Research Master's programme Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Utrecht University, the Netherlands

MSc Thesis Sofia Arsenia Gabriëlla Eleonora van den Brand (5611423)
TITLE: Clinical Prediction Models: A Comparison Between Cross-Validation and Bootstrap Approaches for Internal Validation
May 2021

Supervisors:
Dr. Maarten van Smeden
Prof. Dr. Ben Van Calster

Second grader:
Prof. Dr. Ir. Bernard P. Veldkamp

Preferred journal of publication: Statistics in Medicine
Word count: 5961

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

DOI: xxx/xxxx

MASTER THESIS

Clinical Prediction Models: A Comparison Between Cross-Validation and Bootstrap Approaches for Internal Validation

Sofie A. G. E. van den Brand*

[1]Methodology & Statistics, Utrecht University, The Netherlands

Correspondence
*S.A.G.E. van den Brand, Padualaan 14, 3584 CH Utrecht, the Netherlands. Email: s.a.g.e.vandenbrand@uu.nl

Present Address
Padualaan 14, 3584 CH Utrecht, the Netherlands

Summary

Clinical prediction models are important tools to support medical decision making. Statistical evaluation through internal validation is an essential step before such a model can be implemented in practice. However, with many internal validation strategies available, in particular various bootstrap and cross-validation variants, it is currently unclear which approach achieves the best out-of-sample performance under varying circumstances. In this thesis three extensive simulation studies are performed, evaluating the impact of outcome prevalence, strength of predictors, and type of modeling algorithm (both regression and tree-based). The ability of the approaches to estimate out-of-sample performance is evaluated in terms of Area under ROC curve, calibration slope and intercept, Mean Absolute Prediction Error, Root Mean Squared Prediction Error, $R^2_{\text{Cox-Snell}}$, R^2_{Tjur} and Estimated Calibration Index. The results illustrate that all internal validation approaches show more variability, compared to large sample external validation, for the AUC and R^2 measures, but less for the calibration intercept. Moreover, no single internal validation approach outperforms all others across all investigated scenarios and performance measures. This indicates that the optimal strategy is dependent on the specific setting and evaluation metric used. The results show that Harrell's bootstrap works generally well when using regression-based models. However, when using tree-based models, which showed a high amount of overfitting, Harrell's, .632 and .632+ bootstrap display large overestimation, while 5, 10 and 10x10 fold cross-validation approach the external validation estimates.

KEYWORDS:
Clinical prediction models, bootstrap, cross-validation, internal validation

1 | INTRODUCTION

In healthcare, clinical prediction models aim to support medical decision making by providing estimated risks of presence of disease (diagnosis) and future health status (prognosis) for patients.[1] It is generally acknowledged that the performance of these predictions needs to be evaluated before implementing the models in practice. The first step in evaluating prediction

[0]**Word count: 5961 (including abstract). Supervisors: Maarten van Smeden & Ben Van Calster**

models is internal validation, which uses the data available for model development to estimate predictive performance while avoiding overfitting of the data (i.e. adjust for idiosyncrasies within the data).[2,3] Overfitting can happen for example when obtaining the apparent performance, where all available data is used to both develop and evaluate the model, often resulting in too optimistic performance estimates.[4,5] The second step to evaluate prediction models is external validation, commonly regarded as an estimate of true model performance, which aims to estimate out-of-sample performance. This thesis mainly focuses on the first step, internal validation.

Advocated internal validation strategies are regularly a specific application of the bootstrap or cross-validation.[5] Generally, the bootstrap entails resampling, mirroring the process of drawing a sample from a certain population.[6] Cross-validation on the other hand, splits the data into K mutually exclusive parts, where the model is iteratively trained on $K-1$ parts and tested on the remaining part. In this thesis, the focus lies on the popular variants: 5, 10 and 10x10 fold cross-validation, and Harrell's, .632, and .632+ bootstrap. Each of these approaches aims to evaluate the model development procedure, and thus repeats all steps of model development. Split sample and leave-one-out-cross-validation are not assessed in this thesis. The former is found to be statistically inefficient,[7,8,9] while the latter is too computationally exhaustive.

Previous studies comparing cross-validation and bootstrap have resulted in conflicting recommendations for internal validation. Several studies concluded that bootstrap methods should be used.[7,10,11] For instance, Steyerberg et al.[10] showed in a resampling study that bootstrap approaches led to accurate estimates of model performance using logistic regression models. However, different studies recommended cross-validation approaches[12,13] and yet again others did not advocate one method over the other.[14,15] Moreover, some studies focused on either cross-validation or bootstrap. For example Varoquaux[16] only investigated the limits for cross-validation, while Jiang and Simon[17] mainly focused on bootstrap approaches within high-dimensional data settings. All of these studies differed regarding the modeling algorithms used, sample sizes, event-fractions, dimensionality, and the performance measures assessed. This suggests that the optimal internal validation strategy may depend on the specific setting and performance estimators, but specific guidance is still lacking.

This thesis presents extensive simulation studies to evaluate and compare internal validation strategies in their ability to estimate out-of-sample performance in the same population the model was developed on. In three separate simulation studies, the sample size is varied across all studies, and prevalence, predictor settings, and modeling algorithms used, in each study, respectively. The influence of adjusting these settings, is assessed by estimating performance measures such as: Area under ROC curve (AUC), calibration slope and intercept, Mean Absolute Prediction Error (MAPE), Root Mean Squared Prediction Error (rMSPE), $R^2_{\text{Cox-Snell}}$, R^2_{Tjur} and Estimated Calibration Index. The aim is identify similarities and differences between the internal validation approaches in approximating out-of-sample performance, to guide researchers in choosing an internal validation method depending on their situation at hand.

The remainder of the thesis is structured as follows. In section 2, modeling algorithms, performance measures, and internal validation approaches are introduced. Following, section 3 describes the setup of the simulation studies and their procedure. The results of these studies can be found in section 4. The discussion of the thesis is outlined in section 5.

2 | DEVELOPMENT OF PREDICTION MODELS WITH BINARY OUTCOME

The focus lies on estimating out-of-sample performance measures of prediction models for $\pi(Y = 1|\mathbf{X})$. Here, Y represents a binary outcome and $\mathbf{X}$ a matrix of P candidate predictors ($\mathbf{X} = X_1, X_2, ..., X_P$). modeling strategies are described in section 2.1, the measures of model performance in section 2.2, and internal validation approaches in section 2.3.

2.1 | Modeling algorithms

2.1.1 | Logistic regression using maximum likelihood

Maximum likelihood logistic regression is commonly used for clinical prediction modeling. Assuming only linear effects, the probability for individual i, with ($i = 1, ..., N$) given some set of P candidate predictors without interactions, can be estimated as follows[18]

$$\pi_i = \frac{1}{1 + e^{-(\boldsymbol{\beta}' \boldsymbol{x}_i)}}.$$

Here, $\boldsymbol{\beta}$ is a vector containing an intercept and coefficients. In maximum likelihood logistic regression, these are estimated by ML (i.e. maximizing the log-likelihood)

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{ y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \}.$$

2.1.2 | Penalized regression

In developing clinical prediction models, shrinkage methods are recommended when the number of events is small compared to the number of candidate predictors.[19] A review of different suggested penalized logistic regression models can be found in Pavlou et al.[20] One application of penalized regression is Firth's correction, which penalizes the likelihood function using Jeffreys invariant prior.[21,22]

$$\log(\boldsymbol{\beta}) + 0.5 \log |I(\boldsymbol{\beta})|,$$

where Fisher's information matrix evaluated at $\boldsymbol{\beta}$ is denoted by $I(\boldsymbol{\beta})$. To obtain unbiased predicted probabilities, the intercept can be corrected by using Firth's logistic regression with intercept-correction (FLIC).[21] FLIC implements a maximum likelihood estimate of the intercept, α^*, while taking the linear predictor obtained by Firth's correction, as an offset in a logistic regression model for $\text{logit}(Y = 1) = \alpha^* + \text{offset}(\boldsymbol{\beta}' \boldsymbol{x}_i)$.

Another commonly applied penalized approach is Ridge regression, which shrinks the regression coefficients by maximizing the following penalized log-likelihood function[18,23]

$$\log(\boldsymbol{\beta}) - \lambda \sum_{p=1}^{P} \beta_p^2.$$

Likewise, another method that is often implemented is Least Absolute Shrinkage and Selection Operator (LASSO) regression. Here, the penalty allows regression coefficients to be shrunk towards zero (i.e. it can perform variable selection)[18,23]

$$\log(\boldsymbol{\beta}) - \lambda \sum_{p=1}^{P} |\beta_p|.$$

For Ridge and Lasso, λ determines the amount of shrinkage to be applied, where larger values indicate greater shrinkage.[18] Following Van Calster and colleagues,[23] the tuning of λ to identify the value that minimizes the deviance for both Ridge and Lasso was based on a 10 fold cross-validation approach, using a grid ranging from 0 to 64 with 250 equal sized step increments on the log scale (except for the null-value itself).

