

Original Paper

Large Language Models in Randomized Controlled Trials Design: Observational Study

Liyuan Jin^{1*}, MD; Jasmine Chiat Ling Ong^{1,2*}, PharmD; Kabilan Elangovan³, BE; Yuhe Ke^{1,4}, MBBS; Alexandra Pyle³, PhD; Daniel Shu Wei Ting^{1,3,5}, PhD; Nan Liu^{1,6}, PhD

¹Duke-NUS Medical School, Singapore, Singapore

²Division of Pharmacy, Singapore General Hospital, Singapore, Singapore

³Artificial Intelligence Office, SingHealth, Singapore, Singapore

⁴Department of Anaesthesiology and Perioperative Medicine, Singapore General Hospital, Singapore, Singapore

⁵Singapore National Eye Centre, Singapore, Singapore

⁶NUS Artificial Intelligence Institute, National University of Singapore, Singapore, Singapore

*these authors contributed equally

Corresponding Author:

Nan Liu, PhD
Duke-NUS Medical School
8 College Road
Singapore 169857
Phone: 65 66016503
Email: liu.nan@duke-nus.edu.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 noninterventional, observational study analyzing 20 parallel-arm RCTs, comprising 10 completed and 10 registered 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 2 independent clinical experts by comparing them to clinically validated ground truth data from ClinicalTrials.gov. We have conducted statistical analysis using natural language processing-based methods, including Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-L, and Metric for Evaluation of Translation with Explicit ORdering (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%). Natural language processing statistical analysis reported BLEU=0.04, ROUGE-L=0.20, and METEOR=0.18 on average objective scoring of LLM outputs. Qualitative evaluations showed that LLM-generated designs scored above 2 points and closely matched the original designs in scores across all domains, indicating strong clinical alignment. Specifically, both original and LLM-based designs ranked similarly high in safety, clinical accuracy, and objectivity or bias in published RCTs. Moreover, LLM-based design ranked noninferior to original designs in 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.

J Med Internet Res 2025;27:e67469; doi: [10.2196/67469](https://doi.org/10.2196/67469)

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 (RCTs) 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 comorbidities or late-stage diseases excluded from phase 3 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 the failure of phase 3 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 in investment and taking more than a decade of work to bring a new drug to 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 the 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 natural language processing (NLP) empowered LLMs to generate sophisticated and contextually relevant clinical content. 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, LLM tools are expected to assist clinical practice ranging from basic health care-related administrative work

[17,18], educational chatbots for medical knowledge [19,20], to advanced clinical notes generation [21-23], complex clinical cases diagnosis [24], and patient triaging [25,26].

Recently, there has been increasing interest in LLM applications in clinical trials [27-30]. Generative artificial intelligence 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 RCTs 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 3 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

Overview

We performed an observational, noninterventional study using GPT-4-Turbo-Preview as a state-of-the-art LLM for designing RCTs.

Validation and Testing Datasets

We randomly selected 20 parallel-arm RCTs (phase 3 or 4): 10 completed RCTs, with results published in leading clinical journals (JAMA, Nature Medicine, NEJM, and The Lancet); and 10 RCTs registered on ClinicalTrials.gov. To mitigate the risks of LLMs' pretraining use in such studies, we used studies published or newly registered after January 2024 (after the GPT-4-Turbo-Preview pretraining date of December 2023). Details of the dataset are presented in Table S1 in [Multimedia Appendix 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 or treatment. We then prompted the LLM for the following outputs: eligibility criteria (inclusion and exclusion criteria), recruitment (sex or gender and age), arm or intervention (active and control arms), and outcomes measurement (measurement design and measurement time frame).

Large Language Model

In this 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 Figure S1 and Table S2 in [Multimedia Appendix 1](#), respectively.

Quantitative Evaluation

We quantitatively evaluated the accuracy (degree of agreement) of the LLMs' outputs by comparing them with the clinically defined ground truth. We first collect ground truth for published studies from the publication (cross-examined with the corresponding study from ClinicalTrials.gov), and recent registered 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 the ground truth. For outputs with clinical answers, such as eligibility criteria, active and control arms in intervention, and measurement design in outcome measures, we defined answers as correct if clinically aligned with the ground truth. Specifically, for eligibility criteria designs, the accuracy was determined by the number of matched LLM designs divided by the total number of eligibility criteria listed by LLM.

