Tim Vigers

4/21/19

Evaluating Models

Predicting the probability of developing certain diseases is an important part of biostatistical work, and “multivariable prediction algorithms are among the major advances made” in 20th century medical research. [1] However, risk prediction is a continually developing process, and it can be difficult to evaluate how well potential new risk factors improve the existing models. In general, these kinds of model are evaluated based on ROC AUC, but when adding new markers to models that already discriminate relatively well, the effect of the marker needs to be enormous to meaningfully impact AUC. [1] Because of these limitations of AUC, researchers have looked at alternative methods of evaluating models.

Pencina et al. propose two new methods: the net reclassification improvement (NRI) and integrated discrimination improvement (IDI). NRI is a form of reclassification that treats those who develop the disease separately from those who do not. Subjects are classified using the old model and then the new model, and NRI is calculated as “a sum of differences in proportions of individuals moving up minus the proportion moving down for people who develop events, and the proportion of individuals moving down minus the proportion moving up for people who do not develop events” and its significance can be evaluated with a simple asymptotic test. [1]

IDI is the integral of sensitivity (IS) of the new model minus IS of the old model, minus the difference in integral of one minus specificity (IP) of the two models, which “can also be seen as an integrated difference in Youden’s indices.” [1] Under the null hypothesis, IDI = 0 (no difference between the models), so the significance of IDI can also be evaluated with a simple asymptotic test. The IDI is very similar to the AUC in that both are averages of model sensitivity, but the IDI is weighted differently.

Both the IDI and NRI appear to be more sensitive to new markers than AUC and warrant further investigation in evaluating model fit. Though promising, the two metrics still require calibration testing with real world data (particularly IDI, which may be sensitive to model calibration), and researchers need to investigate optimal cutoffs for what counts as a meaningful improvement (which may be disease-specific).

Questions

1. Have these methods been calibrated or tested with different cohorts/diseases? How well did they perform?
2. My first reaction reading the Pencina paper was in line with the scientists who “have argued that we need to wait for new and better markers” [1] and I’m still not particularly convinced that we should move away from AUC. Why not just lower our threshold for what counts as a meaningful improvement in AUC? Particularly given the fact that IDI is essentially just a transformation of AUC anyway.

References

1. Pencina, M.J., et al., *Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond.* Stat Med, 2008. **27**(2): p. 157-72; discussion 207-12.