

Review Article

Generative Artificial Intelligence for Clinical Anesthesia Decision Support

Petr Novak^{1*}, Pieter Berg² and Eszter Halvorsen³

¹Division of Anesthesia and Intensive Care, Jutland University of Clinical Medicine, Denmark

²Department of Anesthesiology, Iberian University Hospital and Medical School, Spain

³Department of Perioperative Medicine, Jutland University of Clinical Medicine, Denmark

ABSTRACT

Generative artificial intelligence, built on large language model architectures, has moved rapidly from novelty into systematic evaluation across preoperative, intraoperative, and postoperative anesthesia practice, offering a fundamentally different capability than the discriminative machine-learning models anesthesiology has used for hemodynamic and risk prediction over the past decade.

This review synthesizes the emerging literature on generative AI applications in clinical anesthesia decision support, including preoperative anesthetic plan drafting and risk stratification, patient-facing communication, intraoperative documentation support, and simulated case generation for trainee education.

Early case-control and randomized evaluations show that generative models can produce plausible, often accurate preoperative plans and explanations, and can generate measurable efficiency gains in perioperative documentation, but performance is inconsistent for complex comorbidity profiles and numerical dosing tasks, and equity and bias evaluation remains largely aspirational.

We argue that the more rigorously validated discriminative models, particularly for intraoperative hypotension prediction, remain the appropriate tool for narrow numerical tasks, while generative AI's clearest near-term value lies in documentation, communication, and education rather than autonomous clinical recommendation.

We conclude by outlining the evidentiary infrastructure, including multi-center prospective validation and demographic-stratified performance reporting, that the field still needs before generative AI can be safely integrated as more than an assistant to an accountable clinician.

Keywords: generative artificial intelligence; large language models; anesthesia decision support; perioperative medicine; machine learning

Citation: Novak P, Berg P and Halvorsen E (2026) Generative Artificial Intelligence for Clinical Anesthesia Decision Support. *Anesthesiology*. 6(1):01-05.

DOI: 10.17140/2026.vol6.i01.pp01-05

Corresponding Author: Petr Novak

E-mail Address: p.novak@jucm.dk

Article History: Received 08 January 2026 | Accepted 29 January 2026 | Published 24 February 2026

Copyright: © 2026 Petr Novak, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (<https://creativecommons.org/licenses/by/4.0/>).

INTRODUCTION

A patient shows up for a preoperative clinic visit with

three prior anesthetics, a stack of cardiology notes, and a medication list that has changed twice in the past month. The anesthesiologist reviewing the chart has perhaps fifteen minutes to synthesize all of it into a plan. It is exactly this kind of high-volume, document-heavy, time-pressured synthesis task that has drawn the attention of generative artificial intelligence developers to perioperative medicine. Unlike the discriminative machine-learning models that anesthesiology adopted over the past decade — models trained to output a single number, such as a probability of hypotension or a risk score for postoperative delirium — generative models built on large language model (LLM) architectures produce free text, and sometimes free text plus reasoning, in response to open-ended prompts. That capability changes what the technology can plausibly be asked to do: draft an anesthetic plan, summarize a chart, answer a patient's question in plain language, or generate simulated clinical vignettes for trainees, rather than simply flag a threshold crossing on a monitor.

Interest in generative AI for anesthesia has moved quickly from novelty to systematic evaluation. Narrative reviews now describe a broad landscape of proposed and tested applications spanning preoperative assessment, intraoperative support, postoperative care, and medical education [1] [2] [3], and scoping review protocols have been registered to track the literature as it accumulates [4]. At the same time, the discriminative machine-learning tradition in anesthesiology — hemodynamic prediction models being the most mature example — has continued to develop in parallel, and increasingly the two strands are discussed together as complementary rather than competing technologies [5]. This review surveys where generative AI currently sits in clinical anesthesia decision support: what has been tested, what the early evidence shows about accuracy and safety, where it plausibly adds value alongside established predictive models, and what stands between the current experimental literature and dependable clinical deployment.

WHAT GENERATIVE AI ADDS BEYOND DISCRIMINATIVE PREDICTION MODELS

Discriminative models in anesthesiology are narrow by design. The intraoperative hypotension literature illustrates this well: models such as the commercial Hypotension Prediction Index and various academic deep-learning alternatives are trained on high-resolution arterial waveform features to output a probability that mean arterial pressure will fall below a threshold within the next several minutes [6] [7] [8]. A meta-analysis pooling this literature found that machine-learning approaches generally outperform simple hemodynamic thresholds for anticipating intraoperative hypotension, though performance varies considerably by algorithm and validation cohort, and

calibration in external populations is frequently worse than in the derivation set [9]. These are useful, actionable outputs, but they answer one question at a time and require a specific, structured data stream to function.

