

Large Language Models in Randomized Controlled Trials Design: Observational Study

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

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 18

 Figures 19

 Figure 1..... 20

 Multimedia Appendixes 21

 Multimedia Appendix 1..... 22

Large Language Models in Randomized Controlled Trials Design: Observational Study

Liyuan Jin^{1*} MD; Jasmine Chiat Ling Ong^{2, 1*} PharmD; Kabilan Elangovan³ BE; Yuhe Ke^{2, 1} MBBS; Alexandra Pyle² PhD; Daniel Shu Wei Ting^{4, 1} PhD; Nan Liu^{1, 5} PhD

¹ Duke-NUS Medical School Singapore SG

² Department of Anaesthesiology and Perioperative Medicine Singapore General Hospital Singapore SG

³ SingHealth Singapore SG

⁴ Singapore National Eye Centre Singapore SG

⁵ NUS Artificial Intelligence Institute National University of Singapore 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 two independent clinical experts by comparing them to clinically validated ground truth data from ClinicalTrials.gov. We have conducted statistical analysis using natural language processing (NLP) based methods, including BLEU, ROGUE-L and METEOR, for objective scoring on corresponding LLM outputs. Qualitative assessments were performed using Likert scale ratings (1–3) for domains such as safety, clinical accuracy, objectivity or bias, 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%). NLP statistical analysis reported BLEU = 0.04, ROGUE-L = 0.20, and METEOR = 0.18 on average objective scoring of LLM output. Qualitative evaluations showed that LLM-generated designs scored above 2 points and closely matched to original designs scoring across all domains, indicating strong clinical alignment. Specifically, both original and LLM based design ranked similarly higher in safety, clinical accuracy, objectivity or bias in published RCTs. While LLM based design ranked non inferior to original designs in ongoing registered RCTs in multiple domains. 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 visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/preprint/67469>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/67469>

Original Manuscript

Original Paper

Large Language Models in Randomized Controlled Trials Design: Observational Study

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: Observational Study

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 two independent clinical experts by comparing them to clinically validated ground truth data from ClinicalTrials.gov. We have conducted statistical analysis using natural language processing (NLP) based methods, including BLEU, ROGUE-L and METEOR, for objective scoring on corresponding LLM outputs. Qualitative assessments were performed using Likert scale ratings (1–3) for domains such as safety, clinical accuracy, objectivity or bias, 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%). NLP statistical analysis reported BLEU = 0.04, ROGUE-L = 0.20, and METEOR = 0.18 on average objective scoring of LLM output. Qualitative evaluations showed that LLM-generated designs scored above 2 points and closely matched to original designs scoring across all domains, indicating strong clinical alignment. Specifically, both original and LLM based design ranked similarly higher in safety, clinical accuracy, objectivity or bias in published RCTs. While LLM based design ranked non inferior to original designs in ongoing registered RCTs in multiple domains. 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:

GPT-4; LLM-generated clinical trial designs; clinical trial design evaluation; 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[1]. RCT provides a carefully controlled environment to investigate cause-effect relationships between therapeutic intervention and clinical outcomes with a high degree of internal validity[2]. Over the years, landmark RCTs have significantly influenced treatment guidelines and improved global standards of care across various medical disciplines[3-5].

However, despite their scientific rigor in evidence, RCTs face persistent and well-documented criticisms of poor generalizability from fixed eligibility criteria[6], lack of diversification in recruitment[7], and practical implementation concerns[6]. Patients with complex co-morbidities or late-stage diseases 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.

In addition to concerns about representativeness, clinical trials face an alarmingly high failure rate, especially in the later stages of development. High failure rate of clinical trials is a key stumbling block in drug development pipelines. RCTs failure rate has been reported for various reasons[8-10], 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[11]. 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[12]. Reform of clinical research is much needed to accelerate this process.

