Title

Applying Large Language Models in Perioperative Medicine With an Equity-Informed Perspective

Journal

Anesthesia & Analgesia, 143(3)

ISSN

0003-2999

Authors

Burton, Brittany N
Canales, Cecilia
Chang, En
et al.

Publication Date

2026-04-29

DOI

10.1213/ane.0000000000008017

Peer reviewed

HHS Public Access

Author manuscript

Anesth Analg. Author manuscript; available in PMC 2026 June 15.

Applying Large Language Models in Perioperative Medicine With an Equity-Informed Perspective

Brittany N. Burton, MD, MHS, MAS[1], **Cecilia Canales, MD, MS**[1], **En Chang, BS**[2], **Austin Du, MD**[1], **Caitlin Sebastian, BS**[3], **Rodney A. Gabriel, MD, MAS**[4,5]

[1]Department of Anesthesiology & Perioperative Medicine, University of California, Los Angeles, California

[2]School of Medicine, California University of Science and Medicine, Colton, California

[3]David Geffen School of Medicine, University of California, Los Angeles, California

[4]Division of Perioperative Informatics, Department of Anesthesiology, University of California, San Diego, La Jolla, California

[5]Division of Biomedical Informatics, Department of Medicine, University of California, San Diego, La Jolla, California

Large language models (LLMs) are transformer-based artificial intelligence systems capable of representing and generating natural language at scale.[1] Developments accelerated in 2020 with OpenAI's GPT-3, an LLM trained on hundreds of billions of parameters and notable for its capacity for complex language generation and reasoning.[1] Subsequent iterations, including GPT-4 and other contemporary LLMs (eg, Gemini, Llama, Mistral, and DeepSeek), have demonstrated improved contextual understanding and early applications in biomedical and clinical domains.[1] Although subsequent advances have enabled sophisticated biomedical applications, the use of LLMs in clinical medicine, particularly in perioperative settings, remains in an early translational phase. To date, most clinical deployments represent feasibility or pilot studies with limited prospective validation, external generalizability, or evidence of impact on patient outcomes. Nevertheless, emerging evidence suggests that task-specific biomedical LLMs can meaningfully augment *select* clinical functions. For example, adapted clinical LLMs have demonstrated superior performance compared with clinicians in structured summarization of electronic health record (EHR) documentation, with completeness and fewer clinically relevant omissions.[2] In the perioperative environment, LLM-based models applied to unstructured preoperative notes have outperformed traditional natural language processing approaches for postoperative risk prediction, identifying additional patients who may not be captured by standard models.[3]

Despite their strengths, these models inherit biases embedded in human-generated datasets, reflecting structural inequities, language bias, and healthcare disparities. As a result, their

Address correspondence to Brittany N. Burton, MD, MHS, MAS, Department of Anesthesiology & Perioperative Medicine, University of California, Los Angeles, 757 Westwood Plaza, Los Angeles, CA. BBurton@mednet.ucla.edu.

Conflicts of Interest: None.

integration into perioperative medicine presents both an opportunity to improve healthcare delivery and a risk of reinforcing existing inequities. To advance health equity, particularly in perioperative settings where disparities in outcomes, access, and communication persist, LLMs must be evaluated for both technical performance and alignment with equitable patient care objectives. This commentary examines the role of LLMs in perioperative medicine across three domains—clinical workflow optimization, patient engagement and communication, and research acceleration. It also examines limitations and ethical considerations surrounding their responsible use. We further highlight the importance of data curation to ensure these tools serve diverse populations and promote equitable outcomes.

