Elsevier Editorial System(tm) for Urology

Manuscript Draft

Manuscript Number:

Title: Evaluating the Quality of Predictive Models for Classification

Article Type: Letter-to-the-Editor

Section/Category:

Keywords: neural network, discrimination, calibration

Corresponding Author: Dr. Jainn-Shiun Chiu, MD

Corresponding Author's Institution: Buddhist Dalin Tzu Chi General Hospital

First Author: Jainn-Shiun Chiu, MD

Order of Authors: Jainn-Shiun Chiu, MD; Yuh-Feng Wang, MS, MD; Yu-Chuan Li, MD, PhD; Min-Huei Hsu, MS, MD

Manuscript Region of Origin:

Abstract:

Evaluating the Quality of Predictive Models for Classification

Jainn-Shiun Chiu, M.D.,[1] Yuh-Feng Wang, M.S., M.D.,[1] Yu-Chuan Li, M.D., Ph.D.,[2] and Min-Huei Hsu, M.S., M.D.[2]

[1]Department of Nuclear Medicine, Buddhist Dalin Tzu Chi General Hospital, Chiayi County, Taiwan

[2]Graduate Institute of Medical Informatics, Taipei Medical University, Taipei City, Taiwan

Address correspondence to:

Jainn-Shiun Chiu, MD

Department of Nuclear Medicine, Buddhist Dalin Tzu Chi General Hospital

No.2, Minsheng Rd., Dalin Township, Chiayi County 622, Taiwan

Telephone No.: 886-5-2648000 ext. 5712

Fax No.: 886-5-2648508

E-mail address: shiunkle@mail2000.com.tw

TO THE EDITOR:

Artificial neural network (ANN), a computational model composed of nonlinear processing elements arranged in highly interconnected layers with a configuration that simulates a biological nervous system, has been widely used as a predictive model in medicine with the help of advances in computer-assisted analysis. Therefore, the quality of the chosen ANN model is increasingly concerned. To evaluate the quality for classification models in clinical investigation, it would be more appropriate to calculate *discrimination* and *calibration* concurrently.[1] Discrimination is a measure of how well a model to distinguish subjects correctly as two different classes; calibration, goodness-of-fit, evaluates the degree of correspondence between the estimated probabilities produced by a model and the actual observation.

Common measures used in discrimination for prediction include sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), likelihood ratios for positive and negative tests, and area under the receiver-operating characteristic curve (AUROC). Porter CR et al[2] constructed the ANN and logistic regression models to predict prostate biopsy outcome from various clinical data sets. The authors externally validated their models by demonstrating AUROC, specificity, PPV, and NPV on the basis of fixed sensitivity at 0.90, which could not provide better index for the performance of each model. In fact, many researchers used AUROC with the best simultaneous sensitivity and specificity to determine discriminatory power of a model. The sensitivity and specificity at a cut-off value corresponding to the highest accuracy (i.e., minimal false negative and false positive results) should be computed and compared. Also from the perspective of statistics, the authors did not calculate the differences among AUROCs of their models by using statistical method such as pairwise comparison.[3]

On the other hand, even though the AUROC with the best simultaneous sensitivity and specificity was used, a good discrimination has the possibility of poor calibration when classification outputs are transformed monotonically.[4] To avoid this pitfall, calibration using Hosmer-Lemeshow statistic, Pearson *Chi*-square, or misclassification rate should be considered. Additionally, inter-rater agreement with kappa value among models could be adopted to approach the reproducibility and repeatability.[5] In the era of evidence-based medicine, new predictive model should be carefully and critically appraised since arbitrary assessments may lead to wrong conclusions.

REFERENCES

1. Li YC, Liu L, Chiu WT, et al: Neural network modeling for surgical decisions on traumatic brain injury patients. Int J Med Inform **57**: 1-9, 2000.
2. Porter CR, Gamito EJ, Crawford ED, et al: Model to predict prostate biopsy outcome in large screening population with independent validation in referral setting. Urology **65**: 937-941, 2005.
3. McNeil BJ, Hanley JA: Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 4: 137-150, 1984.
4. Dreiseitl S, Ohno-Machado L: Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform **35**: 352-359, 2002.
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics **33**: 159-174, 1977.