Given the immense time, cost, and effort involved in clinical research, there is an urgent need to reform the RCT design process to address aforementioned challenges. Emerging technologies, particularly large language models (LLMs), offer a novel opportunity to address these challenges. LLMs have recently emerged as an efficient tool in various clinical tasks[13] with comparable clinical alignment to human experts[14]. Developments in NLP empowered LLMs to generate sophisticated and contextually relevant clinical contents. Prominent examples including GPT-4, Gemini, Llama 3 and Claude 3.5, have showcased remarkable versatility and clinical performance in highly specialized clinical tasks [15, 16]. As a result, LLMs tools are expected to assist clinical practice ranging from basic healthcare related administrative work [17, 18], educational chatbot for medical knowledge[19, 20], to advance clinical notes generation[21-23], complex clinical cases diagnosis[24], and patient triaging[25, 26].

Recently, there is increasing interest in LLM applications in clinical trials[27-30]. 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[27-29]. 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[31-33].

In this study, we explore the application of LLMs as a tool for designing randomized controlled trials with clinical alignment and broader applicability. By piloting the use of LLMs in trial design, we aim to assess their potential to enhance the generalizability of study outcomes, optimize eligibility criteria, and ultimately reduce the failure rate of phase III clinical trials. This work contributes to the evolving dialogue on the future of clinical research and offers a practical pathway toward more inclusive, efficient, and evidence-driven trial methodologies.

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 **Supplementary eTable 1**.

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 output are presented in **Supplementary eFigure 1** and **eTable 2**, 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)[34], inclusivity and diversity (adapted from United States Food and Drug Administration (FDA) draft guidance to clinical trial design)[7] 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. Mean scores were calculated based on blinded human expert ratings stratified into RCTs (published/ongoing) with designs (ground truths/LLM designs).

Statistical analysis

We employed average, non-weighted NLP based objective scoring, including BLEU, ROGUE-L and METEOR, for LLM outputs.

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 (reviewer 1 – principal clinical pharmacist; reviewer 2 – specialist physician in anesthesia. Both with > 10 years of clinical practice experience) received no compensation for rating.

Results

Our results showing LLM demonstrated 72% accuracy in overall RCT designs (stratified performance over different design domains are presented in **Supplementary eFigure 2**). 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 NLP based methods, including BLEU[35], ROGUE-L[36] and METEOR[37], for corresponding LLM outputs, presented in **Supplementary eTable 3**. Specifically, BLEU[35] measures n-gram precision to evaluate textual similarity, ROGUE-L[36] focuses on sequence recall and fluency by identifying the longest common subsequences, and METEOR[37] assesses semantic alignment and linguistic variability, incorporating synonyms, stemming, and word order. These metrics collectively provide a comprehensive evaluation of the generated outputs against the reference text. Qualitatively, LLM designs produced comparable clinical alignment, as observed in closely matched Likert scales, RCT design compared to ground truth, with Likert scales scoring above 2 points across all domains (**Figure 1**, grading scores were presented in **Supplementary eTable 4**).

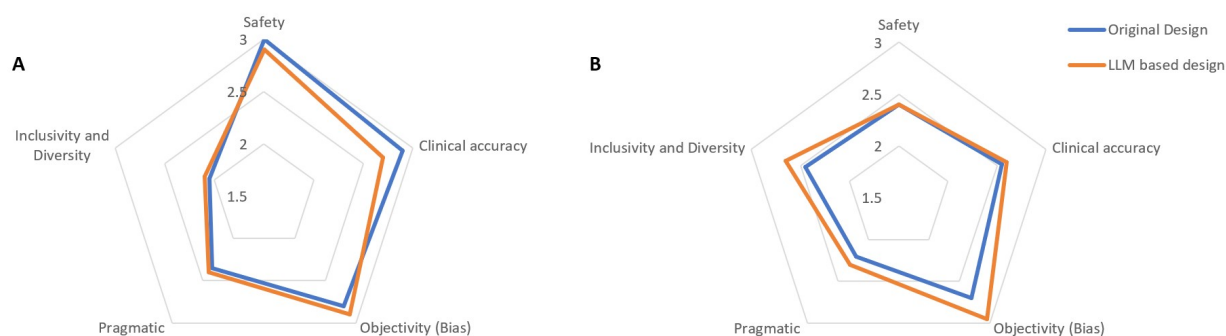


Figure 1. A: Qualitative metrics for 10 published RCTs. B: Qualitative metrics for 10 ongoing registered RCTs.

