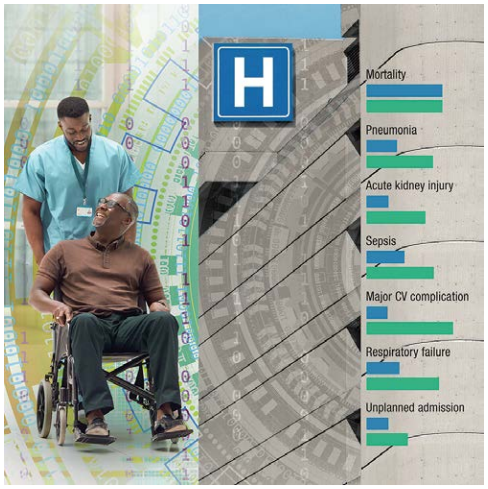
Prediction Algorithms: Is Peer Review Enough?

Laurent G. Glance, M.D., Laszlo Vutskits, M.D., Ph.D., Andrew Davidson, M.B.B.S., M.D., F.A.N.Z.C.A., F.A.H.M.S.

Many risk prediction models have been developed, some of which serve as the foundation for healthcare reform and clinical decision-making. Many of these tools were developed based on diagnoses and procedure codes from the index hospitalization, meaning that most of the information used as the inputs for these prediction models is only available after patient discharge. The inherent disadvantage of this approach is that it does not allow accurate and individualized risk stratification at the time of hospital admission when such an evaluation is of particular clinical relevance.

In this issue, Greenwald *et al.*[1] present an updated version of their Risk Stratification Index. The authors are to be congratulated for creating and validating a robust set of prediction models based on 4,426 International Classification of Diseases codes out of a possible 69,000 diagnostic codes. The authors suggest that the current revision will be more useful than previous versions because it uses International Classification of Diseases codes coded the year before hospital admission, thus making the revised index usable during the index admission. However, the decision to include only International Classification of Diseases codes that are present the year before admission may be both a strength and a limitation, since some patients may develop new diagnoses that are present on admission but not available in historical data. Furthermore, patients may be admitted to hospitals that do not have access to their historical data. Nonetheless, the predictions of the Risk Stratification Index (risk of death, major complications [acute kidney injury, sepsis, respiratory failure], excess length of stay, and unplanned readmission) could be used to guide patient management in important ways. For example, the Risk Stratification Index could be used to (1) identify patients who may benefit from step-down or intensive care, (2) triage surgical patients to match the skill set of anesthesiologists and trainees, (3) guide medical therapy to optimize outcomes, and (4) save costs by reducing the use of unnecessary levels of care.



"Before embedding predictive analytics in the electronic medical record, should we require independent testing to show that they are 'safe' and improve outcomes—or at a minimum, that the models accurately predict outcomes?"

The Risk Stratification Index is one of many prediction models that are now widespread in medicine. Prediction models have become a fundamental driver of healthcare reform and clinical practice. The quality and performance of hospitals and physicians cannot be fairly measured without first adjusting for differences in patient case mix and surgical complexity using risk adjustment models. The Centers for Medicare and Medicaid Services publicly report hospital risk-adjusted outcomes to promote transparency and patient choice. The American College of Surgeons[2] and the Society of Thoracic Surgeons[3] provide their members with nonpublic performance reporting to guide quality improvement. The Centers for Medicare and Medicaid Services is redesigning the healthcare system to deliver higher quality care at a lower cost using pay-for-performance (*e.g.*, Hospital Readmission Reduction Program[4]); episode-based payments (*e.g.*, Comprehensive Care for Joint Replacement,[5] Bundled

Image: A. Johnson, Vivo Visuals Studio.

This editorial accompanies the article on p. 673. This article has an audio podcast.

Accepted for publication October 17, 2022.

Laurent G. Glance, M.D.: Department of Anesthesiology and Perioperative Medicine, University of Rochester School of Medicine, Rochester, New York; and RAND Health, RAND, Boston, Massachusetts.

Laszlo Vutskits, M.D., Ph.D.: Department of Anesthesiology, Pharmacology, Intensive Care, and Emergency Medicine, University Hospitals of Geneva, Geneva, Switzerland; and Geneva Neuroscience Center, University of Geneva, Geneva, Switzerland.

Andrew Davidson, M.B.B.S., M.D., F.A.N.Z.C.A., F.A.H.M.S.: Department of Anesthesia, Royal Children's Hospital, Melbourne, Australia; and Murdoch Children's Research Institute, Melbourne, Australia.