We created a qualitative assessment metric to evaluate both LLM and ground truth designs. This metric comprised 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 3-point Likert Scale (1 is the worst and 3 is the best). For selected registered 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 and registered) with designs (ground truths and LLM designs).

Statistical Analysis

We used average, nonweighted NLP-based objective scoring, including Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-L, and Metric for Evaluation of Translation with Explicit ORdering (METEOR) for LLM outputs.

Ethical Considerations

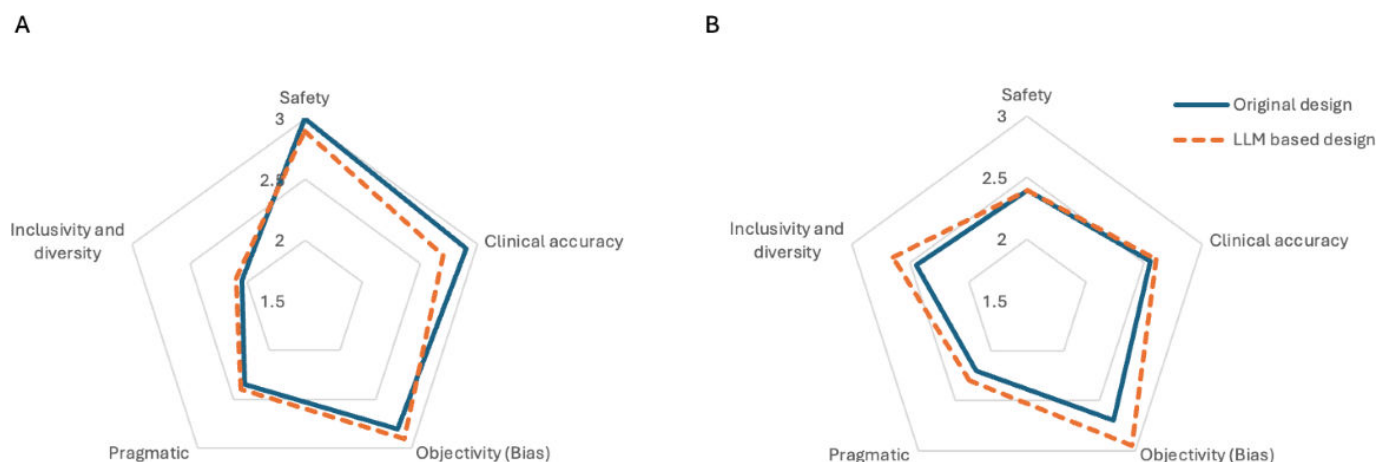
As this study is retrospective in nature and no real patient was involved in the current research, regulatory approval

and informed consent are 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 show that LLM demonstrated 72% accuracy in overall RCT designs (stratified performance across different design domains is presented in Figure S2 in [Multimedia Appendix 1](#)). Specifically, it showed high agreement in Recruitment and Arm or Intervention, with accuracy of 88% and 93%, respectively. However, it demonstrated discrepancies in designing Eligibility Criteria and Outcomes Measurement, with an accuracy of 55% and 53%, respectively. We observed marginal differences in accuracy between LLM outputs and both published RCTs and registered RCTs, except for an improvement in exclusion criteria designs in the latest RCTs. We used statistical analysis using NLP-based methods, including BLEU [35], ROUGE-L [36], and METEOR [37], for corresponding LLM outputs, presented in Table S3 in [Multimedia Appendix 1](#). Specifically, BLEU [35] measures n-gram precision to evaluate textual similarity, ROUGE-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 Table S4 in [Multimedia Appendix 1](#)).

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 or Control Intervention. When assessed qualitatively, we observed marginal differences in the overall clinical accuracy of the LLM design compared with the ground truth, highlighting multiple accepted clinical decisions related to RCT design. Upon qualitative analysis, LLM-based RCT designs closely aligned with 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. In addition, LLM could avoid critical safety and ethical issues identified in the ground truth from the analysis of the selected registered RCTs.

Figure 1. (A) Qualitative metrics for 10 published RCTs. (B) Qualitative metrics for 10 registered RCTs.

