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

DOI: xxx/xxxx

RESEARCH REPORT

The Battle of Internal Validation within Medical Prediction Models: Bootstrap vs. Cross-Validation

Sofie A. G. E. van den Brand*

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

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

Present Address
Padualaan 14, 3584 CH Utrecht

Summary

Proper evaluation of medical prediction models is essential. However, it is unclear which internal validation approach should be used to arrive at the best estimate of out-of-sample performance within different data scenarios for medical prediction models for binary outcomes. This report serves as a motivational example for my eventual thesis. Here, I use a real-world data set to investigate the differences in performance estimation between three bootstrap and three cross-validation approaches for internal validation. While using the same sample and same prediction model, there are differences between the performance measures as estimated by the internal validation approaches. This indicates that the evaluation of a prediction model can be dependent on which internal validation approach is used.

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

1 | INTRODUCTION

Clinical prediction models are important tools to support medical decision making. They provide probabilistic predictions of presence of disease (diagnosis) and future health status (prognosis) for patients. Therefore, proper construction and accompanying evaluation of these models is essential. When one uses the same data to develop and evaluate performance of the prediction model (i.e. apparent performance), performance measures tend to be optimistic [1,2]. When the model is applied to a different data set (i.e. external validation), its performance is generally worse than the apparent performance. However, a separate data set is not always available, or could be too different from the development data, resulting in too pessimistic performance estimates. Therefore, it is necessary that new prediction models are internally validated, aiming to estimate the optimism of predictive performance and adjust for overfitting, using the development data set itself [3,4].

Two of the most advocated internal validation approaches are bootstrap and cross-validation [2]. Bootstrap approaches entail resampling, reflecting the process of drawing a sample from a certain population [5]. Cross-validation approaches split up the data into mutually exclusive parts, where the model is iteratively trained on $k-1$ parts and tested on the remaining part. While some studies suggest that bootstrap is superior to cross-validation [6,7], it is the other way around in another study [8], yet again others do not clearly find one to be superior [9]. Each of these studies differ with respect to the prediction models used, sample sizes, event rates, dimensionality, and the performance measures assessed. Regarding the latter, most studies ignored the assessment of calibration measures, while calibration is known as the Achilles heel of prediction modelling [10].

For this research report I use a real-world data set to investigate the differences in the optimism-corrected performance estimation of two prediction models (Logistic regression by maximum likelihood estimation and by Firth's correction estimation) between several internal validation approaches. This will eventually serve as a motivational example for my master thesis, which

[0]**Word count: 2493 (including abstract). Supervisors: Maarten van Smeden & Ben Van Calster**

addresses the research question: *Which internal validation approach produces the best estimate of out-of-sample performance within different data scenarios for medical prediction models for binary outcomes?* Ultimately, my goal is to equip applied prediction modelers with guidance on choosing an internal validation method, depending on their situation at hand.

The remainder of the research report is built up as follows. Firstly, the approaches and performance measures are explained in detail (section 2). Thereafter, the empirical data set and procedure will be presented in section 3, followed by the results (section 4). Lastly, the findings and their implications are discussed in section 5.

2 | METHODS

2.1 | Internal validation approaches

The focus lies on developing two prediction models for $\Pr(Y = 1|X)$ using all available development data. Here, (Y) represents a binary outcome and (X) a number of predictors. The model performance is evaluated by internal validation approaches, where the training of a model is repeated on different parts of the development data. Thereafter, an optimism-corrected estimate of the performance of the model developed on all data is returned. The internal validation strategies as investigated within this report are described below.

2.1.1 | Cross-validation

Within cross-validation, one of the examined approaches is 10 fold cross-validation. Herein, the data are split up into $k = 10$ mutually exclusive parts, then 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 an optimism-corrected estimate. A second approach is 10x10 fold cross-validation, which can be used to account for possible sampling variation in the splitting of the data. Here, the splitting and train-testing of 10 fold cross-validation is repeated ten times, using a different split for each replication. The mean over all results (100 per performance measure) produces the final optimism-corrected estimates. The last cross-validation approach, leave one out cross-validation, uses a single observation as a test-set and all remaining observations as training-set. After all cases have been used as a test-set, there are as many predictions as there are cases. All predictions are then stacked into a single vector, on which performance measures are calculated.

