

Risk Prediction Tools

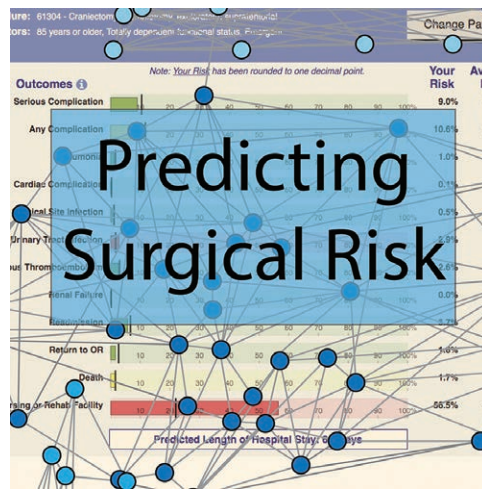
The Need for Greater Transparency

Laurent G. Glance, M.D., Andrew W. Dick, Ph.D., Turner M. Osler, M.D., M.S.

LAST year Amazon captured 40% of online sales partly as a result of accurate personalized predictions designed to help consumers discover what they want using a machine learning algorithm.^{1,2} In health care, we aspire to the same goal: accurate personalized predictions, although in the healthcare arena we are interested in predicting outcomes of medical consequence to our patients rather than their next consumer whim. This revolution in predictive power has changed the nature of online buying, and now seems poised to transform medical practice. Our challenge is to ensure that this transition to personalized medicine is safely implemented. If risk prediction tools are to be used in clinical care, it is essential that they be vetted with the same care as any new drug or medical device.

Liu *et al.* have made an excellent start in this issue of *ANESTHESIOLOGY*³ by describing the estimation of the intrinsic cardiac risk of surgical procedures using the same statistical methodology previously reported for the Universal American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) Surgical Risk Calculator.⁴ It has long been recognized that surgical complexity is a major determinant of the risk of major adverse cardiac events. The Revised Cardiac Risk Index, which is the most widely used cardiac risk prediction tool in clinical practice, simply divides surgical procedures into two categories based on expert opinion: high risk versus not high risk.⁵ The NSQIP Myocardial Infarction or Cardiac Arrest Risk Calculator, on the other hand, groups surgeries into 21 categories (e.g., aortic, bariatric, cardiac, and intestinal). In comparison, the ACS NSQIP Surgical Risk Calculator is much more granular and is based on estimates of procedure-specific risk of major adverse cardiac events for more than 1,500 different Current Procedural Terminology (CPT) codes.⁴

Estimating the surgery-specific risk of major adverse cardiac events for individual procedures, as opposed to broad



“If risk prediction tools are to be used in clinical care, it is essential that they be vetted with the same care as any new drug or medical device.”

more than 1,500 separate CPT codes is an innovative application of hierarchical modeling to a challenging modeling problem. The key finding reported by Liu *et al.* is that there is wide variability in surgery risk for procedures that will not be accurately captured using risk models that categorize procedures based on body regions, or by dividing procedures into just two risk categories. Liu *et al.* report a threefold difference in cardiac risk between open cholecystectomy and a Whipple.³ Both of these surgeries are collapsed into a single category, “high-risk procedures,” in the Revised Cardiac Risk Index. These findings suggest that the venerable Revised Cardiac Risk Index may not be the best risk prediction tool for cardiac risk assessment.

After initial peer review, the accuracy of risk calculators should be independently validated by research groups that did not participate in model development. We encourage Liu *et al.* to release the model equations in the Surgical Risk Calculator for predicting not only cardiac complications, but for all postoperative complications. Releasing the model

categories of surgeries, has face validity. Doing this well for such a large number of separate procedures is not a trivial undertaking. *A priori*, it is unlikely that collapsing surgical procedures into a small number of categories—two in the case of the Revised Cardiac Risk Index, and 21 for the NSQIP Myocardial Infarction or Cardiac Arrest Risk Calculator—results in the most accurate estimates of surgical risk. Alternatively, including more than 1,500 unique CPT codes in a regression model is likely to lead to biased and imprecise estimates of intrinsic cardiac risk because many CPT codes are very uncommon. Instead, Liu *et al.*³ used hierarchical modeling and specified individual CPT codes as a random intercept. Hierarchical modeling will provide more stable estimates for CPT codes with small number of observations than nonhierarchical modeling. The statistical method used to estimate procedure-specific risk for

Image: J. P. Rathmell.

Corresponding articles on pages 247 and 283.

Accepted for publication August 10, 2017. From the Departments of Anesthesiology and Public Health Sciences, University of Rochester School of Medicine, Rochester, New York (L.G.G.); RAND Health, RAND, Boston, Massachusetts (L.G.G., A.W.D.); and the Department of Surgery, University of Vermont, Burlington, Vermont (T.M.O.).

Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. All Rights Reserved. *Anesthesiology* 2018; 128:244-6

equations permits outside developers to evaluate the statistical performance of these risk prediction tools. In the absence of this information, independent investigators have examined the accuracy of the ACS NSQIP Surgical Risk Calculator by comparing observed and predicted patient outcomes obtained using the ACS NSQIP Web-based calculator for small patient cohorts.⁶ This approach will not always yield valid assessments of model performance because of problems with sample size, case-mix homogeneity, and lack of generalizability—as shown in recent work published by Cohen *et al.*⁷ Using the Web-based calculator to externally validate the ACS Surgical Risk Calculator for large patient populations is not practical since submitting data to this Web-based calculator for hundreds of thousands of patients may interfere with clinician real-time access to this widely used resource. We find the ACS claim that “if the Risk Calculator equations were to enter the public domain, the ACS would lose the ability to protect the public from outdated or faulty implementation”⁸ unpersuasive. It should be possible for ACS to provide outside developers with updated model equations, as these become available, to prevent “outdated implementations” of the Risk Calculator.

