REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

PREDICTION OF DEATH ON INTERNATIONAL STROKE TRIAL DATASET WITH THE COMPARISON OF DIFFERENT STATISTICAL METHODS

Alper Umut TOSUN

MASTER OF SCIENCE THESIS

Department of Statistics

Statistics Program

Supervisor

Prof. Dr. Filiz KARAMAN

June, 2022

REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

PREDICTION OF DEATH ON INTERNATIONAL STROKE TRIAL DATASET WITH THE COMPARISON OF DIFFERENT STATISTICAL METHODS

A thesis submitted by Alper Umut TOSUN in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 28.06.2022 in Department of Statistics, Statistics Program.

Prof. Dr. Filiz KARAMAN
Yıldız Technical University
Supervisor

Approved By the Examining Committee

Prof. Dr. Filiz KARAMAN, Supervisor
Yıldız Technical University _____

Assoc. Prof. Dr. Serpil KILIÇ DEPREN, Member
Yıldız Technical University _____

Assoc. Prof. Dr. Seda BAĞDATLI KALKAN, Member
Istanbul Ticaret University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Prediction of Death on International Stroke Trial Dataset With the Comparison of Different Statistical Methods supervised by my supervisor, Prof. Dr. Filiz KARAMAN. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Alper Umut TOSUN

Signature

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Prof. Dr. Filiz KARAMAN, for always helping me and motivating me with her warm attitude.

I am indebted to Prof. Dr. İbrahim Halil TANBOĞA, the person who persuaded and pulled me into the world of statistics. He is a great teacher. I hope I will improve so much and be able to support statistics in future.

Most of all, I am thankful for my family for supporting me throughout my life. I could do nothing without their help.

Alper Umut TOSUN

TABLE OF CONTENTS

LIST OF SYMBOLS

α	Alpha
β	Beta coefficient
df	Degree of freedom
ε	Error term
R^2	R-squared
$\hat{y}$	Predicted value of dependent variable

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AUC	Area Under the Curve
CART	Classification and Regression Trees
CT	Computer Tomography
DT	Decision Tree
EPV	Event per Variable
GAM	Generalized Additive Model
GBM	Gradient Boosting Machine
GLM	Generalized Linear Model
IST	International Stroke Trial
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Logistic Regression
ML	Machine Learning
MSE	Mean Squared Error
NA	Not Available
NN	Neural Network
OLS	Ordinary Least Squares
RCS	Restricted Cubic Spline
RF	Random Forest
ROC	Receiver Operating Characteristic
RSS	Residual Sum of Squares
SVM	Support Vector Machine
XGB	Extreme Gradient Boosting

LIST OF FIGURES

LIST OF TABLES

LIST OF MAPS

ABSTRACT

Prediction of Death on International Stroke Trial Dataset with the Comparison of Different Statistical Methods

Alper Umut TOSUN

Department of Statistics

Master of Science Thesis

Supervisor: Prof. Dr. Filiz KARAMAN

In this study, regression and machine learning methods were applied for prediction of death at sixth month, from stroke patients from different countries. Then, internal and external validities of the models with performance measures were compared.

International Stroke Trial dataset (n=19435) was divided into three after missing values were removed (n=18408): "UK data" (n=5817) including UK and Ireland patients, "non-UK data" (n=9955) including all European patients except UK and Ireland patients, and "non-EU data" (n=2636) including world patients except Europe. Samples with different sizes from UK data created the train sets. Test sets were consisted of non-UK data and non-EU data. The 6 different regression methods applied were logistic regression without restricted cubic splines (RCS), logistic regression with RCS, generalized additive model (GAM), penalized regression models (ridge, lasso and elastic net regression). Five different machine learning methods applied in the study were support vector machines (SVM), random forest (RF), neural network (NN), gradient boosting machines (GBM) and extreme gradient boosting (XGB). The sizes of train sets were consisted of 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500 and 3000, respectively. After event per variables (EPV) were determined for each sample size, regression and machine learning models were developed for each sample size.

Performance of each model was evaluated according to the validities of the model in test sets. R-squared, Brier score, calibration (intercept and slope) and discriminative abilities (AUC) were found and compared as performance measures. As a result, when traditional regression models and machine learning methods were compared, it is observed that in external validations, performances of regression models were better than machine learning methods, especially for higher EPV values. While EPV<30, all models were obviously more unstable in machine learning methods.

Keywords: EPV, regression, machine learning, calibration, validation.

ÖZET

International Stroke Trial Veri Setindeki Ölüm Tahmininin Farklı İstatistiksel Yöntemlerle Kıyaslanması

Alper Umut TOSUN

İstatistik Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Prof. Dr. Filiz KARAMAN

Bu çalışmada, farklı ülkelerdeki akut inme hastalarının altıncı aydaki olay yaşama (mortalite) olasılıklarını tahmin etmek amacıyla regresyon modelleri ve makine öğrenmesi yöntemleri kullanılmıştır. Sonrasında, modellerin performans ölçütleri ile, iç ve dış coğrafi geçerlilikleri (validasyon) değerlendirmeye alınmıştır.

Hastaların verisi (n=19435), eksik veriler çıkartıldıktan sonra (n=18408) üçe ayrılmıştır: Birleşik Krallık ve İrlanda hastaları (Britanya verisi, n=5817), Birleşik Krallık dışında bulunan Avrupa ülkeleri hastaları (Avrupa verisi, n=9955) ve Avrupa dışındaki dünya ülkeleri hastaları (Dünya verisi, n=2636). Britanya hastalarından elde edilen farklı sayılardaki örneklemler öğrenim veri setini oluştururken Avrupa ve Dünya verileri ayrı test veri setlerini oluşturmuştur. Lojistik regresyon, kısıtlı kübik spline (RCS) kullanılarak oluşturulmuş lojistik regresyon, cezalı regresyonlar (ridge, lasso, elastik-net), genelleştirilmiş toplamsal model, makine öğrenmesi yöntemleri (destek vektör makineleri, gradient boosting, rastgele orman, yapay sinir ağları ve XGB) ile öğrenim veri setinde tahmin modelleri geliştirilmiştir. Öğrenim veri seti içerisinde sırasıyla 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500 ve 3000 sayılı örneklemler oluşturularak farklı "değişken başına olay" değerlerine (EPV) ayrılmışlardır. Modeller hem birbirleri arasında hem de EPV değerlerine göre kıyaslanmıştır.

Her bir modelin performansı, test veri setlerindeki geçerliliğine bakılarak değerlendirilmiştir. Performans ölçütleri için R^2, Brier skoru, kalibrasyon (kalibrasyon kestirimi ve kalibrasyon eğimi) ve diskriminasyon ölçütleri (AUC veya C istatistiği) bulunmuş ve karşılaştırılmıştır. Sonuç olarak, makine öğrenme yöntemleri ve geleneksel regresyon yöntemleri kıyaslandığında, özellikle yüksek EPV değerlerinde regresyon modellerinin dış geçerlilik performansları makine öğrenme yöntemlerine göre daha iyi sonuçlar vermiştir. EPV<30 iken, tüm regresyon performanslarının makine öğrenme yöntemlerine kıyasla daha stabil olduğu gözlemlenmiştir.

Anahtar Kelimeler: EPV, regresyon, makine öğrenmesi, kalibrasyon, validasyon.

1
INTRODUCTION

Clinical prediction models combine some characteristics (of the patient, the disease, or treatment) to predict a diagnostic (condition in the present) or prognostic (expected development for the condition) outcome. Typically, a limited number of predictors which are between 2 and 20 are considered [1]. Prediction is an estimation problem. Also, prediction is about testing of hypotheses. When we want to know the most important predictors in a certain disease, or the relation between predictors (e.g., correlation) we need prediction. Statistical models for prediction have traditionally been discerned in main classes: regression, classification, and neural networks. Machine learning, refering to the latter two categories deals with less assumptions [5,1].

For prediction models, the first goal is that predictions need to be reliable for the setting where the data came from internal validity (or reproducibility). External validity (or transportability) is also important when it comes to general applicability of prediction models [1,46].

1.1 Literature Review

There are many articles related to the same data set or methods which were used in this study. Many of the articles were helpful while creating the flow chart of study, visualizing the results, or understanding the main idea behind the work done. Ma and Liu [40] predicted possible outcomes of death (8 categories) of the patients from International Stroke Trial (IST) by focusing on 14 features and around 4000 patients. Supervised and unsupervised machine learning methods were applied. It is demonstrated how Naïve Bayes and polynomial support vector machines leads to maximal outcome prediction accuracy of 40%, and 56% respectively. Comparing to the average predictive accuracy following randomization (12.5%), the best algorithm achieved up to a 4.5 fold prediction accuracy increase, which carried immense clinical utility in improving a patient's chance of survival and quality of life. With the combination of unsupervised learning algorithms such as k-means and supervised outcome data, they also built canonical "profiles" of the most common patients doctors are likely to encounter following initial stroke attack.

Austin and Steyerberg [41] aimed to predict 1-year mortaliy using 11 variables. They conducted an extensive set of empirical analyses to examine the effect of the number of events per variable (EPV) on the relative performance of three different methods for assessing the predictive accuracy of a logistic regression model. Use of apparent performance resulted in an optimistic assessment of model performance until EPV=100. Lowest mean squared error (MSE) was matched with a bootstrap-corrected approach. By split-sample assessment, greatest variability was observed in the prediction model which also resulted too pessimistic an assessment of model performance.

Smeden et al. [39] presented an extensive simulation study in which they studied the influence of EPV, events fraction, number of candidate predictors, the correlations and distributions of candidate predictor variables, area under the ROC curve, and predictor effects on out-of-sample predictive performance of prediction models. The out-of-sample performance (calibration, discrimination, and probability prediction error) of developed prediction models was studied before and after regression shrinkage and variable selection. They found that EPV does not have a strong relation with metrics of predictive performance, and it is not an appropriate criterion for binary prediction model development studies. They suggested that out-of-sample predictive performance can better be approximated by considering the number of predictors, the total sample size and the events fraction.

Członkowska et al. compared Polish patients to patients from other countries in IST to determine why Polish patients had higher mortality than average [32]. Age, gender, presence of atrial fibrillation (AF), conscious level, neurological deficit, cause of death and aspirin use in the 3 days were the variables. A logistic regression analysis revealed that an increased risk of death within 14 days in Poland compared to other IST patients was present among patients younger than 75 years, females, patients who were alert at onset, with partial anterior circulation syndrome or lacunar syndrome and without AF. High fatality in the early stages of ischaemic stroke in Poland seems to be caused by high mortality among patients. It is concluded that with better prognosis, and attention the mortality rate can be lower.

Ploeg et al. aimed to study the predictive performance of different modelling techniques in relation to the effective sample size [23]. The comparison was between 3 modern modelling techniques (SVM, NN, and RF) and 2 classical techniques (logistic regression and CART). They created three large artificial databases with 20 fold, 10 fold and 6 fold

replication of subjects, where dichotomous outcomes according to different underlying models were generated. They applied each modelling technique to increasingly larger development parts (100 repetitions). AUC indicated the performance of each model in the development part and in an independent validation part. Stability in AUC was reached by LR at approximately 20-50 events per variable, followed by CART, SVM, NN and RF models. Optimism decreased with increasing sample sizes and the same ranking of techniques. Since the RF, SVM and NN models showed instability and a high optimism even with >200 events per variable, it can be concluded modern modelling techniques need to be used with very large datasets.

1.2 Objective of the Thesis

The aim of the thesis is to compare the performances of 11 methods in total, including 6 different regression methods and 5 different machine learning methods in external validation. For (almost) each method, 19 different EPV values based on different sample sizes from training set will be created, then models will be developed for different EPV values. Also, each model created based on EPV will be internally and externally validated on test sets. Results of R-squared, Brier score, calibration intercept, calibration slope, and AUC values will be obtained by R programming language and discussed.

1.3 Hypothesis

This study is designed to assess the hypothesis that traditional regression methods will outperform machine learning methods in external validation.

2
GENERAL INFORMATION

2.1 Introduction to Regression Models

Regression models share a general form:

$$response = weight_1 * predictor_1 + weight_2 * predictor_2 + \ldots + weight_k * predictor_k + error\ term \quad (2.1)$$

The variable y to be explained is called the dependent (response) variable. In medical literature, outcome (endpoint) is the name of binary dependent. The factors that explain the dependent variable are called independent variables, and the remaining variables, generically called covariates. The unique function of these covariates is generally to adjust for imbalances that may be present in the levels of the explanatory variable. When the main goal is the identification of the predictors for the dependent variable, every independent variable becomes significant [3].

2.1.1 Linear Regression

The main assumption is the linearity of the relationship between a continuous dependent variable and predictors [3]. For example, if there is a linear relationship between age and 5-year mortality, the average change in mortality per unit change in age can be estimated using linear regression. This estimation task is accomplished by obtaining specific values (estimates) for the 'unknowns' (parameters) of the specific regression function [8].

Ordinary least-squares method (OLS) is the most commonly used method in linear regression. OLS estimates parameters by minimizing the squared discrepancies between observed data on the one hand, and their expected values on the other [3]. Graphically this corresponds to finding a line on the cartesian axes passing through the observed points. The observed measures are regressed by minimizing the distance of the points from the line of the average values [8]. In Figure 2.1, the best fitting line of the response is shown using OLS with an example.

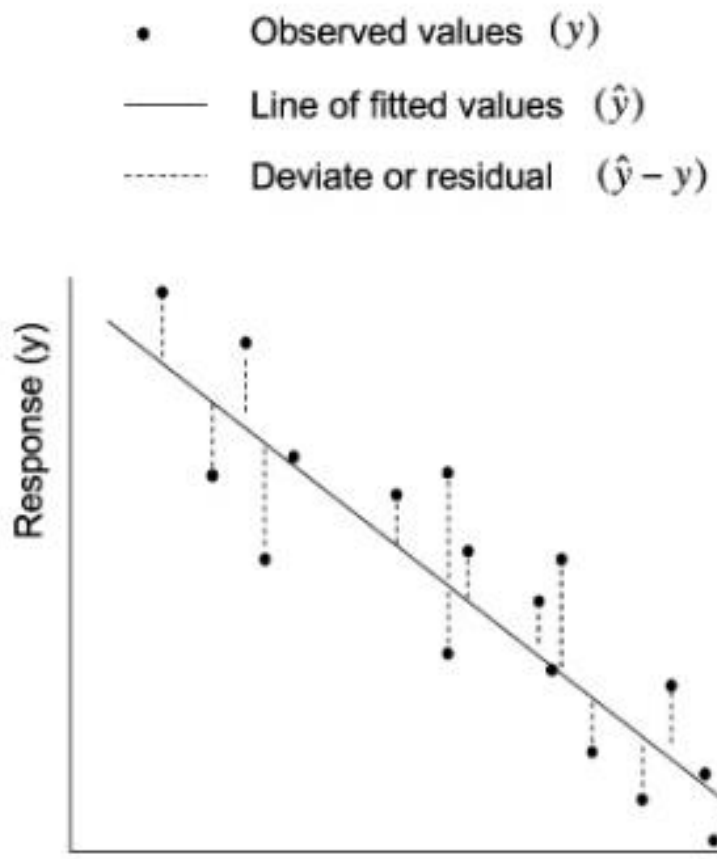


Figure 2.1 Linear model fit with an example by using OLS [8]

2.1.2 Logistic Regression

Logistic regression analyzes the relationship between multiple independent variables and a categorical dependent variable. It estimates the probability of occurrence of an event by fitting data to a logistic curve. There are two types of logistic regression: binary logistic regression and multinomial logistic regression. Binary logistic regression is used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and includes more than two categories, multinomial logistic regression becomes the choice [3].

