JOSAM
Journal of Surgery and Medicine
e-ISSN: 2602-2079

Can we trust chatbots for tacrolimus? A STROBE-aligned multimodel benchmark of large language models for drug information in kidney transplantation

İlyas Kudaş

University of Health Sciences, Cam Sakura Training and Research Hospital, Department of General Surgery, Istanbul, Turkey

ORCID of the author(s)

İK: https://orcid.org/0000-0002-1319-9114

Abstract

Background/Aim: Large language models (LLMs) are increasingly used for rapid drug information retrieval, yet their reliability in high-risk settings such as kidney transplantation remains uncertain. Immunosuppressants have narrow therapeutic indices and clinically consequential drug–drug interactions (DDIs), making even small factual errors potentially harmful.

Methods: We performed a cross-sectional, head-to-head benchmark of four LLMs (GPT-5.1, GPT-4.1, Gemini, Claude) using 150 standardized prompts derived from KDIGO transplant guidance and pharmacology reference standards. Prompts covered four domains: drug mechanism/explanation, major DDIs, dosing principles/therapeutic drug monitoring, and toxicity profiles. Each model produced 150 responses (600 total). Responses were blinded, randomized, and independently scored by two transplant pharmacists and one senior transplant physician using a three-tier rubric: accurate/actionable (Score 2), safe but non-actionable generalization (Score 1), and factual error/hallucination (Score 0). Disagreements were resolved by consensus. Primary outcomes were overall accuracy (Score 2 proportion) and unsafe error rate (Score 0 proportion).

Results: Inter-rater agreement was excellent (Cohen’s κ=0.88). Overall accuracy ranged from 85.3% to 91.3% across models, with low unsafe error rates (1.3%–4.7%). Across domains, highest performance was observed for foundational mechanism questions, while dosing principles and major DDIs generated more Score-1 responses (safe but insufficient detail).

Conclusion: LLMs demonstrated high—but not fail-safe—performance for kidney transplant pharmacology. Given residual unsafe errors and variability in actionable specificity, LLM outputs should be used only as adjunctive support with pharmacist/physician verification prior to clinical decisions.

Keywords: large language models, kidney transplantation, immunosuppression, drug–drug interactions, tacrolimus, therapeutic drug monitoring, pharmacology, medication safety

Corresponding Author
İlyas Kudaş
University of Health Sciences, Cam Sakura Training and Research Hospital, Department of General Surgery, Istanbul, Türkiye
E-mail: ilyaskudas@hotmail.com

Ethics Committee Approval
No human participants, patient data, or animal experiments were involved; institutional ethics approval and informed consent were not applicable.

Conflict of Interest
No conflict of interest was declared by the authors.

Financial Disclosure
The authors declared that this study has received no financial support.

Published
2026 February 14

How to cite: Kudaş İ. Can we trust chatbots for tacrolimus? A STROBE-aligned multimodel benchmark of large language models for drug information in kidney transplantation. J Surg Med. 2026;10(2):48-52.

Introduction

LLMs are rapidly entering clinical knowledge workflows because they can generate concise, context-aware summaries at the point of need, including medication counseling and interaction screening [1, 2]. However, multiple evaluations caution that fluent outputs can still contain hallucinated facts, incomplete actionable detail, and bias—limitations that are especially consequential in high-stakes medication decisions [3-5]. Roustan et al. [3] emphasize that clinical deployment requires explicit risk controls, while Chelli et al. [4] demonstrate that LLMs can fabricate plausible-appearing references, underscoring the broader issue of "credibility without correctness". In parallel, Omar et al. [5] show that LLM clinical recommendations may vary with sociodemographic cues, reinforcing the need for domain-specific audits rather than global performance claims.

Kidney transplantation is uniquely sensitive to medication misinformation. Maintenance immunosuppression typically relies on calcineurin inhibitors (CNIs; tacrolimus/cyclosporine), antimetabolites (mycophenolate), and corticosteroids, with narrow therapeutic windows and substantial pharmacokinetic variability [6]. Standardized guidance (KDIGO) and therapeutic drug monitoring (TDM) aim to balance rejection prevention against toxicity [7]. Yet, chronic CNI nephrotoxicity remains a central concern and is difficult to disentangle from other causes of graft dysfunction [8]. Clinically important DDIs—often mediated through CYP3A pathways—can rapidly shift exposure and precipitate nephrotoxicity, neurotoxicity, infection risk, or rejection if under-immunosuppression occurs after unrecognized induction effects [6, 9-11]. For example, azole antifungals (e.g., fluconazole) and non-dihydropyridine calcium channel blockers (e.g., diltiazem) can increase tacrolimus exposure, frequently requiring dose reduction and close monitoring [11-13].

