

Stata Modules for Calculating Novel Predictive Performance Indices for Logistic Models

Mahnaz Barkhordari,¹ Mojgan Padyab,² Farzad Hadaeigh,³ Fereidoun Azizi,⁴ and Mohammadreza Bozorgmanesh^{3,*}

¹Department of Mathematics, Bandar Abbas Branch, Islamic Azad University, Bandar Abbas, IR Iran

²Centre for Population Studies, Ageing and Living Conditions, Umea University, Umea, Sweden

³Prevention of Metabolic Disorders Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran

⁴Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran

*Corresponding author: Mohammadreza Bozorgmanesh, Prevention of Metabolic Disorders Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, P. O. Box 19395-4763, Tehran, IR Iran. Tel: +98-2122409301-5, Fax: +98-2122402463, E-mail: mr_bozorgmanesh@endocrine.ac.ir

Received 2015 August 5; Revised 2015 November 27; Accepted 2016 January 2.

Abstract

Background: Prediction is a fundamental part of prevention of cardiovascular diseases (CVD). The development of prediction algorithms based on the multivariate regression models loomed several decades ago. Parallel with predictive models development, biomarker researches emerged in an impressively great scale. The key question is how best to assess and quantify the improvement in risk prediction offered by new biomarkers or more basically how to assess the performance of a risk prediction model. Discrimination, calibration, and added predictive value have been recently suggested to be used while comparing the predictive performances of the predictive models' with and without novel biomarkers.

Objectives: Lack of user-friendly statistical software has restricted implementation of novel model assessment methods while examining novel biomarkers. We intended, thus, to develop a user-friendly software that could be used by researchers with few programming skills.

Materials and Methods: We have written a Stata command that is intended to help researchers obtain cut point-free and cut point-based net reclassification improvement index and (NRI) and relative and absolute Integrated discriminatory improvement index (IDI) for logistic-based regression analyses. We applied the commands to a real data on women participating the Tehran lipid and glucose study (TLGS) to examine if information of a family history of premature CVD, waist circumference, and fasting plasma glucose can improve predictive performance of the Framingham's "general CVD risk" algorithm.

Results: The command is added for logistic regression models.

Conclusions: The Stata package provided herein can encourage the use of novel methods in examining predictive capacity of ever-emerging plethora of novel biomarkers.

Keywords: Added Predictive Ability, Calibration, Integrated Discrimination Improvement, Net Reclassification Improvement, Software, Stata

1. Background

Currently risk prediction is an appealing research area (1). In the last two decades, there has been an increasing trend in the discovery of new biomarkers in clinical medicine (2). It is by use of predictive models that people can use their risk factors profile for a certain medical condition to calculate their corresponding risk of developing that event in the future (1). In the view of the current shoot-up in the discovery and emergence of new risk markers, statisticians as well as clinicians will need to tackle the challenge of the assessment of predictive capacities of these biomarkers. Clinically speaking, many predictive models provide risk values (probability of developing a medical condition in the future) that fall into high, low, or intermediate range. While making medical

decision on high-risk and low-risk individuals is somehow straight forward, dealing with intermediate range subclass will be cumbersome (1). As such, enhancement to the extant models has been sought to reclassify individuals more efficiently. Strong biomarkers have been added to relevant models in order to improve their predictive power (3).

Having been frequently making statements that the predictive performance of a model is superior to another, researchers are frequently challenged by statistical reviewers of scientific journals to provide rigorous statistical justification for their statements (4). How best to quantify the improvement in risk prediction offered by these new models? The answer to this question

would play a pivotal role in adopting or rejecting a new risk marker into clinical decision making algorithms (5). Merely demonstrating a statistically significant association of a new biomarker with certain medical condition is not enough (6-9). The performance of prediction models can be assessed using a variety of methods and metrics. Traditionally, a model predictive performance has been assessed from two perspectives, first, discriminatory predictive power and second, calibration. Discriminatory predictive power of a logistic regression model is usually assessed by calculating the area under the receiver operating characteristic curve. The calibration of logistic models is usually tested by calculating Hosmer-Lemeshow χ^2 .

Several new measures have recently been proposed. Among which the most commonly adopted and employed are reclassification tables, net reclassification improvement (NRI), and integrated discrimination improvement (IDI) for binary outcome (2, 5, 7-11).

Commonly-used, user-friendly statistical packages (e.g. SPSS) are yet to provide calculations for novel predictive performances statistics. Many studies, thus, do not make any notice of the novel statistics. Open-source statistical packages can be used to calculate novel statistics. They need, however, some knowledge of programming. This has rendered their usage limited.

