Editorial Commentary: Machine Learning Is Just a Statistical Technique, Not a Mystical Methodology or Peer Review Panacea

Prem N. Ramkumar, M.D., M.B.A., and Riley J. Williams III, M.D.

Abstract: Orthopaedic and sports medicine research surrounding artificial intelligence (AI) has dramatically risen over the last 4 years. Meaningful application and methodologic rigor in the scientific literature are critical to ensure appropriate use of AI. Common but critical errors for those engaging in AI-related research include failure to 1) ensure the question is important and previously unknown or unanswered; 2) establish that AI is necessary to answer the question; and 3) recognize model performance is more commonly a reflection of the data than the AI itself. We must take care to ensure we are not repackaging and internally validating registry data. Instead, we should be critically appraising our data—not the AI-based statistical technique. Without appropriate guardrails surrounding the use of artificial intelligence in Orthopaedic research, there is a risk of repackaging registry data and low-quality research in a recursive peer-reviewed loop.

See related article on page 777

We read with great interest "Machine Learning Can Accurately Predict Overnight Stay, Readmission, and 30-Day: Complications Following Anterior Cruciate Ligament Reconstruction" by Lopez, Gazgalis, Peterson, Confino, Levine, Popkin, and Lynch.[1] The authors are applauded for their efforts in this study.

Upon reviewing the diction, methodologic execution, and reporting of the findings, the zeitgeist of machine learning (ML)—and artificial intelligence (AI) as a whole—can be readily gleaned. Given that the study centers on ML techniques, a nuanced understanding of AI-based reporting is critical. To safeguard meaningful and appropriate use of AI, a systematic and reflective approach is necessary, and the following questions serve as an essential check before engaging in AI-based research:

1. Is the scientific question important or necessary?
2. Has this question been previously answered? If so, are there any grounds to question the previously established answer?
3. Is AI necessary to answer the question?
4. Is the data set sufficient to convincingly answer the question?

We agree that the posed clinical question is relevant. Preoperative knowledge of whether a patient is likely to incur an overnight stay, endure short-term complications, and potentially experience a readmission is important following ACL reconstruction (ACLR) to optimize resource management and set patient expectation. However, the study poses the specific question as to whether or not machine learning (ML) has the potential of predicting specific metrics from a database containing said metrics. Without performing the investigation, it is intuitively known that a ML model can be built. The efficacy of the model, however, is dependent on the quality and quantity of the data—*not* the efficacy of ML modeling itself. Researchers engaging in AI-related research should have the foundational awareness that one cannot appraise the power of ML by simply evaluating its predictive performance as a surrogate in isolation. This concept is often misunderstood. As such, we contend that setting out to evaluate ML by way of assessing its statistical modeling performance with the National Surgical Quality Improvement Program (NSQIP) database for ACLR is not a meaningful question nor an accurate methodology, since the results do not answer the proposed question. Moreover, ML models have been well

Sports Medicine Institute, Hospital for Special Surgery.

The authors report no conflicts of interest in the authorship and publication of this article.

Full ICMJE author disclosure forms are available for this article online, as supplementary material.

0749-8063/22895/$36.00

https://doi.org/10.1016/j.arthro.2022.07.012

established in the literature to be appropriate statistical modeling processes; attempting to again validate ML as a modeling technique at this point is unnecessary. Lopez et al evaluate whether ML models using *this* database for *this* surgery during *this* time period evaluating *these* specific outcome variables are homogeneously consistent enough to be reproducibly modeled. What can be stated is that the data represent fair to good sufficiency as a springboard model for future validation testing in developing a predictive model.

Most of the clinical and technical questions in this study have been answered in the literature across various studies. Boddapati et al. demonstrated that procedure length was independently associated with overnight stay and 30-day readmission rate following ACLR. Comorbidities like obesity, smoking history, and hypertension were again demonstrated to be important risk factors.[3] Kammien et al. described the causes of emergency department (ED) visits within 90 days of ACLR across 81,179 patients and found that 8.3% of patients present to the ED most commonly due to surgical site concerns (39.4%).[4] Additionally, Lu et al. evaluated the NSQIP database between 2007 and 2017 and demonstrated that inpatient ACLRs result in a higher risk of short-term complications.[5] More importantly—and in a separate report evaluating the same database and same procedure over a longer time span (2006-2018)—Lu et al. constructed multiple well-performing ML models that demonstrated predictive efficacy in identifying which ACLR patients are at risk for overnight admission.[6] Of note, no specific models for readmissions and 30-day complications were built.[1] Additionally, no regression models were created for comparison.[1] Thus, the primary advance that Lopez et al. offers is highlighting the difference between ML and logistic regression (LR) modeling for ACLR with the NSQIP database.[1] In the setting of high-quantity or high-quality data, it is established that AI-based modeling techniques will always outperform regression-based techniques.[2] Within the sports literature particularly, the superiority of ML techniques over regression analyses for predictive modeling has been established several times, most notably with large professional databases from the National Hockey League and Major League Baseball.[7,8]