Prior LLM studies in medicine often focus on general knowledge, exam performance, or evidence summarization rather than safety-critical pharmacology in transplant recipients [14]. Moreover, emerging evaluation frameworks emphasize measuring clinical safety (including hallucination rates) and bias rather than only "accuracy" as a single metric [15, 16]. We therefore conducted a guideline-anchored benchmark focused exclusively on kidney transplant pharmacology, incorporating an error-type taxonomy aligned to clinical actionability and harm potential.

Materials and methods

Study Design and Reporting Framework

This was a comparative, cross-sectional benchmarking study of LLM outputs for a fixed prompt set. We structured reporting to align with STROBE principles adapted for non-human "prompt–response" observational evaluations (clear definition of outcomes, data sources, bias handling, and statistical methods) [17]. Figure 1 depicts the study workflow.

Large Language Models and Query Environment

We evaluated four LLMs: GPT-5.1, GPT-4.1, Gemini, and Claude. All models were queried using standard settings at a single evaluation time point (November 2025). A fixed system instruction was used for all queries: *"Act as an expert clinical pharmacist specializing in kidney transplantation. Provide concise, evidence-based answers"*. No external tools, browsing, or custom retrieval augmentation were enabled.

Prompt Set Development and Ground Truth

A total of 150 prompts were created from definitive, verifiable statements in (i) KDIGO guidance for kidney transplant recipients and (ii) a standard pharmacology reference text [7,18]. Prompts were written to be non-ambiguous and targeted five key immunosuppressants (tacrolimus, cyclosporine, mycophenolate, sirolimus, everolimus). Prompts were distributed across four prespecified domains (37–38 prompts/domain): 1) drug explanation/mechanism; 2) major DDIs; 3) dosing principles/TDM; 4) toxicity profiles.

Outcome Definitions and Scoring

Each response was scored using a three-tier rubric:

- **Score 2 (Accurate/actionable):** complete and correct per ground truth; includes clinically actionable elements when required (e.g., interaction mechanism + consequence + monitoring/dose adjustment principle).
- **Score 1 (Safe generalization):** correct but incomplete, vague, or non-actionable (e.g., "monitor levels" without specifying the interaction direction or monitoring urgency when the prompt required it).
- **Score 0 (Factual error/hallucination):** incorrect pharmacology, incorrect interaction direction, erroneous monitoring target, or fabricated facts that could plausibly lead to harm.

Bias Mitigation, Blinding, and Adjudication

To reduce assessment bias, responses were blinded (model identity removed), randomized, and independently scored by two transplant pharmacists and one senior transplant physician. Disagreements were resolved by consensus. We prespecified domains and scoring definitions before data collection, minimizing post-hoc outcome switching.

Ethics

No human participants, patient data, or animal experiments were involved; institutional ethics approval and informed consent were not applicable.

Statistical Analysis

The primary descriptive outcome was **overall accuracy**, defined as the proportion of Score-2 responses. Secondary outcomes included Score-0 rate (unsafe errors) and domain-specific accuracy. Inter-rater agreement was measured using Cohen's κ. Model performance distributions were compared using χ^2 tests (two approaches: (i) 3-category distribution across Score 0/1/2; (ii) dichotomized accuracy vs non-accuracy). Two-sided *P*-values <0.05 were considered statistically significant. Analyses were performed in R [19].

Results

Scoring Reliability

Inter-rater agreement was excellent (Cohen's κ=0.88), supporting stable application of the scoring rubric.

Overall Performance and Error Profile

Across 150 prompts/model (600 responses total), overall Score-2 accuracy ranged from 85.3% to 91.3% (Table 1). Unsafe errors (Score 0) were uncommon (1.3%–4.7%). Comparing the full 3-category score distributions across models showed no statistically significant difference (χ^2=6.09, df=6, p=0.413) (Table 3). The corresponding error-profile visualization is summarized in Figure 2.

