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

Creating a good predictive model is a crucial step in medical research. Both information criteria and internal performance measures are used to select the best model in medical studies. This study investigates the performance of the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Area Under the Receiver Operating Characteristic Curve (AUC), and bootstrapped AUC in selecting the best model by simulating datasets with a known data generating mechanism. We simulated their ability to pick out the data generating model when a researcher is uncertain of the final candidate predictors in the model. Uncertainty can be expressed in candidate predictor inclusion or exclusion. We found a fixed order in the performance of including predictors correctly to the model: AUC, bootstrapped AUC, AIC, and lastly the BIC. This fixed order was reversed for correctly excluding predictors. By increasing sample size, the ability of the methods to include predictors improves, but the ability to exclude does not. The rate of improvement is influenced by different research contexts, such as separability, event rate, number of predictors and the type of predictors. The out-of-sample predictive performance of the methods is also influenced by these factors. We conclude that the AIC and bootstrapped AUC show high performance of selecting the data generating model in both inclusion and exclusion, and thus are suitable methods in general. However, if there are too little events per variable, the ability to pick the generating model is irrelevant, since the resulting prediction models may be inaccurate.

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. As a result, important choices are made regarding treatment and prevention based on these models. 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 the theoretical background that a researcher uses, several candidate models can arise. For the final model selection, two different applications of model selection can be employed.

Firstly, information criteria (IC) estimate the information loss when the probability distribution of the true data generating model is approximated by the probability distribution of a candidate model. By minimizing the 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 creates, on average, non-biased estimates and thus 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, AIC favors more complex models in large sample setting. This means 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[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 or best model in different contexts. So, we are interested in their ability to choose the correct model when the data generating mechanism is known. The ability to choose the correct model is crucial for the model's performance in practice. Testing the success rate of the IC (AIC and BIC) 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. Next to this, we are interested in the impact of choosing the correct model on the predictive out-of-sample performance of the model. The according research question is: How successful are Information Criteria and internal performance measures in choosing the correct prediction model in different simulated contexts? We will also investigate the impact of this ability on the predictive performance of the model. We expect a difference for event rate, sample size, separability in the population, and predictor differences.

2 | METHODS

This study investigates logistic medical prediction models designed for dichotomous risk prediction, such as disease occurrence. Our assumption is that the researcher is certain of a base model but uncertain of the final candidate predictors in the prediction model. As a result, we are interested in testing the ability of IC and internal performance measures to select the exact data generating model. This can be split up in two scenarios: first, the ability to include the candidate predictors that generated the data; second, the ability to exclude the candidate predictors that did not generate the data. By including the uncertain, but correct predictors, the researcher can improve the base model by making it more complex. By excluding the uncertain, but incorrect candidate predictors, the researcher can improve the model by making it more parsimonious.

To investigate the ability of detecting the full data generating mechanism, we simulated a covariate data set and calculated binary outcomes (i.e. diseased or healthy individuals) for both scenarios. To simulate the uncertainty of the candidate predictors in the first scenario, we gave their effects in the model only half the β coefficient of the base model effects. In the second scenario, we set the coefficients of the non-contributing candidate predictors to zero, which effectively leaves us with only the certain base model. The decision whether to include or exclude certain candidate variables is based on the metrics of the IC and internal performance measures between the respective models. An example of this is shown in Figure 1. Note how both models are uncertain of the same predictor and thus compare the same models in their respective analysis. We used several data generating mechanisms, i.e. models, to create the binary outcomes: models with a singular covariate, an interaction between covariates, or a combination of both as candidate predictors. This allowed us to compare the performance of the IC and internal performance measures in a broader set of contexts. To further create different contexts, we varied the number of predictors in the certain base model, the event rate, and the separability in the population. In the next section we discuss the simulation of the data sets.

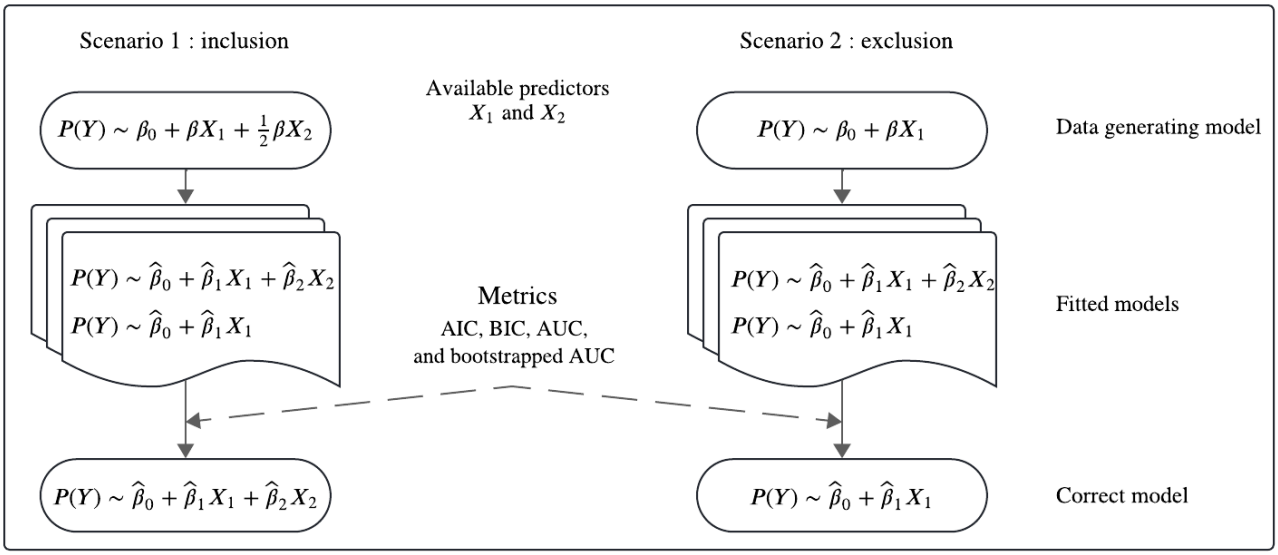

FIGURE 1 The two scenarios for the ability to detect the data generating model. The first scenario the correct model contains the candidate predictor and the second scenario the correct model does not contain the candidate predictor. The IC and internal performance criteria are used to select the final model.

2.1 | Simulating data sets

We generated two population data set of 10,000,000 individuals, characterized by five covariates (X_1, X_2, X_3, X_4, and X_5) and ten covariates ($X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$). We used a correlation of 0.2 between the covariates, to simulate realistic relationships. The population data sets were generated using standard normal distributions in the `mvrnorm` function of the MASS package in R [15]. The population data sets allowed us to generate the binary outcomes for the two scenarios. The mean vector (μ) and covariance matrix (Σ) for the covariates were defined as follows:

$$\boldsymbol{\mu_5} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{5\times 1}, \quad \boldsymbol{\Sigma_5} = \begin{bmatrix} 1 & 0.2 & \dots & 0.2 \\ 0.2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.2 \\ 0.2 & \dots & 0.2 & 1 \end{bmatrix}_{5\times 5}$$

$$\boldsymbol{\mu_{10}} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{10\times 1}, \quad \boldsymbol{\Sigma_{10}} = \begin{bmatrix} 1 & 0.2 & \dots & 0.2 \\ 0.2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.2 \\ 0.2 & \dots & 0.2 & 1 \end{bmatrix}_{10\times 10}$$

Next, we calculated the possible binary outcomes based on this population data set. This was done using a logistic regression model. The value of outcome group was defined Y = 1 if the probability P of the outcome was greater than 0.5, and Y = 0 if P was less than 0.5. The probability of the outcome $P(Y = 1)$ given the covariates was derived as follows, given β_0 and β:

$$P(Y = 1) = \frac{1}{1 + e^{-(model)}}$$

To simulate a range of different contexts, we created several models. The effects with only half the size of the other coefficients represent the uncertain effects in the model of the researcher. These are the models that are used in the first scenario of correct inclusion and the analysis on the left side of Figure 1. The models with five predictors are defined below, where the ten predictor models follow the same structure with the added predictors X_6, X_7, X_8, X_9, and X_{10} in each model:

$$\begin{aligned}
\text{Model } 1 &: \beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_5) \\
\text{Model } 2 &: \beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_1X_2) \\
\text{Model } 3 &: \beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_5) \\
\text{Model } 4 &: \beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_1X_2) \\
\text{Model } 5 &: \beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_1X_2 + 0.5 * X_2X_3)
\end{aligned}$$

When the candidate predictors do not contribute to the data generating model, the certain base models are defined below. These are the models that are used in the second scenario of correct exclusion and the analysis on the right side of Figure 1. The models with five predictors are defined below, where the ten predictor models follow the same structure with the added predictors X_6, X_7, X_8, X_9, and X_{10} in each model:

$$\begin{aligned}
\text{Model } 1 &: \beta_0 + \beta(X_1 + X_2 + X_3 + X_4) \\
\text{Model } 2 &: \beta_0 + \beta(X_1 + X_2 + X_3 + X_4) \\
\text{Model } 3 &: \beta_0 + \beta(X_1 + X_2 + X_3) \\
\text{Model } 4 &: \beta_0 + \beta(X_1 + X_2 + X_3) \\
\text{Model } 5 &: \beta_0 + \beta(X_1 + X_2 + X_3)
\end{aligned}$$

The separability in the population was operationalized by changing β, such that in each model we have the same AUC of 0.6, 0.75, and 0.9 in the population. The β-coefficients were tuned by calculating the AUC in the population by increasing the β of the models by 0.001 per step until the AUC was reached. This was done for the models with the candidate predictors that did contribute to the data generating model. The same coefficients were then used for the data generating models where the candidate predictor did not contribute, to be able to make a direct comparison between the scenarios. However, this means that the data generating models without the candidate predictors have a lower true AUC. The β_0 was set to the value that resulted in a 50% probability of the outcome, which could be calculated by subtracting the mean β_0 of the model. The coefficients (β) of the five and ten predictor models are defined in Table 1.

TABLE 1 β-coefficients of the models for AUC 0.6, 0.75, and 0.9

Predictors	AUC	Model 1	Model 2	Model 3	Model 4	Model 5
5	0.6	0.136	0.114	0.150	0.107	0.124
	0.75	0.391	0.332	0.435	0.312	0.360
	0.9	0.944	0.811	1.056	0.749	0.877
10	0.6	0.073	0.068	0.076	0.071	0.066
	0.75	0.211	0.197	0.221	0.205	0.192
	0.9	0.512	0.479	0.539	0.502	0.467

2.2 | Analysis

For the analysis we compared the model with and without the uncertain candidate predictors. In case of multiple uncertain predictors, we compared all models that contain one or two of the uncertain predictors. The comparison is performed by fitting all the necessary models in a logistic regression with the `lrm` function in the rms package[11]. The ability to select the correct data generating models was analyzed by comparing the IC and internal performance measure metrics on their tested models. Per metric there is a winning model.

The AIC and BIC of the models were calculated in the summary statistics of the `lrm` object. The internal performance measure AUC was also calculated in the summary statistics of the fitted `lrm` object. To correct for the optimism in the apparent AUC in the sample, we bootstrapped the sample and calculated the optimism-corrected 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, using 80 bootstraps. Finally, a validation on a test set, ten times the sample size, was performed and the according AUC was calculated. This could be used for assessing the out-sample performance of the methods.

A range of 50 to 1,000 samples from the population data set with intervals of 50 was used to test the success rate of the IC and internal performance measures. Each sample size was iterated 2000 times, which allowed us to calculate the success rate. The success rate was calculated by dividing the number of times the correct model was selected by the total number of converged iterations. Convergence was defined as the models that did not show non-convergence warnings. We also recorded non-successful, i.e. non-converging, iterations of models. To create different contexts, the simulation was repeated for three different event rates (5%, 20%, and 50%). The event rates were created by under sampling the population with 50% event rate to the desired rate. Visualizations were generated using the package ggplot2[17]. The simulations, analyses and visualizations were performed in R-studio[18] and are available on Github.

3 | RESULTS

In this section we discuss the outcome of the simulation: the ability of IC and internal performance measures to correctly include and exclude candidate predictors to the model. The ability to choose the correct model was defined as the success rate of choosing the data generating model. Success rate was quantified as the proportion of total iterations the data generating model was chosen. We also investigated the impact of this ability on the AUC of the model when validated on the test set. In the figures success rate and test set AUC was plotted against sample size. The results are presented in the following four sub-studies: the effect of separability, event rate, model size, and model differences.

3.1 | Study 1: Separability (AUC)

This study investigated the effect of separability in the population on the ability to correctly include and exclude predictors to the model. We looked at the total trend for all methods and the differences between them. AUC is used to measure the separability in the population, defined as the ability of the model to discriminate between the two classes. The first row of Figure 2 illustrates the success rate of correctly including the last predictor to the $P(Y=1) : \beta_0+\beta(X_1+X_2+X_3+X_4+0.5*X_5)$ model with fixed event rate 50% at AUC 0.6, 0.75 and 0.9. The second row shows the AUC when validated on the test set for the methods. The first row of Figure 3 illustrates the success rate of correctly excluding predictor X_5 from the correct $P(Y=1) : \beta_0+\beta(X_1+X_2+X_3+X_4)$ model with fixed event rate 50% at AUC 0.6, 0.75 and 0.9. Again, the second row shows the AUC when validated on the test set for the methods. Both figures use their respective scale of the population AUC for the y-axis.

The first row of Figure 2 shows an upward trend with sample size for the ability to correctly include predictors to the model in all methods. Higher linear separation in the population resulted in a steeper increase of success rate with increasing sample size for all the methods. Once we look at the methods separately, we see that the AUC performs best in detecting the complex model, followed by the bootstrapped AUC, which is intersected by the AIC and lastly the BIC. The second row of the figure shows the AUC in the test set, where a different pattern than for the success rate is observed: the population with a separability of AUC 0.6, the AUC in the test-set is almost equal for all methods across the range of sample sizes. There is even a slight advantage for the BIC. Even though the last predictor does contribute to the data generating model, this model results to a lower AUC when validated in the test set. The population with a separability of AUC 0.75 shows that the difference between the methods in AUC of the test set becomes larger, which is advantageous for the methods with higher success rate. This resulted in the same order

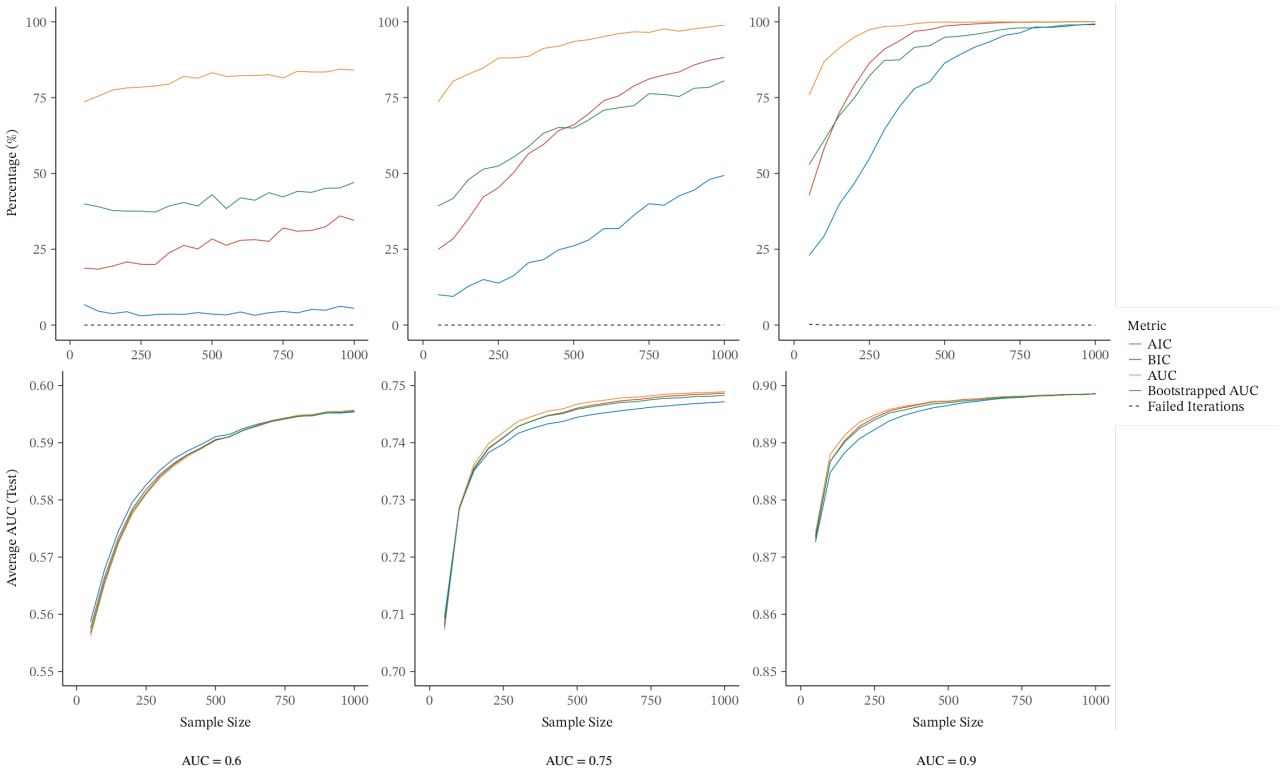

FIGURE 2 Success rate and AUC in the test set of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_5)$ model with event rate 50% and AUC 0.6, 0.75 and 0.9. X_5 is the candidate predictor. Note: different scales on the y-axis for the test set AUC.

of performance as the success rate: the AUC, bootstrapped AUC, AIC, and BIC. In the population with a separability of AUC 0.9 this advantage disappears and the performance difference between the methods decreases with sample size. This is a result of the high success rates in these conditions. The shape of the curves in the second row can be seen as asymptotic to the true AUC in the population. Lower separability in the population has a slower rate of approach to the population AUC asymptote.

The first row of Figure 3 shows that the ability to correctly exclude predictors to the model seems to have no trend with sample size. Better linear separation in the population data resulted in lower success rates for the IC and internal performance measures once sample size is increased. In low sample sizes, the success rate was higher in population with better separability. The following order of success rate is visible: the BIC, AIC, bootstrapped AUC and lastly AUC. The second row of the figure illustrates the AUC when the model is validated in the test set. It shows a similar pattern as before with one important difference: the BIC performs best in the populations with the separability of AUC 0.6, 0.75, and 0.9. In the population with a separability of AUC 0.6, the AUC in the test-set has the biggest difference between all methods, with the BIC performing best. The difference between the methods becomes smaller with increasing separability in the population: the different methods result in the same value of AUC when validated on the test set. The rate of approach to the population asymptote is similar to the first inclusion scenario of this study.

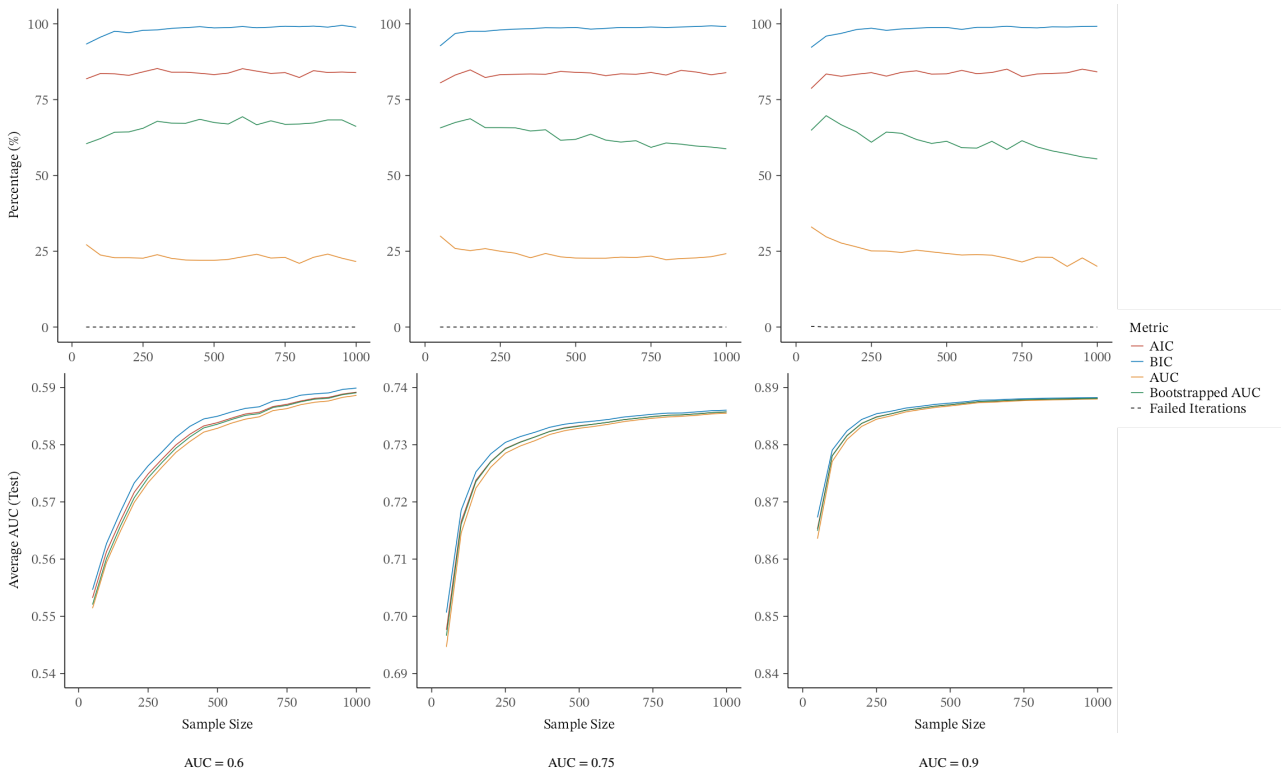

FIGURE 3 Success rate and AUC in the test set of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4)$ model with event rate 50% and AUC 0.6, 0.75 and 0.9. X_5 is the candidate predictor.

3.2 | Study 2: Event rate

This study looked at the effect of the event rate on the ability to correctly include and exclude predictors to the model. The first row of Figure 4 illustrates the success rate to correctly include the last predictor to the $P(Y = 1) : \beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_5)$ model with the separability in the population of AUC 0.9 at event rates 5%, 20%, and 50%. The second row shows the AUC when validated on the test set for the methods. The first row of Figure 5 illustrates the ability to correctly exclude predictor X_5 from the $P(Y = 1) : \beta_0 + \beta(X_1 + X_2 + X_3 + X_4)$ model with the separability in the population of AUC 0.9 at event rates 5%, 20%, and 50%. Again, the second row shows the AUC when validated on the test set for the methods.

The first row of Figure 4 shows that a higher event rate results in a higher overall success rate for the IC and internal performance measures once sample size is increased. The AUC performed best in detecting the complex model, followed by the bootstrapped AUC, which is intersected by the AIC and lastly the BIC. This is the same order as in study 1. All methods show better performance in higher population AUC values and higher sample sizes. A larger number of non-converging models occurred in the 5% event rate compared to the 20% and 50% event rates. The second row of the figure shows the performance in the test set of the methods. The event rate mediates the difference between the methods in the test set: a lower event rate resulted in a larger difference between the methods. As a result, the difference between the methods in the largest event rate becomes negligible, except for the BIC. In all three event rates, the BIC performs the worst, followed by the AIC, bootstrapped AUC, and lastly the AUC. In larger sample sizes, the difference between the methods decreased, especially the between the BIC and other methods. In the 20% and 50% event rates the BIC became similar to the other methods when sample size is high. This is a result of the high success rates in these conditions. Lower event rate in the population has a slower rate of approach to the population AUC asymptote.

