




How large language models can augment perioperative medicine: a daring discourse

Rodney A Gabriel ¹, Edward R Mariano ^{2,3}, Julian McAuley,⁴
Christopher L Wu ⁵

ABSTRACT

Interest in natural language processing, specifically large language models, for clinical applications has exploded in a matter of several months since the introduction of ChatGPT. Large language models are powerful and impressive. It is important that we understand the strengths and limitations of this rapidly evolving technology so that we can brainstorm its future potential in perioperative medicine. In this daring discourse, we discuss the issues with these large language models and how we should proactively think about how to leverage these models into practice to improve patient care, rather than worry that it may take over clinical decision-making. We review three potential major areas in which it may be used to benefit perioperative medicine: (1) clinical decision support and surveillance tools, (2) improved aggregation and analysis of research data related to large retrospective studies and application in predictive modeling, and (3) optimized documentation for quality measurement, monitoring and billing compliance. These large language models are here to stay and, as perioperative providers, we can either adapt to this technology or be curtailed by those who learn to use it well.

INTRODUCTION

The interest in natural language processing (NLP) models for clinical applications exploded in a matter of several months since the introduction of ChatGPT toward the end of 2022.¹ NLP is a branch within the broader field of artificial intelligence (AI) that aims to process text into meaningful ways similar (or better) than a human

would. Interest in its clinical applications is not new and has been significantly studied in the era of using the electronic health record.² While there are several methodological approaches to developing NLP models, it was not until the wide use of transformers—an advanced deep learning modality that incorporates self-attention while processing sequential input (such as natural language)—combined with the increased computational power resources, where applications using large language models have surfaced such as ChatGPT.³ While a deeper exploration of the methodology of transformers is beyond the scope of this daring discourse, it is important, as clinicians and scientists who specialize in perioperative medicine that we understand the strengths and limitations of this rapidly evolving technology.

Large language models

Large language models are NLP models built on various forms of transformers (with potentially billions of weights/parameters) which are pretrained on extremely large sets of data.⁴ Examples of large language models include, but not limited to generative pretrained transformers (GPT) and bidirectional encoder representation from transformers. From these large language models, conversational AI applications have been built including ChatGPT (based on the large language model GPT) and Google Bard (based on the large language model LaMDA). There are several current differences in strengths and limitations between these conversational AI applications including training in non-English languages (Bard is limited to English language), ability to code in programming languages (not yet available with Bard), and historic information the models were trained on (ChatGPT as of now has learned data up to September 2021).⁵ While the technical differences between the innerworkings of each transformer-based language model is beyond the scope of this article, it is important to note a common theme, in that these models were pretrained on millions to even trillions of

text and contains millions to billions of parameters.

Why these large language models are not perfect

As seen with ChatGPT, applications using large language models are powerful and impressive. However, the question remains, what can we do with this given the limitations of healthcare data, specifically related to the privacy-related restrictions and liability concerns? An understanding of their weaknesses is important. One important consideration is that, while these models are trained on massive data sets, they are not trained on institution-specific electronic health record data. The reason is related to patient privacy related restrictions. Therefore, outputs from these models—when used at specific healthcare institutions—would not necessarily be reflective of institution-specific patterns related to its patient population, clinician practices, and documentation style. Thus, the next major concern is the potential inaccuracy of its outputs, which may lead to misapplication of results, errors in clinical judgment, and potential liability issues. Finally, accessing large language models through non-healthcare platforms, including the provision of inputs and submission of queries, may put one's personal data at risk.

Large language models are here to stay!

As we continue to think about appropriate use cases of these large language models, a question worth asking is whether these models will replace clinician decision-making. Should we be concerned that such technology could replace a portion of the clinician workforce? After all, ChatGPT has been demonstrated to pass the medical licensing exam and diagnose rare diseases in a matter of seconds.⁶ In any case, these large language models are here to stay and, as perioperative physicians, we likely must need to adapt to this technology or be curtailed by those who learn to use it well.

