Generation of preoperative anaesthetic plans by ChatGPT-4.0: a mixed-method study

Malek, M.A.; Velzen, M. van; Dahan, A.; Martini, C.; Sitsen, E.; Sarton, E.; Boon, M.

BJA

British Journal of Anaesthesia, 134 (5): 1333–1340 (2025)

doi: 10.1016/j.bja.2024.08.038
Advance Access Publication Date: 14 November 2024
Clinical Investigation

CLINICAL PRACTICE

Generation of preoperative anaesthetic plans by ChatGPT-4.0: a mixed-method study

Michel Abdel Malek*, Monique van Velzen, Albert Dahan, Chris Martini, Elske Sitsen, Elise Sarton and Martijn Boon

Department of Anaesthesiology, Leiden University Medical Centre, Leiden, The Netherlands

*Corresponding author. E-mail: m.abdel_malek@lumc.nl

Abstract

Background: Recent advances in artificial intelligence (AI) have enabled development of natural language algorithms capable of generating coherent texts. We evaluated the quality, validity, and safety of this generative AI in preoperative anaesthetic planning.

Methods: In this exploratory, single-centre, convergent mixed-method study, 10 clinical vignettes were randomly selected, and ChatGPT (OpenAI, 4.0) was prompted to create anaesthetic plans, including cardiopulmonary risk assessment, intraoperative anaesthesia technique, and postoperative management. A quantitative assessment compared these plans with those made by eight senior anaesthesia consultants. A qualitative assessment was performed by an adjudication committee through focus group discussion and thematic analysis. Agreement on cardiopulmonary risk assessment was calculated using weighted Kappa, with descriptive data representation for other outcomes.

Results: ChatGPT anaesthetic plans showed variable agreement with consultants' plans. ChatGPT, the survey panel, and adjudication committee frequently disagreed on cardiopulmonary risk estimation. The ChatGPT answers were repetitive and lacked variety, evidenced by the strong preference for general anaesthesia and absence of locoregional techniques. It also showed inconsistent choices regarding airway management, postoperative analgesia, and medication use. While some differences were not deemed clinically significant, subpar postoperative pain management advice and failure to recommend tracheal intubation for patients at high risk for pulmonary aspiration were considered inappropriate recommendations.

Conclusions: Preoperative anaesthetic plans generated by ChatGPT did not consistently meet minimum clinical standards and were unlikely the result of clinical reasoning. Therefore, ChatGPT is currently not recommended for preoperative planning. Future large language models trained on anaesthesia-specific datasets might improve performance but should undergo vigorous evaluation before use in clinical practice.

Keywords: artificial intelligence; ChatGPT; clinical decision support; generative AI; preoperative anaesthetic planning

Editor's key points

- This exploratory mixed-method study evaluated the capacity of ChatGPT-4.0 to generate preoperative anaesthetic plans for a set of clinical cases compared with a panel of expert anaesthesiologists.
- Preoperative risk evaluation and plans for anaesthetic management, monitoring, and postoperative care were generally coherent, but occasionally failed to meet minimum standards.
- Further development of large language models trained on anaesthesia-specific datasets should improve performance before application to clinical practice.

Received: 16 November 2023; **Accepted:** 20 August 2024

Worldwide, over 300 million surgical procedures are performed annually that need some form of general, neuraxial, or regional anaesthesia.[1] Anaesthesiologists must decide which anaesthetic technique is safest and most suitable for each unique patient–procedure combination. This requires thorough knowledge of the procedure and its risks, the pharmacokinetics and pharmacodynamics of administered drugs, and their influence on human physiology. Appropriate airway strategy and potential intraoperative and postoperative complications are important aspects of preoperative anaesthetic planning. However, because of variations in medical expertise among attending anaesthesia providers, there can be differences in the quality of preoperative anaesthetic preparation.

Recently, there have been major recent advances in the field of natural language processing, in which the generative pretrained transformer (GPT) model has demonstrated remarkable proficiency in interpreting input and generating coherent text.[2] These models could support clinical caregivers in decision pathways and reduce inter-physician variability and ultimately improve outcome.[3,4] In the field of anaesthesia, it might help the physician to identify patients at risk for major intraoperative or postoperative complications, indicate the most appropriate anaesthetic techniques, or flag potential flaws in existing preoperative plans. However, application of GPT for this purpose is new and has not been investigated. We investigated the potential of the large language model (LLM) ChatGPT version 4.0 in estimating perioperative cardiopulmonary risk and generating a preoperative anaesthetic plan for 10 specific patient-procedure scenarios. These plans were compared with plans generated by medical professionals and were qualitatively assessed by an adjudication committee of senior anaesthesiologists.

