Editorial Commentary: Machine Learning in Medicine Requires Clinician Input, Faces Barriers, and High-Quality Evidence Is Required to Demonstrate Improved Patient Outcomes

Ayoosh Pareek, M.D., and R. Kyle Martin, M.D., F.R.C.S.C.

Abstract: Machine learning (ML) and artificial intelligence (AI) may be described as advanced statistical techniques using algorithms to "learn" to evaluate and predict relationships between input and results without explicit human programming, often with high accuracy. The potentials and pitfalls of ML continue to be explored as predictive modeling grows in popularity. While use of and optimism for AI continues to increase in orthopaedic surgery, there remains little high-quality evidence of its ability to improve patient outcome. It is up to us as clinicians to provide context for ML models and guide the use of these technologies to optimize the outcome for our patients. Barriers to widespread adoption of ML include poor quality data, limits to compliant data sharing, few clinicians who are expert in ML statistical techniques, and computing costs including technology, infrastructure, personnel, energy, and updates.

See related article on page 2090

Machine learning (ML) and artificial intelligence (AI) can be described as advanced statistical techniques that use algorithms to model intervariable relationships that may be much more complex than previously possible with traditional statistical techniques. These models can "learn" from the data without explicit programming or significant human directed actions.[1] Just as when the Kaplan-Meier survival curve was first introduced in the 1950s,[2] the use of ML is taking off within the field of orthopaedic surgery with predictive analytics articles appearing with increasing frequency. More than just a fad, there is reason to believe the hype. Throughout society ML has become ubiquitous, with recent breakthroughs in autonomous vehicles, language translation, facial recognition, and Amazon recommender systems. While the medical field tends to lag behind private enterprise innovations due to being purposefully late adopters,[3] the tide is starting to turn. Machine learning is primed to provide us with new tools aimed at improving patient care and optimizing outcomes at a patient-specific level.

It is with great personal and professional interest that we read "Machine Learning Algorithms Predict Achievement of Clinically Significant Outcomes Following Orthopaedic Surgery: A Systematic Review" by Kunze, Krivicich, Clapp, Bodendorfer, Nwachukwu, Chahla, and Nho.[4] In this study, the authors reviewed the literature for ML applications within orthopedic surgery that targeted the prediction of patient-reported outcome (PRO). The PROs of interest were specified as the achievement of minimal clinically important difference (MCID), patient-acceptable symptomatic state (PASS), and substantial clinical benefit (SCB). Of the 18 articles that met the inclusion criteria, Kunze et al. report that prediction of MCID achievement was included in all studies, while a minority also included SCB. None of the studies attempted to predict PASS. The 18 studies came from a variety of orthopedic subspecialties including spine, sports medicine, and both upper and lower extremity arthroplasty. Overall, the authors found that ML algorithms predicted MCID with fair to good performance in most studies, but there was little to no evidence for the prediction of SCB or PASS.

Mayo Clinic (A.P.); University of Minnesota (R.K.M.)

The authors report the following potential conflicts of interest or sources of funding: R.K.M. has received meeting and/or travel support from Arthrex, Smith & Nephew, and Conmed and has received the Norwegian Centennial Chair Seed Grant to his institution. A.P. has received institutional funding from the Norwegian Centennial Chair Seed Grant. Full ICMJE author disclosure forms are available for this article online, as supplementary material.

0749-8063/2281/$36.00
https://doi.org/10.1016/j.arthro.2022.01.026

We commend the authors for their rigorous systematic review incorporating relevant inclusion/exclusion criteria culminating in a concise report on the available literature. Lending credence to the notion that this is a hot topic, it is important to note that all 18 studies included in the systematic review were published within the past 4 years. Interestingly, it also appears that at least 7 of the 18 included studies (∼40%) appear to be from the same group of authors, and half of the 18 include coauthors involved with this systematic review. This highlights the preliminary nature of these studies from a select group of clinician scientists. These studies form the foundation from which our specialty can learn from and improve upon in the future. As familiarity with ML techniques and clinical applications increases among orthopaedic surgeons and researchers, so too will the number of predictive algorithms we see in print. We suspect a similar systematic review performed only a few years from now would include many more studies.

There were a few notable findings reported by Kunze et al. First, only 3 of the 18 included studies presented their findings in line with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines.[5] The TRIPOD checklists for model development and validation encourage uniform and robust reporting that adheres to a minimum standard. Noncompliance with these guidelines limits our understanding of the studies, further complicated by their heterogeneity.[6] Another interesting observation was that preoperative PRO scores were significant predictors of postoperative subjective outcome, consistent with our recent ML evaluation of subjective failure after anterior cruciate ligament reconstruction.[7] This highlights the importance of collecting preoperative PRO from our patients for more than purely academic purposes. It may also suggest there is something about how people answer the surveys that influences their subjective outcome rather than the PRO itself being a true marker of success/failure, a confounding variable.