Dependence on this technology may lead us down a wrong path, in which, in one extreme, would be that our clinical decisions come second to what the language models tell us and to our patients. With the limitations of bias and accuracy from these pretrained models, this may cause more harm than good.⁷ However, these limitations will not prevent healthcare institutions from adapting large language models into practice. Therefore, we should maintain the integrity of human intelligence and, focus on how to use

¹Anesthesiology, University of California San Diego, La Jolla, California, USA

²Anesthesiology and Perioperative Care Service, VA Palo Alto Health Care System, Palo Alto, California, USA

³Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Stanford, California, USA

⁴Computer Science and Engineering, University of California San Diego, La Jolla, California, USA

⁵Anesthesiology, Hospital for Special Surgery, New York City, New York, USA

Correspondence to Dr Rodney A Gabriel, Anesthesiology, University of California San Diego, La Jolla, California, USA; ragabriel@health.ucsd.edu

and fine-tune these NLP applications to augment our clinical performance, especially in the perioperative space. As Samer Narouze eloquently stated in the recent American Society of Regional Anesthesia and Pain Medicine spring meeting in 2023, ‘ChatGPT will not replace clinicians, but clinicians who do not use ChatGPT will be replaced.’

Leveraging large language models to improve perioperative medicine: an example of the future of large language models

As electronic medical record systems adopt large language model technology, such as GPT-4, we need to proactively think how to leverage these models into practice. For perioperative medicine, there are several opportunities, but three potential major areas may serve as initial targets: (1) clinical decision support and surveillance tools; (2) improved aggregation and analysis of research data related to large retrospective studies and application in predictive modeling; and (3) optimized documentation for quality measure, monitoring and billing compliance (figure 1).

Clinical decision support and surveillance are key in this era of big data especially with the concerns of the administrative burden placed on healthcare professionals when dealing with electronic medical records.⁸ For example, tracking pain and opioid outcomes could benefit

from this technology. Large language models may be leveraged to perform a number of tasks to aid in surveillance: first, it may be used to classify text documents based on whether or not they are relevant to opioid use and pain outcomes. Second, it may be used as a text summarizer to provide succinct summaries of clinical notes that healthcare providers can review (particularly when presented with patients with a complex chronic pain history and interventions); and third, it may be used to alert pain providers of specific patient populations that would potentially benefit from personalized interventions—such as preoperative pain consultations, acute pain service managements perioperatively, and/or transitional pain clinic referrals postoperatively.^{9 10} NLP can be used to identify surgical patients who persistently use opioids months after surgery.¹¹ Furthermore, NLP can be leveraged to generate preanesthetic history notes that can help anesthesia providers in preoperative care clinics as they triage and evaluate high volumes of surgical patients.¹² Thus, these large language models can help us identify the right patients faster so we can focus on personalizing care for those who most need it.

Research using large datasets has several limitations, especially due to the observational nature of such studies, non-standardized representations of data, and challenges in controlling for unseen

confounding factors.¹³ Although they attempt inferences across population-sized samples, outcomes data may not always be accurate and may be largely dependent on select structured data points. As already presented, these data analyses may lead researchers to conclusions that will not always be accurate and thus may influence clinical decision-making in the wrong direction. Examples of machine learning models predicting opioid use outcomes after surgery have been reported,¹⁴⁻¹⁷ however, the aforementioned limitations of outcomes data are of concern. To improve accuracy of these types of models, especially as the size of datasets continues to grow, large language models may be used to process clinical documents to generate data points and more accurately classify patients. Thus, large language models can help us advance our research, by allowing researchers to train their models on much larger datasets while potentially reducing the concern of inaccurate training data.

Optimized documentation and billing compliance are essential for developing a new pain medicine practice and maintaining healthy financial infrastructure, and this applies to acute, transitional, and chronic pain programs.⁹ Previously, computer-assisted coding has been used in both rule-based and NLP-driven approaches.^{18 19} Large language models have the potential to glean additional data