2.1.2 | Bootstrap

The first of the bootstrap approaches to be investigated is Harrell's enhanced bootstrap, also called Harrell's bias correction. It uses the following procedure[1]: Generate B bootstrap samples by sampling with replacement from the original data. For this report, B = 2000 bootstrap samples were used for each prediction model. Let θ represent a vector of performance measures. Fit a model and obtain performance measures for each bootstrapped sample: $\theta_{1,boot-B,boot}$. Next, obtain performance measures by applying these models to the original data: $\theta_{1,o-B,o}$, 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 on the original data.

When using the previous bootstrap approach, on average 63.2% of the data in the original sample is resampled 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[11], which is the second approach to be studied here. 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: $\theta_{1,test-B,test}$. Taking the average over the performance measures, results in θ_{test}. With the latter, the optimism-corrected performance estimate can be obtained by $\theta_{.632} = 0.368 * \theta_{app} + 0.632 * \theta_{test}$.

Nonetheless, the approach above tends to be optimistically biased in highly overfit situations[12,1]. Therefore, another approach was proposed to account for this overfitting: .632+ Bootstrap, the last bootstrap approach to be considered. 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. $\hat{\gamma}$ can be defined for the dichotomous classification problem as follows [13]. Let $\hat{p}_1$ be the proportion of the observed responses for which $Y_i = 1$ and $\hat{q}_1$ the proportion of observed predictions that are 1. Then $\hat{\gamma}$ equals $\hat{p}_1(1 - \hat{q}_1) + (1 - \hat{p}_1)\hat{q}_1$. Using the relative overfit a weight, $\hat{w}$, can be derived

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

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

2.2 | Performance measures

The performance measures that are assessed for the report, represented as the vector θ in the section above, are described below. Together, these measures address the model's ability to differentiate between low and high risk subjects (i.e. discrimination) [5] and the reliability of the estimated risks (i.e. calibration) [10].

A first performance measure is the C-statistic, which assesses discriminative performance. It represents the probability that a patient without the event has a lower predicted probability than a randomly chosen patient with the event [5]. The closer this value is to 1, the greater the discriminative performance of the model. The theoretical lower bound of this measure is 0.5, indicating an uninformative model. A second performance measure within this report is Cox-Snell R^2 (R^2_{CS}), a so-called pseudo R^2 which is an overall measure of model performance. It approximates the interpretation of R^2 for linear regression, the proportion of variance explained [14]. 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.

The calibration slope and intercept are used to assess 'weak calibration' [10]. This means that, on average, the risks are not over- or underestimated and that the risk estimates are neither too extreme or moderate. The target values for weak calibration are 1 for the calibration slope (too extreme risks: < 1, too moderate risks: > 1) and 0 for the calibration intercept (overestimation of risk: < 0, underestimation of risk: > 0). The last performance measure assessed, is the Estimated Calibration Index. This can be used to compare the calibration performance between different models [15]. The ECI has a range of 0 to 100, where 0 represents the situation that a model estimates the risks perfectly.

3 | CASE-STUDY

3.1 | Real-world data

The data for this report comes from the International Ovarian Tumor Analysis (IOTA) consortium ($n = 5914$) [16], an international multi-center prospective cohort study. The primary outcome is the classification of adnexal tumors as either benign or malignant (table 1). Data collection occurred in three phases between 1999 and 2012 across 24 centers in 10 countries. Written or oral informed consent from the patients was obtained before the ultrasound scan and surgery.

TABLE 1 Descriptive statistics for primary outcome and predictors within the simplified ADNEX model [15].

	Benign	**Malignant**
Outcome, N	3983	1931
Variable, N (%) or median (IQR)		
Age (years)	42 (32 - 54)	57 (46 - 66)
Clinical center (oncology)	2179 (54.7)	1573 (81.5)
Fractions of solid parts in lesion	0.00 (0.00 - 0.01)	0.35 (0.05 - 1.00)
More than 10 cyst locules	199 (5.0)	272 (14.1)
Acoustic shadows (yes)	676 (17.0)	67 (3.5)
Ascites (yes)	64 (1.6)	656 (34.0)

Abbreviations: IQR: Inter-Quartile Range.

