Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Performance Evaluation of Large Language Model Managing Clinical Trial-Related Adverse Events

Yankang Cui
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Tingling Zhang
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Kexin Zhang
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
xiaowen Tan
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Dian Fu
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Bo Fang
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Aiming Jiang
Changhai Hospital
Jingping Ge
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Zhiling Zhang
Sun Yat-sen University Cancer Center
Shun Zhang
Nanjing Drum Tower Hospital: Nanjing University Medical School Affiliated Nanjing Drum Tower Hospital
Zhichao Hu
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Guangyuan Zhang
Zhongda Hospital Southeast University
Han Guan
First Affiliated Hospital of Bengbu Medical University
Yuzheng Ge
Nanjing First Hospital
Nengwang Yu
Qilu Hospital of Shandong University
Zicheng Xu
Jiangsu Cancer Hospital

Yuanyuan Qu
Shanghai Genitourinary Cancer Institute
Xinning Wang
Affiliated Hospital of Qingdao University
Qingyi Zhu
Second Affiliated Hospital of Nanjing Medical University
Haowei He
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Jie Dong
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Bianjiang Liu
Jiangsu Province Hospital
Hao Tang
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Feng Xu
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Tianyi Shen
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Jie Cao
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Song Xu
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Łukasz Zapała
Medical University of Warsaw
Xuejun Shang
Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command
Le Qu
septsoul@hotmail.com

Nanjing Jinling Hospital: General Hospital of Eastern Theatre Command

---

Article

Keywords: Artificial intelligence, chatbots, clinical trial, adverse events, DeepSeek

Posted Date: June 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8972183/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Artificial intelligence (AI)-based chatbots have emerged as significant sources of medical and general information. However, their accuracy and completeness in addressing clinical trial-related adverse events (ctrAEs) remain understudied.

Methods

This study evaluates the performance of three large language models (GPT-4o, Gemini 1.5 pro and DeepSeek-R1) in responding to 62 clinically validated questions spanning eight clinical trial-related general questions, seven ctrAE categories (gastrointestinal, Central Nervous System, Allergic Reactions, Respiratory, Cardiac, Liver Function, Kidney) and 18 patient-specific scenarios. Expert evaluations of accuracy and completeness were conducted using a 4-point Likert scale (1: least; 4: most).

Results

Overall, all engines performed well in terms of accuracy and completeness. Scores indicating significant inaccuracy or incompleteness (1–2) were not found in DeepSeek, and exceptionally rare for ChatGPT, with only 6 out of 496 answer ratings (1.21%) falling into this category. Of the 62 questions evaluated, all 8 physician raters assigned DeepSeek a score of 4 (fully accurate or complete) for 39 questions related to accuracy and 34 questions related to completeness. In the 18 patient scenarios, DeepSeek's average accuracy score was 3.833 (median = 4), while the average completeness score was 3.799 (median = 4).

Conclusion

DeepSeek exhibits robust capabilities in delivering accurate and comprehensive guidance on ctrAE management, outperforming ChatGPT and Gemini. Its consistent high ratings across diverse clinical contexts suggest potential utility in supporting real-time decision-making for trial-related adverse events. Limitations include reliance on simulated scenarios and rater subjectivity. Future studies should validate these findings in live clinical settings. This work underscores the importance of model-specific validation prior to clinical implementation.

Background

The role of the surgical oncologist has evolved beyond the operating theater. As key members of the multidisciplinary team (MDT), they are increasingly involved in the selection and management of patients enrolled in neoadjuvant, adjuvant, and conversion therapy clinical trials involving novel systemic agents. Consequently, the ability to efficiently recognize and manage treatment-related adverse events (e.g.,

immune-related adverse events from checkpoint inhibitors) has become an essential skill. The management of multi-system adverse drug reactions in clinical trials presents significant challenges for specialists, particularly when dealing with events involving non-specialty domains. While guidelines recommend interdisciplinary consultation, real-time access to expert support remains limited in practice, potentially leading to delayed or suboptimal ctrAEs management(1). Patients frequently demonstrate limited understanding of clinical trials, with surveys revealing that more than 60% lack awareness of potential benefits (such as access to novel therapies) or risks (including unforeseen adverse effects) associated with participation(2). Many participants consequently rely on unverified online sources when making decisions regarding trial enrollment. This knowledge gap creates significant information asymmetry that may adversely affect both treatment adherence and patient safety.

