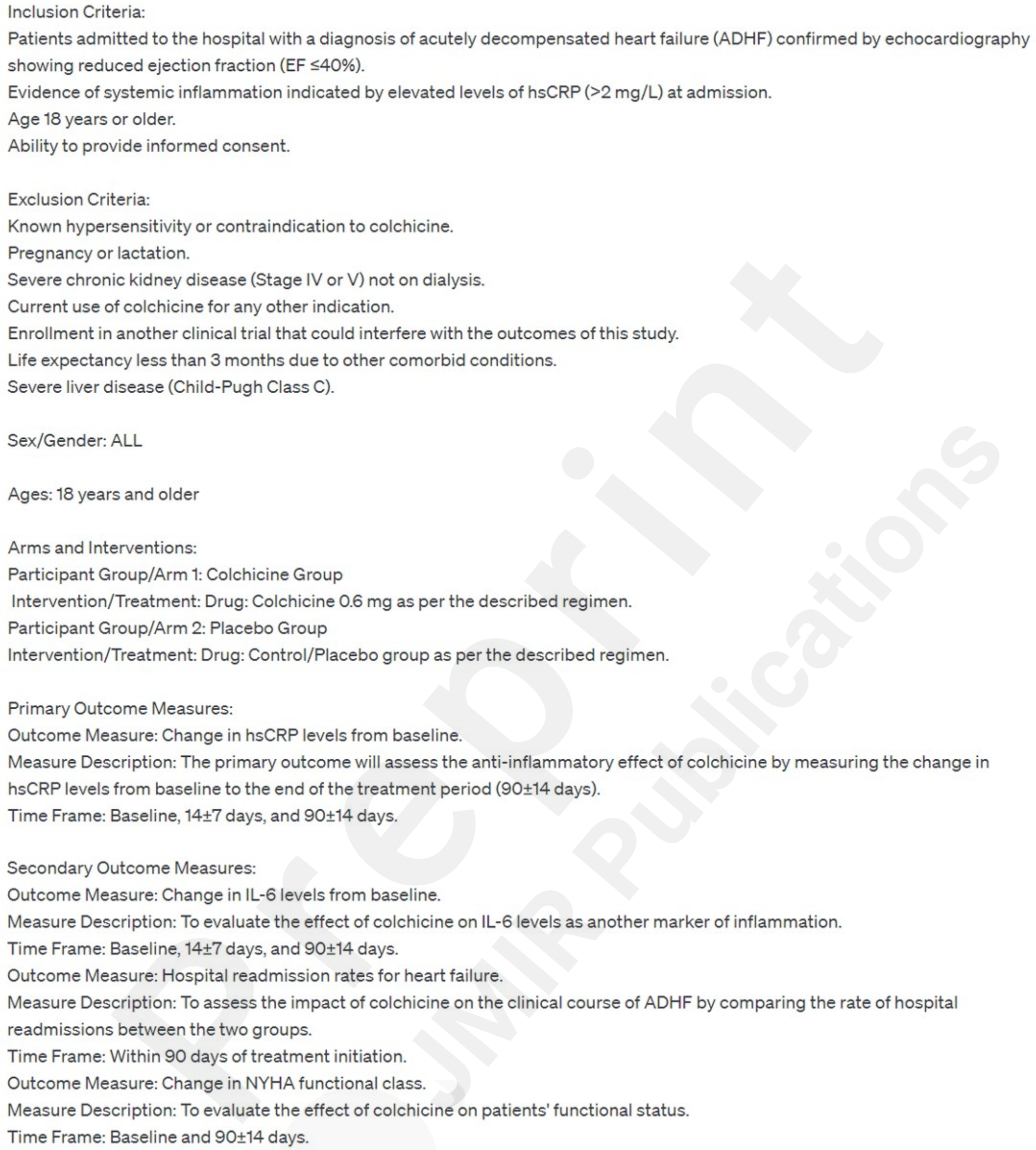
Study Design:
Allocation:
Randomized
Interventional Model:
Parallel Assignment
Masking:
Quadruple (ParticipantCare ProviderInvestigatorOutcomes Assessor)
Primary Purpose:
Treatment