Methods

The study was submitted for review to the institutional ethics board (METC-LDD, study reference no. 2023–062), which waived the need for ethical review.

Design

This mixed-method study aimed to explore the utility of the LLM, general pretrained transformer, version 4.0 (ChatGPT 4.0, version May 23rd 2023; OpenAI), in preoperative anaesthetic planning in a tertiary care centre in the Netherlands. We performed a convergent study design which quantitatively compared preoperative anaesthetic plans generated by GPT-4 for 10 cases with preoperative anaesthetic plans generated by a survey panel of eight anaesthesiologists. In addition, a qualitative assessment was performed by an adjudication committee consisting of senior consultant anaesthesiologists (AD/MS/CM). A flowchart of the study is provided as Figure 1. To prevent GPT from learning from our inputs during the conduct of this study, all prompts were given to OpenAI with the 'Chat History & Training' setting turned off. This prevented the LLM from learning from our previous input and the generated output.

Case vignettes

We collected 20 anaesthetic cases that had taken place in our tertiary centre in 2022 and 2023. The case-mix consisted of all types of surgery, including transplant, paediatric, and cardiothoracic surgery. Cases were required to have a complicating factor, such as a deranged physiology, a complicated medical history, a higher risk of complications, an emergency, or a combination thereof. Neither the researchers nor the adjudication committee or consultants of the survey panel had been clinically involved in the cases. The cases were de-identified and rewritten to 20 uniform vignettes containing relevant information in colloquial medical language to emulate real-world use. To overcome bias for specific types of surgery, 10 out of 20 vignettes were chosen at random for the study using Research Randomizer (version 4.0; Urbaniak and Plous, 2013). See Supplementary material (Supplemental file 1) for the case vignettes.

Prompt

We engineered a prompt to ask ChatGPT to generate a preoperative anaesthetic plan for each case vignette. The prompt consisted of a cardiovascular and pulmonary risk assessment followed by a choice for type of anaesthesia (general, regional, peripheral, procedural sedation, and analgesia, or any of these combined). For each type, choice of anaesthetic and required drug doses were queried. In the case of neuraxial anaesthesia or peripheral nerve blocks, type and level of block and the local anaesthetic were queried. We further asked ChatGPT the plans for monitoring, number of intravenous catheters, necessity of invasive and noninvasive monitoring, and vasopressors and inotropic medication. Lastly, we asked ChatGPT to define appropriate postoperative management, including specific drugs to be used, prophylaxis against postoperative nausea and vomiting, and whether the patient should be closely monitored after surgery (see Supplementary material [Supplemental file 2]). We followed a 'zero-shot' approach, in which the prompt did not contain examples of what to answer. All questions were given as open-ended questions, except for a limited number of questions in which an option list was given to simplify responses (e.g. 'normal, intermediate, or high' for cardiovascular risk assessment). All questions were given in the same prompt. There were no follow-up questions or clarifications requested.

Quantitative assessment

For quantitative assessment of answers, eight randomly selected, blinded senior anaesthesia consultants were asked to complete a survey form that contained the same case vignettes that were presented to ChatGPT. The questions in the survey form for each case vignette were identical and had the same wording and format as the prompt used for ChatGPT (Supplementary material [Supplemental file 3]). These data were captured using an online data capture tool (Castor EDC, Amsterdam, The Netherlands).

Qualitative assessment

For qualitative assessment of answers, an adjudication committee of three experienced clinicians each with >20 yr of clinical experience (AD/CM/MS) held a focus group discussion with the researchers (MAM/MvV/MB). Before the study, the adjudication committee was neither aware of the cases nor the answers provided by ChatGPT. The adjudication committee was familiarised with the case vignettes and asked to formulate an anaesthetic plan for each vignette, weighing risks and options. These plans were used as a reference for discussing and qualitatively evaluating the answers ChatGPT produced for the same cases. The qualitative assessment was a structured discussion, focused on validity, usability,

evaluation of relevance and appropriateness, evaluation of safety and risk, and consistency of answers.

Data analysis

Both qualitative and quantitative results are presented in this mixed-method study. For the qualitative results obtained from the appraisal of ChatGPT answers by the adjudication committee, we conducted a comprehensive thematic analysis of the qualitative data collected. The collected information was carefully synthesised, and common threads and recurring themes regarding the ChatGPT output were identified.

