Thursday 25 July 2024

09:00-12:30 Mini Symposium 2 (Room 2)

STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative – recent progress and foci for the future

Organizers: Willi Sauerbrei and Els Goetghebeur in collaboration with the STRATOS Steering Group

Assessing performance when developing or validating clinical risk prediction models in the era of machine learning

Ben van Calster, Ewout Steyerberg for TG6.

An abundance of performance measures for clinical risk prediction models have been proposed in the statistical and machine learning literature. We aim to provide an overview of contemporary performance measures for models with binary outcomes, motivated by the assessment of the value of the previously developed ADNEX model to predict whether an ovarian tumor is malignant in external validation data (n=894, 49% malignant tumors). We consider five domains of model performance. These include overall measures (e.g. Brier score), measures for discrimination (e.g. AUROC), and measures of calibration (e.g. expected calibration error). When supporting a clinical decision for the patient, a decision threshold on the estimated risk is required to define classification as high versus low risk. The 2x2 table of classification versus outcomes can be described with classification measures (e.g. F1) and clinical utility measures (e.g. net benefit). We discuss 32 common performance measures (9 overall, 3 discrimination, 6 calibration, 11 classification, 3 utility). For each performance domain, matching graphical assessments are available. We define three key desirable characteristics for performance measures: properness (i.e. whether the value of the measure is optimal when the correct risks are used); having an understandable interpretation; and having a clear focus by targeting only one of the five domains. The majority of measures fail for at least one characteristic, while the F1 score fails at all three. All considered classification measures at a given threshold t are improper. A natural requirement is that a performance measure should match the intended use of the model. We discern three common situations. First, when externally validating models that aim to support clinical decision making, it makes sense to assess performance in the following order: discrimination (AUROC), calibration (calibration plot) and clinical utility (net benefit). Second, if a model is merely used for informing/counseling patients about their risk, external validation should focus on calibration. Third, when methodologically comparing multiple models, overall measures are useful. Other measures may be added, if they meet the three key characteristics. In conclusion, we recommend to consider a limited set of key measures to assess performance aspects in relation to the intended use of a prediction model, focusing on (semi-)proper measures with a clear interpretation and focus.