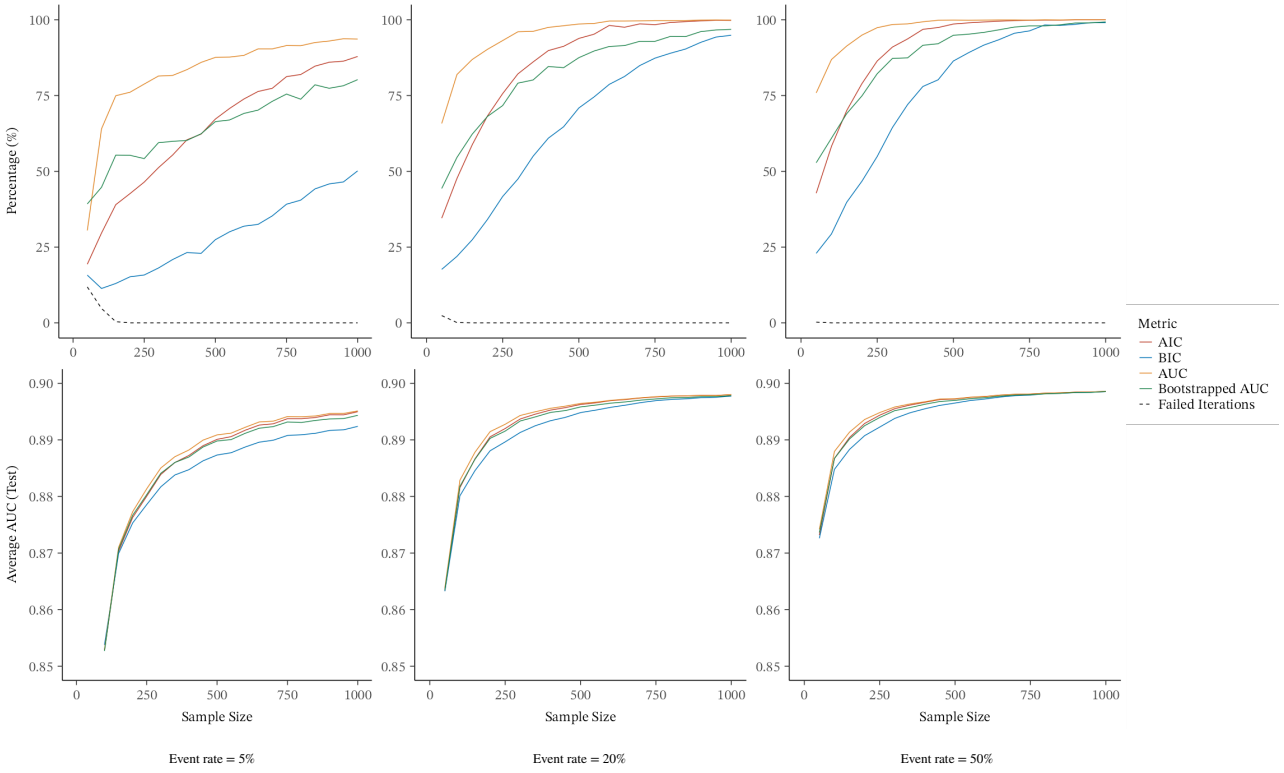

FIGURE 4 Success rate and AUC in the test set of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_5)$ model with AUC 0.9 and event rate 5%, 20% and 50%. X_5 is the candidate predictor

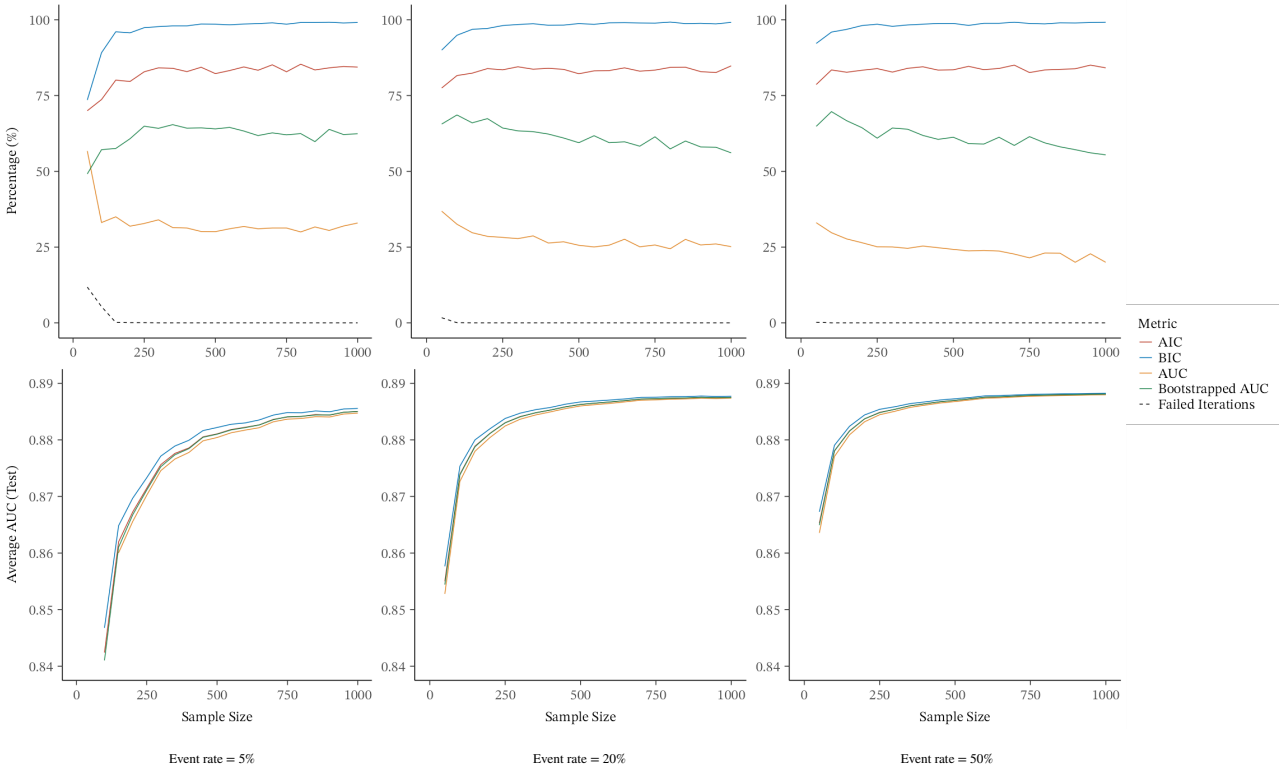

FIGURE 5 Success rate and AUC in the test set of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4)$ model with AUC 0.9 and event rate 5%, 20% and 50%. X_5 is the candidate predictor.

A pattern in the ability to correctly exclude the uncertain predictor X_5 seems non-existent in Figure 5. All three event rates showed a similar slightly downward trend in the success rate of the IC and internal performance measures with sample size. A larger number of non-convergence in the 5% event rate compared to the 20% and 50% event rates can be observed. From highest to lowest success rate: the BIC, AIC, bootstrapped AUC and lastly the AUC. The AUC when validated on the test set in the second row, shows a similar pattern as in study 1. The BIC performs best in low event rates. This advantage becomes negligible in higher event rates and sample sizes. All other methods showed similar performance across the contexts. The asymptote of lower event rates is slower to approach the population AUC than higher event rates.

3.3 | Study 3: Model size

In this study we looked at the effect of the model size on the ability to correctly include and exclude predictors to the model. The first row of Figure 6 illustrates the ability to correctly include the last predictor to the $P(Y=1): \beta_0+\beta(X_1+X_2+X_3+X_4+0.5*X_5)$ model and the $P(Y=1): \beta_0+\beta(X_1+X_2+X_3+X_4+0.5*X_5+X_6...X_{10})$ model with separability in the population of AUC 0.9 and event rate 50%. The second row shows the AUC of the methods when validated on the test set. Figure 7 demonstrates the percentage and AUC in the test set for excluding the uncertain predictor X_5 from the $P(Y=1): \beta_0+\beta(X_1+X_2+X_3+X_4)$ model and the $P(Y=1): \beta_0+\beta(X_1+X_2+X_3+X_4+X_6...X_{10})$ model. This was calculated for a separability in the population of AUC 0.9 and event rate 50%.

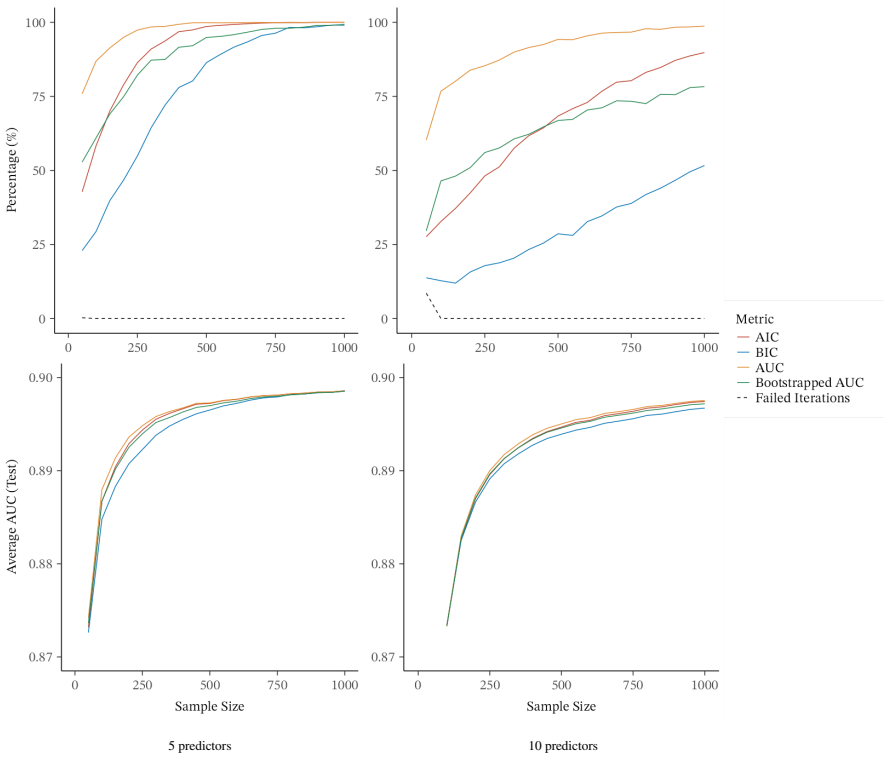

FIGURE 6 Success rate and AUC in the test set of IC and internal performance measures for the $\beta_0+\beta(X_1+X_2+X_3+X_4+0.5*X_5)$ and $\beta_0+\beta(X_1+X_2+X_3+X_4+0.5*X_5+X_6...X_{10})$ model with AUC 0.9 and event rate 50%. X_5 is the candidate predictor.

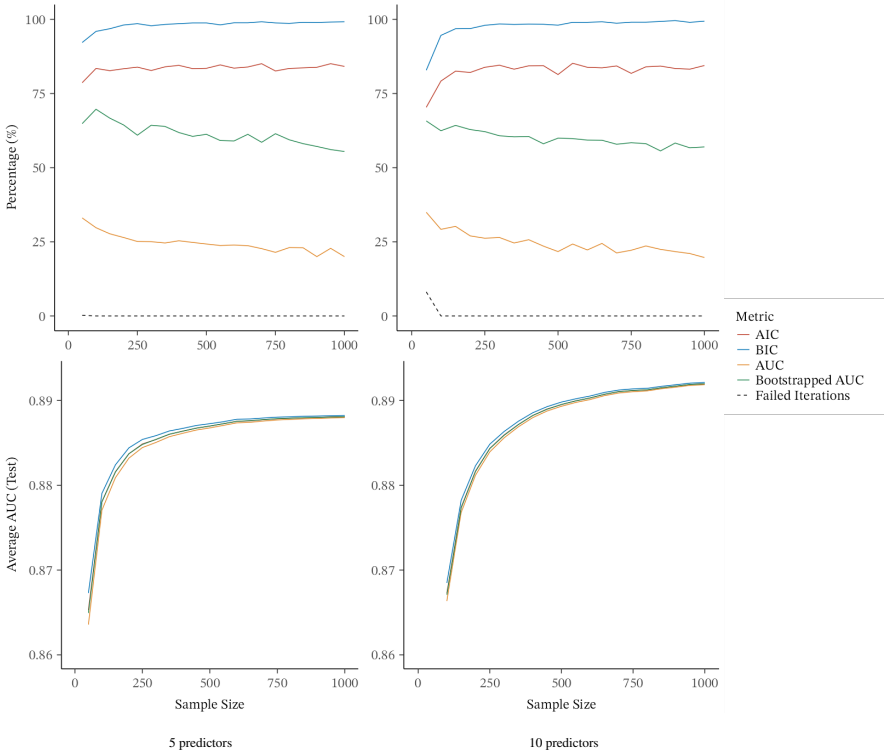

FIGURE 7 Success rate and AUC in the test set of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4)$ and $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + X_6 ... X_{10})$ model with AUC 0.9 and event rate 50%. X_5 is the candidate predictor.

For the first scenario of inclusion Figure 6 shows that the 5-predictor model starts of with higher overall performance rates. It also shows faster increase in performance rates compared to the 10-predictor model. The pattern of the 10-predictor model looks similar to the separability in the population of AUC 0.75 (Figure 3) or an event rate of 5% (Figure 4) in study 1 and study 2, respectively. The same mediating effect on the sample size and event rate applied to model size. However, the candidate predictor in the 10-predictor model had a smaller coefficient compared to the 5-predictor model. This contributed to a slower increase in performance rates. The order of success rate reoccurred: AUC, followed by the bootstrapped AUC, which is intersected by the AIC and lastly the BIC. The second row of Figure 6 shows the AUC of the methods when validated on the test set. The 5-predictor model showed larger differences between the methods compared to the 10-predictor model in low sample sizes. However, the 10-predictor model showed a larger difference between the methods in higher sample sizes. This is the same pattern as for the separability in the population of AUC 0.75 or an event rate of 5%. In both models the BIC performs the worst. The rate of approach to the population AUC is higher in the 5-predictor model, which can be explained by the higher success rates of the methods.

Figure 7 illustrates that both the 5-predictor and the 10-predictor model show a similar pattern in the success rate of the IC and internal performance measures. The 10-predictor model only showed higher non-convergence rates at low sample size. The same pattern of success rate from highest to lowest occurred: BIC, AIC, bootstrapped AUC and lastly AUC. The second row of the figure shows a larger difference between the methods for the 5-predictor model compared to the 10-predictor model. In both models there was a slight advantage for the BIC, preferring the model without the candidate predictor. However, the differences can be viewed as negligible. The rate of approach to the population AUC was higher in the 5-predictor model than in the 10-predictor model. Since both models had similar success rates, the 5-predictor fitted the right coefficients faster. The 10-predictor model also had a higher asymptotic AUC compared to the 5-predictor model, since it had more certain predictors to model the data.

3.4 | Study 4: Model differences

In this study we looked at the effect of different candidate predictors on the ability of IC and internal performance measures to correctly include and exclude predictors to the model. The first row of Figure 8 shows the ability to correctly include the candidate predictor X_5 and X_1X_2 to the $P(Y = 1) : \beta_0+\beta(X_1+X_2+X_3+X_4+0.5*X_5)$ model and $P(Y = 1) : \beta_0+\beta(X_1+X_2+X_3+X_4+0.5*X_1X_2)$ model, respectively. The first row of Figure 9 illustrates the ability to correctly include the two candidate predictors $\{X_4, X_5\}$, $\{X_4, X_1X_2\}$, and $\{X_1X_2, X_2X_3\}$ to the $P(Y = 1) : \beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_5)$ model, the $P(Y = 1) : \beta_0+\beta(X_1+X_2+X_3+0.5*X_4+0.5*X_1X_2)\}$ model, and $P(Y = 1) : \beta_0+\beta(X_1+X_2+X_3+X_4+0.5*X_1X_2+0.5*X_2X_3)$ model. The separability in the population is set to AUC 0.9 and event rate 50%. The second rows of both figures show the AUC when validated on the test set for the methods. The effect of the model size on the ability to exclude the last predictors from the models and the AUC validated on the test set is not illustrated, but shown in Appendix B. In case of one uncertain predictor, the success rate was defined as the proportion of total iterations the data generating model is chosen out of the two models. In case of two uncertain predictors, the success rate was defined as the proportion of total iterations the data generating model is chosen out of the four models. This means that the nested versions, only containing one out of the two uncertain predictors, were also considered incorrect.

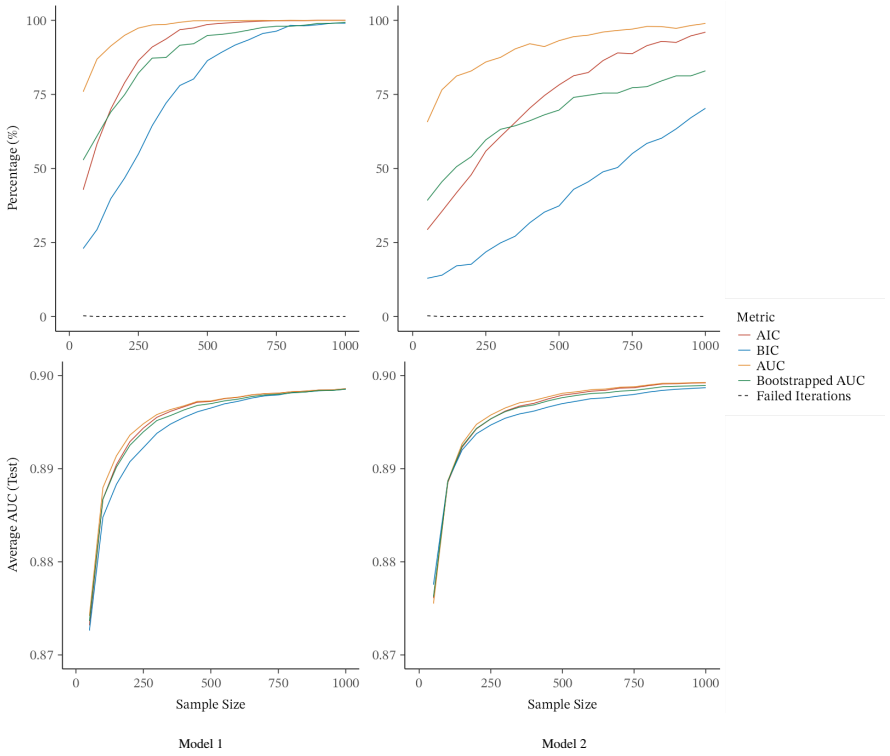

FIGURE 8 Success rate and AUC in the test set of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_5)$ and $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_1X_2)$ model with AUC 0.9 and event rate 50%. X_5 and X_1X_2 are the respective candidate predictors.

The first rows of Figure 8 and Figure 9 show that the ability to correctly include predictors to the model is dependent on the type of uncertain variable we use. The models that contained one or two interaction terms showed a slower increase of success rate with sample size. The overall success rates also started off lower for most methods when an interaction term was present. However, there was no difference in the order of success rate between the IC and internal performance measures in all the models. The consistent order of the AUC, followed by the bootstrapped AUC, which is intersected by the AIC and lastly the BIC persisted. The second rows of the figures show that models containing interaction terms show no differences in the AUC of the test set at low sample. However, a difference appeared at high sample sizes for the BIC, that also lacked behind in success

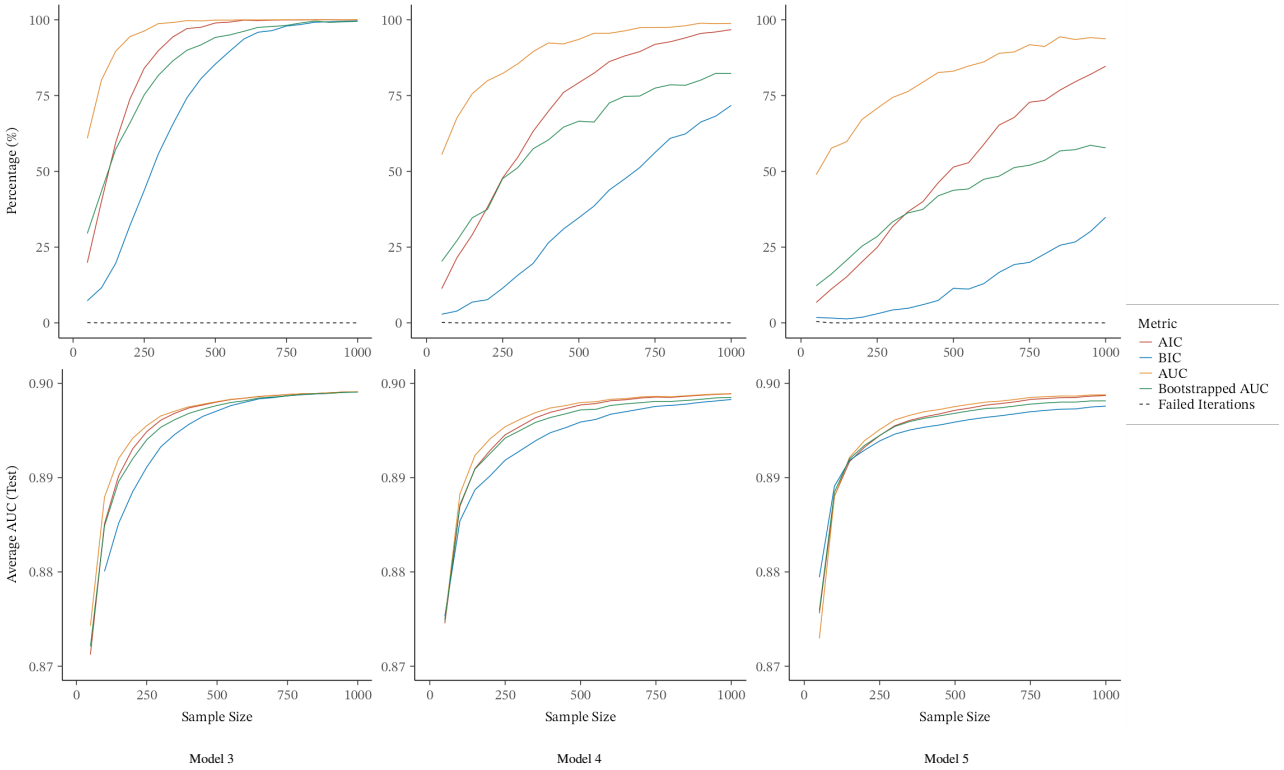

FIGURE 9 Success rate and AUC in the test set of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_5)$, $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_1X_2)$, and $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_1X_2 + 0.5 * X_1X_3)$ model with AUC 0.9 and event rate 50%. $\{X_4, X_5\}$, $\{X_4, X_1X_2\}$, and $\{X_1X_2, X_2X_3\}$ are the respective candidate predictors.

rate. The models without interaction term showed a difference in the AUC of the test set at low sample sizes, but the difference became smaller at high sample sizes. The model with both a normal end predictor and interaction term, had a hybrid pattern of these two patterns. The metric AUC seemed to have the highest AUC when validated in the test set, followed by AIC and bootstrapped AUC. The AUC asymptotes appeared to be the similar for the different candidate predictors.

The ability to correctly exclude the candidate predictor shows no pattern in success rate with sample size. All candidate predictors showed a similar pattern, which is visible in Figure 3 at AUC = 0.9 for the X_5 candidate predictor in the $P(Y = 1) : \beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_5)$ model. The only difference between the candidate predictor models was the height of the AUC asymptotes, due to the size of the predictor coefficients.

4 | DISCUSSION

In the field of medical prediction modelling both IC and internal validation techniques play an important role in the process of model selection. This paper offers a comprehensive overview of the success rate of IC and internal performance measures in choosing the correct model in different contexts. This paper provides additional insights in the effect of different methods on the out-of-sample performance of the model. After simulation data with a multitude of data generating mechanisms under a variety of contexts, we conducted four sub-studies to investigate the impact of the separate contexts on the ability of selecting the correct model and the out-of-sample performance.

We found a consistent order of IC and internal performance measures in both the scenario of including and excluding candidate predictors. In every context where we want to correctly include a predictor to the model, the AUC in the sample performs best, followed by the bootstrapped AUC, which is intersected by the AIC once we increase sample size and lastly the BIC. The difference between AIC and BIC can be explained by their tendency to favor more complex and parsimonious model, respectively [7]. The AUC also tends to prefer more complex models, due to overly optimistic performance measures [13], which can be corrected by the bootstrapped AUC [11 12]. All measures show a monotonic increase in success rate with increasing sample size.

Larger separability and higher event rate accelerate the rate at which the success rate increases with sample size. Model term differences are also of importance, as the rate at which the success rate increases with sample size is lower when an interaction term is present. Next, in every context where we want to correctly exclude a predictor from the model, we see that the BIC performs best, the AIC, followed the bootstrapped AUC and lastly the AUC in the sample. This also fits the expected pattern of preferring more parsimonious over complex models by the methods. A pattern in the success rate with increasing the sample size seems to be non-existent and can even decrease in optimal situations for including a predictor. Since only the base models created the data here, there were no effects for model differences. These factors are of importance when choosing the correct model in a medical prediction model and should be considered when choosing the final model, since the highest out-of-sample performance is the goal.

In our sub-studies regarding the validated AUC of the chosen models by the IC and internal performance measures, we only see slight differences between the contexts. In general, all IC and internal performance measures show similar performance in both their ability to include and exclude candidate predictors. A difference in the AUC of the test set becomes visible in the ability to include predictors in the model: in cases of low sample size, low event rate, high model size, and interaction variables as candidate predictors we see the AUC in the test set diverging between the methods. This is most predominant in the BIC, which under performs in these extreme cases, with the AIC, AUC, and bootstrapped AUC all performing similarly. In those cases, the increase of sample size results in a convergence of the AUC in the test set for all methods, including the BIC. This can be explained, since in these extreme cases the success rate of the BIC is lower than the other methods. This results in choosing a more parsimonious model and thus less explained variance. However, we do need to note that despite that there are small differences on large scale, in general contexts the differences between the methods are small to negligible. In practice only small difference in AUC can be expected between the methods.

Even though both IC and internal performance measures are widely used in medical prediction modelling, it remained unclear how these methods compare in their ability to select the correct or best model in different contexts. Our results show that the AIC and bootstrapped AUC are the best performing methods in selecting the correct model in most general contexts. The AUC in the sample is overly optimistic due to its inability to exclude candidate predictors that do not contribute to the data generating mechanism. In scenarios of low event rate and medium to high separability in the BIC underperforms in selecting the correct model and in the out-of-sample performance. These results, divided in several sub-studies, provide guidance for researchers in choosing the correct model in medical prediction modelling. Decisions about the model selection method should be based on the context of the data, such as the sample size, event rate, population separability, and the model terms.

However, not all contexts are usable for different model selection methods. In cases of low sample size, low event rate, and large model size separation issues, i.e. perfect or near perfect prediction, can occur. These conditions can be summarized as low events per variable (EPV), a common problem in medical prediction modelling, leading to biased estimates of the model parameters[19 20]. In this simulation this occurs through high non-convergence rates and diverging IC and internal performance measure rates at low EPV and high population separability. Even though they don't bias the analysis of success rate, they do have practical implications. Low EPV can lead to overfitting of the model, which can result in poor out-of-sample performance. This is especially important in medical prediction modelling, where the model is used to make decisions about the patient's health. Strategies such as Firth's penalized likelihood[21] or LASSO[22] regression can be used to correct for bias in the model parameters. However, low EPV in combination with correction methods do not offer a guarantee and may also yield poor out-of-sample performance[23]. In these cases, prediction models should be used with caution, or even avoided.

Some limitations do apply to our simulation study. We looked at a finite number of a scenarios, which may not cover all possible scenarios. However, we did cover a wide range of scenarios, which are commonly encountered in medical prediction modelling. We also reduced our scope of the study to only AIC and BIC as IC and AUC and bootstrap AUC as internal performance measures. There are other IC and internal performance measures that could be used in practice, such as the likelihood ratio test, the adjusted R-squared, or the integrated discrimination improvement. However, the AIC and BIC are the most commonly used IC measures in practice and the AUC and bootstrapped AUC are the most commonly used internal performance measures. Future studies could investigate the effect of more specific IC and internal performance measures on the success rate of model selection and out-of-sample performance.

We conclude that the AIC and bootstrapped AUC have the best ability of the IC and internal performance measures in choosing the data generating model in medical prediction modelling. In more general contexts they lead to high ability of correct model selection and high out-of-sample performance. Our recommendation is to use the AIC and bootstrapped AUC in the process of model selection with uncertain candidate predictors. However, we advise strongly against fitting models in scenarios of low EPV, since this can lead to biased estimates of the model parameters and poor out-of-sample performance.