In recent years, large language models (LLMs) such as ChatGPT, Gemini, and DeepSeek have gained widespread adoption in healthcare, with applications spanning clinical decision support, patient education, and research assistance(3). For instance, LLM-powered chatbots can explain trial protocols(4), drug mechanisms(5), and risk-benefit ratios to patients in natural language(6). In professional settings, LLMs are increasingly used to generate ctrAEs management recommendations by synthesizing cross-disciplinary guidelines and literature. However, existing studies predominantly focus on accuracy or general medical knowledge, leaving a critical gap in evaluating LLMs’dual capabilities in patient-facing clinical trial communication and clinician-oriented ctrAEs management(7).

Patients often search for clinical trial information online, but the accuracy and reliability of web-based content vary widely. LLMs may improve understanding by delivering structured responses to complex medical queries. Studies suggest ~ 75% of patients prefer LLM-generated personalized explanations over generic web content(8). Nevertheless, model outputs risk "hallucinations" (unsubstantiated claims) or outdated information, particularly regarding recent trial data(9, 10).

Current LLMs exhibit marked variability in medical question-answering benchmarks (e.g., MultiMedQA, MedMCQA): GPT-4 achieves 86.5% accuracy on MedQA (USMLE-style questions), while Gemini Pro scores 67.0%(11). DeepSeek, optimized for medical corpora, demonstrates superior contextual understanding in domain-specific tasks. However, these evaluations rely on standardized exam questions, lacking real-world testing on patient-centric queries.

This study evaluates ChatGPT-4, Gemini 1.5 Pro, and DeepSeek-R1 in three scenarios: 1.ctrAEs Management Accuracy: Compare model-generated interventions against guidelines using real-world trial cases; 2. Patient Query Response Quality: Simulate common patient questions (e.g., "eligibility criteria for lung cancer trials"), assessing accuracy, understandability, and risk disclosure completeness; 3.Dual-Perspective Satisfaction: Collect experienced clinical researchers utility ratings (Likert scale) and patient feedback on clarity/trustworthiness. As the first study to concurrently assess LLMs in specialist ctrAEs handling and patient education, our findings will inform model optimization for end-to-end clinical trial applications. This study has been reported in line with the STROCSS guidelines(12).

Methods

This cross- sectional study was exempt from institutional review board review given the lack of patient data. Available guidelines for the management of ctrAEs were reviewed. Based on these guidelines, a total of 62 questions were generated by the senior author and refined/approved by other study authors as representative common questions that arise in clinical settings. Five questions each from 7 common ctrAE categories were generated (gastrointestinal, Central Nervous System, Allergic Reactions, Respiratory, Cardiac, Liver Function, Kidney). The top Google (Alphabet, Inc) search queries related to clinical trial in the worldwide from 2004, to October 1, 2024, were identified using Google Trends. This publicly available web tool provides data on search-term volume for the Google search engine over time. The top 25 search queries related to clinical trial were identified and summarized into 9 general questions. All questions were designed as descriptive and open- ended in nature (supplemental file 1), but with clearly defined answers present in available guidelines from international committees with expertize in ctrAEs.

Finalized questions were entered into three chatbots (ChatGPT (V.GPT- 4o), Gemini 1.5 Pro, and DeepSeek-R1) by the first author on March 1, 2025. Answers were provided back to the rating physicians in blind (the physicians don't know which the answer was generated from). Rating physicians were all past or current initiators or main participants who have participated in two or more clinical trial studies. Prior to commencing the scoring process, all 8 participating experts attended a standardized training session. During this session, they received detailed scoring criteria and illustrative examples to ensure consistent evaluation practices. This preparatory phase aimed to align expert interpretations of quality metrics, mitigate subjective biases, and establish a shared framework for objective assessment. All answers were graded by each rater for accuracy and completeness for three chatbots. Accuracy was graded on a 4-point Likert scale, with 1 signifying completely inaccurate, 2 mostly inaccurate, 3 mostly accurate, and 4 accurate. Raters were instructed to grade accuracy results based on guideline content, not personal management style. Similarly, completeness was graded on a 4-point Likert scale, with 1 signifying incomplete, 2 missing multiple pieces of key information, 3 missing one piece of key information, and 4 complete. Raters were instructed to grade based on major pieces of key information rather than minor or optional items.