Discussion

Principal Findings

RCTs serve key roles in clinical practice, and inclusivity has been heavily emphasized by the FDA [38] to ensure consistently high-quality design that is scientifically justifiable. Current results highlight the potential role of 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 such as GPT-4 can perform tasks with little or no task-specific fine-tuning. Extensive pretraining on medically related free texts sets them apart from conventional machine learning or deep learning models, simulating clinical reasoning and inferential skills across diverse disciplines [39], allowing potential integration into sophisticated clinical tasks such as in clinical trial design. We infer that LLM could recommend the most commonly used comparator arms for trials of similar nature and discipline; logical deduction of active intervention dosage regimen based on preclinical or phase 1 and phase 2 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 the actual published design. These design elements often vary widely across different studies and interventions tested in the real world. Qualitatively, the overall safety and clinical accuracy of these reported differences was not compromised significantly. Stronger performance in recruitment and intervention might be partially explained by the fact that LLMs are trained on previous examples of clinical trial designs, with better understanding in predicting sample sizes for inclusion and standard therapeutic intervention regimes. However, inferior performance in eligibility criteria designs and outcomes measurement emphasizes that critical clinical insights are necessary to facilitate clinically relevant clinical trial designs. Overall, LLM-based clinical trial designs might benefit more administrative aspects of clinical trial design, such as formulating standard intervention regimes and determining patient sample size, while further improvements are

necessary to allow designs for highly specialized clinical trial-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 the generalizability of the qualitative findings. Future studies could incorporate more diverse expert raters and a 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 LLMs' capabilities across diverse medical specialties. Future studies with larger sample sizes, expanding LLMs of interest for evaluation, and cost-effectiveness analysis stratified by various medical specialties are necessary. Furthermore, for phase 3 and phase 4 trials, substantial work including prior registration and funding would have been published and would affect the interpretation of this study toward the approach of LLM-based RCT designs. Future studies on LLM design from the initial hypothesis and direct comparison with concurrent human expert designs are necessary. Finally, alternative trial designs such as open-label, crossover, or pragmatic trials were not considered in this study.

Comparison With Prior Work

To identify relevant studies, 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 or Abstract]) AND ("artificial intelligence" [MeSH Terms] OR "generative AI" [Title or Abstract] OR "language model" [Title or Abstract])

AND (2022:2024[pdat]). We restricted the search to articles published in PubMed between January 1, 2022, and April 1, 2024. We screened a total of 575 articles from PubMed and included a final total of 6 publications. We included peer-reviewed articles investigating the performance of generative artificial intelligence models applied in the conduct of clinical trials or RCTs. We excluded review papers and studies that did not report any model performance.

Existing clinical trial-related LLM studies, presented in Table 1, have only focused on preliminary text classification tasks and are mostly limited to last-generation LLMs, such as Bidirectional Encoder Representations from Transformers (BERT) [40]. For instance, performance over eligibility criteria recognition achieved a moderate F_1 -score over BERT-related LLMs [41]. AutoCriteria, leveraging GPT-4 in a zero-shot setting, significantly improved entity extraction

across multiple diseases, highlighting the promise of the latest LLMs [42]. Other efforts include classifying exclusion criteria in cancer trials using BERT, again demonstrating LLM feasibility in clinical tasks [43]. GPT-4 has also been explored for sample size calculation, but observed inconsistencies underscore the need for caution in high-stakes applications [44]. In addition, predictive modeling of trial publication outcomes using BERT demonstrated the utility of LLM in combining structured and unstructured clinical trial data [45]. With rapid advancement in LLM development and taking advantage of LLMs’ accessibility and efficiency as demonstrated in this 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 interventions.

Table 1. Existing large language model applications in clinical trials-related studies.

