Challenge Journal of Perioperative Medicine

Research Article

Evaluation of the content quality of regional anesthesia and postoperative analgesia approaches generated by ChatGPT-4.0 according to surgical incision sites

Müzeyyen Beldağlı [a,*]

[a] *Department of Anesthesiology and Reanimation, Samsun Training and Research Hospital, Samsun, Türkiye*

ABSTRACT

Background: Large language models (LLMs) are increasingly consulted for perioperative decision support, yet their ability to give professional-grade guidance for regional anesthesia and analgesia remains uncertain.

Materials and Methods: In a prospective observational study, we presented eight incision-based figures (Items 2–9) representing common abdominal incisions to ChatGPT-4.0 and requested a regional anesthesia technique and postoperative analgesia plan for each. Five independent anesthesiologists rated each response on Accuracy, Comprehensiveness, and Safety using a 5-point Likert scale. Inter-rater reliability was summarized with Fleiss' κ. One non-incision item (Item 10) was analyzed descriptively and excluded from pooled statistics. Single-shot prompts were used.

Results: Mean ratings were high: Accuracy 4.28, Comprehensiveness 4.30, Safety 4.00 (1–5 scale). Inter-rater agreement was substantial for Safety (κ=0.76) and lower for Accuracy (κ=0.33) and Comprehensiveness (κ=0.31). Two consistent low points emerged: right-lower-quadrant (McBurney/Lanz) incision–Safety mean 3.0 and suprapubic (Pfannenstiel) incision–Accuracy 3.0; Comprehensiveness 3.4; Safety 3.4. When explicitly asked for postoperative plans, the model rarely proposed neuraxial techniques (e.g., epidural), favoring fascial-plane/peripheral strategies.

Conclusions: An LLM produced clinically usable suggestions for common abdominal incisions with strong safety agreement, but performance was not uniform, and neuraxial options were under-recommended. These tools may serve as a helpful adjunct for education and option-generation, yet they should be used with expert oversight and local protocols. Future work should test repeated sampling, prompt standardization, model/tier comparisons, and link recommendations to patient outcomes.

ARTICLE INFO

Article history:
Received – August 5, 2025
Revision requested – September 14, 2025
Revision received – September 26, 2025
Accepted – October 16, 2025

Keywords:
postoperative analgesia
regional analgesia
large language models
ChatGPT 4.0

Citation: Beldağlı M. Evaluation of the content quality of regional anesthesia and postoperative analgesia approaches generated by ChatGPT-4.0 according to surgical incision sites. *Chall J Perioper Med.* **2025**; 3(3):100–105.

1. Introduction

Regional anesthesia has undergone a major transformation, particularly since the integration of ultrasound into anesthetic practice, shifting its focus from neuraxial to peripheral techniques [1,2]. However, many anesthesiologists–especially those primarily engaged in clinical practice–have lacked the advanced anatomical knowledge required to keep pace with this evolution. New techniques are continuously being described, and even the most recent editions of textbooks often fall short of encompassing the rapidly expanding and dynamic body of knowledge in this field [3].

Clinicians and trainees in anesthesiology are constantly seeking educational resources that are accessible, easy to understand, and time-efficient. For a period,

* *Corresponding author*. E-mail address: mbayram_88@hotmail.com (M. Beldağlı)
ISSN: 2980-292X / DOI: https://doi.org/10.20528/cjpm.2025.03.005

YouTube videos seemed to meet this need; however, this approach also produced inconsistent and sometimes unsafe outcomes, as most videos were not created by professionals nor subjected to any form of oversight [4]. Today, large language models (LLMs) such as ChatGPT have begun to fill a similar role, offering rapid, on-demand information to eager learners. Yet, the same concerns regarding accuracy, reliability, and content control inevitably arise with these new tools as well [5].