Greater transparency is likely to lead to better models and more accurate predictions. None of us are as smart as all of us. In 2006, Netflix offered a \$1 million prize to challenge outside developers to improve their recommendation algorithm. Measure developers, such as ACS, can accomplish this for free by placing risk prediction models in the public domain. The benefits of crowd-sourcing are seen in trauma injury scoring, which have led investigators to create injury scoring systems—such as the New Injury Severity Score, International Classification of Disease Injury Severity Score, and Trauma Mortality Prediction Model—that are more accurate than the original Injury Severity Score. Without access to the model equations of existing risk prediction models, it is difficult for measure developers to create better models because they cannot compare the performance of new models to older models. At a minimum, measure developers should provide an anonymized data set with the predictors and outcomes, along with the predictions for each patient. Such a data set would allow other investigators to examine, confirm, and possibly improve upon the results of the original risk prediction tool. We hope that Liu *et al.* will lead the way in facilitating crowd-sourcing of risk prediction tools as they have led the way in the technical aspects of hierarchical modeling of procedures as random effects.

In closing, we applaud the ACS for creating a high-quality platform for measuring and reporting surgical outcomes. This effort by our surgical colleagues has contributed to improved patient outcomes in surgery,⁹ and helps lay the groundwork for precision delivery in perioperative medicine. We also applaud the ACS's commitment to transparency by providing researchers with access to its NSQIP database, which has resulted in thousands of publications. We believe, however,

that it is important that the ACS release either the model equations for the ACS NSQIP Surgical Risk Calculator or an enhanced public use file with model predictions. Unlike Amazon and Netflix, whose business model is to maximize profits, the ACS has an obligation to improve risk prediction by promoting transparency in order to fulfill its fundamental mission of “improving the care of surgical patients.”¹⁰ If Amazon's product recommendation system fails, Amazon's profitability will suffer. If risk prediction tools are inaccurate, patients may suffer. The time is right for all measure developers, including the ACS, to ensure that their models can be directly evaluated and validated by outside investigators. Ultimately, personalized medicine should be based on best-of-class risk prediction tools. It is in everyone's interest to ensure that risk prediction tools are as accurate as possible, most importantly the patients'.

Research Support

This project was supported with funding from the Department of Anesthesiology at the University of Rochester School of Medicine, Rochester, New York.

Competing Interests

The authors are not supported by, nor maintain any financial interest in, any commercial activity that may be associated with the topic of this article.

Correspondence

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

References

1. Kam K: The market is underestimating Amazon. Available at: <https://www.forbes.com/sites/kenkam/2016/05/27/the-market-is-underestimating-amazon/#5a50f78a3939>. Accessed December 5, 2017
2. Arora A: Recommendation engines: How Amazon and Netflix are winning the personalization battle. Available at: <https://www.martechadvisor.com/articles/customer-experience/recommendation-engines-how-amazon-and-netflix-are-winning-the-personalization-battle/>. Accessed December 5, 2017
3. Liu JB, Liu Y, Cohen ME, Ko CY, Sweitzer BJ: Defining the intrinsic cardiac risks of operations to improve preoperative cardiac risk assessments. *ANESTHESIOLOGY* 2018; 128:283–92
4. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, Cohen ME: Development and evaluation of the universal ACS NSQIP surgical risk calculator: A decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg* 2013; 217:833–42.e1–3
5. Lee TH, Marcantonio ER, Mangione CM, Thomas EJ, Polanczyk CA, Cook EF, Sugarbaker DJ, Donaldson MC, Poss R, Ho KK, Ludwig LE, Pedan A, Goldman L: Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation* 1999; 100:1043–9
6. McMillan MT, Allegrini V, Asbun HJ, Ball CG, Bassi C, Beane JD, Behrman SW, Berger AC, Bloomston M, Callery MP, Christein JD, Dickson E, Dixon E, Drebin JA, Fernandez-Del Castillo C, Fisher WE, Fong ZV, Haverick E, Hollis RH, House MG, Hughes SJ, Jamieson NB, Kent TS, Kowalsky SJ, Kunstman JW, Malleo G, McElhany AL, Salem RR, Soares KC, Sprys MH, Valero V III, Watkins AA, Wolfgang CL, Zureikat AH, Vollmer CM Jr: Incorporation of procedure-specific risk

into the ACS-NSQIP Surgical Risk Calculator improves the prediction of morbidity and mortality after pancreatoduodenectomy. *Ann Surg* 2017; 265:978–86

7. 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
8. Liu Y, Cohen ME, Ko CY, Bilimoria KY, Hall BL: Considerations in releasing equations for the American College of Surgeons NSQIP Surgical Risk Calculator: In reply to Wanderer and Ehrenfeld. *J Am Coll Surg* 2016; 223:674–5
9. Cohen ME, Liu Y, Ko CY, Hall BL: Improved surgical outcomes for ACS NSQIP hospitals over time: Evaluation of hospital cohorts with up to 8 years of participation. *Ann Surg* 2016; 263:267–73
10. American College of Surgeons: Inspiring quality: Highest standards, better outcomes. 2017. Available at: <http://www.facs.org/about-acsc>. Accessed December 5, 2017