Prediction Models

— Why Are They Used or Not Used? —

Fumiaki Nakamura, MD, PhD; Michikazu Nakai, PhD

Predicting a patient's prognosis advances medical decision making in clinical settings. Risk prediction models (also called prognostic models, prediction rules, or risk scores) are tools to estimate individual patients' risk or probability by numerical values. Although many prediction models have been published, few have been used in routine clinical settings because of inconvenience and complexity.¹ In this issue of the Journal, Hu et al show how they applied the CHA₂DS₂-VASc score to predict the incidence of atrial fibrillation (AF) in chronic obstructive pulmonary disease (COPD) patients.² Previous studies have reported prediction models for AF based on community cohorts,^{3–5} but most cardiologists are not familiar with these community-based prediction models. It seems reasonable to evaluate individual risk with what we already know, but we must be aware that CHA₂DS₂-VASc score was developed as the model for predicting ischemic stroke in patients with AF, not for incident AF in COPD patients. It is possible that applying the wrong prediction model may cause over- or underestimation of a patient's risk. In order to reduce this, we need to understand the evaluation methods, risk prediction models and the reasons why high-performance

Article p1792

models are complex.

Performance of Prediction Models

We need to assess the performance of risk prediction models more carefully before applying their results in the clinical setting. The performance of risk prediction models is divided into two components: discrimination and calibration.⁶

Discrimination is the ability of the prediction model to distinguish whether high-risk patients are in fact low risk.⁶ A well-known measurement tool of discrimination is the receiver-operating characteristic (ROC) curve or the C-statistic. The C-statistic defines the probability that a randomly selected patient who developed an event had a higher risk score than a patient who had not developed the event. The ROC curve draws the plot of the sensitivity and (1–specificity) for all possible cutoff points (**Figure 1**). When the outcome is a binary event, the area under the ROC curve (AUC) is equivalent to the C-statistic, which ranges from 0.5 to 1, where 0.5 indicates no discriminatory ability and

Figure 1. Example of a receiver-operating characteristic (ROC) curve. The area under the ROC curve (AUC) is equivalent to the C-statistic.

Figure 2. Example of graphical calibration. The gap between the actual and predicted values should be small.

The opinions expressed in this article are not necessarily those of the editors or of the Japanese Circulation Society.

Received October 27, 2017; accepted October 28, 2017; released online November 10, 2017

Center for Cardiovascular Disease Information, National Cerebral and Cardiovascular Center, Suita, Japan

Mailing address: Fumiaki Nakamura, MD, PhD, Center for Cardiovascular Disease Information, National Cerebral and Cardiovascular Center, 5-7-1 Fujishirodai, Suita 565-8565, Japan. E-mail: f-naka@ncvc.go.jp

ISSN-1346-9843 All rights are reserved to the Japanese Circulation Society. For permissions, please e-mail: cj@j-circ.or.jp

1 indicates perfect discriminatory ability. If the C-statistic of a prediction model is low, then it shows that the model does not classify patients to the correct risk category. When dealing with censored data, the ROC curve and AUC are not appropriate to assess discriminatory performance. In such a case, Harrell's C-statistic must be used and interpretation of the C-statistic is equivalent to the AUC. In addition to the C-statistic and AUC, other statistical methods such as integrated discrimination improvement or net reclassification index are also recommended when comparing multiple prediction models for the same outcome.

Calibration is the ability to estimate the accuracy of a model's prediction.⁶ When the calibration of a prediction model is poor, the model over- or underestimates the absolute probability of the outcome event, no matter how good the discrimination of a model is. Therefore, it is important to keep the calibration value as high as possible. We have two methods of measuring the calibration of a model: statistical or graphical. The common statistical method of calculating calibration in the medical research field is the Hosmer-Lemeshow test. This method usually divides by quintile or decile and compares the percentage of both the predicted values and observed values. We can also evaluate the calibration of a prediction model by graph to compare predicted and observed values at different levels (Figure 2). In order to compare the calibrations of different prediction models, Akaike information criteria or Bayesian information criteria can be applied and lower values with either of these indexes suggest better calibration of the model.

Why Is the CHA₂DS₂-VAsC Score Widely Used?

CHA₂DS₂-VAsC score is widely used to determine the initiation of anticoagulant therapy for AF patients to prevent ischemic stroke. However, the original paper reported fairly moderate discriminatory power.⁷ Suzuki et al also stated that the C-statistic of the CHA₂DS₂-VAsC score was 0.671 (95% confidence interval: 0.606–0.736) in pooled data from four Japanese registries.⁸ In general, a prediction model based on a mathematical equation will classify patients more accurately. Recent research shows that ATRIA, which is a more complex risk score, performs better than CHA₂DS₂-VAsC.^{9,10} The ATRIA score is based on regression coefficients and the interaction term of 'prior stroke' is additionally considered, whereas the CHA₂DS₂-VAsC score is not based on regression coefficients, which is technically considered an incorrect method of constructing a prediction models,¹¹ and interaction terms are not considered.

