Testing for Improvement in Prediction Model Performance

Margaret Pepe*, University of Washington

Abstract

New methodology has been proposed in recent years for evaluating the improvement in prediction performance gained by adding a new predictor, Y, to a risk model containing a set of baseline predictors, X, for predicting a binary outcome D. We prove theoretically that null hypotheses concerning improvement in ROC performance, integrated discrimination improvement (IDI), change in population distribution of risk and some risk reclassification measures are all equivalent to the simple null hypothesis that the coefficient for Y is zero in the risk model, $P(D=1|X,Y)$. Therefore, testing for improvement in prediction performance is redundant if Y has already been shown to be a risk factor.

We investigate properties of tests through simulation studies, focusing on the change in the area under the ROC curve (AUC). An unexpected finding is that standard testing procedures that do not adjust for variability in estimated regression coefficients are extremely conservative. This may explain why the AUC is widely considered insensitive to improvements in prediction performance and suggests that the problem of insensitivity is with use of invalid procedures for inference rather than with the measure itself. To avoid redundant testing and use of potentially problematic methods for inference, we recommend that hypothesis testing be limited to evaluation of Y as a risk factor in the risk model, for which methods are well developed and widely available. Analyses of measures of prediction performance should focus on estimation rather than on testing.

We evaluate various estimators and confidence intervals for the improvement in the AUC. We apply the methods to a dataset concerning prediction of renal artery stenosis where serum creatinine level is found to improve prediction beyond a set of demographic and clinical variables.

* Presenting author