2.1.3 | Tree-based models

An alternative to the regression framework are tree-based models, such as Classification And Regression Trees (CART).[24,25] This is a simple approach which splits development data into two child nodes, where the split is based on obtaining the smallest sum of impurities in the nodes. The complexity parameter (cp) determines the size of the tree, where the smaller cp, the more complex a tree is allowed to be.[26,27] Any split that does not improve the fit of the model, expressed in R^2,[27] by cp is not attempted. Within this thesis, the tuning of each decision tree is based on 10 fold cross-validation with a grid search for the complexity parameter ranging from 0 to 0.01 with step increments of 0.00034.

Another tree-based model is Random Forest (RF).[28] For classification, RF combines the results of a large number of decision trees through a majority vote. This reduces the variance between predictions compared to using single trees.[29] In this study, the number of aggregated trees was set to 250. The optimal number of predictors to be possibly split at each node, was tuned by using 10 fold cross-validation over a grid ranging from 1 to the maximum number of predictors within a simulation scenario with step increments of 2.[30]

2.2 | Model performance

Performance of clinical prediction models can be quantified using the measures described below. The ability of a model to differentiate between low and high risk subjects (i.e. discrimination) is often assessed by the area under the ROC-curve (AUC). This represents the probability that a patient without the event has a lower predicted probability than a randomly selected patient

with the event.[6] The closer the AUC is to 1, the greater the discriminative performance of the model. An uninformative model typically has an AUC of 0.5.

The reliability of estimated risks is also an important performance measure, which can be evaluated by the calibration slope and intercept.[31] Together, these measures indicate whether on average, the risks are not over- or underestimated and the risk estimates are neither too extreme nor too moderate. The target value for the calibration slope is 1 (too extreme risks (overfitting): < 1, too moderate risks (underfitting): > 1) and 0 for the calibration intercept (overestimation of risk: < 0, underestimation of risk: > 0). To compare calibration performance between different models, one can use the Estimated Calibration Index (ECI).[32] In this thesis, an adapted version of the ECI is computed. The ECI is the average squared difference between the estimated risks, $\hat{p}_{nj}$, and observed event-fractions, $\hat{o}_{nj}$. Where $\hat{o}_{nj}$ is derived from a flexible calibration model, see Van Hoorde et al. for more details.[32] For n ($n = 1, ..., N$) observations and j ($j = 1, ..., J$) outcome categories the original formula of the ECI,

$$\frac{\sum_{n=1}^{N}\sum_{j=1}^{J}(\hat{p}_{nj} - \hat{o}_{nj})^2}{N * J} * \frac{100 * J}{2},$$

has a range of between 0 and 100. In this thesis, this formula is adapted to have a range between 0 and 1, using the following formula instead,

$$\frac{\sum_{n=1}^{N}\sum_{j=1}^{J}(\hat{p}_{nj} - \hat{o}_{nj})^2}{\sum_{n=1}^{N}\sum_{j=1}^{J}(\hat{p}_{nj} - \bar{Y}_j)^2}.$$

Here, $\bar{Y}_j$ represents the event-fraction of outcome category j. An ECI of 0 represents perfectly estimated risks, while 1 represents a random model, in which the observed proportion is equal to $\bar{Y}_j$.

Commonly used measures to evaluate explanatory power are Cox-Snell R^2 (R^2_{CS}) and Tjur's R^2 (R^2_{Tjur}, i.e. the coefficient of Discrimination).[33,34] Both are so-called pseudo R^2 measures and approximate the interpretation of R^2 for linear regression, the proportion of variance explained. The R^2_{CS} measure takes sample size into account, but it cannot reach the maximum value of 1, because its range is dependent on the event-fraction in the observed data. R^2_{Tjur} can be interpreted as the difference between the averages of estimated risks for events and non-events.[34] For both measures, higher values indicate greater explanatory power and therefore, better performance.

To assess the prediction error, one can use the Mean Squared Prediction Error (MAPE) and the square root of the Mean Squared Prediction Error (rMSPE).[35] The prediction error is based on the distance between the estimated and true probabilities, where the latter is obtained by applying the data generating model to the generated data set. Please note that these measures are therefore only estimable in situations where the data generating mechanism is known, such as simulation studies. The lower the values of both rMSPE and MAPE, the better the model performance.

2.3 | Internal validation approaches

To obtain valid estimates of model performance measures, such as described above, internal validation is critical. Specifically, its aim is to evaluate the model development procedure. In the following, several applications of cross-validation and bootstrap approaches are explained.

2.3.1 | Cross-validation

K-fold cross-validation is a popular approach to perform internal validation of prediction models, with $K = 5$ and $K = 10$ as two particularly common approaches.[18] In K-fold cross validation, the data are split up into K mutually exclusive parts, where all modeling steps are repeated on $K - 1$ parts and the resulting model is tested on the remaining part in an iterative fashion. The performance estimates are calculated on each left-out part and averaged to result in a cross-validation based estimate of performance.[36] An extension of K-fold cross-validation is repeated cross-validation. For instance, in 10x10 fold cross-validation, the 10-fold cross validation is repeated ten times, using a different split for each repetition. The average over the results (100 per performance measure) produces the final estimate of performance. Within this thesis, the folds were stratified by outcome for all cross-validation approaches to reduce experimental variance.[36]

2.3.2 | Bootstrap

One commonly used bootstrap approach is Harrell's enhanced bootstrap, also called Harrell's bias correction.[37,38] It uses the following procedure:[4] Generate B bootstrap samples by sampling with replacement from the original data. For this thesis, B

= 500 bootstrap samples are used for each prediction model. Let θ represent a vector of performance measures. Fit a model and obtain performance measures for each bootstrapped sample: $\theta_{1,boot}$, $\theta_{2,boot}$, ..., $\theta_{B,boot}$. Next, obtain performance measures by applying these models to the original data: $\theta_{1,o}$,$\theta_{2,o}$, ..., $\theta_{B,o}$, where o stands for original data. The estimate of the optimism can then be obtained as

$$\hat{\Lambda} = \frac{1}{B}\sum_{b=1}^{B}(\theta_{b,boot} - \theta_{b,o}).$$

The bias-corrected performance measure is derived as $\theta_{Heb} = \theta_{app}$ - $\hat{\Lambda}$, where the apparent performance, θ_{app}, is obtained from a model that is trained and tested on all of the original data.

When using the previous bootstrap approach, on average 63.2% of the cases in the original sample are sampled at least once within each bootstrap sample. This overlap may give an overestimation of the performance measures. Therefore, Efron proposed the .632 bootstrap approach to account for the overlap.[39] Using those cases that are not sampled in each bootstrap sample as a test set for the B'th prediction model, obtain the performance measures $\theta_{1,test}$, $\theta_{2,test}$, ..., $\theta_{B,test}$ and their average θ_{test}. With the latter, the .632 optimism-corrected performance estimate can be obtained by $\theta_{.632} = 0.368 * \theta_{app} + 0.632 * \theta_{test}$.

However, this approach tends to be optimistically biased in highly overfit situations (i.e. when θ_{app} has large bias).[40,4] Therefore, Efron and Tibshirani[40] proposed the .632+ bootstrap. Here, relative overfit, $\hat{R}$, is derived by

$$\hat{R} = \frac{\theta_{test} - \theta_{app}}{\hat{\gamma} - \theta_{app}},$$

where $\hat{\gamma}$ is the no-information error rate (i.e. the value of a performance measure in a random model). For instance, $\hat{\gamma}_{\text{AUC}}$ is 0.5.[4,41] With the relative overfit, a weight, $\hat{w}$, can be derived

$$\hat{w} = \frac{.632}{1 - .368 * \hat{R}},$$

which can then finally be used to obtain the .632+ estimates $\theta_{.632+} = (1 - \hat{w}) * \theta_{app} + \hat{w} * \theta_{test}$.

3 | SIMULATION STUDIES

To evaluate and compare the differences in performance between internal validation approaches, three simulation studies are conducted, which are explained in detail according to the ADEMP structure in sections 3.1-3.5.[42] The software used and error-handling is discussed in section 3.6. For the sake of manageability of the studies, some assumptions are made regarding data generation, modeling steps and data settings. These assumptions are entwined in what follows.

3.1 | Aim

The overall aim of the three studies was to investigate and compare performance of bootstrap and cross-validation based internal validation approaches. Specifically, to estimate out-of-sample performance of prediction models for binary outcomes using the same target population that was used for model development. The focus lied on commonly encountered low-dimensional data settings, modeling algorithms[43] (section 2.1) and internal validation approaches (section 2.3).

