JAMA Guide to Statistics and Methods

# Evaluating Discrimination of Risk Prediction Models
## The C Statistic

Michael J. Pencina, PhD; Ralph B. D'Agostino Sr, PhD

**Risk prediction models** help clinicians develop personalized treatments for patients. The models generally use variables measured at one time point to estimate the probability of an outcome occurring within a given time in the future. It is essential to assess the performance of a risk prediction model in the setting in which it will be used. This is done by evaluating the model's discrimination and calibration. *Discrimination* refers to the ability of the model to separate individuals who develop events from those who do not. In time-to-event settings, discrimination is the ability of the model to predict who will develop an event earlier and who will develop an event later or not at all. *Calibration* measures how accurately the model's predictions match overall observed event rates.

Related article page 1030

In this issue of *JAMA*, Melgaard et al used the C statistic, a global measure of model discrimination, to assess the ability of the $CHA_2DS_2$-VASc model to predict ischemic stroke, thromboembolism, or death in patients with heart failure and to do so separately for patients who had or did not have atrial fibrillation (AF).[1]

## Use of the Method

### Why Are C Statistics Used?

The C statistic is the probability that, given 2 individuals (one who experiences the outcome of interest and the other who does not or who experiences it later), the model will yield a higher risk for the first patient than for the second. It is a measure of concordance (hence, the name "C statistic") between model-based risk estimates and observed events. C statistics measure the ability of a model to rank patients from high to low risk but do not assess the ability of a model to assign accurate probabilities of an event occurring (that is measured by the model's calibration). C statistics generally range from 0.5 (random concordance) to 1 (perfect concordance).

C statistics can also be thought of as being the area under the plot of sensitivity (proportion of people with events for whom the model predicts are high risk) vs 1 minus specificity (proportion of people without events for whom the model predicts are high risk) for all possible classification thresholds. This plot is called the receiver operating characteristic (ROC) curve, and the C statistic is equal to the area under this curve.[2] For example, in the study by Melgaard et al, $CHA_2DS_2$-VASc scores ranged from a low of 0 (heart failure only) to a high of 5 or higher, depending on the number of comorbidities a patient had. One point on the ROC curve would be when high risk is defined as a $CHA_2DS_2$-VASc score of 1 or higher and low risk as a $CHA_2DS_2$-VASc score of 0. Another point on the curve would be when high risk is defined as a $CHA_2DS_2$-VASc score of 2 or higher and low risk as a $CHA_2DS_2$-VASc score of lower than 2, etc. Each cut point is associated with a different sensitivity and specificity.

It is useful to quantify the performance and clinical value of predictive models using the positive predictive value (PPV; the proportion of patients in whom the model predicts an event will occur who actually have an event) and the negative predictive value (NPV; the proportion of patients whom the model predicts will not have an event who actually do not experience the event). An important measure of a model's misclassification of events is 1 minus NPV, or the proportion of patients the model predicts will not have an event who actually have the event. The PPV and 1 minus NPV can be more informative for individual patients than the sensitivity and specificity because they answer the question "What are this patient's chances of having an event when the model predicts they will or will not have one?" If the event rate is known, then the PPV and NPV can be estimated based on sensitivity and specificity and, hence, the C statistic can be viewed as a summary for both sets of measures.

### What Are the Limitations of the C Statistic?

The C statistic has several limitations. As a single number, it summarizes the discrimination of a model but does not communicate all the information ROC plots contain and lacks direct clinical application. The NPV, PPV, sensitivity, and specificity have more clinical relevance, especially when presented as plots across all meaningful classification thresholds (as is done with ROCs). A weighted sum of sensitivity and specificity (known as the standardized net benefit) can be plotted to assign different penalties to the 2 misclassification errors (predicting an individual who ultimately experiences an event to be at low risk; predicting an individual who does not experience an event to be at high risk) according to the principles of decision analysis.[3,4] In contrast, the C statistic does not effectively balance misclassification errors.[5] In addition, the C statistic is only a measure of discrimination, not calibration, so it provides no information regarding whether the overall magnitude of risk is predicted accurately.