This study aimed to assess how well ChatGPT-4.0, one of the most commonly used free large language models, provides accurate and practical information about regional anesthesia and postoperative analgesia in various surgical settings. To introduce a clinically grounded perspective, we applied an incision-based evaluation framework that allows model performance to be assessed within realistic surgical contexts rather than through generic question sets. As these techniques continue to expand and change, many clinicians now look to such tools for quick guidance and learning support. Our goal was to see if ChatGPT-4.0 could offer explanations that feel trustworthy, complete, and clinically useful.

2. Materials and Methods

2.1. Study design

This prospective observational study was designed to evaluate the validity of data generated by an artificial intelligence tool in the clinical context. Ethical approval was obtained from the Samsun University Non-Interventional Clinical Research Ethics Committee (Decision No: GOKAEK 2025/11/7). The study was conducted in line with the principles of the Declaration of Helsinki. The main purpose was to assess the quality of regional anesthesia and postoperative analgesia suggestions produced by ChatGPT-4.0 based on visual representations of surgical incisions. No patient data or direct clinical intervention was involved.

2.2. Evaluator selection

Twenty anesthesiologists experienced in fascial plane blocks and active in national and international regional anesthesia meetings were invited by email. Written consent was obtained from those who agreed, and five anesthesiologists participated as evaluators. All evaluators were senior anesthesiologists recognized for their expertise in regional anesthesia, each having conducted training sessions, contributed to international publications, and held active roles within the National Society of Regional Anesthesia. To avoid bias, evaluators were not included as study authors and did not take part in data analysis. All evaluations were performed independently, and their identities were kept confidential. Evaluators were blinded to each other's assessments, and no formal calibration session was conducted; however, a brief orientation was provided to ensure consistent understanding of the scoring criteria.

2.3. Preparation of figures and ChatGPT-4.0 responses

We prepared several visuals representing different surgical incision sites to simulate common surgical approaches (Fig. 1). Each image was presented to ChatGPT-4.0, asking for a suitable regional anesthesia method and postoperative analgesia recommendation. The model's written responses were saved without any editing or modification.

2.4. Incision items presented to ChatGPT

- Item 2: Median midline laparotomy (vertical midline).
- Item 3: Left lower quadrant incision.
- Item 4: Right lower quadrant (McBurney/Lanz region).
- Item 5: Right upper quadrant (Kocher-type) incision.
- Item 6: Suprapubic (Pfannenstiel) incision.
- Item 7: Suprapubic (Pfannenstiel)–second scenario.
- Item 8: Paramedian vertical incision (just off the midline).
- Item 9: Transverse midline laparotomy.

2.5. Evaluation of the responses

Each response was independently scored by the five evaluators under three criteria:

1. Scientific accuracy
2. Comprehensiveness
3. Patient safety

A five-point Likert scale (1 = very poor, 5 = excellent) was used. Evaluators rated the content separately to maintain objectivity and prevent mutual influence.

2.6. Statistical analysis

Descriptive statistics and data handling were performed in IBM SPSS Statistics v22; inter-rater agreement (Fleiss' κ with $\bar{P}$ and P_e) and all figures were generated with custom Python scripts, and the heatmap layout/caption was prepared with assistance from ChatGPT. The primary analysis was restricted to incision-based items (Items 2–9). Five anesthesiologists rated each item on Accuracy, Comprehensiveness, and Safety using a 5-point Likert scale (1–5); item×criterion results are summarized as mean±SD, with overall item means and rater-level means also reported. Inter-rater reliability was estimated per criterion across Items 2–9 using Fleiss' κ with five categories, treating Likert responses as ordinal for κ and approximately interval for means/SDs. Item 10 (non-incision) was analyzed descriptively and excluded from pooled estimates; a sensitivity check including Item 10 produced only modest shifts and did not change interpretation. No formal hypothesis tests were prespecified; where applicable, p-values would be two-sided (α=0.05). κ values are interpreted by conventional thresholds (≤0.20 poor; 0.21–0.40 fair; 0.41–0.60 moderate; 0.61–0.80 substantial; ≥0.81 almost perfect) and emphasized for magnitude rather than statistical significance.

