Check for updates

EDITORIAL COMMENTARY

Arthroscopy
The Journal of Arthroscopic and Related Surgery
AANA ADVANCING THE SCOPE
Practice-Changing Innovation

Editorial Commentary: Advancing the Value of Artificial Intelligence in Health Care Means Rethinking What We Are Measuring It Against

Krish S. Sardesai, B.A., Brian J. Cole, M.D., M.B.A.,
and Kyle N. Kunze, M.D., Associate Editor

Abstract: Despite an increasing emphasis on value-based health care, inappropriate resource utilization remains prevalent, with unnecessary referrals for imaging and surgical procedures. Recent studies have explored the use of large language models (LLMs) as copilots for streamlining data acquisition and optimizing decision-making with respect to these areas of patient care. Emerging data suggest that patient questionnaires, when parsed by LLMs, may allow for decision-making regarding the appropriate next steps in evaluation and treatment that is comparable to human providers. While encouraging, such use cases merit caution about the biases inherent in generalized LLMs whose recommendations fundamentally do not reflect informed clinical decision making but rather a byproduct of models trained on vast amounts of data from unverified sources that include medical information. Further, the question remains: should we be comparing LLM performance to the outcomes of the current health care system we function within or the outcomes that are necessary to advance patient care moving forward? When researchers choose to repurpose LLMs for these use cases, they must consider which options provide the greatest value.

Arthroscopy. 2026;00:1-4

Musculoskeletal care is the leading driver of health care spending in the United States and bears a high responsibility for wasteful resource allocation.[1-3] Preoperative imaging represents a major avenue of wasteful spending; estimates report that 76%-84% of preconsultation knee magnetic resonance imaging (MRI) are clinically unnecessary.[3] When proceeding to surgery, operating room resources account for one-third of all expenditures.[1] Prior cross-sectional studies in have showed that waste in this setting is derived from discrepancies between estimated and true operating times, physical medical waste (including general and hazardous), unnecessary surgical trays, and wasted implants and instruments.[4] Advancements in artificial intelligence (AI) and large language models (LLMs)—which can process large amounts of unstructured data to provide new insights about care delivery—may allow for the identification of specific areas in which current practices are not cost effective and contribute to decreasing value.[5,6]

In their article entitled "Large Language Model Predicts Surgeon Recommendations for Imaging and Surgery for Patients Presenting for Knee and Shoulder Complaints With 70% and 81% Accuracy Using Previsit Questionnaire Responses,"[7] Halvorson, Keeley, Niknam, Zack, Majumdar, Feeley, Zhang, and Lansdown investigated the accuracy of LLMs to parse free-text previsit patient questionnaires and make recommendations on imaging and surgical treatment based on questionnaires alone. Specifically, the recommendations of a pretrained GPT4o model (OpenAI, San Francisco, CA) were compared against a surgeon's final treatment recommendation among a cohort of 1141 new surgical patients (554 knee, 587 shoulder) that presented to their institution. For the imaging component of the investigation, the LLM was prompted to determine whether an urgent MRI was needed after being given a 14-question patient questionnaire where patients described their injury onset, mechanism, functional limitations, pain, and demographic

Midwest Orthopaedics at Rush, Chicago, Illinois, U.S.A.

View this article online at wileyonlinelibrary.com. DOI: 10.1002/arj.70038

© 2026 the Arthroscopy Association of North America.

factors. This LLM had 70% accuracy (knee 75% and shoulder 66%) and had a sensitivity and specificity of 83% and 64%, respectively. Accuracy was highly variable by injury, with high accuracy for injuries to the anterior cruciate ligament, meniscus, and rotator cuff (80%-94%) and low accuracy for osteoarthritis (54%-66%). In the second arm of the study, the LLM was prompted to determine whether urgent surgery was needed. The LLM was more accurate (81% vs 69%) and sensitive (88% vs 32%) when provided with an MRI report than when making decisions based on questionnaires alone. The authors suggest that pretrained LLMs are sufficiently accurate in predicting whether an "urgent MRI" of the knee/shoulder is necessary prior to recommending treatment and for surgical treatment.

