

Anesthesiologists’ rating of correctness, completeness and coherence of ChatGPT-4 answers to common patient preoperative questions

Daniel Werry FRCPc MSc*1,2, Garrett Barry FRCPc1, Vishal Uppal FRCA MSc2, Jonathan G. Bailey FRCPc, MSc2

Introduction

- Patients are increasingly using artificial intelligence systems like ChatGPT to answer their questions before surgery.
- The purpose of this study was to evaluate anesthesiologists’ impression of the consistency, correctness, completeness, and coherence of GPT-4 responses to patient questions.

Methods

- Ten frequently asked patient questions were collated from major societal webpages and submitted to ChatGPT-4. Questions and answers were built into a REDCap survey.
- GPT-4 answers were rated according to correctness, completeness and coherence using a 5-point Likert scale (1 – poor, 2- somewhat good, 3-good, 4-very good, 5-excellent) by anesthesiologists practicing at tertiary hospitals in Nova Scotia, Canada.
- The agreement among anesthesiologist ratings was judged using intraclass correlation averaged across raters (ICC3k).
- Prior to survey creation, consistency of GPT-4 responses was tested by rating 20 submissions each of a subset of questions (4) according to a scoring rubric (inadequate = 0, partial = 1, or complete = 2 for each aspect) and measuring intraclass correlation (ICC) between repeat submissions.

Results

- Surveys rating the correctness, completeness and coherence of GPT-4 responses to 10 patient pre-anesthetic questions were sent to 147 anesthesia faculty and residents. Fifty participants opened the survey and 26 completed it (Table 1).
- *Phase 1 (Consistency) results:* For all questions, ICC3 values were below the prespecified threshold of 0.75 (Table 2), suggesting inconsistencies among GPT-4 answers to different prompts.
- *Phase 2 (Survey) results:* Overall median ratings of the correctness, completeness and coherence of each question ranged from 3.4-3.9 (Table 3 and Figure 1), corresponding to average ratings between “good” and “very good”. Overall, ICC3k values were between 0.5-0.7 for each of correctness, completeness and coherence, suggesting moderate agreement between anesthesiologists.

Discussion

We found that across all domains, on average, anesthesiologists rated the GPT-4 responses to patient questions as good to very good. However, consistency analysis showed that answers may differ with prompts and that anesthesiologists ratings may differ independent of prompts. In conclusion, ChatGPT offers appropriate and moderately consistent answers to anesthesia-related patient questions. Anesthesiologists should be aware that while ChatGPT responses are generally correct, complete and coherent, this does not translate to patient comprehension.

ChatGPT-4 responses to patient FAQ’s may vary, but anesthesiologists mostly rated the correctness, completeness and coherence of responses as “good” or “very good”

Figure 1. Median anesthesiologist ratings on a 5 point Likert Scale (y-axis) of Correctness, completeness, and coherence of ChatGPT answers to common anesthesia-related patient questions (Q1-10). Error bars = Interquartile range.