In linear regression, the outcome is continuous. It can have any one of an infinite number of possible values. In logistic regression, the outcome can only have a limited number of possible values. In Figure 2.2, linear and logistic regression model fits on given binary data are compared.

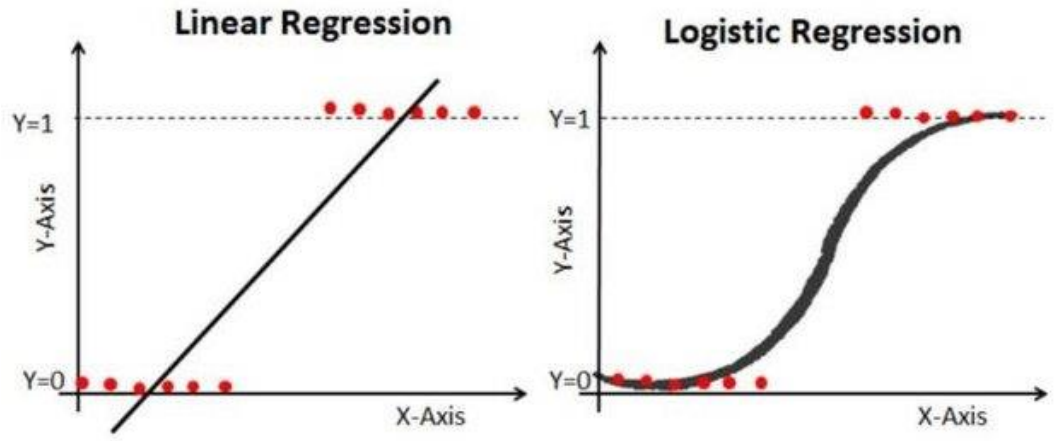


Figure 2.2 Linear regression versus logistic regression [15]

For binary illness measures, logistic regression is the most used modeling procedure used to analyze epidemiologic data. To understand the reason for the popularity of logistic regression, we first should know about the logistic function which describes the mathematical form on which the logistic model is based [9]. The values of this function f(z) is shown in Figure 2.3. On the right side of the figure, when z is +∞, f(z) equals 1. When z is -∞, f(z) then equals 0.

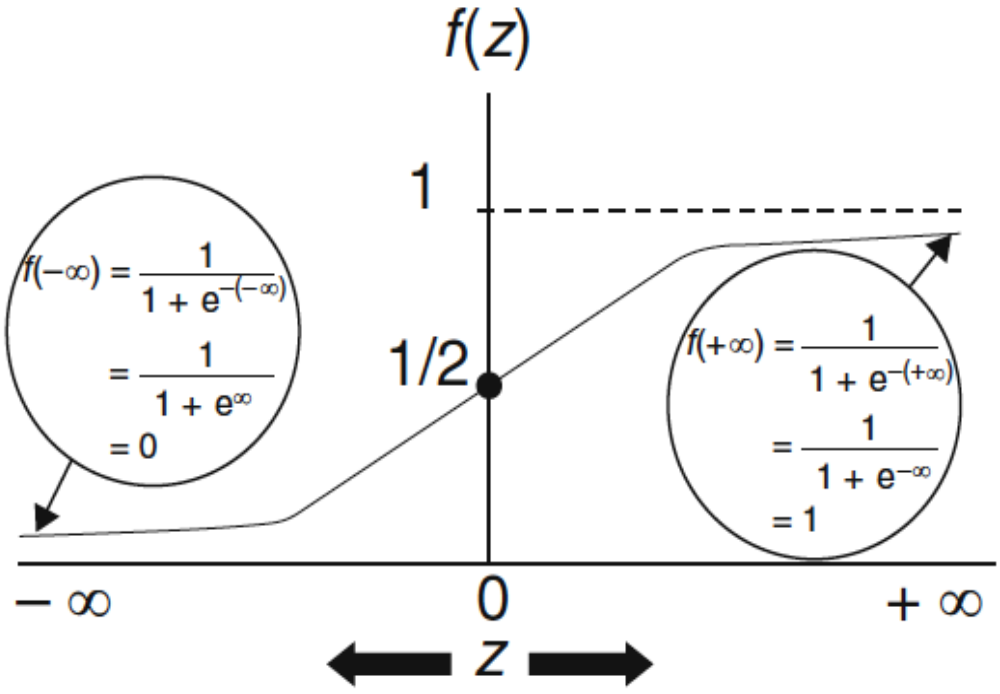


Figure 2.3 Logistic function [9]

In epidemiologic terms, such probability gives the risk of an individual getting a disease. Moreover, the logistic model is built to ensure for any risk estimate we have, it will be always between 0 and 1. For other possible models, this is not always valid [9].

As shown in the Equation (2.2), z is an index that combines Xs.

$$z = \alpha + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + \beta_k * X_k \tag{2.2}$$

By substituting the right hand-side of the formula given in (2.2), we get the expression of f(z) as given in Equation (2.3):

$$f(z) = \frac{1}{1 + e^{-(\alpha + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + \beta_k * X_k)}}, or \ \frac{1}{1 + e^{(-z)}} \tag{2.3}$$

If we put the expression of f(z) in epidemiologic context as in Equation (2.4), we may say we determine the disease status according to observed independent variables. Disease status "y" is either 1 if "with disease" or 0 if "without disease".

$$P\big(y = 1 | X_1, X_2, \ldots X_k\big) = \frac{1}{1 + e^{-(\alpha + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + \beta_k * X_k)}} \tag{2.4}$$

For notational convenience, the probability statement which is the left side of Equation (2.4), is denoted simply as P(X), as given in Equation (2.5) [9]:

$$P\big(y = 1 | X_1, X_2, \ldots X_k\big) = P(X) \tag{2.5}$$

Thence the logistic model may be written as in Equation (2.6):

$$P(X) = \frac{1}{1 + e^{-(\alpha + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + \beta_k * X_k)}} \tag{2.6}$$

2.1.3 Splines, Cubic Splines and Restricted Cubic Splines

Spline functions which are piecewise polynomials are used for fitting curves. They are polynomials within intervals of X that are connected across different intervals of X. Splines have primarily been used in the physical sciences, to approximate a wide variety of functions [2]. The use of splines for regression is similar to a draftsman's spline. A set of knots divide the range of values of the predictor. Separate regression lines or curves are fit between the knots. Number of the knots and their positions should be chosen. The degree of polynomial between the knots should also be determined. Fitting the polynomials are not enough. They should also join smoothly. Thus, a spline function is a set of smoothly joined piecewise polynomials. "Smoothly joined" means that for

polynomials of degree *n*, both the spline function and its first *n-1* derivatives are continuous at the knots [10].

Fitting a polynomial of degree *n* requires $k + n + 1$ regression parameters (including the intercept) if *k* knots are used. In practice, cubic splines which are polynomial of degree 3 are usually used, requiring $k + 3$ coefficients in addition to the intercept, compared to only 1 coefficient to fit a linear model. With this smallest degree of polynomial, sufficient flexibity is provided for fitting data. No high degree of freedom as in order splines is needed. Cubic splines are smooth in appearance because of first and second derivatives being continuous at the knots. Restricted cubic splines are used to avoid poor behaviour of cubic splines at the two tails. If the splines of a cubic spline are constrained to be linear in the two tails, it is called restricted cubic spline. Restriction does not only provide a better fit to the data, but also has the effect of reducing the degrees of freedom. In a regression analysis with restricted cubic splines, $k - 1$ degrees of freedom (the linear variable *x* and $k - 2$ piecewise cubic variables) are applied in addition to the intercept [10].

Table 2.1 Procedures to deal with continuous predictors in prediction models [1]

Procedure	**Characteristics**	**Recommendation**
Dichotomization	Simple, easy interpretation	Bad idea
More categories	Categories capture prognostic information better, but are not smooth; sensitive to choose of cut-points and hence instable	Primarily for illustration, comparison with published evidence
Linear	Simple	Often reasonable as a start
Transformations	Log, square root, inverse, exponent, etc.	May provide robust summaries of non-linearity
Restricted cubic splines	Flexible functions with robust behaviour at the tails of predictor distributions	Flexible descriptions of non-linearity
Fractional polynomials	Flexible combinations of polynomials; behaviour in the tails may be unstable	Flexible descriptions of non- linearity

Restricted cubic spline is one of the procedures to deal with continuous predictors in prediction modelling as given in Table 2.1. Specification of a spline uses the following notation as in (2.7):

$$u+ = u \text{ if } u > 0;$$
$$u+ = 0 \text{ if } u \leq 0 \qquad (2.7)$$

If the k knots are placed at $t_1 < t_2 < \ldots < t_k$, then for a continuous variable *x*, a set of $(k-2)$ new variables are created as in Equation (2.8):

$$x_i = (x - t_i)_+^3 - \frac{(x - t_{k-1})_+^3 (t_k - t_i)}{t_k - t_{k-1}} + \frac{(x - t_k)_+^3 (t_{k-1} - t_i)}{t_k - t_{k-1}}, i = 1, \ldots, k-2 \quad (2.8)$$

Thence, the original continuous predictor has been enhanced by introducing a set of new variables, which are linear in the regression coefficients. So the model can be fitted using the regular regression procedures, and conclusions can be drawn as usual. Particularly, non-linearity can be tested by comparing the log-likelihood of the model containing the new variables with the log-likelihood of a model containing *x* as a linear variable. Restricted cubic splines can be used in any type of regression [2,10].

2.2 Penalized Regression Models

The main aim of penalized estimation is to reduce the variance and to improve prediction. One of the popular methods is Ridge regression. In Ridge, sum of squares of regression coefficients are shrunk towards zero. Least Absolute Shrinkage and Selection Operator, or "LASSO" is the other popular one which shrinks the sum of absolute values of regression coefficients toward zero.

Elastic net is also another popular method which is a combination of Ridge and LASSO. These methods were principally developed to deal with high dimensional correlated data. More recently, penalized estimation techniques are also discussed for applicability in low dimension data [12].

2.2.1 Ridge

A model is fit to a set of training data in ridge regression. Penalized RSS is minimized. A penalty (*t*) is imposed to constraint the size of coefficients. Bias is introduced into the model with this constraint. More stable coefficients are produced in the model by applying a bias-variance tradeoff. The parameter *t* which controls the degree of shrinkage is always greater than or equal to 0. In this model, y_i is the elemental abundance in sample *i* and x_{ij} is the intensity at wavelength *j* of sample *i*. Larger *t* means greater shrinkage, and greater bias. Tuning *t* is the key to obtain optimal bias and variance balance in the model [5,11]. The ridge penalty is defined as in Equation (2.9):

$$\hat{\beta}^{ridge} = argmin_\beta \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, subject\ to \sum_{j=1}^{p} \beta_j^2 \leq t \qquad (2.9)$$

2.2.2 Lasso

Main procedure in LASSO is to shrink some of the coefficients, and to set most other ones to zero. This is known as the sparsity principle, where it is assumed that a smaller subset of predictor variables drives prediction results. Other coefficients can be excluded from the model without losing a remarkable performance. At the end, interpretability of the model increases by turning a large model into a sparse model. The lasso is related to backward-stepwise selection, a process starting with the full model and deleting the least effective predictor sequentially. Even if backward-stepwise selection produces a sparse interpretable model, the model in process may not have reliable predictive power. This is because coefficients are excluded from the model in an iterative process, with no reconsideration of inclusion. Thence, coefficients that are actually important for the model might be dropped in the early process by mistake. Ridge does not produce a sparse model, conversely. Ridge just shrinks the coefficients, and is more stable. To provide a sparse, nearly stable model, the lasso merges the desirable features of backward-stepwise selection and ridge regression [5,11].

In the following model, y_i is the elemental abundance in sample *i* and x_{ij} is the intensity at wavelength *j* of sample *i.* The lasso penalty is defined as in Equation (2.10):

$$\hat{\beta}^{lasso} = argmin_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 , subject\ to\ \sum_{j=1}^{p} |\beta_j| \leq t \qquad (2.10)$$

The shrinkage parameter *t* controls the amount of shrinkage in the model as explained before. Lasso does not perform well in the "*p>>N"* case, even though it is sparse. Besides that, for a group of highly correlated variables, the lasso indiscriminately chooses a variable from that group, then leaves out potentially valuable information for model predictions [5,11].

2.2.3 Elastic Net

Elastic net regression is an amalgam of lasso and ridge regression. The sparse properties of lasso regression and the stability of ridge regression in the *p>>N* case are retained.

The coefficients of the model have the following form as given in (2.11), subject to choices of α and *t* constraints; y_i is the elemental abundance in sample *i* and x_{ij} is the intensity at wavelength *j* of sample *i* [5,11]:

$$\hat{\beta}^{elastic.net} = argmin_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

$$subject\ to \sum_{j=1}^{p} \left(\alpha |\beta_j| + (1-\alpha)\beta_j^2 \right) \leq t \qquad (2.11)$$

Averaging of highly correlated features are promoted by the second term of the elastic net penalty. In the elastic net regression, information about all the features is retained in the cluster by averaging them, unlike the lasso, which selects a feature indiscriminately from a cluster of highly correlated features to represent in the model. The sparse nature of the model is controlled by the first term in the penalty, which promotes a sparse solution in the feature coefficients [5,11].

2.3 Generalized Additive Models

Generalized Additive Model (GAM) is a flexible and widely used method for conducting non-linear regression analysis. It relaxes the usual parametric assumptions and enables us to expose structure in the relationship between the response and explanatory variable in exponential family [14,20,21].

A non-linear f_j is fit to each X_j, so non-linear relationships that standard linear regression will miss can be automatically modeled in GAM. Thus, there is no need to manually try out different transformations on each variable separately [6].

Since the model is additive, the effect of each X_j on Y individually can be examined while holding all of the other variables fixed. Therefore, GAMs provide a handy representation if we are interested in inference. Moreover, non-linear fits can potentially make more accurate predictions for the target Y [6].

The primary limitation of GAMs is that the model is restricted to be additive. Significant interactions can be missed with many variables. However, interaction terms can be manually added to the GAM model by including additional predictors of the form $X_j \times X_k$ as with linear regression [6].

To understand the structure of GAM, we can start with simple multiple regression formula as given in Equation (2.12):

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_1 x_1 + \epsilon_i \qquad (2.12)$$

As in the Equation (2.13), instead of multiplying each variable $x1, x2, \ldots, xn$ with coefficients $\beta 1, \beta 2, \ldots, \beta n$, some functions $f1, \ldots, fn$ can be used to model the relationship between predictors and target variable:

$$\hat{y} = f_1(x_1) + \cdots + f_n(x_n) + \epsilon_i \tag{2.13}$$

Thence, non-linear functions add possibility to model each predictor's contribution to the response variable [6].