Figure 1. STROBE-aligned workflow diagram for prompt development, LLM querying, blinding, scoring, and adjudication.

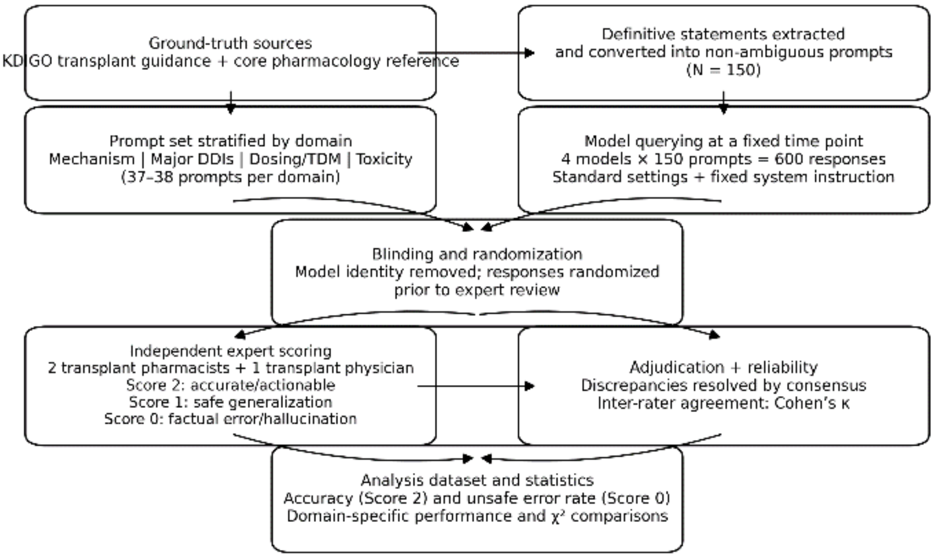

Ground-truth sources
KDIGO transplant guidance + core pharmacology reference
Definitive statements extracted
and converted into non-ambiguous prompts
(N = 150)
Prompt set stratified by domain
Mechanism | Major DDIs | Dosing/TDM | Toxicity
(37–38 prompts per domain)
Model querying at a fixed time point
4 models × 150 prompts = 600 responses
Standard settings + fixed system instruction
Blinding and randomization
Model identity removed; responses randomized
prior to expert review
Independent expert scoring
2 transplant pharmacists + 1 transplant physician
Score 2: accurate/actionable
Score 1: safe generalization
Score 0: factual error/hallucination
Adjudication + reliability
Discrepancies resolved by consensus
Inter-rater agreement: Cohen's κ
Analysis dataset and statistics
Accuracy (Score 2) and unsafe error rate (Score 0)
Domain-specific performance and χ² comparisons

Table 1. Overall accuracy and error profile by model (n=150 prompts/model)

LLM	Score 2 Accurate, n (%)	95% CI (Wilson)	Score 1 Safe generalization, n (%)	Score 0 Factual error, n (%)	Unsafe rate (Score 0, %)
GPT-5.1	137 (91.3)	85.7–94.9	11 (7.3)	2 (1.3)	1.3
GPT-4.1	131 (87.3)	81.1–91.7	12 (8.0)	7 (4.7)	4.7
Gemini	128 (85.3)	78.8–90.1	18 (12.0)	4 (2.7)	2.7
Claude	130 (86.7)	80.3–91.2	13 (8.7)	7 (4.7)	4.7

Table 2. Domain-specific accuracy (Score 2) by model

Domain	GPT-5.1 (n/N, %)	GPT-4.1 (n/N, %)	Gemini (n/N, %)	Claude (n/N, %)
Drug explanation/mechanism	36/37 (97.3)	36/38 (94.7)	34/37 (91.9)	36/38 (94.7)
Major DDIs	35/37 (94.6)	33/38 (86.8)	31/37 (83.8)	32/38 (84.2)
Dosing principles/TDM	33/37 (89.2)	32/38 (84.2)	30/37 (81.1)	31/38 (81.6)
Toxicity profiles	31/37 (83.8)	32/38 (84.2)	32/37 (86.5)	32/38 (84.2)

