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

DOI: xxx/xxxx

RESEARCH ARTICLE

Decoding Predictive Performance: 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, 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.

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].

2 | BACKGROUND

The traditional approach to medical prediction models often 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.

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. We will follow a mini-thesis structure that only look at one simple data generating mechanism. The results of this study will be used to inform the design of the thesis.

3 | METHODS

In this study, we focus on prediction models designed for dichotomous risk prediction, such as disease occurence. To investigate the succes rate of IC and internal performance measures, we simulated a population dataset and sampled from this dataset. The simulations, analyses and visualizations were performed in R-studio[15] and are available on Github .

3.1 | Data generating mechanism

We generated two population datasets, each with a different data generating mechanism. We simulated 10,000,000 individuals per model, each characterized by two or three covariates (X1, X2, and X3). The covariatesof the two models were independently derived from the same standard normal distributions. The mean vector (μ) and the covariance matrix (Σ) for these distributions were defined as follows:

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

The probability of each individual belonging to the outcome group was determined based on their covariate values. This was done using a logistic regression model, which defined the probability of the outcome $Y = 1$ as a function of the covariates:

$$\text{Model } 1 : P(Y = 1 | X_1, X_2) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}} \tag{1}$$

$$\text{Model } 2 : P(Y = 1 | X_1, X_2, X_3) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}} \tag{2}$$

In model 1, the coefficients were set as $\beta_0 = 1.65, \beta_1 = 0.8, \beta_2 = 0.8$. In model 2, the coefficients were set as $\beta_0 = 2, \beta_1 = 0.8, \beta_2 = 0.8, \beta_3 = 0.4$. The outcome for each individual was then generated following a Bernoulli distribution.

3.2 | Analysis

In this study, we analyzed the success rate by comparing the true model to a competing model. The true model was defined as the model that generated the data: $(Y = 1 | X_1, X_2)$ or $(Y = 1 | X_1, X_2, X3)$. The competing model is defined by the opposing model: $(Y = 1|X_1, X_2, X3)$ or $(Y = 1|X_1, X_2)$, respectively. The competing model is therefore misspecified and can thus be used to investigate the ability of the ICs and internal performance measures to choose the correct model. We are effectively analyzing the ability to correctly add or remove covariates from the model. The success rate was quantified as the proportion of instances where the true model was correctly chosen as the best performing model.

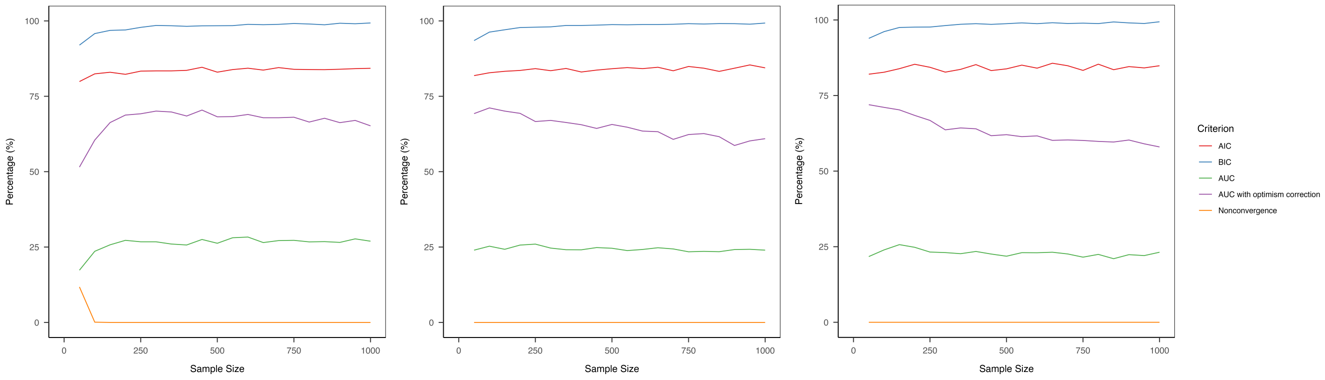
For our internal performance evaluation, we specifically used the AUC, calculated by fitting the model using `lrm` function of the rms package. To correct for the optimism in the apparent AUC in the sample, we bootstrap the sample and calculate the optimism-corrected AUC. The optimism-corrected AUC is calculated by subtracting the average bootstrapped optimism from the apparent AUC. The optimism is calculated by taking the difference between the AUC in the bootstrap sample and the AUC in the test sample[16]. This is done by the `validate` function in the rms package. For this purpose, a bootstrap with 80 resamples was performed in each iteration. The `lrm` function also provides the BIC and AIC of the fitted model.