Grades were summarized using means, medians, and ranges for each chatbot overall and for each ctrAE category. Completeness and accuracy scores were compared between chatbots using the Kruskal-Wallis test, followed by Dunn's test for multiple comparisons. Inter-rater agreement was assessed with Kendall's coefficient of concordance, given the presence of more than two raters and non normal distribution ratings. SPSS Statistics 26 and GraphPad Prism 8 were used to analyze.

Results

All chatbots were evaluated for accuracy and completeness across 62 questions spanning nine categories (see Supplemental File 1). Overall performance was relatively high for all models: ChatGPT

achieved mean scores of 3.702 (accuracy) and 3.629 (completeness); Gemini scored 3.395 (accuracy) and 3.373(completeness); while DeepSeek attained the highest scores in both metrics (accuracy: mean=3.806; completeness: mean=3.764). Kendall's correlation coefficients for accuracy and completeness are presented in Table 1. Kendall’s coefficient is a conservative measure of agreement. While a value of 0.2-0.4 is often categorized as “fair”, it still indicates a statistically significant degree of association beyond chance. Our panel of raters was intentionally composed of experts from diverse sub-fields, which we believe enhances the validity of our study. However, this diversity naturally introduces a variety of viewpoints, which can depress measures of agreement like Kendall’s coefficient while capturing a richer, more comprehensive assessment. Statistical analysis revealed that DeepSeek's ratings were significantly higher than both GPT-4 and Gemini for both accuracy and completeness ($p<0.001$).

We input the top nine Google search queries related to clinical trials into three chatbots (ChatGPT, Gemini, and DeepSeek). Eight physician raters evaluated the responses, with a maximum of 36 ratings possible per category. DeepSeek demonstrated superior performance in completeness compared to both Gemini and ChatGPT, with mean and median (IQR) 34.75 and 35(34.25-35) (table 2). In terms of accuracy, DeepSeek achieved comparable results to GPT-4, with similarly median scores in the 34-35 range.

We analyzed scores stratified by category, pooling responses across five questions per category (maximum score: 20 per category). For both accuracy and completeness, mean and median scores across categories ranged from 18 to 20 for chatbots except Gemini (Table 2). Notably, DeepSeek outperformed Gemini in completeness scores for both gastrointestinal and allergic reaction categories. The median scores for DeepSeek were significantly higher than those for Gemini in both the Gastrointestinal Issues category (19 [IQR 18–20] vs. 17 [IQR 15.25–18]; $p = 0.0091$) and the Allergic Reaction category (19.5 [IQR 18.25–20] vs. 17.5 [IQR 16–19]; $p = 0.0316$). The Kruskal Wallis test showed significant differences among the three groups in cardiovascular issues for completeness ($p=0.045$) and kidney dysfunction for accuracy($p=0.039$) , but Dunn's post hoc test did not identify specific difference groups (adjusted $p>0.05$), possibly due to insufficient sample size or small inter group differences.

Across 18 patient-specific scenarios, DeepSeek demonstrated superior performance in both accuracy and completeness. Its median accuracy score was 66.5 (IQR 65.25-67.75; mean 66.25), significantly higher than that of GPT-4 ($p = 0.0431$) and Gemini ($p = 0.0007$). Similarly, its median completeness score was 67.5 (IQR 64.25-69; mean 66.75), which was also significantly higher than both GPT-4 ($p = 0.0475$) and Gemini ($p = 0.0037$). Meanwhile, GPT-4 outperformed Gemini, achieving median scores of 61.5 (IQR 59.25-64.5; mean 61.38) for accuracy and 62.5 (IQR 57.25-64.75; mean 61.5) for completeness.