The STATA could be counted among the softwares that provide users with an environment where new statistical analysis, which has not incorporated into the original software, can be performed by users who have some knowledge of programming. Furthermore, user-developed modules could be incorporated into the original software and utilized by researchers with no programming skills, provided that the module has been developed in the standard format of the STATA.

2. Objectives

We wrote, therefore, “adperd” command, in the STATA standard format, for calculating NRI and IDI for logistic regression models. The module is easy to run and can be used by researchers with few knowledge of programming. The help file has also been written in the standard STATA format and can be found in the same as was other statistical modules. Examples, using real data, have been incorporated in the program and can be found in the standard STATA format.

3. Materials and Methods

3.1. Added Predictive Ability for Logistic Models

3.1.1. Integrated Discrimination Improvement

IDI has been defined by Pencina et al. (2) as:

$$(1) \quad IDI = (IS_{new} - IS_{old}) - (IP_{new} - IP_{old})$$

Where IS is integral sensitivity over all cutoff values and IP is integral “one minus specificity,” “new” refers to the prediction model incorporating the new biomarker and “old” refers to the prediction model that does not. The arguments in the first presentation assumed nested models. However, IDI is applicable to situation where aim is to compare any two models with possibly different predictors and different analytic techniques, as far as predicted probabilities are calibrated to the same incidence or prevalence rate (12). Pencina et al. (2) provided the following estimator for IDI:

$$(2) \quad \text{Absolute IDI} = \left(\bar{P}_{new,events} - \bar{P}_{old,events} \right) - \left(\bar{P}_{new,nonevent} - \bar{P}_{old,nonevent} \right)$$

\bar{p} is the average estimated probability of an event. An average is taken for people in sample who experienced an event (“event”) and for those who did not (“nonevents”). Rearranging terms we obtain:

$$(3) \quad \text{Absolute IDI} = \left(\bar{P}_{new,events} - \bar{P}_{new,nonevent} \right) - \left(\bar{P}_{old,event} - \bar{P}_{old,nonevent} \right)$$

Which is the difference in discrimination slope proposed by Caputo et al. (13). The magnitude of discrimination slopes and their difference tends to be small (14, 15). This could be more conspicuous when the incidence or prevalence of the event of interest is relatively low (16). Considering the definition mentioned above, one could define relative IDI as the increase in discrimination slopes divided by the slope of the old model. As such, relative IDI could be estimated as follows:

$$(4) \quad \text{Relative IDI} = \frac{\left(\bar{P}_{new,events} - \bar{P}_{new,nonevent} \right) - \left(\bar{P}_{old,event} - \bar{P}_{old,nonevent} \right)}{\left(\bar{P}_{old,event} - \bar{P}_{old,nonevent} \right)}$$

Hereafter, we refer to IDI as absolute IDI.

3.1.2. Net Reclassification Improvement

To obtain NRI, predicted probability based on the basic (old or without new biomarker or new risk factor) and enhanced (new or with new biomarker or new risk factor) are classified into three categories (17); these two cross-classification are then cross-tabulated. The reclassification of people who develop and those who do not develop events is to be considered separately. Any ‘upwards’ movement across classes for those with the event (i.e. event group) implies improvement; whereas, any ‘downwards’ movement across classes indicates worse reclassification. The interpretation is opposite for those who do not develop event (the event nonevent group). The NRI will be a sum differences in proportion of individuals moved up minus proportion of those who moved down

among event subjects, and the proportion of individuals moved down minus proportion of those who moved up among nonevent subjects. The NRI as such quantifies the improvement in reclassification. If assign 1 for each upwards movement and -1 for each downwards movement, and 0 for no movement in categories, the NRI can be estimated as:

$$(5) \text{ Cutpoint – based NRI} = \frac{\sum_{i \text{ in events}} v(i)}{\text{events}} - \frac{\sum_{i \text{ in nonevents}} v(i)}{\text{non events}}$$

where $v(i)$ is the above-defined movement indicator.

In general, it is not recommended to use more than three categories unless they are already established and there are rigorous justifications to do so. For situation where finer portioning is needed, Pencina et al. (17) have suggested the cutpoint-free NRI. The definition of cutpoint-free NRI remains consistent with cutpoint-based formula defined above with the only difference in the meaning of the upwards and downwards reclassification (17).

3.2. Commands

3.2.1. “Adpred” Command

3.2.1.1. Syntax

Adpred depvar oldrisk newrisk, cutpoint (numlist).

Depvar represents dependent binary outcome variable.

Oldrisk is the variable that represents the risk calculated based on the baseline model.

Newrisk is the variable that represents the risk calculated based on the enhanced model.

Numlist is the list of risk cutoff points for cutpoint-based NRI.

3.2.1.2. Description

“adpred” calculates absolute and relative IDI, as well as cutpoint-based and cutpoint-free NRI. For cutpoint-based NRI to be calculated cutoff-points of risk should be specified by users.