WORKFLOW OPTIMIZATION IN PERIOPERATIVE MEDICINE

Healthcare-focused LLMs trained on biomedical data have primarily been applied to optimize clinical workflows, which include supporting preoperative decision-making, information extraction from unstructured records, risk stratification, and predictive modeling. These capabilities reduce manual effort, improve standardization, and support quality improvement initiatives across clinical settings.[1] LLM-based extraction and summarization of clinical data can meaningfully enhance clinical documentation quality, analytic performance for risk stratification, and workflow efficiency.[1,2] One perioperative-specific study that illustrates early feasibility of LLM-supported workflows is the implementation of the Perioperative AI Chatbot (PEACH).[4] Its primary objective was to support preoperative decision-making by providing guideline-consistent recommendations to care teams. PEACH demonstrated strong performance, achieving 96.7% overall accuracy that increased to 97.9% after a system update. The model showed only 1 to 2 deviations across 240 cases, improved clinical decision-making in 95% of cases, and demonstrated high inter-rater reliability. Subgroup analysis showed documentation time savings of 5.8 minutes per moderate-complexity case, and 4.6 minutes for experienced clinicians.[4] Although these findings demonstrate efficiency gains, future iterations of tools like PEACH should assess whether these time savings and decision-support benefits are equitably distributed across patient populations. PEACH was developed and evaluated within the Singapore healthcare system, where national data governance structures and regulatory frameworks differ from those in the United States. Therefore, translation of similar LLM-enabled perioperative tools into US clinical practice would require alignment with US regulatory requirements governing protected health information, including compliance with the Health Insurance Portability and Accountability Act (HIPAA), institutional data use agreements, and health system-specific information security policies.[5] These considerations underscore that, beyond technical feasibility, privacy regulation and data governance are vital determinants of scalability and adoption of LLM-based clinical tools within US perioperative settings.

Recent work by Alba et al analyzed approximately 85,000 preoperative clinical notes and demonstrated that LLM-based models outperformed traditional natural language processing methods in predicting common postoperative complications, such as acute kidney injury and pneumonia, identifying up to 39 additional patients at risk per 100 beyond those identified by standard approaches.[3] Similarly, Chung et al assessed the performance of a general-domain LLM, GPT-4 Turbo, across eight perioperative prediction tasks using retrospective EHR data, with ≥1000 cases per prediction task.[6] These tasks included American Society of

Anesthesiologists (ASA) physical status classification, intensive care unit (ICU) and hospital admission, inpatient mortality, and predictions of length of stay for ICU and hospital admission. The model achieved F1 scores of 0.81 for ICU admission and 0.86 for inpatient mortality, indicating reasonable accuracy for risk stratification. However, duration-based predictions were less reliable, with mean absolute errors of approximately 4.5 days for hospitalization and 1.1 days for ICU stay, which suggests limitations of general-purpose LLM in timeline prediction and discharge planning.[6]

Emerging implementation efforts have focused on preoperative documentation support, structured risk stratification from unstructured clinical notes, and guideline-consistent decision assistance within anesthesia preoperative workflows.[3,4,6] These efforts have been advanced through small-scale pilot studies embedded within real clinical environments and supported by multidisciplinary stakeholder engagement, including anesthesiology, perioperative informatics, and health system leadership. Successful deployments have relied on human-in-the-loop designs, institutional governance structures, and prospective evaluation of safety and usability, rather than autonomous model operation.[4,5,7] These approaches represent critical steps for accelerating adoption while preserving clinician oversight, enabling iterative refinement, and laying the groundwork for future evaluation of equity-focused outcomes across diverse perioperative populations. LLMs are most likely to be adopted with minimal workflow disruption when implemented as embedded, assistive tools within existing EHR processes (eg, documentation support or preoperative risk summarization), rather than as autonomous decision-makers.[1,4,5]

PATIENT ENGAGEMENT AND COMMUNICATION

LLMs may improve patient experience by improving access to reliable, tailored medical information and reducing psychological discomfort. In a study evaluating LLM responses to radiation oncology patient care questions, the LLM output matched or exceeded expert answers in 94% of responses; however, the authors also identified a small number of potentially harmful responses, highlighting important safety considerations.[8] Although patient education materials are recommended to be written at a sixth-grade reading level, several medical associations report that existing patient-facing information materials exceed this threshold.[9] LLMs can assist in improving the readability and accessibility of such materials.[9]

Beyond readability, LLMs show promise in clinical communication. In an evaluation of ChatGPT's cardiovascular prevention advice, cardiologists assessed responses against established clinical guidelines and judged 84% to be appropriate for patient-facing communication, supporting the potential role of LLMs as drafting or communication support tools rather than autonomous decision-makers.[10] LLMs like GPT-4 have also demonstrated utility in translating complex pathology reports to patients.[11] LLMs are capable of providing not just accurate information, but also empathetic responses. ChatGPT has outperformed humans in emotional awareness interactions and rapidly improved to near-perfect ratings on validated measures.[12]