The authors should be applauded for a well-conducted study that explores a new use case with a leading closed-source LLM that leverages new insight into the potential for AI as a prognostic clinical tool. Studies such as this contribute to new understandings surrounding the performance of LLMs for clinically relevant use cases and avoid the redundancy of investigations that are observed in the literature.[8] However, there are several important considerations when interpreting the results of this study. First, there is bias when using a surgeon's final determination as the gold standard for ordering imaging or proceeding to surgery. In interpreting the accuracy/sensitivity/specificity of the LLM and its external validity, we are at the behest of these surgeons' individual philosophy on imaging, risk aversion, and subconscious bias. The fact that this is a single-institution study of seven surgeons' patient-base also compromises the external validity of these conclusions. There are external factors—such as variability in insurance prior authorization and institution-specific order sets for common injuries—which may influence whether a pathology merits imaging and/or surgery.[9,10] It is difficult to control for these factors in a single-institution study. Second, though the authors conclude that sufficient accuracy was observed, the accuracies and overall performance in real-world practice would be clinically unacceptable. A 30% error rate in determining the need for urgent MRI would not be feasible and limits the justification for using an LLM compared with expert physician. Third, when examining the surgical model, the authors report a low sensitivity (32%) when the model was imaging naïve, only exhibiting a sensitivity >80% when augmented with imaging, but this large gap presents a paradoxical conflict when interpreting these results. The rationale for the study is to reduce unnecessary MRI studies and reports, but when considering the >50% rise in sensitivity with image-augmentation, we cannot truly determine how many of these MRIs would have been deemed "unnecessary" or "wasteful"—and to what extent these wasteful MRI helped the LLM—when, by virtue of their imaging model, they might not have been ordered. This is important as, strikingly, Petron et al. reported orthopaedic surgeons are likely to agree with the ordering of only 12% of the MRI they receive preconsult.[11] For the current study, we may accept the success of the LLM in reducing unnecessary imaging while understanding that the sensitivity of the surgical model remains unacceptably low unless supplemented with similarly wasteful imaging reports—at least to some extent.

Finally, there are issues when using a model like GPT4o, which is trained on a large body of unknown datasets and, as such, is not a model specific to orthopaedic surgery and was not developed in such a way that it was intended to be used for medical purposes. There are alternatives to help enhance an LLM's vertical integration into a specialty like orthopaedic surgery, such as retrieval-augmented generation. This would allow a specific model to be customized by being curated with specific data without retraining the entire model, such that when a query is fed to the LLM (e.g., "*Is urgent MRI needed?*"), the model can vectorize both the query and the supplemental data (e.g., AAOS imaging guidelines and high-quality orthopaedic literature) before passing through the LLM's layers. This, in turn, has been showed to provide more accurate answers that are trustworthy.[12] Another strategy could consist of utilizing convolutional neural networks, which, in this circumstance, would involve freezing most of the model's pretrained layers and retraining only final classification layers with labeled data (e.g., MRI reports labeled as necessary or unnecessary) to help increase accuracy of models with respect to a specific discipline, like orthoapedic surgery.[13]

Despite these limitations, the data presented by Halvorson et al. serve as a valuable example for how LLMs can augment, but not replace, clinical decision-making. Their study provides conclusions that are impressive considering the LLM never accesses a patient's chart, lacks a physical exam, and does not have a surgeon's advantage of leveraging decades of experience prior to making decisions. However, the model may have ultimately been limited by not having access to this data, as evidenced by the relatively limited performance across various use cases, and making decisions without this data does not mimic real-world clinical practice. With a field as interdisciplinary and resource dependent as orthopaedic surgery, it is important to assess the value of LLMs in parsing unstructured data and providing clinical insights.

We are left to consider two essential questions: what are we measuring AI performance against and why? In this study, it was measured against the processes and outcomes within a health care system that is already established (surgeon decision making and treatment recommendations in an academic practice), but the model still showed lower performance than humans. We should

15263231, 0, Downloaded from https://arthroscopyjournals.onlinelibrary.wiley.com/doi/10.1002/arj.70038 by Change Healthcare Llc, Wiley Online Library on [10/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Arthroscopy The Journal of Arthroscopic and Related Surgery AANA ADVANCING THE SCOPE Practice-Changing Innovation

be careful in using this benchmarking and rather aim to compare performance to the outcomes that we need to advance health care. Ultimately, this study reaffirmed that there is a clinically unacceptable error rate in management when LLMs are utilized for patient care.[13,14] Utilization of LLMs at this stage for similar use cases would only widen the gap that currently exist—although patients could gain recommendations faster and with less data points, many recommendations would be incorrect when juxtaposed to experienced clinical decision making, resulting in more waste. What we should be measuring against are the following questions, which are more challenging but likely more valuable:

(1) Will specialty-specific LLMs with curated training data be more useful than generalized, rapidly changing LLMs that are publicly available?
(2) What are the health-outcomes and financial consequences of implicating LLMs for immediate decision making and triaging as compared to traditional clinical outpatient workflows or emergency room triage?
(3) Can patient-facing LLMs be used before specialty visits to help reduce the burden of inappropriate imaging acquisition and treatment pathways?