3.2 | Data generating mechanism

The simulation studies were divided into several scenarios, which were simulated 500 times each (i.e. runs). For all scenarios, the AUC of the data generating model (AUC_{dgm}) stayed constant at 0.75 and for each run of a scenario, a development data set was generated. Depending on the scenario, the sample size of the development data set was either half ($\text{N}_{\text{min}}/2$), twice ($\text{N}_{\text{min}} * 2$), or the minimal sample size itself (N_{min}), as calculated using the criteria proposed by Riley and colleagues.[44] The R^2_{CS} needed for the sample size calculations was based on an approximation of the AUC_{dgm} and expected event-fraction, as suggested by Riley et al.[45] A schematic overview of the studies and scenarios is depicted in Table 1.

Candidate predictor data were simulated from a multivariate normal distribution, where the covariance matrix corresponds to a correlation matrix with a fixed correlation of 0.2 between the predictors (i.e. equal pairwise correlation for all predictors). The regression coefficients of the data generating model were approximated numerically to ensure a prespecified event-fraction and AUC_{dgm}.[35] Details of this numerical procedure can be found in the supplemental material file within in the GitHub repository accompanying this thesis.

The relations between predictors and outcomes were generated as linear effects in a logistic regression model without interactions or non-linearities. Compared to real data applications, in which binary predictors occur frequently, using only continuous predictors may have lowered the occurrence of both data separation and overfit data. Regarding model performance, modeling algorithms that assume a logistic function are expected to have an advantage compared to other models. However, the goal is explicitly not to compare performance between models.

TABLE 1 Design simulation studies

Simulation settings	Study 1	Study 2	Study 3
AUC_{dgm}	0.75	0.75	0.75
Candidate predictors	10	6, 30	20
Event-fraction	0.05, 0.2, 0.5	0.2	0.2
Percentage of noise variables	20%	0%, 50%	20%
Models	ML, ML_{AIC}	ML, ML_{AIC}	ML, F, Ridge, Lasso, CART, RF
Sample sizes			
$N_{min}/2$	1064, 330, 219	198, 990	660
N_{min}	2128, 660, 438	396, 1980	1320
$N_{min} * 2$	4256, 1320, 876	792, 3960	2640
Number of scenarios	18	24	18

Note. Abbreviations: ML = Logistic regression using maximum likelihood, ML_{AIC} = Logistic regression using maximum likelihood with backwards selection based on AIC criterion ($p < .157$), F = Logistic regression using Firth's correction, CART = Classification and Regression Trees, RF = Random Forest.

For study 1 and 3, each scenario will generally contain 30% strong predictors, 50% weak predictors and 20% noise predictors. For all studies, a strong predictor was defined as 3 times the regression coefficient of a weaker predictor (on the log-odds ratio scale), noise variables were generated to have a regression coefficient of 0. For each of the 60 scenarios, a separate independent data set from the same data generating mechanism served as a validation data set to quantify the out-of-sample performance. The size of each independent validation set was calculated as $50 * \frac{100}{\text{event-fraction}}$, where the event-fraction was dependent on the scenario.[46,47,48]

3.2.1 | Study 1: Influence of event-fraction

In study 1, the focus was on the impact of varying the event-fraction (i.e. prevalence of the outcome): 0.05, 0.2, 0.5 and sample sizes ranging from 219 to 4256, while keeping constant the number of candidate predictors at 10. In this study, the maximum likelihood logistic regression model was developed both with and without backwards elimination (AIC-based selection, $p < .157$), resulting in a total of 18 scenarios.

3.2.2 | Study 2: Influence of dimensionality and noise

Study 2 focused on the dimensionality and noise of predictors. Here, the number of predictors varied between 6 and 30; the percentage of noise variables within candidate predictors between 0% and 50%; and sample sizes ranged from 198 to 3960. The expected event-fraction stayed constant at 0.2. In case of 50% noise variables, the distribution of candidate predictor effects was as follows: 1 strong, 2 weak and 3 noise predictors for every 6 predictors. Predictor effects for scenarios with 0% were set to 1 strong and 5 weak predictors for every 6 predictors. Maximum likelihood logistic regression model was developed both with and without backwards elimination ($p < .157$), resulting in a total of 24 scenarios.

3.2.3 | Study 3: Influence of modeling algorithms

In study 3, the variability between types of regression or tree-based models was studied. The type of modeling algorithms to be studied, which were selected according to those most commonly used in clinical prediction modeling.[43] Tuning hyperparameters

TABLE 2 Occurrence of errors during the simulation studies and their consequences

	Instances (%)		Consequences
Generated development data sets	30,000	(100%)	
Bootstrap samples drawn	15,000,000	(100%)	
Folds created for cross-validation	3,450,000	(100%)	
Total number of models fitted*	18,480,000	(100%)	
Separation detected	8,325	(0.045%)	When separation was detected no action was taken, since all metrics could still be obtained.
Probabilities of exactly 0 or 1 occured	40,151	(2.17%)	The values of these were changed to probabilities of 0.000001 or 0.999999, respectively.
Estimation problems ECI	306	(0.0017%)	
Number of models with predictor selection	9,240,000	(50% of total)	
No predictors selected	121	(0.001%)	Calibration slope can not be estimated and will thus be replaced by the largest calibration slope within the scenario.
No predictors selected apparent model	29	(0.0003%)	No metrics could be estimated for the .632+ bootstrap approach and were thus treated as missing data.
Number of models with hyperparameter tuning*	3,696,000	(25% of total)	
< 8 events detected	0	(0%)	Use LOOCV for the tuning of hyperparameters instead of 10-fold cross-validation.

Note. Abbreviations: ECI = Estimated Calibration Index, LOOCV = Leave One Out Cross-Validation. Probabilities of exactly 0 or 1 should not be confused with separation within maximum likelihood estimation, as this specific error only occurred within tree-based models. *The total number of models fitted and number of models with hyperparameter tuning excludes the models created during hyperparameter optimization.

of the machine learning algorithms was based on 10-fold cross validation with grid search to maximize the accuracy. Details on tuning of hyperparameters can be found in section 2.1. Again the sample size varied (660 - 2640), while the event-rate and number of candidate predictors were held constant at 0.2 and 20, respectively. Within study 3, 18 scenarios were investigated.

3.3 | Estimands

The predictive performance metrics for the simulation studies were the model performance measures, as described in section 2.2: the Area Under the ROC Curve (AUC); calibration slope and intercept; ECI; R^2_{CS} and R^2_{Tjur}; Mean Absolute Prediction Error (MAPE) and root Mean Squared Prediction Error (rMSPE). These metrics were obtained for all of the methods described next.

3.4 | Methods

In total, eight methods were investigated within the simulation studies, of which six were internal validation methods: 5, 10 and 10x10-fold cross-validation,[18] and Harrell's enhanced, .632, and .632+ bootstrap.[38,39,40] Apparent validation was also assessed which used the same development data set to both train and test a prediction model. Serving as a reference for all aforementioned approaches, external validation was also assessed, by applying the prediction model derived within the apparent validation procedure to a large scale independent validation set.

3.5 | Performance measures

Of primary interest was the difference between the metrics obtained within the internal and external validation approaches. For each simulation run, the apparent, internal (6 methods) and external validation metrics were extracted and their median differences summarized the simulation results. The empirical standard deviations were also calculated. For the calibration slope, the Inter Quartile Range (IQR) was calculated instead of standard deviation, due to expected skewness in the distribution of slope values in smaller sample sizes.[35]

3.6 | Software and error-handling

All simulations and subsequent analyses were performed using R (version 4.0.3).[49] The simulations were carried out on a high-performance computing facility. `pmsampsize` was utilized for the sample size calculations.[50] With `MASS`, data were drawn from a multivariate distribution.[51] For Firth's correction, the `logistf` package was implemented,[52] while `glmnet` and `glmnetUtils` were used for both Lasso and Ridge regression.[53,54] CART models were built with `rpart`[27], and random forest models with `ranger`,[30] using `caret`[55] for the tuning of the hyperparameters. The `tidyverse` package[56] and accompanying packages were used for data wrangling and visualization. Regarding the latter, `facetscales` was used to manually set the scales within the facets of the box plots.[57] Some of the tables were created with `kableExtra`.[58] Details on specific package versions can be found within the requirements file available in the GitHub repository.

The estimation errors encountered within the simulations and the handling thereof, have been summarized in Table 2. Only within creating boxplots as depicted below, the upper boundaries of the calibration slope results were winsorized at 10. However, raw scores were used to calculate the median differences. A table depicting the instances of errors, specified per scenario, and the handling thereof, can be found within the supplementary material in the GitHub repository.