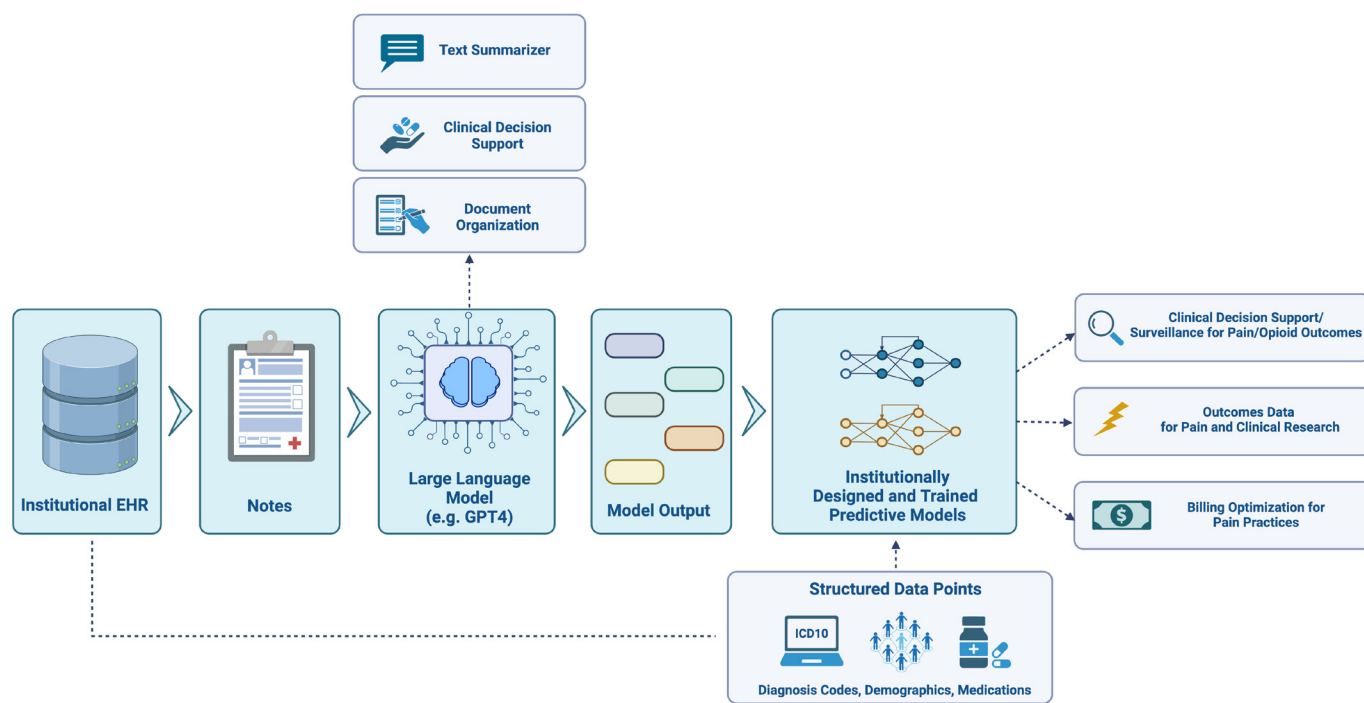


Figure 1 Schematic illustrating the potential integration of large language models into electronic medical record systems and how it may be used to aid in tasks related to perioperative medicine. EHR, electronic health record; GPT, generative pretrained transformers; ICD, International Classification of Diseases.

points such as indicators of quality from text-based documentation, which are becoming more relevant in value-based healthcare. Future studies need to further investigate the application of such models within the various types of pain medicine practices and determine whether or not the use of large language models truly improves documentation performance and/or compliance. Thus, we may be able to leverage these models to increase our financial gains within our practices. However, the same technologies could also be used by insurers and may lead to the widening of healthcare inequities and lack of access to care. For example, large insurance carriers are already leveraging AI to more quickly analyze and deny claims.²⁰ The insurance industry can save billions of dollars through automated reviews of claims; a system called PXDX used by Cigna has been reported to facilitate 50 medical reviews every 10s.^{21 22} Large hospital and physician groups that have resources to use AI to boost billing may be able to keep up with insurers and will have an advantage compared with smaller groups caring for underserved populations that do not have this resource.

DISCUSSION

The future of AI in medicine will likely depend on large language model technology, especially as electronic medical record systems begin to adopt its technology. We should brainstorm on how it can make us better and more efficient perioperative providers. In this daring discourse, we highlighted the limitations of large language models but then ventured into how we can leverage this technology for the benefit of patient care. Dependence of this technology has serious concerns related to its potential negative downstream effects on clinical decision-making as well as compromising the privacy of electronic medical record data. Despite this, we know it will play an integral role in medicine. While we discussed three potential areas—clinical decision support, advancement in research using electronic medical record data, and optimizing documentation and billing compliance—there are likely several other areas that need to be explored. A high-level understanding of its strengths/limitations while collaborating with technical experts

in the field of NLP is key to successfully operationalizing and researching the use of NLP in perioperative medicine.