The expansion of patient portal messaging, while associated with improved survival,[13] has increased clinician workload and burnout.[14] LLMs offer a compelling alternative to physician- or nurse-managed inboxes. ChatGPT has proven effective at triaging portal messages in nephrology clinics, correctly triaging 93% of the patient messages and only under-estimating the priority of 4% of messages.[15] In some settings, LLMs have demonstrated performance comparable to or exceeding nurse-led communication.[16] A site-specific prompt engineering chatbot (SSPEC) deployed in medical center reception sites resolved patient queries more efficiently than nurse-led communication with higher patient satisfaction and reduced negative patient emotions.[16] In the perioperative setting, LLMs may streamline patient check-ins, improve understanding and adherence to pre- and postoperative instructions, and potentially decrease cancellations and readmissions.[17,18]

RESEARCH ACCELERATION

LLMs may support equity-focused perioperative research by easing traditional barriers of time, expertise, and resource constraints. Although inequity in perioperative outcomes across race, ethnicity, socioeconomic status, and language is well documented, investigating these inequities often requires integrating heterogenous data sources and applying theory-informed, context-sensitive analytic approaches.[19] LLMs can streamline processes such as synthesizing conceptual frameworks, extracting relevant information from the EHR, conducting literature reviews, and formatting manuscripts for publications. By assisting with these tasks, LLMs can help researchers navigate the nuanced and equity-sensitive dimensions of perioperative disparities research.

LLMs have now evolved to conduct sophisticated analyses at scale. For instance, ChatGPT-4 summarized 350 medical research abstracts in about 15 minutes and produced summaries that were 70% shorter than the original manuscript, while maintaining high quality, accuracy, and minimal bias on a blinded evaluation.[20] In another pilot study, ChatGPT-4 organized 113 primary papers into a 6000-word thematic review with tables and figures in about 25 minutes, achieving 85% coverage accuracy and plagiarism of <2%.[21] Although these outcomes reflect the potential to update perioperative equity review, they do not account for the time and expertise required for prompt development, iterative refinement, human validation and error correction. Prompt development itself represents a resource-intensive component of LLM deployment, requiring domain expertise, iterative testing, and careful constraint design.[1,21,22] Prompt framing can influence model outputs, including which clinical factors are emphasized or omitted, and may introduce bias if underlying assumptions are not examined. Unequal access to prompt engineering expertise may therefore be an additional source of inequity in LLM-enabled research and clinical workflows. These complexities highlight that although LLMs can enhance human workflows, they are not yet capable of functioning independently.

Recent studies have also demonstrated use of LLMs in automating components of systematic reviews. In a study of 3230 citations across a seven-language dataset, a GPT-4 pipeline matched human accuracy in title and abstract screening, full-text review, and data extraction.[22] With engineered prompts, inter-rater agreement for full-text screening reached Cohen's $\kappa > 0.80$.[22] Automation can accelerate systematic reviews and broaden inclusivity,

but decisions on complex issues of equity, bias, and representation still require human oversight to avoid embedding bias or overlooking subgroup disparities.

LLM-generated feedback can significantly improve the clarity, syntax, and coherence of scientific writing.[23] A large-scale evaluation of GPT-4's peer-review-style commentary demonstrated concordance with human reviewer feedback 30% to 39% of cases, comparable to human–human concordance. Over half of the authors surveyed rated the artificial intelligence critiques helpful or very helpful.[24] These findings suggest the potential of LLMs in supporting grant writing, institutional review board protocols, and manuscript preparation. Despite producing coherent text, LLMs may generate outputs that contain factual inaccuracies, hallucinations, or embedded biases, highlighting the necessity of human validation. Equity-focused integration requires acknowledgment of these limitations, deliberate mitigation strategies, and sustained human over-sight. Professional societies and leading biomedical journals have issued formal policies governing the use of LLMs in research and publication processes, generally allowing limited use for language editing or drafting assistance while prohibiting LLMs from authorship and requiring transparent disclosure of artificial intelligence involvement.[5] These evolving positions reinforce the importance of human account-ability, disclosure, and governance as LLMs become more integrated into scientific workflows.