This analysis procedure was repeated for three different event rates (5%, 20%, and 50%). Per event rate, we calculated the success rate for two different ICs and internal performance techniques for a sample size ranging from 50 to 1,000 with intervals of 50. Per sample size, 3,500 iterations were performed. The success rate was calculated for each combination of event rate and sample size. The nonconvergence of the model fit was also recorded. Visualizations are generated using the package ggplot2[17].

4 | RESULTS

4.1 | Model 1: $(Y = 1|X_1, X_2)$

The result of the analysis is shown in figure 1. It can be seen that BIC has the highest overall success rate in all three scenario's. After that AIC has the highest success rate, followed by the optimism corrected AUC. The AUC has the lowest success rate in all three scenario's. There is no trend for sample size, with a fairly constant success rate across all three scenario's.



(a) 5% disease occurence.

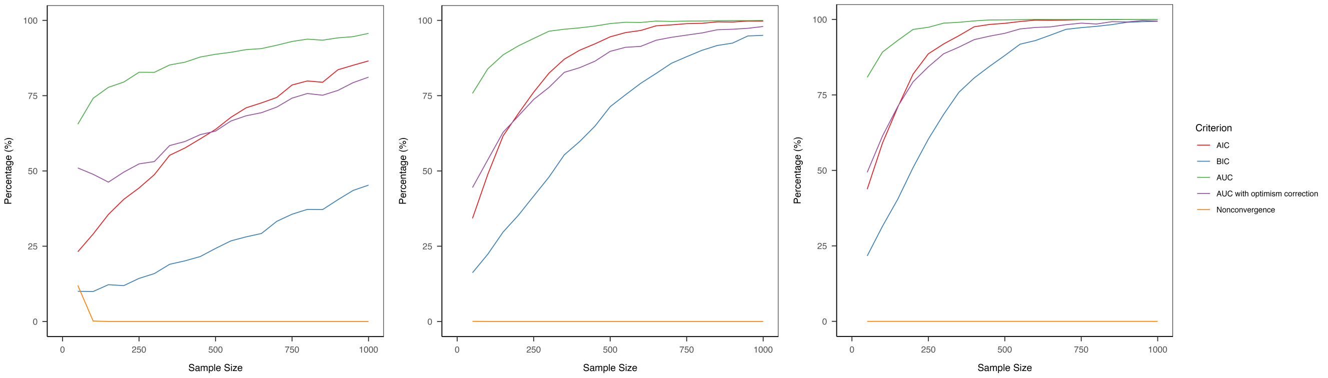
(b) 20% disease occurence.

(c) 50% disease occurence.

FIGURE 1 Success rate of ICs and internal performance measures in choosing the correct model.

4.2 | Model 2: $(Y = 1|X_1, X_2, X3)$

The result of the analysis is shown in figure 2. It can be seen that AUC has the highest overall success rate in all three scenario's. The performance of bootstrapped AUC and AIC is comparable across context. BIC has the lowest overall success rate in all three scenario's, performing worse in the low event rate scenario. There is clear trend for sample size, with a higher success rate for larger sample sizes: a convergence to the data generating model.



(a) 5% disease occurence.

(b) 20% disease occurence.

(c) 50% disease occurrence.

FIGURE 2 Success rate of ICs and internal performance measures in choosing the correct model.

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

16. 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

17. Wickham H. Getting Started with Ggplot2. In: Wickham H. , ed. *Ggplot2: Elegant Graphics for Data Analysis*, , Use R! Cham: Springer International Publishing, 2016:11–31