2.4 Introduction to Machine Learning

Machine learning, which is an area of artificial intelligence that enables modeling by learning from data has been the subject of many studies in recent years. Firstly in 1959, Arthur Samuel introduced the concept of machine learning [22]. Classification, understanding the relationship between variables in the data set, image processing, clustering and estimation can be done by using machine learning methods. Machine learning methods are used in various fields such as health, economy, security, meteorology, and fine arts.

2.4.1 Supervised Learning

Learning from training data is the procedure in supervised learning. Pairs of input objects and desired outputs create training data. The output of the function can be a continuous value. Also, it can predict a class label of the input object. The duty of the supervised learner is to predict the value of the function for any valid input object after experiencing a number of training examples. The learner has to generalize from the presented data to unknown situations in a reasonable way to achieve this [24].

2.4.2 Unsupervised Learning

Manual labels of inputs are not used in unsupervised learning. It is different from supervised learning where learning is observed using a set of human prepared examples, like regression or classification. Here we are only given the Xs and some feedback function on our performance. The training set of vectors we have are without function values in unsupervised learning. Partitioning the training set into subsets is the problem in this instance [24].

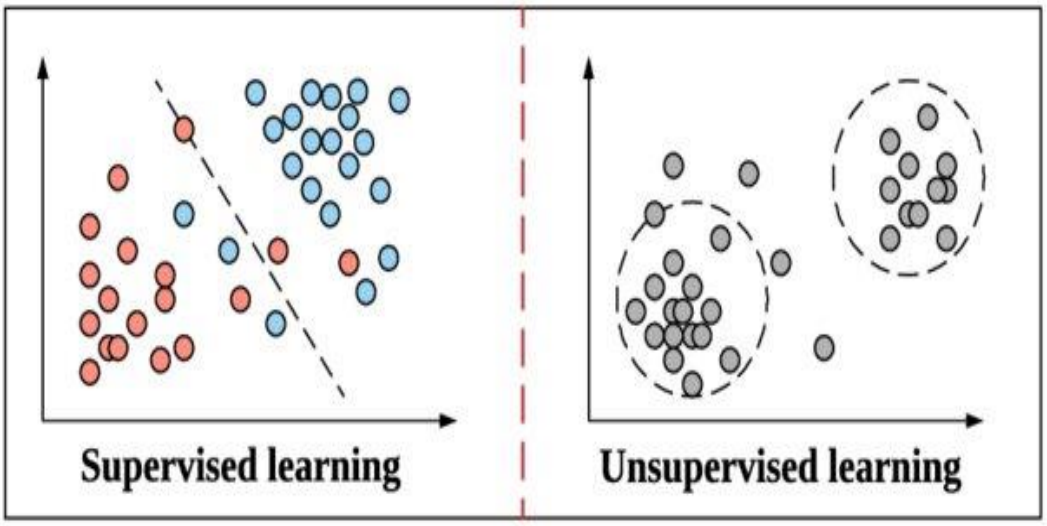


Figure 2.4 Comparison between supervised and unsupervised learning [13]

2.5 Machine Learning Algorithms

Machine learning algorithms are mathematical model mapping methods used to learn or uncover underlying patterns embedded in the data.

2.5.1 Support Vector Machine

Support Vector Machine (SVM) which is used mostly for pattern recognition is known as one of best machine learning algorithms. SVM has been applied to many pattern classification problems such as image recognition, speech recognition, text categorization, face detection and faulty card detection, etc. It is a supervised machine learning algorithm. SVM has great ability to generalize the problem, which is important in statistical learning.

The object of SVM is to find such an optimal hyperplane that separates the data as accurately as possible. For example, it tries to find such a surface that is as distant as possible from the samples of each class. Selecting samples which are closest to the other class is the critical thing about this kind of classification. These samples are named "Support Vectors". Since the separating hyperplane is optimum, the algorithm is made to have the maximum generalization ability. Separating hyperplane should be located in the middle of the two hyperplanes from different classes to be optimum. Two hyperplanes are constructed by joining support vectors from different classes. Thus, optimal separating hyperplane is placed in the middle of two hyperplanes [27].

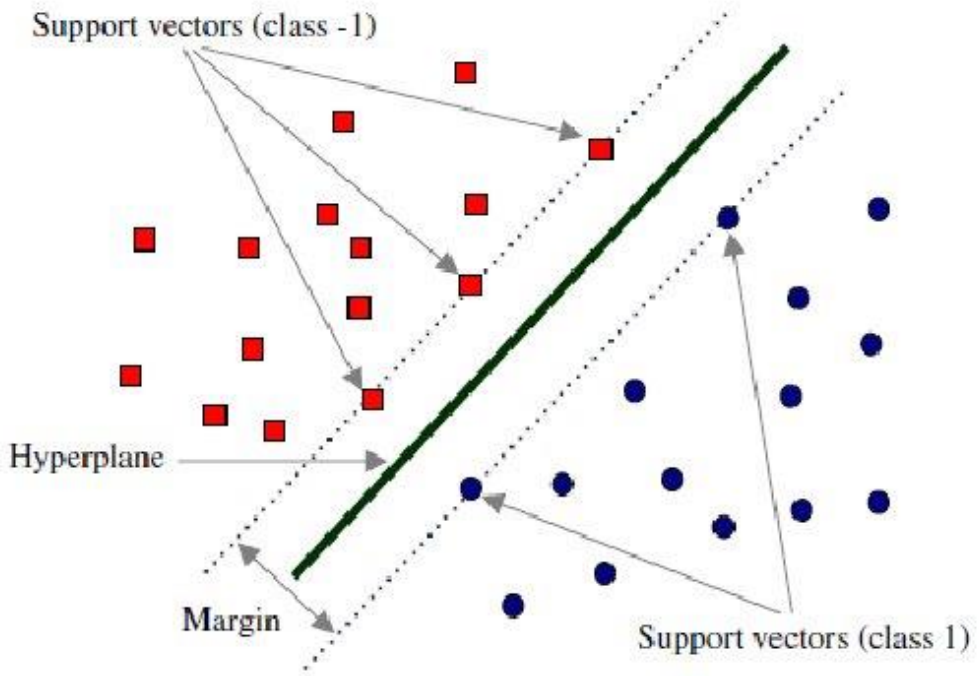


Figure 2.5 Support vector machine [16]

2.5.2 Neural Network

The artificial neural network (ANN) is inspired by the biological brain, which consists of an abstracted model of interconnected neurons. The model is used to solve computer-based application problems in various fields such as statistics, technology or economics. The neural network is a research subject of neuro informatics, also a part of the artificial intelligence. Before solving problems, neural networks need to be trained [26].

Standard neural networks are 'multilayer perceptrons'. These start with an input layer that contains the predictors (input neurons), and end with output layer to generate predictions (output neurons). Hidden layers are placed between input and output neurons. Neurons are activated by input from neurons of the previous layer. This activation is like regression, because neurons evaluate all input signals with weights to come to an output signal to the next layer. The networks can be built with different gradients of complication.

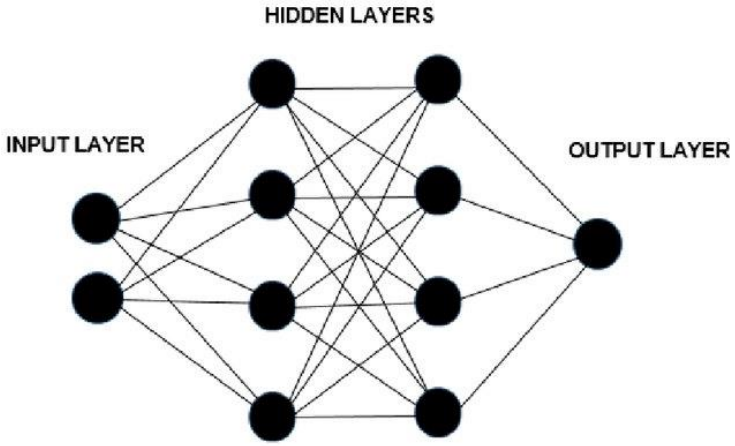


Figure 2.6 Structure of a neural network [19]

2.5.3 Random Forest

Random Forest is an ensemble learning method formed by combination of trees. The predictive power of the model is increased in this method. B bootstrap samples are created from the data and a single tree with full size is grown from each bootstrap sample. Depending on the problem of interest, classification trees or regression trees can be applied. Random Forest acts on the variance of the model, and provides an advantage for problems with multiple independent variables. Clearly, when the tree is being grown, only a random subset of the original features is considered as candidates for a split for each bootstrap sample. The process of randomization is repeated for each split. Thence, each variable is given the same chance of being included in a split. Furthermore, this process guarantees that the trees from the bootstrap samples will be less similar to each other which in return reduces the variance and helps avoiding overfitting problems when the predictions from them are averaged [25].

The splitting rules applied for classification and regression trees can also be applied for random forests. Later with the estimates obtained from each tree, the final estimates are achieved, and interpretations can be made [25].

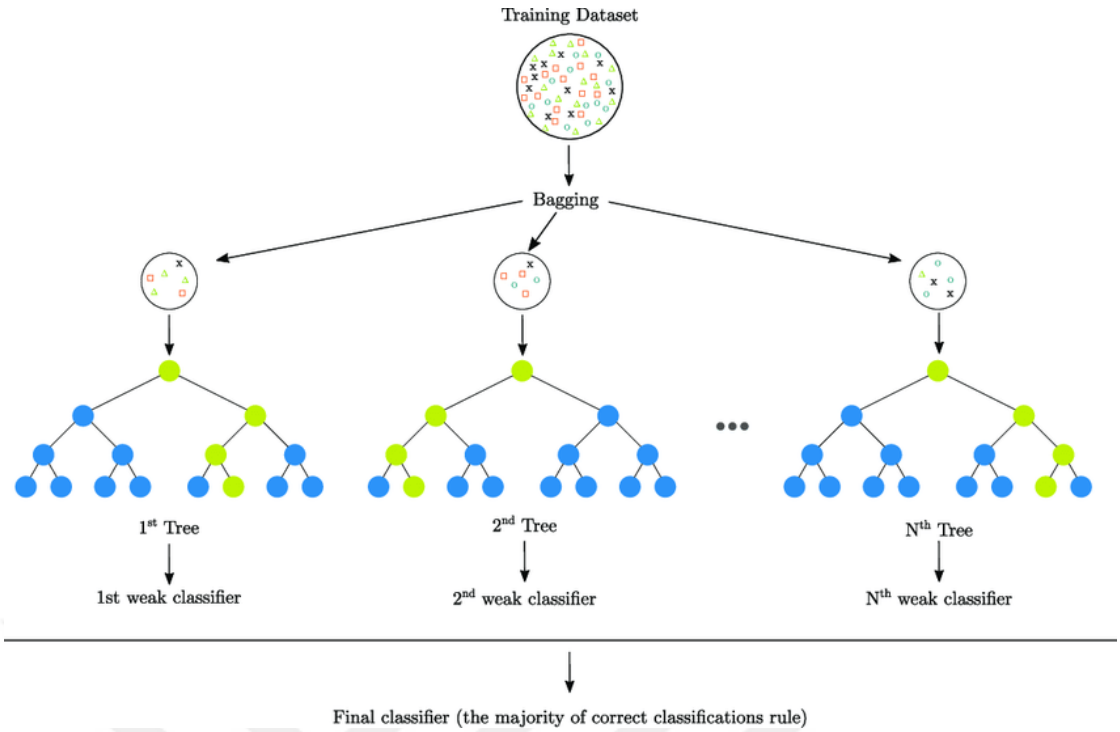


Figure 2.7 RF procedure [18]

2.5.4 Gradient Boosting Machines

Boosting generally is defined as increasing performance. It is a sequential ensemble learning technique to convert a weak hypothesis or weak learners into strong learners to increase model accuracy in machine learning. Decisions from multiple machine learning models in ensemble learning process are combined to reduce errors and improve prediction when compared to a single machine learning model [28].

Gradient boosting is a special case of boosting algorithm where a gradient descent algorithm minimizes the errors and creates a model in the form of weak prediction models e.g. decision trees. The main difference between boosting and gradient boosting is how models are updated from wrong predictions. Gradient boosting adjusts weights by using the "Gradient Descent" algorithm. Gradient Descent updates weights, and iteratively optimizes the loss of the model. Loss refers to the difference between the predicted value and actual value. While MSE is used in regression, logarithmic loss is used in classification. Gradient boosting uses "Additive Modeling" strategy. A new decision tree is added to a model to minimize the loss using gradient descent. Existing trees which slow down the rate of overfitting in the model remain untouched. The output of the new tree and the existing trees are combined until the loss is minimized below a threshold or trees reach a specified limit [5,28].

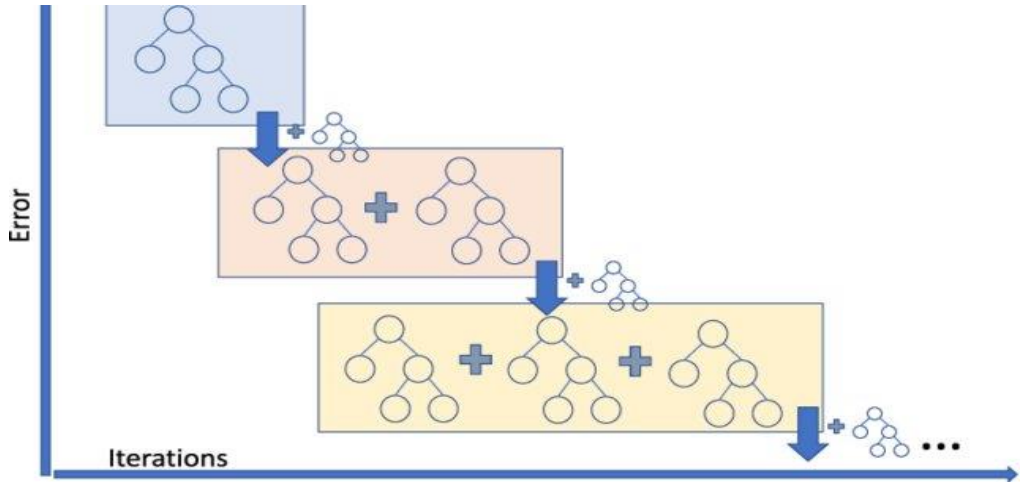


Figure 2.8 GBM procedure [17]

2.5.5 Extreme Gradient Boosting

Extreme Gradient Boosting (XGB) was introduced by Chen et al. [29] to improve the performance and speed of gradient-boosted decision trees. The XGBoost algorithm aims to obtain powerful predictors by using serial dataset training processes to combine weak predictors (Decision Trees). In the sequential modeling process of XGBoost, each decision tree (DT) depends on the result of previous tree to build an improved predictor [30]. Owing to the XGBoost algorithm, overfitting can be prevented during the modeling process by adding a regularization term . The modeling process is more rapid [31,43].

The XGBoost algorithm is sparse aware, and has the ability of using memory resources optimally and effectively dealing with missing values during the training process [43].