4 | RESULTS

Within both study 1 and 2, there was little difference between the scenarios with and without predictor selection, see figures S1-S4, S7 and S8 within supplementary figures. Therefore, the results of these studies mainly focus on the results of the scenarios using the ML_{AIC} model (Sections 4.1 and 4.2). Within study 3, the results of Firth's correction, Ridge and LASSO, and Random Forest are depicted first (section 4.3). The scenarios using CART suffered from many estimation errors (see supplementary material in the GitHub repository). Therefore, a separate section is devoted comparing its results to ML, section 4.3.1. For each study separately, the predictive performance metrics are depicted using box plots and the median differences through nested loop plots. For the sake of readability, outliers are not depicted within the box plots of this thesis. Box plots for all predictive performance metrics (Figures S1-S6) and the median differences with error bars based on the empirical standard deviation (or IQR in case of the calibration slopes, Figures S7-S10) are available in the GitHub repository.

Overarching all studies, with increasing sample size, the variability within each metric decreased, as expected (see Figure 1, 3 and 5). However, for the AUC and R^2 metrics, the internal validation strategies showed more variability than large scale external validation, while for the calibration intercept they showed less variability (see S1 for example). The amount of variability was dependent on specific scenario settings. Therefore, the next sections focus on the study and scenario specific results.

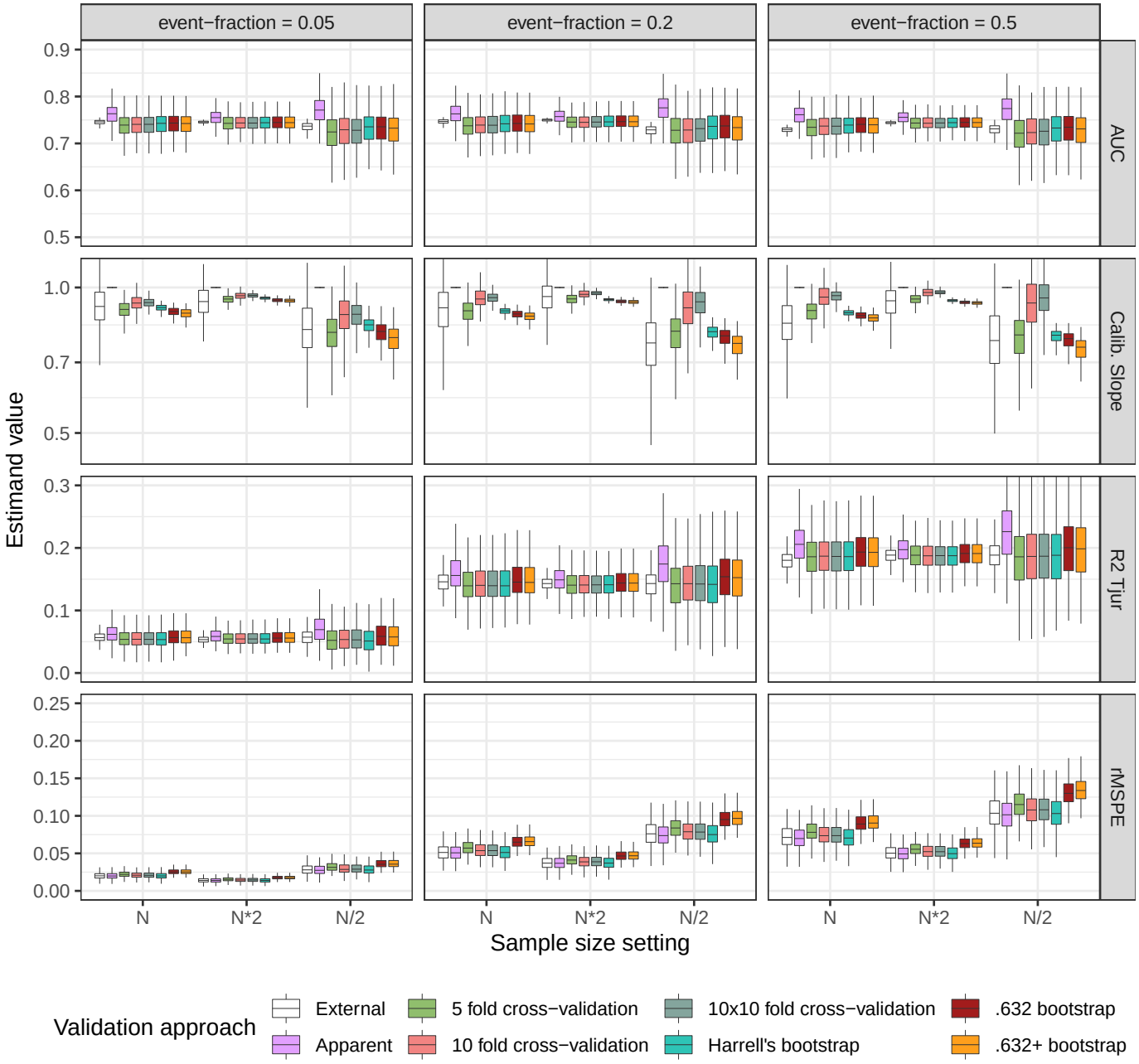

FIGURE 1 Box plots depicting the AUC, Calibration Slope (y-axis on log base 10 scale), R^2_{Tjur} and rMSPE metrics for all validation approaches of the scenarios within study 1 using the ML_{AIC} model over 500 runs. Each group of box plots belongs to a certain sample size setting, wherein the colors represent the validation approaches. The length of the whiskers is determined by 1.5 times the IQR.

4.1 | Study 1: Influence of event-fraction

Only slight variation was visible between the internal validation approaches in estimating the AUC, while obvious differences were observed within the calibration slopes (Figure 1). With decreasing sample sizes, these differences became more pronounced. In Figure 2, both 10 and 10x10 fold cross-validation generally show overestimation of the calibration slope, compared to the large scale external validation. The median calibration slope of external validation suggested that the estimated risks were overly extreme (i.e. slope < 1), however both 10 and 10x10 cross-validation were not able to portray the severity of extreme

risk estimates. In scenario 7 of Figure 2, the median difference between the calibration slopes of 10x10 fold cross-validation and external validation reached 0.24, while the difference was close to zero for .632+ bootstrap.