Despite the exponential increase in recent orthopaedic ML application, several barriers to widespread adoption of ML within the specialty remain. First, in this age of digital economy, data is oil. For a strong ML study, it is important to maximize data quality, quantity, and usability. Deficiency in any of these can seriously limit the utility of the resultant model. The approach with the most potential for strong algorithm development with real-world generalizability uses an organized multicenter collaborative effort as opposed to single-center or single-surgeon databases. Unfortunately, data sharing can be extremely challenging in health care due, in part, to patient and privacy concerns. Second, even with a large repository of high-quality labeled data, there remains a paucity of experts capable of bridging the statistical and clinical gap. While there are many ML experts capable of creating computational models, these models may have limited applicability if constructed without the experience of a clinician, and many clinicians are far more familiar with the traditional statistical approach. A third barrier includes the literal costs of ML. Machine learning and AI tend to require processing units, which the every-day workplace computer may not be suitable for, presenting expenses related to the storage space and computational cost. Additionally, ML models tend to be costly in terms of literal electric energy, which is a finite and limiting resource.[8]

In addition to the aforementioned limitations, there is also a “technical debt” to ML models, which is rarely discussed in health care, but represents a practical consideration.[9] This refers to the considerations that must be accounted for when not only creating ML models, but also their maintenance with regards to technology, infrastructure, computing costs, and personnel. Machine learning models require significant effort to create, while doing so with rigorous consideration of the eventual clinical utility is even more difficult. As noted by Kunze et al., only one of the included studies in their systematic review was externally validated, likely because external validation is time consuming and requires access to a similar, but still high-quality data. Clinicians may also mistrust prediction algorithms developed using the novel ML approach, as these models may be prone to errors and bias, which has been frequently reported.[10] Gaining trust and encouraging clinical application of ML models demands a delicate balance between maximizing model accuracy and reporting clear, interpretable findings. Lastly, the technical debt in ML is not only limited to creating and deploying models, but also with their upkeep. Just as we upgrade our phones and computers to meet our evolving needs, these models need new data and algorithms to maintain accuracy as our patients, surgical techniques, and indications change. There is a significant opportunity cost that must be considered when creating and maintaining these models, but this also represents a strength of the ML approach, as algorithms can be made to learn and further refine prediction accuracy as new data are acquired over time.

We feel that prediction discrimination, typically reported as area under the curve (AUC) or concordance (C-index), deserves its own brief discussion. Traditional teaching tells us that discrimination of greater than .9 is excellent, .8-.9 is good, .7-.8 is fair, and less than .7 is poor.[11] While this may be true of other industries or ML applications, it does not hold for most orthopaedic clinical prediction modeling. In other words, a model with an AUC of .95 may not be inherently more clinically useful than one with an AUC of .67. The reason for this is based on the fact that the reference standard, in this case, the chosen subjective outcome measurement tool, is itself imperfect. In fact, a clinical prediction model with a high AUC of greater than .8 or .9 often represents algorithm design flaws related to data mismanagement or poor implementation (overfitting) of the ML algorithms.[12] Rather, most clinically useful ML studies in healthcare report discrimination between .65-.8.[13] Just as with the *P* value, one needs to consider the whole picture and not confuse statistical significance for clinical significance.[14]

We believe ML is here to stay, and rightfully so, as these applications are increasingly becoming integrated within healthcare.[15] We eagerly await the next generation of ML developments in clinical orthopedic surgery, and expect future models to build upon and improve the preliminary studies currently in print.[16] Next-level ML approaches, including unsupervised learning, computer vision, and natural language processing hold additional potential, and significant clinician input will be required to maximize the benefit to our patients.

References

1. Martin RK, Ley C, Pareek A, Groll A, Tischer T, Seil R. Artificial intelligence and machine learning: An introduction for orthopaedic surgeons. *Knee Surg Sports Traumatol Arthrosc* 2022;30:361-364.
2. D'Arrigo G, Leonardis D, Abd ElHafeez S, Fusaro M, Tripepi G, Roumeliotis S. Methods to analyse time-to-event data: The Kaplan-Meier survival curve. *Oxid Med Cell Longev* 2021;2021:2290120.
3. Berwick DM. Disseminating innovations in health care. *JAMA* 2003;289:1969-1975.
4. Kunze KN, Krivicich LM, Clapp IM, et al. Machine learning algorithms predict achievement of clinically significant outcomes following orthopaedic surgery: A systematic review. *Arthroscopy* 2022;38:2090-2105.
5. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *J Clin Epidemiol* 2015;68:134-143.
6. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J Med Internet Res* 2016;18:e323.
7. Martin RK, Wastvedt S, Pareek A, et al. Predicting subjective failure of ACL reconstruction: A machine learning analysis of the Norwegian Knee Ligament Register and patient-reported outcomes [published online January 11, 2022]. *J ISAKOS*. https://doi.org/10.1016/j.jisako.2021.12.005
8. Hao K. Training a single AI model can emit as much carbon as five cars in their lifetimes. MIT Technol Rev, https://cacm.acm.org/careers/237345-training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/fulltext. Accessed June 7, 2019.
9. Sculley D, Holt G, Golovin D, et al. *Hidden technical debt in machine learning systems* 2015;2503-2511.
10. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ* 2021;375:n2281.
11. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240:1285-1293.
12. Kernbach JM, Staartjes VE. Foundations of machine learning-based clinical prediction modeling: Part II. Generalization and overfitting. *Acta Neurochir Suppl* 2022;134:15-21.
13. Youngstrom EA. A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *J Pediatr Psychol* 2014;39:204-221.
14. Pareek AP. Value: Purpose, power, and potential pitfalls. AAOS Now, https://www.aaos.org/aaosnow/all-articles/2019/aug/research/research02/. Accessed August, 2019.
15. Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med* 2020;104:101822.
16. Martin RK, Pareek A, Krych AJ, Maradit Kremers H, Engebretsen L. Machine learning in sports medicine: Need for improvement. *J ISAKOS* 2021;6:1-2.