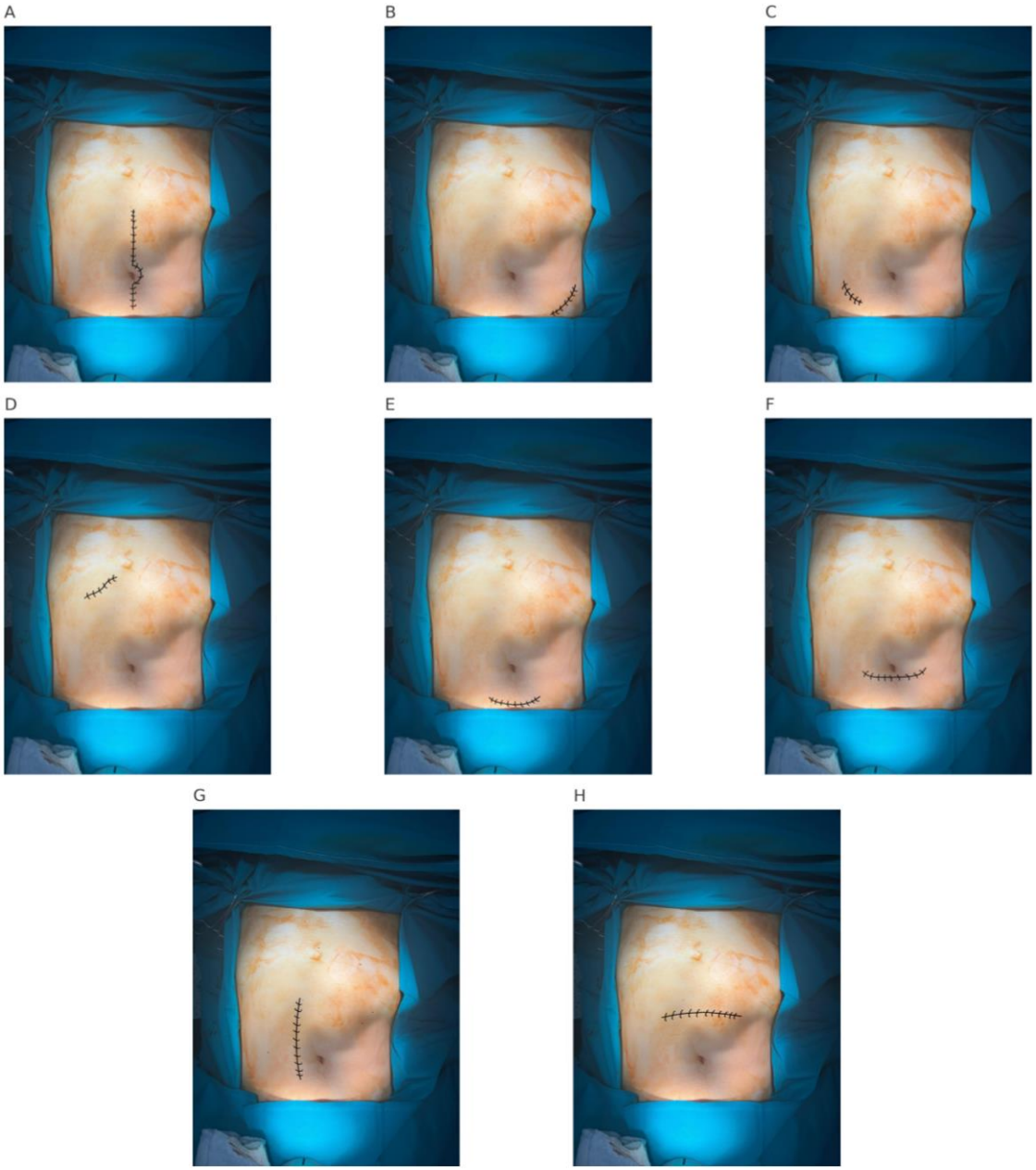

Fig. 1. Images presented to ChatGPT. Panels A–H correspond to the manuscript items in order: (a) Item 2; (b) Item 3; (c) Item 4; (d) Item 5; (e) Item 6; (f) Item 7; (g) Item 8; (h) Item 9.