Studies	LLM ^a application	LLM ^a base model	Testing dataset sample size	Evaluation metrics used	Model performance
A comparative study of pretrained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora [41]	Eligibility screening	BERT ^b	470/230/1000	F_1 -score	0.72/0.84/0.62
AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models [42]	Eligibility screening	GPT-4 ^c	180 trials	F_1 -score	0.90
Text classification of cancer clinical trial eligibility criteria [43]	Eligibility screening	BERT ^b	764 trials	ACC ^d	0.27-0.95
ChatGPT for sample size calculation in sports medicine and exercise sciences: a cautionary note [44]	Sample size calculation	GPT 4 ^c	4 trials	ACC	0.75
Medical text classification based on the discriminative pretraining model and prompt-tuning [46]	Assist trial outcome measurement	BERT ^b	5127 outcome entities	ACC	0.86
Predicting publication of clinical trials using structured and unstructured data: model development and validation study [45]	Trial outcome prediction	BERT ^b	76,950 trials	F_1 -score	0.70

^aLLM: large language models
^bBERT: Bidirectional Encoder Representations from Transformers
^cGPT: Generative Pre-trained Transformer 4
^dACC: 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, multipronged approach is required. Other factors include the lack of education on the benefits of

participating in clinical trials, patient trust, and the lack of incentives to participate [47]. 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 nonsubject matter experts, including scientific review committees and ethics boards. Moving forward, the development of LLM-based agentic artificial

intelligence workflows could further improve the utility and performance of LLMs in this application. Specialized LLM agents can be developed and incorporated into a multistep “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 [48].

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 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.

Acknowledgments

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.

Data Availability

Data are supplied in supporting files available for download along with the published manuscript.

Authors' Contributions

LJ and JCLO contributed equally to this work. DSWT and NL were responsible for conceptualization. LJ, JCLO, and KE carried out the methodology and investigation. JCLO and YK performed the formal analysis and validation. The original draft was written by LJ, JCLO, KE, YK, and AP. LJ, JCLO, KE, and NL reviewed and edited the manuscript. DSWT and NL supervised the project. Project administration was carried out by NL.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supporting files on study design and evaluations.

[[PDF File \(Adobe File\), 1132 KB-Multimedia Appendix 1](#)]

References

1. Bothwell LE, Podolsky SH. The emergence of the randomized, controlled trial. *N Engl J Med*. Aug 11, 2016;375(6):501-504. [doi: [10.1056/NEJMp1604635](#)] [Medline: [27509097](#)]
2. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. *Lancet*. Apr 14, 2025;405(10489):1633-1640. [doi: [10.1016/S0140-6736\(25\)00672-5](#)] [Medline: [40245901](#)]
3. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK Prospective Diabetes Study (UKPDS) Group. *Lancet*. Sep 12, 1998;352(9131):837-853. [doi: [10.1016/S0140-6736\(98\)07019-6](#)] [Medline: [9742976](#)]
4. Kass MA, Heuer DK, Higginbotham EJ, 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. *Arch Ophthalmol*. Jun 2002;120(6):701-713. [doi: [10.1001/archophth.120.6.701](#)] [Medline: [12049574](#)]
5. Wykoff CC, Abreu F, Adamis AP, 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. *Lancet*. Feb 19, 2022;399(10326):741-755. [doi: [10.1016/S0140-6736\(22\)00018-6](#)] [Medline: [35085503](#)]
6. Nichol AD, Bailey M, Cooper DJ, POLAR, EPO Investigators. Challenging issues in randomised controlled trials. *Injury*. Jul 2010;41 Suppl 1:S20-3. [doi: [10.1016/j.injury.2010.03.033](#)] [Medline: [20413119](#)]
7. Gray DM II, Nolan TS, Gregory J, Joseph JJ. Diversity in clinical trials: an opportunity and imperative for community engagement. *Lancet Gastroenterol Hepatol*. Aug 2021;6(8):605-607. [doi: [10.1016/S2468-1253\(21\)00228-4](#)] [Medline: [34246352](#)]
8. Stensland KD, DePorto K, Ryan J, et al. Estimating the rate and reasons of clinical trial failure in urologic oncology. *Urol Oncol*. Mar 2021;39(3):154-160. [doi: [10.1016/j.urolonc.2020.10.070](#)] [Medline: [33257221](#)]
9. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics*. Apr 1, 2019;20(2):273-286. [doi: [10.1093/biostatistics/kxx069](#)] [Medline: [29394327](#)]