While XGBoost is a robust machine learning algorithm for both regression and classification problems, it unfortunately gives less accurate results when it works with imbalanced label data for classification problems. Imbalance occurs if one or more classes have lower proportions than the other classes in training dataset [33,43].

2.6 Sample Size Considerations for Prediction Models: EPV

In binary logistic regression models, the number of subjects in the smaller of two outcome groups (number of events) relative to the number of regression coefficients estimated (excluding intercept) is an essential measure for performance. This ratio is known as Events Per Variable (EPV). Models with lower EPV have been causing poor performance [37]. Researchers determined that 10 should be the minimum EPV value for binary logistic regression. We should aim for at least 100 events as a minimum for reliable estimation of the average risk, and at least 10-20 EPV if event rate is less than 20%. For

the higher event rate, the bigger EPV values are recommended [1]. Calculation of EPV is given in Equation (2.14) [41]:

$$\textbf{Number of variables} * \left(\frac{\textbf{EPV}}{\textbf{Event rate}}\right) = \textbf{Sample size} \qquad (\mathbf{2.14})$$

2.7 Evaluation of Performance

While developing or validating a prediction model, we want to quantify the quality of the predictions from the model. In other words, we want to see the model performance. Firstly, we would like to quantify how close the measured predictions to the actual ones. Later, more specific questions may be asked about calibration and discrimination properties of the model which are especially relevant for predicting binary outcomes in individual patients [1].

2.7.1 Overall Performance Measures

The distance between predicted ($\hat{y}$) and actual outcome (y) is a central to quantify overall model performance from a statistical perspective. The distance or residual (y- $\hat{y}$), is for continuous outcomes. For binary outcomes, $\hat{y}$ is equal to the predicted probability p. These distances between observed and predicted outcomes are related to the concept of "goodness-of-fit" of a model, and the amount of variability that is explained. The smaller distance means a better model [1].

2.7.1.1 R Squared

The explained variation R squared (R^2) is an overall measure quantifying the amount of information in the model of a given data. It is a measure of goodness of fit, in other words it shows how well the data fits our model. In linear and generalized linear models, and for all types of regression models, R^2 is a useful guide.

R^2 is the most commonly used performance measure when the outcome is continuous. In Nagelkerge's approach, R^2 is based on logarithmic scoring rule as in (2.15):

$$(y-1) * (\log(1-p)) + y * \log(p) \qquad (2.15)$$

The logarithm of predictions "p" is compared to the actual outcome "y". For binary data, if patients with event is log(p), then patients without event will be log(1-*p*) [1].

2.7.1.2 Brier Score

Brier score is based on quadratic scoring rule that combines discrimination and calibration. The formula of Brier score is as given in (2.16):

$$(y - p)^2 \tag{2.16}$$

The outcome is *y* and *p* is the prediction. Also, the formula of Brier score is as in (2.17):

$$y * (1 - p)^2 + (1 - y) * p^2 \tag{2.17}$$

For example, if the probability of snow for today is 0.8 and if it snows today, then the Brier score will be (0.8-1)*(0.8-1) = 0.04 since *p*=0.8 and *y*=1. If the prediction were equal to the probability, then the result would be 0. So, best possible Brier score is 0.

Brier score can be scaled in maximum score as given in (2.18):

$$Brier_{scaled} = 1 - \left(\frac{Brier}{Brier_{\max}}\right) \tag{2.18}$$

Equation (2.18) is scaled after maximum Brier score is as given in (2.19):

$$Brier_{max} = mean(p) * \big(1 - mean(p)\big)^2 + \big(1 - mean(p)\big) * mean(p)^2 \tag{2.19}$$

Here mean (*p*) indicates the average probability. Scaled Brier score has better interpretability and ranges from 0 to 100%.

As seen in Figure 2.9 [1], logarithmic (solid line) and quadratic (dashed line) scores are presented where *y*=1 for "event" and *y*=0 for "no event". It is shown that logarithmic score severely penalizes false predictions close to 0 or 100%.

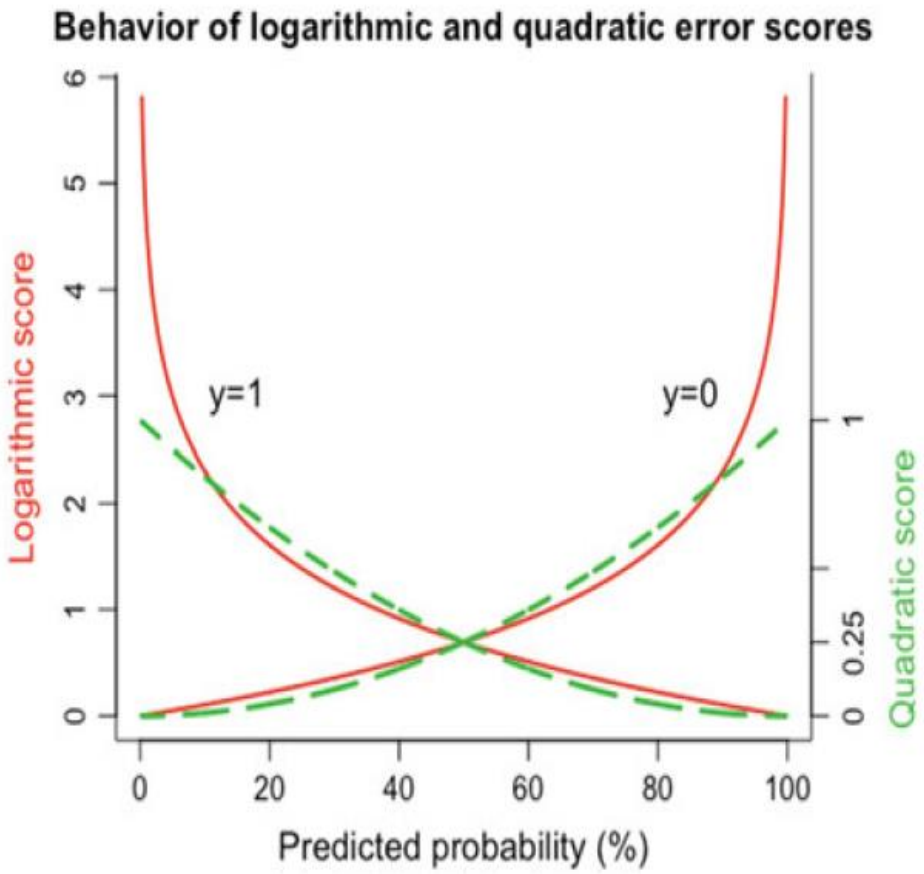


Figure 2.9 Logarithmic and quadratic error scores [1].

2.7.2 Discriminative Ability

Binary outcomes with and without event need discrimination. One of the most commonly used performance is concordance statistic (*c*).

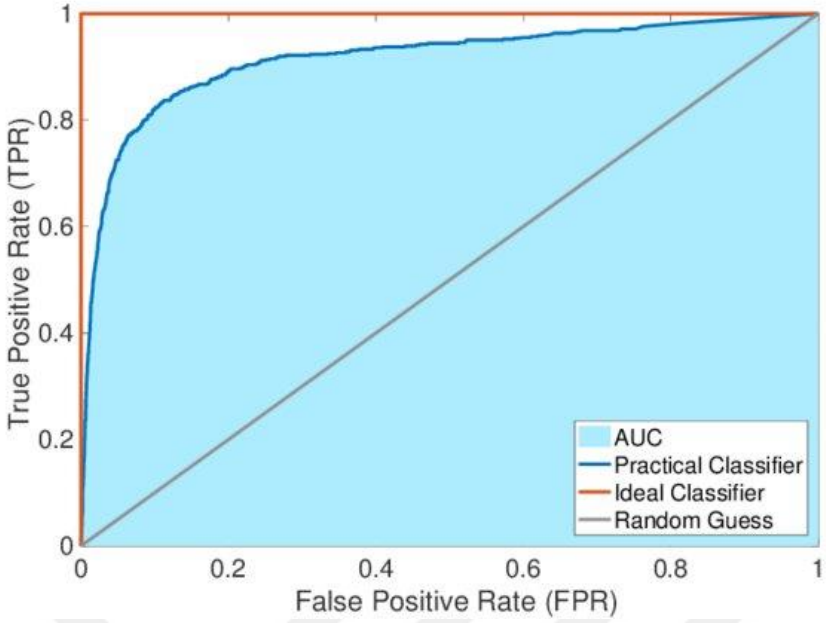


Figure 2.10 AUC and ROC [42]

For binary outcomes, *c* is identical to the area under the receiver operating characteristic (ROC) curve (AUC). The ROC curve shows the true positive rate (sensitivity) against false positive rate (1-specifity). The *c*-statistic as given in Figure 2.10, ranges from 0.5 to 1 where 0.5 indicates that model predictions are totally random and 1 indicates perfect discrimination.

2.7.3 Calibration

Calibration means the agreement between the predicted probability of the outcome versus the observed. If we predict 50% probability of a disease in a patient, we expect to see 50 out of 100 patients with ~50% probability. While weaker forms of calibration just require the average predicted risk (mean calibration) or the average prediction effects (weak calibration) to be correct, strong calibration requires the equality of event rate and predicted risk for every covariate pattern. Calibration is shown graphically, while the observed values on the *y*-axis with 0 or 1 (event/no event) and the predictions on the *x*-axis ranging between 0 and 100% [1,35].

A perfect calibration occurs on the 45-degree line. Calibration plot also gives an idea about discrimination. A better discriminating model has more spread between observed proportions per group.

The concept of weak calibration is associated to the average strength of the predictor effects. We write as in (2.20) for linear regression:

$$y_{new} = a + b_{overall} * \hat{y} \tag{2.20}$$

For generalized linear models, the formula becomes as given in (2.21):

$$f(y_{new}) = a + b_{overall} * linear\ predictor \tag{2.21}$$

Here the linear predictor is the merger of regression coefficients from the model and the predictor values in the new data. For y_{new}, a link function f is used like logit in logistic regression. $b_{overall}$ is named the calibration slope, while a is named the intercept. Calibration slope equals 1 with apparent validation, since this yields the best fit on the data under study with either least squares or maximum likelihood methods. The calibration slope which is less than 1 in internal validation reflects the amount of shrinkage that is required for a model. It shows how much reduction on the effects of predictors on average is needed to make the model well calibrated for new patients from the original population. Thus, the calibration slope can be used as a shrinkage factor for model adjustment in future use. In external validation, the calibration slope pictures both the combined effect of overfitting on the development data and true differences in the effects of predictors [1].

For the assessment of 'mean calibration' (or 'calibration-in-the-large'), the average predicted risk is compared with the overall event rate. When the average predicted risk is higher than the overall event rate, risk is overestimated by the algorithm. On the contrary, underestimation occurs when the observed event rate is higher than the average predicted risk [38]. Figure 2.11 shows calibration plots with overestimation (thin, dashed line), underestimation (thick, dashed line) and ideal calibration (line in the middle):

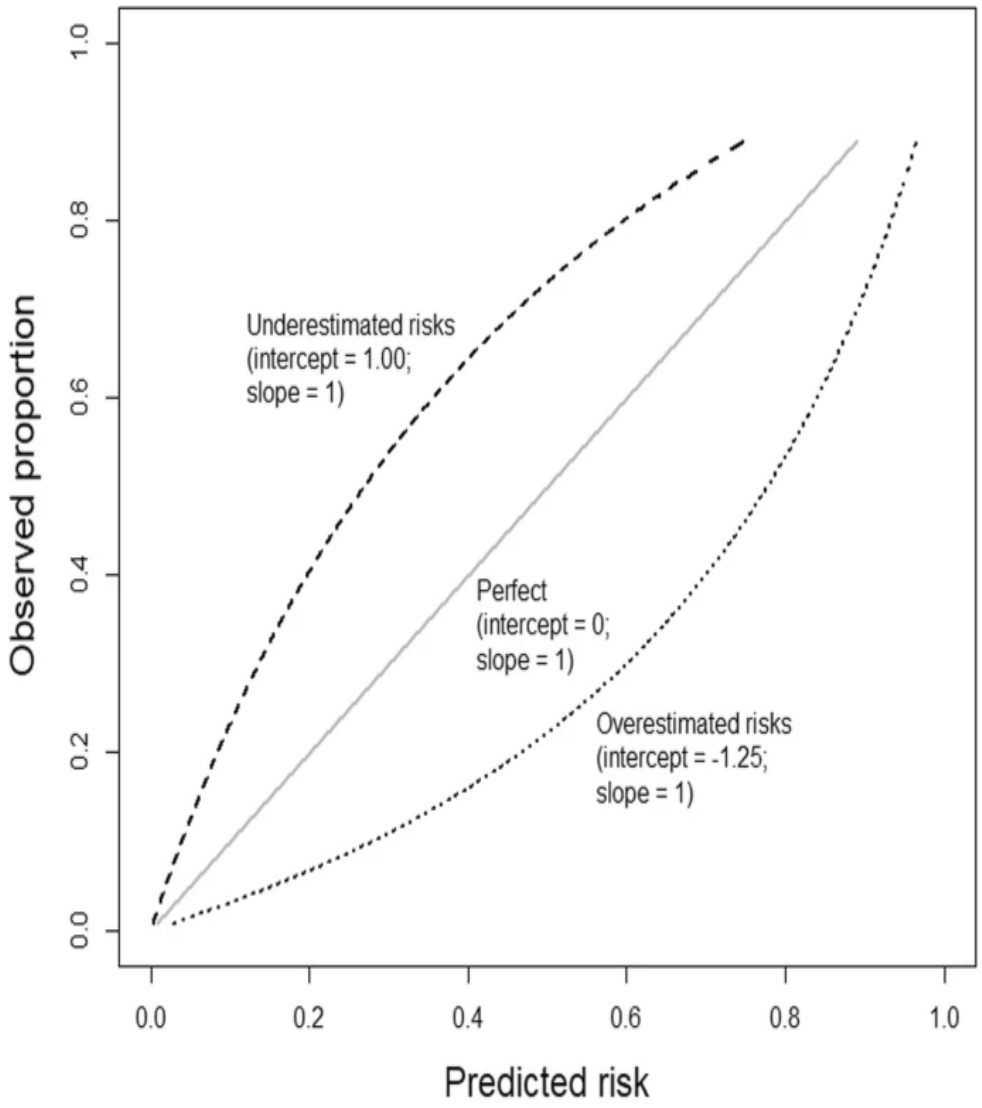


Figure 2.11 Calibration plots [38].

If the model does not overestimate or underestimate risk and does not give overly extreme (too close to 0 and 1) or modest (too close to disease prevalence or incidence) risk estimates, it is called weak calibration. It can be assessed by the calibration intercept and calibration slope. The calibration slope which evaluates the spread of the estimated risks has a target value of 1. If slope is less than 1, estimated risks are too extreme. For example, risk is much higher for patients who are at high risk and much lower for patients who are at low risk. If slope is over 1, it means risk estimates are too moderate. The calibration intercept which has a target value of 0 is an assessment of calibration-in-the-large. Negative intercept means overestimation, while positive intercept means underestimation [38].

