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

DOI: xxx/xxxx

RESEARCH REPORT

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

Word count: 2284

KEYWORDS

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

1 | INTRODUCTION

In public health, prediction models are used to target preventive interventions to persons at high risk of having or developing a disease. Based on these medical prediction models, important choices are made regarding treatment and prevention. Furthermore, in clinical practice prediction models may be used to inform patients and their doctors on the probability of a diagnosis or prognostic outcome. Therefore it is crucial that the models are accurate and reliable[1]. Modelling choices, such as the choice of predictors, the functional form of the predictors, and the choice of the model itself influence this accuracy and reliability. Making valid modelling choices can therefore benefit public health and clinical practice[2]. The traditional approach to medical prediction models often uses logistic regression[3]. Based on field specific background theory several candidate models can arise. For the final model selection, two methodologies can be employed.

Firstly, information criteria (IC) 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 data generating model has, by definition, the highest out-of-sample performance in the population. Examples are the Akaike's information criterion (AIC[5]) and the Bayesian information criterion (BIC[6]), which are both commonly used for model selection in health studies. Both the AIC and BIC minimize the Kullback-Leibler divergence by using the likelihood of the model as the goodness-of-fit term, but differ in their penalty term. The penalty term incorporates dimension k (number of parameters) and n (sample size): the BIC employs a complexity penalization of $k \log n$ as opposed to $2k$ employed by the AIC. Consequently, the BIC tends to choose fitted models that are more parsimonious than those favored by the AIC[7]. Conversely, the AIC favours more complex models in large sample setting. An important note is that IC only provide information about the relative quality between models that use the same likelihood function and data set[8].

Secondly, internal model performance measures are used to calculate the within sample performance. They can also be used to estimate the out-of-sample performance for the candidate models in the population. Performance can in this case be defined as the model's ability to correctly classify a person as healthy or diseased. Several internal performance measures are available for an estimate of the out-of-sample model performance. The Area Under the receiver operated Curve (AUC) analysis is developed

for predictive model selection[9] and is widely adopted in clinical science to assess the model's sensitivity and specificity trade-off[10]. Bootstrapping is a preferred technique for assessing prediction models using performance measures such as the AUC[11 12], since using the cases from the original analysis sample results in an overly optimistic performance estimate[13]. The AUC is also criticized for being a semi-proper scoring rule[14], meaning that the best performance can be attained by a misspecified model.

Concluding, IC intend to approximate the data generating model and internal validation techniques are developed to estimate the out-of-sample performance. In reality both aim to choose a model with the highest out-of-sample performance. However, it remains unclear how these methods compare in selecting the correct model in different contexts. So, we are interested in their ability to choose the correct model when the data generating mechanism is known. Testing the success rate of the IC and performance measures (AUC and bootstrapped AUC) 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 is: How successful are Information Criteria and internal performance measures in choosing the correct prediction model in different simulated contexts? A difference is expected for sample size, event rate and size of the coefficients.

We will follow a mini-thesis structure in this research report. We limit ourselves by only looking at a small scale simulation; i.e. without high performance computing. This setup and the according results will serve as a guideline for the final thesis. The structure of this report is as follows: first, we will describe the methods and data generating mechanism. Second, we will present the results of the simulation study. Third, we will discuss the results and finally, we will conclude with a summary and recommendations for the final thesis.

2 | METHODS

This study investigates medical prediction models designed for dichotomous risk prediction, such as disease occurrence. To investigate the success rate of IC and internal performance measures, we simulated two population data sets and sampled from these data sets. The simulations, analyses and visualizations were performed in R-studio[15] and are available on Github.

2.1 | Data generating mechanism

We generated two population data sets, each with a different data generating mechanism. We simulated 10,000,000 individuals per data set, characterized by three covariates (X1, X2, and X3). The covariates of the two models were independently derived using 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 $P(Y = 1)$ given the covariates $\beta_1, \beta_2, \beta_3$:

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

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

The coefficients (β) of the models are defined in Table 1. The outcome group was defined Y = 1 if the probability P of the outcome was greater than 0.5, and Y = 0 if the probability P of the outcome was less than 0.5.

TABLE 1 Coefficients of Model 1 and Model 2

Coefficient	Model 1	Model 2
β_0	1.65	2
β_1	0.8	0.8
β_2	0.8	0.8
β_3	0	0.4

2.2 | Analysis

In this study, we analyzed the success rate by comparing the true model to one single competing model, for the sake of simplicity and computing power. The true model was defined as the model that generated the data: firstly $(Y = 1|X_1, X_2)$ and then $(Y = 1|X_1, X_2, X3)$. 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 IC and internal performance measures to choose the correct model. This is effectively analyzing the ability to correctly add or leave out predictors to the model. The success rate was quantified as the proportion of instances where the true model was correctly chosen; the true model had the lowest IC or highest internal performance measure.