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[38] 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. Extensive pre-training on medically related free texts set them apart from conventional machine learning or deep learning models, simulating clinical reasoning and inferential skills across diverse disciplines[39, 40], allowing potential integration into sophisticated clinical tasks such as in clinical trial design. We infer that LLM could recommend 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 interventions tested in real-world. Qualitatively, the overall safety and clinical accuracy of these reported differences was not compromised significantly. Stronger performance over recruitment and intervention might be partially explained as LLMs are trained with previous examples of clinical trial designs with better understanding in predicting sample sizes for inclusion and standard therapeutic intervention regime etc. However, inferior performance in eligibility criteria designs and outcomes measurement emphasizes critical clinical insights are necessary to facilitate clinically relevant clinical trial designs. Overall, LLMs based clinical trials

designs might benefit more administrative aspects of clinical trials design such as formulating standard intervention regime and patient sample size while further improvements are necessary to allow designs for highly specialized clinical related domains. 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 has 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. Although both human reviewers were experienced clinicians, the lack of a broader multidisciplinary review panel may limit generalizability of the qualitative findings. Future studies could incorporate more diverse expert raters and certified medical board. 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. Future study with larger sample size, expanding LLMs of interests for evaluating, and cost-effectiveness analysis stratified to various medical specialties are necessary. Furthermore, for phase III/IV trials, substantial works including prior registration and fundings would have been published and would affect interpretation of current study towards approach of LLM RCT designs. Future studies on LLMs design from initial hypothesis and direct comparison with concurrent human expert designs are necessary. 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[41]. For instance, performance over eligibility criteria recognition achieved moderate F1 scores over BERT related LLMs[42]. AutoCriteria, leveraging GPT-4 in a zero-shot setting, significantly improved entity extraction across multiple diseases, highlighting the promise of latest LLMs[43]. Other efforts include classifying exclusion criteria in cancer trials using BERT, again demonstrating LLM feasibility clinical task[44]. GPT-4 has also been explored for sample-size calculation but observed inconsistencies underscores the need for caution in high-stakes applications[45]. Additionally, predictive modeling of trial publication outcomes using BERT demonstrated the utility of LLM in combining structured and unstructured clinical trial data[46]. 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.

Studies	LLM Application	LLM Base Model	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[42]	Eligibility Screening	BERT	470/ 230/ 1000	F1	0.72/ 0.84/ 0.62
AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models[43]	Eligibility Screening	GPT 4	180 Trials	F1	0.90
Text Classification of Cancer Clinical Trial Eligibility Criteria[44]	Eligibility Screening	BERT	764 Trials	ACC	0.27-0.95
ChatGPT for Sample-Size Calculation in Sports Medicine and Exercise Sciences: A Cautionary Note[45]	Sample Size Calculation	GPT 4	4 Trials	ACC	0.75
Medical text classification based on the discriminative pre-training model and prompt-tuning[47]	Assist Trial Outcome Measurement	BERT	5127 Outcome entities	ACC	0.86
Predicting Publication of Clinical Trials Using Structured and Unstructured Data: Model Development and Validation Study[46]	Trial Outcome Prediction	BERT	76,950 Trials	F1	0.70