To evaluate quantitative data, we used Krippendorff's alpha to calculate agreement in risk assessment and Fleiss Kappa for the other (ordinal) data. Descriptive analyses were used for the other domains when appropriate.

Model accession

ChatGPT-4 model version date 23rd May, 2023 was used.

Results

This study was conducted between May and June 2023. Surveys were completed by all eight anaesthesiologists of the survey panel and all data were eligible for analysis. Figure 1 contains a flowchart of the study and Table 1 presents a summary of the quantitative and qualitative assessments. The results of both analyses are discussed per domain below.

Risk assessment

Agreement between the survey panel and ChatGPT-4 with regard to perioperative cardiac and pulmonary risk assessment was fair and poor, respectively (Cohen's kappa 0.41 and 0.13, respectively). Risk estimates by ChatGPT for either postoperative cardiac or pulmonary complications were adjudicated to be adequate in only three of the 10 cases. In the other cases, risk estimates were considered to be either too high or too low. For instance, in two cases for cardiac surgery (Case 9 elective coronary artery bypass grafting and case 10 urgent repair of type A aortic dissection), the risk for postoperative cardiac complications (Case 9) or pulmonary complications (PPC, Case 10) were estimated to be intermediate by ChatGPT, while the majority of survey panellists and the adjudication committee judged this risk to be high. Discrepancies were also noted in non-cardiothoracic surgery cases (Table 1). Overall agreement within the group of panellists was good (Krippendorff's alpha 0.76).