The `lrm` function of the rms package is used to get the AIC and BIC of the model[11]. For our internal performance evaluation, we specifically used the AUC, calculated by fitting the model using `lrm` function. To correct for the optimism in the apparent AUC in the sample, we bootstrapped the sample and calculated the optimism-corrected AUC. The optimism-corrected AUC was calculated by subtracting the average bootstrapped optimism from the apparent AUC. The optimism was calculated by taking the difference between the AUC in the bootstrap sample and the AUC in the test sample[16]. This was be done by the `validate` function in the rms package.

This analysis procedure was repeated for three different event rates (5%, 20%, and 50%). This was done by under sampling the full population to the according event rate. The sample size varied from 50 to 1,000 with intervals of 50. The combination of event rate and sample size form the contexts in the simulation. Per context 3,500 samples were drawn. For each data point the success rate of choosing the correct model was calculated based on the IC and internal performance measures. The optimism-corrected AUC used 80 bootstrap samples. The nonconvergence of the model fit was also recorded. In case of nonconvergence in one model, both models were considered nonconvergent. This means the percentage is only based on the models that did converge. Visualizations were generated using the package ggplot2[17].

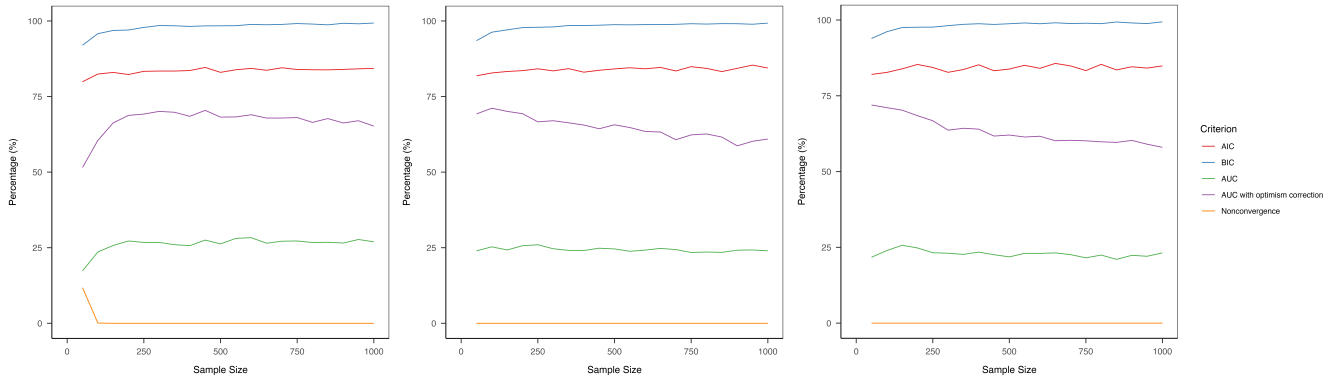
3 | RESULTS

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

The analysis results of model 1 are shown in Figure 1. The BIC has the highest overall success rate in all three contexts. The AIC has the second highest success rate, followed by the optimism corrected AUC. The AUC has the lowest success rate in all three contexts. There is no clear trend for sample size or event rate, with a fairly constant downward success rate across all three contexts. Nonconvergence is only a problem in the lowest sample size and event rate.

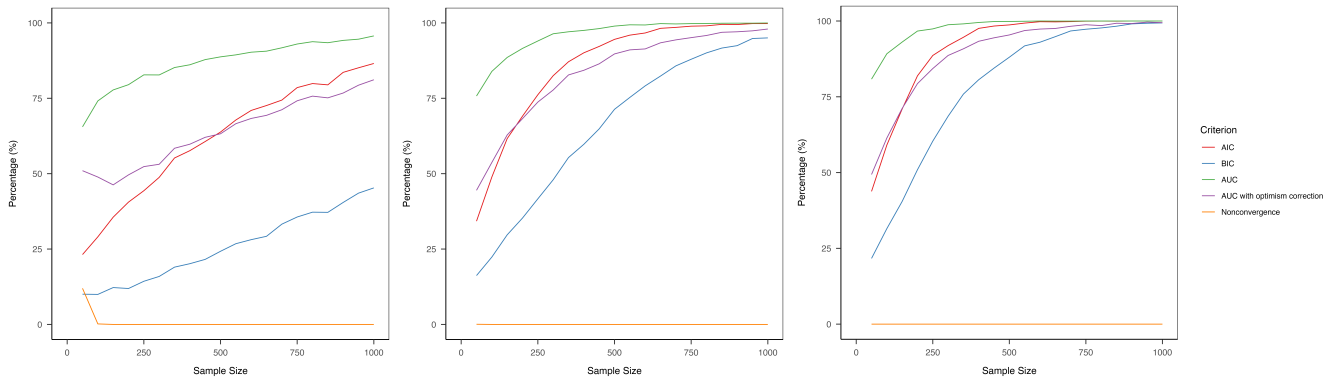
3.2 | Model 2: $(Y = 1|X_1, X_2, X3)$

The analysis results of model 2 are shown in Figure 2. The AUC has the highest overall success rate in all three contexts. The performance of bootstrapped AUC and AIC is comparable across context. The BIC has the lowest overall success rate in all three contexts, performing worse in the low event rate context. There is clear trend for sample size, with a higher success rate for larger sample sizes: a convergence towards the data generating model. Nonconvergence is only a problem in the lowest sample size and event rate.