2.8 Validation of Prediction Models

Prediction models are expected to provide valid outcome predictions for new patients. Validation is needed for predictive modeling. Internal validity means reproducibility of the predictive model, while external validity is transportability, or generalizability of the predictive model in plausibly related populations [1].

2.8.1 Internal Validation

The model we develop within a sample of patients represents the underlying population. The process of determining the reproducibility of this model is internal validation. It shows how valid this model is for the population. One of the most common internal validation techniques is apparent validation. 100% of the available data is used to develop the model, and 100% of the data is used to test the model. This procedure gives optimistic performance estimate [47,1].

The sample is randomly divided if we want to apply split-sample validation. Typically, splits are as 50:50 or 2:1. The model is developed in one part of the sample and tested in the other part. For large samples, split-sample validation may work well. However, apparent validation already performs well for large samples. Also, splitting may be unlucky for the investigator. For example, the mean age of the patients in one split may differ from the other which is a very important factor for prediction. To sum up, split-sample is not recommended as the best method [1]. In Figure 2.12, an example of split-sampling procedure is shown.

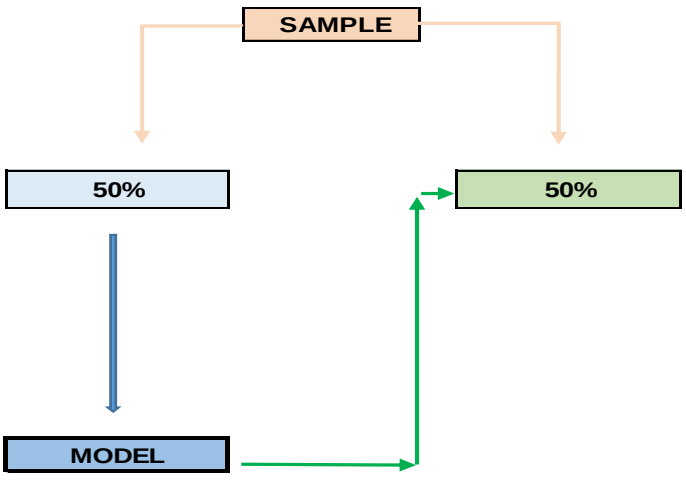


Figure 2.12 Split-sample validation

In cross-validation, more stability is aimed compared to split-sample method. Patients are randomly divided into k equal sized folds, or groups. In one-fold, the model is validated while in (k-1) folds, the model is developed [1,6].

Bootstrap is another resampling method like cross-validation. The main difference here is that samples are drawn with replacement. Bootstrap samples are of the same size as the original sample. Bootstrap with at least 500 repetitions is considered feasible [1]. An example of sampling in bootstrap process is given in Figure 2.13:

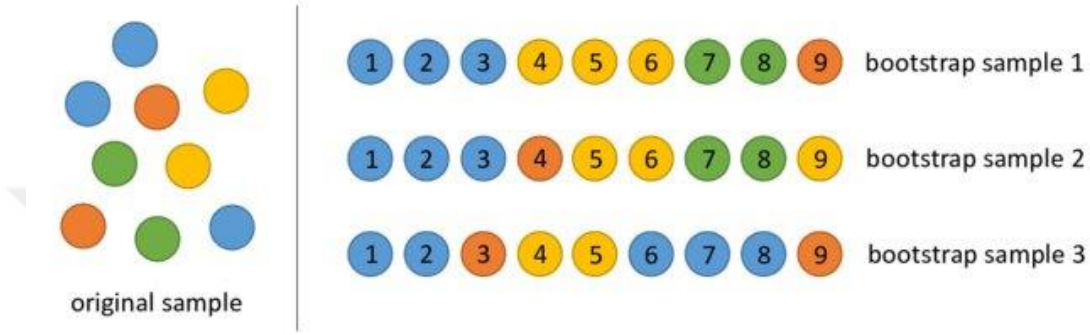


Figure 2.13 Bootstrap approach [36]

2.8.2 External Validation

General applicability of a model is provided by external validation. Even though external validation is similar to the structure of split-sample validation, external validation studies may address aspects of historic (temporal), geographical (spatial), methodological and spectrum transportability [1,48].

Temporal validation refers to performance when a model is tested in different historical periods. The model is developed in earlier treated patients and validated in recently treated patients. It is not a random split. Evaluation of predictive models according to site or hospitals is possible with geographic validation. However, validation samples may get small. Geographical validation is not a random split, too. Fully independent validation is the external validation by independent investigators. Other investigators may use slightly different definitions of predictors, outcome, and study patients that were differently selected compared to the development setting [1,48].

3

MATERIALS AND METHODS

3.1 Data Source

The International Stroke Trial (IST) by the collaborative group Peter Sandercock and friends [4] is one of the biggest randomized trials in acute stroke.

Table 3.1. Definition and numerical features of variables in IST (n=18408)

Variables	**Definition**	**Percentage or Distribution** *(1st-Quartile, Median, Mean, 3rd-Quartile)*
DALIVE	Discharged alive from hospital	53% (alive), 47% (dead)
AGE	Age in years	65.0 73.0 71.8 80.0
SEX	M=male; F=female	53% (M), 47% (F)
RDELAY	Delay between stroke and randomization in hours	9.00 19.00 20.06 29.00
RCONSC	Conscious state at randomization	77% (fully alert), 23% (drowsy and unconscious)
RATRIAL	Atrial fibrillation (Y/N)	17%
RSBP	Systolic blood pressure at randomization (mmHg)	140.0 160.0 160.2 180.0
RXASP	Trial aspirin allocated (Y/N)	50% (aspirin), 50% (heparin)
DDIAGHA	Haemorrhagic stroke	3%
RVISINF	Infarct visible on CT (Y/N)	33%

19435 patients with acute ischaemic stroke between 1991 and 1996 are included. The aim of the trial was to establish whether early administration of aspirin, heparin, both or neither influenced the clinical course of acute ischaemic stroke.

3.2 Choice of Variables

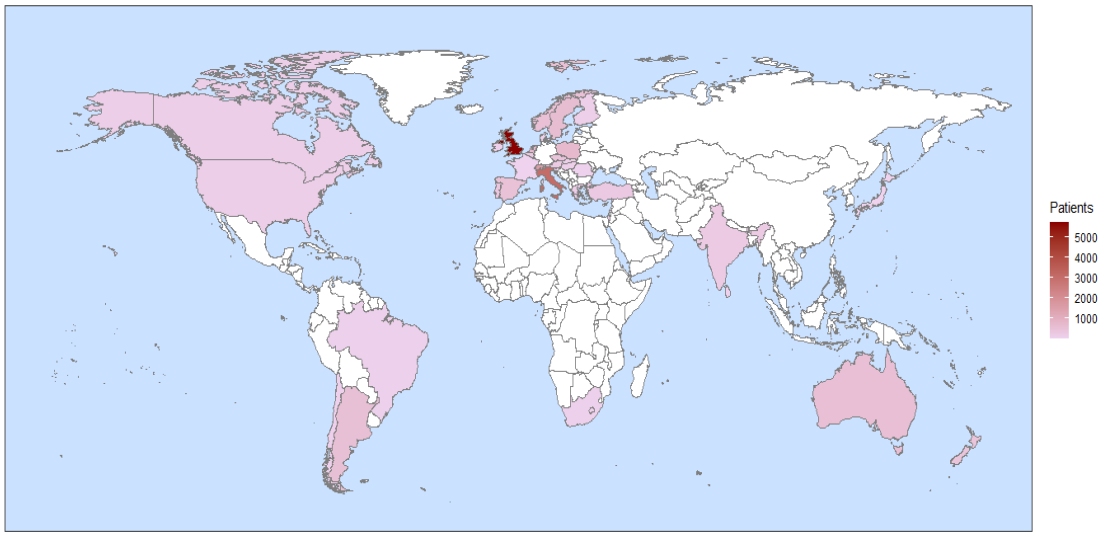
Nine variables are used to predict the mortality of the patients in the model which consists of three continuous variables (RSBP, AGE, RDELAY) and six categorical variables as given with details in Table 3.1.

Missing values of 1027 were removed since the data is already large enough. 18408 became the new number of rows of our data. For the variable RCONSC, drowsy and unconscious patients were combined to make it binary.

R version 4.1.2 was used for the whole process of the study. Codes are available in the Appendix section.

3.3 Sample Creation

The patients in the IST are from all around the world. Map 3.1 shows the countries according to the numbers of patients.



Map 3.1 International Stroke Trial patients in world map

The main aim of the study was geographical validation. International Stroke Trial dataset (n=19435) was divided into three after missing values were removed (n=18408): "UK data" (n=5817) including UK and Ireland patients, "non-UK data" (n=9955) including all European patients except UK and Ireland patients, and "non-EU data" (n=2636) including world patients except all European patients. After UK data was sampled, EPV of samples

were determined. Table 3.2 shows the percentages and distributions of the variables from IST samples.

Table 3.2 Numerical features of variables from IST samples

Variables	**Percentage or Distribution**		
	(UK, n=5817)	**(NON-UK, n=9955)**	**(NON-EU, n=2636)**
DALIVE	40% (alive), 60% (dead)	57% (alive), 43% (dead)	70% (alive), 30% (dead)
AGE	67.00 75.00 73.79 82.00	65.00 73.00 71.58 80.00	60.75 70.00 68.29 77.00
SEX	51% (M), 49% (F)	53% (M), 47% (F)	58% (M), 42% (F)
RDELAY	10.00 21.00 21.46 31.00	8.00 17.00 18.58 27.00	12.00 24.00 22.58 32.00
RCONSC	73% (alert), 27% (other)	78% (alert), 22% (other)	78% (alert), 22% (other)
RATRIAL	21%	17%	10%
RSBP	140.0 160.0 158.5 180.0	140.0 160.0 162.3 180.0	140.0 150.5 156.0 170.0
RXASP	50%	50%	50%
DDIAGHA	3%	2%	2%
RVISINF	27%	34%	42%

To create prediction models, sub-samples were randomly created from UK data with 19 different sample sizes: 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, and 3000. For GAM, only 17 models were created since models did not work in low (50 and 100) sample sizes. EPV values were determined for each of the sub-sample. For example, EPV for the sub-sample with size 1000 is as given in (3.1):

$$1000 * \frac{0.596}{9} = 66.22 \tag{3.1}$$

In the Equation (3.1), 9 is the number of variables to predict DALIVE. 1000 is the sample size. 0.596 is the event rate. Lowest value for EPV became 3.31, while 198.67 was the highest. EPV values for each sample size in the study is given in Table 3.3.

Table 3.3 EPV values for each sample size used in this study

Sample	**EPV**
50	3.311110455

100	6.622222386
150	9.933332842
200	13.24444477
250	16.55555228
300	19.86667158
350	23.17777614
400	26.48889545
450	29.8
500	33.11110455
600	39.73332842
700	46.35555228
800	52.97777614
900	59.6
1000	66.22222386
1500	99.33332842
2000	132.4444477
2500	165.5555228
3000	198.6667158

3.4 Modeling Steps

For each sample, 6 different regression models and 5 different machine learning models were created for prediction of death at sixth month in IST. UK data models except the first were internally validated in UK data. All samples were externally validated both in non-UK and non-EU data. Figure 3.1 shows the validation processes. In R, val.prob.ci.2 function was used for calibration plots [7].

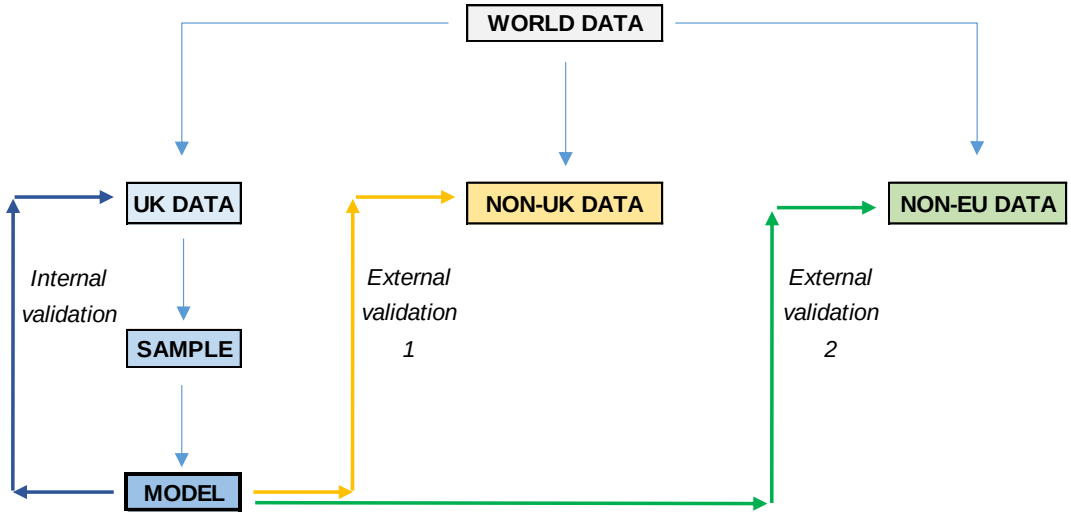


Figure 3.1 Internal and external validation based on geographic validation

3.5 Regression Techniques

The regression techniques which were compared with the machine learning (ML) algorithms included standard logistic regression, logistic regression with restricted cubic splines, generalized additive models (GAM), and penalized regression including ridge, lasso and elastic net.

Lower sample models with size 50 and 100 did not work, so they were extracted for GAM. Smoothing splines specified by s(x) were used with continuous predictors as mentioned in software code in appendix. For logistic regression with RCS, 5 knots were used with continuous predictors. Alpha values for lasso, elastic net and ridge were set to 1, 0.5 and 0, respectively.

3.6 Machine Learning Algorithms

The parameters which are not mentioned here were set to default values. In gradient boosting machines, 100 trees were grown. For random forests, mtry was chosen 2, and 500 classification trees were grown. There is no difference than default settings for "gbm" functions in R having the same name as libraries.

In extreme gradient boosting, maximum depth of the trees was 6. The task was chosen logistic regression for binary classification. Output was as probability. Number of rounds was set to 70 for training. While creating the XGB model, the smallest number in the rounds was chosen since we needed the step with lowest test RMSE.

While creating neural network model, size and decay should be specified. Size is the number of units in hidden layer and decay is the regularization parameter to avoid over-fitting. In this study, size was 10 and decay was 5% for NN. Maximum number of iterations was 750. For support vector machines, kernlab package was used in R programming language. C-classification (C=10) and radial basis function (sigma=0.1) were applied.

4

RESULTS AND DISCUSSION

4.1 Results

For the plots in this section, x-axis represents EPV (or event per degree of freedom) values, while y-axis represents the model performances. Regression models and machine learning models were given in different figures in this section.

