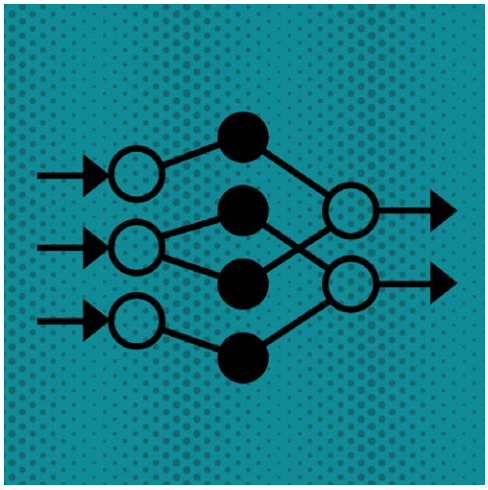
Machine Learning Comes of Age

Local Impact *versus* National Generalizability

Michael L. Burns, M.D., Ph.D., Sachin Kheterpal, M.D., M.B.A.

Machine learning, a subfield of artificial intelligence, is an increasingly popular topic within medicine. Evangelists of machine learning hope that it will revolutionize health care. While machine learning may still be in the "hype" phase of excitement, we are beginning to see applications within perioperative medicine with potential perioperative clinical impact.[1] Addressing meaningful problems that may decrease patient harm, improve quality of life, or reduce administrative burden is an important goal when implementing machine learning in health care.



"Readers should demand positive and negative predictive values when assessing whether or not to implement the latest trend in prediction science."

In this issue of Anesthesiology, Mišić *et al.*[2] evaluate various machine learning techniques for predicting 30-day postoperative readmissions. Hospital readmissions are costly and common events that are the target of healthcare improvement and policy change initiatives, but there are broader implications in the article by Mišić *et al.* All anesthesiologists should note that this work calls into question the purported value of: (1) advanced model diagnostics that are difficult to interpret; (2) using thousands of data elements to predict outcomes *versus* parsimonious approaches; (3) focusing on multicenter "generalizability" of prediction models rather than just optimizing for future predictions at a given hospital; and (4) advanced machine learning algorithms *versus* classic techniques.

The authors are to be commended for focusing on simplicity and interpretability. Despite the novelty of machine learning, traditional statistical methods such as positive and negative predictive values can be used to compare machine learning against traditional modeling techniques. Positive predictive value is the probability that the model in question correctly labels a true positive event (*e.g.*, how many patients with a "positive" model result for readmission actually experience a readmission?). Similarly, negative predictive value is the probability the model correctly labels negative results. While the perioperative literature is filled with c-statistics, positive and negative predictive values reflect the clinicians' perspective. The authors use these intuitive measures when evaluating their results. This focus on clinical value demonstrates that while different models may have similar c-statistics, the positive predictive value can vary significantly. Readers should demand positive and negative predictive values when assessing whether or not to implement the latest trend in prediction science.

Next, the authors demonstrate that massive volumes of data may not result in better predictions than reasonably comprehensive data streams. Two indices are commonly used for predicting hospital readmissions: the LACE (Length of Stay, Acuity of admissions, Comorbidities, Emergency department visits)[3] and HOSPITAL (Hemoglobin level at discharge, discharged from Oncology service, Sodium levels at discharge, any ICD-9 coded Procedure performed during hospital stay, Index admission Type, number of hospital Admissions during the previous year, Length of stay) scoring models.[4] Both LACE and HOSPITAL use patient electronic health record data to calculate a continuous score to predict an unplanned hospital readmission within 30 days of patient discharge. These indices focus on slightly different patient populations. HOSPITAL focuses on patients discharged from medical services, while LACE was developed for use in both medical and surgical patients. Both depend upon data available at the time of discharge, limiting the ability to identify patients at high risk of readmission early in their

Image: J. P. Rathmell/The Noun Project.

This editorial accompanies the article on p. 968.

Accepted for publication January 13, 2020. Published online first on March 23, 2020. From the Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, Michigan.

 Anesthesiology 2020; 132:939–41. DOI: 10.1097/ALN.0000000000003223

<zdoi;. DOI: 10.1097/ALN.0000000000003223>

stays. In the current work, the authors considered five types of data that are available throughout a patient admission: general (patient demographics, diagnoses, and surgery duration), laboratory testing, medications administered, provider teams, and surgeon billing codes. They observe an improvement over both LACE and HOSPITAL scoring indices, boosting c-statistics from 0.73 to 0.87 using the machine learning approaches. Surprisingly, their results showed that adding features such as medications and provider team did not meaningfully improve model performance beyond basic patient demographics, diagnoses, and laboratory values. These few data types appear to incorporate the predictive value of the many other healthcare data elements generated during an inpatient stay and at discharge. The authors demonstrate that efforts to aggregate many different data sources into machine learning models to predict clinical events may only provide minimal incremental value.