□

APPENDIX

A SCENARIO 1: INCLUDING PREDICTORS

In the following figures we show all the results of the first scenario, where we want to correctly include the candidate predictors in the model. The models that create the data are shown in the method section. These figures show the success rate of selecting the correct model for the IC and internal performance measures. The figures also show the test AUC of the IC and internal performance measures for the different models. The test AUC is calculated as the AUC of the model selected by the IC and internal performance measures when validated on the test set.

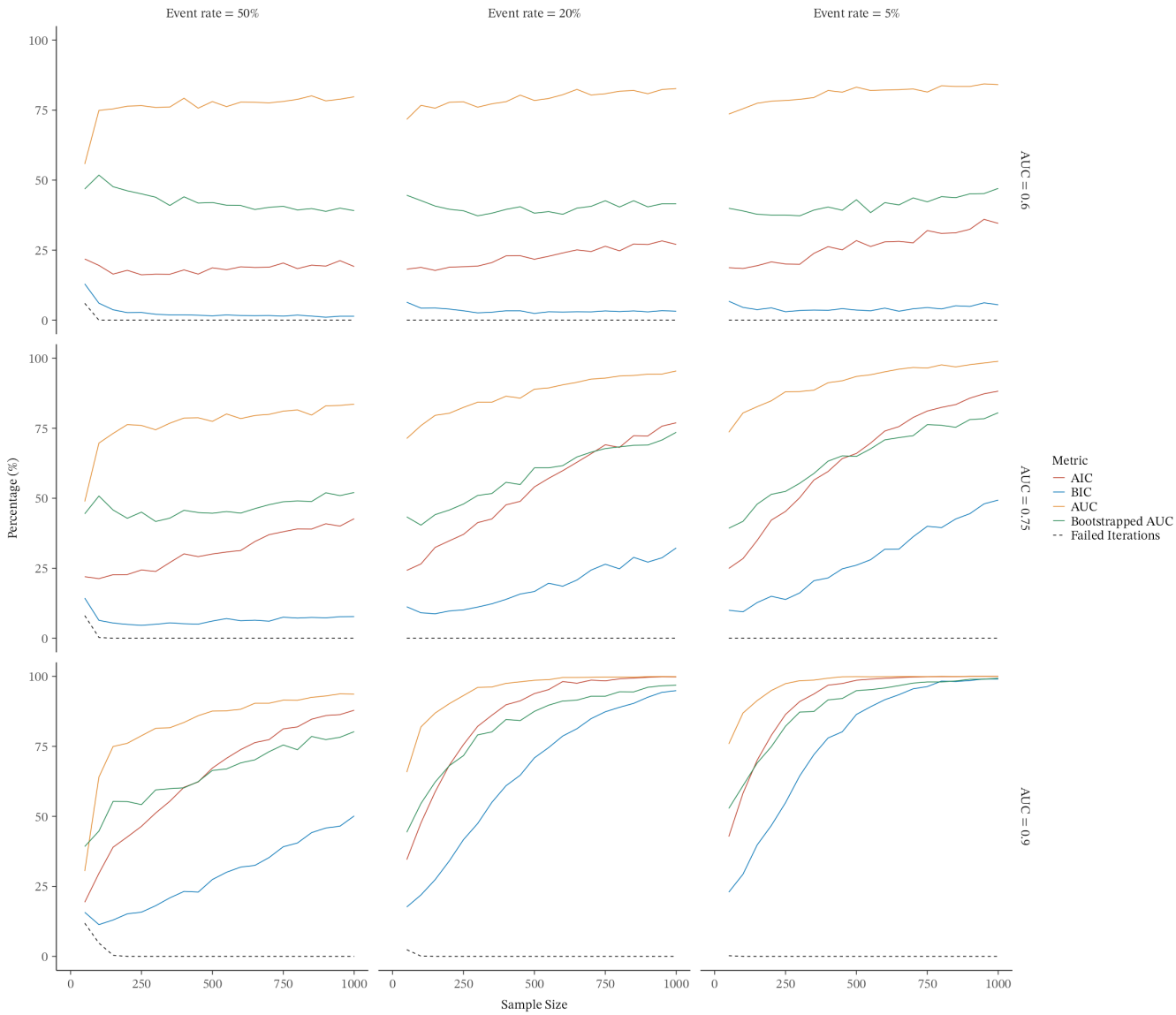

FIGURE A1 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_5)$ model with X_5 as candidate predictor.

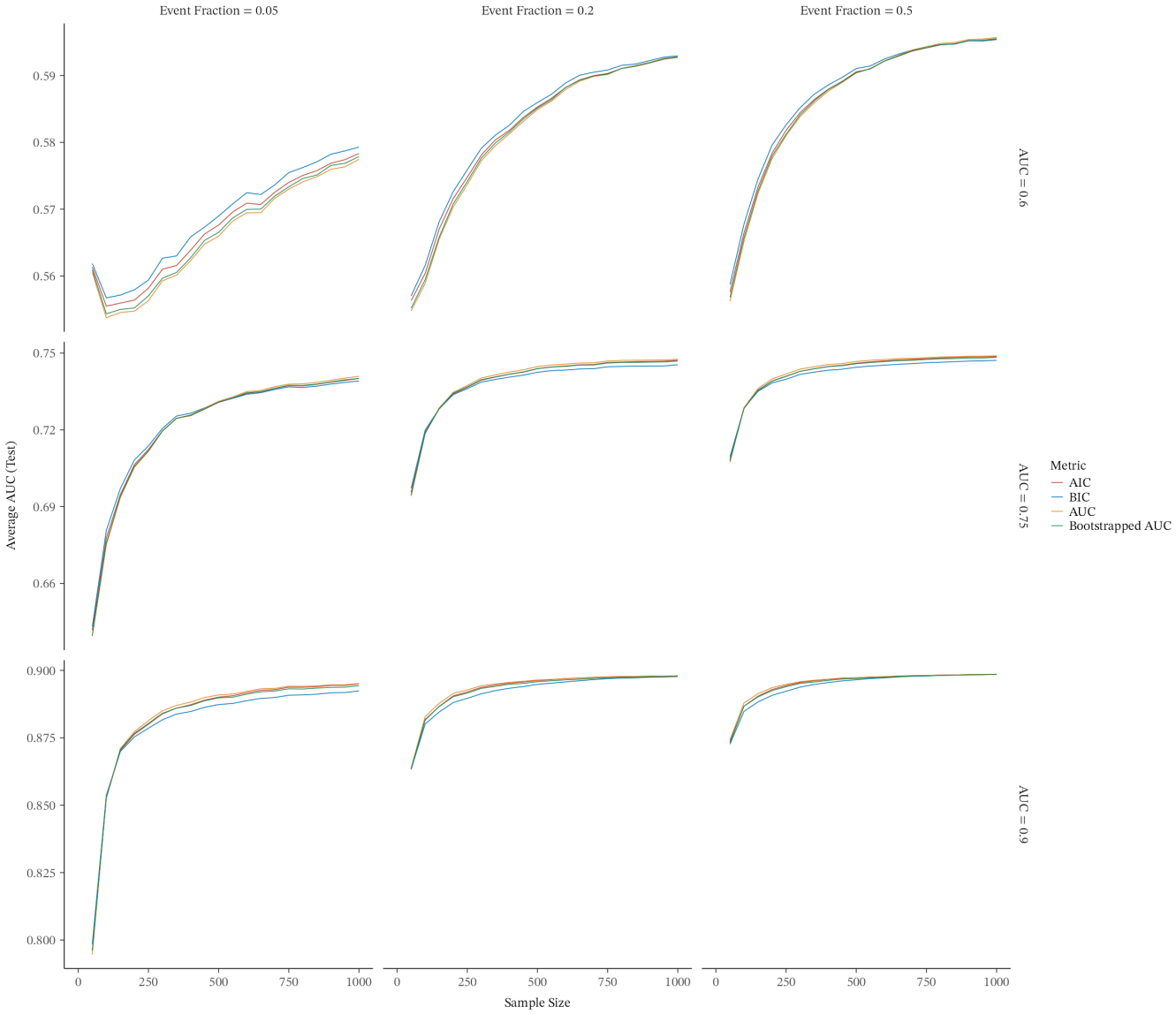

FIGURE A2 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_5)$ model with X_5 as the candidate predictor.

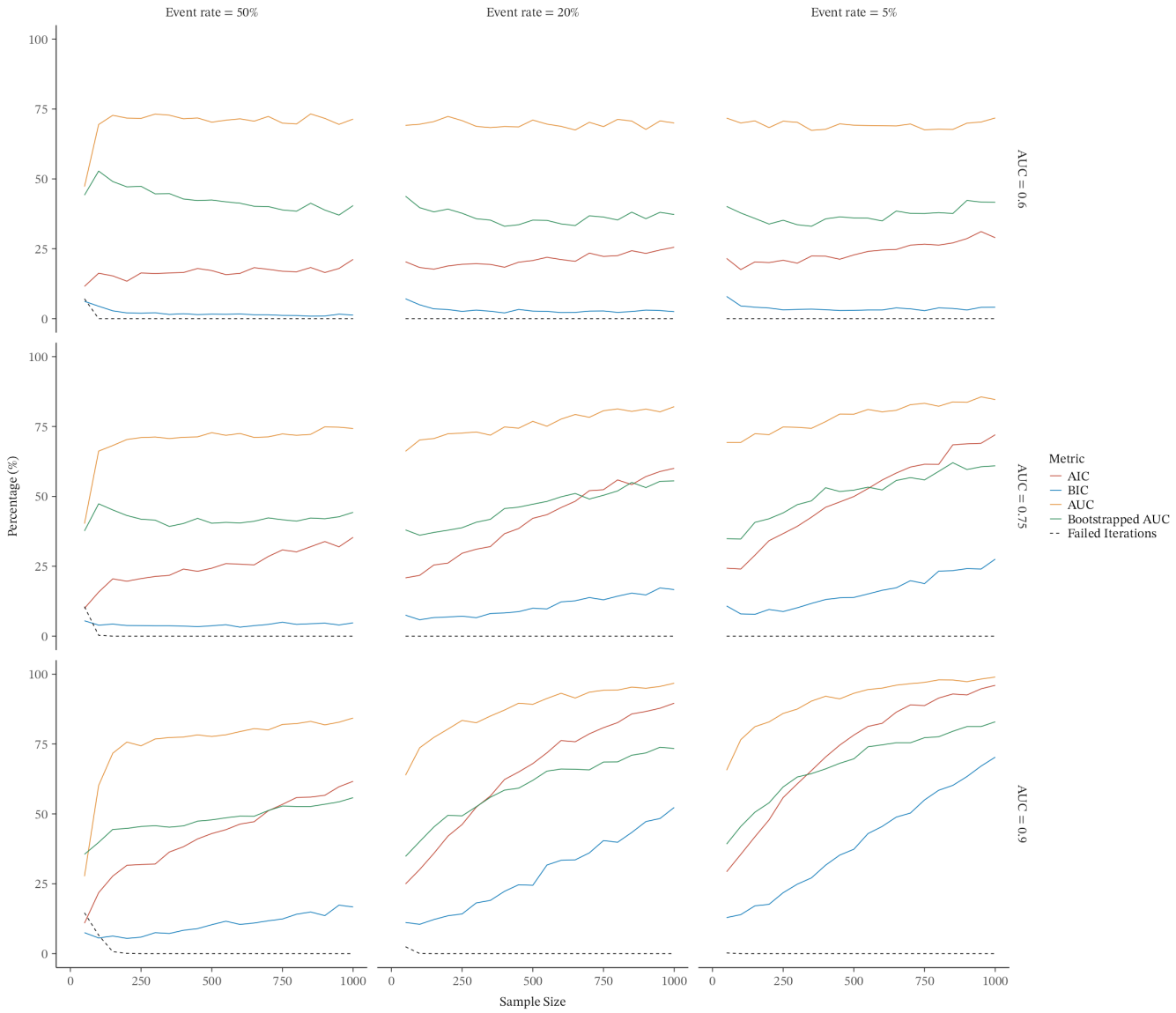

FIGURE A3 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_1X_2)$ model with X_1X_2 as the candidate predictor.

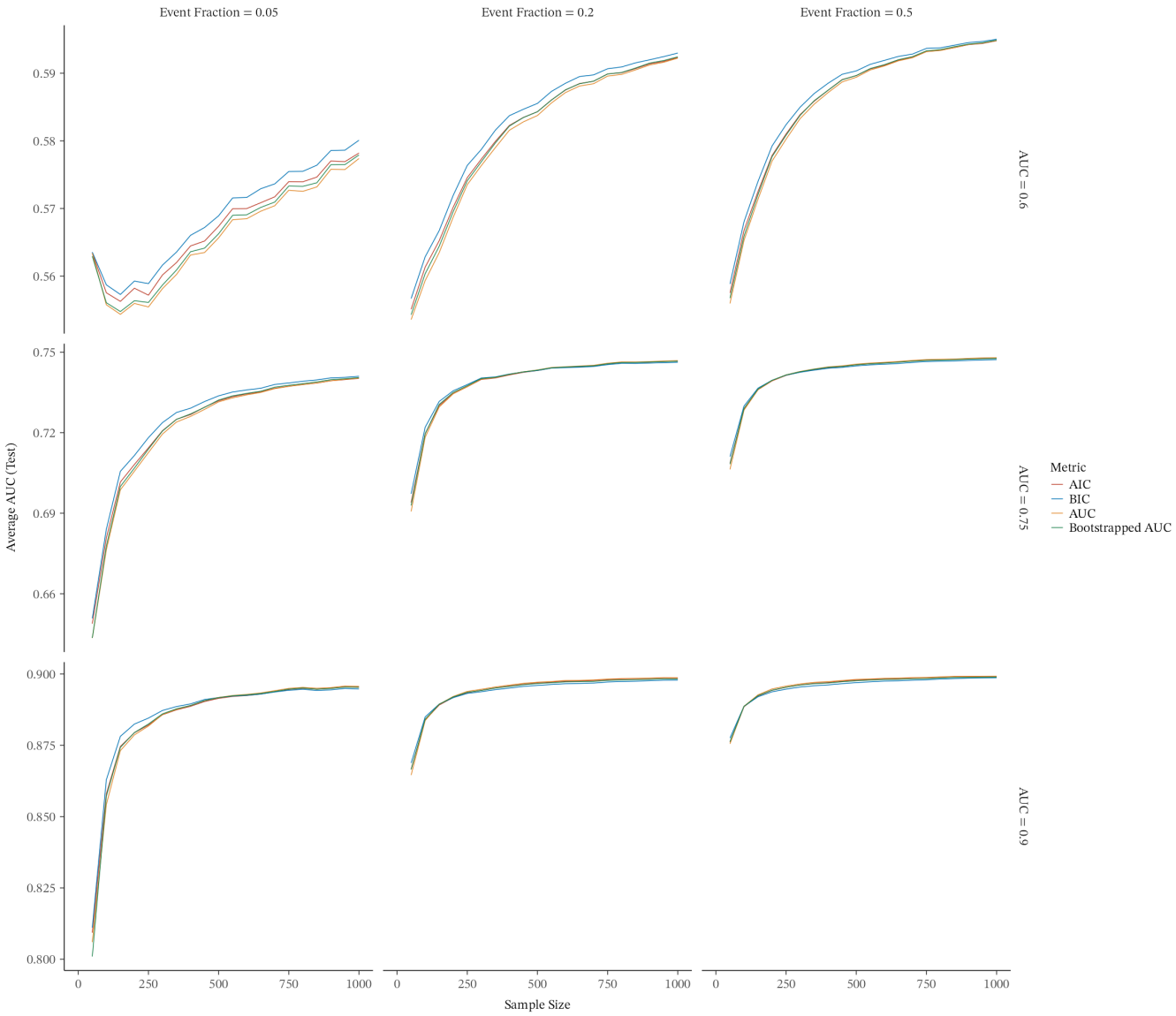

FIGURE A4 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_1X_2)$ model with X_1X_2 as the candidate predictor.

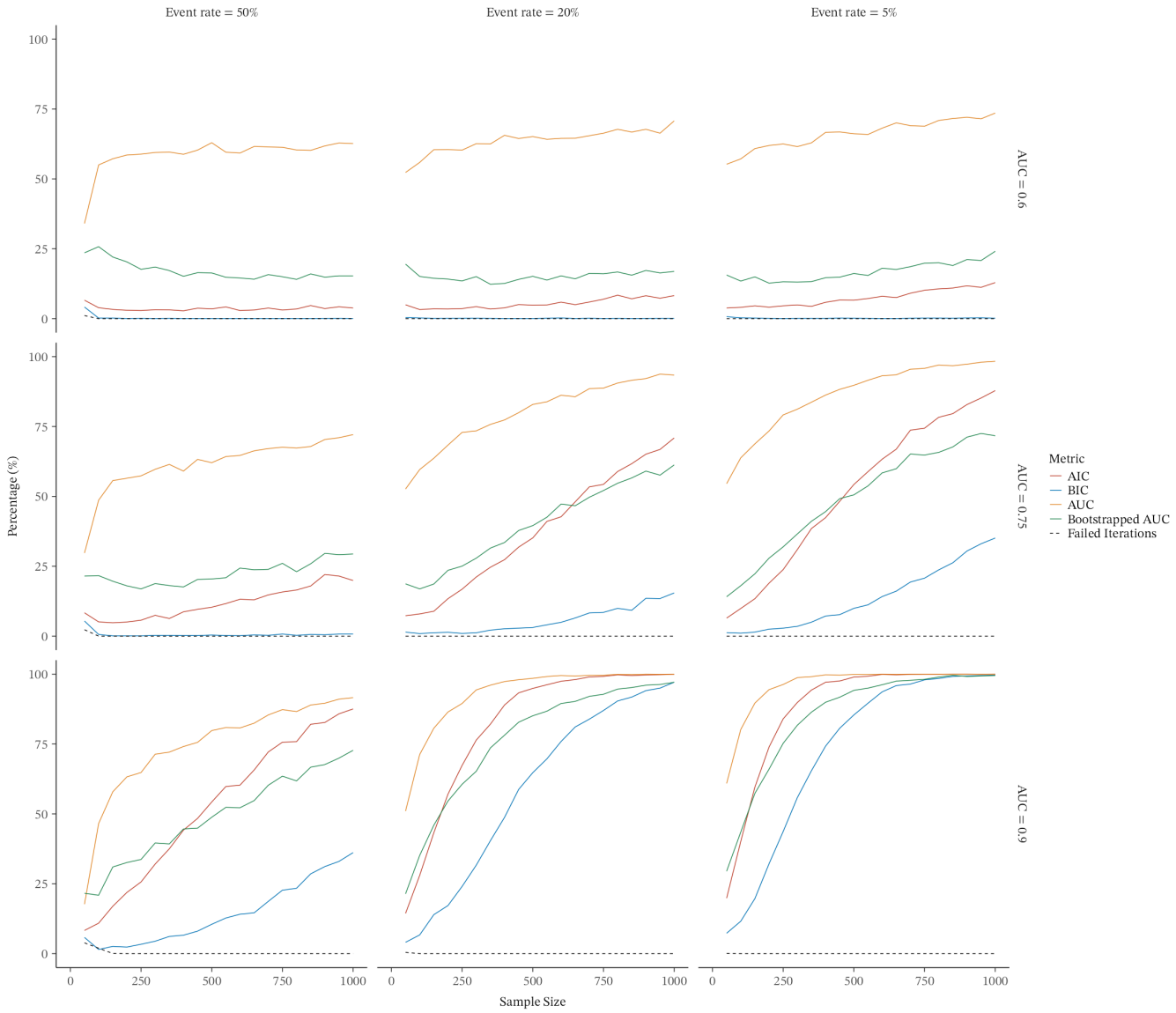

FIGURE A5 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_5)$ model with X_4 and X_5 as the candidate predictors.

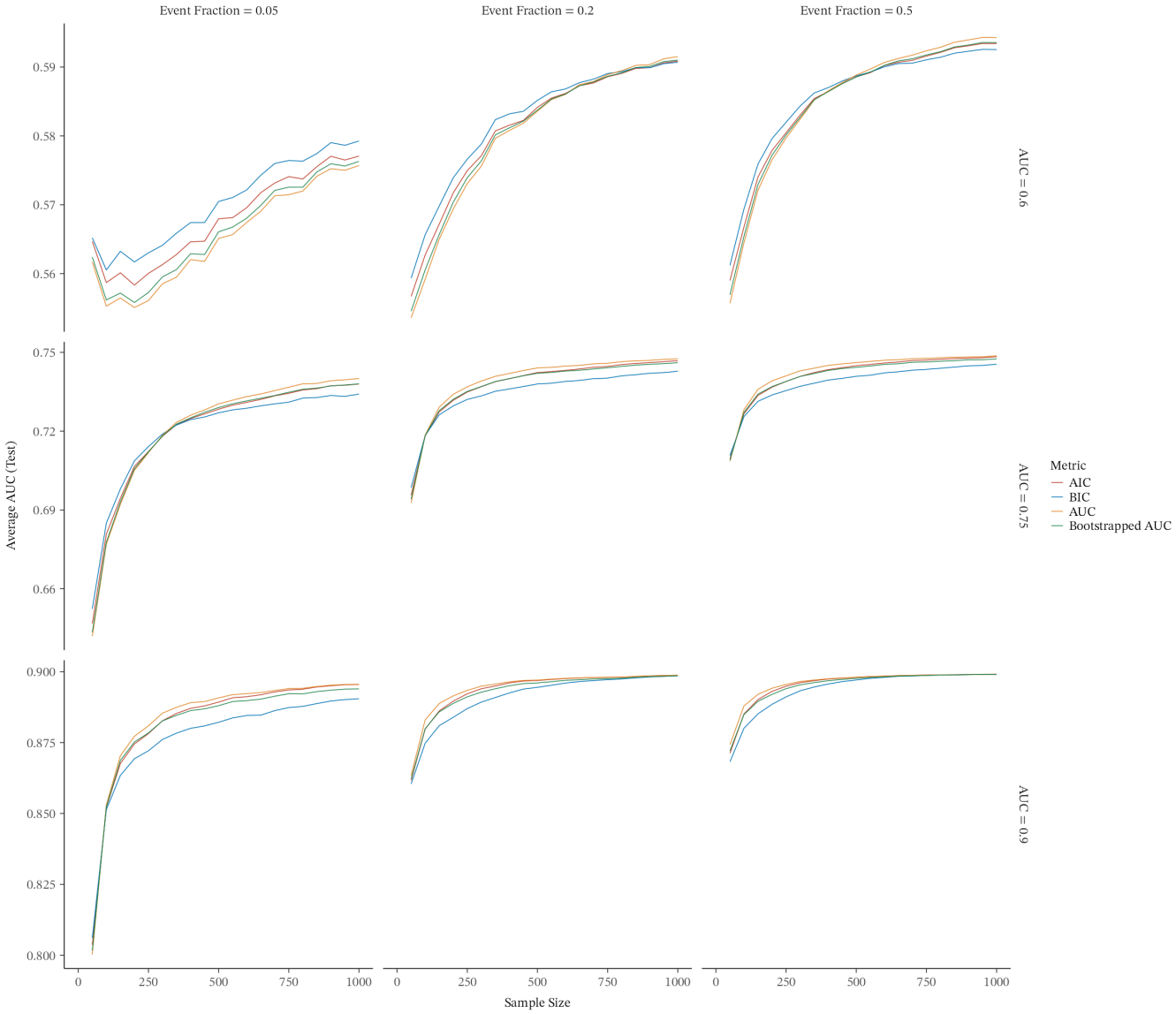

FIGURE A6 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_5)$ model with X_4 and X_5 as the candidate predictors.

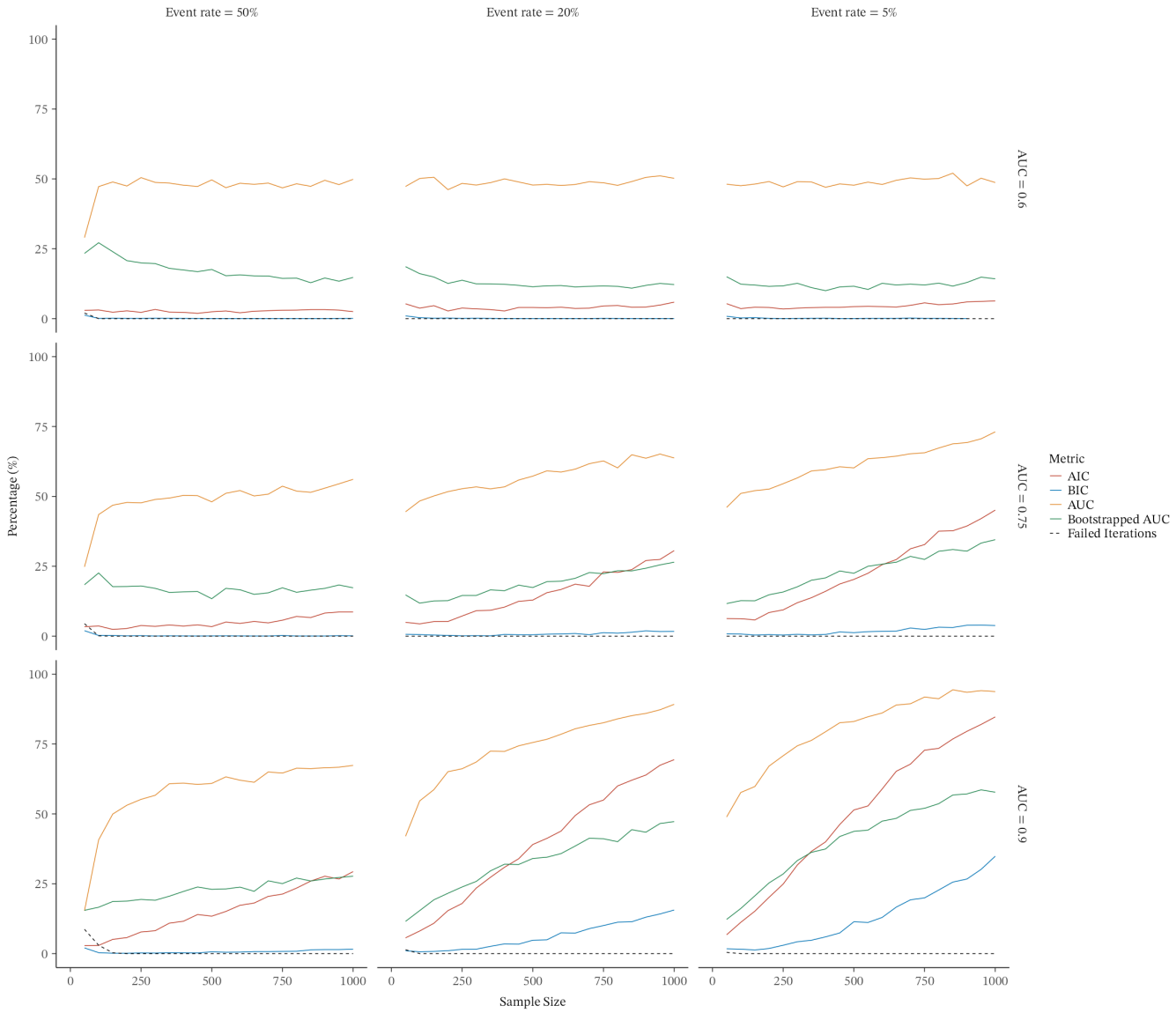

FIGURE A7 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_1X_2 + 0.5 * X_2X_3)$ model with X_1X_2 and X_2X_3 as the candidate predictors.