Conditions:
NASH - Nonalcoholic Steatohepatitis

Intervention / Treatment:
Drug: MGL-3196
Tablet
Other Names:
Resmetirom
Drug: Placebo
Matching Tablets

You are tasked to design:
Inclusion Criteria:
Exclusion Criteria:
Sex/Gender: (Male/Female/ALL)
Ages:
Arms and Interventions: Participant Group/Arm and its Intervention/Treatment
Primary Outcome Measures: Outcome Measure with its Measure Description and Time Frame
Secondary Outcome Measures: Outcome Measure with its Measure Description and Time Frame

eFigure 2: Example output of LLM based RCT design.



Inclusion Criteria:
Patients admitted to the hospital with a diagnosis of acutely decompensated heart failure (ADHF) confirmed by echocardiography showing reduced ejection fraction (EF ≤40%).
Evidence of systemic inflammation indicated by elevated levels of hsCRP (>2 mg/L) at admission.
Age 18 years or older.
Ability to provide informed consent.

Exclusion Criteria:
Known hypersensitivity or contraindication to colchicine.
Pregnancy or lactation.
Severe chronic kidney disease (Stage IV or V) not on dialysis.
Current use of colchicine for any other indication.
Enrollment in another clinical trial that could interfere with the outcomes of this study.
Life expectancy less than 3 months due to other comorbid conditions.
Severe liver disease (Child-Pugh Class C).

Sex/Gender: ALL

Ages: 18 years and older

Arms and Interventions:
Participant Group/Arm 1: Colchicine Group
Intervention/Treatment: Drug: Colchicine 0.6 mg as per the described regimen.
Participant Group/Arm 2: Placebo Group
Intervention/Treatment: Drug: Control/Placebo group as per the described regimen.

Primary Outcome Measures:
Outcome Measure: Change in hsCRP levels from baseline.
Measure Description: The primary outcome will assess the anti-inflammatory effect of colchicine by measuring the change in hsCRP levels from baseline to the end of the treatment period (90±14 days).
Time Frame: Baseline, 14±7 days, and 90±14 days.

Secondary Outcome Measures:
Outcome Measure: Change in IL-6 levels from baseline.
Measure Description: To evaluate the effect of colchicine on IL-6 levels as another marker of inflammation.
Time Frame: Baseline, 14±7 days, and 90±14 days.
Outcome Measure: Hospital readmission rates for heart failure.
Measure Description: To assess the impact of colchicine on the clinical course of ADHF by comparing the rate of hospital readmissions between the two groups.
Time Frame: Within 90 days of treatment initiation.
Outcome Measure: Change in NYHA functional class.
Measure Description: To evaluate the effect of colchicine on patients' functional status.
Time Frame: Baseline and 90±14 days.

Supplementary Text

eTable 1: Selected published RCTs (No. 1-10) and ongoing newly registered RCTs (No. 11-20).

NO	JOURNAL	PUBLISH YEAR	RCT NAME	SPECIALITY	PHASE	REGISTRY NUMBER	FIRST POSTED DATE	TRIAL TYPE
1	NEJM	2024/2	A Phase 3 Study to Evaluate the Efficacy and Safety of MGL-3196 (Resmetirom) in Patients With NASH and Fibrosis (MAESTRO-NASH)	Gastroenterology	III	NCT03900429	NA	Placebo-control
2	NEJM	2024/1	Testosterone Treatment and Fractures in Men with Hypogonadism	Endocrine	IV	NCT03518034	NA	Placebo-control
3	NEJM	2024/1	Azithromycin during Routine Well-Infant Visits to Prevent Death	Paediatric	IV	NCT03676764	NA	Placebo-control
4	NEJM	2024/1	Efficacy and Safety of Acoramidis in Transthyretin Amyloid Cardiomyopathy	Cardiology	III	NCT03860935	NA	Placebo-control
5	JAMA	2024/1	Continued Treatment With Tirzepatide for Maintenance of Weight Reduction in Adults With Obesity The SURMOUNT-4 Randomized Clinical Trial	Endocrine	III	NCT04660643	NA	Placebo-control
6	The Lancet	2024/2	Clinical Efficacy of Typhoid Conjugate Vaccine (Vi-TCV) Among Children Age 9 Months Through 12 Years in Blantyre, Malawi	Infectious Disease	III	NCT03299426	NA	Placebo-control
7	The Lancet	2024/1	Chemoprevention for malaria with monthly intermittent preventive treatment with dihydroartemisinin–piperaquine in pregnant women living with HIV on daily co-trimoxazole in Kenya and Malawi: a randomised, double-blind, placebo-controlled trial	Infectious Disease	III	NCT04158713	NA	Placebo-control
8	The Lancet	2024/1	A Phase III Study of Safety and Efficacy of Ligelizumab in the	Dermatology	III	NCT03580369	NA	Placebo-control