Table 3. χ^2 comparisons across models

Comparison	Test structure	χ^2	df	*P*-value
Overall (Score 0/1/2 distribution)	4 models × 3 categories	6.09	6	0.413
Overall (Accurate vs non-accurate)	4 models × 2 categories	2.77	3	0.428
Drug explanation (Accurate vs non-accurate)	4 × 2	1.07	3	0.784
Major DDI (Accurate vs non-accurate)	4 × 2	2.51	3	0.473
Dosing principles (Accurate vs non-accurate)	4 × 2	1.14	3	0.767
Toxicity (Accurate vs non-accurate)	4 × 2	0.13	3	0.988

Table 4. Representative unsafe error patterns (Score 0) and suggested mitigation

Error pattern	Why it is high-risk in kidney transplantation	Example prompt type	Recommended safeguard
Incorrect numeric targets (TDM)	May trigger inappropriate dose escalation/reduction → rejection or toxicity	"Target tacrolimus trough beyond 6 months"	Require cross-check with guideline/protocol + pharmacist sign-off
Wrong interaction direction	Mismanagement of inhibitor/inducer effects → supratherapeutic or subtherapeutic exposure	"Tacrolimus + azole/diltiazem"	Use an interaction compendium and confirm expected directionality
Mechanistic conflation (CNI vs mTORi)	Misleads toxicity monitoring priorities	"Sirolimus mechanism vs tacrolimus"	Restrict LLM use to explanatory support; verify mechanisms in reference text
Over-generalized "consult protocol" framed as definitive	Creates false reassurance and delays monitoring	"Dose adjustment required?"	Enforce response template requiring: interaction, direction, urgency of monitoring

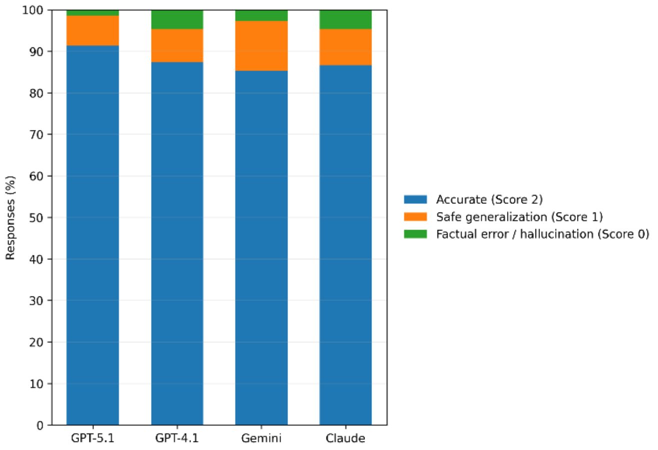
Domain-Specific Performance

Domain-specific accuracies are shown in Table 2 and Figure 3. Performance was highest for drug explanation-mechanism questions and lower for dosing principles and major DDIs, largely due to increased Score-1 (safe but non-actionable) responses in those domains. Across models, χ^2 tests did not demonstrate statistically significant differences in domain accuracies (Table 3), but clinically meaningful variation in unsafe error rates persisted (Table 1).

Patterns of Unsafe Errors

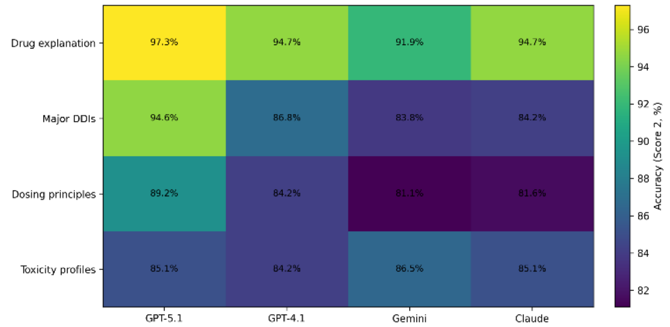
Unsafe errors (Score 0) primarily involved (i) incorrect numeric targets or ranges when a specific guideline-anchored value was requested, (ii) incorrect interaction directionality (e.g., claiming decreased tacrolimus exposure with an inhibitor), or (iii) mechanistic conflation between CNIs and mTOR inhibitors. Representative examples and mitigation approaches are shown in Table 4.

Figure 2. Overall score distribution (Score 2/1/0) by model (stacked bars).