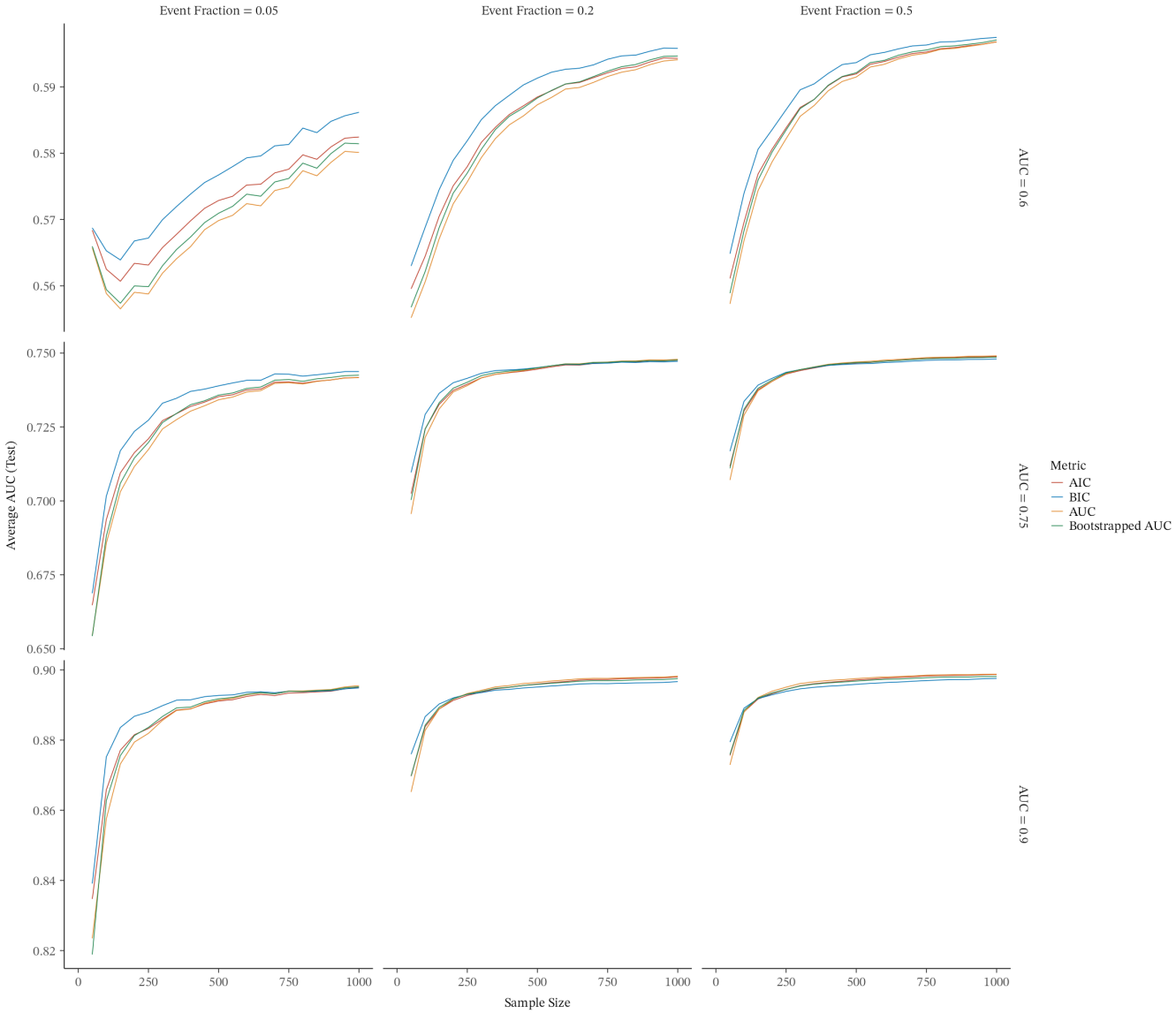

FIGURE A8 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_1X_2 + 0.5 * X_2X_3)$ model with X_1X_2 and X_2X_3 as the candidate predictors.

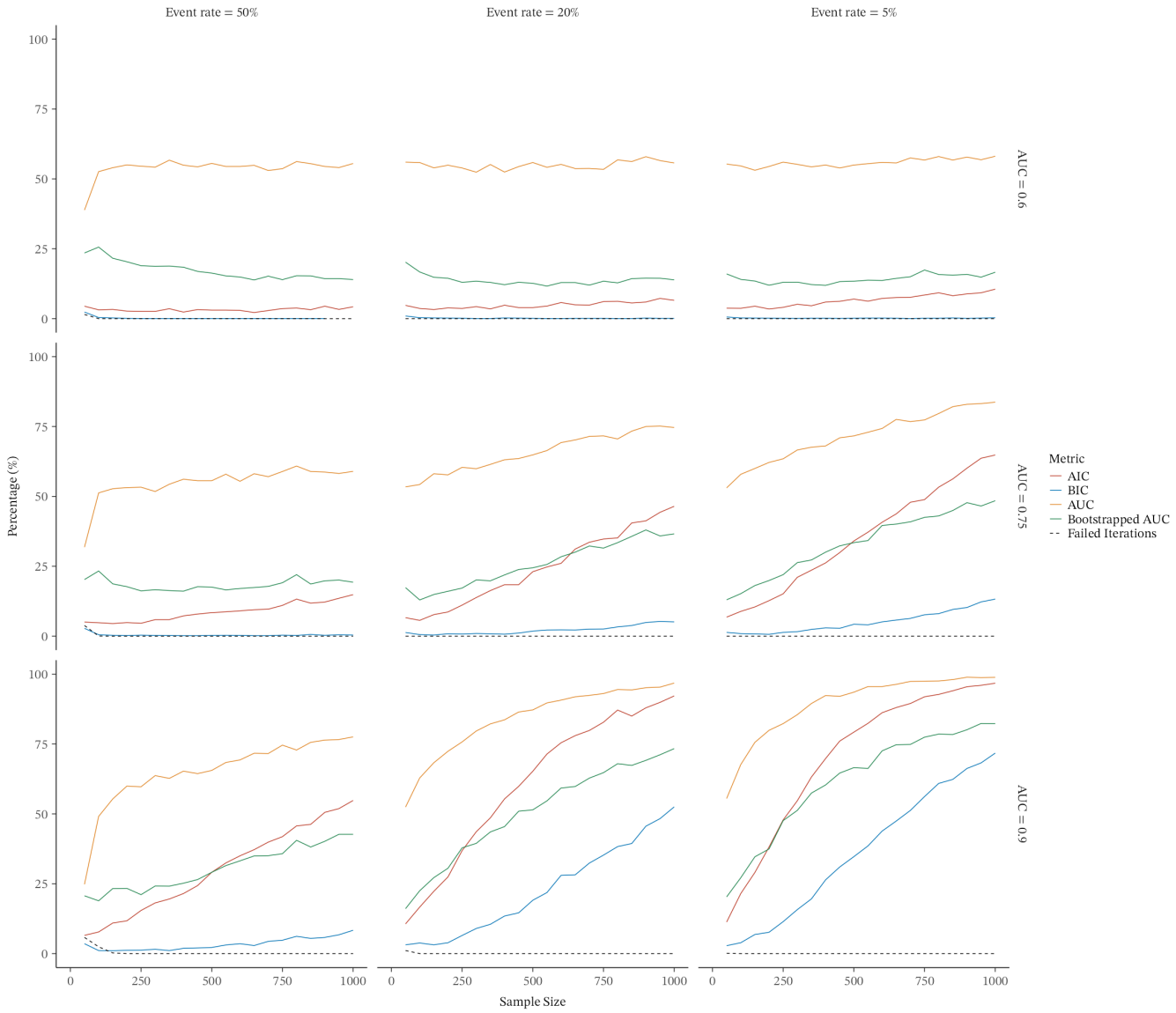

FIGURE A9 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X4 + 0.5 * X_1X_2)$ model with X_4 and X_1X_2 as the candidate predictors.

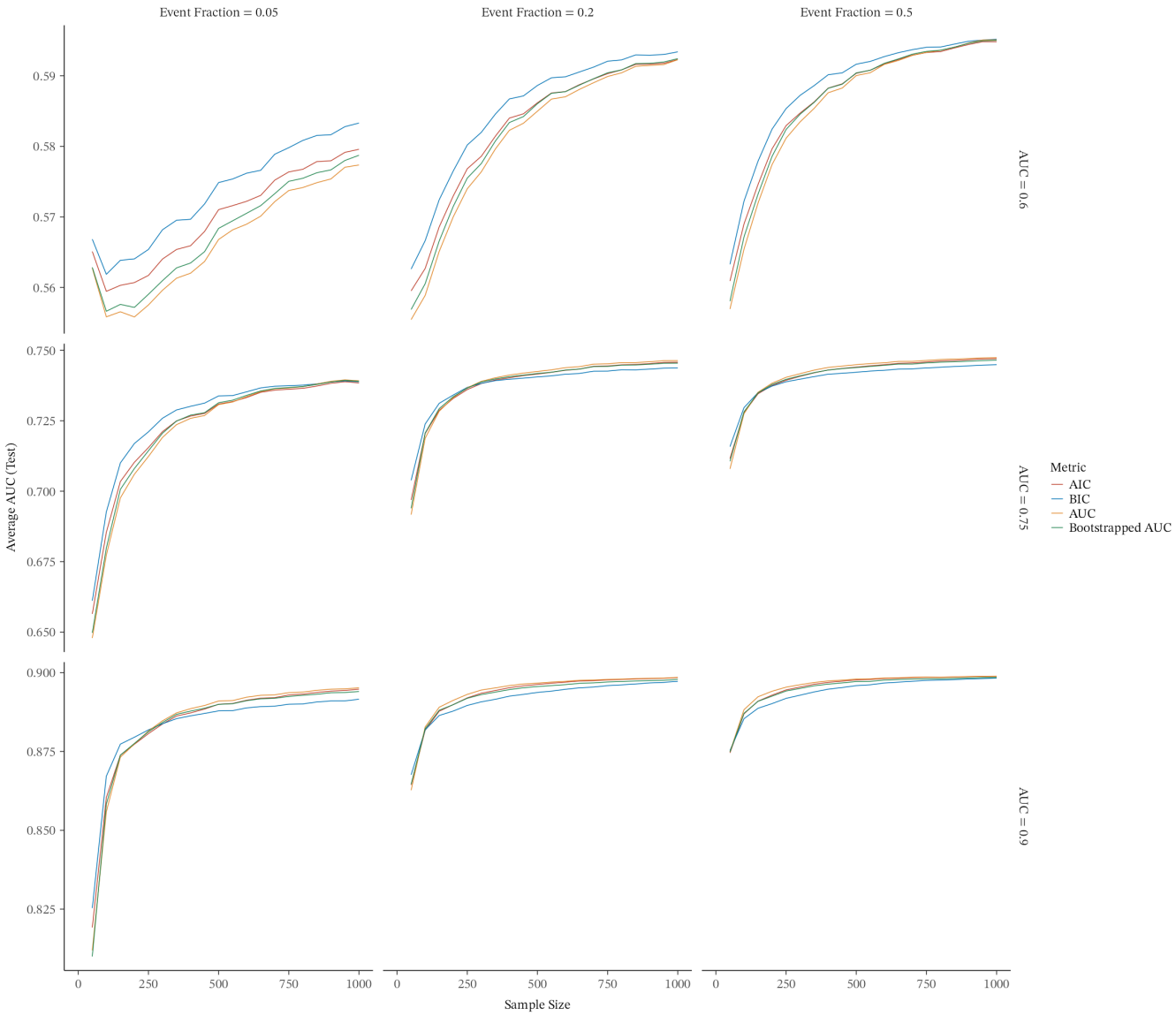

FIGURE A10 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_1X_2)$ model with X_4 and X_1X_2 as the candidate predictors.

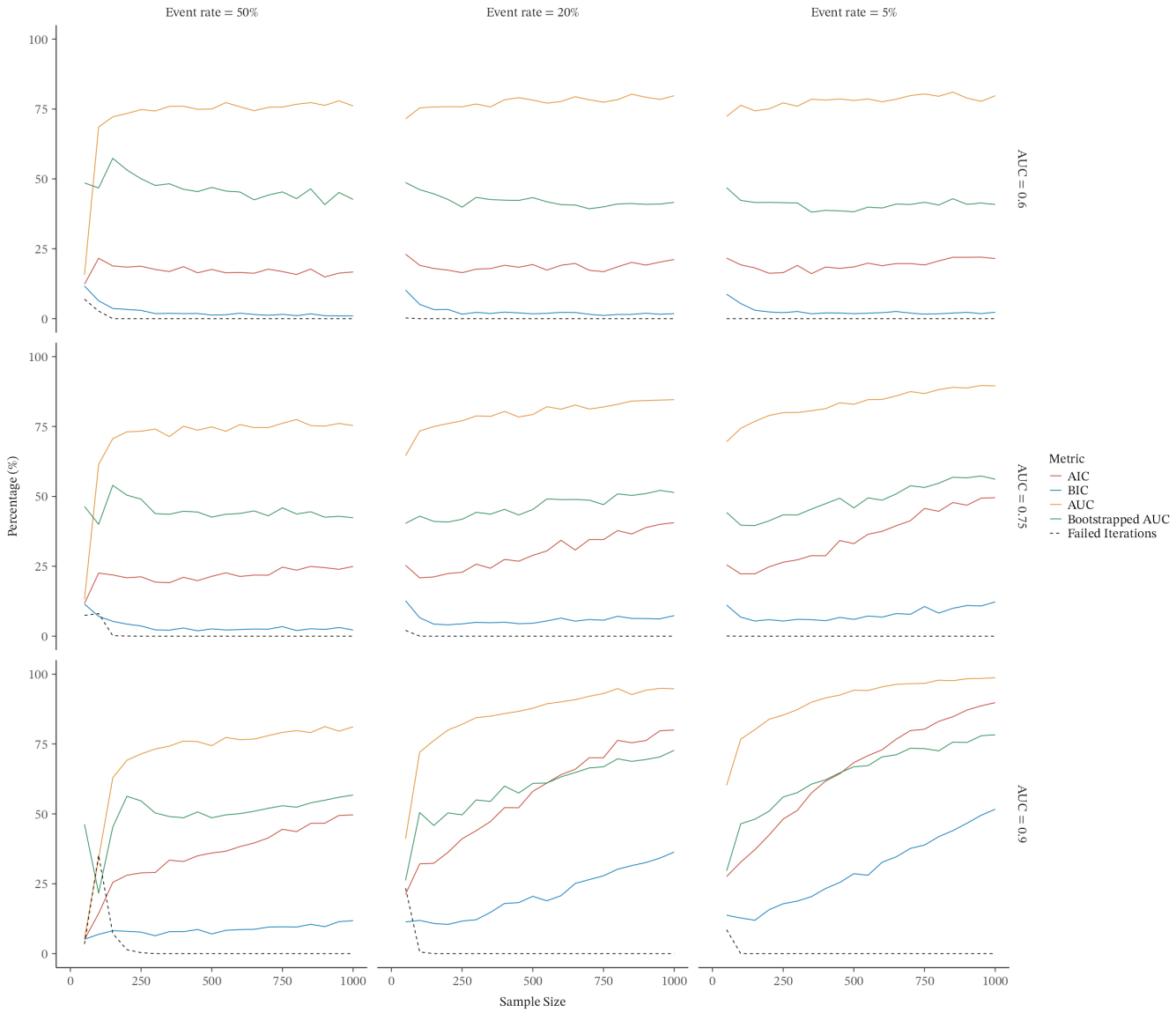

FIGURE A11 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_5 + X_6 ... X_{10})$ model with X_5 as the candidate predictor.

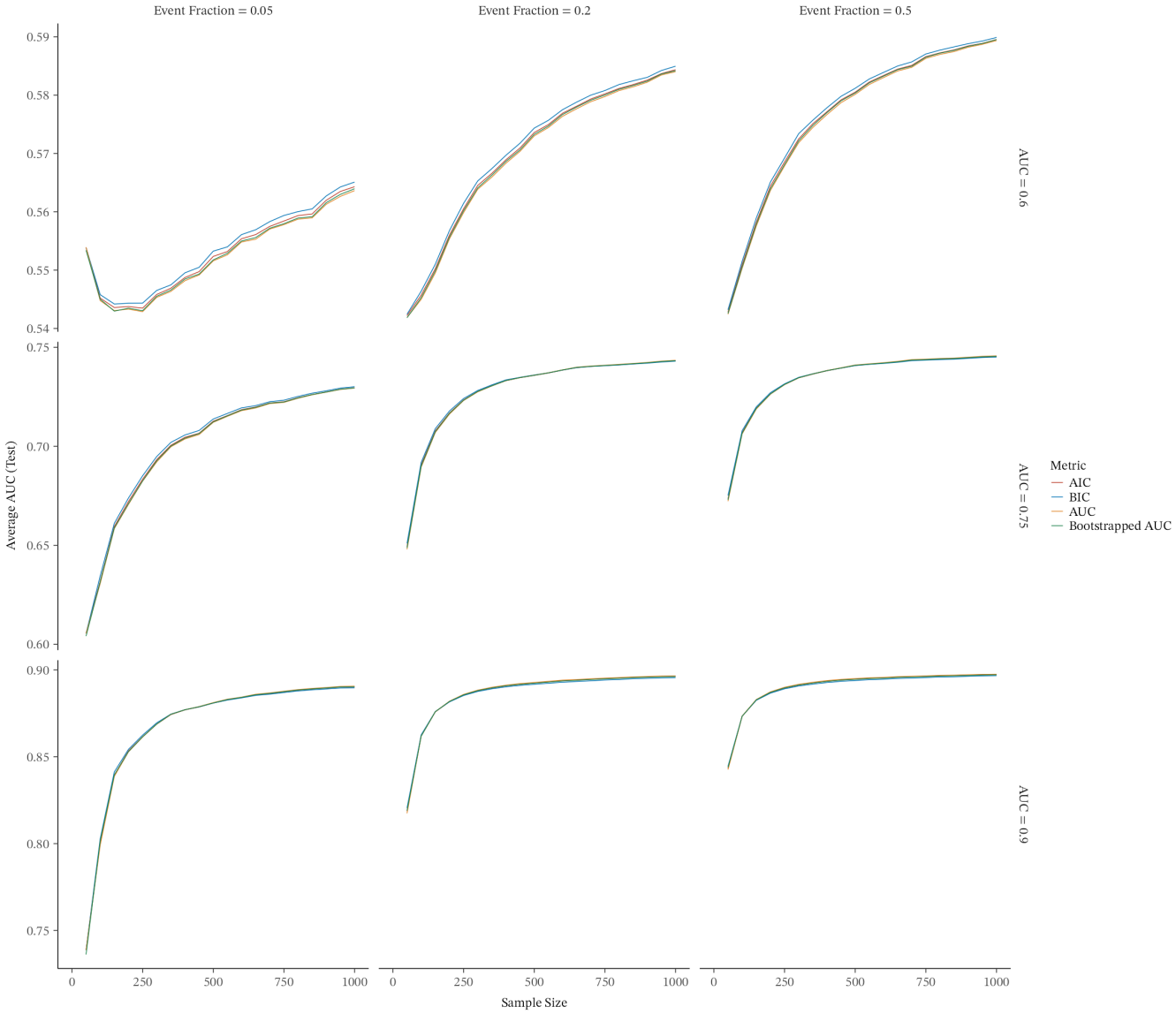

FIGURE A12 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_5 + X_6 ... X_{10})$ model with X_5 as the candidate predictor.

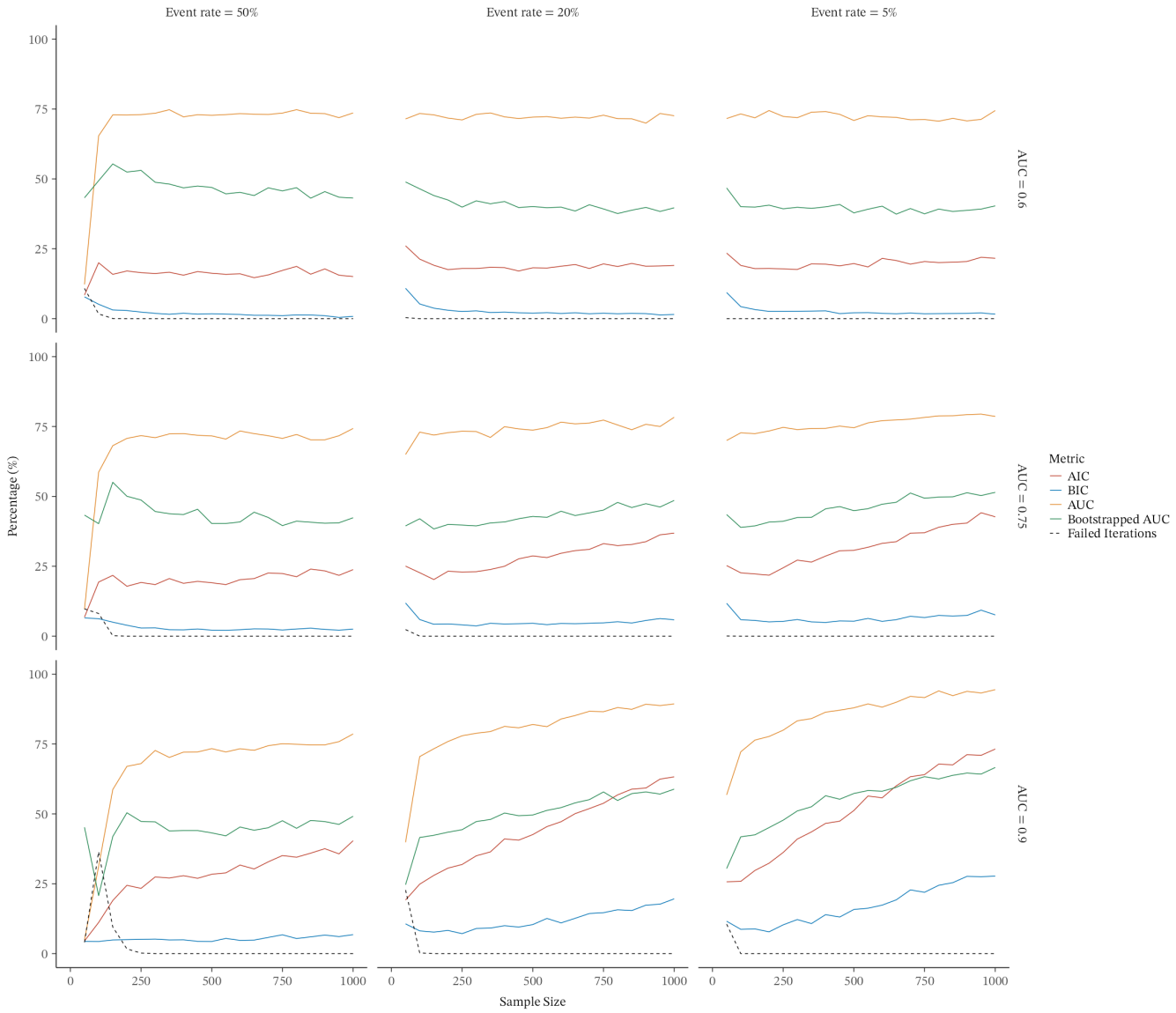

FIGURE A13 Success rate of IC and internal performance measures for the $\beta_0+\beta(X_1+X_2+X_3+X_4+0.5*X_1X_2+X_6...X_{10})$ model with X_1X_2 as the candidate predictor.