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

This study contributes significantly to the existing literature by providing empirical data on the accuracy and clinical alignment of LLMs specifically in the context of RCT design. Unlike previous studies, which primarily focus on preliminary text classification tasks, our research applied LLMs to the comprehensive design of RCTs, including elements such as eligibility criteria, recruitment strategies, and intervention arms. Our findings demonstrate that LLMs can replicate existing RCT designs with reasonable accuracy and add value by enhancing the diversity and pragmatism of trial designs. This is crucial in addressing common pitfalls in RCT generalizability and participant diversity. Various factors affect and influence clinical trial accessibility, and a comprehensive, multi-pronged approach is required. Factors include the lack of education on benefits of participating in clinical trials, patient trust and lack of incentives to participate[48]. The design of the clinical trial may inadvertently pose a barrier to entry. Clinical trials often exclude certain populations to a greater

extent than others, such as patients with late-stage organ dysfunction.

Amid the growing interest in the use of LLMs to accelerate clinical trial processes, there is still a paucity of tools developed to improve the overall quality and inclusivity of clinical trials. Our study demonstrated that LLM is capable of assisting in trial design, encompassing elements of “best practices in clinical trial designs”. This can serve as a good reference point for non-subject matter experts including scientific review committees and ethics board. Moving forward, the development of LLM-based agentic AI workflows could further improve utility and performance of LLMs in this application. Specialized LLM agents can be developed and incorporated into a multi-step “checklist” approach to perform critical review and evaluation of various domains of a clinical trial design. Multiagent conversations have been shown to improve LLM output accuracy and mitigate cognitive bias[49].

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 LLMs 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

This work was supported by the Duke-NUS Signature Research Program, funded by the Ministry of Health, Singapore. The funder had no role in study design, conduct, data analysis and interpretation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Singapore Ministry of Health.

Author Contributions

Liyuan Jin and Jasmine Chiat Ling Ong contributed equally to this work.

Conceptualization: Daniel Shu Wei Ting, Nan Liu

Methodology/Investigation: Liyuan Jin, Jasmine Chiat Ling Ong, Kabilan Elangovan

Formal Analysis/validation: Jasmine Chiat Ling Ong, Yuhe Ke

Writing – Original Draft: Liyuan Jin, Jasmine Chiat Ling Ong, Kabilan Elangovan, Yuhe Ke
Alexandra Pyle

Writing – Review & Editing: Liyuan Jin, Jasmine Chiat Ling Ong, Kabilan Elangovan, Nan Liu

Supervision: Daniel Shu Wei Ting, Nan Liu

Project Administration: Nan Liu

Conflicts of Interest

All authors declare no relevant conflicts of interest.

Data Availability Statement

All data generated or analyzed during this study are included in this published article and its supplementary document.