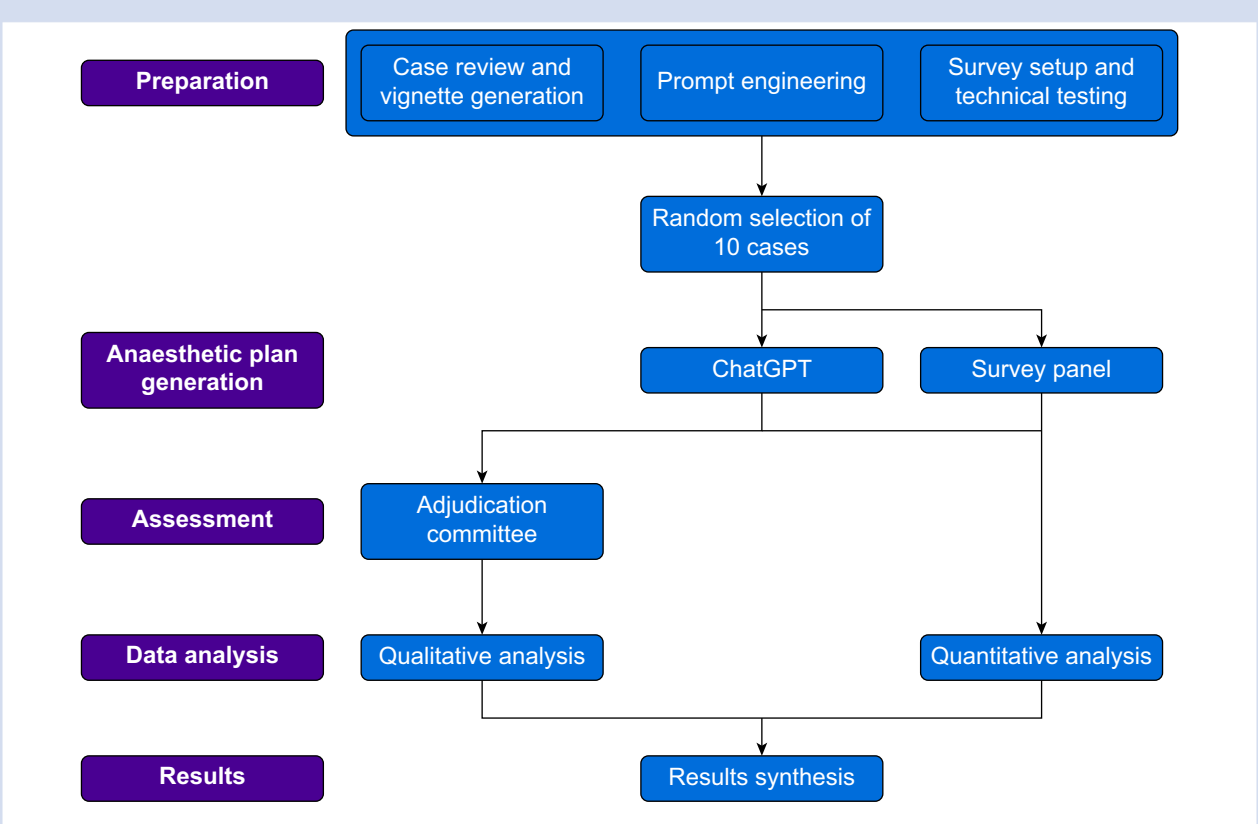

Fig 1. Flowchart of the study design. The study began with the preparation phase, which involved compiling 20 clinical vignettes of real-world cases, prompt engineering, and setting up a structured survey. Ten cases were then randomly selected and presented to ChatGPT to generate anaesthetic plans, and to a survey panel of eight senior anaesthesia consultants. The outputs from ChatGPT were then assessed through separate qualitative and quantitative assessments. The qualitative phase involved focus group discussion with an adjudication committee and a thematic analysis. The quantitative phase compared the results with the outcomes of the structured survey by the panel. The results from both assessments were then converged in the results phase.

Table 1 The quantitative and qualitative assessment of GPT-4. This table presents the qualitative assessments and quantitative comparisons of 10 clinical vignettes, with the output of GPT-4 juxtaposed against the survey panel and adjudication committee. For each case, the results from both the survey panel and the adjudication committee were recorded. Assessments covered three domains: risk assessment, anaesthesia technique, and postoperative management, each subdivided by category. To illustrate the distribution of responses in the risk assessment domain, the number of panellists who contributed to each answer is provided. In anaesthesia technique, majority opinions from the survey panel are indicated with an asterisk (*). In cases of negative adjudication, comments from the adjudication committee were included, with references to relevant anaesthesia guidelines. CABG, coronary artery bypass grafting; Combined, combined technique; CVC, central venous catheter; EA, epidural anaesthesia; GA, general anaesthesia; GPT, generative pretrained transformer; IAC, intra-arterial cannula; MCU, medium care unit; O, advice was judged appropriate; PONV, postoperative nausea and vomiting; RA, regional anaesthesia; SA, spinal anaesthesia; SAD, supraglottic airway device; TOE transoesophageal echocardiogram; TT, tracheal tube; VL, videolaryngoscopy; X, advice was judged suboptimal or inappropriate (see comments below). *At least 6 out of 8 survey panellists chose this option. ① Paracetamol. ② NSAID. ③ Oral opioid. ④ Parenteral opioid. ⑤ Regional or neuraxial anaesthesia. ⑥ Patient-controlled analgesia. ⑦ Ondansetron. ⑧ Dexamethasone. ⑨ Metoclopramide.

		Risk assessment	Anaesthesia technique			Postoperative management		
		Cardiac (*n*)/pulmonary (*n*)	Type of anaesthesia	Airway device	Invasive monitoring options	Pain management options	PONV prophylaxis	Destination
1 Adult Elective Shoulder prosthesis	Survey panel	Normal (7) intermediate (1)/intermediate (8)	GA + RA*	TT*	None*	① ② ③ ④ ⑤	⑦* ⑧	Ward*
	GPT	Intermediate/high	GA	TT	None	① ④ ⑥	⑦	MCU/ICU
	Adjudication	X [f]	X [a]	O	O	O	O	O
2 Adult elective transthoracic oesophagectomy	Survey panel	Intermediate (8)/ intermediate (8)	GA + EA*	TT*	IAC*	① ② ⑤ ⑥*	⑦* ⑧	MCU/ICU*
	GPT	High/high	GA	TT	CVC IAC	① ② ④ ⑥	⑦	MCU/ICU
	Adjudication	O	X[1]	O	O	O	O	O
3 Adult elective open gastrectomy	Survey panel	Intermediate (5) high (3)/ intermediate (5) high (3)	GA + EA*	TT*	None*	① ② ⑤ ⑥*	⑦* ⑧	MCU/ICU*
	GPT	High/high	GA	TT	IAC	① ③	⑦	MCU/ICU
	Adjudication	O	X[1]	O	O	X[B]	O	O
4 Adult urgent repeat laparotomy after liver transplant	Survey panel	Intermediate (5) high (3)/ normal (3) intermediate (5)	GA*	*In situ* tracheostomy	IAC* CVC	① ④ ⑥	⑦ ⑧	MCU/ICU
	GPT	High/high	GA	*In situ* tracheostomy	IAC CVC	① ④ ⑥	⑦	MCU/ICU
	Adjudication	X[f]	O	O	X[c]	O	O	O
5 Adult urgent Caesarean delivery in a preeclamptic patient	Survey panel	Normal (4) intermediate (3) high (1)/normal (5) intermediate (2) high (1)	GA (5) SA (3)	TT	None*	①* ② ④*	⑦ ⑨	Ward*
	GPT	High/normal	GA	TT	IAC	① ② ④ ⑥	⑦	MCU/ICU
	Adjudication	X[f]	X[d]/X[e]	O	O	O	O	O
6 Adult urgent laparoscopic resection of ectopic pregnancy	Survey panel	Intermediate (4) or high (4)/ normal (6) intermediate(2)	GA*	TT*	None*	①* ③ ④	⑦ ⑧	Ward*
	GPT	High/intermediate	GA	SAD	IAC CVC	① ④ ⑥	⑦	MCU/ICU
	Adjudication	X[f]	O	X[2]/X[h]	X[i]	O	O	O