### Why Did the Authors Use C Statistics in Their Study?

Melgaard et al[1] sought to determine if the $CHA_2DS_2$-VASc score could predict occurrences of ischemic stroke, thromboembolism, or death among patients who have heart failure with and without AF. The authors used the C statistic to determine how well the model could distinguish between patients who would or would not develop each of the 3 end points they studied. The C statistic yielded the probability that a randomly selected patient who had an event had a risk score that was higher than a randomly selected patient who did not have an event.

### How Should the Findings Be Interpreted?

The value of the C statistic depends not only on the model under investigation (ie, $CHA_2DS_2$-VASc score) but also on the distribution of risk factors in the sample to which it is applied. For example, if age is an important risk factor, the same model can appear to perform

much better when applied to a sample with a wide age range compared with a sample with a narrow age range.

The C statistics reported by Melgaard et al[1] range from 0.62 to 0.71 and do not appear impressive (considering that a C statistic of 0.5 represents random concordance). This might be due to limitations of the model; eg, if there were an insufficient number of predictors or the predictors had been dichotomized for simplicity. The nationwide nature of the data used by Melgaard et al suggests that the unimpressive values of the C statistic cannot be attributed to narrow ranges of risk factors in the analyzed cohort. Rather, it might suggest inherent limitations in the ability to discriminate between patients with heart failure who will and will not die or develop ischemic stroke or thromboembolism.

The C statistic analysis suggested that the $CHA_2DS_2$-VASc model performed similarly among heart failure patients with and without AF (C statistics between 0.62 and 0.71 among patients with AF and 0.63 to 0.69 among patients without AF). An additional insight emerges from NPV analysis looking at misclassification of events occurring at 5 years, however. Between 19% and 27% of patients without AF who were predicted to be at low risk actually had 1 of the 3 events and thus were misclassified, yielding an NPV of 73% to 82%. Between 24% and 39% of patients with AF whom the model classified as low risk had major events, yielding an NPV of 61% to 76%. Because there was less misclassification among patients without AF who were predicted to be at low risk, a $CHA_2DS_2$-VASc score of 0 is a better determinant of long-term low risk among patients without AF than patients with AF. This aspect of the model performance is not apparent when looking at C statistics alone.

## Caveats to Consider When Using C Statistics to Assess Predictive Model Performance

Special extensions of the C statistic need to be used when applying it to time-to-event data[6] and competing-risk settings.[7] Furthermore, there exist several appealing single-number alternatives to the C statistic. They include the discrimination slope, the Brier score, or the difference between sensitivity and 1 minus specificity evaluated at the event rate.[3]

The C statistic provides an important but limited assessment of the performance of a predictive model and is most useful as a familiar first-glance summary. The evaluation of the discriminating value of a risk model should be supplemented with other statistical and clinical measures. Graphical summaries of model calibration and clinical consequences of adopted decisions are particularly useful.[8]


### ARTICLE INFORMATION

**Author Affiliations:** Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, Durham, North Carolina (Pencina); Department of Mathematics and Statistics, Boston University, Boston, Massachusetts (D'Agostino).

**Corresponding Author:** Michael J. Pencina, PhD, Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, 2400 Pratt St, Durham, NC 27705 (michael.pencina@duke.edu).

**Section Editors:** Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

**Conflict of Interest Disclosures:** The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.



### REFERENCES

1. Melgaard L, Gorst-Rasmussen A, Lane DA, Rasmussen LH, Larsen TB, Lip GYH. 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*. doi:10.1001/jama.2015.10725.

2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.

3. Pepe MS, Janes H. Methods for evaluating prediction performance of biomarkers and tests. In: Lee M-LT, Gail M, Pfeiffer R, Satten G, Cai T, Gandy A, eds. *Risk Assessment and Evaluation of Predictions*. New York, NY: Springer; 2013:107-142.

4. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574.

5. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn*. 2009;77:103-123.

6. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*. 2004;23(13):2109-2123.

7. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med*. 2013;32(30):5381-5397.

8. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.