LLMS AND DISPARITIES IN CLINICAL CARE

LLMs are trained on pre-defined textual data and are therefore vulnerable to the limitations within that training corpora. In clinical contexts, disparities can be introduced and perpetuated through several mechanisms: (1) unequal representation and data missingness in the EHR, particularly among minoritized groups; (2) biases and inaccuracies in provider documentation of medical, family, and social histories; (3) language variation, including dialects or nonstandard English, which may be underrepresented in training data; (4) reliance on structural proxies (eg, costs or access measures) that substitute for clinical need and reproduce systemic inequities; (5) automation bias, where clinicians over-rely on model outputs; (6) limited capture of social determinants of health, reducing contextual accuracy; and (7) aggregate evaluation metrics, which may mask subgroup-level disparities. Each of these points represents an entryway for bias to be encoded, learned, and reproduced by LLMs in perioperative settings (Figure).

Label Bias

Label bias refers to systematic error in outcome or target variables used to train or evaluate models. It occurs when labels reflect clinician judgment or care processes rather than true underlying disease severity.[7,25,26] In perioperative medicine, the American Society of Anesthesiologists (ASA) physical status classification is a commonly used label that is subject to inter-clinician and institutional variability not fully explained by comorbidity burden. As ASA class is frequently used as both an input and outcome in perioperative modeling, LLMs trained on clinical text containing these labels may learn practice-pattern bias rather than physiologic risk, with potential downstream implications for predictive reliability across patient populations.

Inequitable Data Representation

Minoritized groups are often underrepresented in training data or associated with higher rates of data missingness in the EHR, commonly referred to as health data disparities, leading to poorer model accuracy and reliability for these subgroups.[27] In analytic terms, data missingness refers not simply to absent values, but to the underlying mechanisms that determine whether data are recorded, including whether missingness occurs at random or reflects systematic patterns tied to clinical workflows, access to care, or documentation practices.[28] In EHRs, social history variables are often missing in systematic ways rather than at random, reflecting inconsistent screening practices, time constraints, and variable documentation priorities across clinical workflows and patient populations.[28] Such nonrandom missingness can propagate disparities by systematically compromising data quality for groups whose clinical information is less completely captured, which undermines the validity and predictive performance of models.[28] In addition, stereotypes and myths that mirror human cognitive biases may also be embedded within clinical text used for training, further reinforcing inequitable outputs.[29]

Bias in Clinical Documentation and Language Variation

Language-related variation presents another pathway for inequity. Patient–provider language discordance, particularly when patients use dialects, colloquialisms, or nonstandard English, may result in incomplete or inaccurate documentation. These discrepancies propagate into training data,[30] and LLMs trained primarily on standard English may subsequently generate suboptimal recommendations.[31] Without explicit validation across diverse linguistic inputs, language variation represents a systematic source of disparity.

Structural Proxies and the Reproduction of Systemic Inequities

Furthermore, if LLMs are trained on datasets where access or cost are proxies for “need,” these models may reproduce structural inequities. For example, one widely cited algorithm used in US healthcare under-referred Black patients because it optimized on costs rather than illness burden.[26] In perioperative medicine, bias may be introduced when outcomes such as ICU admission, postoperative monitoring intensity, or use of advanced analgesic techniques are operationalized as indicators of disease severity, despite their dependence on institutional resources, payer mix, and clinician practice patterns. Decisions about ICU admission, postoperative monitoring, or use of advanced analgesic techniques are influenced by hospital resources, staffing, insurance coverage, and institutional protocols, in addition to disease severity. Models trained on these labels may preferentially learn patterns of care delivery rather than underlying surgical risk, potentially disadvantaging patients treated in under-resourced settings.