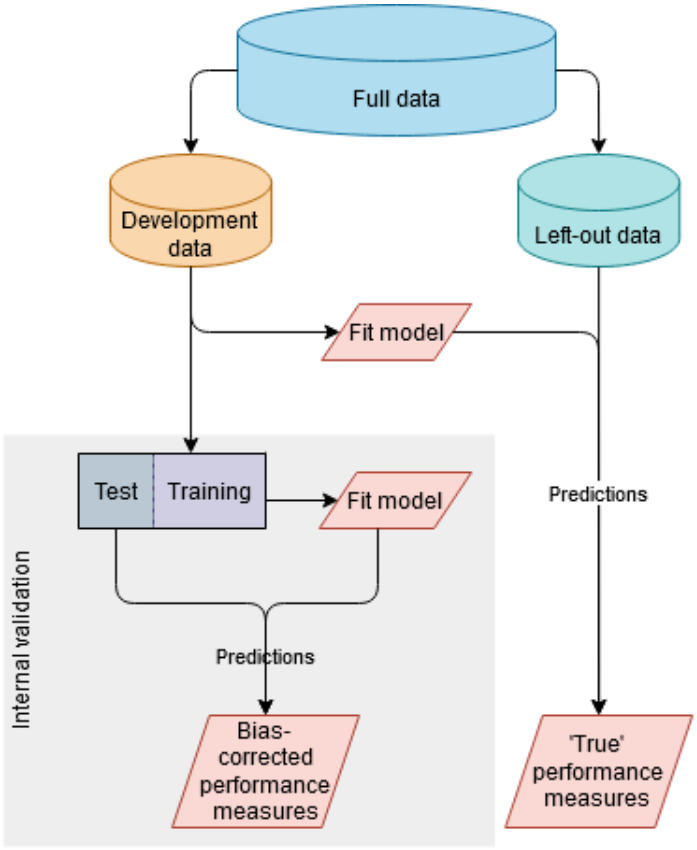

FIGURE 1 Flowchart of the approach taken to arrive at the performance measures.

Patients were excluded from the study when the suspected mass was a physiological cyst; trans-vaginal ultra sonography was refused; the patient was pregnant at the time of presentation; or when the mass was surgically removed more than 120 days after the ultrasound scan. The patients that were included, were referred to a participating centre for ultrasound examination due to a possible adnexal mass [17].

3.2 | Procedure

Multiple prediction models have already been developed to classify these adnexal masses. The ADNEX model is one of them [17], and known to perform very well (C-statistic = 0.94) [18]. For this report, only the variables as specified in table 1 are used as predictors within two prediction models. These models are logistic regression by maximum likelihood (ML) estimation and by Firth's correction (F). Regarding the latter a correction for the intercept is applied, so that predicted probabilities become unbiased [19,20].

The following procedure is used to arrive at the performance estimates (see figure 1). After splitting the data into the development data and the left-out data using random sampling, a model is trained on only the development part ($n = 211$). The size of the development data is based on sample size calculations to arrive at a preferred C-statistic of .75 [21,5]. Next, model performance is assessed on the left-out data, which mimics external validation to serve as the 'true' performance of the model. Bias-corrected performance measures are obtained for each of the internal validation approaches (gray area of figure 1). To account for possible sampling error within the initial split of the data, the process displayed in figure 1 is repeated 100 times. For each prediction model, this results in a 100 sets of both biased-corrected and 'true' performance measures.

These analyses are performed using Rstudio [22] `R version 4.0.3`. For Firth's correction the `logistf` package is implemented. The following packages are also used: `tidyverse, cvAUC, janitor, tableone, foreign, xtable, pracma, GridExtra` and `MASS`.

4 | RESULTS

Two alterations from the planned methods as explained above have been made. Firstly, while running the code to obtain the performance measures for the prediction model using maximum likelihood, warnings and errors arose, indicating non-convergence of the algorithm. In those cases, the events were perfectly separable from the non-events (i.e. separation)[23] which occurred due to the small sample size. Therefore, the results for this specific model are based on a larger sample size of $n = 550$. Secondly, the relative overfitting for the .632+ bootstrap method ($\hat{R}$, section 2), was altered slightly so that it could not become more than 1. In some cases of the ECI, the difference between θ_{app} and θ_{test} caused the $\hat{R}$ to become larger than 1.