3. Results

Eight incision-based items (Items 2–9) were analyzed. For each visual prompt, ChatGPT-4.0 was asked to suggest a regional anesthesia method and a postoperative analgesia plan. Five anesthesiologists independently rated each response on a five-point Likert scale (1–5) under three domains: Accuracy, Comprehensiveness, and Safety. Inter-rater reliability was quantified using Fleiss' kappa (κ). Item 10 differed in format and was treated as exploratory; it is summarized descriptively and excluded from pooled statistics.

3.1. Domain-level performance

Mean scores were high across domains: Accuracy 4.28, Comprehensiveness 4.30, and Safety 4.00. Inter-rater agreement was substantial for Safety (κ=0.76), while Accuracy (κ=0.33) and Comprehensiveness (κ=0.31) showed lower, fair-to-moderate agreement (Table 1).

3.2. Item-level highlights

Overall scores were highest for Item 2 (mean 4.93) and lowest for Item 6 (mean 3.27). By domain: Accuracy peaked at Item 2 (5.00) and was lowest at Item 6 (3.00); Comprehensiveness peaked at Item 3 (5.00) and was lowest at Item 6 (3.40); Safety was highest at Item 2 (5.00) and lowest at Item 4 (3.00) (Table 2).

3.3. Interpretation

Safety judgments were the most consistent among raters, suggesting clearer shared thresholds for risk and

feasibility. Comprehensiveness varied more, indicating differences in how breadth and depth of content were weighed. Accuracy fell between these two patterns. Exploratory analysis of Item 10 did not materially change the overall interpretation (Table 3).

Table 1. Domain-level summary (primary: Items 2–9).

Domain	Mean across items & raters	Fleiss' κ	$\bar{P}$ (observed agreement)	P_e (chance agreement)
Accuracy	4.275	0.332	0.60	0.401
Comprehensiveness	4.300	0.310	0.60	0.420
Safety	4.000	0.760	0.85	0.375

Interpretation (κ): ≤0.20 poor, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, ≥0.81 almost perfect.

Table 2. Per-item summary (primary: Items 2–9).

Item	Accuracy Mean	Accuracy SD	Comprehensiveness Mean	Comprehensiveness SD	Safety Mean	Safety SD	Overall Mean	Overall SD
2	5.000	0	4.800	0.447	5.000	0	4.933	0.258
3	4.600	0.548	5.000	0	5.000	0	4.867	0.352
4	4.400	0.548	3.800	0.447	3.000	0	3.733	0.704
5	4.200	0.447	4.400	0.548	4.000	0	4.200	0.414
6	3.000	0	3.400	0.548	3.400	0.548	3.267	0.458
7	4.200	0.447	4.400	0.548	4.000	0	4.200	0.414
8	4.400	0.548	4.000	0	3.600	0.548	4	0.535
9	4.400	0.548	4.600	0.548	4.000	0	4.333	0.488

Table 3. Rater-level overall means.

Rater	Overall mean across items & domains
Rater 1	4.167
Rater 2	4.208
Rater 3	4.333
Rater 4	4.083
Rater 5	4.167

Across incision-based items, average ratings were high (Fig 2). Item 2 (median midline) and Item 3 (LLQ) reached 5.0 in all three criteria. Item 5 (RUQ/Kocher), Item 7 (Pfannenstiel), Item 8 (paramedian), and Item 9 (transverse midline) clustered between 4.0–4.6. Two clear low points emerged: Item 4—Safety mean 3.0 for the right-lower-quadrant (McBurney/Lanz) incision, and Item 6—Accuracy 3.0; Comprehensiveness 3.4; Safety 3.4 for the suprapubic (Pfannenstiel) incision. These dips align with the agreement results—Safety had the highest κ, indicating that lower safety scores likely reflect shared caution rather than rater noise.