It has been reported that in the current state of AI development, we are in a specialist-generalist paradox. In other words, AI models outperform specialists on specific tasks, but AI still lags far behind experts in less controlled (generalized) settings when the boundaries become less clear. This has implications for patient-facing LLMs of which the majority consist of off-the-shelf proprietary models. With a generalist model using patient input, this may lead to incorrect recommendations. Specialty-curated LLMs have shown promised for decreasing medical error rates, but these studies have been performed by clinicians themselves and not by patients. Finally, the health-outcomes and financial consequences of implementing LLMs in real-world clinical scenarios remains largely unknown. Although conceptually this use case would increase workflow efficiency and be cost-effective, well-designed studies are needed to answer this question. Forward-thinking strategies in AI research aimed at measuring against the gaps that currently exist, as opposed to against outcomes that we have, will increase value and accelerate progress in musculoskeletal health care.

DISCLOSURES

The authors (B.J.C., K.N.K.) declare the following financial interests/personal relationships which may be considered as potential competing interests. B.J.C. reports a relationship with Aesculap/B.Braun (research support), the *American Journal of Sports Medicine* (editorial or governing board), Arthrex (IP royalties, paid consultant, research support), the Arthroscopy Association of North America (board or committee member), Bandgrip (stock or stock options), Elsevier Publishing (IP royalties), the Journal of the *American Academy of Orthopaedic Surgeons* (editorial or governing board), JRF Ortho (other financial or material support), the National Institutes of Health (NIAMS & NICHD) (research support), *Operative Techniques in Sports Medicine* (publishing royalties and financial or material support), and Ossio (stock or stock options). K.N.K. reports a relationship with AllaiHealth that includes consulting or advisory and equity or stocks and serves on the Editorial Board - *Arthroscopy* Journal. The other author (K.S.S.) declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

ORCID

Kyle N. Kunze https://orcid.org/0000-0002-0363-3482

REFERENCES

1. Chen KJ, Rascoe A, Su CA, et al. Value challenge: A bottoms-up approach to minimizing cost and waste in orthopaedic surgery. *JBJS Open Access.* 2023;8:e22.
2. Coale M, Schiffman B, Iannuzzi N, Huang J. Magnetic resonance imaging for elbow pathology: Overused by both orthopedic surgeons and primary care providers. *JSES Int.* 2022;6:1062-1066.
3. Mohammed HT, Yoon S, Hupel T, Payson LA. Unnecessary ordering of magnetic resonance imaging of the knee: A retrospective chart review of referrals to orthopedic surgeons. *PLoS One.* 2020;15:e0241645.
4. Childers CP, Maggard-Gibbons M. Understanding costs of care in the operating room. *JAMA Surg.* 2018;153:e176233.
5. Zhang C, Liu S, Zhou X, et al. Examining the role of large language models in orthopedics: Systematic review. *J Med Internet Res.* 2024;26:e59607.
6. Kunze KN, Nwachukwu BU, Cote MP, Ramkumar PN. Large language models applied to health care tasks may improve clinical efficiency, value of care rendered, research, and medical education. *Arthroscopy.* 2025;41:547-556.
7. Halvorson RT, Keeley T, Niknam K, et al. Large language model predicts surgeon recommendations for imaging and surgery for patients presenting for knee and shoulder complaints with 70% and 81% accuracy using previsit questionnaire responses. *Arthroscopy.* 2026;xx:xx-xx.
8. Kunze KN, Gerhold C, Dave U, et al. Large language model use cases in health care research are redundant and often lack appropriate methodological conduct: A scoping review and call for improved practices. *Arthroscopy.* 2025;41:4928-4945.e2.
9. Harrer S, Hedden K, Mikaeili S, et al. Magnetic resonance imaging prior authorizations for orthopaedic care are negatively affected by medicaid insurance status. *J Am Acad Orthop Surg.* 2025;33:e244-e252.

15263231, 0, Downloaded from https://arthroscopyjournals.onlinelibrary.wiley.com/doi/10.1002/arj.70038 by Change Healthcare Llc, Wiley Online Library on [10/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10. van Beek EJR, Kuhl C, Anzai Y, et al. Value of MRI in medicine: More than just another test? *J Magn Reson Imaging.* 2019;49:e14-e25.
11. Petron DJ, Greis PE, Aoki SK, et al. Use of knee magnetic resonance imaging by primary care physicians in patients aged 40 years and older. *Sports Health.* 2010;2:385-390.
12. Ke YH, Jin L, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digit Med.* 2025;8:187.
13. Jia H, Zhang J, Ma K, Qiao X, Ren L, Shi X. Application of convolutional neural networks in medical images: a bibliometric analysis. *Quant Imaging Med Surg.* 2024;14:3501-3518.
14. Nwachukwu BU, Varady NH, Allen AA, et al. Currently available large language models do not provide musculoskeletal treatment recommendations that are concordant with evidence-based clinical practice guidelines. *Arthroscopy.* 2025;41:263-275.e6.

15263231, 0, Downloaded from https://arthroscopyjournals.onlinelibrary.wiley.com/doi/10.1002/arj.70038 by Change Healthcare Llc, Wiley Online Library on [10/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License