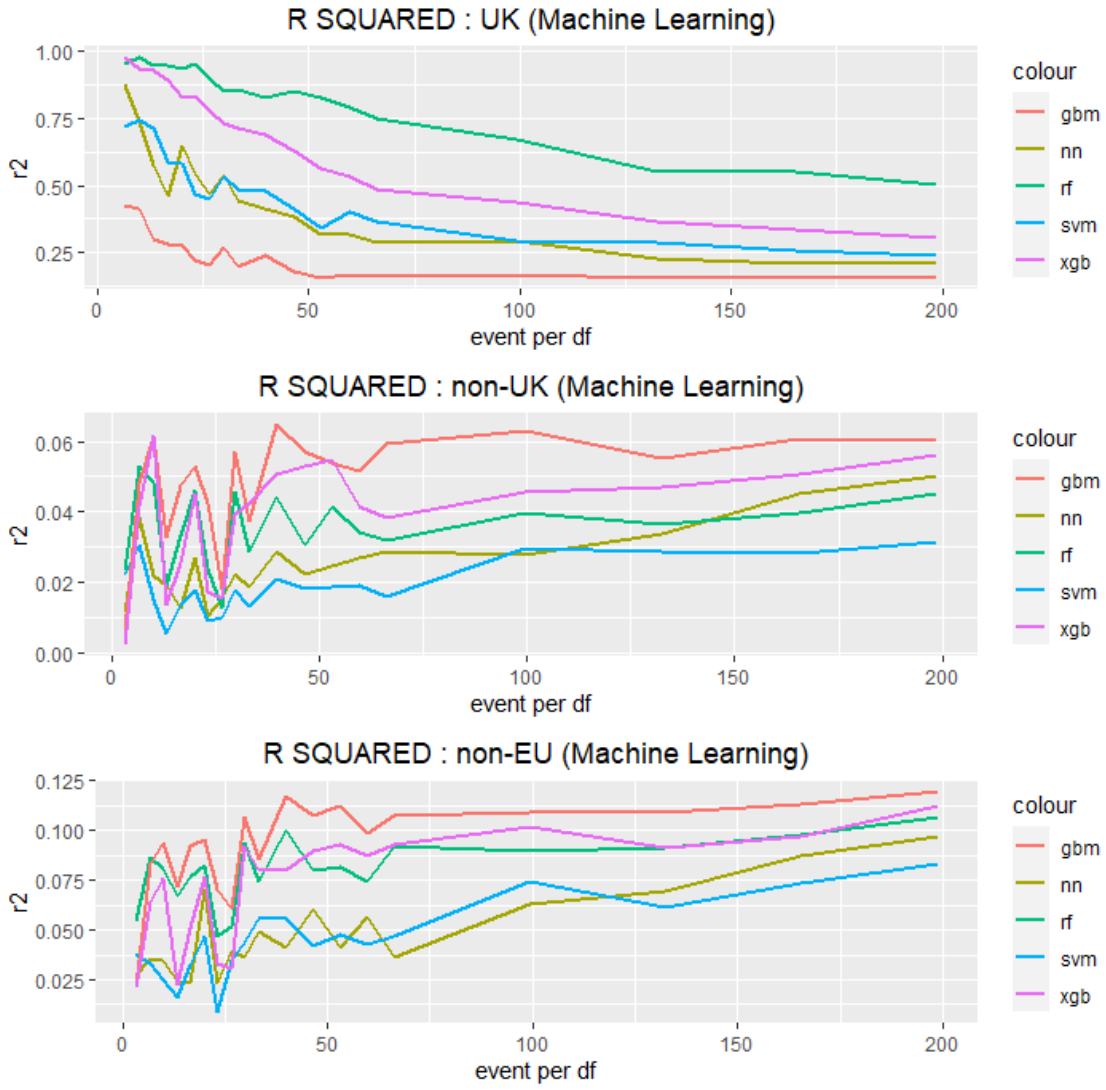


Figure 4.1 Results of R squared in machine learning models

4.1.1 Results of R-Squared

Figure 4.1 and Figure 4.2 show the comparison of R squared values between models.

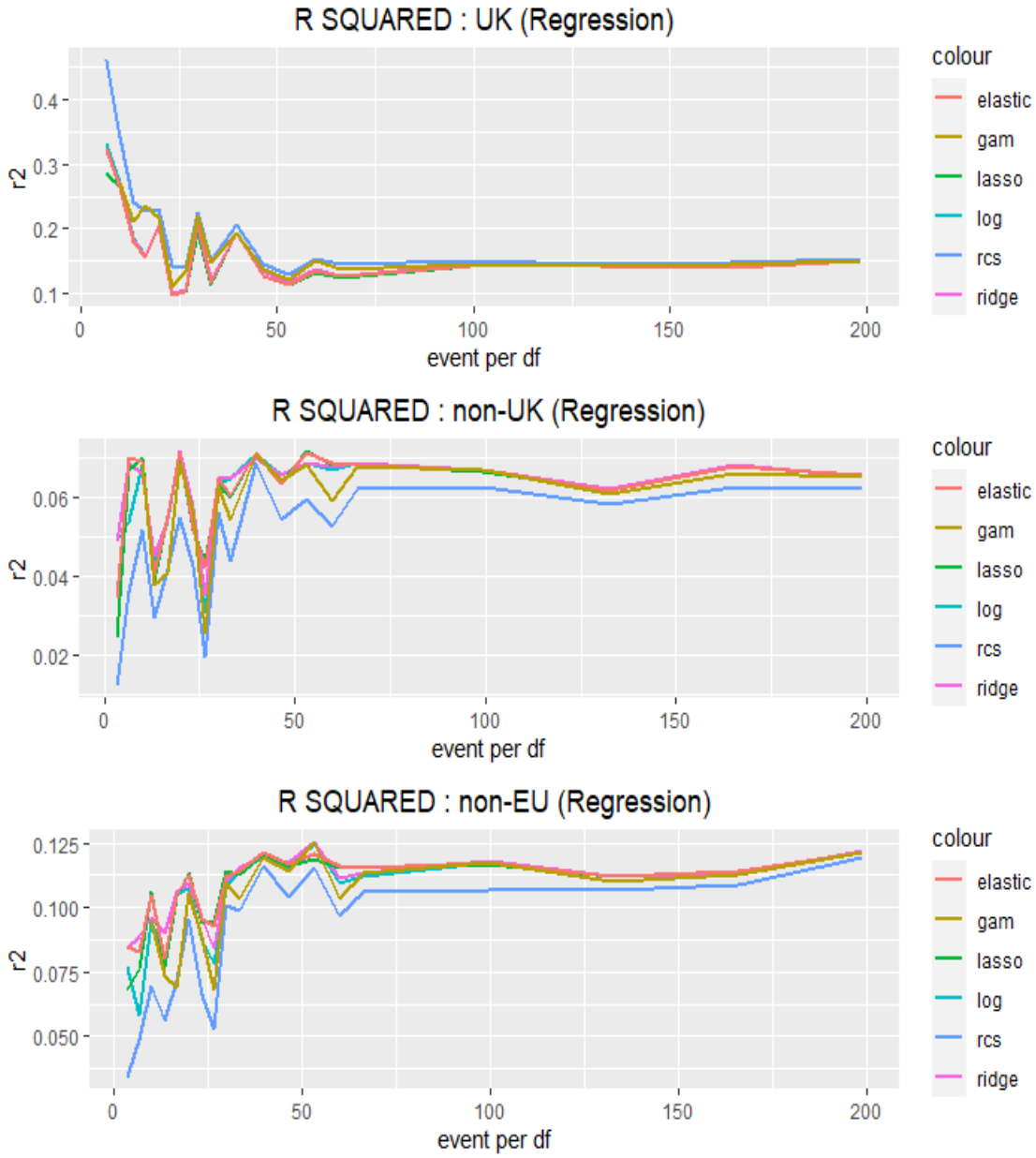


Figure 4.2 Results of R squared in regression models

For EPV>30, all regression models have stable R squared. R squared was between 0.100-0.125 for non-EU, and between 0.06-0.07 for non-UK external validation. Excluding GBM, machine learning methods have lower R squared values. R squared for non-UK was below 0.1, and for non-EU it was below 0.05. For EPV<30, almost all models have unstability.

4.1.2 Results of Brier Score

Figure 4.3 and Figure 4.4 show comparison of Brier score values between models. In Appendix section, all models were given together. For all the figures of regression results, "rcs" shows the logistic regression with restricted cubic splines curves, while "log" shows logistic regression without restricted cubic splines curves. Internal validation results, non-UK external validation results and non-EU external validation results are located at the top, middle and bottom, respectively for all the plots of this section.

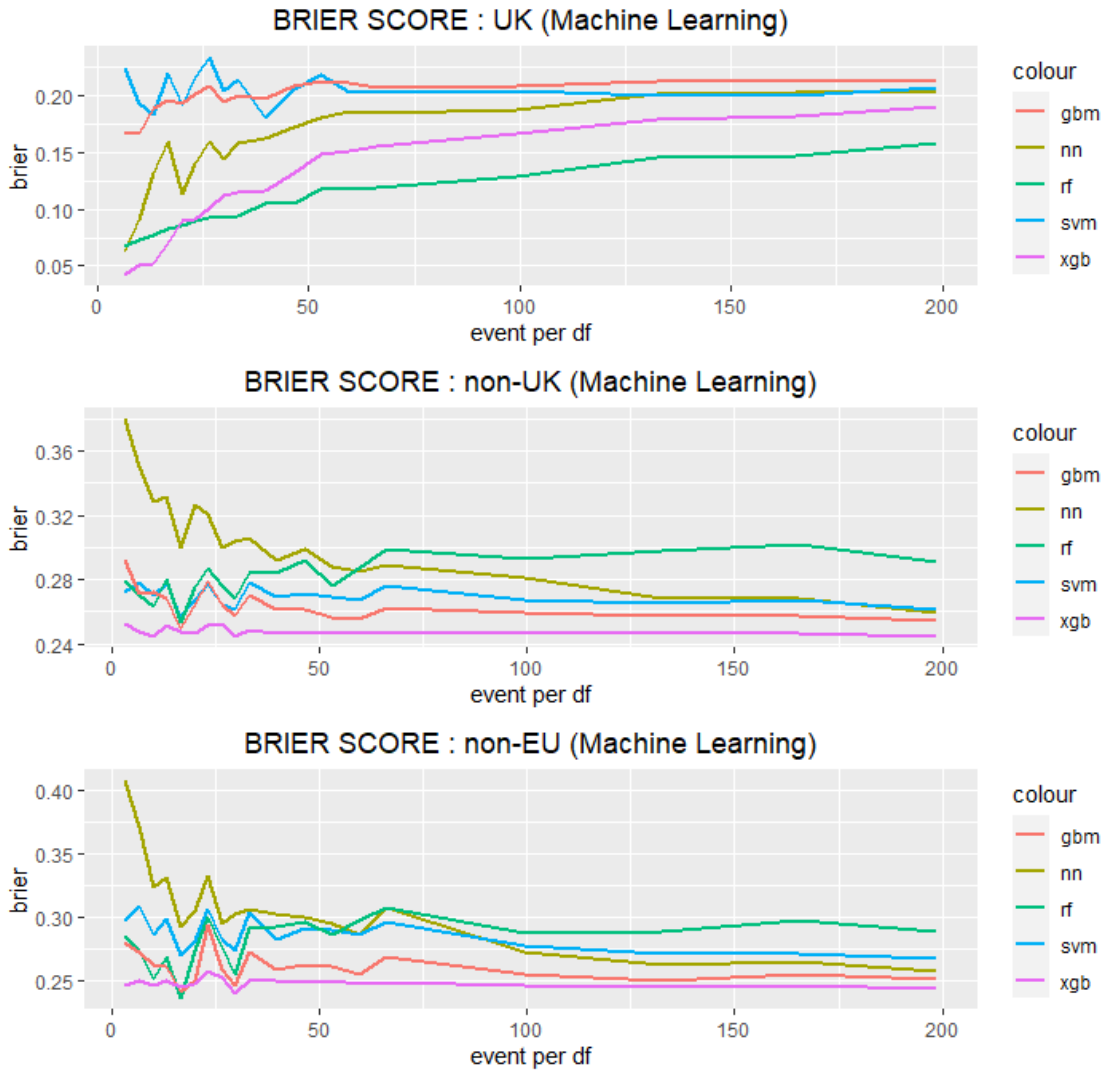


Figure 4.3 Results of Brier score in machine learning models

According to Brier score results, XGB performed best, followed by regression methods which were similar to each other. GBM performed best in all machine learning models. For most of the models, no major change in Brier score was observed with the increase in EPV values. For highest EPV values, regression models were between 0.24-0.26, while machine learning models were between 0.24-0.29 according to external validation results.

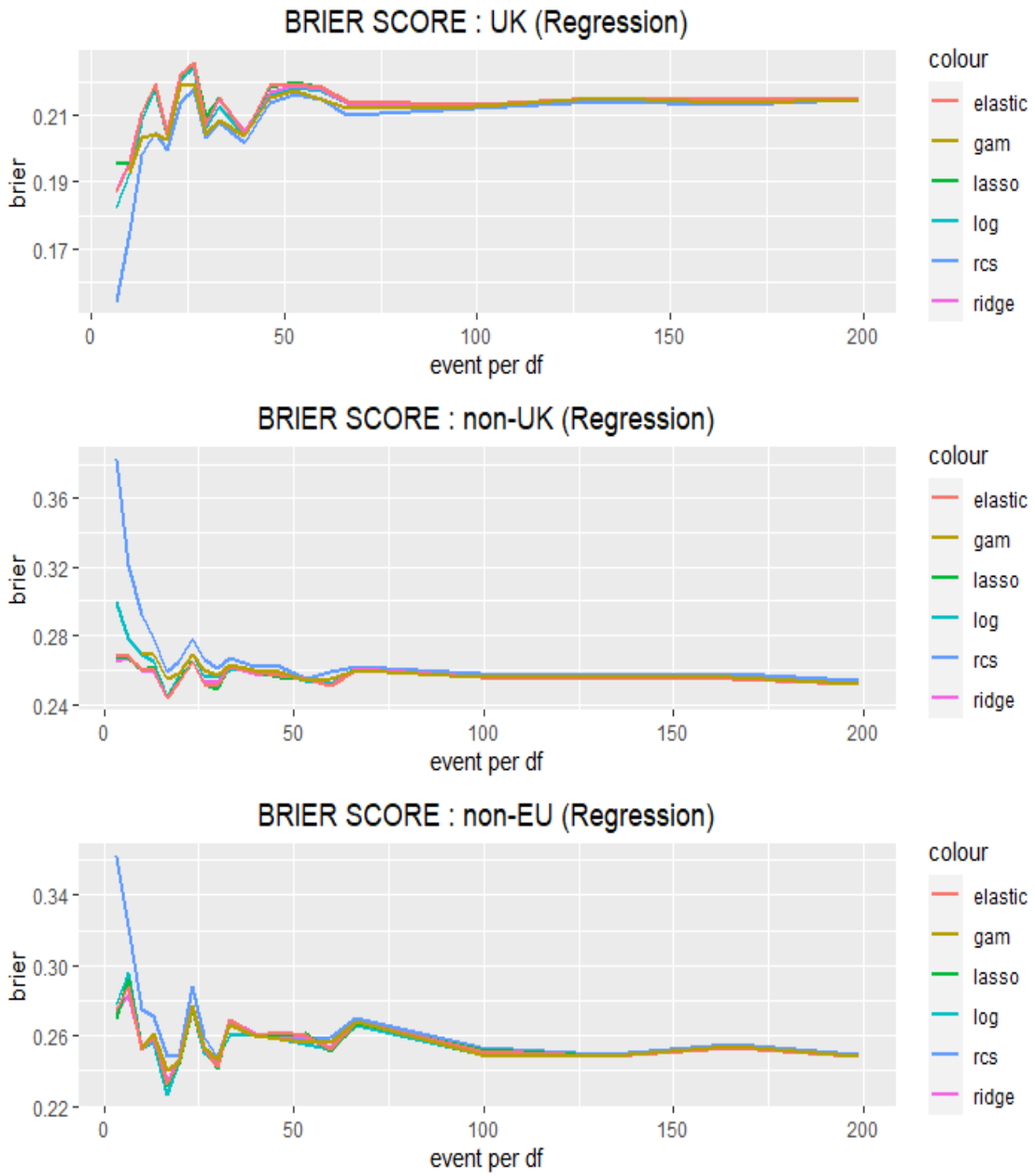


Figure 4.4 Results of Brier score in regression models

4.1.3 Results of AUC

For EPV>30, all regression models have stable and high AUC levels in external validations. Machine learning methods have lower and unstable AUC levels excluding GBM. For EPV<30, almost all models seem to be unstable.