The study question was specifically whether or not a machine learning model is capable of reliably predicting these metrics (overnight stay, readmission, 30-day complications) after ACLR and whether it would outperform logistic regression (LR). Thus, an AI-based technique was de facto necessary to answer the question regardless of the validity or necessity of the study question. Moreover, the data were more than sufficient to underscore that ML techniques outperform LR techniques based on the consistently improved AUCs with an artificial neural network. Thus, Lopez et al redemonstrate that ML outperforms LR when all conditions are equal.[1] From a clinical standpoint, it is unlikely that an administrative database, including a time period beginning a decade ago (2012-2018), is reflective of reality today. When comparing differences in the prevalence of overnight stay following ACLR between the Lu et al. (11.3%, 2006-2018) and Lopez et al. (10.2%, 2012-2018) studies using the same databases, it is clear that fewer patients are staying overnight as we continue to improve perioperative pain modalities and optimize same-day discharge protocols. From a fidelity standpoint, NSQIP data are reported to be accurate to three significant figures with routine auditing for data quality and reports inter-rater disagreement of <2% for all variables.[9] Full methodologic evaluation of Lopez et al. is limited without disclosure of the full source code.[2]

Although the authors should be applauded for demonstrating the superiority of ML to LR techniques, the study underscores several pitfalls associated with AI-related research. First, registry data are ripe for repackaging using ML techniques without critical advancement.[2] This study is very similar to the work done by Lu et al.,[6] whose group demonstrated the viability of ML for modeling the NSQIP database after ACLR. Using this logic, we could assert that every study that once applied regression modeling can be repurposed using ML techniques—while failing to add meaning. Second, as mentioned, the premise of the title and report itself claims to evaluate ML, when in reality the methodology merely reflects the data evaluated. ML is simply another statistical technique, but this error continues to propagate in the literature, as studies persist in describing what "machine learning models predict,"[10-12] rather than what the data predict. To claim otherwise would be tantamount to asserting "Student's *t*-test accurately detects differences" in every study that has demonstrated a *P* value less than .05. Third and finally, the allure of ML is exciting and novel, but adhering to basic principles of answering meaningful questions with appropriate data is the safest way to avoid propagating hype and safeguard methodologic rigor.

References

1. Lopez C, Gazgalis A, Peterson JR, Confino JE, Levine WN, Popkin CA, Lynch TS. Machine learning can accurately predict overnight stay, readmission, and 30-day: Complications following anterior cruciate ligament reconstruction. *Arthroscopy* 2023;39:777-786.
2. Ramkumar PN, Pang M, Polisetty T, Helm JM, Karnuta JM. Meaningless applications and misguided methodologies in artificial intelligence-related orthopaedic research propagates hype over hope. *Arthroscopy* 2022;38:2761-2766.
3. Boddapati V, Fu MC, Nwachukwu BU, et al. Procedure length is independently associated with overnight hospital stay and 30-day readmission following anterior cruciate

ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc* 2020;28:432-438.
4. Kammien AJ, Zhu JR, Gouzoulis MJ, et al. Emergency department visits within 90 days of anterior cruciate ligament reconstruction. *Orthop J Sports Med* 2022;10: 23259671221083586.
5. Lu Y, Lavoie-Gagne O, Khazi Z, Patel BH, Mascarenhas R, Forsythe B. Inpatient admission following anterior cruciate ligament reconstruction is associated with higher postoperative complications. *Knee Surg Sports Traumatol Arthrosc* 2020;28:2486-2493.
6. Lu Y, Forlenza E, Cohn MR, et al. Machine learning can reliably identify patients at risk of overnight hospital admission following anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc* 2021;29:2958-2966.
7. Luu BC, Wright AL, Haeberle HS, et al. Machine learning outperforms logistic regression analysis to predict next-season NHL player injury: An analysis of 2322 players from 2007 to 2017. *Orthop J Sports Med* 2020;8: 2325967120953404.
8. Karnuta JM, Luu BC, Haeberle HS, et al. Machine learning outperforms regression analysis to predict next-season major league baseball player injuries: Epidemiology and validation of 13,982 player-years from performance and injury profile trends. *Orthop J Sports Med* 2020;8:2325967120963046:2000-2017.
9. Khuri SF, Henderson WG, Daley J, et al. Successful implementation of the Department of Veterans Affairs' National Surgical Quality Improvement Program in the private sector: The Patient Safety in Surgery study. *Ann Surg* 2008;248:329-336.
10. Shapira J, Peskin B, Norman D. Editorial commentary: Machine learning can indicate hip arthroscopy procedures, predict postoperative improvement, and estimate costs. *Arthroscopy* 2022;38:2217-2218.
11. Kunze KN, Krivicich LM, Clapp IM, et al. Machine learning algorithms predict achievement of clinically significant outcomes after orthopaedic surgery: A systematic review. *Arthroscopy* 2022;38:2090-2105.
12. Grazal CF, Anderson AB, Booth GJ, Geiger PG, Forsberg JA, Balazs GC. A machine-learning algorithm to predict the likelihood of prolonged opioid use following arthroscopic hip surgery. *Arthroscopy* 2022;38: 839-847.e2.