Continued

Table 1 Continued

		Risk assessment	Anaesthesia technique			Postoperative management		
		Cardiac (*n*)/pulmonary (*n*)	Type of anaesthesia	Airway device	Invasive monitoring options	Pain management options	PONV prophylaxis	Destination
7 Child urgent laparoscopic appendectomy	Survey panel	Normal (8)/normal (8)	GA*	TT*	None*	①* ②* ③	⑦*	Ward*
	GPT	Normal/normal	GA	SAD	None	①	⑦	Ward
	Adjudication	O	O	X[2]	O	X[b]	X[k]	O
8 Child urgent mastoidectomy	Survey panel	Normal (8)/normal (7) intermediate (1)	GA*	TT*	None*	①* ②	⑦* ⑧	Ward*
	GPT	Normal/high	GA	Uncuffed TT	None	① ②	⑦	Ward
	Adjudication	X[f]	O	X[l]	O	O	X[k]	O
9 Adult elective CABG	Survey panel	Intermediate (2) high (6)/ normal (1) intermediate (7)	GA*	TT*	IAC* CVC*	①* ② ④*	⑦*	ICU*
	GPT	Normal/high	GA	TT	IAC CVC	① ② ④ ⑥	⑦	ICU
	Adjudication	X[f]	O	O	O	O	O	O
10 Adult urgent type A aortic dissection repair	Survey panel	Intermediate (3) high (5)/ intermediate (3) high (5)	GA*	TT*	IAC* CVD* TEE Cerebral SvO_2	① ④	⑦	ICU*
	GPT	High/intermediate	GA	TT	'Consider IAC and CVC' 'Consider TOE' No cerebral SvO_2	① ④	⑦	ICU
	Adjudication	X[f]	O	O	X[m]	O	O	O

Guidelines

1 Enhanced Recovery After Surgery (ERAS) guideline for open upper abdominal surgery: opioid-sparing analgesia strategies, including regional analgesia techniques, should be implemented in a context of a multimodal analgesic regimen. Thoracic epidural analgesia is strongly advised.[5]

2 Omission of rapid sequence induction and intubation in patients at risk for regurgitation is associated with increased risk for aspiration.[6]

Remarks Adjudication Committee

a Absence of regional or multimodal analgesia strategy was deemed suboptimal

b The postoperative analgesic plan was deemed inadequate for the expected postoperative analgesic requirement

c The placement of a pulmonary artery catheter was deemed excessive and not indicated

d The administration of fentanyl during a rapid sequence induction for a Caesarean is not ideal because of potential respiratory depression of the neonate, however may be justified in case of severe hypertension in pre-eclamptic patient

e The advice to have noradrenaline stand-by at induction for a pre-eclamptic patient with severe hypertension is potentially reflective of flawed reasoning, as the goal should be to avoid aggravation of hypertension.

f Risk estimate by generative pretrained transformer (GPT) was considered inappropriate

h The use of a supraglottic airway device (SAD) in patients presenting with an elevated risk of aspiration was deemed unsafe. Additionally, the administration of rocuronium during induction for the placement of an SAD is uncommon practice

i The placement of a central venous pressure monitoring was deemed excessive and not indicated

k 'The advised prophylaxis against postoperative nausea and vomiting, was deemed not indicated but not necessarily inappropriate'

l The use of an uncuffed tube in this case was deemed outdated

m The response 'an arterial line, central venous catheter, and transoesophageal echocardiography could be considered' was regarded as overly cautious and therefore inappropriate

Anaesthesia technique and airway management

We observed variability in anaesthesia technique amongst anaesthesiologists in the survey panel, either choosing general anaesthesia, neuraxial or regional anaesthesia, or a combination thereof, whereas ChatGPT solely advised general anaesthesia without use of supplemental regional or neuraxial analgesia, or multimodal analgesic options. The adjudication committee regarded the ChatGPT advice subpar in Case 1 (total shoulder prosthesis), Case 2 (transthoracic oesophagectomy), and Case 3 (open gastrectomy). In the first case, the lack of both regional anaesthesia and absence of multimodal analgesic alternatives was estimated to be insufficient. In the latter two cases, ChatGPT actively advised against supplemental neuraxial analgesic techniques in the absence of absolute contraindications, again without advising multimodal alternatives.