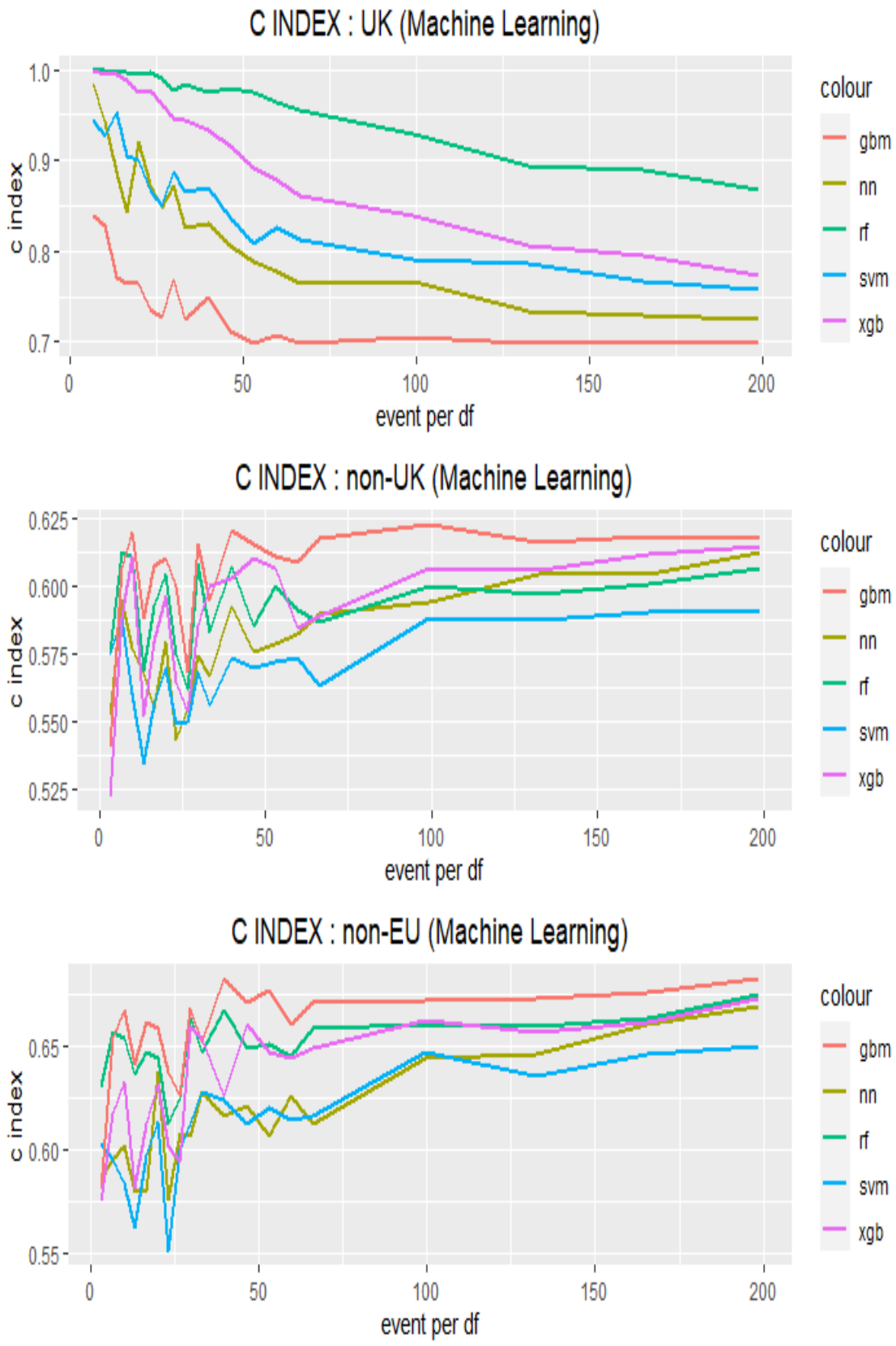


Figure 4.5 Results of AUC (C index) in machine learning models

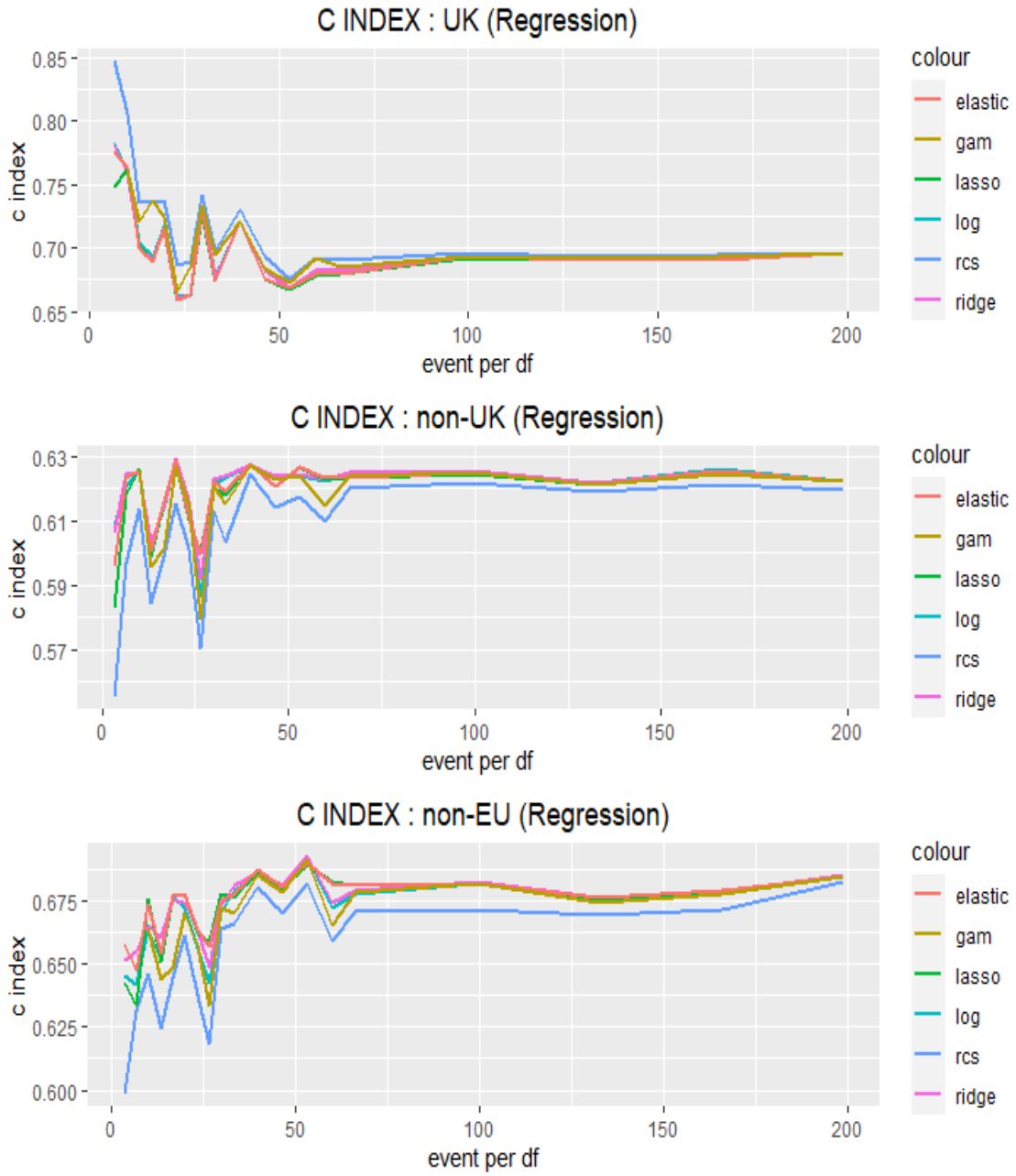


Figure 4.6 Results of AUC (C index) in regression models

4.1.4 Results of Calibration

For all regression models and machine learning methods excluding SVM and XGB, slight miscalibration was observed. Calibration-in-the-large is lower in XGB. According to calibration plots, probability of death localized around 0.4-0.6 for XGB in external validations. For EPV>50, traditional regression models are more stable than machine learning methods according to calibration slopes. XGB obviously has higher slope values than other machine learning methods.