Generative models are different in kind. Built on transformer architectures trained on enormous text corpora, they accept unstructured input — a paragraph of history, a stack of consult notes, a patient's own question typed in plain language — and produce unstructured output tailored to that input. This flexibility is precisely why the anesthesia literature has begun testing them for tasks that resist a single numeric output: drafting a preoperative anesthetic plan that weighs multiple comorbidities against a proposed procedure, explaining a regional anesthesia technique to an anxious patient, or summarizing a lengthy electronic health record into a one-paragraph handoff. The tradeoff, discussed at length in recent explainability-focused reviews, is that generative outputs are harder to audit than a probability score: a hypotension model's calibration curve can be plotted, but a paragraph of generated clinical reasoning has to be read and judged, case by case, for factual accuracy and internal consistency [5] [10].

PREOPERATIVE ASSESSMENT AND ANESTHETIC PLANNING

The preoperative encounter has become the most heavily studied application. In a case-control evaluation of 150 simulated patient presentations, ChatGPT was tasked with performing day-surgery suitability and pre-anesthesia risk assessment; the model performed creditably on straightforward cases but showed more variable judgment as comorbidity complexity increased, and its recommendations required anesthesiologist verification rather than being usable unreviewed [11]. A separate mixed-method study asked GPT-4.0 to generate full preoperative anesthetic plans and had anesthesiologists rate them against plans written by clinicians; the generated plans were often judged complete and reasonably well-organized, but graders identified omissions and occasional overconfident recommendations that would have gone uncorrected without expert review [12].

Cardiac risk stratification before noncardiac surgery is a natural stress test for this technology because it forces the model to integrate guideline-based reasoning (functional capacity, revised cardiac risk index criteria, biomarker thresholds) with patient-specific detail. An exploratory study of large language models applied to preoperative assessment and planning for cardiac patients undergoing noncardiac surgery found the models capable of reproducing much of the standard risk-stratification logic, but noted gaps in handling ambiguous or borderline functional-capacity data — exactly the scenarios where clinician judgment matters most [13]. Equity-focused analyses have pushed the discussion

further, arguing that if LLM-drafted plans are adopted into practice, they need explicit evaluation for whether recommendation quality varies across patient demographics and language proficiency, since training data biases can be reproduced or amplified in generated text in ways that are not obvious from a single case review [14].

Beyond planning, generative models are also being tested as patient-facing communication tools during preoperative counseling. A multi-rater quality assessment of ChatGPT-4 as a digital patient-education resource for anesthesia found the generated explanations generally accurate and readable, though not uniformly calibrated to the reading level or emotional register that a real preoperative conversation demands [15], and a parallel study asking anesthesiologists to grade ChatGPT's answers to patients' preoperative questions found similarly mixed marks — high on correctness for common questions, more uneven on completeness for nuanced or less-common ones [16].

INTRAOPERATIVE AND POSTOPERATIVE APPLICATIONS

Intraoperative use of generative AI remains far less mature than preoperative use, in large part because the intraoperative environment already has a mature discriminative-model ecosystem built around continuous physiologic waveforms rather than free text. The hemodynamic-prediction literature described above continues to be the dominant intraoperative decision-support tool in practice [6] [7]. Where generative AI has been proposed intraoperatively, it tends to appear in a supporting role: summarizing case events for handoff documentation, drafting operative or anesthesia records from structured monitor data, or acting as a natural-language interface layered on top of an existing prediction model so that a probability output is translated into a plain-language recommendation a clinician can act on quickly [10] [1].

Postoperatively and across the broader perioperative pathway, a randomized crossover trial evaluated the clinical and economic impact of deploying a large language model to support perioperative medicine tasks and found measurable efficiency gains — less time spent on documentation and information synthesis — without a demonstrated increase in adverse outcomes over the trial period, though the authors were careful to note that the trial was not powered to detect rare safety events [17]. This is one of the few controlled trials in the space rather than a retrospective or simulated-case evaluation, and it illustrates both the promise (measurable time savings on cognitively demanding documentation tasks) and the current limits of the evidence base (small trials, short follow-up, outcome measures oriented toward efficiency rather than long-term clinical endpoints).