Several questions received unanimous ratings of 4 from all eight reviewers. Specifically, DeepSeek achieved 39/62 (62.9%) for accuracy and 34/62 (54.8%) for completeness, whereas GPT-4 scored significantly lower, with 15/62 (24.2%) for accuracy and 2/62 (3.2%) for completeness ($p < 0.001$).

Scores indicating significant inaccuracy or incompleteness (1–2) were not found in DeepSeek, and exceptionally rare for ChatGPT, with only 6 out of 496 answer ratings (1.21%) falling into this category. High ratings (4) were rare for Gemini, appearing in only 3/62 (4.8%) of responses for accuracy and 2/62 (3.2%) for completeness.

Discussion

This study systematically evaluated the performance of three mainstream LLMs (ChatGPT, Gemini, and DeepSeek) in managing adverse drug reactions during clinical trials. The results demonstrate that DeepSeek exhibits superior accuracy and satisfaction scores compared to ChatGPT and Gemini, particularly in resolving both general inquiries and specialized clinical questions. The use of simulated scenarios and predefined questions in this study was a deliberate design choice to establish a controlled and reproducible benchmark for initial LLM performance. This approach allows for the isolation of model capabilities from the confounding variables inherent in messy, real-world clinical data, providing a clear proof-of-concept of their potential utility in an idealized setting.

While our analysis revealed a statistically significant difference in completeness scores between DeepSeek and GPT-4o, it is essential to consider whether this difference translates into a clinically meaningful improvement. Although the absolute difference in scores was modest, in the context of [e.g., gastrointestinal issues, allergic reactions], even a small increase in information completeness could be critical. For instance, the omission of a single key symptom or drug interaction could have substantial downstream consequences. Therefore, we posit that the observed improvement, while statistically modest, may hold practical value by reducing the risk of critical omissions. This discrepancy between the models may stem from the following mechanisms:

1.Complexity in Cross-System CtrAEs Management: CtrAEs involving respiratory, hematologic, and other systems in clinical trials are highly heterogeneous, requiring integration of pharmacological mechanisms and clinical expertise. DeepSeek exhibited stronger logical coherence in analyzing multi-system pathophysiological relationships. For example, it accurately distinguished molecular mechanisms between dose-dependent toxicities and immune-mediated hypersensitivity reactions and provided targeted management recommendations. In contrast, ChatGPT and Gemini often confused dose-dependent toxicities with immune-mediated allergies in complex cases, reducing the clinical applicability of their recommendations(13).

2. Deep Integration of Professional Medical Knowledge: DeepSeek's algorithmic architecture may more tightly integrate evidence-based medical databases and the latest clinical trial guidelines, enabling it to prioritize authoritative consensus when recommending management strategies(14). For instance, for chemotherapy-induced myelosuppression, DeepSeek not only suggested granulocyte colony-stimulating factor (G-CSF) use but also adjusted dosing based on patient-specific risk stratification (e.g., ECOG performance status). Other models frequently overlooked critical risk-benefit trade-offs in similar scenarios.

3. Balanced Dynamic Risk Assessment and Evidence-Based Decision-Making: In multi-system ctrAEs cases, DeepSeek demonstrated precise judgment in balancing "treatment urgency" and "evidence hierarchy." For example, when addressing immune checkpoint inhibitor-associated pneumonitis, its recommendations clearly differentiated between mild (monotherapy with steroids) and severe (combination therapy with intravenous immunoglobulin) treatment protocols, citing National Comprehensive Cancer Network (NCCN) guidelines to support tiered management. Comparatively, other models overrelied on single-study conclusions and lacked adaptability to evolving clinical practices.

4.Optimized User Interaction Design: The disparity in satisfaction ratings further reflects the clinical usability of response structures. DeepSeek adopted a three-part framework—"mechanistic analysis → priority ranking → management workflow"—aligning with clinicians' decision-making pathways. For example, in hepatotoxicity analysis, its outputs followed the logical sequence of "liver enzyme abnormalities → discontinuation criteria → alternative drug selection," whereas competing models often provided fragmented information, increasing cognitive load for clinical application(15).

