Assessing the risk of individual patients for experiencing specific clinical outcomes is essential for clinicians to make decisions. In practice, many of these clinical outcomes are rare, which makes the risk assessment process like finding needles in a haystack. Machine learning (ML) aims to develop computational algorithms from massive data to extract informative signals and holds great potentials in this scenario. Recently, ML algorithms have demonstrated the capability to achieve superior performance compared with clinically used risk calculators in outcome prediction with real world patient data, such as claims or electronic health records (EHRs).

In the study from Herrin et al, ML models were trained to predict the risk of gastrointestinal bleed (GIB) in patients who used antithrombotic drugs. This is a critical problem, given the severity of GIB and widespread use of antithrombotic drugs among patients receiving cardiac care, as stated in the study by Herrin et al. Although there are several clinically used risk calculators for GIB, they were mostly derived from data with limited sample sizes and did not take complicated patient contexts, such as drug interactions and comorbidity histories, into consideration. This makes ML models promising for producing better risk predictions, especially considering the fact that GIB is also a rare outcome.

Herrin et al trained the ML models for GIB risk prediction from the OptumLabs Data Warehouse (OLDW), which is a national, large-scale claims database. The deidentified EHRs of more than 300,000 patients using antithrombotic drugs were included in the study. Three ML models, regularized Cox regression (RegCox), extreme gradient boosting (XGBoost), and random survival forest (RSF), were investigated to predict the risk of individual patients experiencing GIB at 6 and 12 months after the first time they were prescribed with antithrombotic drugs. A total of 32 variables, including demographic characteristics, condition groups, comorbidities, and medications, were collected from the baseline period to profile the patients and served as the inputs of the ML models. As a reference, the HAS-BLED score (ie, hypertension, abnormal kidney or liver function, stroke, history of factors associated with presence of bleeding, labile international normalized ratio, older age [>65 years], and use of drugs or alcohol concomitantly), which is a clinically used GIB risk score, was also calculated for each patient. Herrin et al found that HAS-BLED achieved a prediction of GIB risk performance area under the curve (AUC) of 0.6 for 6 months and an AUC of 0.57 for 12 months. RegCox and XGBoost achieved an AUC of 0.67 value on both risks, while the performance of RSF was an AUC of 0.62 at 6 months and an AUC of 0.6 at 12 months. These numbers quantitatively showed that ML models performed better than HAS-BLED score. However, what are the practical implications?

First, the quantitative prediction improvements for ML models over HAS-BLED were marginal based on the AUC values. Herrin et al also discussed how more sophisticated machine learning models and other types of patient information, such as imaging, may further boost the ML model performance. While quantitative performance is absolutely important, there is no perfect threshold of AUC value that means that the model can be used in clinical practice by exceeding such threshold. Moreover, for rare outcomes, AUC can be misleading because of the extreme imbalance between the patients with and without the outcome event. In the study by Herrin et al, the positive predictive values (PPVs) for all ML algorithms and HAS-BLED were low on the validation set (approximately 0.04), which indicates there was a large number of false positives. The optimal cutoff points for all 3
ML algorithms were set very low owing to class imbalance. This also happens in other rare outcome prediction settings, such as for suicide attempts. In one of my recent studies on suicide risk prediction for children and adolescent from EHRs, the ML model achieved an AUC of more than 0.85, but the PPV value was less than 0.1. This implies that when conducting model comparative effectiveness studies on predicting rare outcomes, we should not consider AUC values only. Other performance metrics, such as sensitivity, specificity, and PPV, are also important for us to comprehensively understand the model behaviors.

Second, quantitative performance metrics are not the only factors to determine the clinical utility of an algorithm. Model interpretability or explainability is also crucial. The role of these models in clinical practice is decision support, rather than making decisions. Therefore, clinicians prefer to use models that they can understand and that align well with their own experience and knowledge. This is an important reason why scorecard-type risk calculators, like HAS-BLED, are popular in clinical practice, despite the fact that their quantitative performances may not be high. Although ML models can achieve superior performance, they are usually complex and thus sacrifice interpretability. The study by Herrin et al demonstrated the importance of scores of input variables returned by RegCox, but how much they added to the risk factors used in HAS-BLED was unclear. In addition, HAS-BLED score is simply the sum of individual risk factor scores, while the relationships among the input variables and the GIB risk in the 3 ML models were not as straightforward. This may hinder their clinical utilities.

Third, the sample size used in the study by Herrin et al makes it one of the largest studies of this kind. Large sample size is critical for the problem itself (for capturing more cases in rare outcome prediction) and the ML models (for reliable training). While the study by Herrin et al was made possible through OLDW, it emphasizes the importance of collaborative and open data science. Different than the data in other real-world problems, clinical data are sensitive in nature because they contain protected health information. Therefore, large collaborative clinical research consortiums with rigorous data use and transmission controls would be crucial, and the FAIR (findability, accessibility, interoperability, reusability) principle provides a general guideline for scientific data management and stewardship. In addition, privacy-preserving ML mechanisms are getting more and more attentions in medicine recently. As an example, federated learning, which aims at training an ML model collectively from multiple local data hosting sites without transmitting their raw data, holds great potential in this scenario.

Last but not least, as Herrin et al pointed out, prospective evaluation is essential for assessing the real clinical impact of these risk prediction models. Many factors other than the model itself, such as patient status, clinician behavior, and clinics operation, may impact the model deployment process and associated clinical outcomes. Only after the model is deployed in real clinical workflows, we can comprehensively understand whether and how it can improve the patient outcomes.

The study by Herrin et al demonstrated on a large real-world patient claims data set that ML models can perform better than clinically used risk predictor tools on GIB, which implies the great potential of ML on predicting rare clinical outcomes. This is a good start. Many other factors, including more comprehensive performance evaluation metrics, model interpretability, and data quantity need to be considered for assessing the potential clinical impact of these models. More importantly, efforts on prospective evaluations on clinical ML models with implementation science are critical and urgently needed.
REFERENCES