Social Determinants of Health

Omission of social determinants of health in training data limits a model’s ability to capture non-clinical factors associated with perioperative risk, resulting in systematic gaps in risk representation rather than random error. This limitation disproportionately dis-advantages patients whose outcomes are strongly influenced by socioeconomic factors.[25] Performance evaluation further complicates this issue, as reliance on aggregate metrics such as overall

model accuracy can undermine subgroup-level disparities when equity-relevant endpoints are not explicitly assessed.[32]

Automation Bias

Finally, automation bias may perpetuate inequities when clinicians defer to model outputs without sufficient independent verification, allowing biased or incomplete recommendations to influence clinical decision-making.[7,29] In perioperative settings, reliance on algorithmic decision-support outputs under conditions of time pressure and cognitive load has been shown to influence clinician judgment in ways that reflect underlying data and design biases.[7,29] When integrated into perioperative workflows, over-reliance on such systems may differentially affect vulnerable patient populations if models are trained on data that encode existing practice patterns, thereby reinforcing inequities in perioperative risk assessment and management.[7]

RESPONSIBLE USE OF LLMS IN CLINICAL MEDICINE

Data Collection and Inclusion Across Perioperative Settings

Equitable application of LLMs in perioperative medicine depends on how and what data are included in the EHR.[25,27] Datasets used to train and evaluate perioperative LLMs should reflect the patient populations and clinical environments represented within health systems, including affiliated community hospitals, urban or rural safety-net sites, and regional referral centers, rather than being limited to tertiary academic campuses alone.[1,27] Perioperative documentation practices, workflow structures, resource availability, and transfer patterns differ across these settings and directly shape the clinical text and structured data from which EHR-based LLMs learn. Explicitly defining the intended use and target population of an LLM is therefore essential, as models developed within a single institutional context may not generalize across affiliated sites or clinical settings within academic networks.[1,27]

Inclusion of safety-net and community-based academic affiliates improves representation of clinical complexity, language variation, payer mix, and social risk that influence perioperative decision-making and outcomes.[19,27] Community engagement within academic health systems, including partnerships with local clinical leaders and patient advisory groups, can further enhance participation, trust, and completeness of EHR data, particularly for sociodemographic variables that are inconsistently documented.[19,27] Use of self-reported and disaggregated sociodemographic data, rather than aggregated or proxy-based measures, supports more accurate characterization of perioperative risk and reduces the potential for obfuscating subgroup differences when developing and validating EHR-based LLM applications.[19,25]

Data Curation, Variable Definition, Aggregation, and Governance

After data collection, equitable performance of perioperative LLMs depends on data curation—the systematic processes of cleaning, structuring, and validating datasets—which are essential for preserving representativeness across patient populations.[27] Clear and consistent definitions of clinical and sociodemographic variables are necessary to ensure comparability across sites and time, particularly for variables related to race, ethnicity,

language, and social risk.[25] Data recoding decisions should be evaluated for their impact on minority group representation, as excessive recoding or aggregation may mask meaningful differences in perioperative risk, care pathways, and outcomes. Patterns of missing data should be examined across demographic and institutional strata and interpreted in the context of documentation practices, language concordance, and structural barriers. Data governance frameworks that emphasize transparency in preprocessing decisions, inclusion and exclusion criteria, and subgroup representation are essential to ensure that curated datasets support LLM applications that generalize across diverse perioperative settings and contribute to equitable clinical care delivery.[27,32]

Transparency, Performance Evaluation, and Generalizability

Transparency of data use must be prioritized by model developers. This requires not only describing the data sources but also reporting rates of missingness and the representation of subpopulations. For example, model developers should indicate what proportion of the dataset includes patients from lower socioeconomic groups, racial and ethnic minorities, or other historically underrepresented populations. Providing demographic tables stratified by race and ethnicity, language preference, or insurance status would allow clinicians and health systems to evaluate representativeness relative to their target population. Transparency further requires quantifying subgroup-level missingness (eg, higher rates of absent social history fields among Medicaid patients compared to privately insured patients), since these patterns can propagate bias. Exclusion criteria and preprocessing steps (such as discarding incomplete records) should likewise be disclosed, as these decisions may disproportionately affect vulnerable groups.