5. Limitations and Future Directions: Despite DeepSeek's potential advantages, limitations persist: First, our study relied on simulated adverse event scenarios rather than authentic, historical AE reports. While this approach allowed for a controlled evaluation with, it may lack the complexity and noise inherent in real-world clinical data. Spontaneous real-world reports often contain misspellings, ambiguous abbreviations, and incomplete information, which could challenge the models differently. Consequently, the performance metrics reported here might represent a 'best-case scenario.' The generalizability of our findings to real-world pharmacovigilance workflows needs to be confirmed through future studies utilizing datasets of genuine AE reports. Second, our findings are inherently tied to the specific versions of the LLMs we evaluated. The field of AI is rapidly evolving, with model providers frequently releasing updates that can significantly alter performance, capabilities, and even output tendencies. Therefore, our results represent a snapshot in time. The performance metrics we reported may not be reproducible if the same study is conducted with future iterations of these models. This highlights a broader challenge in the evaluation of LLMs and underscores the need for continuous benchmarking and validation in applied settings like pharmacovigilance. Third, our analysis was constrained to a focused set of seven common and critical adverse event categories (from the ctrAE group). This narrow scope was a deliberate choice to enable a more in-depth analysis of model performance within these specific domains. However, it necessarily limits the broader applicability of our conclusions. The models' ability to handle other types of AEs, particularly rarer or more complex multi-system events, remains unexplored. Future research should aim to expand this scope to include a more comprehensive and diverse range of adverse events to fully assess the utility of LLMs in pharmacovigilance.

Conclusions

This study conducts the first systematic evaluation of three mainstream LLMs (DeepSeek, ChatGPT and Gemini) in managing ctrAEs. The results demonstrate that DeepSeek significantly outperforms other models in both accuracy and user satisfaction when addressing cross-disciplinary clinical questions,

confirming the unique value of domain-optimized models in supporting complex medical decision-making. These findings provide a feasible AI-assisted solution for specialist physicians managing ctrAEs outside their primary expertise. Additional research and validation are needed to explore their potential for enhancing clinical trial safety and reducing cross-disciplinary consultation costs.

Abbreviations

Artificial intelligence
AI
clinical trial-related adverse events
ctrAEs
large language models
LLMs
granulocyte colony-stimulating factor
G-CSF
National Comprehensive Cancer Network
NCCN
Interquartile Range
IQR

Declarations

Ethics approval and consent to participate: All methods in this study were carried out in accordance with the Declaration of Helsinki. The Institutional Review Board of Nanjing Jinling Hospital reviewed the study protocol and determined that it qualifies for an exemption from full ethical review, as it solely involved expert evaluators providing professional assessments. Nevertheless, informed consent was obtained from all expert evaluators prior to their participation, ensuring they were fully aware of the study's purpose and how their evaluations would be used.

Consent for publication Informed consent for publication was obtained from all expert evaluators participating in this study.

Availability of data and materials All data generated or analysed during this study are included in this published article [and its supplementary information files].

Competing interests: The authors declare that they have no competing interests.

Funding: Not applicable.

Author contributions: LQ, JPG, DF, AJ, SX, SZ, ZLZ, and BJL scored the answers. YKC, TLZ, XJS, JD, and LQ conceived and designed the study. XWT, TYS, FX, HT, HWH, YKC, and BF performed the literature search. GYZ, HG, YZG, NWY and ZCX collected the data. YKC, KXZ, DF, JC, and YYQ analyzed the data

and drafted the tables. YKC,XNW, QYZ and ZJW drafted the manuscript. ZCH, ŁZ, and LQ revised the manuscript. All authors read and approved the final manuscript.

Data availability: All data generated or analysed during this study are included in this published article [and its supplementary information files].

Acknowledgements: Not applicable.