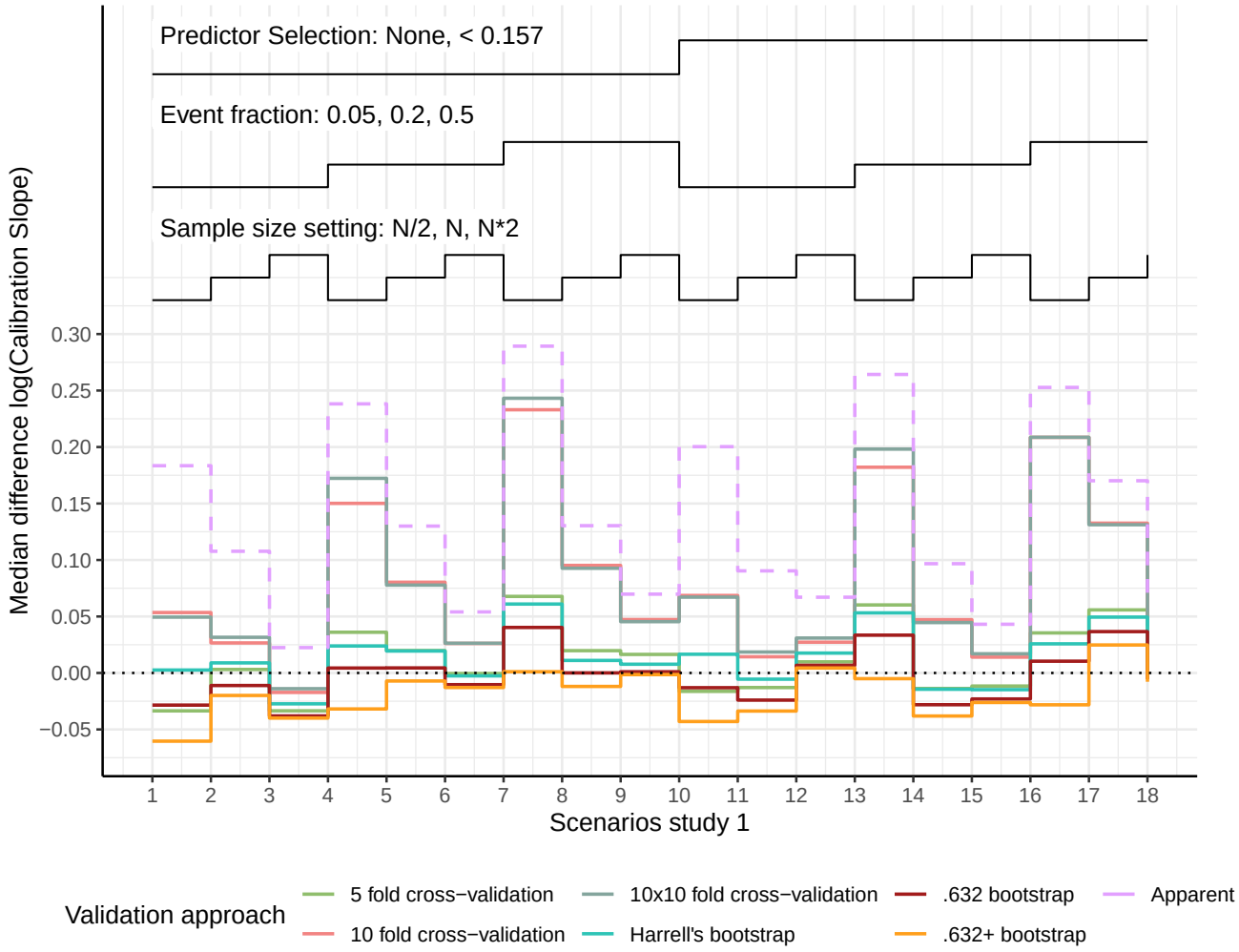

FIGURE 2 Nested loop plot depicting the median differences of the log(Calibration slopes) between the external and each of the internal validation approaches and apparent approach for study 1. The graph is based on 500 runs, using the ML and ML_{AIC} model. Each step represents a single scenario, where the color depicts the validation approach used. The dotted line shows the target value of a median difference of 0.

When the event-fraction increased, both R^2_{Tjur} and root Mean Squared Prediction Error (rMSPE) increased simultaneously regarding their estimates and variability (Figure 1). With increasing event-fraction, there is more variance to be explained and a bigger spread in estimated risks, which in turn might lead to bigger prediction errors. Regarding the rMSPE, .632 and .632+ bootstrap approaches and 5 fold cross-validation (albeit less profound) overestimated the prediction error, while the other methods approached the external validation in both their variability and estimates.

4.2 | Study 2: Influence of dimensionality and noise

The estimations of out-of-sample performance differed little between the scenarios with and without noise predictors. Therefore, only the results with 50% noise predictors are depicted within Figure 3. Similar to study 1, the calibration slopes showed some variability between the internal validation approaches. Regarding the scenarios using only 6 candidate predictors and the smallest sample size settings ($N_{min}/2 = 198$), 10 fold and especially 10x10 fold cross-validation showed a large amount of both

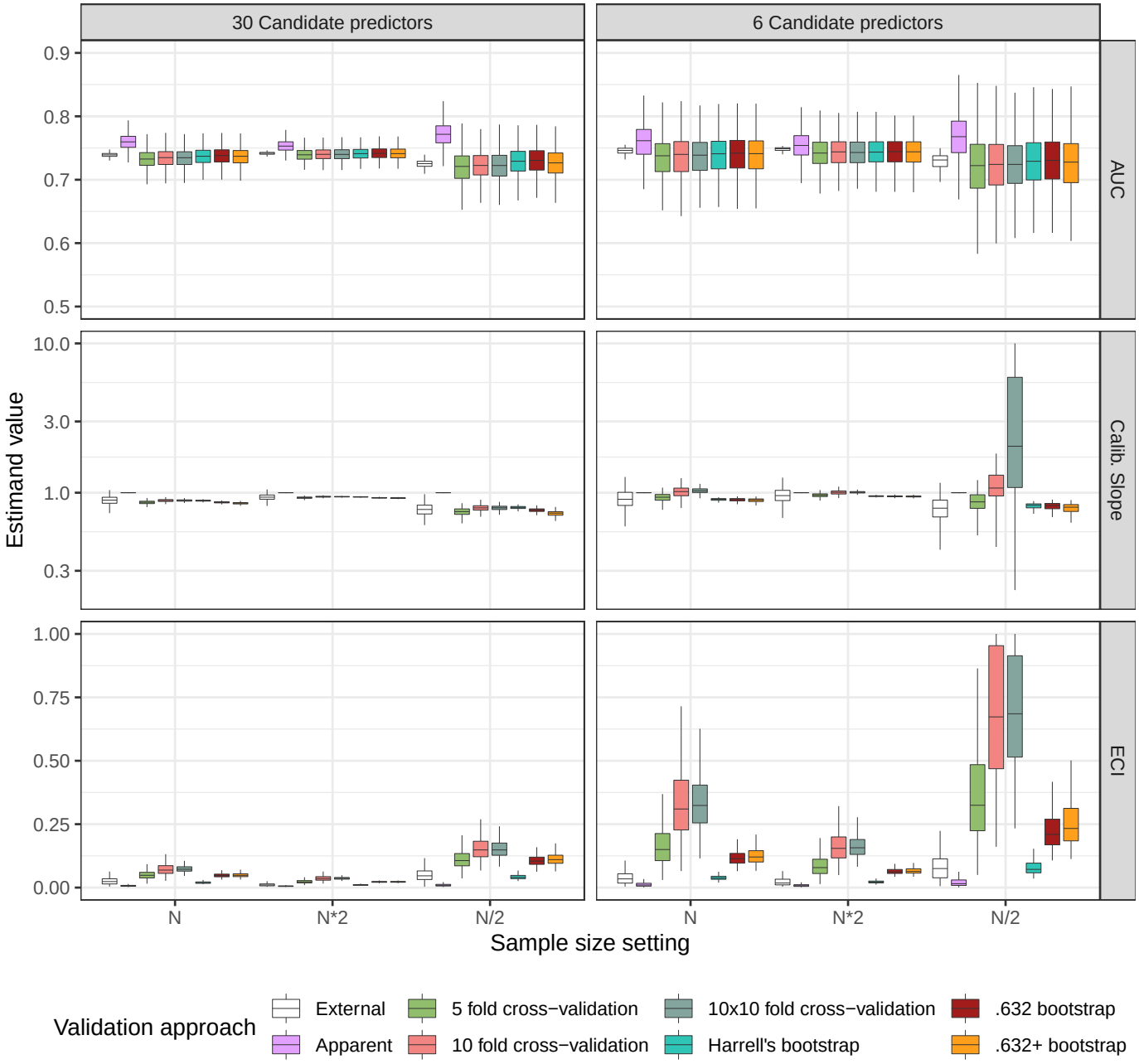

FIGURE 3 Box plots depicting the AUC, Calibration Slope (y-axis on log base 10 scale) and ECI for all validation approaches. Here, the scenarios of study 2 with 50% noise predictors using the ML_{AIC} model are portrayed over 500 runs. Each group of box plots belongs to a certain sample size setting, wherein the colors represent the validation approaches. The length of the whiskers is determined by 1.5 times the IQR.

overestimation and variability in contrast to the bootstrap results (Figure 3). In these instances, the results from each fold were tested on a small subset, in which the model may have performed poorly. Since the results of all test-sets were averaged, the final results were sensitive to these outlying values. 5 fold cross-validation suffered less from these issues, since the test-sets were larger. Generally, all approaches showed more variability in the scenarios with 6 candidate predictors (right column) than with 30 candidate predictors (left column). See S3 and S4 for the box plots of all metrics in the GitHub repository accompanying the thesis.

Some interesting results were seen for the ECI as well (Figure 4). Contrary to the estimations of the calibration slope, only Harrell's bootstrap approached the externally validated ECI, and did so quite consistently as well. In general, bootstrap approaches, and especially Harrell's bootstrap, are recommended for data sets with sample size smaller than the recommended minimum[44] and a small number of candidate predictors when using maximum likelihood logistic regression.