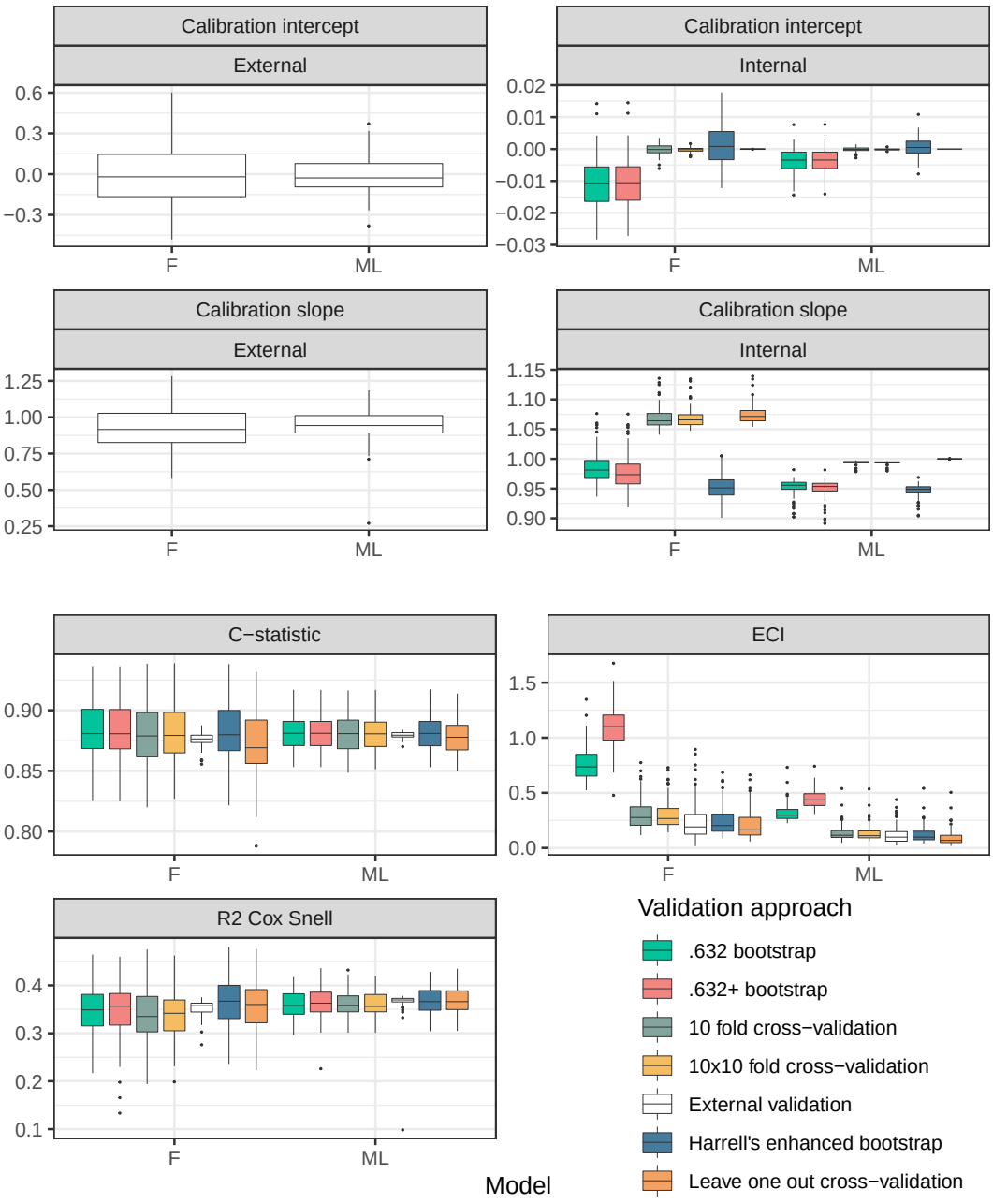

FIGURE 2 Boxplots of the performance measures using six internal validation approaches over 100 trials for each prediction model. Model estimation using Firth's correction (F) is based on development sample $n = 211$, maximum likelihood (ML) on $n = 550$. The white boxplots represent the 'true' performance measure obtained by external validation.