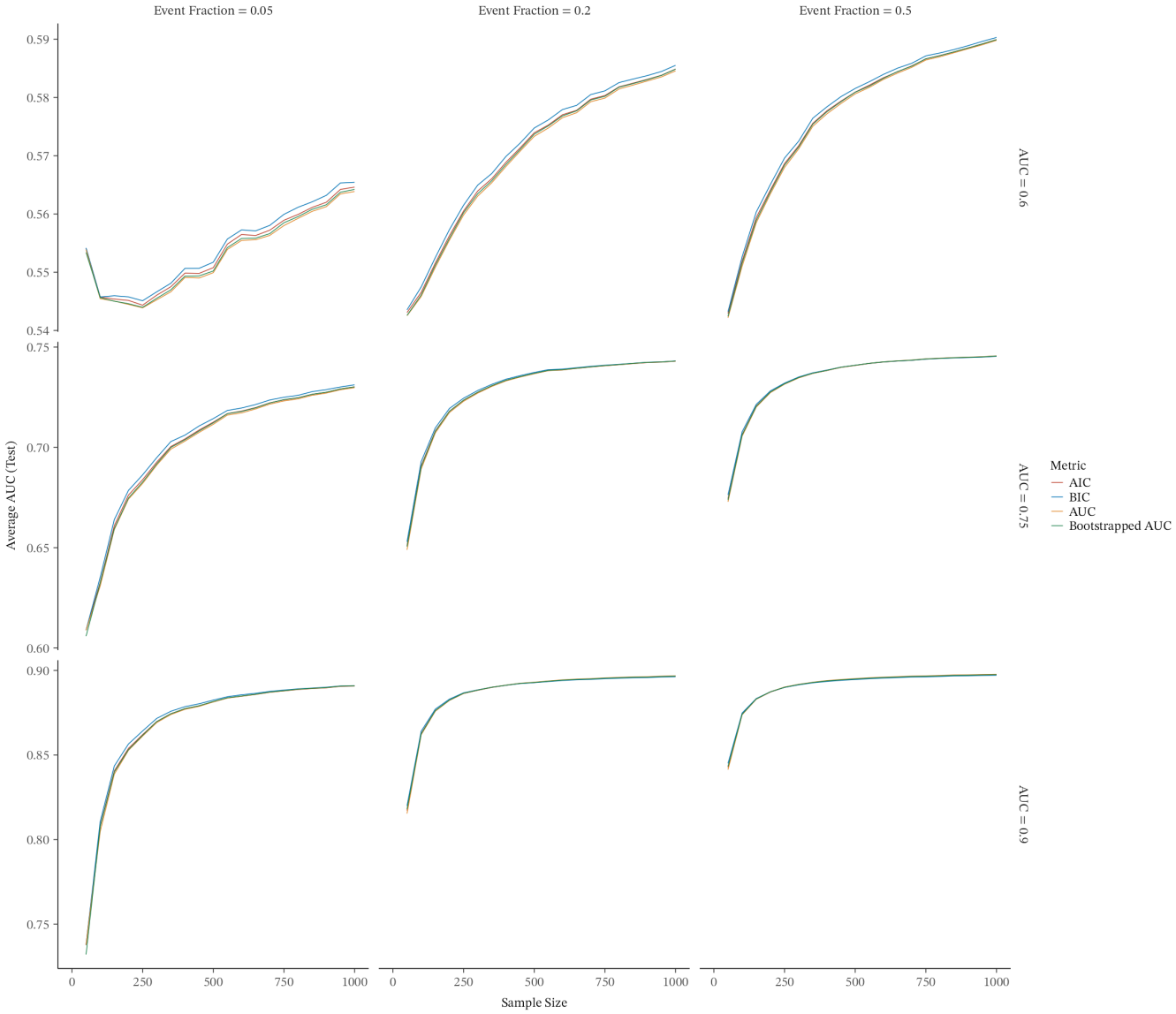

FIGURE A14 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + 0.5 * X_1X_2 + X_6 ... X_{10})$ model with X_1X_2 as the candidate predictor.

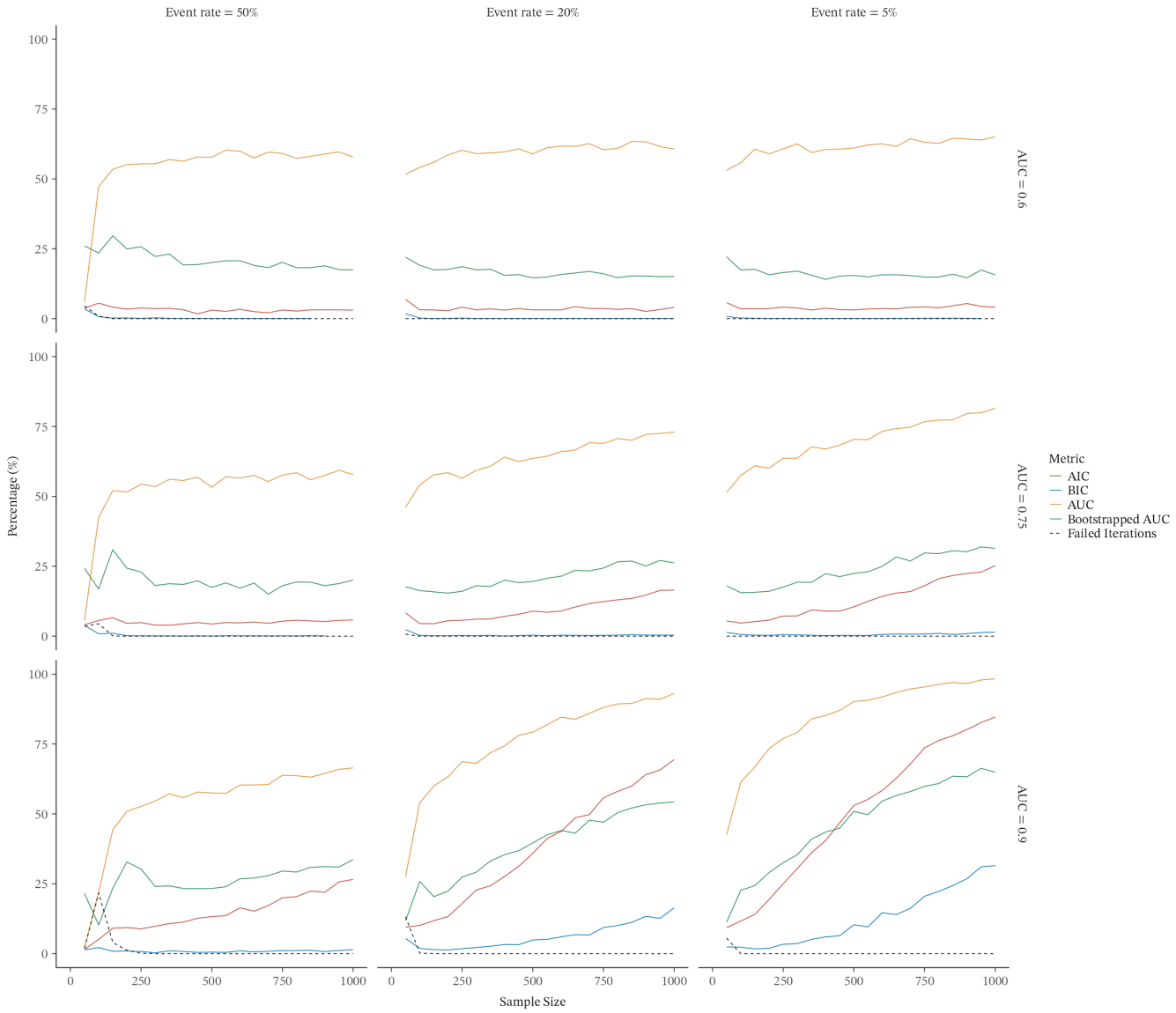

FIGURE A15 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_5 + X_6 ... X_{10})$ model with X_4 and X_5 as the candidate predictors.

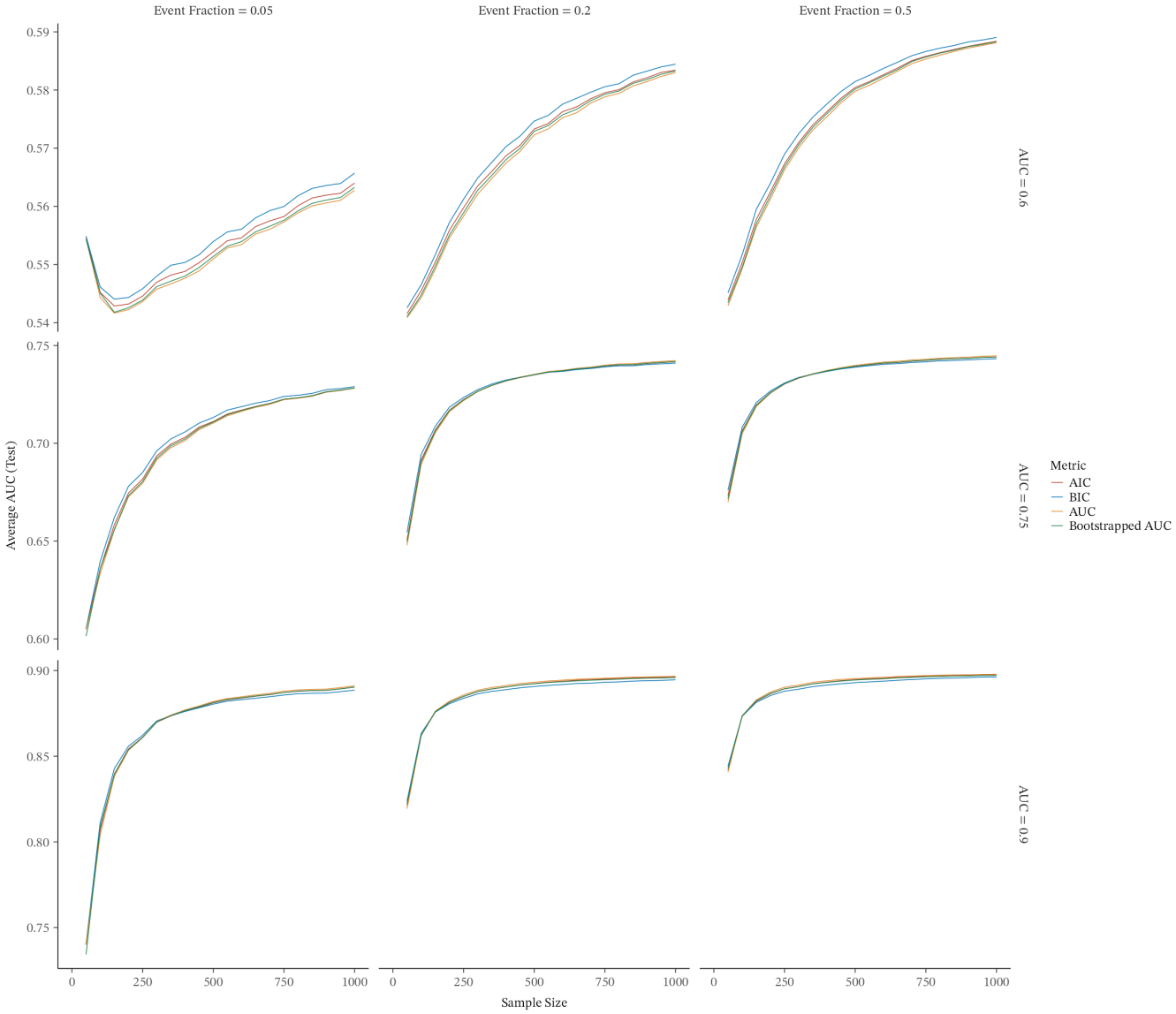

FIGURE A16 Test AUC of IC and internal performance measures for the $\beta_0+\beta(X_1+X_2+X_3+0.5*X_4+0.5*X_5+X_6...X_{10})$ model with X_4 and X_5 as the candidate predictors.

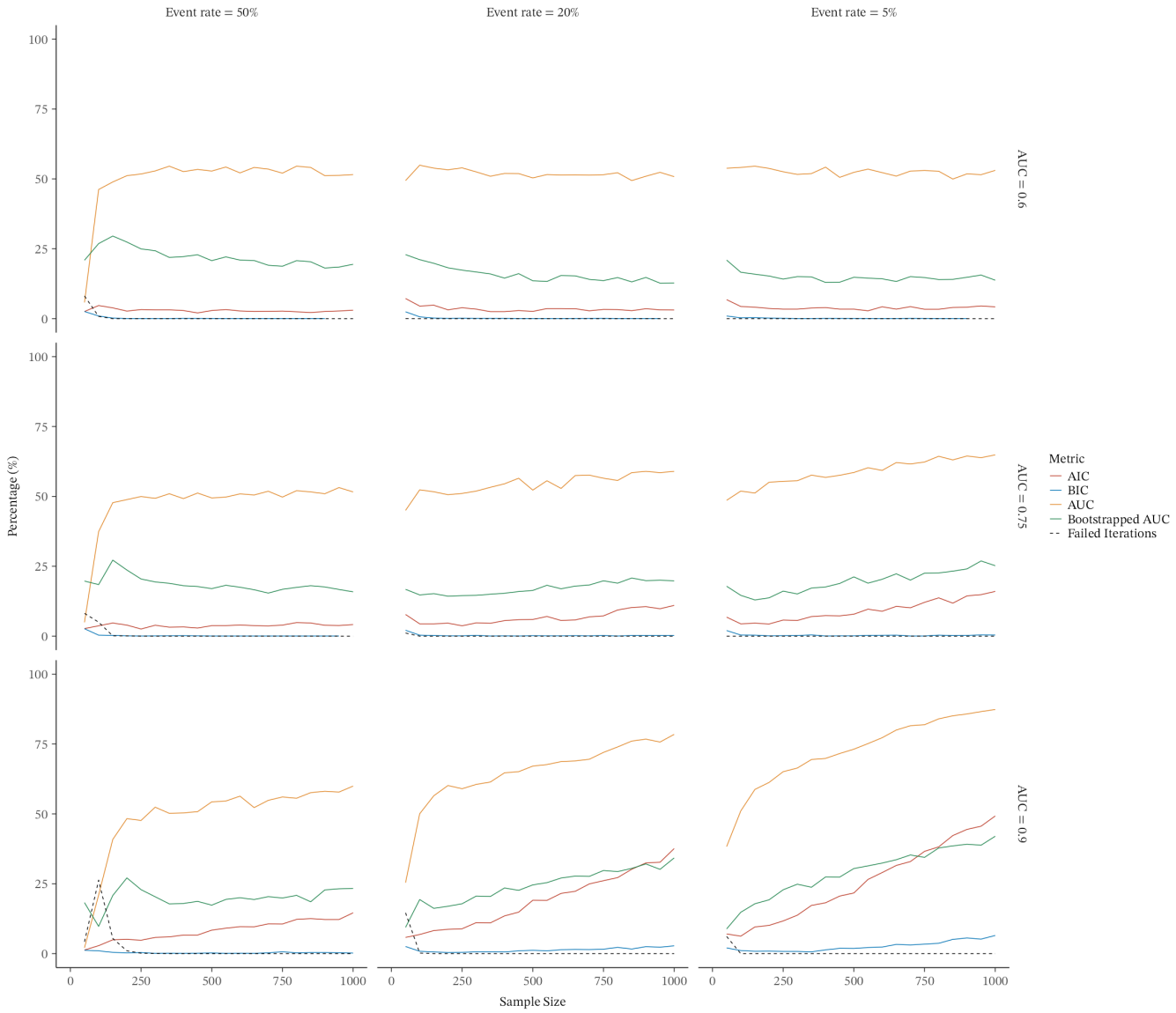

FIGURE A17 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_1X_2 + 0.5 * X_2X_3 + X_6...X_{10})$ model with X_1X_2 and X_2X_3 as the candidate predictors.

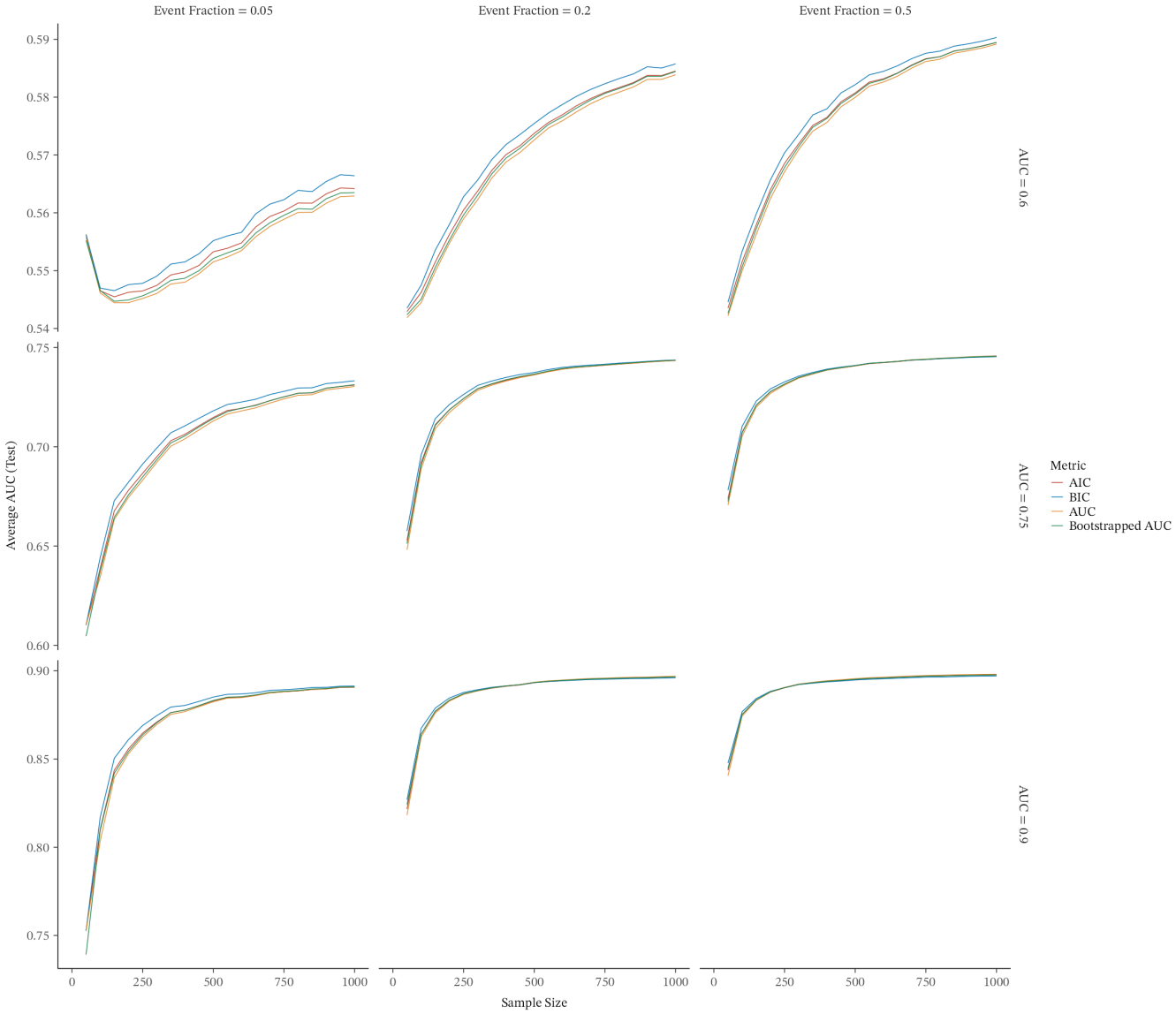

FIGURE A18 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_1X_2 + 0.5 * X_2X_3 + X_6...X_{10})$ model with X_1X_2 and X_2X_3 as the candidate predictors.

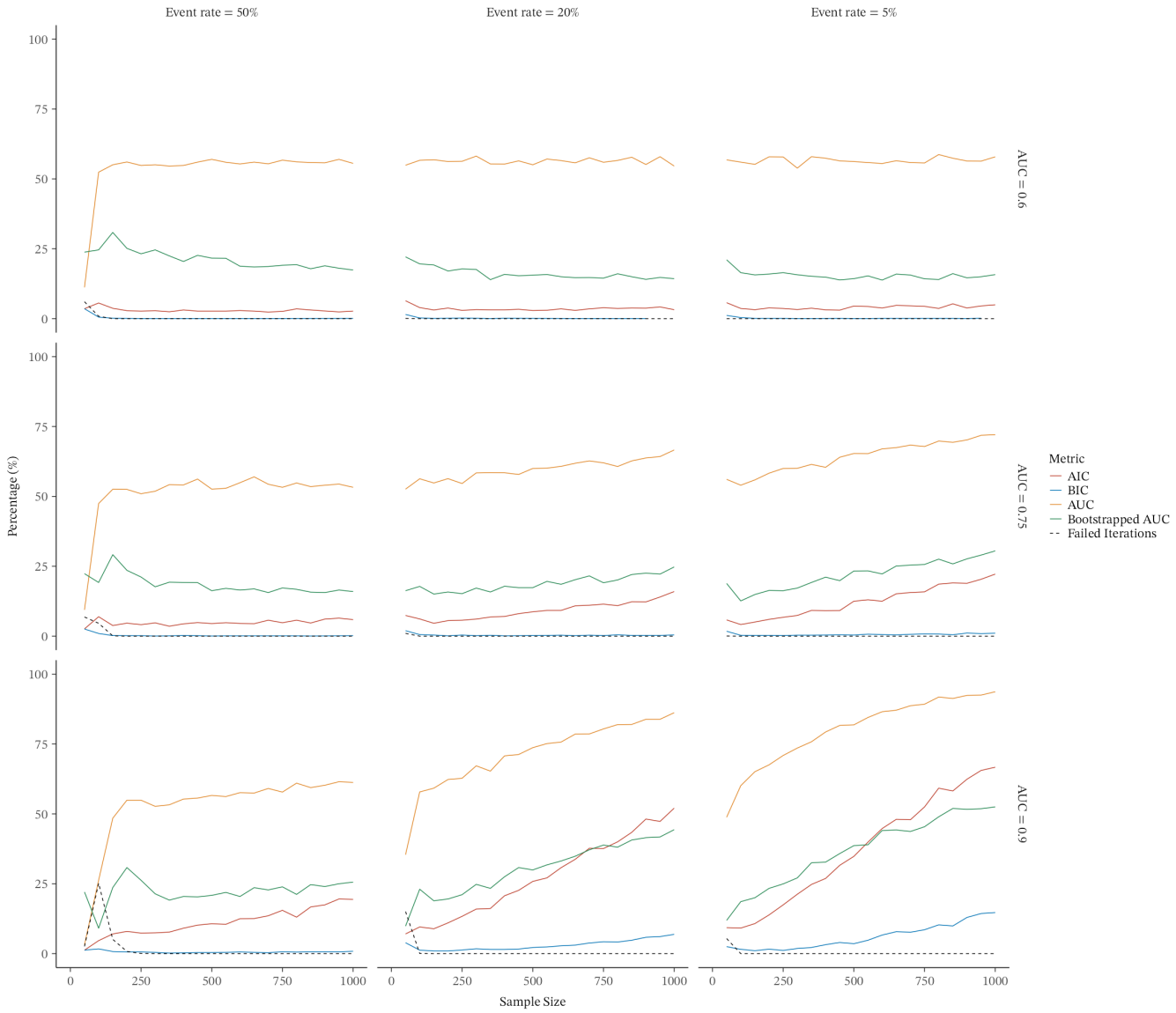

FIGURE A19 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_1X_2 + X_6...X_{10})$ model with X_4 and X_1X_2 as the candidate predictors.

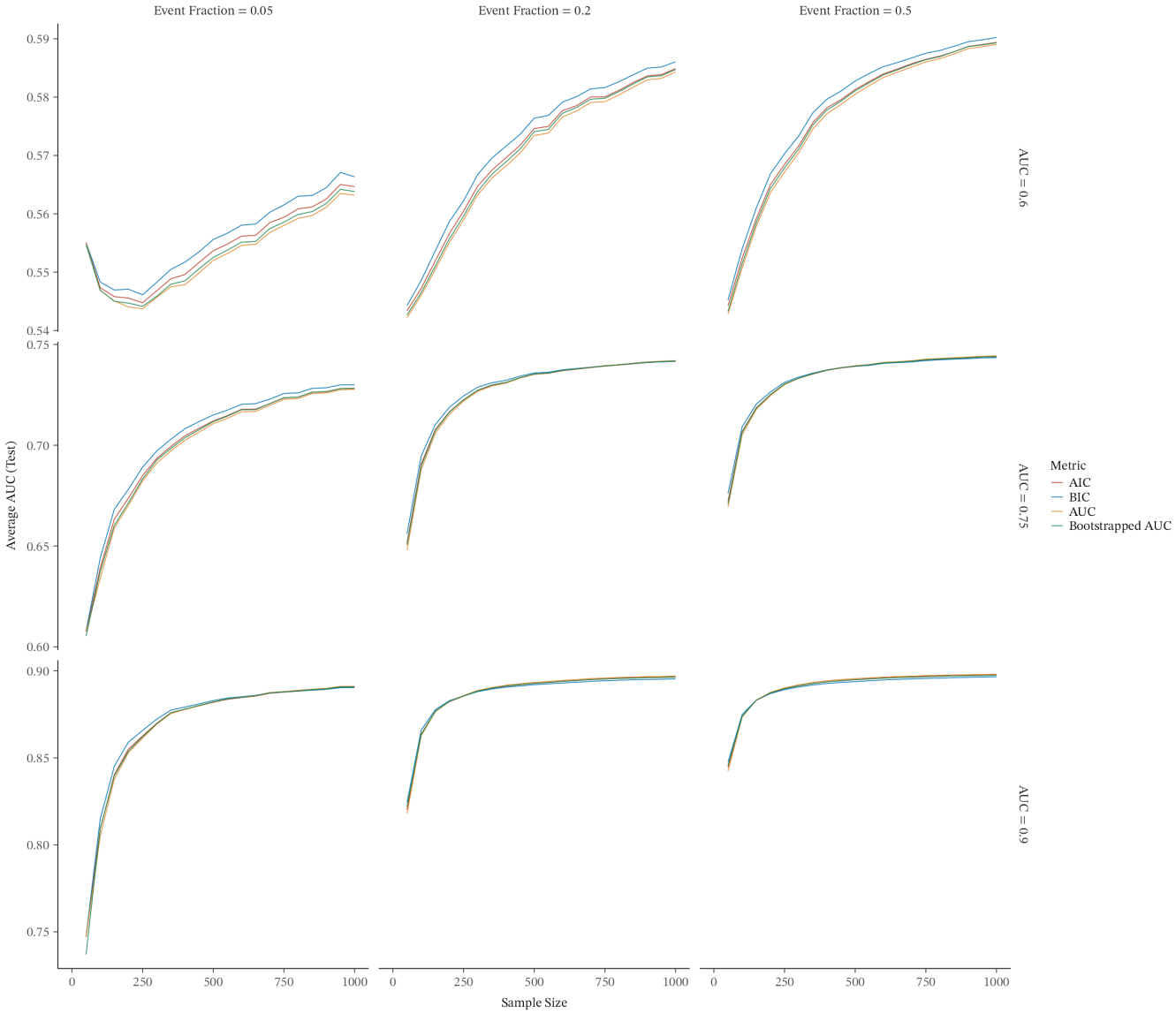

FIGURE A20 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + 0.5 * X_4 + 0.5 * X_1X_2 + X_6...X_{10})$ model with X_4 and X_1X_2 as the candidate predictors.

B SCENARIO 2: EXCLUDING PREDICTORS

In the following figures we show all the results of the second scenario, where we want to correctly exclude the candidate predictors in the model. The models that create the data are shown in the method section. These figures show the success rate of selecting the correct model for the IC and internal performance measures. The figures also show the test AUC of the IC and internal performance measures for the different models. The test AUC is calculated as the AUC of the model selected by the IC and internal performance measures when validated on the test set.