4. Discussion

In this study, we found that ChatGPT-4.0 generally produced satisfactory recommendations for abdominal incisions across accuracy, comprehensiveness, and safety. However, performance clearly dipped in two scenarios: the right-lower-quadrant (McBurney/Lanz) incision—especially on Safety—and the suprapubic (Pfannenstiel) incision, where scores were lower across multiple criteria. Notably, when we explicitly asked for postoperative analgesia plans, the model rarely proposed neuraxial techniques at all, favoring peripheral/fascial-plane blocks. Taken together with the substantial inter-rater agreement on Safety, these findings suggest not random variability but a consistent, clinically meaningful concern in these specific contexts and a systematic omission of neuraxial options.

In abdominal surgery, contemporary ERAS pathways increasingly reserve—rather than routinely use—neuraxial analgesia (e.g., thoracic epidural), particularly for laparoscopic cases and fast-track recovery protocols [6]. This shift reflects practical concerns about urinary retention, hemodynamic instability, and catheter logistics, alongside wider adoption of multimodal systemic analgesia and fascial-plane blocks (TAP, QL, ESP) that can support early mobilization [7,8]. Consistent with this trend, ChatGPT-4.0 likewise tended to de-emphasize neuraxial options in its postoperative plans, favoring peripheral/fascial-plane techniques. While this alignment with modern protocols is reassuring, it also raises the possibility of under-recommending neuraxial techniques in selected open procedures where epidural analgesia may still offer meaningful benefit.