Most importantly, this study raises important questions about the value of assessing generalizability of machine learning models for use outside of the institution from which the training data was derived. Both HOSPITAL and LACE used logistic regression in model development and tested generalizability by comparison across multiple hospitals. The HOSPITAL readmission scoring system was developed using data from a single healthcare institution and later validated with an international multicenter study.[5] LACE was developed using data from 11 hospitals across five cities in Ontario, Canada. The generalizability of specific indices and models is usually considered a primary feature when evaluating a model. With most studies, there exists a large gap between model cross-center generalizability and local accuracy. While human pathophysiology should be similar from one hospital to the next, clinical processes and hospital structures of care will vary, making models such as those created for predicting readmission difficult to generalize.

However, it is important to consider the purpose of the model: risk adjustment for multicenter comparison *versus* optimal local performance to change individual patient care. The desired purpose may drive whether a nationally validated model *versus* a locally curated and temporally validated model is desired. If the purpose is to change individual patient care at a specific hospital, a generalizable methodology and temporal validation may be the correct path forward. In temporal validation, a model is evaluated by testing its performance using data from a time period after the derivation cohort. In the study by Mišić *et al.*,[2] the authors used data from 2013 to 2016 to develop three machine learning models and compared to existing LACE and HOSPITAL at their home institution. They demonstrate clearly better performance, which may be expected given that the models were "tuned" using local data. Importantly, the models are then evaluated using 2017 and 2018 data at their local hospital and perform well, correctly identifying 39% of readmissions. Temporal validation should be considered an important validation methodology if the purpose of model development is direct point of care change. If the goal of a model is to compare hospitals, then generalizability across locations is important.

Finally, a downside of machine learning is that as models become more sophisticated, their interpretability and reproducibility worsens, challenging implementation within healthcare systems. While popular techniques such as random forest and gradient boosted trees were considered, the more classic machine learning technique of L1 logistic regression demonstrated superior performance. L1 logistic regression builds upon classic logistic regression by handling a larger number of candidate independent variables without risking overfitting. Moving forward, there seems to be limited value to overcomplicating models; adequate model performance and clinical value may be achieved using parsimonious data fields and interpretable models. The ideal scenario of creating models generalizable across all hospitals may be realized by generating site-specific models using reusable techniques. This portends a healthcare industry where hospital quality improvement staffing teams will include data scientists capable of implementing publicly available machine learning models. The use of publicly available techniques that are adapted locally may decrease the need for advanced skills in data extraction, model development, and model tuning.

There is a lot of excitement surrounding the use of machine learning in medicine. It is critical to understand how to apply these technologies and, as the work of Mišić *et al.* suggests, creating standard methodology is an important first step. While much work remains to realize the potential of improved prediction,[6] the framework for use of machine learning in perioperative medicine is beginning to take shape.

Competing Interests

Drs. Burns and Kheterpal are listed as a co-inventors on a patent application 62/791,257 entitled "Automated System And Method For Assigning Billing Codes To Medical Procedures" related to the use of machine learning techniques for anesthesia procedure billing.

Correspondence

Address correspondence to Dr. Kheterpal: sachinkh@med.umich.edu

References

1. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, Rinehart J, Cannesson M: Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. Anesthesiology 2018; 129:663–74
2. Mišić VV, Gabel E, Hofer I, Kumar R, Mahajan M: Machine learning prediction of postoperative emergency department hospital readmission. Anesthesiology 2020; 132:968–80

3. Walraven C van, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, Austin PC, Forster AJ: Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ 2010; 182:551–7
4. Donzé J, Aujesky D, Williams D, Schnipper JL: Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA Intern Med 2013; 173:632–8
5. Donzé JD, Williams MV, Robinson EJ, Zimlichman E, Aujesky D, Vasilevskis EE, Kripalani S, Metlay JP, Wallington T, Fletcher GS, Auerbach AD, Schnipper JL: International validity of the HOSPITAL score to predict 30-day potentially avoidable hospital readmissions. JAMA Intern Med 2016; 176:496–502
6. Finkelstein A, Zhou A, Taubman S, Doyle J: Health care hotspotting — a randomized, controlled trial NEJM 2020; 382:152–62