SPECIAL POPULATIONS: OBSTETRIC AND PEDIATRIC ANESTHESIA

Subspecialty literature has begun to look specifically at populations where the stakes of a generic, population-trained model are amplified by physiologic distinctiveness. A review of artificial intelligence and decision-support tools in obstetric anesthesia catalogs opportunities — triage support for peripartum hemorrhage risk, drafting of neuraxial anesthesia counseling material, synthesis of complex antenatal records — while cautioning that pregnancy-specific physiology and the fetal consideration are poorly represented in the general medical text corpora most LLMs are trained on, raising the risk that generated recommendations default to non-pregnant reasoning unless specifically prompted otherwise [18].

Pediatric anesthesia raises an analogous concern from a different angle: weight-based dosing, developmental physiology, and family communication needs that differ qualitatively from adult practice. A review of generative AI applications in pediatric anesthesia describes uses in trainee education (generating realistic simulated cases across the full range of pediatric ages and comorbidities), in family-facing communication (translating a proposed anesthetic plan into age-appropriate, reassurance-oriented language for parents), and cautiously in clinical support, while emphasizing that weight-based and age-based dosing calculations generated by a language model need independent verification against a validated source before use, since numerical reasoning is a documented weak point of transformer-based text generators [19].

EDUCATION, ETHICS, AND PROFESSIONAL IDENTITY

A recurring theme across the literature is that the clearest near-term value of generative AI in anesthesiology may lie in education rather than direct patient care. Because LLMs can generate an essentially unlimited supply of varied, realistic clinical vignettes on demand, they are well suited to producing practice questions, simulated preoperative interviews, and case discussions for residents and student registrars — uses that carry much lower risk than direct clinical deployment because a generated training case that contains a subtle error can be corrected by a supervising educator before it does harm [20] [19]. This same body of work argues that generative AI's arrival changes what anesthesiology training programs need to teach: alongside airway management and regional technique, trainees increasingly need to be taught how to verify, and when to distrust, an AI-generated recommendation, since uncritical acceptance of a fluent but wrong answer is arguably a more dangerous failure mode than the model being wrong in the first place [20].

Ethical and professional-identity questions follow

directly from that observation. Reviews addressing ethical integration of generative AI into perioperative practice converge on a small number of recurring concerns: the risk of automation bias, where a clinician defers to a fluent AI-generated plan without independently verifying it; the opacity of how a given model arrived at a recommendation, which complicates informed consent when patients ask how a plan was derived; data privacy when patient information is entered into commercial LLM interfaces that may retain or use that data for further training; and the medicolegal question of where responsibility sits when a clinician acts on an AI-generated recommendation that later proves wrong [20] [14] [5]. None of these concerns are unique to anesthesiology, but the combination of high acuity, time pressure, and reliance on rapid synthesis of complex data makes anesthesia practice an unusually exacting test case for them.

TOWARD DEPLOYMENT: WHAT THE EVIDENCE STILL LACKS

Read together, the current literature supports a fairly specific conclusion: generative AI is genuinely useful today for tasks that are documentation-heavy, time-consuming, and forgiving of a verification step — drafting, summarizing, translating clinical information into patient-friendly language, and generating educational content — while it is not yet validated for autonomous clinical decision-making in the way that, say, a calibrated hypotension-prediction algorithm has been validated for a specific, narrow warning task. Several structural gaps explain why. Most published evaluations are small, single-center, and use simulated rather than real patient encounters, which limits confidence that reported performance will generalize [4]. Randomized evidence is sparse; the crossover trial of LLM deployment in perioperative medicine is a notable exception, but a single trial with efficiency-oriented endpoints cannot settle questions about rare-event safety [17]. Equity and bias evaluation is still mostly aspirational rather than empirical, with reviews calling for demographic-stratified performance analysis that has not yet been widely reported for anesthesia-specific tasks [14]. And integration with the existing discriminative-model ecosystem — how a generative interface should sit alongside a validated hemodynamic-prediction algorithm rather than replace it — remains largely undescribed in the literature, even though combining the two seems like the more plausible deployment path than either technology working in isolation [5] [10].

Closing these gaps will likely require the same kind of infrastructure that the discriminative-model literature took a decade to build: multi-center prospective validation, standardized reporting of accuracy and failure modes stratified by patient subgroup, and clear institutional policy on what a clinician is expected to independently verify before acting on a generated

recommendation. Until that infrastructure exists, the responsible framing for generative AI in anesthesia practice is as an assistant to a clinician who remains accountable for the final decision, not as a decision-maker in its own right.