References

1. Basile AO, Yahi A, Tatonetti NP. Artificial Intelligence for Drug Toxicity and Safety. Trends Pharmacol Sci. 2019;40(9):624-35.
2. Iskander R, Moyer H, Vigneault K, Mahmud SM, Kimmelman J. Survival Benefit Associated With Participation in Clinical Trials of Anticancer Drugs: A Systematic Review and Meta-Analysis. Jama. 2024;331(24):2105-13.
3. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
4. Ruan J, Su Q, Chen Z, Huang J, Li Y. CPRS: a clinical protocol recommendation system based on LLMs. Int J Med Inform. 2025;195:105746.
5. Liu XH, Lu ZH, Wang T, Liu F. Large language models facilitating modern molecular biology and novel drug development. Front Pharmacol. 2024;15:1458739.
6. Choi JY, Yoo TK. Development of a novel scoring system for glaucoma risk based on demographic and laboratory factors using ChatGPT-4. Med Biol Eng Comput. 2025;63(1):75-87.
7. Lai H, Ge L, Sun M, Pan B, Huang J, Hou L, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA Netw Open. 2024;7(5):e2412687.
8. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025;31(3):943-50.
9. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024;630(8017):625-30.
10. Verspoor K. 'Fighting fire with fire' - using LLMs to combat LLM hallucinations. Nature. 2024;630(8017):569-70.
11. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns (N Y). 2024;5(3):100943.
12. Agha RA, Mathew G, Rashid R, Kerwan A, Al-Jabir A, Sohrabi C, et al. Revised Strengthening the reporting of cohort, cross-sectional and case-control studies in surgery (STROCSS) Guideline: An update for the age of Artificial Intelligence. Prem J Sci. 2025;10:100081.
13. Chen YQ, Yu T, Song ZQ, Wang CY, Luo JT, Xiao Y, et al. Application of Large Language Models in Drug-Induced Osteotoxicity Prediction. J Chem Inf Model. 2025.

14. Peng Y, Malin BA, Rousseau JF, Wang Y, Xu Z, Xu X, et al. From GPT to DeepSeek: Significant gaps remain in realizing AI in healthcare. J Biomed Inform. 2025;163:104791.
15. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models. Int J Med Inform. 2025;198:105871.

Tables

Table 1. Kendall's correlation coefficients for accuracy and completeness

Kendall's correlation coefficients	GPT-4o	Gemini	DeepSeek-R1
Accuracy	0.278**	0.287**	0.203*
Completeness	0.211**	0.222**	0.271**