Twitter Edward R Mariano @EMARIANOMD and Christopher L Wu @@ChrisWuMD

Contributors All the authors were involved with concept design and preparation of initial and final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests RG's institution has received funding and/or product for research purposes from Epimed, Infutronic, SPR Therapeutics, Merck, and Precision Genetics. RG is a consultant for Avanos. ERM, JM, and CLW have no financial conflicts of interest to disclose.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

© American Society of Regional Anesthesia & Pain Medicine 2023. No commercial re-use. See rights and permissions. Published by BMJ.



To cite Gabriel RA, Mariano ER, McAuley J, et al. *Reg Anesth Pain Med* 2023;**48**:575–577.

Received 27 April 2023
Accepted 7 June 2023
Published Online First 19 June 2023

Reg Anesth Pain Med 2023;**48**:575–577.
doi:10.1136/rapm-2023-104637

ORCID iDs

Rodney A Gabriel <http://orcid.org/0000-0003-4443-0021>

Edward R Mariano <http://orcid.org/0000-0003-2735-248X>

Christopher L Wu <http://orcid.org/0000-0002-4484-0787>

REFERENCES

- Will ChatGPT transform healthcare. *Nat Med* 2023;29:505–6.
- Hossain E, Rana R, Higgins N, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Comput Biol Med* 2023;155:106649.
- Shen Y, Heacock L, Elias J, et al. Chatgpt and other large language models are double-edged swords. *Radiology* 2023;307:e230163.
- Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med* 2022;5:194.
- Johnson A. Bard vs Chatgpt: the major difference between the AI chat tools, explained. 2023.
- Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.
- Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine* 2023;90:104512.
- Downing NL, Bates DW, Longhurst CA. Physician burnout in the electronic health record era: are we ignoring the real cause? *Ann Intern Med* 2018;169:50–1.
- Sun EC, Mariano ER, Narouze S, et al. Making a business plan for starting a transitional pain service within the US healthcare system. *Reg Anesth Pain Med* 2021;46:727–31.
- Tighe PJ, Lucas SD, Edwards DA, et al. Use of machine-learning classifiers to predict requests for preoperative acute pain service consultation. *Pain Med* 2012;13:1347–57.
- Gabriel RA, Park BH, Simpson S. Utilizing natural language processing and machine learning to identify persistent opioid use from postoperative surgical documents. abstract. American Society of Regional Anesthesia and Pain Medicine Annual Conference; 2023
- Suh HS, Tully JL, Meineke MN, et al. Identification of preanesthetic history elements by a natural language processing engine. *Anesth Analg* 2022;135:1162–71.
- Kaji AH, Rademaker AW, Hyslop T. Tips for analyzing large data SETS from the JAMA surgery statistical editors. *JAMA Surg* 2018;153:508–9.
- Gabriel RA, Harjai B, Prasad RS, et al. Machine learning approach to predicting persistent opioid use following lower extremity joint arthroplasty. *Reg Anesth Pain Med* 2022;47:313–9.
- Giladi AM, Shipp MM, Sanghavi KK, et al. Patient-reported data augment health record data for prediction models of persistent opioid use after elective upper extremity surgery. *Plast Reconstr Surg* 2023; Publish Ahead of Print.
- Yen H-K, Ogink PT, Huang C-C, et al. A machine learning algorithm for predicting prolonged postoperative opioid prescription after lumbar disc Herniation surgery. An external validation study using 1,316 patients from a Taiwanese cohort. *Spine J* 2022;22:1119–30.
- Debbi EM, Krell EC, Sapountzis N, et al. Predicting postdischarge opioid consumption after total hip and knee arthroplasty in the opioid Naive patient. *J Arthroplasty* 2022;37.
- Chen P-F, Wang S-M, Liao W-C, et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. *JMIR Med Inform* 2021;9:e23230.
- Venkatesh KP, Raza MM, Kvedar JC. Automating the overburdened clinical coding system: challenges and next steps. *NPJ Digit Med* 2023;6:16.
- Kaushik K, Bhardwaj A, Dwivedi AD, et al. Machine learning-based regression framework to predict health insurance premiums. *Int J Environ Res Public Health* 2022;19:7898.
- Rosenthal E. Denials of health insurance claims are rising—and getting weirder. 2023. Available: url: <https://www.fiercehealthcare.com/payers/denials-health-insurance-claims-are-rising-and-getting-weirder> [Accessed 30 May 2023].
- RuckerP, MillerMD. How Cigna saves millions by having its doctors reject claims without reading them. 2023. Available: url: <https://www.propublica.org/article/cigna-pxdx-medical-health-insurance-rejection-claims> [Accessed 30 May 2023].