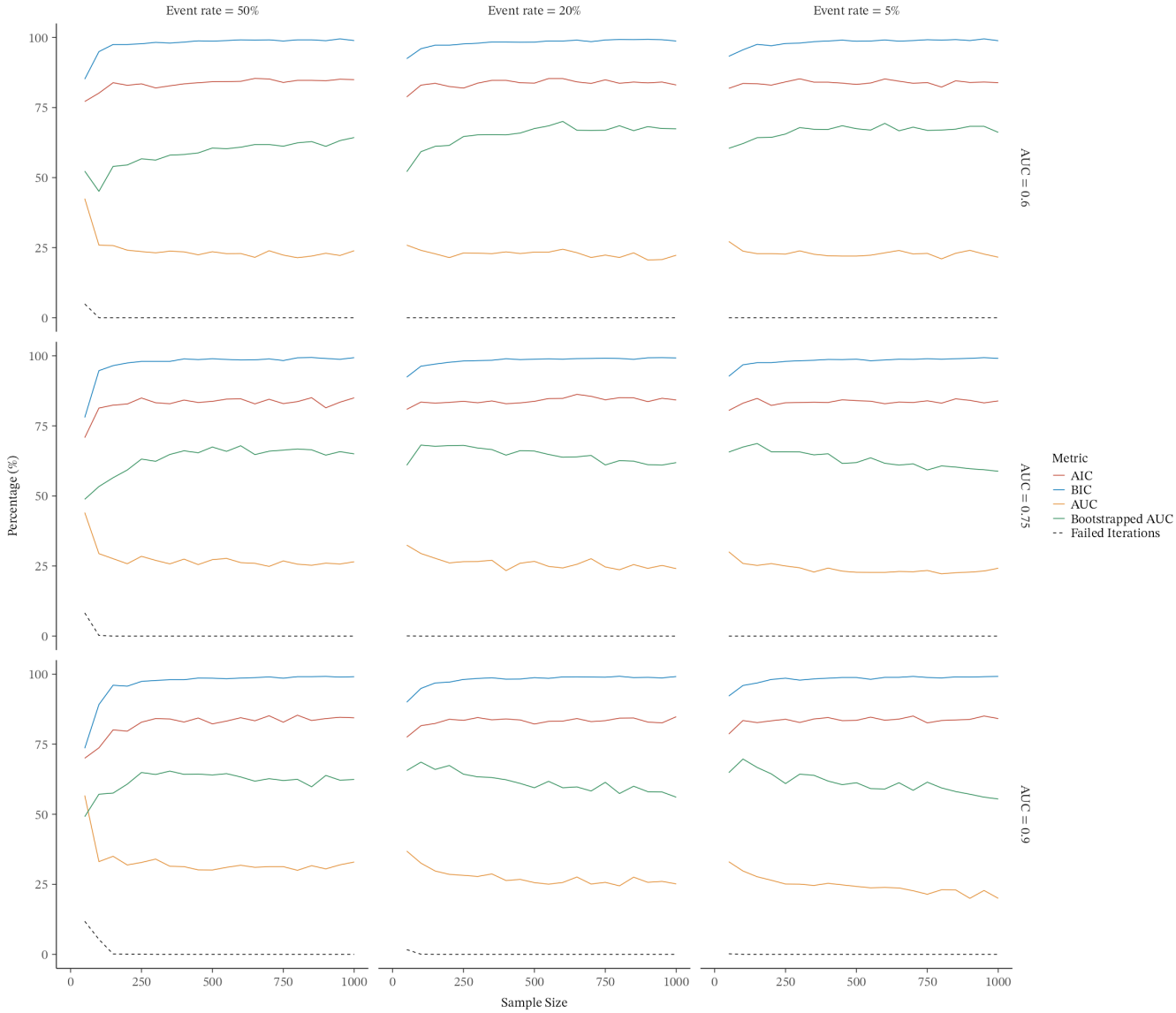

FIGURE B21 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4)$ model with X_5 as the candidate predictor.

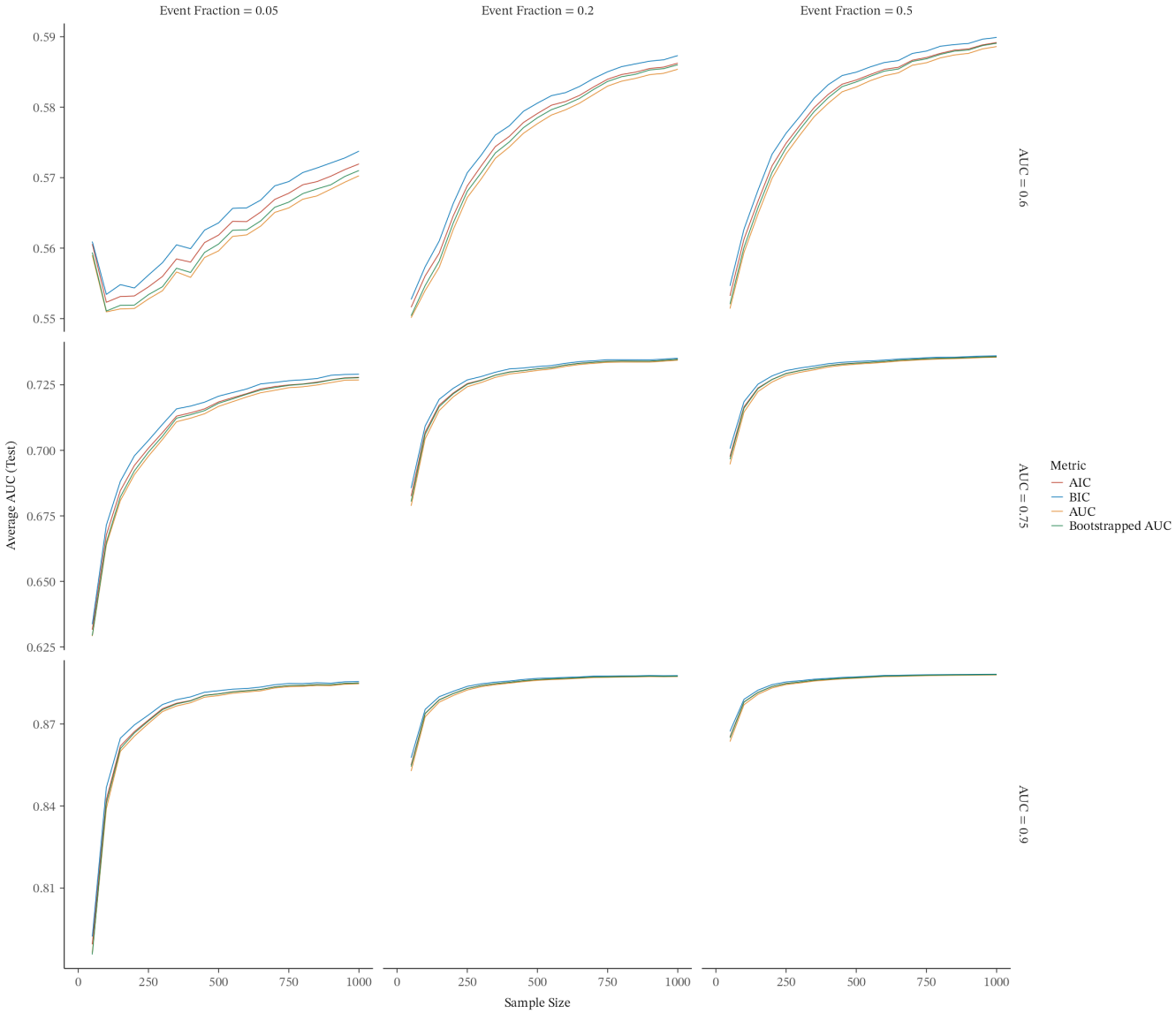

FIGURE B22 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4)$ model with X_5 as the candidate predictor.

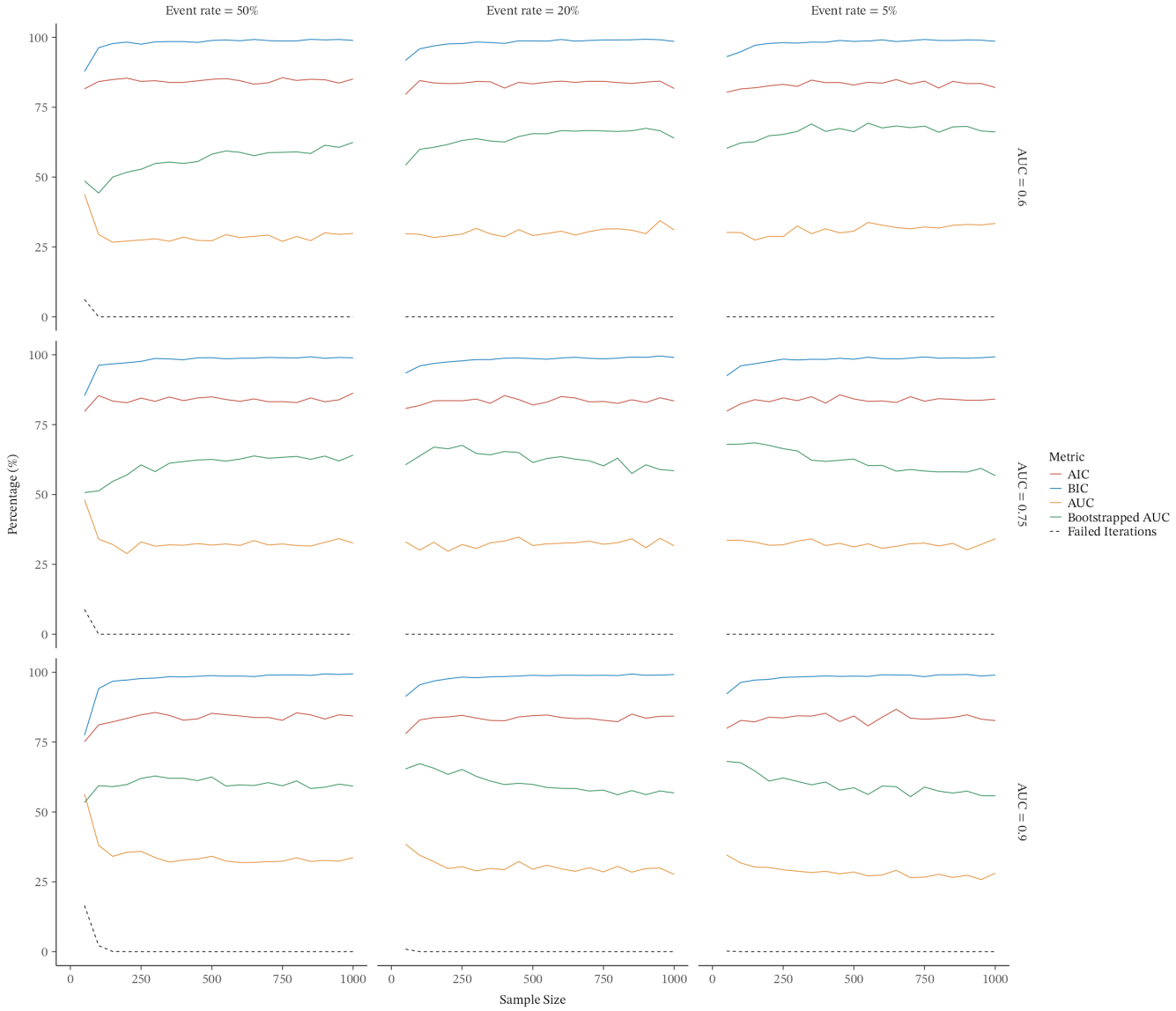

FIGURE B23 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4)$ model with X_1X_2 as the candidate predictor.

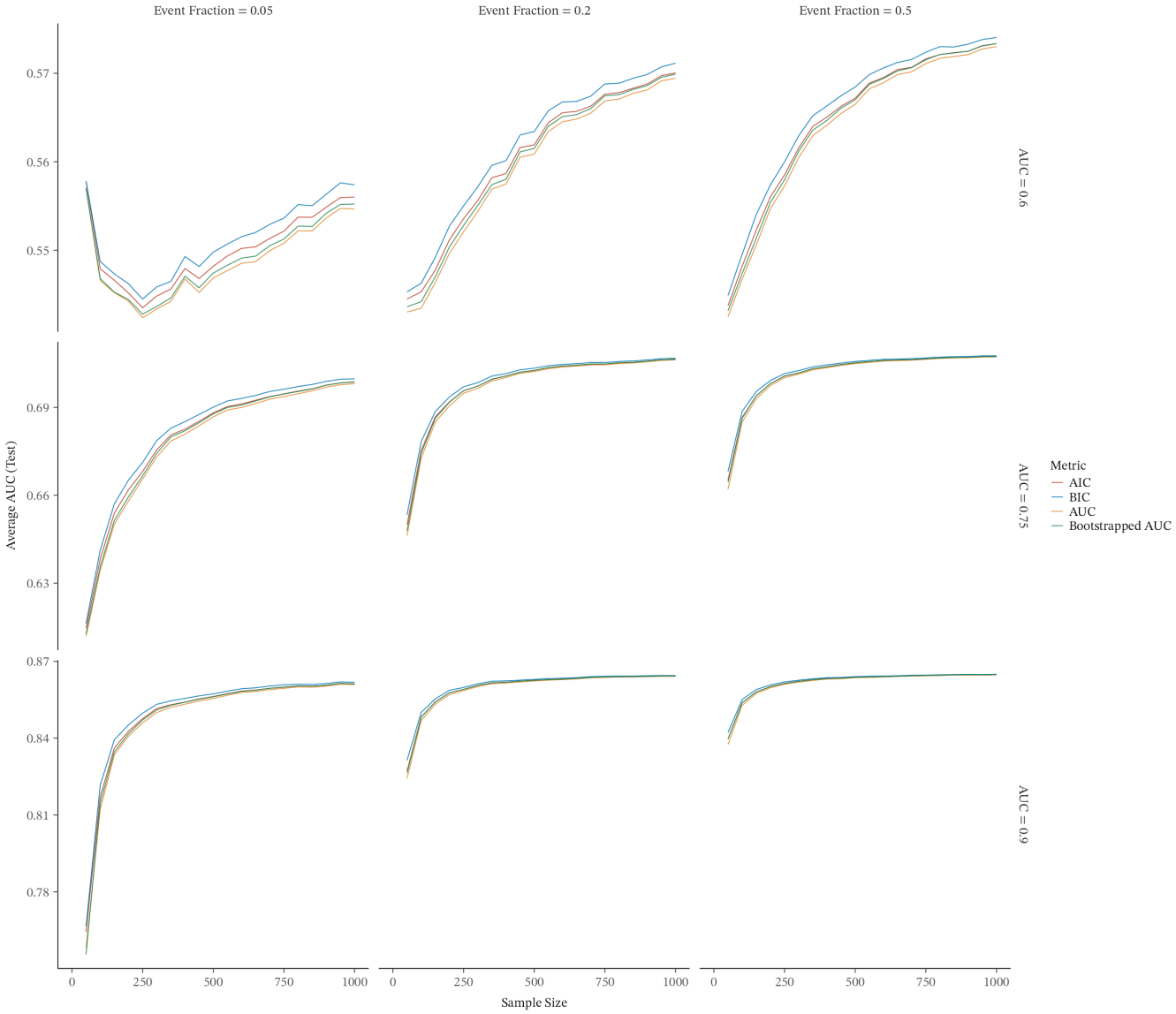

FIGURE B24 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4)$ model with X_1X_2 as the candidate predictor.

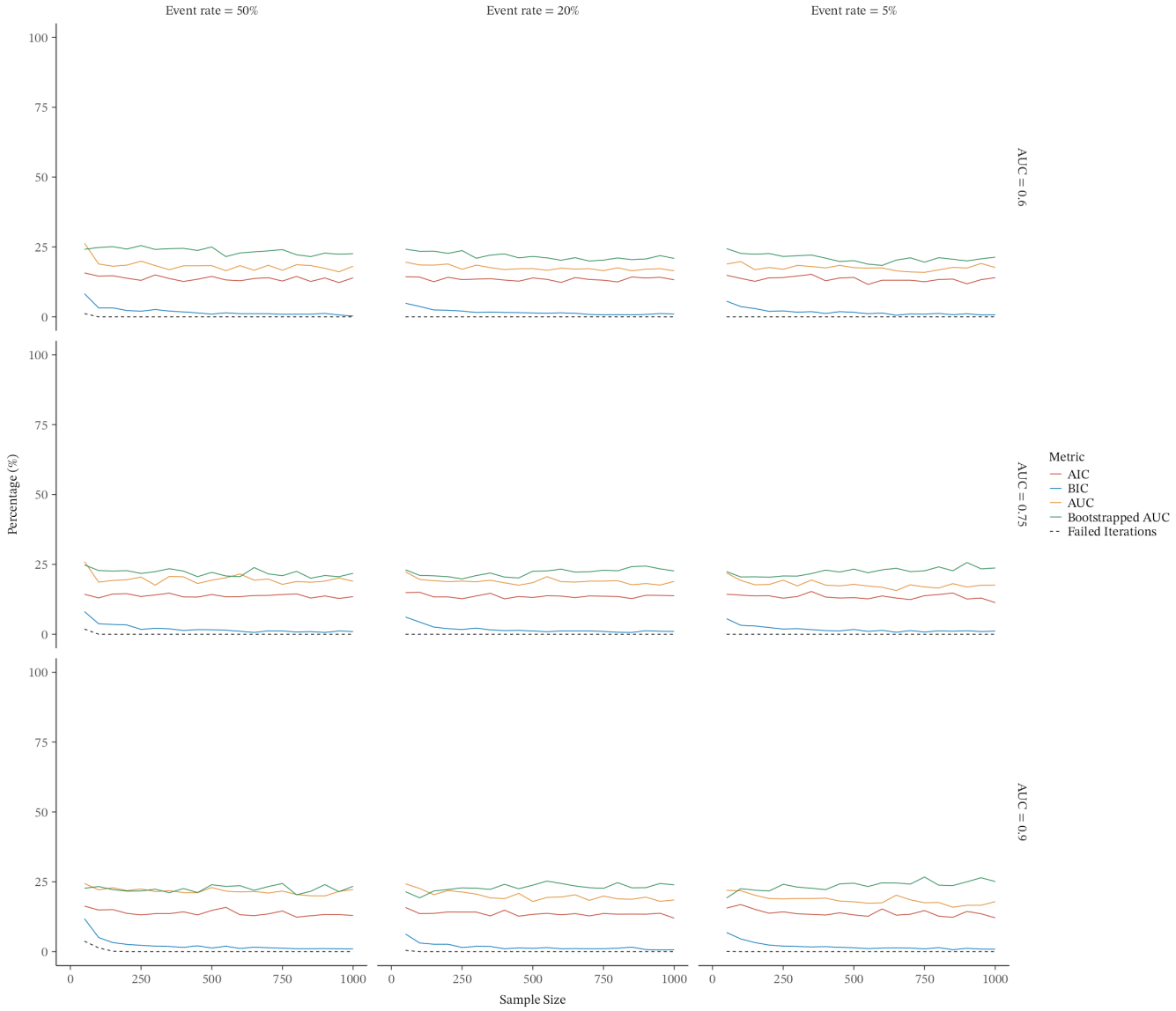

FIGURE B25 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3)$ model with X_4 and X_5 as the candidate predictors.

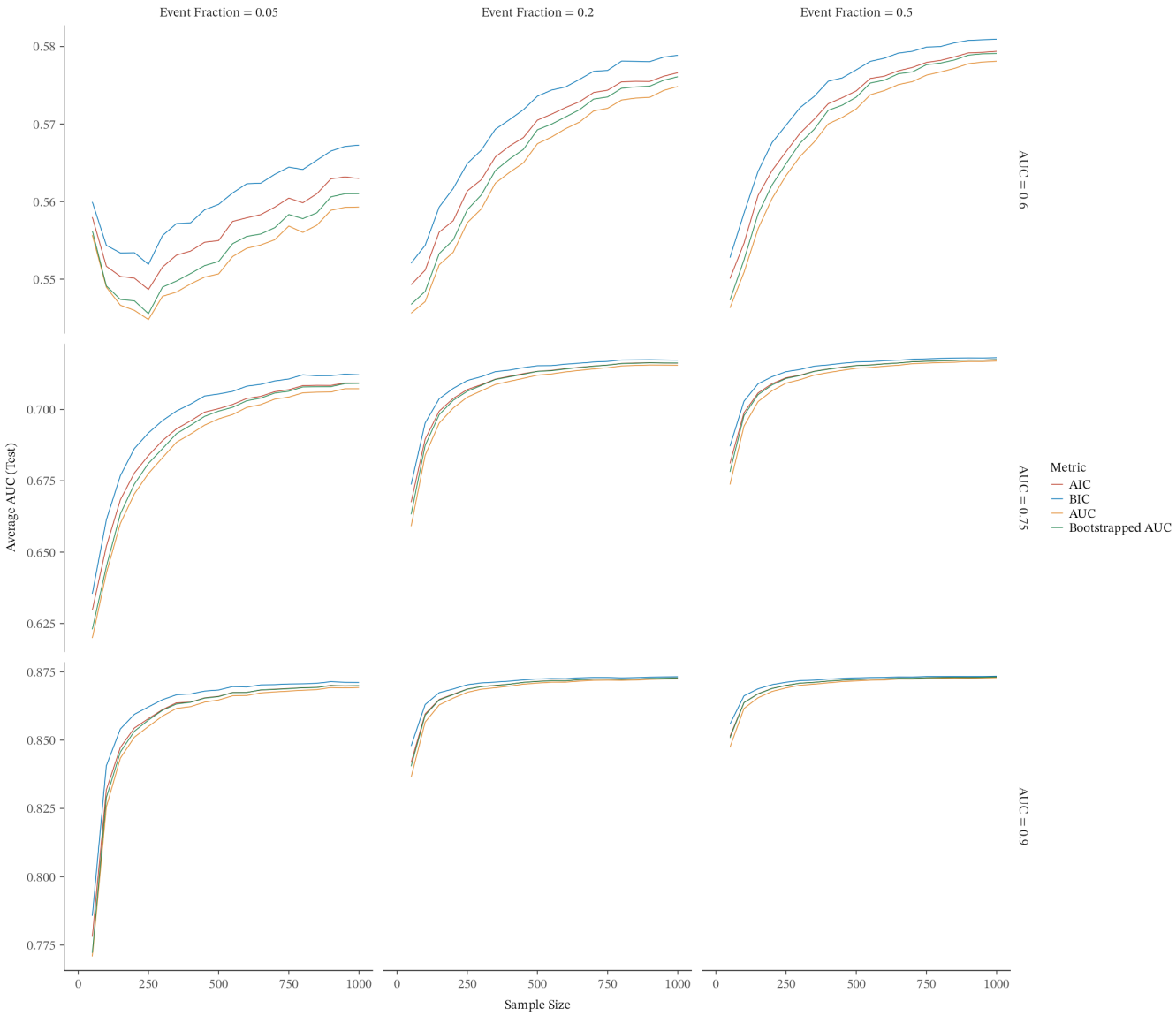

FIGURE B26 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3)$ model with X_4 and X_5 as the candidate predictors.

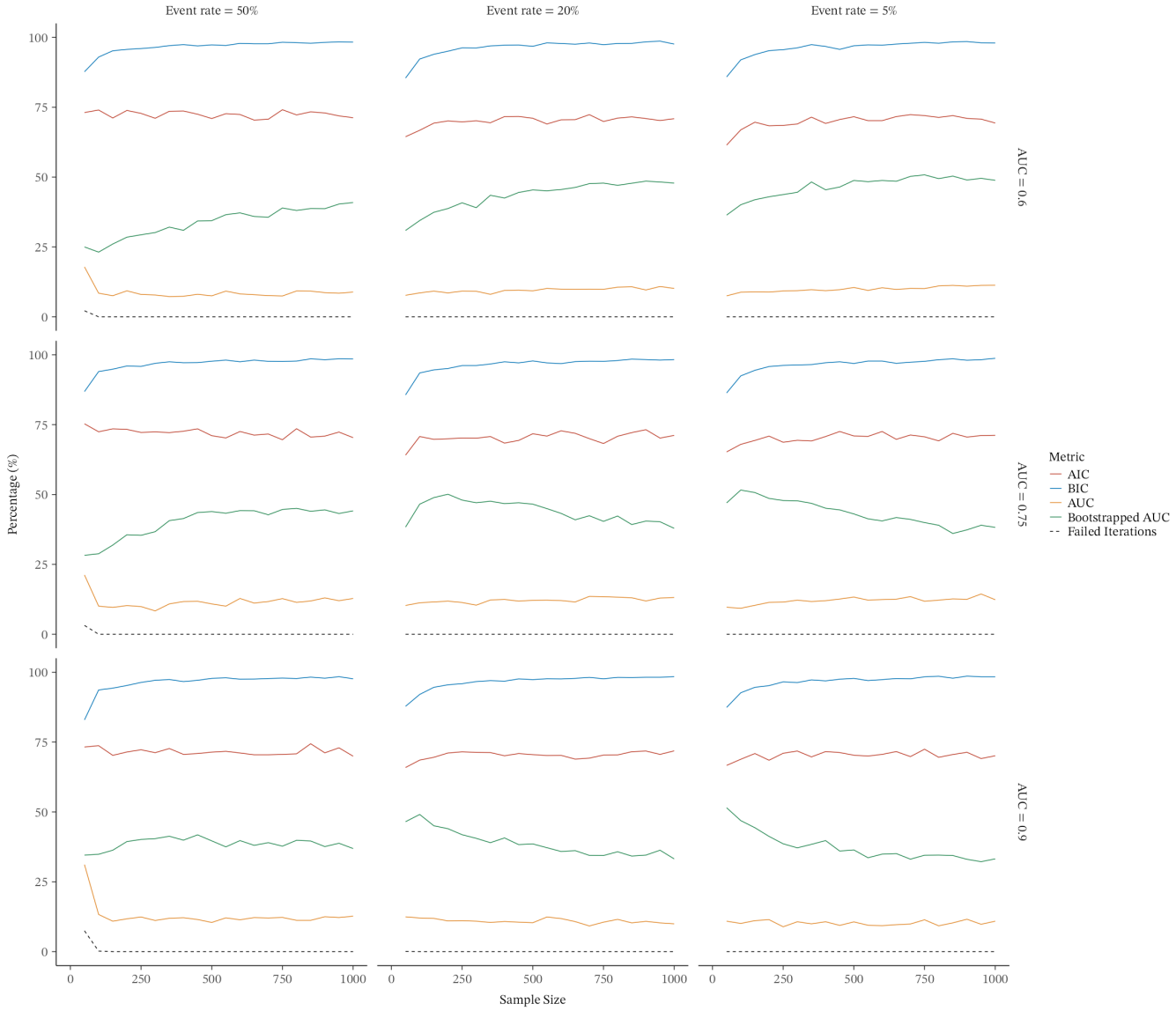

FIGURE B27 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3)$ model with X_1X_2 and X_2X_3 as the candidate predictors.