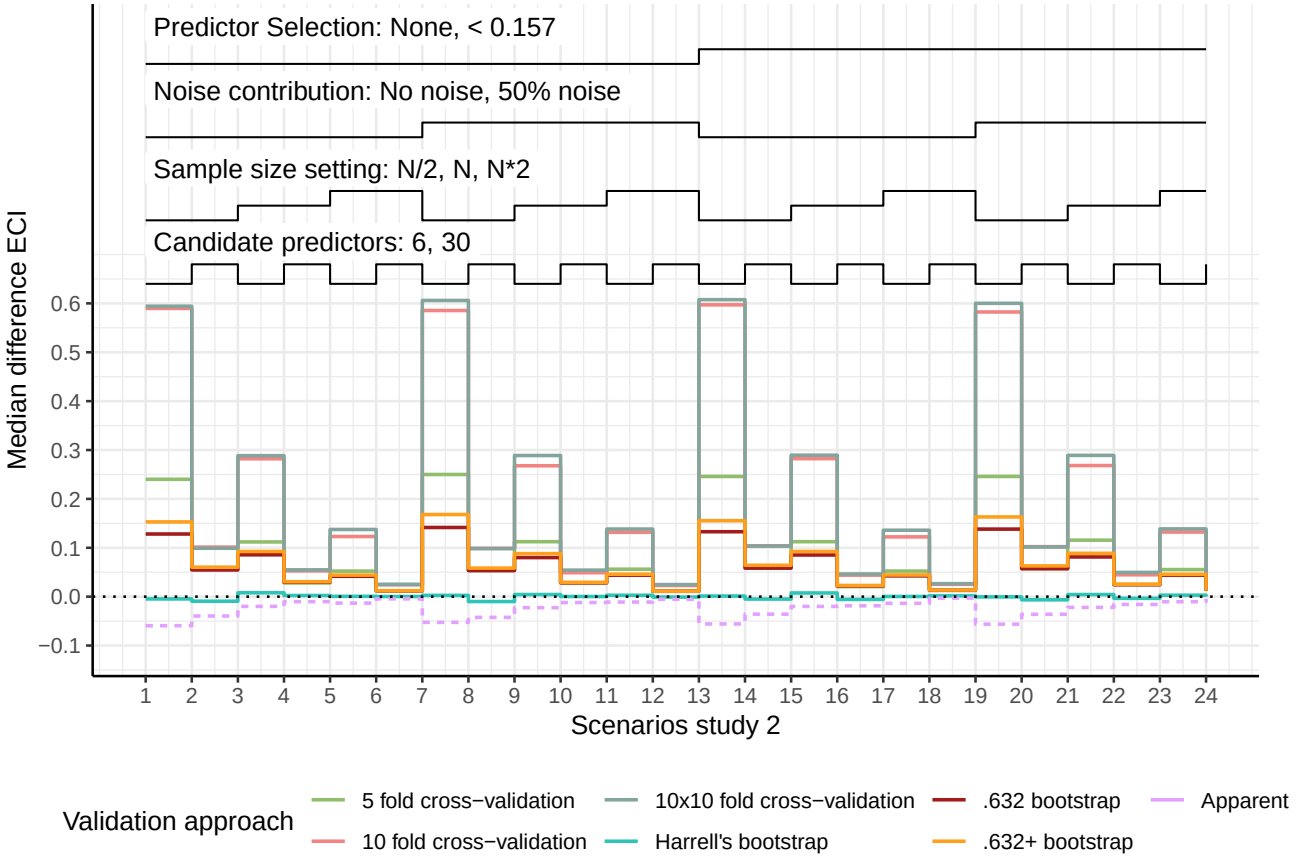

FIGURE 4 Nested loop plot depicting the median differences of the ECI between the external and each of the internal validation approaches and apparent approach for study 2, using logistic regression with ML estimation (with and without backwards elimination). The graph is based on 500 runs. Each step represents a single scenario, where the color represents the validation approach used. The dotted line shows the target value of a median difference of 0.

4.3 | Study 3: Influence of modeling algorithms

The results of Firth were comparable to those of ML, and Ridge performed similar to Lasso (for a comparison, see S5 and S6). Similarly to the previous studies, albeit less profound, both Firth and Lasso showed some variability between the validation approaches regarding the calibration slope. Moreover, two interesting results can be observed within the Estimated Calibration Index (ECI, Figure 5). Firstly, when using Ridge regression, there was more variability within all internal validation approaches, compared to logistic regression using Firth's correction. Secondly, the apparent results were too pessimistic (lower values of ECI indicate better performance), while one expects the apparent approach to be too optimistic, as is visible within Firth's correction. Similar to study 2, Harrell's bootstrap approach appears to approach the external validation results most consistently and accurately across all metrics.

However, when looking at the results of Random Forest, this is not the case (Figure 5). Both apparent and all bootstrap approaches showed severe overestimation of the model performance regarding AUC, R^2_{Tjur} (and R^2_{CS}) and prediction errors.

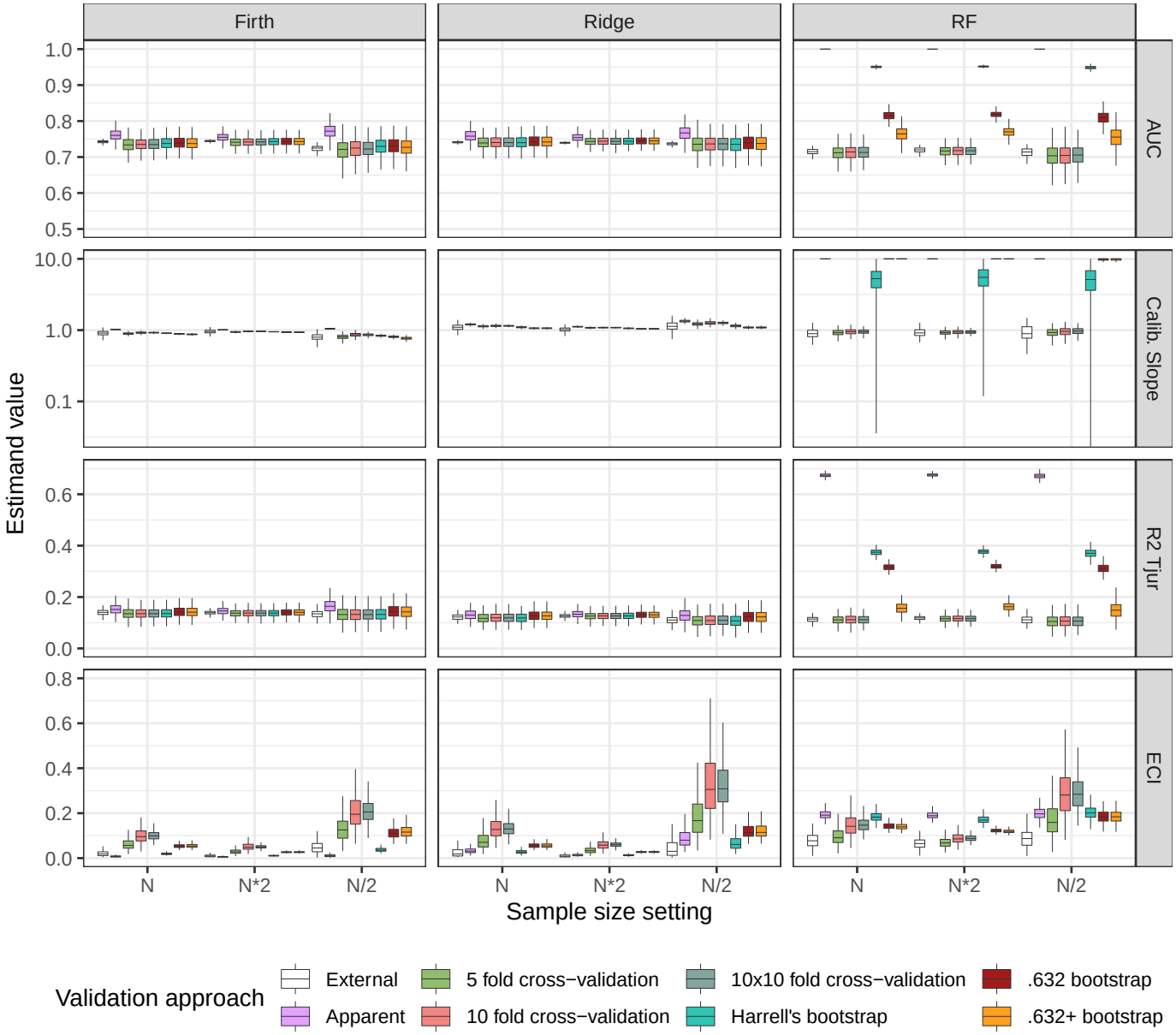

FIGURE 5 Box plots depicting the AUC, Calibration Slope (y-axis on log base 10 scale), R^2_{Tjur} and ECI for all validation approaches, for the scenarios of study 3 using Firth and Ridge regression and random forest over 500 runs. Each group of box plots belongs to a certain sample size setting, wherein the colors represent the validation approaches. The length of the whiskers is determined by 1.5 times the IQR.

Although, for R^2_{CS} and rMSPE the .632+ bootstrap approach range comes closer to those of the external validation approach, see S6. Concerning the calibration slope, apparent and bootstrap approaches suggest that the model is severely underfitting the risks, while the external estimates approach 1 (i.e. calibration slope > 1 indicates that the risks are too low for high risk patients, but too high for low risk patients). Contrarily to the results of the regression models, cross-validation approaches performed better in estimating out-of-sample performance for random forest than bootstrap approaches.

4.3.1 | CART

The results from CART suffered from some estimation issues, mostly because the results of the CART models returned probabilities that were exactly 0 or 1 (see Table 2). Table 3 shows the results for both CART and ML models, where the latter is meant as a reference.