Regarding airway management, survey panellists uniformly chose tracheal intubation in all cases requiring general anaesthesia, whereas ChatGPT advised a supraglottic airway device in Case 6 (urgent laparoscopic resection of ectopic pregnancy) and Case 7 (urgent laparoscopic removal of appendix) (Table 1). These responses were considered inappropriate by the adjudication committee because of the elevated risk for pulmonary aspiration of gastric contents. When choosing a supraglottic airway device, ChatGPT also recommended use of a neuromuscular blocker at induction of anaesthesia, which the adjudication committee judged as uncommon practice.

Propofol was the preferred induction hypnotic of the survey participants and of ChatGPT. Propofol was also the preferred maintenance hypnotic by the panellists *vs* a preference of inhalation anaesthetics by ChatGPT (sevoflurane or isoflurane). Similarly, the choice of intraoperative opioid differed between panellists (sufentanil, 64%; fentanyl 24%; remifentanil 12%) and ChatGPT (fentanyl, 70%; remifentanil 30%). Rocuronium was the most selected neuromuscular blocking agent by the panellists and ChatGPT; succinylcholine was chosen only occasionally by a panellist (3%) for rapid sequence induction. The adjudication committee reviewed the medication advice by ChatGPT as generally appropriate.

Monitoring and vasoactive drugs

Both ChatGPT and survey panellists advised standard American Society of Anesthesiologists (ASA) monitoring for all patients, however ChatGPT more frequently opted for invasive monitoring techniques than the survey panellists (Table 1). In Case 6 (urgent laparoscopic resection of ectopic pregnancy; haemodynamic class 1–2 shock), ChatGPT advised use of an arterial catheter and central venous catheter, the latter adjudicated as defensive, but not necessarily wrong. In Case 4 (repeat laparotomy after liver transplantation; no mention of pulmonary artery hypertension), ChatGPT advised to consider inserting a pulmonary artery catheter, which was also adjudicated as very defensive. In Case 10 (urgent repair for type A aortic dissection), ChatGPT also advised to consider inserting an arterial catheter, central venous catheter, and transoesophageal cardiac ultrasound, which was adjudicated as too mild by the committee. ChatGPT did not mention the option of frontal cerebral tissue oximetry in this case. With the exception of one example, when ChatGPT gave the incorrect advice of having norepinephrine readily available in the case of a pre-eclamptic patient with a severe hypertension emergency (Case 4), the advice with regard to vasopressors was largely appropriate.

Postoperative management

Both ChatGPT and the anaesthesiologist of the survey panel advised a combination of paracetamol, an NSAID, and opioid analgesia for postoperative pain relief. ChatGPT predominantly chose patient-controlled analgesia with morphine or hydromorphone, where the panellists predominantly chose subcutaneous morphine or oral oxycodone in cases where no additional neuraxial or regional analgesia was selected. In contrast, less invasive choices by ChatGPT were noted in Case 3 (open gastrectomy), where ChatGPT opted for paracetamol and an oral opioid, and in Case 7 (laparoscopic appendectomy) where it only advised postoperative paracetamol. These postoperative analgesic plans were adjudicated as insufficient. Prescription of antiemetics was mostly congruent between the survey panel and ChatGPT, with ondansetron with or without dexamethasone being advised in most cases. Although the adjudication committee deemed antiemetics not indicated in cases 7 and 8, the antiemetic plans in general were reviewed as appropriate.

Postoperative destination

Postoperative destination was mostly in agreement between the panellists and ChatGPT, except for Case 1 (shoulder prosthesis) where ChatGPT advised a monitored Intensive Care Unit/Medium Care Unit (ICU/MCU) bed *vs* ward discharge by the panellists. The adjudication committee judged the postoperative destination of Case 1 by ChatGPT as very defensive.