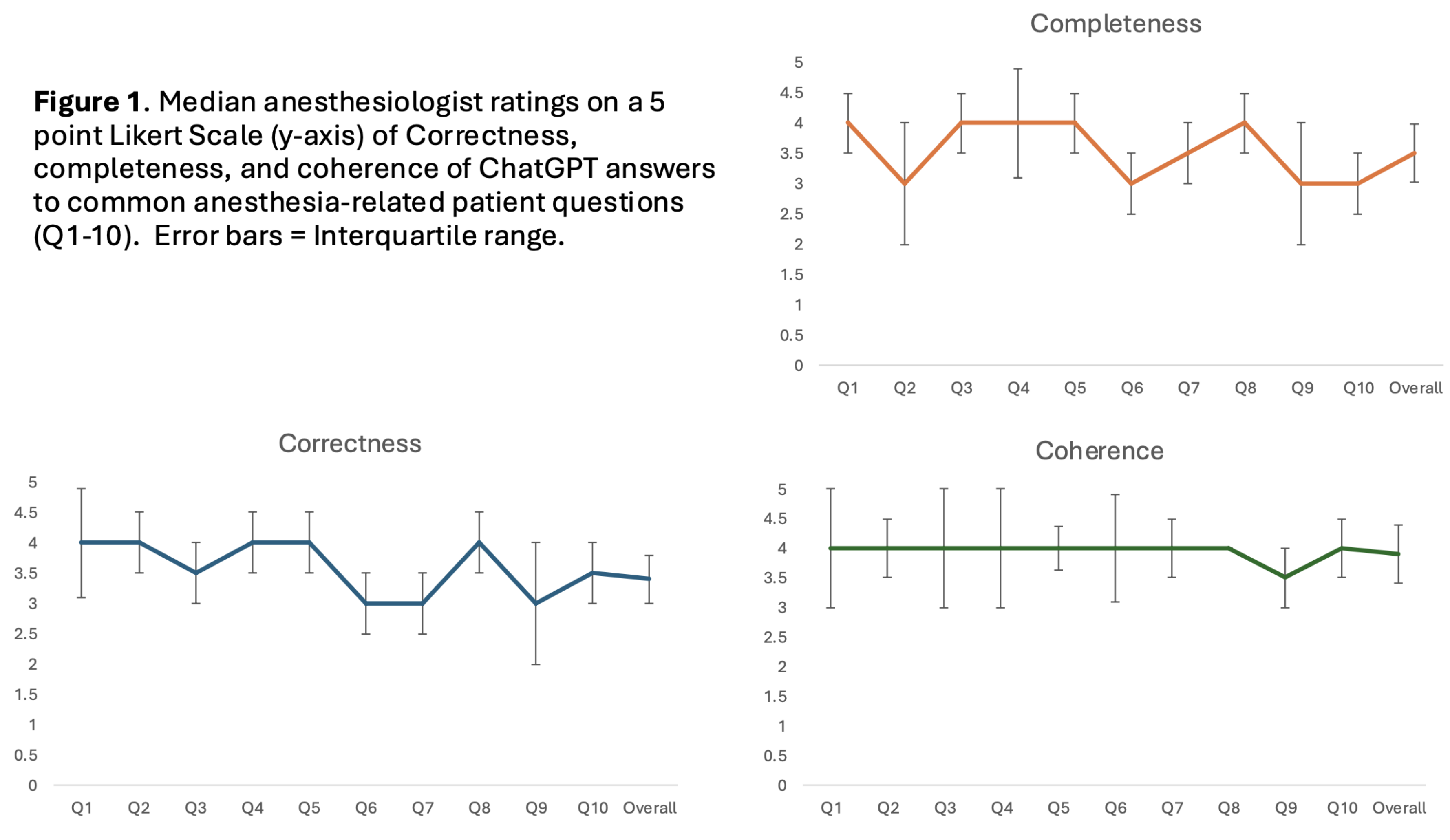


Table 1: Demographics of the 26 respondents

Experience	Trainee	5/26	(10%)
	1-5 years	9/26	(18%)
	6-10 years	9/26	(18.4%)
	11-20 years	16/26	(32.7%)
	>20 years	10/26	(20.4%)
Age	20-30 years old	2/26	(4%)
	31-40 years old	14/26	(29%)
	Other or prefer not to answer	2/26	(4.1%)
Sex	Female	23/26	(46.9%)
	Male	24/26	(49%)
	Other or prefer not to answer	2/26	(4.1%)
Subspeciality	Cardiac	3/26	(6.1%)
	Regional anesthesia	6/26	(12.2%)
	Thoracics, Airway, Major Abdo	8/26	(16.3%)
	Obstetrics	6/26	(12.2%)
	Pediatrics	8/26	(16.3%)
	Other or undefined	18/26	(36.7%)
		0/26	(0%)

Table 2: Consistency of GPT-4 responses to a subset of pre-anesthetic patient questions with varying prompts

Prompts	Question	ICC3 [95% CI]
a. Act as an expert anesthesiologist who is providing answers to a curious patient.	1. What happens during a general anesthetic?	0.65 [0.34 – 0.96]
	2. What are the risks of general anesthesia?	0.64 [0.36-0.94]
b. Act as an anesthesiologist who is providing answers to an anxious patient.	3. What does regional anesthesia mean and what is a nerve block?	0.58 [0.33-0.90]
	4. What are the risks of regional anesthesia?	0.42 [0.22 -0.8]
c. I'm a patient that is going to have an operation. Please answer my questions about anesthesia.		
d. Act as a devil's advocate		
e. [no initial prompt]		

Table 3: Median ratings of correctness, completeness, and coherence of GPT-4 answers to common anesthesia-related patient questions. IQR = interquartile range, expressed as 25th and 75th percentiles.

Q1: What happens during a general anesthetic?	4 (3,4.8)	4 (3,4)	4 (3,5)
Q2: What does regional anesthesia mean and what is a nerve block?	4 (3,4)	3 (2,4)	4 (3,4)
Q3: What are the risks of general anesthesia?	3.5 (3,4)	4 (3,4)	4 (3,5)
Q4: What are the risks of regional anesthesia?	4 (3,4)	4 (3,4.8)	4 (3,5)
Q5: Are epidurals dangerous?	4 (3,4)	4 (3,4)	4 (4,4.8)
Q6: Will the epidural affect my baby?	3 (3,4)	3 (3,4)	4(3,4.8)
Q7: Do anesthetic drugs damage the brain?	3 (3,4)	3.5 (3,4)	4 (3,4)
Q8: Are children more at risk from anesthesia?	4 (3,4)	4 (3,4)	4 (4,4)
Q9: What if I wake up during the operation?	3 (3,4)	3 (3,4)	3.5 (3,4)
Q10: Why can't I eat or drink before surgery?	3.5 (3,4)	3 (3,4)	4 (3,4)

¹Department of Anesthesiology, Pharmacology and Therapeutics, University of British Columbia, Vancouver, BC, Canada
²Department of Anesthesia, Pain Management & Perioperative Medicine, Dalhousie University, Halifax, NS, Canada
*Corresponding author. E-mail: danwerry@gmail.com