TABLE 3 Results of CART and ML within Study 3

Model	AUC	Calibration Slope	R^2_{CS}	ECI	rMSPE
Approach	Mean (*sd*)	Median (IQR)	Mean (*sd*)	Mean (*sd*)	Mean (*sd*)
CART					
5 fold cv	0.603 (0.023)	0.468 (0.408 - 0.531)	-0.017 (0.028)	0.401 (0.238)	0.149 (0.005)
10 fold cv	0.601 (0.024)	0.536 (0.485 - 0.590)	-0.005 (0.023)	0.483 (0.298)	0.145 (0.004)
10x10 fold cv	0.6 (0.021)	0.544 (0.503 - 0.586)	-0.007 (0.013)	0.704 (0.343)	0.146 (0.004)
Harrell's bootstrap	0.539 (0.049)	0.442 (0.431 - 0.451)	-0.238 (0.038)	0.152 (0.242)	0.136 (0.007)
.632 bootstrap	0.627 (0.021)	0.437 (0.432 - 0.443)	-0.44 (0.053)	0.501 (0.138)	0.201 (0.005)
.632+ bootstrap	0.628 (0.014)	0.163 (0.151 - 0.175)	-0.743 (0.077)	0.631 (0.053)	0.232 (0.008)
Apparent	0.642 (0.048)	1 (1 - 1)	0.083 (0.031)	0.073 (0.239)	0.139 (0.007)
External	0.597 (0.031)	0.565 (0.494 - 0.624)	0.002 (0.026)	0.283 (0.196)	0.143 (0.007)
ML					
5 fold cv	0.741 (0.013)	0.93 (0.918 - 0.94)	0.116 (0.013)	0.031 (0.012)	0.041 (0.005)
10 fold cv	0.742 (0.012)	0.948 (0.939 - 0.956)	0.116 (0.012)	0.052 (0.017)	0.039 (0.005)
10x10 fold cv	0.742 (0.012)	0.949 (0.944 - 0.952)	0.116 (0.012)	0.051 (0.008)	0.039 (0.005)
Harrell's bootstrap	0.743 (0.012)	0.939 (0.935 - 0.943)	0.118 (0.012)	0.012 (0.003)	0.037 (0.006)
.632 bootstrap	0.743 (0.012)	0.927 (0.922 - 0.932)	0.117 (0.012)	0.028 (0.004)	0.046 (0.005)
.632+ bootstrap	0.743 (0.012)	0.924 (0.919 - 0.929)	0.116 (0.012)	0.028 (0.004)	0.047 (0.005)
Apparent	0.755 (0.012)	1 (1 - 1)	0.131 (0.012)	0.006 (0.003)	0.037 (0.006)
External	0.741 (0.002)	0.93 (0.89 - 0.969)	0.117 (0.003)	0.013 (0.007)	0.037 (0.006)

Note. Abbreviations: CART = Classification and Regression Trees, ML = Logistic regression using maximum likelihood without predictor selection, cv = cross-validation, AUC = Area under the (ROC) Curve, *sd* = standard deviation, IQR = Inter Quartile Range, ECI = Estimated Calibration Index, rMSPE = Root Mean Squared Prediction Error. The table is based on the scenarios with the largest sample size (ML: $N = 4256$, CART: $N = 2640$).

First and foremost, it is obvious that the CART model performed worse than logistic regression (on data that were generated under a logistic regression model). Similarly to the results of Random Forest, the cross-validation methods approached the external estimates quite closely (except for the ECI). However, especially .632+ bootstrap showed quite some deviance from the externally validated results for the calibration slope, R^2_{CS} and ECI. This can be explained by going back to its calculations (see section 2.3.2). When the difference between θ_{test} and θ_{app} becomes larger, the relative overfit, $\hat{R}$, approached 1. Consequently, the weight of the apparent performance became smaller, resulting in too pessimistic out-of-sample predictive performance. This is remarkable since the purpose of this method was to improve the .632 approach in conditions in which the .632 was optimistically biased in highly overfit situations. Here, the internal validation approaches are pessimistically biased in a highly overfit situation. As one can see, the results showed that in this special case, the .632+ performs worse than the .632 bootstrap.

Furthermore, notice the amount of negative values within R^2_{CS}, while its theoretical range goes from 0 to < 1. Negative values can appear the log-likelihood of the intercept-only model is bigger than that of the final model (i.e. the final model was worse than the intercept-only model). Meaning that random models performed better than the CART models that were built within this scenario.

5 | DISCUSSION

The aim of this thesis was to assess and compare performance of bootstrap and cross-validation based internal validation approaches. Specifically, to estimate out-of-sample performance of prediction models for binary outcomes using the same target population that was used for model development. To do so, simulation studies were performed on commonly encountered low-dimensional data settings, modeling algorithms and internal validation approaches. The key findings can be summarized in the following points.

Generally, internal validation approaches showed more variability than external validation when estimating the out-of-sample AUC, ECI and R^2 metrics, but less for the calibration intercept. Here, too little variability painted a too optimistic picture of the model performance, while too much variability might either result in needless concern, or too much optimism. The differences in variability between internal and the large sample external validation underscore that they represent two different steps within prediction model evaluation: model development and out-of-sample performance. Study 1 showed that an increase in event-fraction also increased the variability of internal validation approaches in estimating prediction error and explanatory power. Moreover, bootstrap .632 and .632+ tend to increasingly overestimate these metrics as well. Within this study, both Harrell's bootstrap and 5-fold cross-validation consistently approached the external validation performance closest. The results of study 2 indicated that for sample sizes smaller than the recommended minimal sample size (as defined by Riley et al.[44]) and only few candidate predictors, both 10 and 10x10 fold cross-validation are not suitable for internal validation. Both approaches are susceptible to outlying results within the folds, which may occur often within these settings. Perhaps using the median instead of the average over the results might provide more reliable estimates in these cases. Moreover, Harrell's bootstrap method was most efficient (i.e. consistent and relatively unbiased) in estimating out-of-sample performance measures across all scenarios of this study (Figure 3 and 4). In study 3, it became clear that differences between performance estimates derived from different internal validation approaches became more pronounced when using tree-based models, compared to regression-based. Within the results of random forest, the apparent performance was the furthest removed from the external results on many metrics. The bootstrapped estimates take the apparent performance into account, to some degree, inducing a large amount of overestimation. Moreover, the results of CART left much to be desired from a model performance perspective. Nevertheless, it did provide insight into some odd behavior of the .632+ bootstrap approach, where it performed worse than the method it was supposed to improve, .632 bootstrap in pessimistically biased highly overfit situations (section 4.3.1). On the whole, confirming the results of Kim et al.[13], cross-validation approaches are better suited to estimate out-of-sample performance than bootstrap approaches when using tree-based models.

The results presented here are consistent with previous studies advocating bootstrap approaches.[7,10,11] Of the three studies, only Tantithamthavorn et al.[11] used a random forest classifier aside from regression-based models, for which they also found .632 and Harrell's bootstrap to be unstable in estimating the calibration slope. In line with the suggestions by Van Calster et al.,[31] one should not rely on only one model performance measure, as some of the results indicated moderate discrimination, but unreliable risk estimates (Figure 3). For example, Mondol et al.[7] suggested bootstrap approaches, especially .632+, to be the most efficient when validating a prediction model with rare outcome (0.05 event-fraction) when assessing the AUC and calibration slope. The results regarding the AUC were indeed supported, but when looking at prediction error and explanatory power, both .632 and .632+ were the least efficient of all internal validation approaches. These findings concerning .632+ bootstrap, support those of of Jiang and Simon,[17] who also found its estimates to be upwardly biased in small sample cases with an event-fraction of 0.5.

Some caution is advised when considering these results. First of all, the assumptions made while generating the simulation data should be taken into account (e.g. the relations between predictors were linear without any interactions). It would be interesting to use real clinical data to compare the approaches, similar to Tantithamthavorn et al.[11] have done for defect prediction models. Moreover, one should be aware that performance measures such as the prediction errors can not be estimated in real data situation, as the data generating mechanism not known. Secondly, the tuning of hyperparameters was based on the commonly applied 10-fold cross-validation, while a different approach might have yielded better performance. Thirdly, it might be interesting to also expand the study with more internal validation approaches (such as leave-pair-out cross-validation or regular or out-of-sample bootstrap), machine learning based models and performance measures. For instance, leave-pair-out cross-validation might be interesting in light of the findings regarding the performance 10 fold and 10x10 fold cross-validation in small sample and low-dimensionality data settings. Moreover, this study focused only on low-dimensional data settings, while it might be interesting to see the performance of internal validation approaches across different $P > N$ scenarios.

Summarizing, the goal of this thesis was to equip applied prediction modelers with guidance on choosing an internal validation method, depending on their situation at hand. Previous studies mostly advocated the use of bootstrap methods for internal

validation, however this thesis has shown that this is not always the best approach. The simulation studies support the use of bootstrap, especially Harrell's bootstrap, for internal validation when using regression modeling, while showing its limitations with tree-based models. For these latter models, cross-validation approaches are better suited. Studies in which the sample size is smaller than the minimum required and where only a small number of candidate predictors exist, should refrain from using cross-validation approaches with many folds. Moreover, the relevance of externally validating clinical prediction models is accentuated, due to the differences observed in variability between both external and internal validation. The sample used for external validation allows for evaluating a model in a more variable sample, while internal validation can only use and re-use the same limited sample. They may therefore lack the ability to detect the whole spectrum of performance existing within prediction models (the extremely good and the bad). This study shows that deciding on an internal validation approach should be a careful process, as the results of different approaches may lead to different conclusions about the validity of a prediction model.