3.2.1.3. Option

“Cut (numlist)” gives the numbers that present cut-points of risk based on the old model at which new model is to be evaluated.

4. Results

4.1. Example

4.1.1. Study Population

We used a real data set of the Tehran lipid and glucose study (TLGS) to predict incident cardiovascular diseases

(CVD). Detailed descriptions of the TLGS have been reported elsewhere (18, 19). We used data on 4 052 women with complete data on covariates, contributing to a 42 659 person-year follow-up. The median follow-up time was 11.5 years, at the time of the current study.

4.1.2. Statistical Analysis

In the analysis of outcomes (CVD), CVD we used the logistic regression model. The baseline model was developed based on the age, smoking, systolic blood pressure, use of anti-hypertensive drugs, total and HDL cholesterol, smoking, and diabetes. Improved model was developed by adding family history of premature CVD to the baseline model to basic logistic model and a family history of premature CVD, waist circumference, and fasting plasma glucose to the basic logistic model.

We set the statistical significance level at a two-tailed type I error of 0.05 and used Stata version 12.0 (Stata Corp, College Station, Texas USA) for all statistical analyses.

4.1.3. Assessment of Model Performance

We generally use several criteria to compare the overall predictive values of alternative models. However, for the current paper to be succinct and more focused on the STATA module we have curtailed other measures and herein are presenting added predictive capacity.

4.1.4. Added Predictive Capacity

Absolute and relative IDI and cut-point-based and cutpoint-free NRI were used as measures of predictive ability added to the baseline model by paraclinical parameters (2). Bootstrapping method was implemented in order to obtain 95% confidence intervals (95% CIs).

Improvement in the discriminative capacity of prediction models can be quantified in numerous ways. A natural approach takes the difference in discrimination metrics between the models with and without the new predictor. The Δ AUC (difference in the area under the receiver operating characteristics curves of the prediction model with and without new marker) is produced in this manner and so is the IDI, defined as a difference in discrimination slopes. The relative IDI can be calculated as the ratio of IDI over the discrimination slope of the model without the new predictor (generally referred to as the baseline model). Integrated discrimination improvement can be seen as continuous version of NRI with probability differences used instead of categories. Alternatively, it can be defined as a difference in discrimination slopes. Discrimination slope in the binary context is defined as difference of mean predicted probabilities of events and nonevent. The cutpoint-free NRI, is obtained when a study is focused on the relative increase in the predicted probabilities for individuals who have experienced events and the decrease for individuals who do not.

Table 1. Predictive Performances of the Basic Framingham's "General Cardiovascular Diseases Risk" Algorithm vs. Enhanced Model

Added Predictive Values	Basic Model vs. Enhanced Model ^a	P Value
Absolute IDI	0.0060 (-0.0001 - 0.0121)	0.055
Relative IDI	0.0480 (-0.0032 - 0.0991)	0.066
Cutpoint-Based NRI	0.0048 (-0.0383 - 0.0479)	0.827
Cutpoint-Free NRI	0.2732 (0.1561 - 0.3903)	0.000

IDI, integrated discriminatory improvement index; NRI, net reclassification improvement index.

^aValues are expressed as statistics (95% CIs).

Tables 1 presents the novel analysis obtained from Stata (please see also appendix for further details). As is evident the table, adding the family history of premature CVD to the baseline model does not improve the prediction capacity of the baseline model in terms of IDI or NRI. Herein, Pvalues are pertinent to testing the null hypothesis that the magnitude of the increase in the predictive capacity of the baseline model conferred by a novel marker added to the baseline model is zero. All P-values are greater than 0.05 that is to say the improvement observed in the predictive capacity of the enhanced model based on the IDI and NRI is not statistically meaningful.

The Framingham's "general CVD risk" algorithm incorporated age, systolic blood pressure, using blood pressure lowering drugs, total and high-density lipoprotein cholesterol, smoking, and diabetes. The enhanced model was developed by adding a family history of premature CVD, waist circumference, and fasting plasma glucose to the basic Framingham's "general CVD risk" algorithm components.

5. Discussion

When examining the clinical relevance of a new risk biomarker or when examining if the predictive power of a currently available predictive model can be augmented by new biomarker (s), NRI and IDI can be very informative. We hope by using the packages provided herein novel analysis will be more extensively utilized in studies aimed at examining the predictive ability of prediction models or clinical usefulness of new biomarkers.

Acknowledgments

We thank Michael J. Pencina, PhD, department of mathematics and statistics, Boston university, Boston, MA, for statistical help and the constructive comments. We express our appreciation to the participants of district-13 of Tehran for their enthusiastic support in this study.