Discussion

ChatGPT is a chatbot that leverages a multimodal LLM to generate general purpose language. Since the introduction of ChatGPT, rapid growth and adoption by the public, including applications in medicine, has been seen. We assessed the proficiency of ChatGPT-4.0 in producing preoperative anaesthetic plans for 10 clinical cases of varying medical complexity. These plans were compared with plans produced by a panel of eight experienced anaesthesiologists and qualitatively assessed by an adjudication committee. In general, plans generated by ChatGPT were relatively coherent and reasonably adequate on many clinical domains, but failed to consistently produce flawless plans in all domains; evidence of flawed reasoning or some inadequacy of plans were adjudicated to be present in all cases. In addition, ChatGPT responses were repetitive and lacked the variety expected in clinical practice for such cases, as evidenced by a strong preference to advise general anaesthesia, and an absence of neuraxial or regional anaesthetic options. This suggests that plans formulated by ChatGPT are unlikely to be the result of sound clinical reasoning.

Disagreements between plans generated by ChatGPT and plans generated by the survey panel or review by the adjudication committee were not always considered clinically important. For example, choices of hypnotic or opioid agents, choices for invasive or noninvasive monitoring options and advice for antiemetic prophylaxis by ChatGPT were mostly judged to be adequate by the adjudication committee. Notable differences were seen in cardiac and pulmonary risk assessment. In part, discrepancies observed between ChatGPT, the survey panel, and the adjudication committee can be explained by the various risk estimation tools that are available for this purpose, yielding varying risk estimates. In

addition, differences between qualitative and quantitative testing in this study might also have contributed to this finding. However, plans in which the recommendations made by ChatGPT diverged from accepted clinical standards or guidelines were concerning. For instance, ChatGPT failed to follow Enhanced Recovery After Surgery (ERAS) recommendations for postoperative epidural analgesia in patients scheduled for upper abdominal surgery,[5] even actively advising against its use in these cases, in the absence of contraindications and without providing alternative multimodal analgesia alternatives. In addition, the ChatGPT plan for postoperative analgesia for these specific cases did not contain advice for patient-controlled analgesia with opioids, which would have been appropriate. Finally, in two other cases, ChatGPT failed to advise tracheal intubation for patients at increased risk for pulmonary aspiration. In the 4th National Audit Report (NAP4),[6] omission of rapid sequence induction and intubation in patients at risk was recognised as a risk for patient harm or death from aspiration. As such, anaesthesia caregivers risk providing substandard patient care if they blindly adhere to instructions from ChatGPT.

To the best of our knowledge, this is the first study to assess the utility of generative artificial intelligence in language-based tasks in clinical anaesthesia. However, use of LLMs in nonclinical anaesthesia queries has been studied. For instance, ChatGPT-3 was able to answer 40–60% of questions of the anaesthesia exam of the Royal College of Anaesthetists correctly.[7] Similar results were observed with ChatGPT-4 for the American Board of Anesthesiology exam (80% correct answers).[8] Although promising, both studies noted limitations of these models, including the limited and repetitive nature of generated answers and incorrect handling of basic arithmetic, leading to errors and imprecise responses.[9] Our study supports these findings and shows that ChatGPT-4 should currently not be used for assistance in preoperative anaesthetic planning.

There are several limitations to consider. Firstly, this study was conducted in a single tertiary centre in the Netherlands. We are aware of institutional preferences and potential bias for certain medications or approaches. However, we feel that our general finding that ChatGPT-4 should currently not be used for clinical assistance is based on universal standards of good clinical practice that are not consistently met by ChatGPT. Nevertheless, as new models are developed, rapid advancements in this area could lead to improvements in clinical reasoning by future models.

Importantly, this study did not assess the reasoning behind specific choices by ChatGPT-4, nor did we allow for follow-up questions. Studies show that in follow-up questions, ChatGPT-4 can elaborate on its choices and the reasoning behind it.[7,8] Yet, expressing disagreement or scepticism in follow-up questions can lead to wavering in favour of incorrect answers.[10] However, we wanted to examine the out-of-the-box capabilities without priming of ChatGPT-4.

Secondly, ChatGPT has a factor of randomness controlled by a parameter called 'temperature' that is built in the model. The temperature ranges from 0 to 1 and governs the degree of randomness in predicting the next word when generating text, where 0 will generate deterministic results and 1 generates creative but unpredictable results. This degree of randomness is introduced to generate more interesting, lively, and anthropomorphic results, thus entering the same prompt multiple times will generate different responses. We used ChatGPT-4.0, which has the default temperature set to 0.7. Not surprisingly, when we re-entered one of the cases, ChatGPT would generate a complete opposite approach. Unfortunately, the temperature parameter was only adjustable through an application programming interface (API) to which the access was limited by OpenAI because of the overwhelming demand. In addition, accessing a large language model (LLM) through an API is not the common case for the average clinician seeking assistance from an LLM, as this requires specific technical knowledge. Nonetheless, further study of the influence of temperature on the consistency and reliability of responses generated for clinical use is warranted.[11]