 Anesthesiology 2022; 137:661–3. DOI: 10.1097/ALN.0000000000004421

<zdoi;. DOI: 10.1097/ALN.0000000000004421>

Payments for Care Improvement[6]); and accountable care organizations (Medicare Shared Savings Program[7]). How much the Centers for Medicare and Medicaid Services pay hospitals depends on their risk-adjusted performance. Prediction models are also used to guide clinical decision-making: (CHA_2DS_2-VASc score for atrial fibrillation stroke risk[8]) and risk stratification before surgery (American College of Surgeons Surgical Risk Calculator[9]).

However, to be useful, a prediction model must accurately predict outcomes. Hospital performance is quantified by comparing its performance (*e.g.*, the observed mortality rate) to its predicted performance (*e.g.*, the expected [predicted] mortality rate). Suppose a prediction model does not accurately predict outcomes. In that case, patients may unintentionally be guided to select low-performance hospitals, the Centers for Medicare and Medicaid Services may penalize high-performance hospitals while rewarding low-performance hospitals, and clinicians may decide to place high-risk patients on the ward immediately after surgery. The performance of prediction models can be assessed using standard statistical criteria (*e.g.*, model discrimination, model calibration) in patient samples that are independent of the sample used to create the prediction model. Using best-in-class prediction models is important because whether a hospital is classified as either a high- or low-quality hospital can depend on which prediction model is used for risk adjustment and not just on the intrinsic quality of the hospital.[10] Similarly, the decision to pursue invasive testing before surgery is also a function of which prediction model is used for risk stratification.[11]

Before the Centers for Medicare and Medicaid Services use a performance measure for quality reporting or value-based purchasing, the measure and the underlying risk adjustment methodology must first be evaluated and endorsed by the National Quality Forum.[12] There is no formal mechanism to evaluate and endorse most prediction models before they are used clinically. Although the Food and Drug Administration is responsible for the regulation of "Software as a Medical Device," it has only recently issued guidance that prediction models like the Epic Sepsis Model (developed by the commercial electronic health record vendor Epic) should be subject to regulatory review.[13] This prediction model, which is widely used at hundreds of U.S. hospitals without first undergoing independent validation, was recently shown to miss 67% of patients with sepsis.[14] Before embedding predictive analytics in the electronic medical record, should we require independent testing to show that they are "safe" and improve outcomes—or at a minimum, that the models accurately predict outcomes? We believe that the answer is a resounding "yes." There are currently best practices for the reporting of prediction models.[15] However, peer review should only be the first step before a prediction model is used to guide clinical care.

Some critics of these algorithms point out that the code's details are often kept proprietary and not published with the article.[16] Anesthesiology encourages authors to describe the code in sufficient detail so that readers can consider whether the fundamentals of the algorithm are built on robust designs and data. What is the right balance between protection of intellectual property *versus* transparency? We believe that journals need to ask more from authors. In the absence of widespread regulation of prediction models, journals are the first and only line of defense to ensure that valid prediction models are disseminated to front-line clinicians. We propose that journals strongly encourage developers to make code available to outside researchers to allow the independent evaluation of prediction models. Alternatively, developers could provide a working version of the prediction model (without sharing proprietary code) to allow independent validation. Greenwald *et al.*[1] are to be commended for providing code for each of the Risk Stratification Index models along with sample data. It is worth noting that if the algorithms are already patented, then the details of the algorithm may already be in the public domain as part of the regulatory requirement for obtaining the patent. Journals should also promote the validation of prediction models by publishing the work of independent teams who evaluate and replicate published models. One challenge in the evaluation of these models is that they are likely to be regularly updated or tweaked. It would be hoped that these models are upgraded as new validation data emerge, as medical practice changes, and as we have access to new types of data (for example, physiologic data from wearables). Like a new phone, the software behind the prediction scores may—and perhaps should—be upgraded every year. Whatever process we have for rigorous evaluation will need to be nimble enough to accommodate regular upgrades.

Last, some worry about the ethical implications of these predictive algorithms. Instead of being used to identify patients who need escalated care, could they instead be used to identify patients who would be denied care because they are predicted to have an increased risk of major complications, extended length of stay, and readmission? Could insurance companies and hospitals use them to selectively avoid patients deemed to be at too high a financial or reputational risk? Race, ethnicity, and insurance status (*e.g.*, Medicaid coverage) are frequently a proxy for unmeasured disease severity. Will including race and ethnicity in prediction models unfairly disadvantage vulnerable populations by encouraging hospitals to selectively avoid these vulnerable individuals?