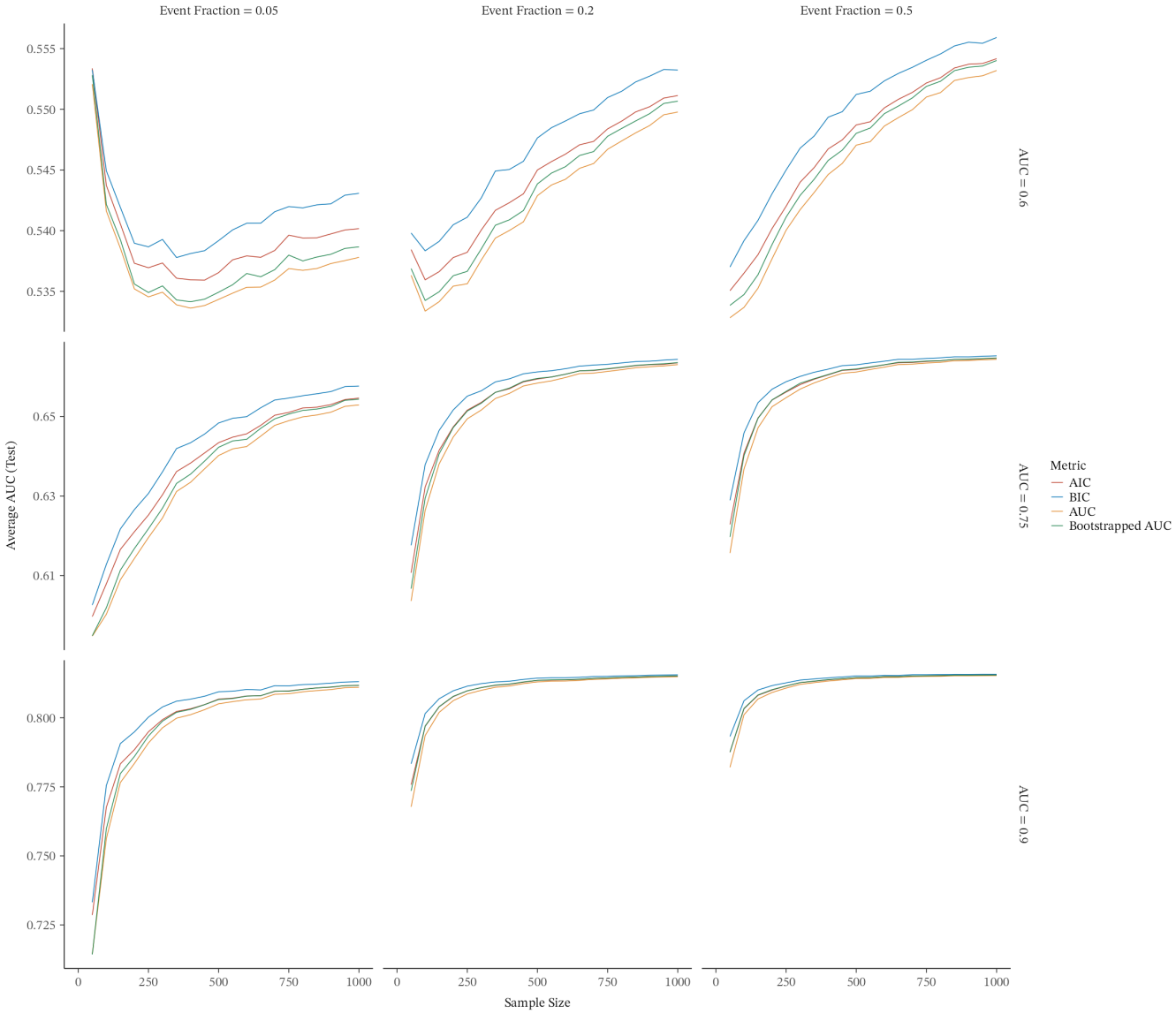

FIGURE B28 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3)$ model with X_1X_2 and X_2X_3 as the candidate predictors.

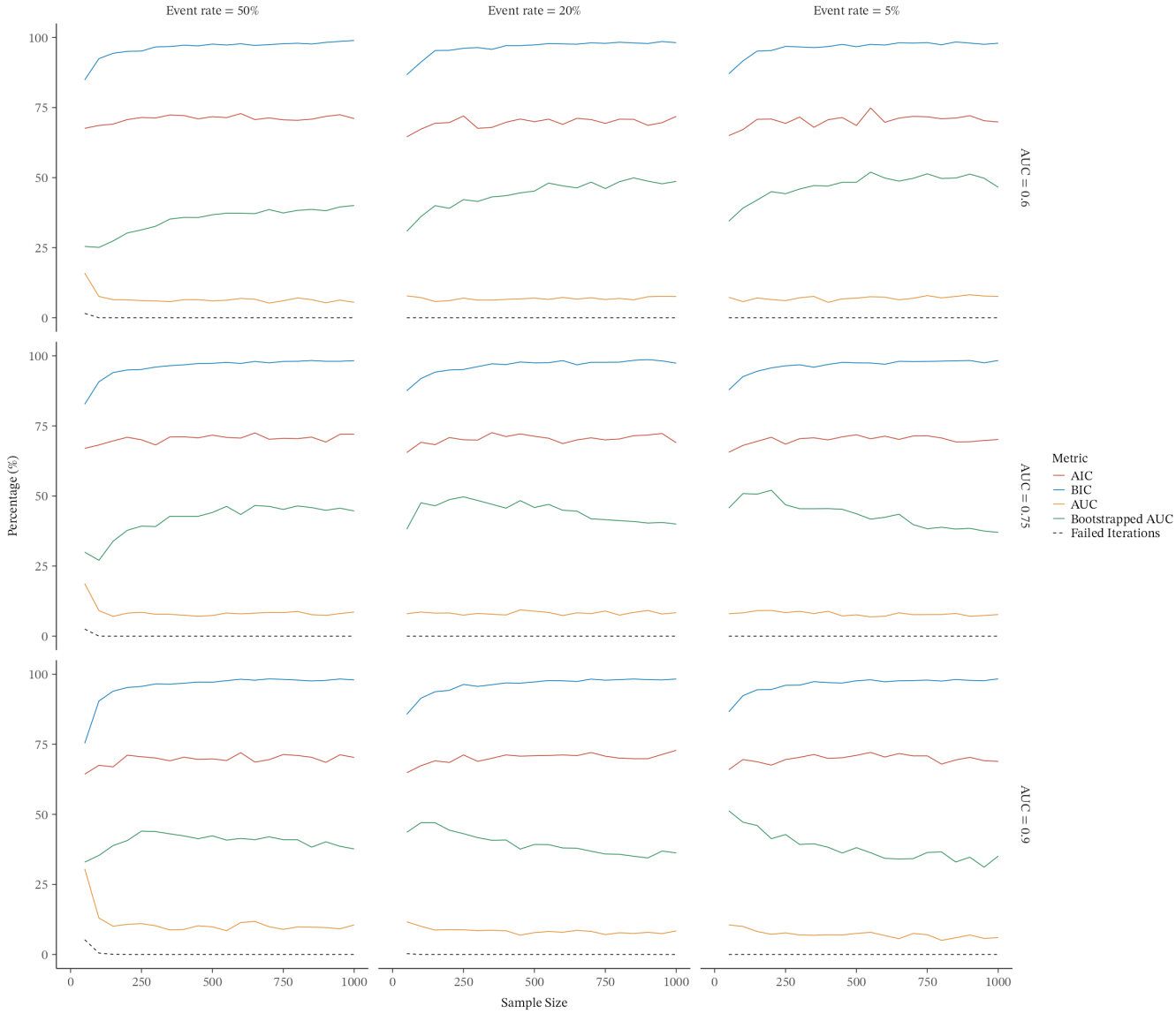

FIGURE B29 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3)$ model with $X4$ and X_1X_2 as the candidate predictors.

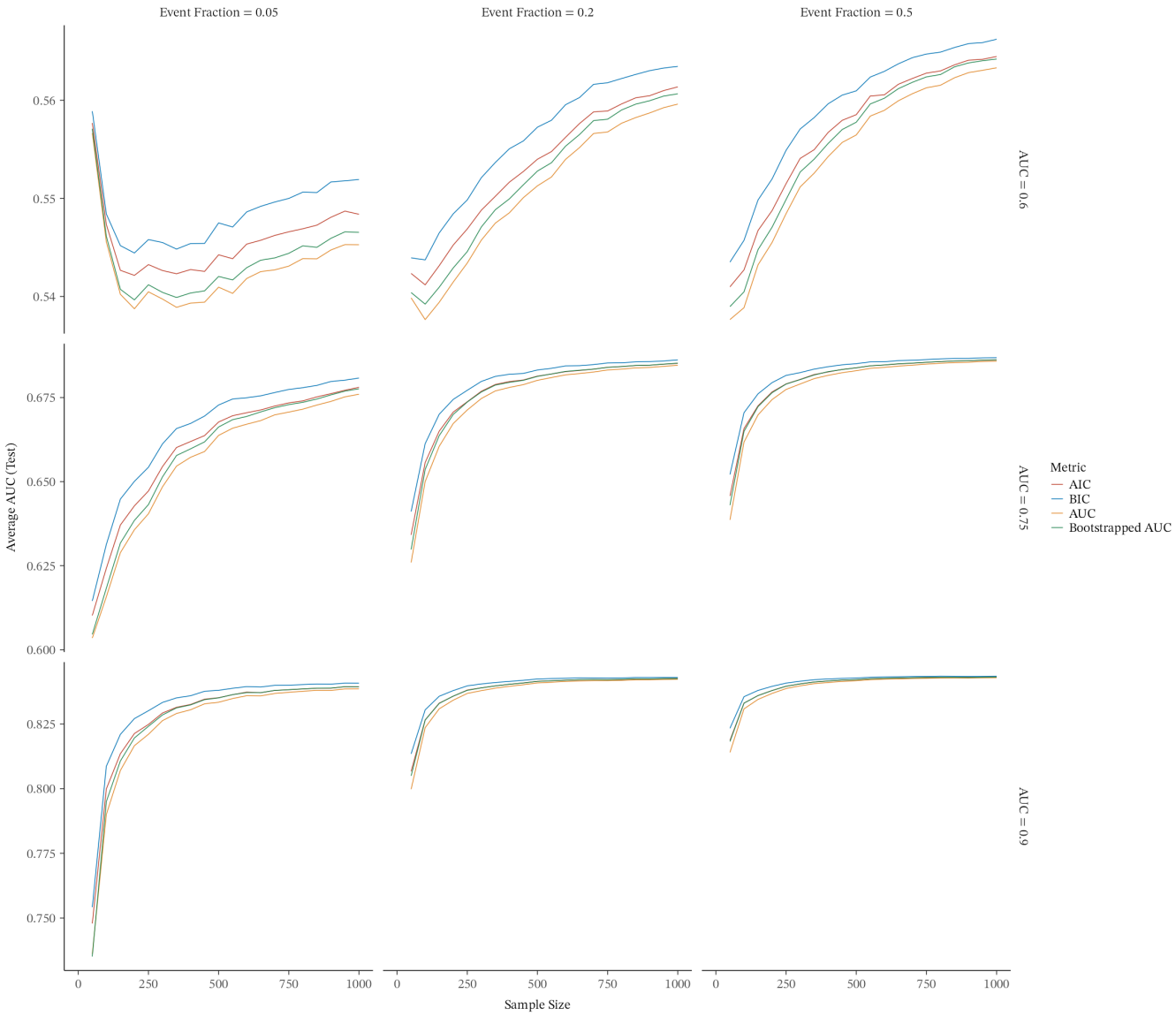

FIGURE B30 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3)$ model with X_4 and X_1X_2 as the candidate predictors.

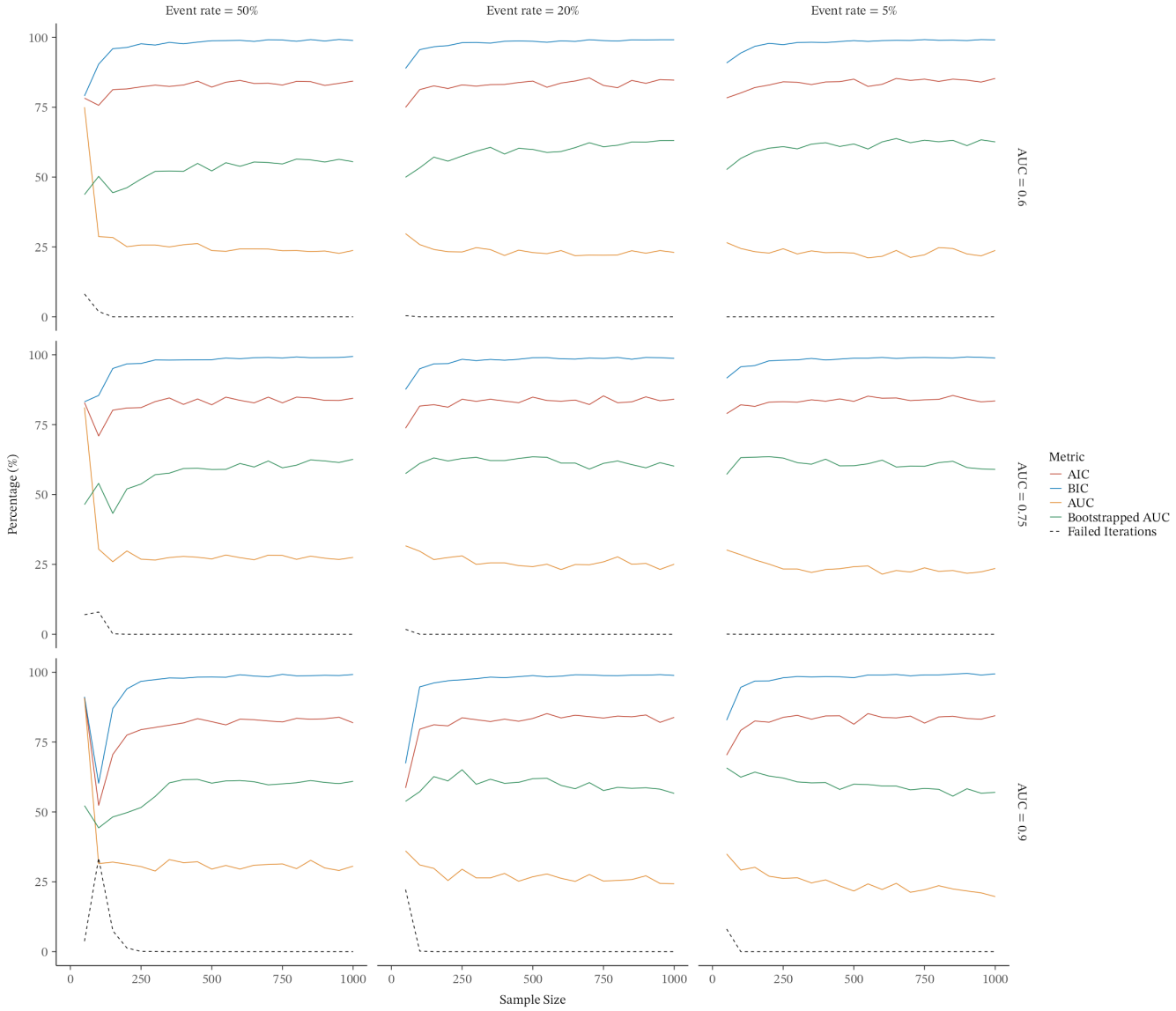

FIGURE B31 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + X_6 ... X_{10})$ model with X_5 as the candidate predictor.

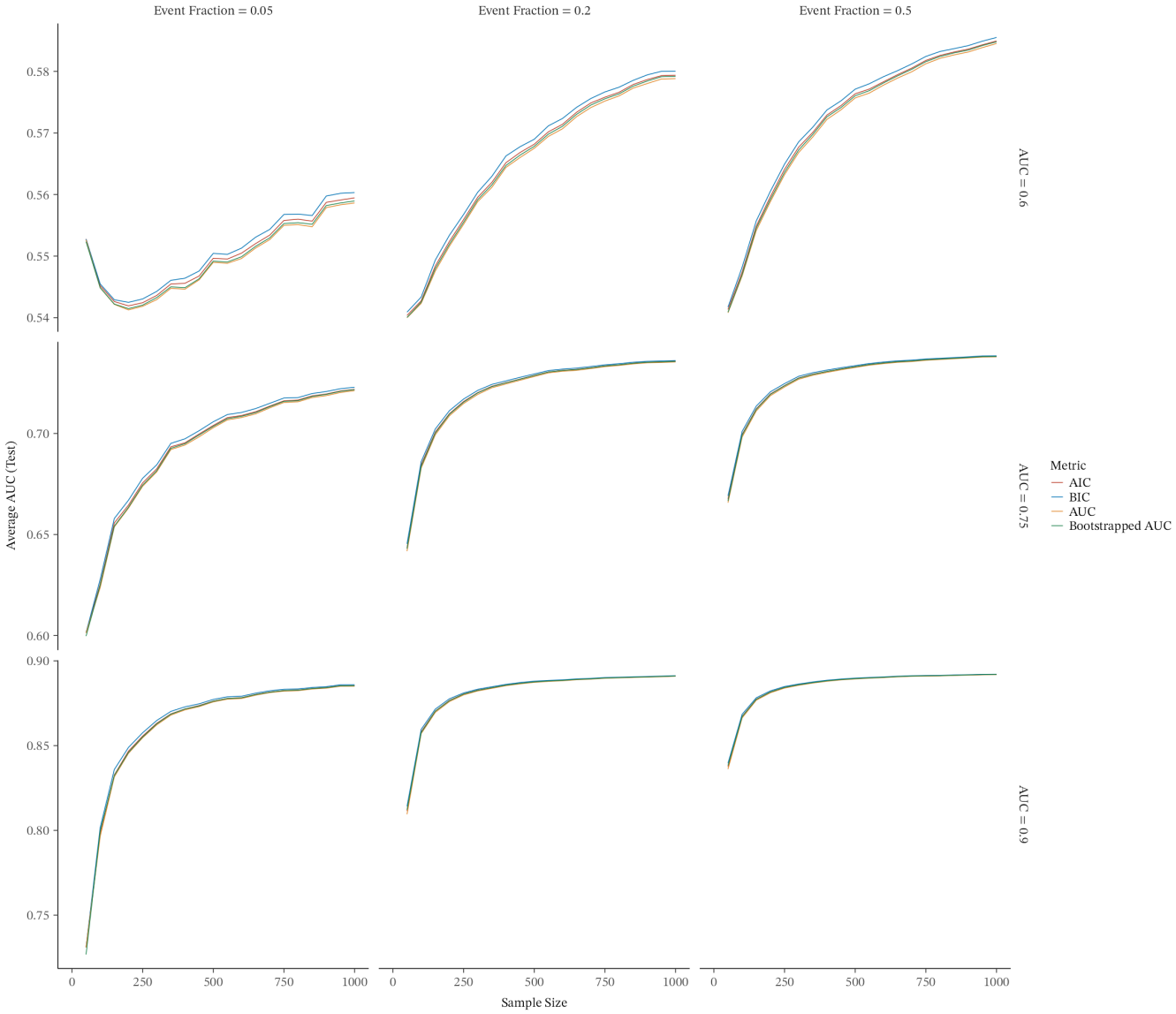

FIGURE B32 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + X_6...X_{10})$ model with X_5 as the candidate predictor.

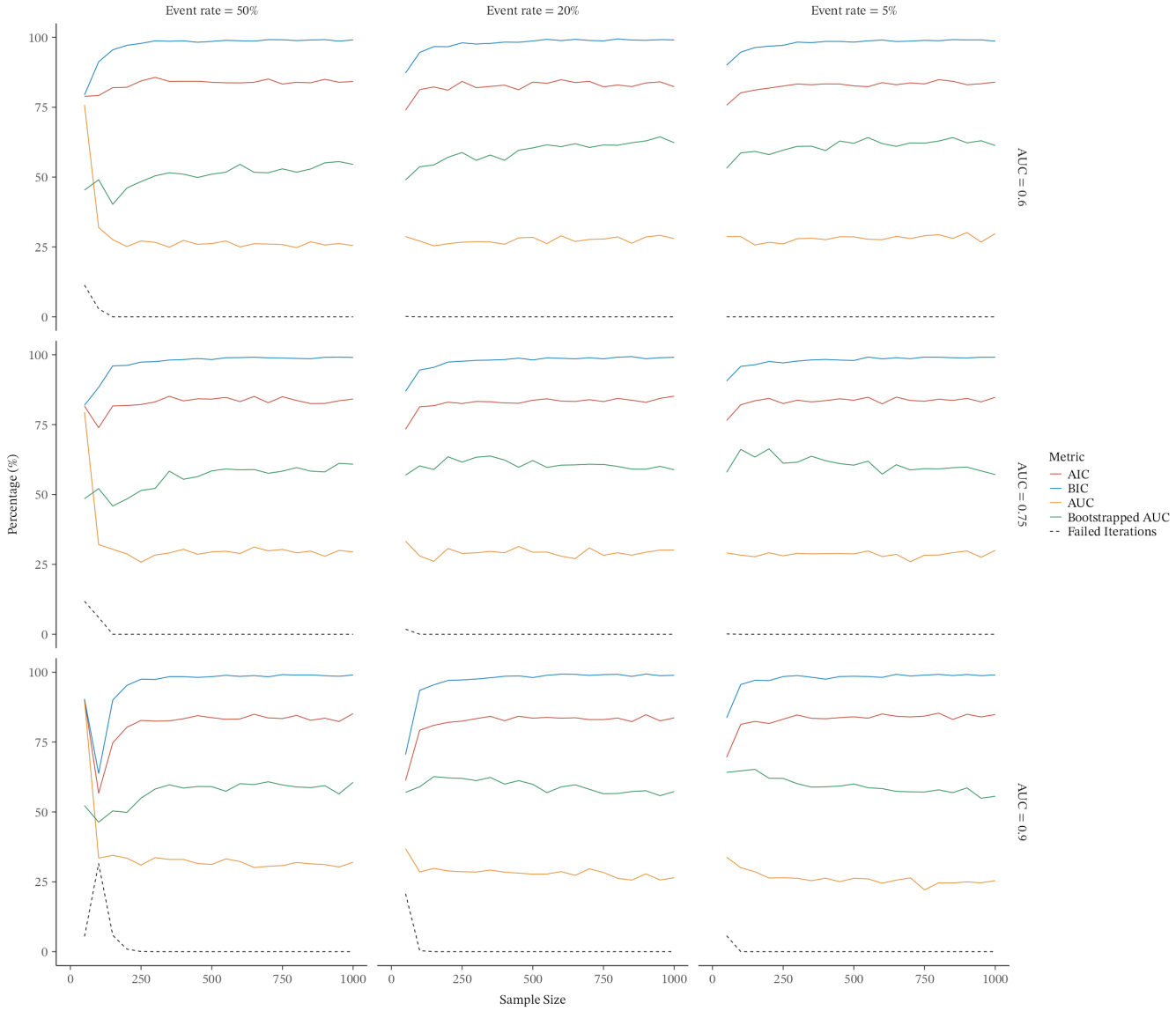

FIGURE B33 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + X_6 ... X_{10})$ model with $X_1 X_2$ as the candidate predictor.

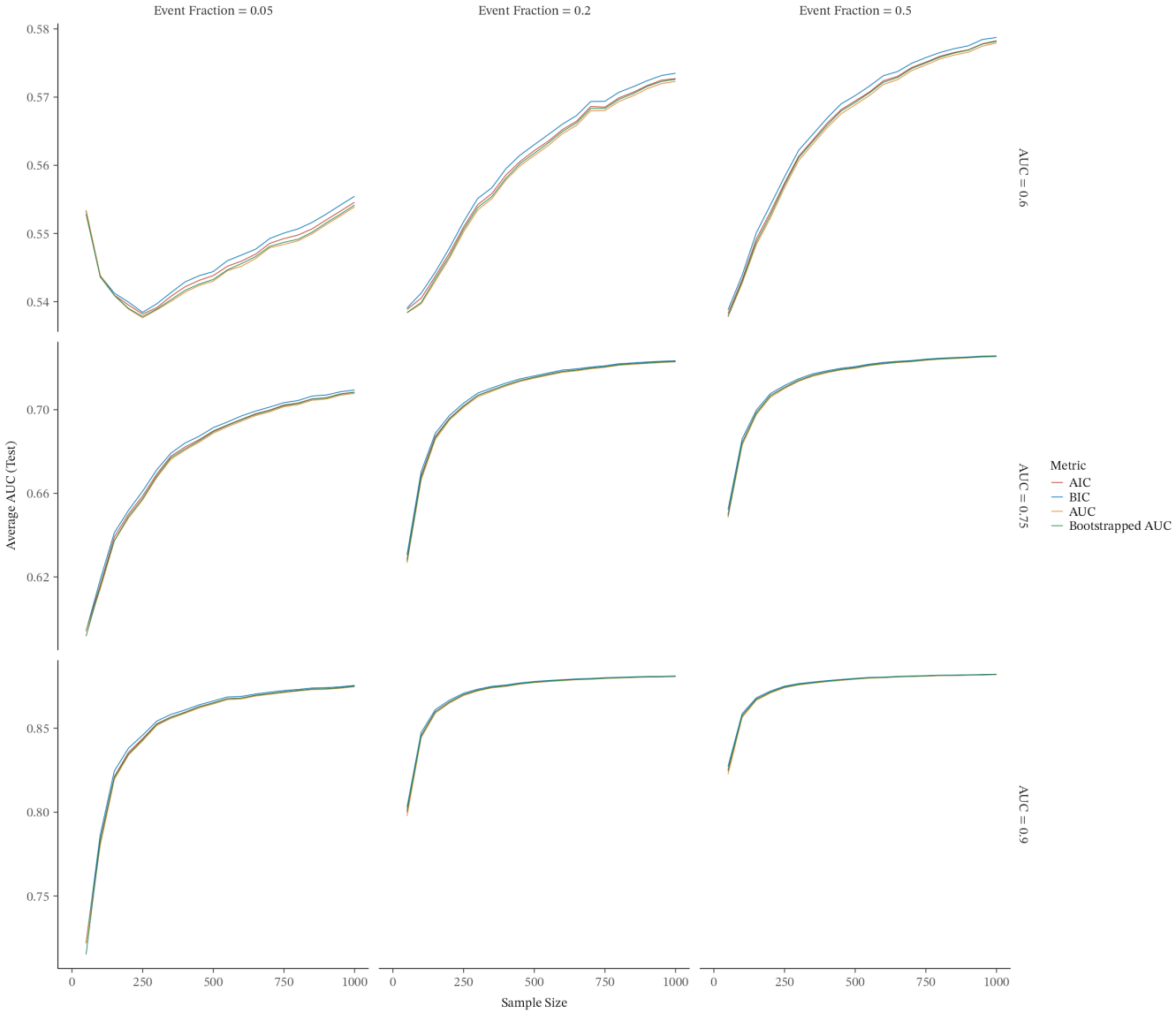

FIGURE B34 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_4 + X_6...X_{10})$ model with X_1X_2 as the candidate predictor.