10. Hwang TJ, Carpenter D, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS. Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA Intern Med.* Dec 1, 2016;176(12):1826-1833. [doi: [10.1001/jamainternmed.2016.6008](https://doi.org/10.1001/jamainternmed.2016.6008)] [Medline: [27723879](https://pubmed.ncbi.nlm.nih.gov/27723879/)]
11. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. *Trends Pharmacol Sci.* Aug 2019;40(8):577-591. [doi: [10.1016/j.tips.2019.05.005](https://doi.org/10.1016/j.tips.2019.05.005)] [Medline: [31326235](https://pubmed.ncbi.nlm.nih.gov/31326235/)]
12. Hutson M. How AI is being used to accelerate clinical trials. *Nature New Biol.* Mar 2024;627(8003):S2-S5. [doi: [10.1038/d41586-024-00753-x](https://doi.org/10.1038/d41586-024-00753-x)] [Medline: [38480968](https://pubmed.ncbi.nlm.nih.gov/38480968/)]
13. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* Aug 2023;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
14. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature New Biol.* Aug 2023;620(7972):172-180. [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
15. Ke YH, Jin L, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *NPJ Digit Med.* Apr 5, 2025;8(1):187. [doi: [10.1038/s41746-025-01519-z](https://doi.org/10.1038/s41746-025-01519-z)] [Medline: [40185842](https://pubmed.ncbi.nlm.nih.gov/40185842/)]
16. Lim DYZ, Ke YH, Sng GGR, Tung JYM, Chai JX, Abdullah HR. Large language models in anaesthesiology: use of ChatGPT for American Society of Anesthesiologists physical status classification. *Br J Anaesth.* Sep 2023;131(3):e73-e75. [doi: [10.1016/j.bja.2023.06.052](https://doi.org/10.1016/j.bja.2023.06.052)] [Medline: [37474421](https://pubmed.ncbi.nlm.nih.gov/37474421/)]
17. Karakas C, Brock D, Lakhotia A. Leveraging ChatGPT in the pediatric neurology clinic: practical considerations for use to improve efficiency and outcomes. *Pediatr Neurol.* Nov 2023;148:157-163. [doi: [10.1016/j.pediatrneurol.2023.08.035](https://doi.org/10.1016/j.pediatrneurol.2023.08.035)] [Medline: [37725885](https://pubmed.ncbi.nlm.nih.gov/37725885/)]
18. Ong JCL, Chen MH, Ng N, et al. A scoping review on generative AI and large language models in mitigating medication related harm. *NPJ Digit Med.* Mar 28, 2025;8(1):182. [doi: [10.1038/s41746-025-01565-7](https://doi.org/10.1038/s41746-025-01565-7)] [Medline: [40155703](https://pubmed.ncbi.nlm.nih.gov/40155703/)]
19. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: performance of ChatGPT on a PES medical examination. *Cardiol J.* 2024;31(3):442-450. [doi: [10.5603/cj.97517](https://doi.org/10.5603/cj.97517)] [Medline: [37830257](https://pubmed.ncbi.nlm.nih.gov/37830257/)]
20. 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 Med Educ.* Oct 17, 2023;23(1):772. [doi: [10.1186/s12909-023-04752-w](https://doi.org/10.1186/s12909-023-04752-w)] [Medline: [37848913](https://pubmed.ncbi.nlm.nih.gov/37848913/)]