*p<0.05;**p<0.01

Table 2. Scores for accuracy and completeness for each engine in each category

Accuracy									Completeness							
	GPT-4o	Gemini	DeepSeek -R1	overall	total p	GPT vs. Gemini	GPT vs. DeepSeek	Gemini vs. DeepSeek		GPT-4o	Gemini	DeepSeek -R1	overall	total p	GPT vs. Gemini	GPT vs. DeepSeek	Gemini vs. DeepSeek
General					0.0128*	0.9088	0.2717	0.0101*	General					0.0032*	0.99	0.0372*	0.0037*
mean(SD)	33.38(2.72)	32.88(1.126)	35.25(0.71)	33.83(1.97)					mean(SD)	32.13 (2.167)	31.13 (1.885)	34.75(0.886)	32.67 (2.278)				
Median(IQR)	34(31.5-35.75)	32.5(32-33.75)	35(35-36)	34(32.25-35)					Median(IQR)	31.5 (31-33.75)	32 (29.25-32.75)	35 (34.25-35)	33 (31-35)				
Gastrointestinal Issues					0.0909	0.2887	0.99	0.1166	Gastrointestinal Issues					0.0124*	0.4155	0.4155	0.0091*
mean(SD)	18.75 (0.89)	17.75 (1.28)	19(0.93)	18.5 (1.14)					mean(SD)	18.13 (0.99)	16.88 (1.46)	19(0.93)	18 (1.41)				
Median(IQR)	19 (18.25-19)	17.5 (17-18.75)	19(18-20)	19 (18-19)					Median(IQR)	18 (17.25-18.75)	17 (15.25-18)	19(18-20)	18 (17-19)				
Central Nervous System					0.0681	0.5336	0.9999	0.0632	Central Nervous System					0.2111	0.9999	0.5887	0.2756
mean(SD)	18.88 (0.99)	18 (1.31)	19.38(0.74)	18.75 (1.15)					mean(SD)	18.5 (1.60)	18.25(1.28)	19.25(1.75)	18.67(1.15)				
Median(IQR)	19(18.25-19.75)	18 (17-19)	19.5(19-20)	19 (18-20)					Median(IQR)	19(17-20)	18(17-19.75)	20(19.25-20)	19(17-20)				
Allergic Reactions					0.0764	0.9687	0.6084	0.0711	Allergic Reactions					0.0292*	0.9999	0.1704	0.0316*
mean(SD)	19.13(0.83)	17.38(2.45)	19.75(0.46)	18.75(1.78)					mean(SD)	18.13(0.99)	17.38(1.60)	19.25(0.89)	18.25(1.39)				
Median(IQR)	19(18.25-20)	17(15.25-20)	20(19.25-20)	19.5(18-20)					Median(IQR)	18(17.25-18.75)	17.5(16-19)	19.5(18.25-20)	18(17.25-19)				
Respiratory Issues					0.081	0.2262	0.9999	0.115	Respiratory Issues					0.0811	0.6572	0.9391	0.0757
mean(SD)	19(0.93)	17.63 (1.60)	19.13(1.13)	18.58 (1.38)					mean(SD)	18.63 (1.06)	17.75 (1.49)	19.13(0.64)	18.5 (1.22)				
Median(IQR)	19 (18-20)	18 (16.25-18.75)	19.5(18.25-20)	19 (18-20)					Median(IQR)	18.5 (18-19.75)	18 (17-18.75)	19(19-19.75)	19 (18-19)				
Cardiovascular Issues					0.1207	0.5347	0.9999	0.1304	Cardiovascular Issues					0.0453*	0.1218	0.9999	0.0739
mean(SD)	18.88 (0.99)	17.75 (1.75)	19.25(0.89)	18.63 (1.38)					mean(SD)	18.75 (1.49)	17.38 (1.06)	18.88(0.83)	18.33(1.31)				
Median(IQR)	18.5 (18-20)	18 (16.25-19.5)	19.5(18.25-20)	18.5 (18-20)					Median(IQR)	19.5(17-20)	17.5(16.25-18)	19(18-19.75)	18(17-19.75)				
Liver Function Abnormalities					0.067	0.1701	0.9999	0.1053	Liver Function Abnormalities					0.3046	0.9999	0.9774	0.3853
mean(SD)	18.75(1.04)	17.75(0.71)	18.88(1.13)	18.46(1.06)					mean(SD)	18.13(0.83)	17.63(1.30)	18.63(1.19)	18.13(1.15)				
Median(IQR)	19 (18-19.75)	18 (17-18)	19(18-20)	18 (18-20)					Median(IQR)	18(18-18)	18(17-18.75)	19(17.25-19.75)	18(17.25-19)				
Kidney Dysfunction					0.0329*	0.0744	0.9999	0.0676	Kidney Dysfunction					0.0571	0.164	0.9999	0.0843
mean(SD)	19.25(0.71)	17.88(1.25)	19.25(0.89)	18.79(1.14)					mean(SD)	19(1.07)	17.63(1.51)	19.13(1.13)	18.58(1.38)				
Median(IQR)	19(19-20)	18(17-18.75)	19.5(18.25-20)	19(18-20)					Median(IQR)	19(18.25-20)	18(16.25-19)	19.5(18.25-20)	19(18-20)				
Patient Specific Scenario					0.0009*	0.6628	0.0431*	0.0007*	Patient Specific Scenario					0.0036*	0.9999	0.0475*	0.0037*
mean(SD)	61.38(3.503)	58.88(2.748)	66.25(2.121)	62.17(4.146)					mean(SD)	61.5(4.036)	59.75(4.83)	66.75(2.435)	62.67(4.806)				
Median(IQR)	61.5(59.25-64.5)	59.5(56.5-61)	66.5(65.25-67.75)	62(59.25-66)					Median(IQR)	62.5(57.25-64.75)	61(54.75-62.75)	67.5(64.25-69)	63(61-66)				
*p<0.05																	

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementalfile.pdf