9-12-2013

Net Reclassification Index: a Misleading Measure of Prediction Improvement

Margaret Sullivan Pepe
Fred Hutchinson Cancer Rsrch Center, mspepe@uw.edu

Holly Janes
Fred Hutchinson Cancer Rsrch Center, hjanes@scharp.org

Kathleen F. Kerr
University of Washington, katiek@u.washington.edu

Bruce M. Psaty
University of Washington, psaty@u.washington.edu

Methods

We used a published simulated dataset that is available online (http://labs.fhcrc.org/pepe/dabs/datasets.html). The data comprise a cohort of 10,000 subjects followed for 10 years for cardiovascular events. Each subject has a covariate called the baseline risk score, and 5 markers labeled Y, M_1, M_2, M_3, M_4. By design, Y is predictive while M_1-M_4 are not. We used likelihood-ratio statistics to evaluate the significance of adding Y or M_1-M_4 to the baseline score.

For model development, we selected subjects randomly with probability .04 (n=419 were selected). Using this training dataset, we fit three logistic regression models (Table 1), a baseline model and two expanded models.

The development and assessment of risk models in the same dataset leads to optimistic estimates of prediction performance. An independent validation dataset is generally considered ideal for obtaining an unbiased assessment of the performance of risk models derived from a model development dataset. We used the remaining 9581 subjects as an independent validation dataset. We calculated three statistics to compare the prediction performance of the baseline model with the expanded models: the ΔAUC statistic, the category-free NRI[2] and a two-category NRI[1] with threshold at 1% risk. Standard p-values were calculated.[1,2]

We also repeated the exercise 100 times, each time randomly selecting 419 subjects for model fitting and assessing performance with the remaining 9581 validation set subjects.

Results

In the entire cohort, likelihood-ratio statistics confirm that Y is predictive (p<0.001). M_1-M_4 are not significant (p=0.503).

Table 2 shows the *ΔAUC* and NRI statistics calculated in the validation dataset. All measures correctly indicate that Y improves prediction. Because M_1-M_4 are not associated with outcome, these markers simply add random noise to the baseline model: as a result the risk model with M_1-M_4 is actually *less* predictive than the baseline model. The *ΔAUC* measure, having a negative sign, corroborates this fact (p < 0.001). However, both *NRI* statistics are positive (p<.001) and provide statistically significant support for the incorrect conclusion that prediction is improved by M_1-M_4 .

Of the 100 repetitions of this exercise, the category-free NRI and two-category NRI statistics were significantly positive in 62% and 82% of validation datasets, respectively. In contrast, the *ΔAUC* was statistically significantly positive in 0%.

Discussion

In the illustrative dataset, the *ΔAUC* correctly indicated that including M_1-M_4 in the model reduces prediction performance while the *NRI* statistics erroneously indicated that M_1-M_4 add predictive information. This phenomenon was not an anomaly of one specific choice for training and test datasets. Moreover, extensive simulation studies and mathematical arguments presented recently show that this is a general phenomenon.[4,5] The problem is more severe when training sets are small and several candidate predictors are studied.

The *ΔAUC* statistic is considered an "insensitive" measure, possibly due to its narrow scale or use of invalid conservative p-values in training datasets.[6] However, the *NRI* suffers a more serious problem: being "too sensitive" even to non-existent improvements in prediction. We recommend avoiding use of the NRI in practice.

References

1. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr , Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med.* 2008;27(2):157-172.
2. Pencina MJ, D'Agostino Sr RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med.* 2011;30(1):11-21.
3. Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: A scientific statement from the American Heart Association. *Circulation.* 2009;119(17):2408-2416.
4. Hilden J, Gerds T. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index [published online ahead of print April 2, 2013]. *Stat Med. doi: 10.1002/sim.5804.*
5. Pepe MS, Fan J, Feng Z, Gerds T, Hilden J. The Net Reclassification Index (NRI): a misleading measure of prediction improvement with mis-calibrated or over fit models.

UW Biostatistics Working Paper Series. Working Paper 392

http://biostats.bepress.com/uwbiostat/paper392

6. Pepe, MS, Kerr, KF, Longton, G and Wang, Z. Testing for improvement in prediction model performance [published online ahead of print April 30, 2013]. *Stat Med.* doi: 10.1002/sim.5727

Table 1. Log odds ratios (OR) for models fit using the model development training dataset (*n*=419). The expanded models included the baseline score and either *Y* or the four marker panel: M_1-M_4 .

Predictors	**Baseline**	**Y added**	**M_1-M_4 added Log(OR)**
baseline score	2.03	1.99	2.09
Y	—	0.81	—
M_1	—	—	–0.21
M_2	—	—	–0.57
M_3	—	—	–0.30
M_4	—	—	0.20
intercept	–4.39	–4.66	–4.64

Table 2. Independent validation data (n=9581) estimates of measures of improved prediction for models fit with a training data set (Table 1).

	Markers Added to the Baseline Risk Score	
Performance Measure	**M1**	**M2--M5**
Δ AUC	0.034	-.012
P-value	<.001	<.001
Category-free NRI	0.690	0.129
P-value	<.001	<.001
Two-category NRI	0.065	0.023
P-value	<.001	<.001