CONCLUSION

Generative artificial intelligence has moved into anesthesiology faster than the evidence needed to support routine clinical use has been generated, which is not unusual for a disruptive technology but does require honesty about where the field currently stands. The strongest current evidence favors documentation, summarization, patient communication, and education — tasks where a fluent but occasionally wrong output can be caught by a human reviewer before it reaches a patient. The weakest current evidence, and the area demanding the most caution, is autonomous or semi-autonomous clinical recommendation in complex, high-stakes scenarios such as cardiac risk stratification, obstetric emergencies, or pediatric dosing, where the training data underlying general-purpose language models does not obviously capture the physiologic nuance those situations require. The discriminative machine-learning models anesthesiology already relies on, particularly for intraoperative hemodynamic prediction, remain the more rigorously validated technology for the narrow tasks they were built for, and the most plausible near-term path forward pairs the two: a validated predictive model doing the narrow numerical work, and a generative layer translating that output, along with the rest of the clinical context, into language a busy clinician can act on quickly. Realizing that combination safely will depend on the field building the same evidentiary rigor around generative tools — prospective validation, bias auditing, and clear rules for human verification — that it has spent the last decade building around predictive ones.

DECLARATIONS

Acknowledgements The authors thank their respective departments for administrative support during the preparation of this manuscript.

Funding This review received no external funding.

Conflict of Interest The authors declare no conflicts of interest.

Author Contributions PN conceived the review and drafted the initial manuscript. PB performed the literature synthesis and critically revised the manuscript for intellectual content. EH contributed to manuscript structure, verified references, and approved the final version. All authors read and approved the final manuscript.

Ethics Statement This is a literature review and did not involve new human or animal data; no ethics approval was required.

Data Availability No new data were generated or analyzed for this review. All sources discussed are publicly available in the cited literature.

REFERENCES

1. Anesthesia transformed: AI pioneering a new era in perioperative medicine (2025) *Anesthesiology and Perioperative Science*, DOI:10.1007/s44254-025-00091-9.
2. Applications and challenges of large language models in anesthesiology: narrative review and future perspectives (2025) *Anesthesiology and Perioperative Science*, DOI:10.1007/s44254-025-00156-9.
3. Large language models in perioperative medicine: applications and future prospects: a narrative review (2025) *Canadian Journal of Anesthesia*, DOI:10.1007/s12630-025-02980-w.
4. Large language models and their current use in perioperative medicine: a scoping review protocol (2025) PMID:41017313.
5. Explainable artificial intelligence in anesthesiology prediction models: bridging the gap from black-box algorithms to clinical decision support (2026) *Frontiers in Medicine*.
6. Deep learning models for the prediction of intraoperative hypotension (2021) *British Journal of Anaesthesia*, PMID:33558051.
7. Intraoperative Hypotension Prediction Based on Features Automatically Generated Within an Interpretable Deep Learning Model (2023) *IEEE Journal of Biomedical and Health Informatics*, PMID:37220057.
8. Development of a prediction model for hypotension after induction of anesthesia using machine learning (2020) PMID:32298292.
9. Predictive ability of hypotension prediction index and machine learning methods in intraoperative hypotension: a systematic review and meta-analysis (2024) PMC.
10. Generative AI models: the next anaesthetic agent? (2026) PMC.
11. The performance of ChatGPT in day surgery and pre-anesthesia risk assessment: a case-control study of 150 simulated patient presentations (2024) *Perioperative Medicine*, PMID:39574189.
12. Generation of preoperative anaesthetic plans by ChatGPT-4.0: a mixed-method study (2024) *British Journal of Anaesthesia*.
13. An exploratory study of large language models in preoperative anesthesia assessment and planning for cardiac patients undergoing noncardiac surgery (2026) PMC.
14. Applying Large Language Models in Perioperative Medicine With an Equity-Informed Perspective (2026) *Anesthesia & Analgesia*, PMC:PMC13267585.
15. Evaluating ChatGPT-4 as a digital patient education tool in anesthesia: A multi-rater quality assessment (2026) DOI:10.1177/20552076261420876.
16. Anesthesiologists' Assessment of the Correctness, Completeness, and Coherence of ChatGPT Answers to Patient Preoperative Questions (2026) PMC.
17. Clinical and economic impact of a large language model in perioperative medicine: a randomized crossover trial (2025) *npj Digital Medicine*, PMC:PMC12280117.
18. Artificial intelligence and decision support tools in obstetric anaesthesia: opportunities, challenges, future (2026) *Frontiers in Anesthesiology*, DOI:10.3389/fanes.2026.1893965.
19. Harnessing Generative Artificial Intelligence in Pediatric Anesthesia: Enhancing Learning, Patient Care, and Family Communication (2025) PMC (PMC12343195).
20. Generative AI in perioperative medicine and anesthesiology: ethical integration, educational innovation, and the future of clinical professionalism (2025) *Anesthesia & Analgesia* (perspective).