Please join the Biometric Colloquium

Ben van Calster

Department of Development and Regeneration
KU Leuven, Leuven, Belgium

Performance evaluation of predictive AI models to support medical decisions: overview and guidance

June 18th, 2025 at 9:00 am

Hörsaalzentrum AKH,
Kursraum 25, Ebene 8
Medical University of Vienna, 1090 Wien
Host: Georg Heinze

Abstract:

The scientific literature is packed with measures to assess performance of predictive artificial intelligence (AI) models, and with papers criticizing or supporting one specific measure. We aim to assess the merits of 32 key classic and contemporary performance measures when validating models with a binary outcome that are intended to be used in medical practice to support clinical decision-making. As a case study, we externally validate the ADNEX model for ovarian cancer diagnosis.

We classify the performance measures and accompanying plots in five performance domains (discrimination, calibration, overall, classification, and clinical utility). The first four domains focus on statistical performance, the fifth domain on decision-analytic performance. We argue that performance measures should have two key characteristics: (1) properness (whether the measure's expected value is optimized when it is calculated using the correct probabilities), and (2) focus (whether the measure reflects either a purely statistical aspect of performance, or reflects decision-analytic performance by properly considering misclassification costs).

Fifteen measures violate one or both key characteristics, such that we warn against their use. Interestingly, all classification measures (including F1) are improper for a clinically relevant decision threshold *t*. We recommend to always report the following core set of measures/plots: a risk distribution plot, the c statistic (AUROC), a calibration plot, and a clinical utility measure such as net benefit or expected cost with a decision curve.