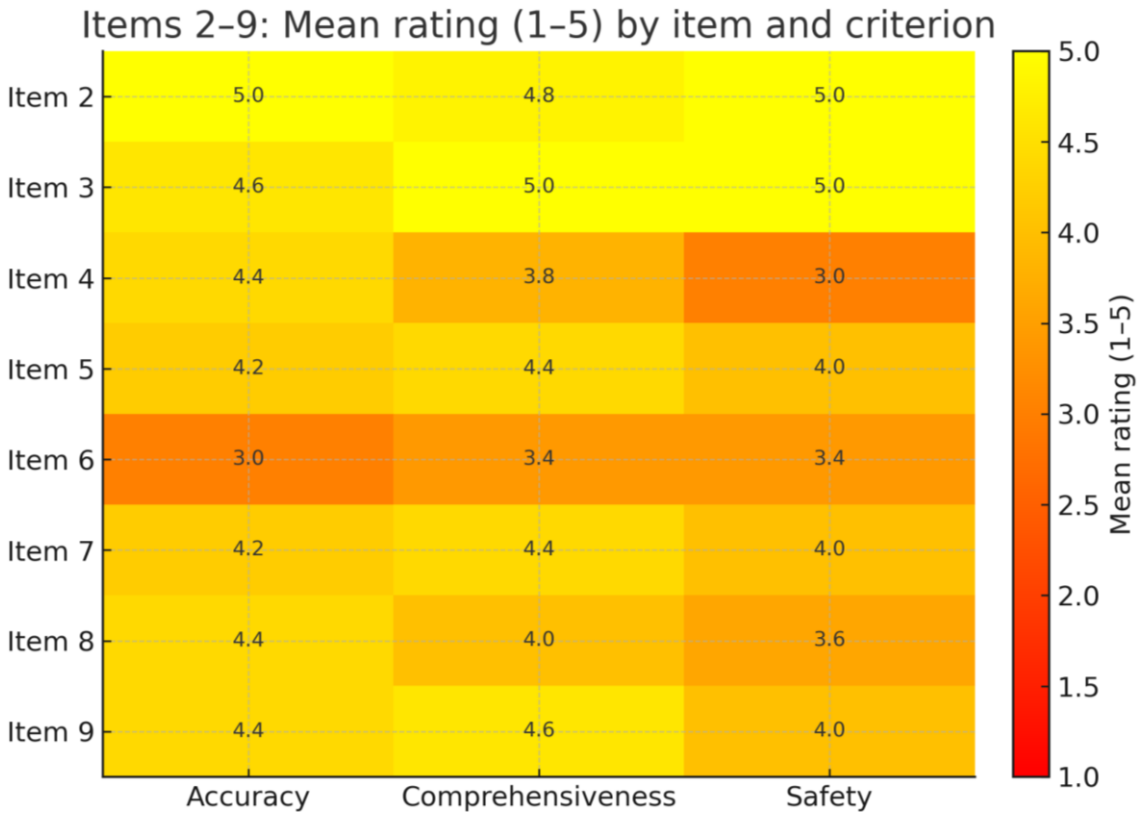

Fig. 2. Heat-map of mean ratings (1–5) for ChatGPT-4.0 recommendations across eight incision-based items (Items 2–9) and three criteria (Accuracy, Comprehensiveness, Safety). Cells show the mean of five anesthesiologist ratings; the colour scale maps red = 1 (low) to yellow = 5 (high).

Beyond lay-facing evaluations of LLMs—where readability, basic factual accuracy, and patient safety are usually the endpoints—judging whether a model can advise an anesthesiologist on analgesia is a different task [9,10]. It requires professional judgment about anatomy, block selection, contraindications (e.g., coagulopathy, infection risk), neuraxial vs. peripheral trade-offs, multimodal rescue options, and feasibility in real operating pathways. Our incision-based design, the use of independent expert raters, and criterion-specific scoring with inter-rater agreement move the assessment squarely into that professional space. The results show that while ChatGPT-4.0 often delivers clinically usable suggestions, it also exhibits context-specific gaps (e.g., RLQ safety, Pfannenstiel underperformance, and under-recommendation of neuraxial options). In short, the study successfully differentiates where the model's guidance aligns with professional standards and where caution or augmentation is warranted for real-world anesthetic pain management.

This evaluation has several constraints. First, we used a single-shot prompt per scenario; because LLM outputs are stochastic and time-varying, repeated queries at different times (or with seed control) might have yielded greater response variability. Second, we queried only one product configuration (ChatGPT-4.0) from a single account; we did not compare free vs. paid tiers, alternate accounts/devices (which may be subject to silent A/B tests), or other LLMs—so generalisability across platforms is uncertain. Third, the visual stimuli were simulated incision images, not real operative photos or ultrasound-based views; important contextual cues (e.g., intraoperative findings, comorbidities, coagulation status) were intentionally withheld, which may have constrained the model's recommendations (notably for neuraxial options). Fourth, we used one prompt phrasing and did not test few-shot/system-prompt strategies; prompt engineering can materially shift outputs. Fifth, the rater panel was small (n=5) and composed of fascial-plane block experts; while this aligns with our focus, it may limit applicability to broader anesthesiology practice. Sixth, outcomes were scored on a five-point Likert scale and agreement summarised with Fleiss' κ; given the ordinal nature of the scale and the limited number of primary items (n=8), these statistics should be interpreted with caution. Confidence intervals or bootstrapped estimates could not be calculated due to the small sample size, and this limitation should be considered when interpreting the reliability of the findings. Finally, we assessed content quality, not clinical effectiveness; no patient-level analgesic outcomes, adverse events, or workflow metrics were captured, so real-world impact remains to be demonstrated.

5. Conclusions

This prospective, expert-rated study found that a large language model generally provided clinically usable regional anesthesia and postoperative analgesia suggestions for common abdominal incisions, with substantial agreement on Safety. Performance was not uniform: recommendations for the right-lower-quadrant (McBur-

ney/Lanz) and suprapubic (Pfannenstiel) incisions scored lower—especially on Safety—and postoperative plans rarely included neuraxial techniques, which may mirror ERAS trends yet risk underuse in selected open cases. Overall, such models can serve as a helpful adjunct for option-generation and education, but they are not a substitute for anesthesiologist judgment and local protocols. Responsible use requires expert oversight and clear guardrails; future work should examine repeated sampling, prompt standardization, version/tier comparisons, and linkage of recommendations to patient outcomes. Beyond these findings, the proposed methodology offers a practical framework that may guide future AI evaluation standards and support responsible integration of AI tools into anesthesia education.