However, physicians tend to prefer a concise model rather than a complex model even when the complex model has higher discriminatory power. Kappen et al¹² point out 4 perceptual barriers to using risk prediction models in clinical practice: (1) "the predicted outcome is not the main area of attention for physicians"; (2) "the decision-making process of physicians is intuitive rather than analytical"; (3) "the probabilistic knowledge of the outcome is difficult to use in decision making"; and (4) "a prediction model does not weigh the benefits and risks of prophylactic drugs with regard to the patient's comorbidity". This may be why the CHA₂DS₂-VAsC score is widely used.

Prediction Models in the Precision Medicine Era

Precision medicine is a revolutionary approach that takes into account individual differences in lifestyle, environment

and biology beyond traditional personalized medicine.¹³ The CHA₂DS₂-VAsC score is classified as non-sufficient for precision medicine because of its moderate discriminatory power. We must effectively use computer-based complex models such as machine learning, which is beyond our intuition, to accomplish precision medicine.¹⁴ However, prediction models only suggest the probability of an event. It is still up to us to make the decision. Thus, in the precision medicine era, physicians must understand the results of prediction models, and, using this tool, must be able to communicate these results to patients to assist them their own decision making.¹⁵

References

- Wyatt JC, Altman DG. Commentary: Prognostic models: Clinically useful or quickly forgotten? *BMJ* 1995; **311**: 1539–1541.
- Hu WS, Lin CL. Use of CHA₂DS₂-VAsC score to predict new-onset atrial fibrillation in chronic obstructive pulmonary disease patients: Large-scale longitudinal study. *Circ J* 2017; **81**: 1792–1797.
- Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): A community-based cohort study. *Lancet* 2009; **373**: 739–745.
- Chamberlain AM, Agarwal SK, Folsom AR, Soliman EZ, Chambless LE, Crow R, et al. A clinical risk score for atrial fibrillation in a biracial prospective cohort (from the Atherosclerosis Risk in Communities [ARIC] study). *Am J Cardiol* 2011; **107**: 85–91.
- Kokubo Y, Watanabe M, Higashiyama A, Nakao YM, Kusano K, Miyamoto Y. Development of a basic risk score for incident atrial fibrillation in a Japanese general population: The Suita Study. *Circ J* 2017; **81**: 1580–1588.
- Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA* 2017; **318**: 1377–1384.
- Lip GY, Nieuwlaar R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: The Euro Heart Survey on Atrial Fibrillation. *Chest* 2010; **137**: 263–272.
- Suzuki S, Yamashita T, Okumura K, Atarashi H, Akao M, Ogawa H, et al. Incidence of ischemic stroke in Japanese patients with atrial fibrillation not receiving anticoagulation therapy: Pooled analysis of the Shinken Database, J-RHYTHM Registry, and Fushimi AF Registry. *Circ J* 2015; **79**: 432–438.
- van den Ham HA, Klungel OH, Singer DE, Leufkens HG, van Staa TP. Comparative performance of ATRIA, CHADS₂, and CHA₂DS₂-VAsC risk scores predicting stroke in patients with atrial fibrillation: Results from a national primary care database. *J Am Coll Cardiol* 2015; **66**: 1851–1859.
- Aspberg S, Chang Y, Atterman A, Bottai M, Go AS, Singer DE. Comparison of the ATRIA, CHADS₂, and CHA₂DS₂-VAsC stroke risk scores in predicting ischaemic stroke in a large Swedish cohort of patients with atrial fibrillation. *Eur Heart J* 2016; **37**: 3203–3210.
- Harrell F. Regression coefficients and scoring rules. *J Clin Epidemiol* 1996; **49**: 819.
- Kappen TH, van Loon K, Kappen MA, van Wolfswinkel L, Vergouwe Y, van Klei WA, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. *J Clin Epidemiol* 2016; **70**: 136–145.
- National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. The National Academies Collection: Reports funded by National Institutes of Health. Washington (DC): National Academies Press (US); 2011.
- Deo RC. Machine learning in medicine. *Circulation* 2015; **132**: 1920–1930.
- Houser SR. The American Heart Association's New Institute for Precision Cardiovascular Medicine. *Circulation* 2016; **134**: 1913–1914.