21. Waisberg E, Ong J, Masalkhi M, et al. GPT-4 and ophthalmology operative notes. *Ann Biomed Eng.* Nov 2023;51(11):2353-2355. [doi: [10.1007/s10439-023-03263-5](https://doi.org/10.1007/s10439-023-03263-5)] [Medline: [37266720](https://pubmed.ncbi.nlm.nih.gov/37266720/)]
22. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. *Radiology.* Jun 2023;307(5):e231259. [doi: [10.1148/radiol.231259](https://doi.org/10.1148/radiol.231259)] [Medline: [37367439](https://pubmed.ncbi.nlm.nih.gov/37367439/)]
23. Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. *Cureus.* Apr 2023;15(4):e37589. [doi: [10.7759/cureus.37589](https://doi.org/10.7759/cureus.37589)] [Medline: [37197105](https://pubmed.ncbi.nlm.nih.gov/37197105/)]
24. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA.* Jul 3, 2023;330(1):78-80. [doi: [10.1001/jama.2023.8288](https://doi.org/10.1001/jama.2023.8288)] [Medline: [37318797](https://pubmed.ncbi.nlm.nih.gov/37318797/)]
25. Waisberg E, Ong J, Zaman N, et al. GPT-4 for triaging ophthalmic symptoms. *Eye (Lond).* Dec 2023;37(18):3874-3875. [doi: [10.1038/s41433-023-02595-9](https://doi.org/10.1038/s41433-023-02595-9)] [Medline: [37231187](https://pubmed.ncbi.nlm.nih.gov/37231187/)]
26. Lim G, Elangovan K, Jin L. Vision language models in ophthalmology. *Curr Opin Ophthalmol.* Nov 1, 2024;35(6):487-493. [doi: [10.1097/ICU.0000000000001089](https://doi.org/10.1097/ICU.0000000000001089)] [Medline: [39259649](https://pubmed.ncbi.nlm.nih.gov/39259649/)]
27. Ghim JL, Ahn S. Transforming clinical trials: the emerging roles of large language models. *Transl Clin Pharmacol.* Sep 2023;31(3):131-138. [doi: [10.12793/tcp.2023.31.e16](https://doi.org/10.12793/tcp.2023.31.e16)] [Medline: [37810626](https://pubmed.ncbi.nlm.nih.gov/37810626/)]
28. Wong C, et al. Scaling clinical trial matching using large language models: a case study in oncology. Presented at: Machine Learning for Healthcare Conference; Aug 11-12, 2023; Columbia University. URL: <https://proceedings.mlr.press/v219/wong23a.html> [Accessed 2025-08-25]
29. Jin Q, Wang Z, Floudas CS, et al. Matching patients to clinical trials with large language models. *arXiv.* Preprint posted online on Jul 27, 2023. [doi: [10.48550/arXiv.2307.15051](https://doi.org/10.48550/arXiv.2307.15051)]
30. Tayebi Arasteh S, Han T, Lotfinia M, et al. Large language models streamline automated machine learning for clinical studies. *Nat Commun.* Feb 21, 2024;15(1):1603. [doi: [10.1038/s41467-024-45879-8](https://doi.org/10.1038/s41467-024-45879-8)] [Medline: [38383555](https://pubmed.ncbi.nlm.nih.gov/38383555/)]
31. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ.* Mar 23, 2010;340:c869. [doi: [10.1136/bmj.c869](https://doi.org/10.1136/bmj.c869)] [Medline: [20332511](https://pubmed.ncbi.nlm.nih.gov/20332511/)]
32. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *J Pharmacol Pharmacother.* Jul 2010;1(2):100-107. [doi: [10.4103/0976-500X.72352](https://doi.org/10.4103/0976-500X.72352)] [Medline: [21350618](https://pubmed.ncbi.nlm.nih.gov/21350618/)]