Risk prediction models have become embedded in the healthcare system. They are the centerpiece of healthcare reform and will play an increasingly important role in clinical decision-making. Anesthesiology welcomes manuscripts that help our readers understand the important features of these algorithms and especially those studies that provide more evidence for when and where they can improve the outcomes for our patients.

Research Support

Supported by the Department of Anesthesiology and Perioperative Medicine at the University of Rochester School of Medicine and Dentistry (Rochester, New York).

Competing Interests

Drs. Vutskits and Davidson are Editors for Anesthesiology. Dr. Glance is a member of the Scientific Methods Panel for the National Quality Forum and an Associate Editor for Anesthesiology. Dr. Glance reports funding from the National Institutes of Health (Bethesda, Maryland; grant Nos. R01AG074492 and RO1NRO16865).

Correspondence

Address correspondence to Dr. Glance: laurent_glance@urmc.rochester.edu

References

1. Greenwald S, Chamoun GF, Chamoun NG, Clain D, Hong Z, Jordan R, Manberg PJ, Maheshwari K, Sessler DI: Risk Stratification Index 3.0, a broad set of models for predicting adverse events during and after hospital admission. Anesthesiology 2022; 137:673–86
2. Hall BL, Hamilton BH, Richards K, Bilimoria KY, Cohen ME, Ko CY: Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: An evaluation of all participating hospitals. Ann Surg 2009; 250:363–76
3. Bowdish ME, D'Agostino RS, Thourani VH, Schwann TA, Krohn C, Desai N, Shahian DM, Fernandez FG, Badhwar V: STS Adult Cardiac Surgery Database: 2021 update on outcomes, quality, and research. Ann Thorac Surg 2021; 111:1770–80
4. Desai NR, Ross JS, Kwon JY, Herrin J, Dharmarajan K, Bernheim SM, Krumholz HM, Horwitz LI: Association between hospital penalty status under the hospital readmission reduction program and readmission rates for target and nontarget conditions. JAMA 2016; 316:2647–56
5. Finkelstein A, Ji Y, Mahoney N, Skinner J: Mandatory Medicare bundled payment program for lower extremity joint replacement and discharge to institutional postacute care: Interim analysis of the first year of a 5-year randomized trial. JAMA 2018; 320:892–900
6. Joynt Maddox KE, Orav EJ, Zheng J, Epstein AM: Year 1 of the bundled payments for care improvement—Advanced model. N Engl J Med 2021; 385:618–27
7. McWilliams JM, Hatfield LA, Landon BE, Hamed P, Chernew ME: Medicare spending after 3 years of the Medicare Shared Savings Program. N Engl J Med 2018; 379:1139–49
8. Melgaard L, Gorst-Rasmussen A, Lane DA, Rasmussen LH, Larsen TB, Lip GY: Assessment of the CHA_2DS_2-VASc score in predicting ischemic stroke, thromboembolism, and death in patients with heart failure with and without atrial fibrillation. JAMA 2015; 314:1030–8
9. Cohen ME, Liu Y, Ko CY, Hall BL: An examination of American College of Surgeons NSQIP Surgical Risk Calculator accuracy. J Am Coll Surg 2017; 224:787–95.e1
10. Iezzoni LI: The risks of risk adjustment. JAMA 1997; 278:1600–7
11. Glance LG, Faden E, Dutton RP, Lustik SJ, Li Y, Eaton MP, Dick AW: Impact of the choice of risk model for identifying low-risk patients using the 2014 American College of Cardiology/American Heart Association perioperative guidelines. Anesthesiology 2018; 129:889–900
12. Glance LG, Joynt Maddox K, Johnson K, Nerenz D, Cella D, Borah B, Kunisch J, Kurlansky P, Perloff J, Stoto M, Walters R, White S, Lin Z: National Quality Forum guidelines for evaluating the scientific acceptability of risk-adjusted clinical outcome measures: A report from the national quality forum scientific methods panel. Ann Surg 2020; 271:1048–55
13. Ross C: In New Guidance, FDA Says AI Tools to Warn of Sepsis Should Be Regulated as Devices. 2022. Available at: https://www.statnews.com/2022/09/27/health-fda-artificial-intelligence-guidance-sepsis/. Accessed October 12, 2022
14. Wong A, Cao J, Lyons PG, Dutta S, Major VJ, Otles E, Singh K: Quantification of sepsis model alerts in 24 US hospitals before and during the COVID-19 pandemic. JAMA Netw Open 2021; 4:e2135286
15. Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med 2015; 162:735–6
16. Wanderer JP, Ehrenfeld JM: Toward external validation and routine clinical use of the American College of Surgeons NSQIP Surgical Risk Calculator. J Am Coll Surg 2016; 223:674