References

1. Bothwell, L.E. and S.H. Podolsky, *The emergence of the randomized, controlled trial*. New England Journal of Medicine, 2016. **375**(6): p. 501-504.
2. Hopewell, S., et al., *CONSORT 2025 statement: updated guideline for reporting randomised trials*. The Lancet, 2025.
3. Group, U.P.D.S., *Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33)*. The lancet, 1998. **352**(9131): p. 837-853.
4. Kass, M.A., et al., *The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma*. Archives of ophthalmology, 2002. **120**(6): p. 701-713.
5. Wykoff, C.C., et al., *Efficacy, durability, and safety of intravitreal faricimab with extended dosing up to every 16 weeks in patients with diabetic macular oedema (YOSEMITE and RHINE): two randomised, double-masked, phase 3 trials*. The Lancet, 2022. **399**(10326): p. 741-755.
6. Nichol, A., et al., *Challenging issues in randomised controlled trials*. Injury, 2010. **41**: p. S20-S23.
7. Gray, D.M., et al., *Diversity in clinical trials: an opportunity and imperative for community engagement*. The Lancet Gastroenterology & Hepatology, 2021. **6**(8): p. 605-607.
8. Stensland, K.D., et al. *Estimating the rate and reasons of clinical trial failure in urologic oncology*. in *Urologic oncology: seminars and original investigations*. 2021. Elsevier.
9. Wong, C.H., K.W. Siah, and A.W. Lo, *Estimation of clinical trial success rates and related parameters*. Biostatistics, 2019. **20**(2): p. 273-286.
10. Pretorius, S. and A. Grignolo, *Phase III trial failures: Costly, but preventable*. 2016.
11. *Artificial Intelligence for Clinical Trial Design: Trends in Pharmacological Sciences*. 2024.
12. Hutson, M., *How AI is being used to accelerate clinical trials*. Nature, 2024. **627**(8003).
13. Thirunavukarasu, A.J., et al., *Large language models in medicine*. Nature medicine, 2023. **29**(8): p. 1930-1940.
14. Singhal, K., et al., *Large language models encode clinical knowledge*. Nature, 2023. **620**(7972): p. 172-180.
15. Ke, Y.H., et al., *Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness*. npj Digital Medicine, 2025. **8**(1): p. 187.
16. Lim, D.Y.Z., et al., *Large language models in anaesthesiology: use of ChatGPT for American Society of Anesthesiologists physical status classification*. British Journal of Anaesthesia, 2023. **131**(3): p. e73-e75.
17. Karakas, C., D. Brock, and A. Lakhota, *Leveraging ChatGPT in the Pediatric Neurology Clinic: Practical Considerations for Use to Improve Efficiency and Outcomes*. Pediatric Neurology, 2023. **148**: p. 157-163.
18. Ong, J.C.L., et al., *A scoping review on generative AI and large language models in mitigating medication related harm*. npj Digital Medicine, 2025. **8**(1): p. 182.
19. Wójcik, S., et al., *Reshaping medical education: Performance of ChatGPT on a PES medical examination*. Cardiology Journal, 2023.
20. Klang, E., 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**.
21. Waisberg, E., et al., *GPT-4 and ophthalmology operative notes*. Annals of Biomedical Engineering, 2023: p. 1-3.
22. Sun, Z., et al., *Evaluating GPT-4 on impressions generation in radiology reports*. Radiology,

2023. **307**(5): p. e231259.
23. Zhou, Z., *Evaluation of ChatGPT's capabilities in medical report generation*. Cureus, 2023. **15**(4).
 24. Kanjee, Z., B. Crowe, and A. Rodman, *Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge*. JAMA, 2023.
 25. Waisberg, E., et al., *GPT-4 for triaging ophthalmic symptoms*. Eye, 2023: p. 1-2.
 26. Lim, G., K. Elangovan, and L. Jin, *Vision language models in ophthalmology*. Curr Opin Ophthalmol, 2024. **35**(6): p. 487-493.
 27. Ghim, J.-L. and S. Ahn, *Transforming clinical trials: the emerging roles of large language models*. Translational and Clinical Pharmacology, 2023. **31**(3): p. 131.
 28. Wong, C., et al. *Scaling clinical trial matching using large language models: A case study in oncology*. in *Machine Learning for Healthcare Conference*. 2023. PMLR.
 29. Jin, Q., et al., *Matching patients to clinical trials with large language models*. ArXiv, 2023.
 30. Tayebi Arasteh, S., et al., *Large language models streamline automated machine learning for clinical studies*. Nature Communications, 2024. **15**(1): p. 1603.
 31. Moher, D., et al., *CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials*. Bmj, 2010. **340**.
 32. Schulz, K.F., D.G. Altman, and D. Moher, *CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials*. Journal of Pharmacology and pharmacotherapeutics, 2010. **1**(2): p. 100-107.
 33. Chan, A.-W., et al., *SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials*. Bmj, 2013. **346**.
 34. Loudon, K., et al., *The PRECIS-2 tool: designing trials that are fit for purpose*. bmj, 2015. **350**.
 35. Papineni, K., et al. *Bleu: a method for automatic evaluation of machine translation*. in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002.
 36. Lin, C.-Y. *Rouge: A package for automatic evaluation of summaries*. in *Text summarization branches out*. 2004.
 37. Banerjee, S. and A. Lavie. *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments*. in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005.
 38. Food and D. Administration, *Evaluating inclusion and exclusion criteria in clinical trials*. 2020.
 39. Singhal, K., et al., *Large language models encode clinical knowledge*. Nature, 2023.
 40. Wei, J., et al., *Emergent Abilities of Large Language Models*. 2022.
 41. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805, 2018.
 42. Li, J., et al., *A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora*. BMC Medical Informatics and Decision Making, 2022. **22**(Suppl 3): p. 235.
 43. Datta, S., et al., *AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models*. Journal of the American Medical Informatics Association, 2024. **31**(2): p. 375-385.
 44. Yang, Y., et al. *Text classification of cancer clinical trial eligibility criteria*. in *AMIA Annual Symposium Proceedings*. 2024.
 45. Methnani, J., et al., *ChatGPT for sample-size calculation in sports medicine and exercise sciences: A cautionary note*. International Journal of Sports Physiology and Performance, 2023. **18**(10): p. 1219-1223.