Finally, we raise the issue that the source of training (i.e. the text corpus on which ChatGPT is trained), might influence outcome by various mechanisms. ChatGPT-4 generated preoperative anaesthetic plans that could be correct in general but lacked case-specific risk–benefit considerations that, in atypical cases, can lead to valid decisions by physicians that deviate from current guidelines or literature. The second mechanism is that literature published by anaesthesia research groups defines the text-corpus, and thus a selection bias might occur in the literature that is included in the training corpus. Additionally, it is unknown whether the model was trained on real-world clinical notes, and thus whether use of colloquial medical language was covered. The third mechanism involves the relevance of the training data, leading to either outdated recommendations or recommendations blinded to new developments in perioperative medicine. For example, ChatGPT-4.0 recommended use of an uncuffed tube in a paediatric anaesthesia case. Finally, prompt engineering is critical to the performance of LLMs and the validity of the results, so optimisation of prompting strategies needs further study.

In conclusion, this exploratory single-centre mixed-method study qualitatively and quantitatively evaluated the capacity of ChatGPT-4.0 to generate preoperative anaesthetic plans for a variety of clinical cases, with a focus on preoperative risk evaluation, anaesthetic management and monitoring, and postoperative care. Although generally coherent, the majority of preoperative plans fell short in one or more areas, and occasionally failed to meet minimum standards of good clinical care. In addition, plans by ChatGPT-4.0 were repetitive in nature and lacked variety, suggesting that they were unlikely to be the product of clinical reasoning. Future LLMs trained on anaesthesia-specific datasets should have improved performance, but should be subjected to rigorous evaluation before being used for assistance in clinical practice.

Authors' contributions

Study conception and planning: MAM, MvV, MB
Data acquisition: MAM, MvV
Data analysis and interpretation: all authors.
All authors participated in the revision of the manuscript, gave final approval of the version to be published, and agreed to be accountable for all aspects of the work thereby ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

We thank the following anaesthesiologists who have contributed to the online survey of cases:
J. van der Bos, V. Chopra, H. Helmerhorst, T. Holl, M. Niesters, P. Okkerse, M. Schooneveldt, and R. van der Steeg.

Declaration of interest

MAM is founder of medical data platform Delphyr, which is unrelated to this work. All other authors declare that they have no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bja.2024.08.038.

References

1. Meara JG, Leather AJM, Hagander L, et al. Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. *Lancet* 2015; **386**: 569–624
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023; **29**: 1930–40
3. Bivard A, Churilov L, Parsons M. Artificial intelligence for decision support in acute stroke — current roles and potential. *Nat Rev Neurol* 2020; **16**: 575–85
4. Tyler NS, Mosquera-Lopez CM, Wilson LM, et al. An artificial intelligence decision support system for the management of type 1 diabetes. *Nat Metab* 2020; **2**: 612–9
5. Feldheiser A, Aziz O, Baldini G, et al. Enhanced Recovery after Surgery (ERAS) for gastrointestinal surgery, part 2: consensus statement for anaesthesia practice. *Acta Anaesthesiol Scand* 2016; **60**: 289–334
6. Cook TM, Woodall N, Frerk C. Fourth national Audit project. Major complications of airway management in the UK: results of the fourth national Audit project of the royal College of Anaesthetists and the difficult airway society. Part 1: anaesthesia. *Br J Anaesth* 2011; **106**: 617–31
7. Aldridge MJ, Penders R. Artificial intelligence and anaesthesia examinations: exploring ChatGPT as a prelude to the future. *Br J Anaesth* 2023; **131**: e36–7
8. Shay D, Kumar B, Bellamy D, et al. Assessment of ChatGPT success with specialty medical knowledge using anaesthesiology board examination practice questions. *Br J Anaesth* 2023; **131**: e31–4
9. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023; **388**: 1233–9
10. Xie Q, Wang Z, Feng Y, Xia R. Ask again, then fail: large language models' vacillations in judgement. *[Preprint.] arXiv Adv Access* 2024. https://doi.org/10.48550/arXiv.2310.02174. published on 11 June
11. Perlis RH, Fihn SD. Evaluating the application of large language models in clinical research contexts. *JAMA Netw Open* 2023; **6**, e2335924

Handling Editor: Hugh C Hemmings Jr