Performance reporting should extend beyond overall metrics. Subgroup-specific results (eg, AUROC for each specific population) are critical to identifying disparities. Where performance gaps exist, mitigation strategies such as data reweighting, augmentation, or subgroup-specific thresholds should be considered before deployment.[33] Of course, models can change as data and patient care evolve—thus, it is essential to consider drift detection specifically for each sub-group. Finally, the generalizability of a model should be proven before full-scale implementation into clinical practice. Specifically, models should be externally validated in patient populations and clinical settings that reflect their intended use, with particular attention to differences in payer mix and demographic composition relative to the training data. Once operationalized, prospective studies should include equity endpoints, be powered to detect subgroup effects, and report equity-focused secondary outcomes.

LIMITATIONS AND ETHICAL CONSIDERATIONS

Despite their growing utility, the use of LLMs to support advances in health equity in perioperative medicine presents important limitations and ethical considerations. Cultural biases are well-known in natural language processing models.[24] Hutchinson et al show that LLMs trained on large corpora replicate harmful stereotypes, for example, associating disability with violence, deviance, or dependency.[34] Sampling bias is especially relevant for patients of non-English linguistic backgrounds since most foundational LLM models were

trained on data written in English. Specifically, prominent LLMs have been found to encode non-English languages into English, undermining predictions when processing inputs from non-English speaking patients or providers.[35] In perioperative settings, inaccuracies can compromise preoperative evaluations in advance of indicated procedures.

Another challenge involves the interpretability and trustworthiness of LLM outputs. For instance, confabulation (also referred to as hallucination) is a well-characterized error made by LLMs in which content is generated without basis in proven facts.[5] Currently, no standardized safeguards exist to reliably detect and mitigate confabulation in LLM outputs. As trust in artificial intelligence-constructed text and data increases, latent errors may continue to be propagated downstream. This risk is compounded by existing "copy-forward" practices in EHRs, which already carry forward errors in documentation. Many anesthesia departments have encountered barriers to establishing a reliable data infrastructure to feed into machine learning models.[36] The effect of poor data infrastructure on the quality of LLM outputs has the potential to disproportionately affect health systems in under-resourced areas, where information technology staff and hardware are scarce.

The integration of LLMs in perioperative medicine also raises important questions regarding informed consent and transparency in patient care. Patients have the right to know when LLMs are used to influence aspects of anesthetic management yet conveying this information in an accessible and meaningful manner remains challenging. Uncertainty surrounding how LLMs process sensitive and potentially identifiable health information raises important concerns about data privacy and patient protection. European governments have previously issued warnings or banned consumer chatbots altogether due to similar data privacy concerns.[37] In healthcare settings, stakeholders incentivized to lower risk (eg, health insurers, integrated health systems) could exploit these vulnerabilities, potentially exacerbating harm to already marginalized groups if given access to sensitive patient information.

CONCLUSIONS

The integration of LLMs into perioperative medicine offers ways to enhance clinical workflows, improve patient–provider communication and engagement, and accelerate research. However, without deliberate governance and equity-centered design, these technologies risk reinforcing existing disparities, generating inaccurate and misleading information, and undermining trust in clinical and scientific processes. Responsible implementation will require inclusive data practices, transparency in clinical applications, and sustained human oversight to ensure that these technologies contribute to more equitable perioperative care rather than reinforce disparities.

Funding:

B. N. Burton was supported by the National Institutes of Health T32 Training Grant (GM148369) through the Department of Anesthesiology and Perioperative Medicine at the University of California, Los Angeles. **This manuscript was handled by:** Adam J. Milam, MD, PhD.