ACKNOWLEDGMENTS

I would like to thank Maarten van Smeden and Ben Van Calster, for their unrelenting support during their supervision of the thesis. Moreover, I acknowledge the helpful suggestions and feedback of Gary Collins, Richard Riley, Georg Heinze and Laure Wynants on the initial simulation protocol. Lastly, I would like to thank my peers, Paulina, Cassandra, Marco, Ruben and Zoë, for your comments and suggestions. Ethical approval for the thesis project was received from Faculty Support Office Ethics Committee: 20-0138.

ORCID

Sofie A.G.E. van den Brand https://orcid.org/0000-0002-2408-3336

Conflict of interest

The author declares no potential conflict of interests.

SUPPORTING INFORMATION

The following supporting information is available as part of the online article: The complete research archive, including a description of error handling, details on the optimization all files and code used for the creation of this thesis (i.e. the GitHub repository).

REFERENCES

1. Smeden vM, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. *Journal of Clinical Epidemiology* 2021; 132: 142–145. doi: 10.1016/j.jclinepi.2021.01.009

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

3. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. 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

4. Iba K, Shinozaki T, Maruo K, Noma H. Re-evaluation of the comparative effectiveness of bootstrap-based optimism correction methods in the development of multivariable clinical prediction models. *BMC Medical Research Methodology* 2021; 21(1): 1–14.

5. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine* 2015; 162(1): W1–W73. doi: 10.7326/M14-0698

6. Steyerberg EW. *Clinical Prediction Models. Statistics for Biology and Health. 2nd edition*. Springer . 2019.

7. Mondol MH, Rahman MS. A comparison of internal validation methods for validating predictive models for binary data with rare events. *Journal of Statistical Research* 2018; 51(2): 131–144. doi: 10.47302/jsr.2017510203

8. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research* 2017; 26(2): 796–808. doi: 10.1177/0962280214558972

9. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models : I . Development, internal validation, and assessing the incremental value of a new (bio) marker. *BMJ Heart* 2012. doi: 10.1136/heartjnl-2011-301246

10. Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, 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

11. Tantithamthavorn C, McIntosh S, Hassan AE, Matsumoto K. An Empirical Comparison of Model Validation Techniques for Defect Prediction Models. *IEEE Transactions on Software Engineering* 2017; 43(1): 1–18. doi: 10.1109/TSE.2016.2584050

12. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence* 1995.

13. Kim JH. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis* 2009; 53(11): 3735–3745.

14. Smith GCS, Seaman SR, Wood AM, Royston P, White IR. Practice of Epidemiology Correcting for Optimistic Prediction in Small Data Sets. *American Journal of Epidemiology* 2014; 180(3): 318–324. doi: 10.1093/aje/kwu140

15. Lyons MB, Keith DA, Phinn SR, Mason TJ, Elith J. A comparison of resampling methods for remote sensing classification and accuracy assessment. *Remote Sensing of Environment* 2018; 208(February): 145–153. doi: 10.1016/j.rse.2018.02.026

16. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* 2018; 180: 68–77.

17. Jiang W, Simon R. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Statistics in medicine* 2007; 26(29): 5320–5334.

18. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media . 2009.

19. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events relative to the number of predictors ,. 2015: 1–5. doi: 10.1136/bmj.h3868

20. Pavlou M, Ambler G, Seaman S, De iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine* 2016; 35(7): 1159–1177. doi: 10.1002/sim.6782

21. Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth's logistic regression with rare events: accurate effect estimates and predictions?. *Statistics in Medicine* 2017; 36(14): 2302–2317. doi: 10.1002/sim.7273

22. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 2002; 21(16): 2409–2419. doi: 10.1002/sim.1047

23. Van Calster B, Valentin L, Froyman W, et al. Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study. *BMJ (Clinical research ed.)* 2020; 370: m2614. doi: 10.1136/bmj.m2614

24. Loh WY. Fifty years of classification and regression trees. *International Statistical Review* 2014; 82(3): 329–348. doi: 10.1111/insr.12016

25. Breiman L, Friedman J, Olshen R, Stone C. *Classification And Regression Trees*. Wadsworth Belmont California . 1984.

26. Speybroeck N. Classification and regression trees. *International Journal of Public Health* 2012; 57(1): 243–246. doi: 10.1007/s00038-011-0315-z

27. Therneau T, Atkinson B. rpart: Recursive Partitioning and Regression Trees. 2019.

28. Breiman L. Random forests. *Machine Learning* 2001(45): 5–32.

29. Probst P, Boulesteix AL. To tune or not to tune the number of trees in random forest?. *arXiv* 2017; 18: 1–18.

30. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 2017; 77(1): 1–17. doi: 10.18637/jss.v077.i01

31. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. *BMC Medicine* 2019; 17(1): 1–7. doi: 10.1186/s12916-019-1466-7

32. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *Journal of Biomedical Informatics* 2015; 54: 283–293. doi: 10.1016/j.jbi.2014.12.016

33. Tabachnick G, Fidell LS. *Using Multivariate Statistics sixth edition*. Pearson . 2013.

34. Tjur T. Coefficients of determination in logistic regression models - A new proposal: The coefficient of discrimination. *American Statistician* 2009; 63(4): 366–372. doi: 10.1198/tast.2009.08210

35. Smeden vM, Moons KG, Groot dJA, 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

36. Forman G, Scholz M. Apples-to-apples in cross-validation studies. *ACM SIGKDD Explorations Newsletter* 2010; 12(1): 49–57. doi: 10.1145/1882471.1882479

37. Iba K, Shinozaki T, Maruo K, Noma H. Re-evaluation of the comparative effectiveness of bootstrap-based optimism correction methods in the development of multivariable clinical prediction models. arXiv; 2020.

38. Harrell FE, Lee KL, Mark DB. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine* 1996; 15: 361–387. doi: 10.1002/0470023678.ch2b(i)

39. Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* 1983; 78(382): 316–331. doi: 10.1080/01621459.1983.10477973

40. Efron B, Tibshirani R, Efron B, Tibshirani R. Improvements on Cross-Validation : The . 632 + Bootstrap Method Improvements on Cross-Validation : The . 632 + Bootstrap Method. *Journal of American Statistical Association* 1997; 92(438): 548–560.

41. Wahl S, Boulesteix AL, Zierer A, Thorand B, Van De Wiel MA. Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. *BMC medical research methodology* 2016; 16(1): 1–18.

42. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; 38(11): 2074–2102. doi: 10.1002/sim.8086

43. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* 2019; 110: 12–22. doi: 10.1016/j.jclinepi.2019.02.004

44. Riley RD, Ensor J, Snell KI, et al. Calculating the sample size required for developing a clinical prediction model. *The BMJ* 2020; 368(March): 1–12. doi: 10.1136/bmj.m441

45. Riley RD, Van Calster B, Collins GS. A note on estimating the Cox-Snell R 2 from a reported C statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome. *Statistics in Medicine* 2021; 40(4): 859–864.

46. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology* 2016; 74: 167–176. doi: 10.1016/j.jclinepi.2015.12.005

47. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of clinical epidemiology* 2005; 58(5): 475–483.

48. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in medicine* 2016; 35(2): 214–226.

49. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2020.

50. Ensor J, Martin EC, Riley RD. pmsampsize: Calculates the Minimum Sample Size Required for Developing a Multivariable Prediction Model. 2020.

51. Venables WN, Ripley BD. *Modern Applied Statistics with S*. New York: Springer. fourth ed. 2002. ISBN 0-387-95457-0.

52. Heinze G, Ploner M, Jiricka L. logistf: Firth's Bias-Reduced Logistic Regression. 2020.

53. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* 2011; 39(5): 1–13.

54. Ooi H. glmnetUtils: Utilities for 'Glmnet'. 2021.

55. Kuhn M. caret: Classification and Regression Training. 2020.

56. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *Journal of Open Source Software* 2019; 4(43): 1686. doi: 10.21105/joss.01686

57. Oller Moreno S. facetscales: facet grid with different scales per facet. 2021. R package version 0.1.0.9000.

58. Zhu H. kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. 2020.

How to cite this article: van den Brand, S.A.G.E. (2021), Clinical Prediction Models: A Comparison Between Cross-Validation and Bootstrap Approaches for Internal Validation, *Statistics in Medicine*, *2021;xx:x–x*.