			Treatment of CSU in Adolescents and Adults Inadequately Controlled With H1-antihistamines					
9	The Lancet	2024/1	Efficacy and safety of the muscarinic receptor agonist KarXT (xanomeline–trospium) in schizophrenia (EMERGENT-2) in the USA: results from a randomised, double-blind, placebo-controlled, flexible-dose phase 3 trial	Psychiatric	III	NCT04659161	NA	Placebo-control
10	Nature Medicine	2024/1	First-line talazoparib with enzalutamide in HRR-deficient metastatic castration-resistant prostate cancer: the phase 3 TALAPRO-2 trial	Oncology	III	NCT03395197	NA	Placebo-control
11	NA	NA	Colchicine in Acutely Decompensated HFREF	Cardiology	IV	NCT06286423	2/29/2024	Placebo-control
12	NA	NA	Clinical Trial of the Efficacy and Safety of Raphamin in Prevention of Recurrences of Chronic Bacterial Cystitis	Infectious Disease	III	NCT06284265	2/28/2024	Placebo-control
13	NA	NA	A Phase 3 Study of LNK01001 Capsule in Moderately to Severely Active Rheumatoid Arthritis	Rheumatology	III	NCT06276998	2/26/2024	Placebo-control
14	NA	NA	Comparison of Postoperative Pain Score Between Perioperative Intravenous Ketamine and Placebo in Patients Undergoing Unilateral Total Knee Arthroplasty Under General Anesthesia	Anesthesia	IV	NCT06267638	2/20/2024	Placebo-control
15	NA	NA	Clinical Trial of the Efficacy and Safety of Raphamin in Combined Treatment of Community-acquired Pneumonia	Infectious Disease	III	NCT06263881	2/16/2024	Placebo-control
16	NA	NA	A Study of Guselkumab in Pediatric	Gastroenterolog	III	NCT0626016	2/15/202	Placebo

			Participants With Moderately to Severely Active Ulcerative Colitis (QUASAR Jr)	y		3	4	-control
17	NA	NA	A Study to Evaluate Mavacamten in Adolescents With Symptomatic Obstructive Hypertrophic Cardiomyopathy	Cardiology	III	NCT06253221	2/12/2024	Placebo-control
18	NA	NA	A Study to Investigate the Effects of PT027 (Budesonide/Albuterol Sulfate) Metered-dose Inhaler Compared With Placebo on Exercise-Induced Bronchoconstriction in Adult Patients With Asthma (BREATH)	Respiratory	III	NCT06245551	2/7/2024	Placebo-control
19	NA	NA	Perfenidone in Type 2 Diabetic Patients With Diabetic Neuropathy (PenDaNt)	Endocrine	IV	NCT06224790	1/25/2024	Placebo-control
20	NA	NA	A Study to Assess Long-term Safety of Fezolinetant Given to Japanese Women Going Through Menopause (Starlight 3)	Obstetrics and gynaecology	III	NCT06206421	1/16/2024	Placebo-control

eTable 2: Averaged objective scoring of LLM output, with NLP based statistical analytical scoring

	TOTAL SCORE	ELIGIBILITY	RECRUITMENT	STUDY ARMS	OUTCOME MEASURES
BLEU	0.04478295	0.04132682	0.25689486	0.07282843	0.0313343
ROUGE-L	0.19870998	0.18931406	0.61798858	0.28428445	0.17543501
METEOR	0.17572606	0.19684976	0.51055223	0.27311108	0.15082139