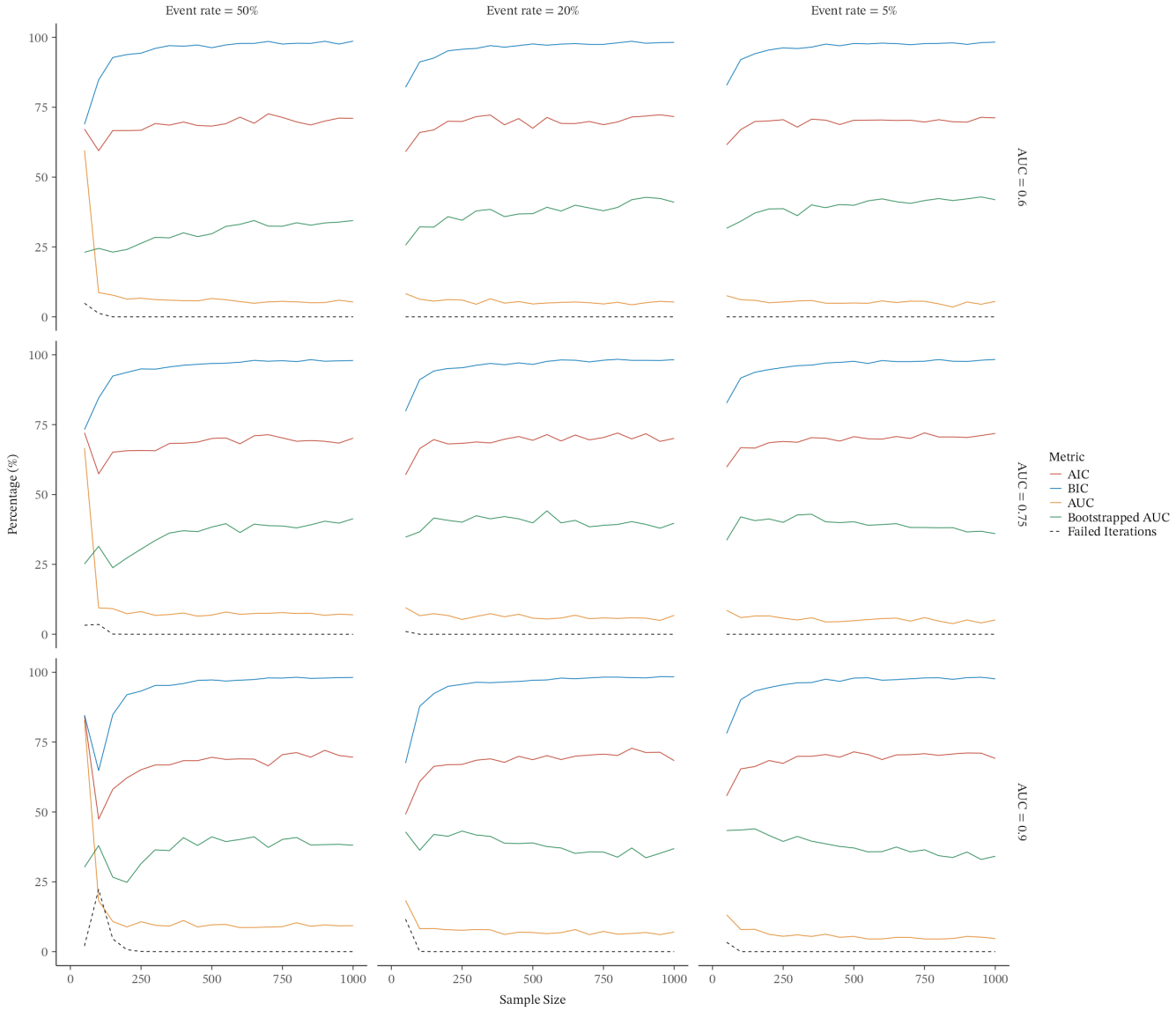

FIGURE B35 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_6 ... X_{10})$ model with X_4 and X_5 as the candidate predictors.

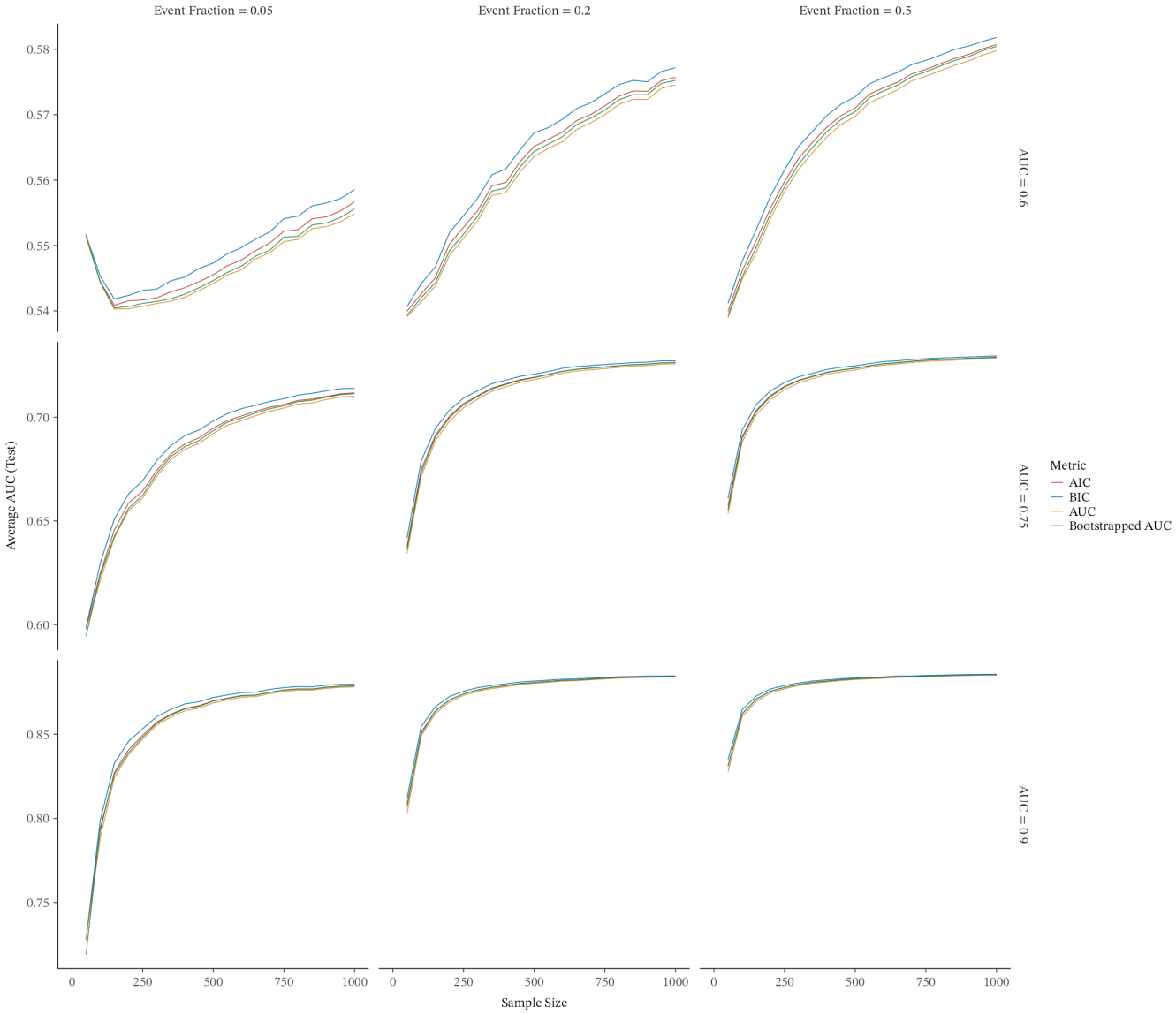

FIGURE B36 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_6...X_{10})$ model with X_4 and X_5 as the candidate predictors.

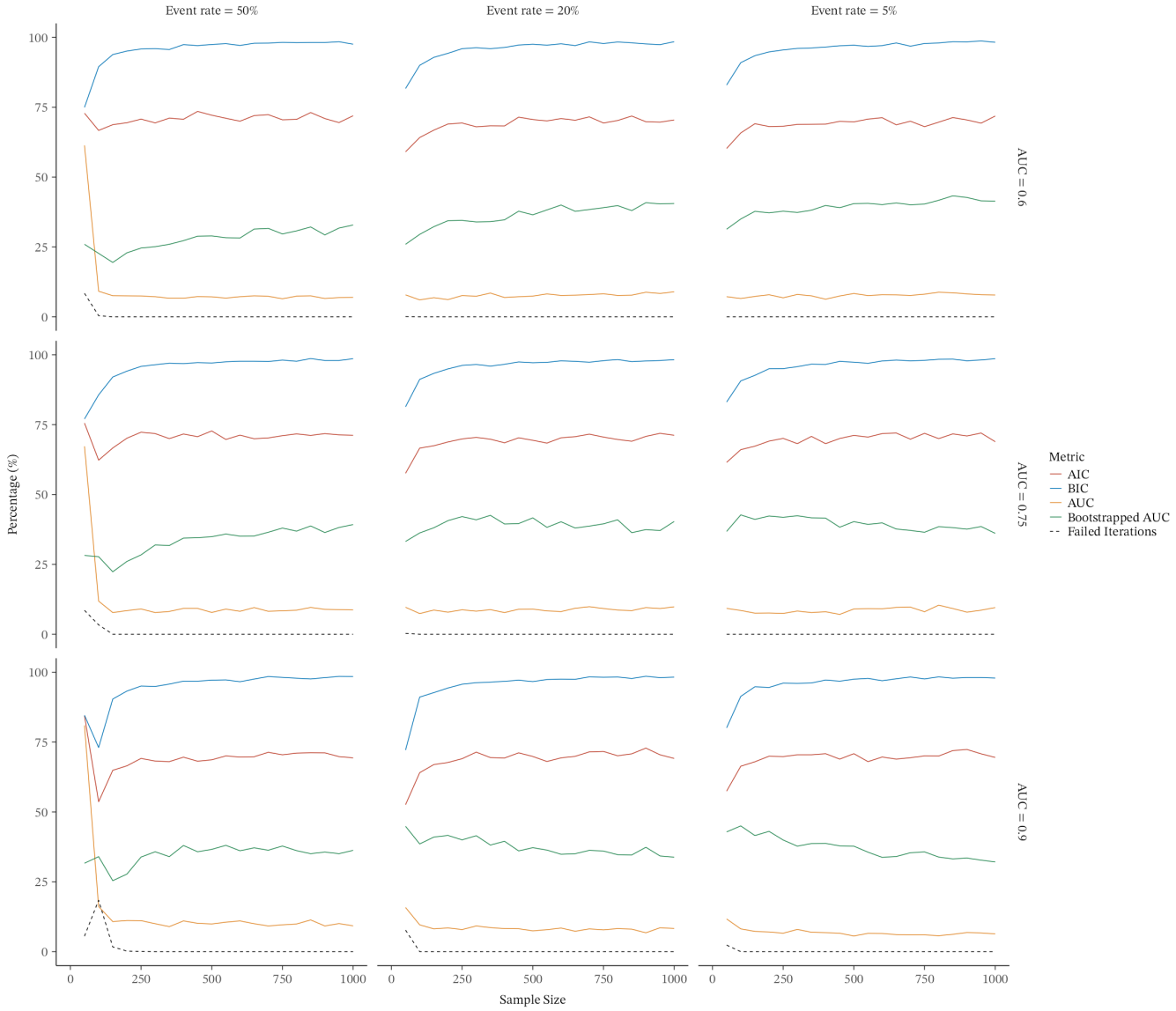

FIGURE B37 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_6 ... X_{10})$ model with X_1X_2 and X_2X_3 as the candidate predictors.

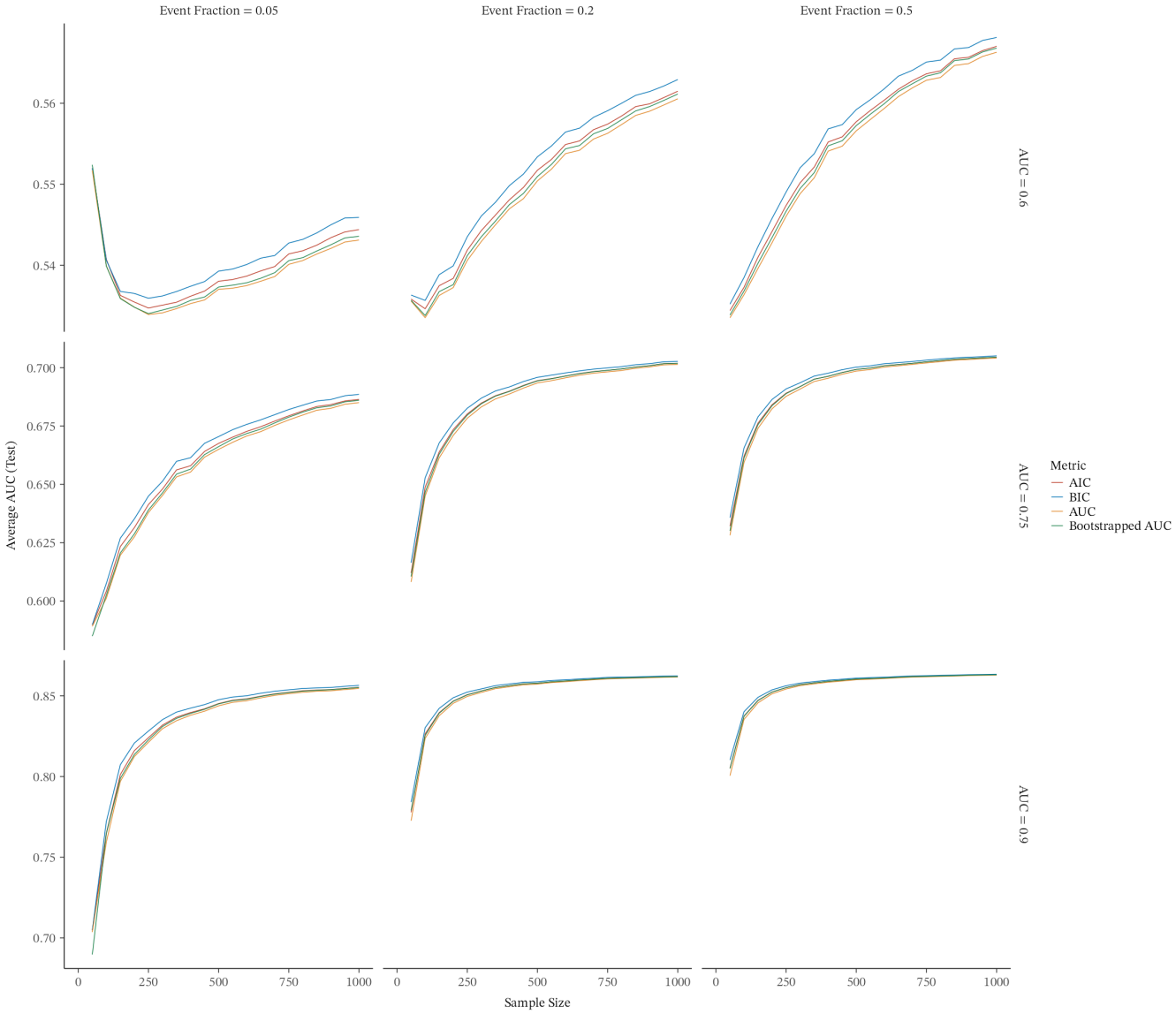

FIGURE B38 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_6...X_{10})$ model with X_1X_2 and X_2X_3 as the candidate predictors.

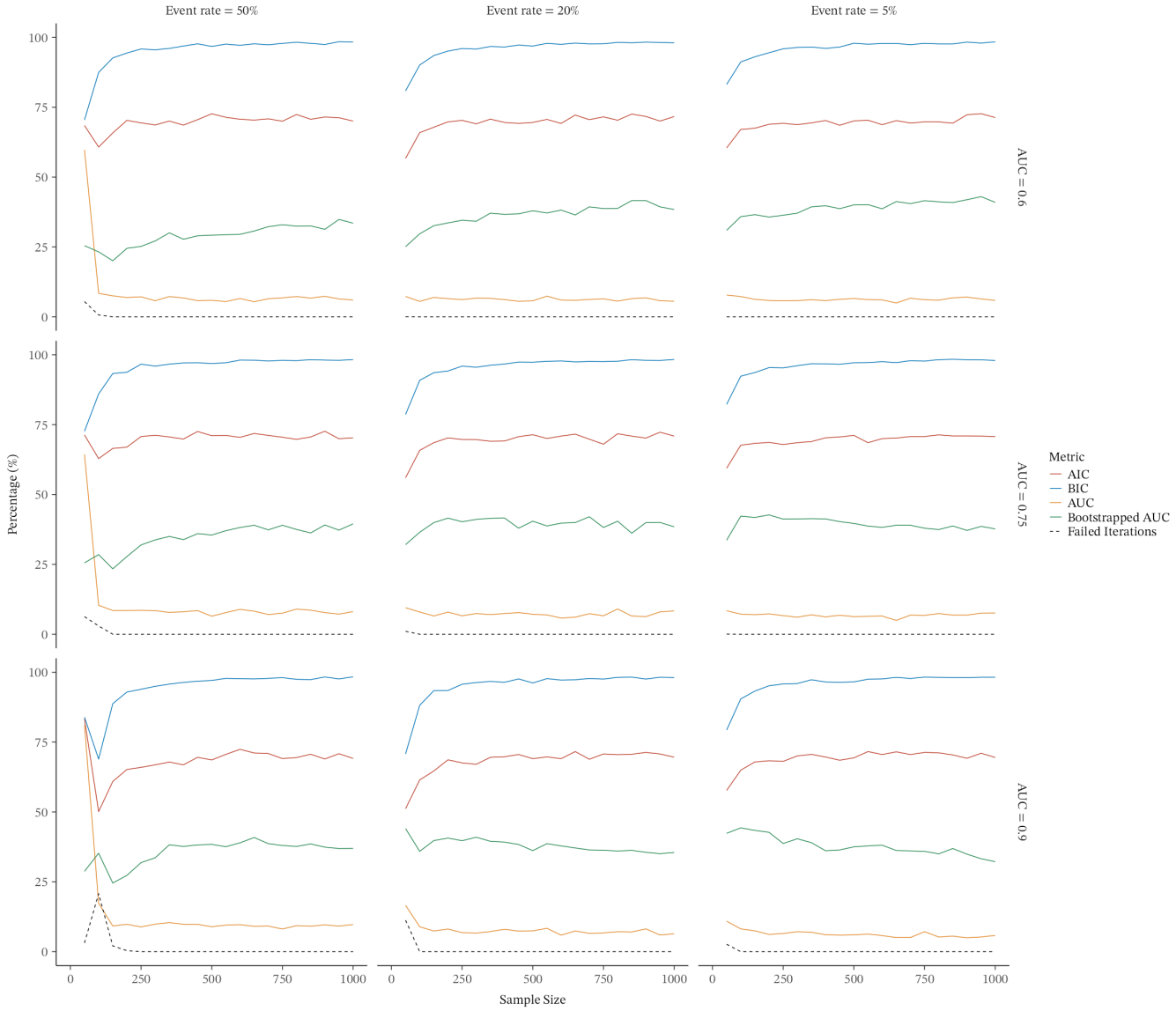

FIGURE B39 Success rate of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_6 ... X_{10})$ model with X_4 and X_1X_2 as the candidate predictors.

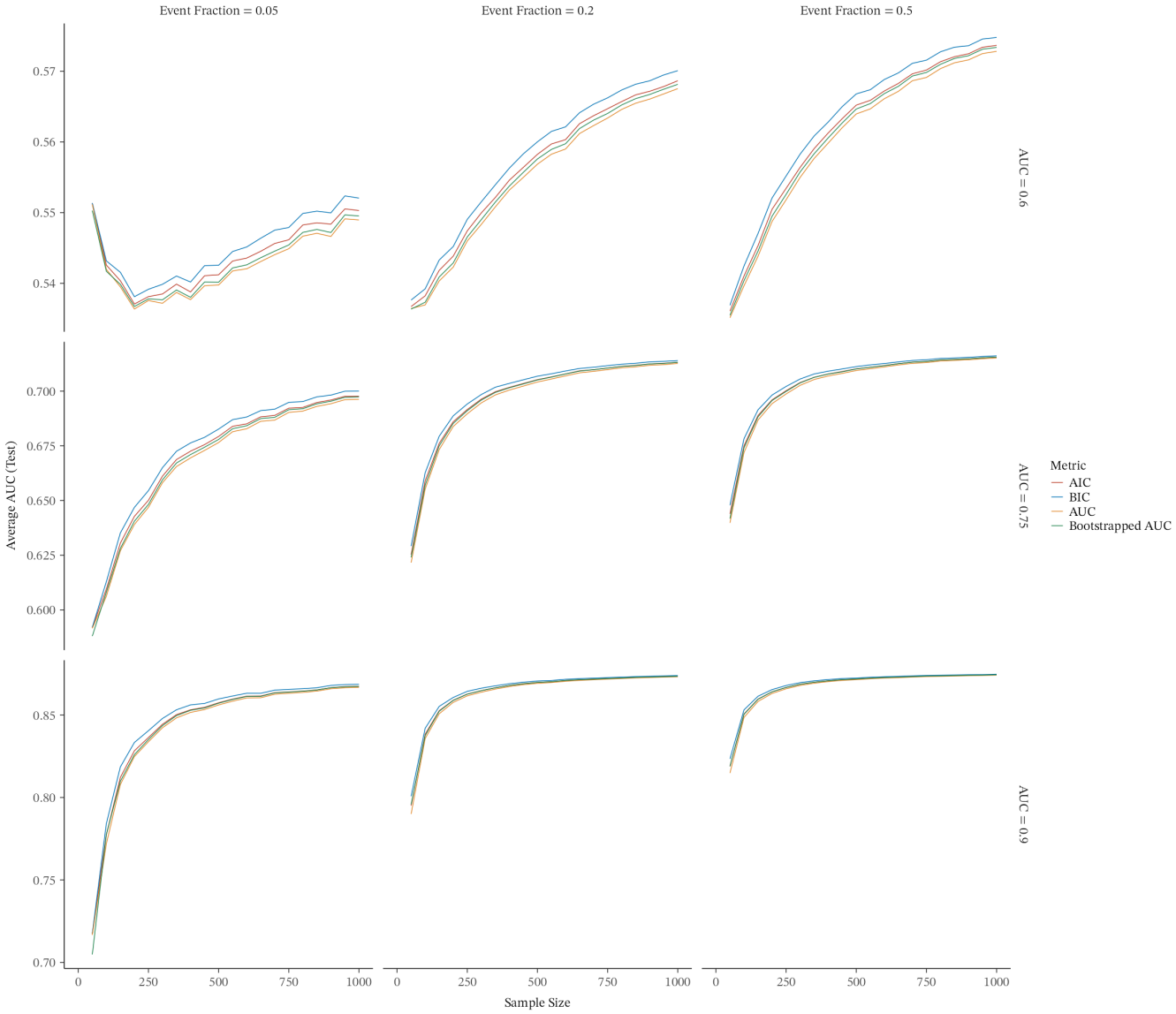

FIGURE B40 Test AUC of IC and internal performance measures for the $\beta_0 + \beta(X_1 + X_2 + X_3 + X_6...X_{10})$ model with X_4 and X_1X_2 as the candidate predictors.

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. Ripley B, Venables B, Bates DM, et al. Package 'Mass'. *Cran r.* 2013;538:113–120.

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

19. Gart JJ, Zweifel JR. On the Bias of Various Estimators of the Logit and Its Variance with Application to Quantal Bioassay. *Biometrika.* 1967:181–187.

20. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in Odds Ratios by Logistic Regression Modelling and Sample Size. *BMC Medical Research Methodology.* 2009;9(1):56. doi: 10.1186/1471-2288-9-56

21. Firth D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika.* 1993;80(1):27–38. doi: 10.1093/biomet/80.1.27

22. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological).* 1996;58(1):267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

23. van Smeden M, Moons KG, de Groot JA, et al. Sample Size for Binary Logistic Prediction Models: Beyond Events per Variable Criteria. *Statistical Methods in Medical Research.* 2019;28(8):2455–2474. doi: 10.1177/0962280218784726