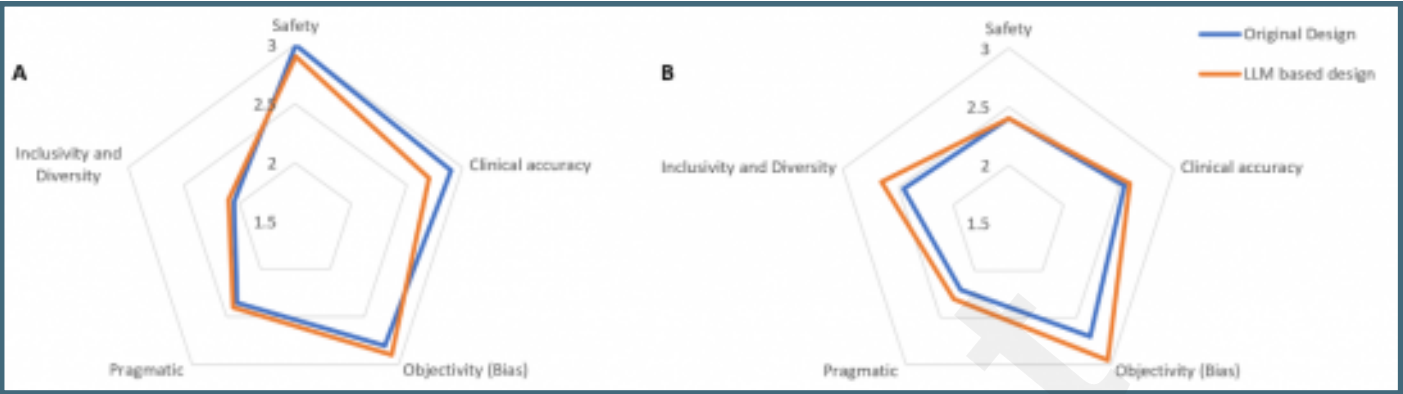
46. Wang, S., et al., *Predicting publication of clinical trials using structured and unstructured data: model development and validation study*. Journal of Medical Internet Research, 2022. **24**(12): p. e38859.
47. Wang, Y., et al., *Medical text classification based on the discriminative pre-training model and prompt-tuning*. Digital Health, 2023. **9**: p. 20552076231193213.
48. Bodicoat, D.H., et al., *Promoting inclusion in clinical trials-a rapid review of the literature and recommendations for action*. Trials, 2021. **22**(1): p. 880.
49. Ke, Y., et al., *Mitigating Cognitive Biases in Clinical Decision-Making Through Multi-Agent Conversations Using Large Language Models: Simulation Study*. J Med Internet Res, 2024. **26**: p. e59439.



Supplementary Files

Figures

A: Qualitative metrics for 10 published RCTs. B: Qualitative metrics for 10 ongoing registered RCTs.



Multimedia Appendixes

Supplementary materials.

URL: <http://asset.jmir.pub/assets/4c061963b365610cd90501c30ba11428.pdf>