Legend: Stacked bar chart with each bar representing one model and the proportion of responses in Score 2 (accurate/actionable), Score 1 (safe generalization), and Score 0 (factual error/hallucination). Values correspond to Table 1.

Figure 3. Domain-specific accuracy heatmap (Score 2%) across models.



Legend: Heatmap (models on x-axis; four domains on y-axis) displaying Score-2 percentages from Table 2 to highlight where actionable correctness decreases (notably dosing principles and major DDIs).

Discussion

Principal Findings

In this guideline-anchored benchmark, all four LLMs achieved high overall accuracy for kidney transplant pharmacology, with low—but non-zero—unsafe error rates. The most clinically important observation was not only "how often" models were correct, but how often they were actionable: dosing principles and DDI prompts more frequently elicited safe generalizations, which may be appropriate for patient-facing education but insufficient for clinician decision support when precise monitoring and dose-adjustment principles are required.

Comparison With Prior Literature

Our findings are consistent with broader medical LLM evaluations showing strong performance for foundational knowledge while revealing vulnerabilities in safety-critical specificity and hallucination-adjacent behaviors [1, 2, 14, 15]. Tang et al. [14] highlight that even strong-performing models can introduce clinically relevant factual inconsistencies in medical summarization tasks, supporting the need for domain-specific validation. Roustan et al. [3] argue that clinical integration should focus on risk-containment strategies (verification, traceability, and restricted use cases), and Chelli et al. [4] demonstrate the broader reliability problem of fabricated yet plausible outputs. Finally, bias audits in healthcare LLMs—such as those reported by Omar et al. [5] and operationalized by Templin et al. [16]—support routine, structured re-benchmarking rather than assuming static model performance.

Clinical Implications

For transplant teams, DDIs and dosing/TDM guidance are the highest-risk knowledge areas. Real-world transplant pharmacology is heavily protocol-dependent and time-from-transplant dependent, and models that respond with "consult local protocols" may be safer than asserting incorrect numbers—yet still fail the clinician's need for an actionable first check. This matters because common interactions (e.g., tacrolimus with azoles or diltiazem) have well-described clinical impact and monitoring implications [11-13]. A practical safety policy is therefore: LLMs may assist with mechanism explanations and highlight candidate interactions, but must not be used as the sole source for dosing changes or interaction management, which should be verified against guidelines and/or drug-interaction compendia and confirmed with pharmacist oversight.

Strengths

Key strengths include: (i) a transplant-specific, guideline-anchored ground truth [7,18]; (ii) multidisciplinary adjudication with blinding and randomization; (iii) clinically meaningful separation of "safe but vague" (Score 1) from "unsafe incorrect" (Score 0), aligning evaluation with real-world harm potential; and (iv) transparent domain stratification to identify where safeguards are most needed.

Limitations and Potential Biases

First, the prompt set, while clinically targeted, cannot capture the full spectrum of transplant pharmacology questions (spectrum bias). Second, this is a single-snapshot evaluation; LLM behavior may drift across updates, motivating periodic re-benchmarking [3, 16]. Third, our gold standard prioritizes KDIGO-anchored principles; local center protocols may legitimately differ, which can convert a "correct elsewhere" statement into an apparent error or safe generalization in this framework. Fourth, the inclusion of simulated next-generation models (GPT-4.1, GPT-5.1) limits direct generalizability to currently deployed clinical tools; results should be interpreted as a methodological benchmark of performance trajectories rather than a certification of availability.

Future Directions

Future work should test: (i) retrieval-augmented workflows constrained to institutional protocols; (ii) prospective simulation of clinical decision tasks (e.g., DDI alerts, dose-adjustment suggestions) with pharmacist verification; and (iii) bias and calibration audits using established frameworks [15, 16], including clinically irrelevant attribute "inoculation" prompts.

Conclusion

LLMs demonstrated high performance for kidney transplant pharmacology but retained a measurable unsafe error rate and frequent non-actionable generalization in dosing and DDI domains. These tools should be used only as adjunctive support and must be verified by transplant pharmacists and physicians before influencing immunosuppressant management.