33. Chan AW, Tetzlaff JM, Gøtzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ*. Jan 8, 2013;346:e7586. [doi: [10.1136/bmj.e7586](https://doi.org/10.1136/bmj.e7586)] [Medline: [23303884](https://pubmed.ncbi.nlm.nih.gov/23303884/)]
34. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ*. May 8, 2015;350:h2147. [doi: [10.1136/bmj.h2147](https://doi.org/10.1136/bmj.h2147)] [Medline: [25956159](https://pubmed.ncbi.nlm.nih.gov/25956159/)]
35. Papineni K, Roukos S, Ward T, Zhu W. Bleu: a method for automatic evaluation of machine translation. Presented at: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; Jul 6, 2002; Philadelphia, PA. [doi: [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135)]
36. Lin CY. Rouge: a package for automatic evaluation of summaries. Presented at: In Proceedings of the Workshop on Text Summarization Branches Out; Barcelona, Spain. 2004.URL: <https://aclanthology.org/W04-1013/> [Accessed 2025-08-25]
37. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. Presented at: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization; Jun 2005; Ann Arbor, MI. URL: <https://aclanthology.org/W05-0909/> [Accessed 2025-08-25]
38. Evaluating inclusion and exclusion criteria in clinical trials. U.S. Food and Drug Administration; 2020. URL: <https://www.fda.gov/media/134754/download> [Accessed 2025-08-25]
39. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. *arXiv*. Preprint posted online on Oct 26, 2022. [doi: [10.48550/arXiv.2206.07682](https://doi.org/10.48550/arXiv.2206.07682)]
40. Devlin J, Chang M, Lee K, Toutanova K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv*. Preprint posted online on May 24, 2019. [doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805)]
41. Li J, Wei Q, Ghiasvand O, et al. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC Med Inform Decis Mak*. Sep 6, 2022;22(Suppl 3):235. [doi: [10.1186/s12911-022-01967-7](https://doi.org/10.1186/s12911-022-01967-7)] [Medline: [36068551](https://pubmed.ncbi.nlm.nih.gov/36068551/)]
42. Datta S, Lee K, Paek H, et al. AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *J Am Med Inform Assoc*. Jan 18, 2024;31(2):375-385. [doi: [10.1093/jamia/ocad218](https://doi.org/10.1093/jamia/ocad218)] [Medline: [37952206](https://pubmed.ncbi.nlm.nih.gov/37952206/)]
43. Yang Y, Jayaraj S, Ludmir E, Roberts K. Text classification of cancer clinical trial eligibility criteria. *AMIA Annu Symp Proc*. 2023;2023:1304-1313. [Medline: [38222417](https://pubmed.ncbi.nlm.nih.gov/38222417/)]
44. Methnani J, Latiri I, Dergaa I, Chamari K, Ben Saad H. ChatGPT for sample-size calculation in sports medicine and exercise sciences: a cautionary note. *Int J Sports Physiol Perform*. Oct 1, 2023;18(10):1219-1223. [doi: [10.1123/ijsspp.2023-0109](https://doi.org/10.1123/ijsspp.2023-0109)] [Medline: [37536678](https://pubmed.ncbi.nlm.nih.gov/37536678/)]
45. Wang S, Šuster S, Baldwin T, Verspoor K. Predicting publication of clinical trials using structured and unstructured data: model development and validation study. *J Med Internet Res*. Dec 23, 2022;24(12):e38859. [doi: [10.2196/38859](https://doi.org/10.2196/38859)] [Medline: [36563029](https://pubmed.ncbi.nlm.nih.gov/36563029/)]
46. Wang Y, Wang Y, Peng Z, Zhang F, Zhou L, Yang F. Medical text classification based on the discriminative pre-training model and prompt-tuning. *Digit Health*. 2023;9:20552076231193213. [doi: [10.1177/20552076231193213](https://doi.org/10.1177/20552076231193213)] [Medline: [37559830](https://pubmed.ncbi.nlm.nih.gov/37559830/)]
47. Bodicoat DH, Routen AC, Willis A, et al. Promoting inclusion in clinical trials-a rapid review of the literature and recommendations for action. *Trials*. Dec 4, 2021;22(1):880. [doi: [10.1186/s13063-021-05849-7](https://doi.org/10.1186/s13063-021-05849-7)] [Medline: [34863265](https://pubmed.ncbi.nlm.nih.gov/34863265/)]
48. Ke Y, Yang R, Lie SA, et al. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. *J Med Internet Res*. Nov 19, 2024;26:e59439. [doi: [10.2196/59439](https://doi.org/10.2196/59439)] [Medline: [39561363](https://pubmed.ncbi.nlm.nih.gov/39561363/)]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BLEU: bilingual evaluation understudy
FDA: US Food and Drug Administration
LLM: large language models
METEOR: Metric for Evaluation of Translation with Explicit Ordering
NLP: natural language processing
RCT: randomized controlled trial
ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Edited by Javad Sarvestan; peer-reviewed by Dingqiao Wang, Neil Desai; submitted 12.10.2024; final revised version received 26.04.2025; accepted 28.04.2025; published 03.09.2025

*Please cite as:**Jin L, Ong JCL, Elangovan K, Ke Y, Pyle A, Ting DSW, Liu N**Large Language Models in Randomized Controlled Trials Design: Observational Study**J Med Internet Res 2025;27:e67469*URL: <https://www.jmir.org/2025/1/e67469>doi: [10.2196/67469](https://doi.org/10.2196/67469)

© Liyuan Jin, Jasmine Chiat Ling Ong, Kabilan Elangovan, Yuhe Ke, Alexandra Pyle, Daniel Shu Wei Ting, Nan Liu. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 03.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.