REFERENCES

1. Mbadjeu Hondjeu AR, Zhao ZY, Newton L, et al. Large language models in perioperative medicine-applications and future prospects: a narrative review. Can J Anaesth. 2025;72:1000–1014. [PubMed: 40490617]
2. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30:1134–1142. [PubMed: 38413730]
3. Alba C, Xue B, Abraham J, Kannampallil T, Lu C. The foundational capabilities of large language models in predicting postoperative risks using clinical notes. NPJ Digit Med. 2025;8:95. [PubMed: 39934379]
4. Ke YH, Yang Ong BS, Jin L, et al. Clinical and economic impact of a large language model in perioperative medicine: a randomized crossover trial. NPJ Digit Med. 2025;8:462. [PubMed: 40691284]
5. Daneshvar N, Pandita D, Erickson S, Snyder Sulmasy L, DeCamp M; ACP Medical Informatics Committee and the Ethics, Professionalism and Human Rights CommitteeCommittee and the Ethics P and HRC. Artificial intelligence in the provision of health care: an American College of Physicians Policy Position Paper. Ann Intern Med. 2024;177:964–967. [PubMed: 38830215]
6. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large language model capabilities in perioperative risk prediction and prognostication. JAMA Surg. 2024;159:928–937. [PubMed: 38837145]
7. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci. 2021;4:123–144. [PubMed: 34396058]
8. Yalamanchili A, Sengupta B, Song J, et al. Quality of large language model responses to radiation oncology patient care questions. JAMA Netw Open. 2024;7:e244630. [PubMed: 38564215]
9. Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J. Enhancing the readability of online patient education materials using large language models: cross-sectional study. J Med Internet Res. 2025;27:e69955. [PubMed: 40465378]
10. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. 2023;329:842–844. [PubMed: 36735264]
11. Yang X, Xiao Y, Liu D, et al. Enhancing doctor-patient communication using large language models for pathology report interpretation. BMC Med Inform Decis Mak. 2025;25:36. [PubMed: 39849504]
12. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol. 2023;14:1199058. [PubMed: 37303897]
13. Alexander J, Beatty A. Association of patient portal messaging with survival among radiation oncology patients. Int J Radiat Oncol Biol Phys. 2024;120:627–638. [PubMed: 38723754]
14. Baxter SL, Saseendrakumar BR, Cheung M, et al. Association of electronic health record inbasket message characteristics with physician burnout. JAMA Netw Open. 2022;5:e2244363. [PubMed: 36449288]
15. Pham JH, Thongprayoon C, Miao J, et al. Large language model triaging of simulated nephrology patient inbox messages. Front Artif Intell. 2024;7:1452469. [PubMed: 39315245]
16. Wan P, Huang Z, Tang W, et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. Nat Med. 2024;30:2878–2885. [PubMed: 39009780]
17. Lin SJ, Sun CY, Chen DN, et al. Perioperative application of chatbots: a systematic review and meta-analysis. BMJ Health Care Inform. 2024;31:e100985.
18. Rainey JP, Blackburn BE, McCutcheon C, et al. Comparative outcomes with an artificial intelligence-powered short message service chatbot after total joint arthroplasty. Surg Technol Int. 2025;45:sti45/1858.
19. Bradley AS, Bonner TJ, Youssef MR, et al. Perioperative health care disparities in the united states: a systematic review. Anesth Analg. 2025;141:1297.
20. Hake J, Crowley M, Coy A, et al. Quality, accuracy, and bias in ChatGPT-based summarization of medical abstracts. Ann Fam Med. 2024;22:113–120. [PubMed: 38527823]

21. Wang ZP, Bhandary P, Wang Y, Moore JH. Using GPT-4 to write a scientific review article: a pilot evaluation study. BioData Min. 2024;17:16. [PubMed: 38890715]
22. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods. 2024;15:616–626. [PubMed: 38484744]
23. Giglio AD, Costa MUP. The use of artificial intelligence to improve the scientific writing of non-native English speakers. Rev Assoc Med Bras. 2023;69:e20230560. [PubMed: 37729376]
24. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356:183–186. [PubMed: 28408601]
25. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169:866–872. [PubMed: 30508424]
26. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447–453. [PubMed: 31649194]
27. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Dig Health. 2024;3:e0000651.
28. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20:144–151. [PubMed: 22733976]
29. Wang J, Redelmeier DA. Cognitive biases and artificial intelligence. NEJM AI. 2024;1.
30. Lee Y, Chang CH, Yang CC. Enhancing patient-physician communication: simulating African American Vernacular English in medical diagnostics with large language models. J Healthc Inform Res. 2025;9:119–153. [PubMed: 40309129]
31. Blodgett SL, Barocas S, DauméWallach H III. Language (technology) is power: a critical survey of "bias" in NLP. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2020:5454–5476.
32. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng. 2023;7:719–742. [PubMed: 37380750]
33. Xu J, Xiao Y, Wang WH, et al. Algorithmic fairness in computational medicine. EBioMedicine. 2022;84:104250. [PubMed: 36084616]
34. Hutchinson B, Prabhakaran V, Denton E, Webster K, Zhong Y, Denuyl S. Social biases in NLP models as barriers for persons with disabilities. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; . 2020:5491–5501.
35. Wendler C, Geiger A, Lu Y, et al. Do Llamas work in English? On the latent language of multilingual transformers. arXiv preprint. 2024;2402.10588. https://arxiv.org/abs/2402.10588.
36. Epstein RH, Hofer IS, Salari V, Gabel E. Successful implementation of a perioperative data warehouse using another hospital's published specification from epic's electronic health record system. Anesth Analg. 2021;132:465–474. [PubMed: 32332291]
37. McCallum S. ChatGPT banned in Italy over privacy concerns. BBC News. March 31, 2023. Accessed August 5, 2025. https://www.bbc.com/news/technology-65139406.