Footnotes

Authors' Contribution: Mohammadreza Bozorgmanesh programmed the codes, performed and interpreted the statistical analyses, and helped drafting the paper. Mahnaz Barkhordari helped programming codes and drafted the manuscript. Mojgan Padyab helped pro-

gramming codes. Farzad Hadaegh and Feridoun Azizi helped drafting the paper.

Funding/Support: This study was supported by grant No. 121 from the national research council of the Islamic Republic of Iran.

References

1. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. *Am J Epidemiol*. 2011;**173**(11):1327-35. doi: 10.1093/aje/kwr013. [PubMed: 21555714]
2. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, 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):157-72. doi: 10.1002/sim.2929. [PubMed: 17569110]
3. Yadegari H, Bozorgmanesh M, Hadaegh F, Azizi F. Non-linear contribution of glucose measures to cardiovascular diseases and mortality: reclassifying the Framingham's risk categories: a decade follow-up from the Tehran lipid and glucose study. *Int J Cardiol*. 2013;**167**(4):1486-94. doi: 10.1016/j.ijcard.2012.04.053. [PubMed: 22578948]
4. Wilson PW, D'Agostino R. Sr., Bhatt DL, Eagle K, Pencina MJ, Smith SC, et al. An international model to predict recurrent cardiovascular disease. *Am J Med*. 2012;**125**(7):695-703 e1. doi: 10.1016/j.amjmed.2012.01.014. [PubMed: 22727237]
5. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. Germany: Springer; 2009.
6. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;**115**(7):928-35. doi: 10.1161/CIRCULATIONAHA.106.672402. [PubMed: 17309939]
7. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem*. 2008;**54**(1):17-23. doi: 10.1373/clinchem.2007.096529. [PubMed: 18024533]
8. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med*. 2006;**145**(1):21-9. [PubMed: 16818925]
9. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. *Biom J*. 2011;**53**(2):237-58. doi: 10.1002/bimj.201000078. [PubMed: 21294152]
10. Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. *J Am Coll Cardiol*. 2009;**54**(14):1209-27. doi: 10.1016/j.jacc.2009.07.020. [PubMed: 19778661]
11. Pencina MJ, D'Agostino RB, Vasan RS. Statistical methods for assessment of added usefulness of new biomarkers. *Clin Chem Lab Med*. 2010;**48**(12):1703-11. doi: 10.1515/CCLM.2010.340. [PubMed: 20716010]
12. Polak JF, Meisner A, Pencina MJ, Wolf PA, D'Agostino RB. Variations in common carotid artery intima-media thickness during the cardiac cycle: implications for cardiovascular risk assessment. *J Am Soc Echocardiogr*. 2012;**25**(9):1023-8. doi: 10.1016/j.echo.2012.05.007. [PubMed: 22721828]
13. Caputo RP, Goel A, Pencina M, Cohen DJ, Kleiman NS, Yen CH, et al. Impact of drug eluting stent length on outcomes of percutaneous coronary intervention (from the EVENT registry). *Am J Cardiol*. 2012;**110**(3):350-5. doi: 10.1016/j.amjcard.2012.03.031. [PubMed: 22560770]

14. Pencina MJ, D'Agostino RB. Sr., Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. *Stat Med.* 2012;**31**(15):1543-53. doi: 10.1002/sim.4508. [PubMed: 22344892]
15. Novack V, Pencina M, Cohen DJ, Kleiman NS, Yen CH, Saucedo JF, et al. Troponin criteria for myocardial infarction after percutaneous coronary intervention. *Arch Intern Med.* 2012;**172**(6):502-8. doi:10.1001/archinternmed.2011.2275. [PubMed: 22371874]
16. Widera C, Pencina MJ, Meisner A, Kempf T, Bethmann K, Marquardt I, et al. Adjustment of the GRACE score by growth differentiation factor 15 enables a more accurate appreciation of risk in non-ST-elevation acute coronary syndrome. *Eur Heart J.* 2012;**33**(9):1095-104. doi: 10.1093/eurheartj/ehr444. [PubMed: 22199121]
17. Pencina MJ, D'Agostino RB. Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med.* 2011;**30**(1):11-21. doi: 10.1002/sim.4085. [PubMed: 21204120]
18. Azizi F, Ghanbarian A, Momenan AA, Hadaegh F, Mirmiran P, Hedayati M, et al. Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase II. *Trials.* 2009;**10**:5. doi: 10.1186/1745-6215-10-5. [PubMed: 19166627]
19. Bozorgmanesh M, Hadaegh F, Azizi F. Predictive accuracy of the 'Framingham's general CVD algorithm' in a Middle Eastern population: Tehran Lipid and Glucose Study. *Int J Clin Pract.* 2011;**65**(3):264-73. doi: 10.1111/j.1742-1241.2010.02529.x. [PubMed: 21314863]