

RESEARCH ARTICLE

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

H. van de Beek^{*1,2}

¹Methods and Statistics, Utrecht University, the Netherlands

²Julius 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, the Netherlands

This is a generic template designed for use by multiple journals, which includes several options for customization. Please refer the author guidelines and author LaTeX manuscript preparation document for the journal to which you are submitting in order to confirm that your manuscript will comply with the journal's requirements. Please replace this text with your abstract. This is sample abstract text just for the template display purpose.

KEY WORDS

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 outcome¹. 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 crucial².

2 | BACKGROUND

The traditional approach to medical prediction models uses logistic regression³. 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 divergence⁴) 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 (AIC⁵) and the Bayesian information criterion (BIC⁶), 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 AIC⁷. 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 set⁸.

Several internal performance measures are available for an estimate of out-of-sample model performance. The AUROC analysis is developed for predictive model selection⁹ and is widely adopted in clinical science to assess the model sensitivity and specificity trade-off¹⁰. 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 estimate¹³. The AUROC is also criticized for being a semi-proper scoring rule¹⁴, 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 Snell¹⁵, Nagelkerke¹⁶, and McFadden¹⁷. Import to note is that R^2 measures increase monotonically with increasing number of covariates even if they have no prognostic value at all¹⁸.

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.

3 | METHODS

In this study, we focus on prediction models designed for dichotomous risk prediction, such as disease occurrence. To investigate the success rate of IC and internal performance measures, we simulated a population dataset and sampled from this dataset. The simulations were performed in R-studio¹⁹.

3.1 | Data generating mechanism

We simulated data for a setting in which there were 3 covariates (X_1 – X_3). These covariates were simulated from independent standard normal distributions ($\mu = 0, \sigma = 1$).

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{bmatrix}$$

All three affected the outcome. The outcome was generated from a logistic regression model with the following formula:

$$P(Y = 1|X_1, X_2, X_3) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}} \quad (1)$$

The coefficients were chosen to be $\beta_0 = 0, \beta_1 = 0.8, \beta_2 = 0.8, \beta_3 = 0.4$. The outcome was generated from a Bernoulli distribution with probability $P(Y = 1|X_1, X_2, X_3)$ with a cutoff value of 0.5. This results in a true C-statistic or AUROC of 0.80. A population of 2.000.000 respondents was created accordingly.

3.2 | Analysis

Success rate was analyzed by comparing the true model with the model chosen by the IC and internal performance measures. The true model was defined as the model that generated the data. The ICs and internal performance measures were calculated for all possible models. The success rate was calculated as the proportion of times the true model was chosen. The model with the lowest IC (AIC and BIC) and highest internal performance measure was chosen. AUROC is used as the internal performance

measure. The AUROC was calculated using the R package pROC²⁰. The difference between these performances is an estimate of the optimism in the apparent performance. This difference is averaged to obtain a stable estimate of the optimism. The optimism is subtracted from the apparent performance to estimate the internally validated performance: estimated performance = apparent performance – average(bootstrap performance – test performance)²¹. A bootstrap of 80 was performed on each iteration. 2500 iterations were performed.

4 | RESULTS

For all sample sizes, the success rate was calculated. This is shown in plot 1

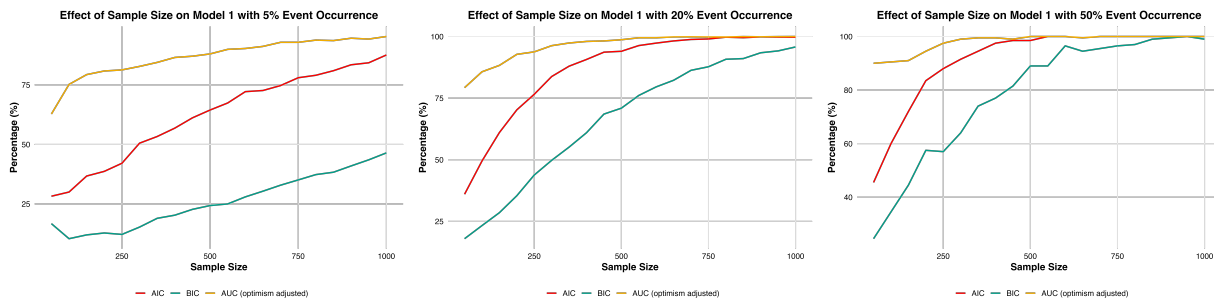


FIGURE 1 Success Rates

5 | DISCUSSION

This is a generic template designed for use by multiple journals

6 | CONCLUSIONS

This is a generic template designed for use by multiple journals

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 Statistics Cham: 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.
20. Robin X, Turck N, Hainard A, et al. pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinformatics*. 2011;12(1):77. doi: 10.1186/1471-2105-12-77
21. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal Validation of Predictive Models: Efficiency of Some Procedures for Logistic Regression Analysis. *Journal of Clinical Epidemiology*. 2001;54(8):774–781. doi: 10.1016/S0895-4356(01)00341-9