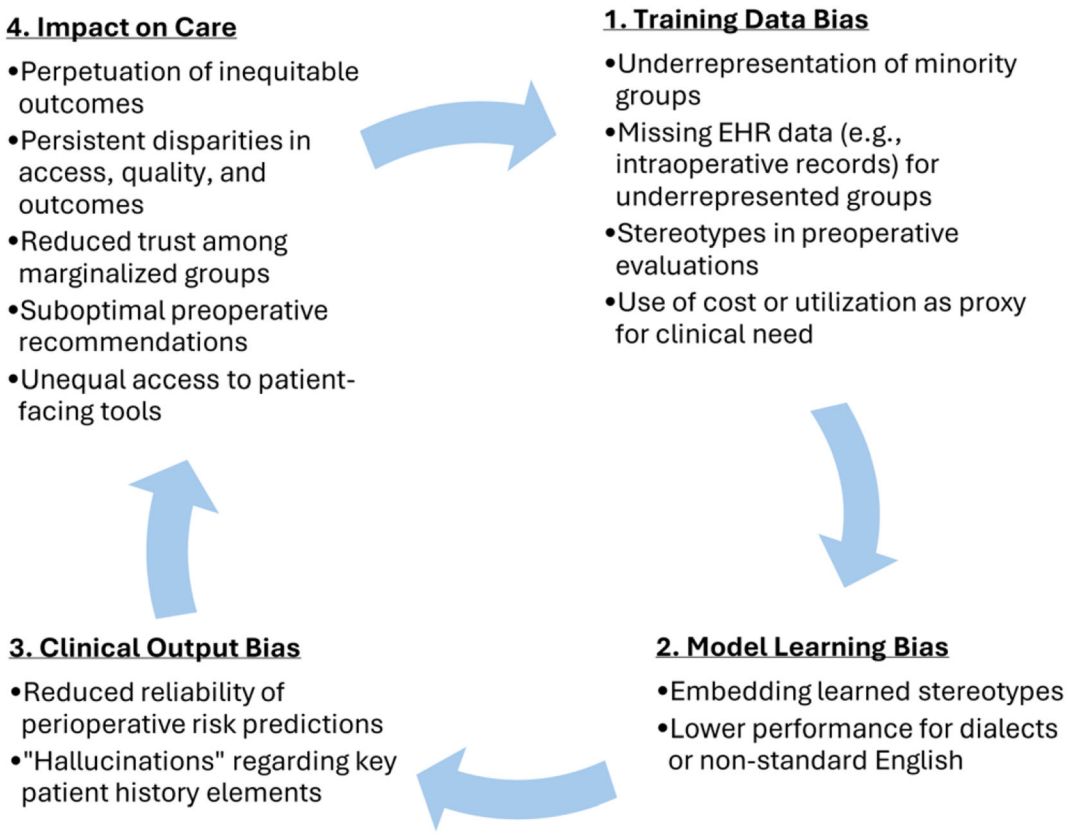

Figure.

How LLMs can perpetuate disparities in care a self-reinforcing loop in which biased training data and model learning lead to biased clinical outputs and downstream disparities; these disparities feed back into future training data. LLMs indicates large language models.