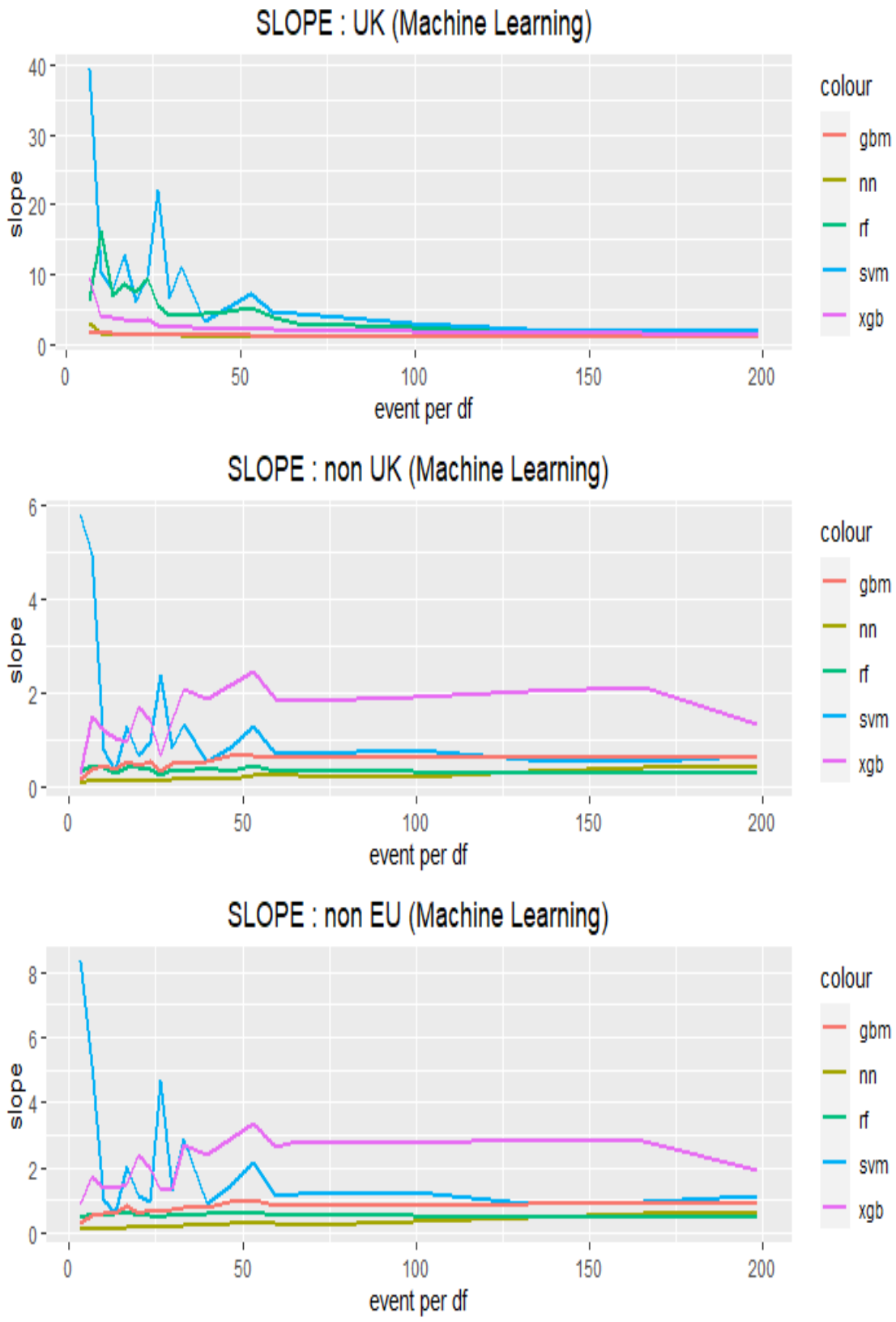


Figure 4.7 Calibration slope in machine learning models

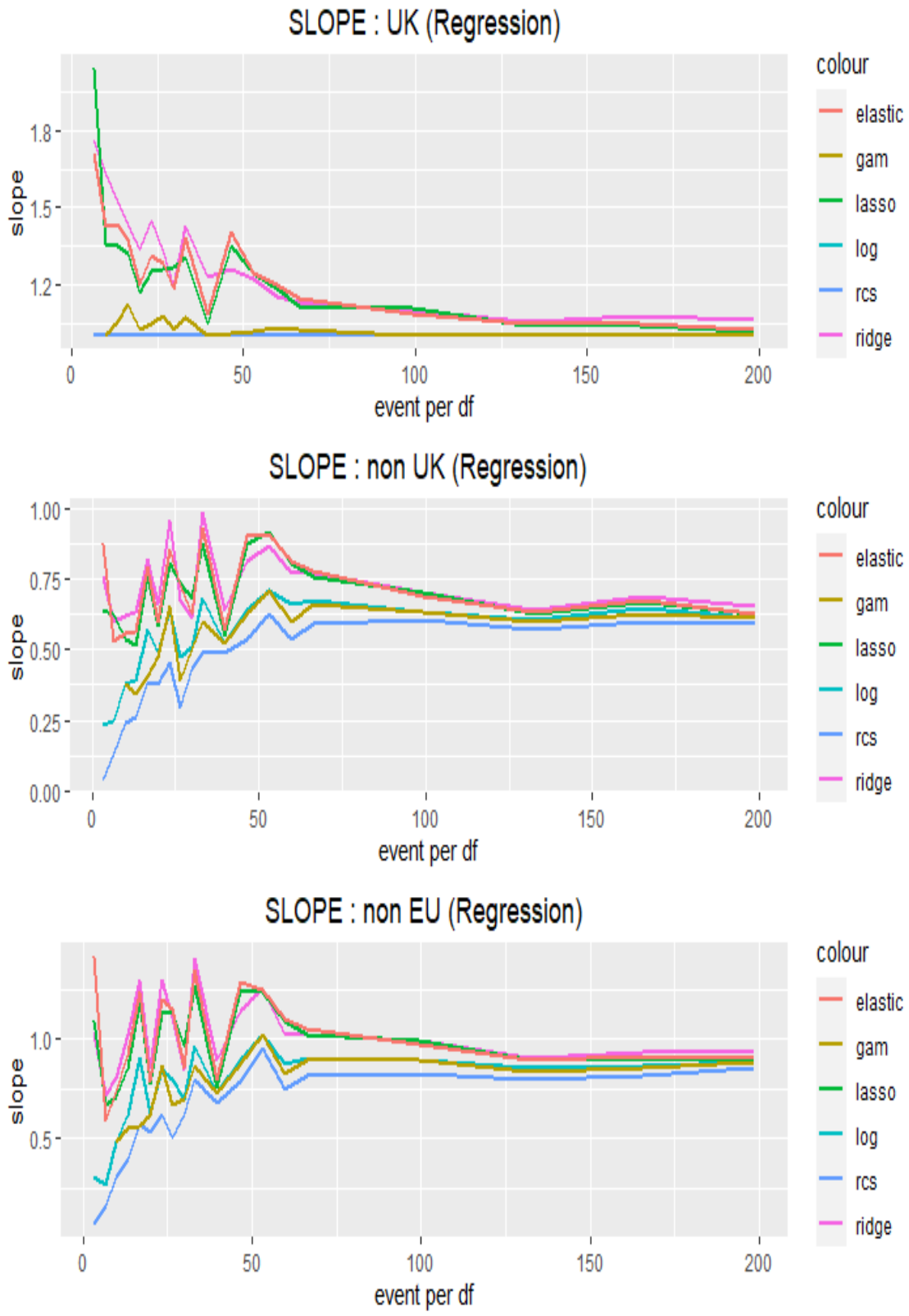


Figure 4.8 Calibration slope in regression models

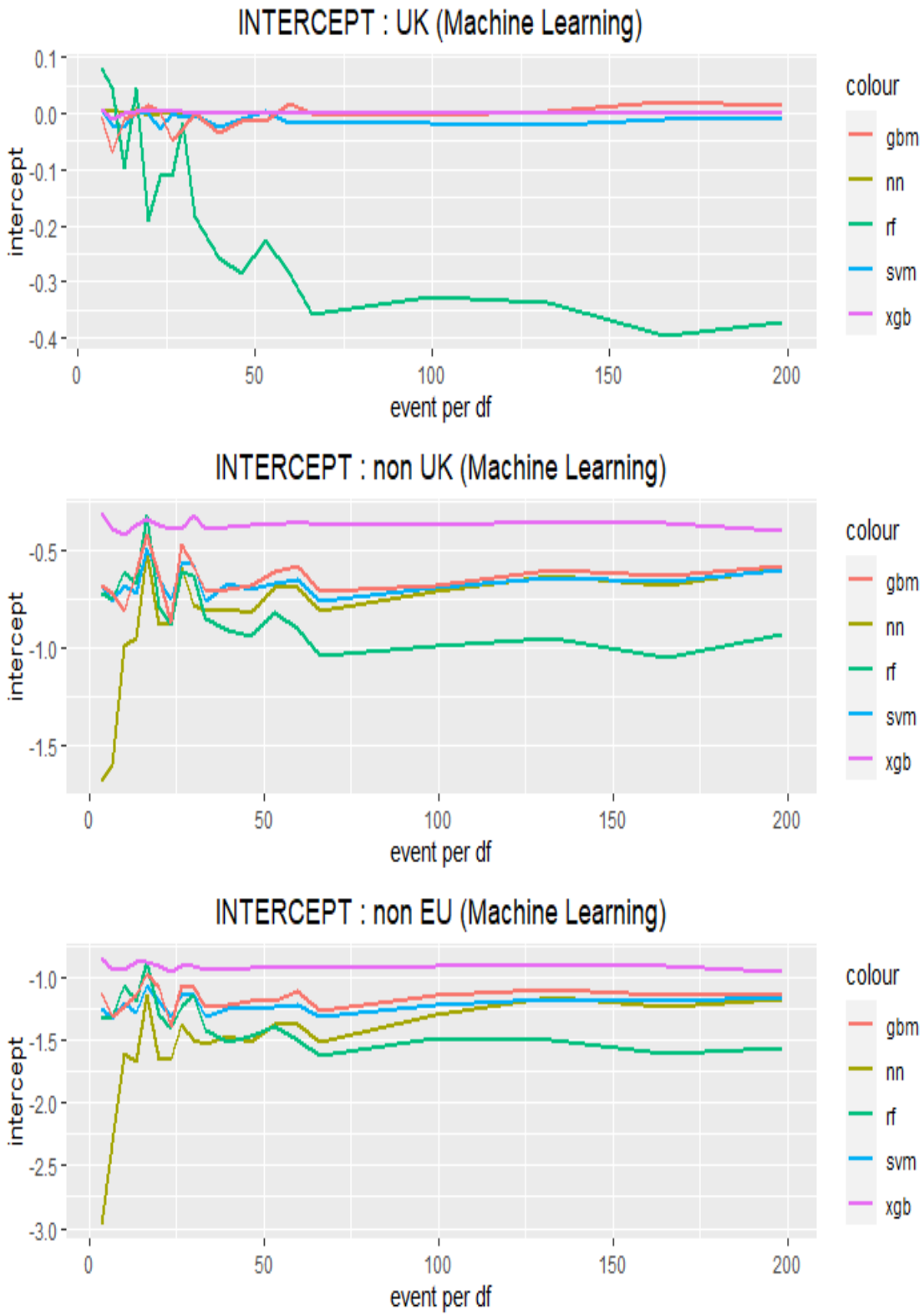


Figure 4.9 Calibration intercept in machine learning models

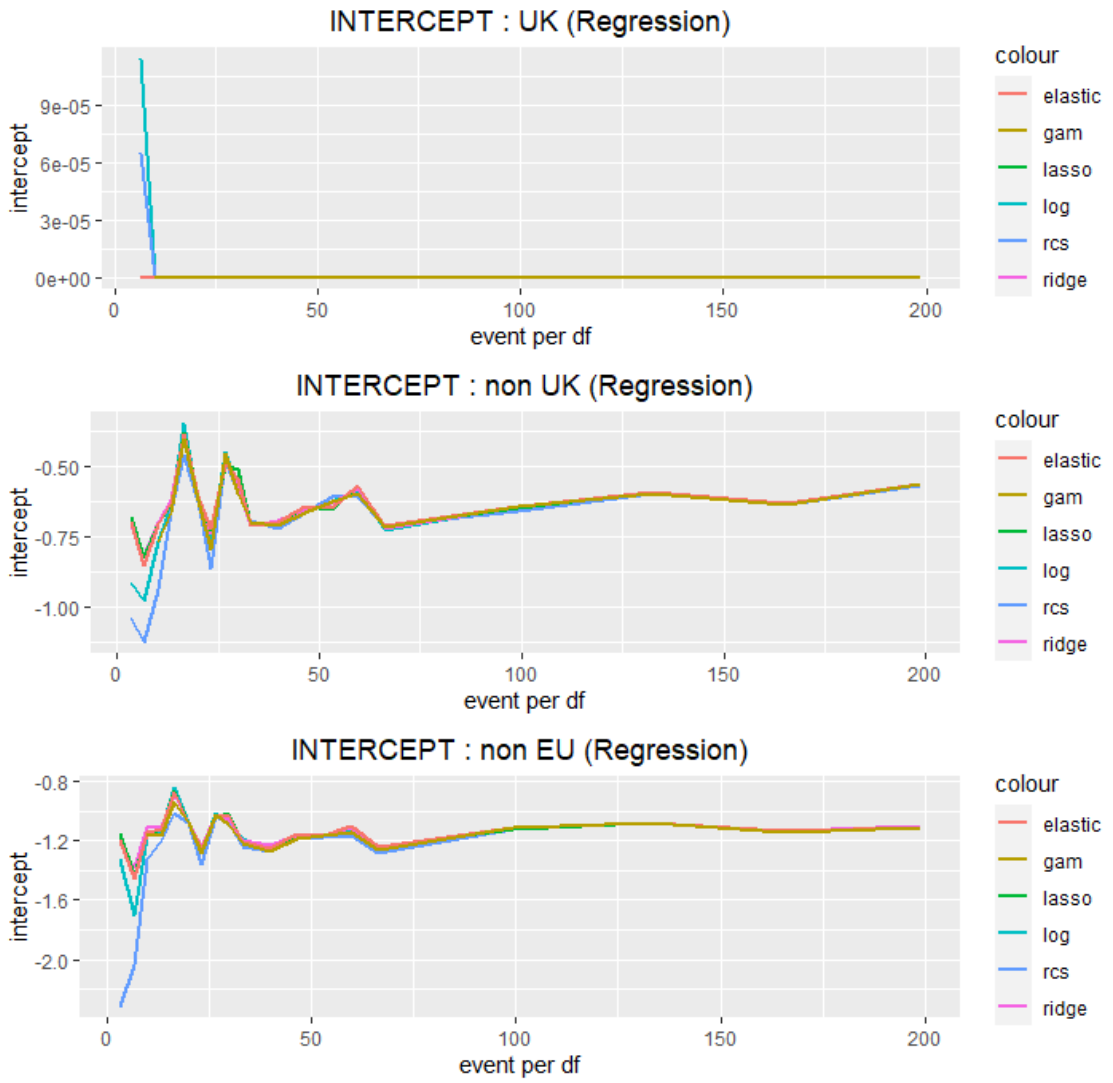


Figure 4.10 Calibration intercept in regression models

4.2 Discussion

This study firstly aimed to compare traditional regression methods and machine learning methods in external validation. Model creation was based on different sample sizes between 50 and 3000, to see the differences not only between models, but also between sample sizes. Performances were assessed in terms of R-squared, Brier score, discrimination and calibration. EPV values are calculated for each sample size, moreover, they are given in the x-axis of the plots. According to Steyerberg [1] and Harrell [2], EPV has great importance in clinical studies. Models with EPV below 10 should not be preferred. In order to see the differences of the performances of the models with lower EPV values (where EPV<30) better, sample sizes were increased by 50 until 500. The increment was 100 from 500 to 1000, and 500 from 1000 to 3000. Tables of the results, plots of the results (of all models together), calibration plots of models (not all) and

programming codes are given in the appendix section. Plots of the results given in this section are separated into two: Machine learning results and regression results.

Missing values were removed, since they were ~10% of the size of IST data. After the removal, splitting IST data was based on geographical setting. Thence, splits were not random. In the programming setting, sample creation from UK data was random ("set.seed(4)" was applied used during all the study).

RF was the relatively underperformed algorithm in the study. In internal validation, R-squared and AUC results of RF increased unexpectedly for higher EPV values, whereas these performances were one of the worst among all methods in external validations. Calibration slope of RF was below 1 showing the sign of overfitting [34].

The parameters of NN were changed to give healthy results in all samples, because default options for NN did not work for all the samples regardless of the sample size. If EPV values were much higher, NN could maybe perform better. GBM and XGB gave the best performances among machine learning methods. Only GBM and XGB could sometimes compete regression methods. On the other hand, SVM has been one of the worst performing three methods in the study when all performance measures are considered.

Regression models show more stability especially for higher EPV (>30) for all performance measures. Ridge performed best overall, followed by elastic net regression. Restricted cubic splines did not improve the results of logistic regression. GAM needed more sample size to operate, thus two models with sample sizes 50 and 100 were not used.

In addition to parameter adjustment, usage of small number of predictors (9 predictors) may be the reason for machine learning methods to perform worse than regression methods. It is suggested that flexible machine learning methods perform better than regression techniques when large numbers of predictors are counted (high dimensional data) [44, 45]. Besides, research [23] suggests that machine learning requires more data than regression.

REFERENCES

[1] E. W. Steyerberg, *Clinical Prediction Models*. Springer, 2019.

[2] F. E. Harrell, *Regression Modeling Strategies.* Springer, 2015.

[3] E. Núñez, E. W. Steyerberg, and J. Núñez, "Regression modelling strategies," *Revista Española de Cardiología*, vol. 64, no. 6, pp. 501–507, Jun. 2011, doi: 10.1016/j.recesp.2011.01.019.

[4] P. Sandercock, M. Niewada, and A. Czlonkowska. "International Stroke Trial database (version 2)", [dataset]. University of Edinburgh. Department of Clinical Neurosciences, 2011. Available: https://doi.org/10.7488/ds/104

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer, New York, 2009.

[6] G. James, D. Witten, T. J. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning : with applications in R*. New York: Springer, 2013.

[7] B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M. J. Pencina, and E. W. Steyerberg, "A calibration hierarchy for risk models was defined: from utopia to empirical data," *Journal of Clinical Epidemiology*, vol. 74, pp. 167–176, Jun. 2016, doi: 10.1016/j.jclinepi.2015.12.005.

[8] P. Ravani, P. Parfrey, V. Gadag, F. Malberti, and B. Barrett, "Clinical research of kidney diseases III: principles of regression and modelling," *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association*, vol. 22, no. 12, pp. 3422–3430, Dec. 2007, doi: 10.1093/ndt/gfm777.

[9] D. G. Kleinbaum and M. Klein, *Logistic Regression : A Self-Learning Text*. New York, Springer, 2010.

[10] R. Croxford, "Restricted Cubic Spline Regression: A Brief Introduction", 2016.

[11] M. V. Ozanne, "Comparison of Shrunken Regression Methods for Major Elemental Analysis of Rocks Using Laser-Induced Breakdown Spectroscopy (LIBS)," 2012. Accessed: Apr. 26, 2022. [Online]. Available: https://ida.mtholyoke.edu/bitstream/handle/10166/1041/Thesis%20Final%20Draft.pdf?sequence=1

[12] C. S. Göbl, L. Bozkurt, A. Tura, G. Pacini, A. Kautzky-Willer, and M. Mittlböck, "Application of Penalized Regression Techniques in Modelling Insulin Sensitivity by Correlated Metabolic Parameters," *PLOS ONE*, vol. 10, no. 11, p. e0141524, Nov. 2015, doi: 10.1371/journal.pone.0141524.

[13] B. Qian *et al.*, "Orchestrating the Development Lifecycle of Machine Learning-based IoT Applications," *ACM Computing Surveys*, vol. 53, no. 4, pp. 1–47, Sep. 2020, doi: 10.1145/3398020.

[14] H. S. Kapoor, K. Jain, and S. K. Sharma, "Generalized Additive Model for Evaluation of Premium for Diabetic Patients," *Journal of Advances in Applied Mathematics*, vol. 1, no. 3, Jul. 2016, doi: 10.22606/jaam.2016.13002.

[15] A. Nikhil, and B. Techstudent. “A Predictive Analytic study on stock market trend by supervised machine learning algorithms,” 2020.

[16] C. Ulas and M. Çetin, “Incorporation of a language model into a brain computer interface based speller through HMMs,” *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, doi: 10.1109/ICASSP.2013.6637828.

[17] I. Baturynska and K. