Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Decoding Predictive Power: A Simulation Study on Information Criteria vs. Internal Performance Measures

H. van de Beek*[1,2]

[1]Methods and Statistics, Utrecht University, the Netherlands

[2]Julius Center for Health Sciences and Primary Care, UMC Utrecht, the Netherlands

Correspondence
*Hidde van de Beek, Utrecht University.
Email: h.vandebeek@uu.nl

Present address
Utrecht University, Utrecht, the Netherlands

-

KEYWORDS
Simulation study, Information criteria, Internal performance measures, Predictive power, Model selection

1 | INTRODUCTION

In public health, prediction models have the ability to target preventive interventions to persons at high risk of having or developing a disease (prognosis and diagnosis). In clinical practice, prediction models may inform patients and their doctors on the probability of a diagnosis or prognostic outcome[1]. Identification of patients at high risk can be based on a combination of risk factors, risk indicators or other predictors (e.g. a particular patient characteristic, bio-marker or test result). The performance of these prediction models impacts public health and clinical practice, thus making correct modelling choices crucial[2].

The traditional approach to medical prediction models uses logistic regression[3]. For model selection, two methodologies are often employed. Information Criteria (IC) -or model selection methods- estimate the information loss when the probability distribution of the true model is approximated by the probability distribution of a candidate model. By minimizing this discrepancy (Kullback-Leibler divergence[4]) between these distributions, the goal is to select the model that represents the data generating mechanism. The second methodology is the procedure of using internal model performance measures for estimating the out-of-sample performance of the proposed candidate models. Model performance includes discriminatory ability, calibration and overall accuracy. Commonly used techniques are Area under the Receiver Operated Curve (AUROC) and R^2, based on which the best performing proposed model can be selected.

Information criteria, like the Akaike's information criterion (AIC[5]) and the Bayesian information criterion (BIC[6]), are both commonly used for model selection in health studies. The AIC and BIC share the same goodness-of-fit term, but the penalty terms differ based on the manner in which the dimension k (number of parameters) and n (sample size) is incorporated: BIC employs a complexity penalization of $k \log n$ as opposed to $2k$ of the AIC. Consequently, BIC tends to choose fitted models that are more parsimonious than those favored by AIC[7]. Conversely, the AIC favours more complex models in large sample setting. Information criteria only provide information about the relative quality between models using the same likelihood function and data set[8].

Several internal performance measures are available for an estimate of out-of-sample model performance. The AUROC analysis is developed for predictive model selection[9] and is widely adopted in clinical science to assess the model sensitivity and specificity trade-off[10]. Bootstrapping is a preferred technique for assessing prediction models using performance measures such as AUROC[11] [12], since using the cases from the original analysis sample results in an overly optimistic performance estimate[13].

The AUROC is also criticized for being a semi-proper scoring rule[14], which means that the best performance can be attained by a misspecified model. Another performance measure of prediction models is the R^2. In ordinary least-squares regression it is interpretable as the proportion of outcome variation, which can be explained by the predictors. Pseudo R^2 indices have been developed for logistic regression, such as methods of Cox and Snell[15], Nagelkerke[16], and McFadden[17]. Import to note is that R^2 measures increase monotonically with increasing number of covariates even if they have no prognostic value at all[18].

Concluding, ICs and internal performance measures intend to approximate the data generating model and have the best out-of-sample performance, aiming for the highest predictive power in the total population. We are interested in their ability to choose the correct model when the data generating mechanism is known. Testing the IC success rate and performance measures (AUROC and R^2) for choosing the correct model in different simulation contexts, such as sample size and data quality, may yield information for future prediction modelling choices. The according research question will be: How successful are Information Criteria and internal performance measures in choosing the correct prediction model in different simulated contexts? We expect a difference in these factors and a specific use and integration of the techniques in real life application.

2 | ANALYTICAL PLAN

The current project will investigate the research question by simulating a range of data sets that differ in context/population (i.e. disease occurence and discriminatory ability). From these populations different sample sizes will be used to see the effect on success rates of selecting the correct model. For these parameters optimal values will be calculated. First these success rates will be evaluated using IC. In the second part internal performance measures will be evaluated for their contextual performance. The population simulations and analysis will be done using High Performance Computers at the University Medical Center Utrecht. We will use Rstudio for version controlled **R** files of data generation, visualization and analyses[19].