Acknowledgements

The Author thanks to the five independent anesthesiologists who served as expert raters for their time and thoughtful evaluations. It is also acknowledged that the use of ChatGPT-4.0 for assistance with figure caption wording and minor language polishing; all clinical content, analyses, and the final manuscript were authored and verified by the investigators. Finally, the Author is grateful to the academics and educators in Türkiye who have guided regional anesthesia practice, whose teaching and scholarship provided the intellectual backdrop for this study.

Funding

The author received no financial support for the research, authorship, and/or publication of this manuscript.

Conflict of Interest

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this manuscript.

Data Availability

The datasets created and/or analyzed during the current study are not publicly available, but are available from the corresponding author upon reasonable request.

Ethics Approval and Consent to Participate

This study was approved by the ethics committee of Samsun University Clinical Research Ethics Committee (Reference Number: GOKAEK 2025/11/7; Date: 28/05/2025). Written informed consent was obtained from the participants. All methods were performed in accordance with relevant guidelines and regulations.

Author Contributions

The author confirms sole responsibility for all aspects of the study including: conceptualization, methodology, formal analysis, investigation, data curation, visualization, writing – original draft, and writing – review & editing.

REFERENCES

1. Yamamoto T, Schindler E. Regional anesthesia as part of enhanced recovery strategies in pediatric cardiac surgery. *Curr Opin Anaesthesiol*. **2023**;36(3):324–333.
2. Ahiskalioglu A, Yayik AM, Celik EC, et al. The shining star of the last decade in regional anesthesia part I: Interfascial plane blocks for breast, thoracic, and orthopedic surgery. *Eurasian J Med*. **2022**;54(Suppl 1):97–105.
3. Yayik AM, Celik EC, Aydin ME, et al. The shining star of the last decade in regional anesthesia part II: Interfascial plane blocks for cardiac, abdominal, and spine surgery. *Eurasian J Med*. **2023**;55(Suppl 1):9–20.
4. Nelms MW, Javidan A, Chin KJ, et al. YouTube as a source of education in perioperative anesthesia for patients and trainees: A systematic review. *Can J Anaesth*. **2024**;71(9):1238–1250.
5. Gul S, Erdemir I, Hanci V, Aydogmus E, Erkoc YS. How artificial intelligence can provide information about subdural hematoma: Assessment of readability, reliability, and quality of ChatGPT, BARD, and Perplexity responses. *Medicine (Baltimore)*. **2024**;103(18):e38009.
6. Wagemans MF, Scholten WK, Hollmann MW, Kuipers AH. Epidural anesthesia is no longer the standard of care in abdominal surgery with ERAS: What are the alternatives?. *Minerva Anestesiol*. **2020**;86(10):1079–1088.
7. Roofthooft E, Joshi GP, Rawal N, Van de Velde M, PROSPECT Working Group of the European Society of Regional Anaesthesia and Pain Therapy and supported by the Obstetric Anaesthetists' Association. PROSPECT guideline for elective caesarean section: Updated systematic review and procedure-specific postoperative pain management recommendations. *Anaesthesia*. **2021**;76(5):665–680.
8. Lirk P, Thiry J, Bonnet MP, Joshi GP, Bonnet F. Pain management after laparoscopic hysterectomy: Systematic review of literature and PROSPECT recommendations. *Reg Anesth Pain Med*. **2019**;44(4):425–436.
9. Ismaiel N, Nguyen TP, Guo N, Carvalho B, Sultan P, study collaborators. The evaluation of the performance of ChatGPT in the management of labor analgesia. *J Clin Anesth*. **2024**;98:111582.
10. Meyer MKR, Kandathil CK, Davis SJ, et al. Evaluation of rhinoplasty information from ChatGPT, Gemini, and Claude for readability and accuracy. *Aesthetic Plast Surg*. **2024**;49:1868–1873.