(a) 5% event rate. (b) 20% event rate. (c) 50% event rate.

FIGURE 1 Success rate of IC and internal performance measures in choosing the correct model when $P(Y = 1|X_1, X_2)$ generates the data.



(a) 5% event rate. (b) 20% event rate. (c) 50% disease occurrence.

FIGURE 2 Success rate of IC and internal performance measures in choosing the correct model when $P(Y = 1|X_1, X_2, X_3)$ generates the data.

4 | DISCUSSION

In the field of medical prediction modelling there are several methods for choosing the final model. Both IC and internal validation techniques play an important role in this process. However, it is unclear how these methods compare to each other in selecting the correct model in different contexts. In this mini thesis we used a simple data generating mechanism to simulate data and adjusted sample size event occurrence in order to evaluate how successful IC and internal performance measures are in choosing the correct prediction model in different simulated contexts. The experiments performed in this study provide insight into an aspect that has been missing in the literature up to now.

Reflecting on the experiments, we start with analyzing the general patterns: the first experiment, only using two out of three covariates, shows no convergence towards the data generating mechanism. Specifically, the experiment shows a constant and slightly downward success rate in both the IC and internal validation techniques when the sample size increases. There is no difference between the three event rates. In the second experiment however, when all three covariates are involved in the data generating mechanism, there is convergence towards the true model. Namely, the success rate increases for both the IC and internal validation techniques when larger samples are used. The event rate increases the rate of convergence towards the data generating model.

Next, we look at the IC and internal validation techniques separately. The IC show quite different performance: the AIC seems to be the most stable choice in the different contexts. It performs second best in both experiments and has a high overall success rate. The BIC shows a more varied performance across the different contexts. It performs best in the first experiment, but worst in the second experiment. The internal validation techniques show an almost identical pattern: The optimism corrected AUC performs similar to the AIC in the first experiment, but performs slightly worse in the second experiment. The apparent

-non optimism corrected- AUC also performs varied. It has the highest success rate in the second experiment, but the lowest in the first experiment. This is opposite to the pattern of the BIC. The characteristics of the BIC and AUC can be explained. The penalty terms of the BIC leads it to prefer the more parsimonious model, which is beneficial in the first experiment [7]. However, in the second experiment the BIC is too conservative and doesn't choose the correct model. In most cases the AUC is too optimistic, which is a beneficial characteristic in the second experiment, but not in the first experiment [13]. These characteristics make them impractical in a real life context: we want to use a model selection method that performs well in various contexts.

However, we can only draw limited conclusions from the results in this mini thesis. Although the results in this study show that the AIC performs best overall, the data generating mechanisms are very simple and do not reflect the complexity of real world data. We used coefficients with an arbitrary and singular size, which is of course also not the case in different medical research areas. The use of different coefficients or more complex data generating mechanisms might give an advantage to other model selection methods.

The results also show that nonconvergence is a problem in the lowest sample size and event occurrence. It is not affected by the data generating mechanism. Nonconvergence in the rms package is a case of perfect separation, which is a known problem in logistic regression. This however is not a surprise, as this problem is more likely to occur in the context of low sample size and event occurrence [18]. There also appears to be a slight bump in the success rate for this context, which implies that this definition of nonconvergence might not be strict enough: a case of near perfect separation might be the remaining problem. This makes the problem of nonconvergence more complex, because it is not clear how to define near perfect separation. The problem of the remaining nonconvergence needs to be addressed in the final thesis. If this problem cannot be solved, interpretation in these contexts may be unreliable.

5 | CONCLUSION

We can conclude that the choice of model selection method, different IC and internal validation techniques, can have a large impact on final model selection. This is particularly crucial in medical prediction modelling, as the selected model informs significant clinical or life altering health interventions. In the final thesis we will need to use more complex and diverse data generating mechanisms to simulate data. The introduction of interaction variables, categorical variables, and different variance-covariance matrices can make the data more true to nature. This will result in a better understanding of the performance of IC and internal validation techniques in different contexts, allowing us to give better advice for medical prediction modelling.

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

18. van Smeden M, de Groot JAH, Moons KGM, et al. No Rationale for 1 Variable per 10 Events Criterion for Binary Logistic Regression Analysis. *BMC Medical Research Methodology.* 2016;16(1):163. doi: 10.1186/s12874-016-0267-3