3 | METHODS

In our research, we focused on prediction models designed for dichotomous risk prediction. We used a simulation study to investigate the effects of imbalance correction methods across 18 unique data-generating scenarios (section 2.2).

For each scenario, we compared the out-of-sample predictive performance (i.e., model performance on data not used to train the model) of prediction models developed using a two-step procedure (section 2.3). This procedure consisted of an imbalance correction step in which the data were pre-processed, and a model training step in which the pre-processed data were used to train a machine learning model.

All code used to implement the simulation study, and process the results is made publicly available (section 2.6). Ethical approval for this research was granted by the Ethical Review Board of the Faculty of Social and Behavioural Sciences at Utrecht University and is filed under number 23-1780.

3.1 | Data generation

References

1. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and Prognostic Research: Application and Impact of Prognostic Models in Clinical Practice. *BMJ.* 2009;338:b606. doi: 10.1136/bmj.b606

2. Wessler BS, Paulus J, Lundquist CM, et al. Tufts PACE Clinical Predictive Model Registry: Update 1990 through 2015. *Diagnostic and Prognostic Research.* 2017;1(1):20. doi: 10.1186/s41512-017-0021-2

3. Steyerberg E. Applications of Prediction Models. In: Steyerberg E., ed. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, , Statistics for Biology and Health. New York, NY: Springer, 2009:11–31

4. Kullback S, Leibler RA. On Information and Sufficiency. *The Annals of Mathematical Statistics.* 1951;22(1):79–86.

5. Akaike H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control.* 1974;19(6):716–723. doi: 10.1109/TAC.1974.1100705

6. Schwarz G. Estimating the Dimension of a Model. *The Annals of Statistics.* 1978;6(2):461–464.

7. Neath AA, Cavanaugh JE. The Bayesian Information Criterion: Background, Derivation, and Applications. *WIREs Computational Statistics.* 2012;4(2):199–203. doi: 10.1002/wics.199

8. Chowdhury MZI, Turin TC. Variable Selection Strategies and Its Importance in Clinical Prediction Modelling. *Family Medicine and Community Health.* 2020;8(1):e000262. doi: 10.1136/fmch-2019-000262

9. Pepe MS. An Interpretation for the ROC Curve and Inference Using GLM Procedures. *Biometrics.* 2000;56(2):352–359. doi: 10.1111/j.0006-341X.2000.00352.x

10. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD). *Circulation.* 2015;131(2):211–219. doi: 10.1161/CIRCULATIONAHA.114.014508

11. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in StatisticsCham: Springer International Publishing, 2015

12. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM. Internal and External Validation of Predictive Models: A Simulation Study of Bias and Precision in Small Samples. *Journal of Clinical Epidemiology.* 2003;56(5):441–447. doi: 10.1016/S0895-4356(03)00047-7

13. LeDell E, Petersen M, van der Laan M. Computationally Efficient Confidence Intervals for Cross-Validated Area under the ROC Curve Estimates. *Electronic journal of statistics.* 2015;9(1):1583–1607. doi: 10.1214/15-EJS1035

14. Zhou QM, Zhe L, Brooke RJ, Hudson MM, Yuan Y. A Relationship between the Incremental Values of Area under the ROC Curve and of Area under the Precision-Recall Curve. *Diagnostic and Prognostic Research.* 2021;5(1):13. doi: 10.1186/s41512-021-00102-w

15. Cox DR, Snell EJ. *Analysis of Binary Data, Second Edition*. CRC Press, 1989.

16. Nagelkerke NJ. A Note on a General Definition of the Coefficient of Determination. *biometrika.* 1991;78(3):691–692.

17. McFadden D. Conditional Logit Analysis of Qualitative Choice Behavior. 1973.

18. Mittlböck M, Heinzl H. A Note on R2 Measures for Poisson and Logistic Regression Models When Both Models Are Applicable. *Journal of Clinical Epidemiology.* 2001;54(1):99–103. doi: 10.1016/S0895-4356(00)00292-4

19. R Core Team . R: A Language and Environment for Statistical Computing. tech. rep., R Foundation for Statistical Computing; Vienna, Austria: 2023.