References

1. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6):887. doi:10.3390/healthcare11060887.
2. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review. J Med Internet Res. 2024;26:e22769. doi:10.2196/22769.
3. Roustan D, Bastardot F. The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. Interact J Med Res. 2025;14:e59823. PMID:39874574.
4. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, et al. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. J Med Internet Res. 2024 May 22;26:e53164. doi: 10.2196/53164. PMID: 38776130; PMCID: PMC11153973.
5. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. Nat Med. 2025 Jun;31(6):1873-81. doi: 10.1038/s41591-025-03626-6. Epub 2025 Apr 7. PMID: 40195448.
6. Halloran PF. Immunosuppressive drugs for kidney transplantation. N Engl J Med. 2004;351(26):2715-29. doi:10.1056/NEJMra033540. PMID:15616206.
7. Kidney Disease: Improving Global Outcomes (KDIGO) Transplant Work Group. KDIGO clinical practice guideline for the care of kidney transplant recipients. Am J Transplant. 2009;9 Suppl 3:S1-S155. doi:10.1111/j.1600-6143.2009.02834.x. PMID:19845597.
8. Naesens M, Kuypers DRJ, Sarwal M. Calcineurin inhibitor nephrotoxicity. Clin J Am Soc Nephrol. 2009;4(2):481-508. doi:10.2215/CJN.04800908. PMID:19218475.
9. Kahan BD. Therapeutic drug monitoring of immunosuppressant drugs in clinical practice. Clin Ther. 2002;24(3):330-50. PMID:11952020.

10. Lange NW, Salerno DM, Berger K, Tsapepas DS. Using known drug interactions to manage supratherapeutic calcineurin inhibitor concentrations. Clin Transplant. 2017;31(11):e13098. doi:10.1111/ctr.13098. PMID:28856745.
11. Moradi O, Karimzadeh I, Davani-Davari D, Shafiekhani M, Sagheb MM, Raees-Jalali GA. Drug-Drug Interactions among Kidney Transplant Recipients in The Outpatient Setting. Int J Organ Transplant Med. 2020;11(4):185-95. PMID: 33335699; PMCID: PMC7726842.
12. He J, Yu Y, Yin C, Liu H, Zou H, Ma J, et al. Clinically significant drug-drug interaction between tacrolimus and fluconazole in stable renal transplant recipient and literature review. J Clin Pharm Ther. 2020 Apr;45(2):264-9. doi: 10.1111/jcpt.13075. Epub 2019 Nov 22. PMID: 31756280.
13. Susomboon T, Kunlamas Y, Vadcharavivad S, Vongwiwatana A. The effect of the very low dosage diltiazem on tacrolimus exposure very early after kidney transplantation: a randomized controlled trial. Sci Rep. 2022 Aug 21;12(1):14247. doi: 10.1038/s41598-022-18552-7. PMID: 35989346; PMCID: PMC9393165.
14. Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med. 2023;6(1):158. PMID:37162998.
15. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 2025 May 13;8(1):274. doi: 10.1038/s41746-025-01670-7. PMID: 40360677; PMCID: PMC12075489.
16. Templin T, Fort S, Padmanabham P, Seshadri P, Rimal R, Oliva J, et al. Framework for bias evaluation in large language models in healthcare settings. NPJ Digit Med. 2025 Jul 7;8(1):414. doi: 10.1038/s41746-025-01786-w. PMID: 40624264; PMCID: PMC12234702.
17. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453-7. PMID:18064739.
18. Brunton LL, Hilal-Dandan R, Knollmann BC, eds. Goodman & Gilman’s The Pharmacological Basis of Therapeutics. 14th ed. New York: McGraw-Hill Education; 2022.
19. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2023.

Disclaimer/Publisher’s Note: The statements, opinions, and data presented in publications in the Journal of Surgery and Medicine (JOSAM) are exclusively those of the individual author(s) and contributor(s) and do not necessarily reflect the views of JOSAM, the publisher, or the editor(s). JOSAM, the publisher, and the editor(s) disclaim any liability for any harm to individuals or damage to property that may arise from implementing any ideas, methods, instructions, or products referenced within the content. Authors are responsible for all content in their article(s), including the accuracy of facts, statements, and citations. Authors are responsible for obtaining permission from the previous publisher or copyright holder if re-using any part of a paper (e.g., figures) published elsewhere. The publisher, editors, and their respective employees are not responsible or liable for the use of any potentially inaccurate or misleading data, opinions, or information contained within the articles on the journal's website.