4.1 | Comparison validation approaches

Figure 2 depicts boxplots of the performance measures obtained within each model. The results of both the 'true' (i.e. external validation) and bias-corrected (i.e. internal validation) approaches are displayed. When applying different internal validation approaches to estimate the predictive performance of the same model obtained on the same data, the estimates should be about equal. However, the results of this study do show differences between the estimations. Within the ECI for example, both bootstrap .632 and .632+ estimate the ECI to be higher than all other approaches. As explained in section 2, these approaches take a weighted version of θ_{test}, into account. This can be very different from the apparent performance, leading to overly-corrected estimates. For the calibration intercept, bootstrap .632 and .632+ estimates again deviate from the other approaches. Moreover, all the bootstrap approaches show more variation across the 100 trials than cross-validation. Especially LOOCV shows little to no variation. For the calibration slope, cross-validation approaches center around 1.07 (risks are too moderate) while most of the bootstrap results stay below 1 (risks are too extreme). For the C-statistic, the results for LOOCV are slightly lower than the other approaches.

In general, figure 2 shows that there is less variation within the results using maximum likelihood (ML) estimation compared to Firth's correction (F). The whiskers of all validation approaches across all performance measures obtained by maximum likelihood, are shorter than those of Firth's correction. The models show about equal 'true' performance, although it was expected that Firth's correction would perform worse, due to its smaller sample size. For the C-statistic, the mean 'true' performance across all trials was C-statistic = .880 for ML and C-statistic = .876 for Firth's correction. For R^2_{CS}, the results differ between the validation approaches when Firth's correction is used, but are almost equal when using maximum likelihood estimation. When comparing the internal and external validation of both the calibration slope and intercept, there is more variation within external validation, than what is estimated by internal validation. For the C-statistic and R^2_{CS}, it is the other way around.

5 | DISCUSSION

For this research report, I evaluated both bootstrap and cross-validation approaches using a real-life data set. While these approaches should be similar in their estimations of several performance measures, the results within this report showed that there are obvious differences. This means that the estimation of model performance can be dependent on the internal validation procedure. All the while proper evaluation of such a model is essential within model development.

Nevertheless, some shortcomings of the present findings should be mentioned. First of all, the sample sizes of both prediction models were different, which clouds the comparison of performance between the two models. Moreover, the results as shown in the section above, show the variation of results between the trials, however uncertainty within each trial is not expressed. Finally, as a side note, no model selection procedures were used within the internal validation approaches, while in actual prediction modelling they should be used.

These shortcomings will be further addressed in my eventual master thesis. Moreover, to understand why one approach outperforms the other in certain conditions, I will use simulation studies. Herein, several data scenarios will be created varying in sample size, dimensionality and event rate, using a predefined discriminative performance of C-statistic = 0.75. Moreover, both machine learning models and different types of logistic regression models will be compared to each other as well. Eventually, the results of these simulations may provide the knowledge to guide applied prediction modelers on choosing an internal validation method, depending on their situation at hand.

5.1 | Ethical consent

Ethical consent for the retrospective use of IOTA data as an illustrative example for statistical studies has been approved by Ethics committee Research UZ/KU Leuven (reference S4709) and by FERB under number 20-0490.

References

1. 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.

2. 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

3. 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

4. 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

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

6. 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

7. 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

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

9. 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

10. 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

11. 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

12. 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.

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

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

15. 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

16. Timmerman D, Valentin L, Bourne T. *IOTA: The International Ovarian Tumor Analysis*. IOTA-group; : 1999.

17. Van Calster B, Van Hoorde K, Valentin L, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: Prospective multicentre diagnostic study. *The BMJ* 2014; 349(October): 1–14. doi: 10.1136/bmj.g5920

18. 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

19. Firth D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* 1993; 80(1): 27–38.

20. 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

21. 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

22. RStudio Team . *RStudio: Integrated Development Environment for R*. RStudio, PBC.; Boston, MA: 2020.

23. Van Smeden M, De Groot JA, Moons KG, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology* 2016; 16(1): 1–12. doi: 10.1186/s12874-016-0267-3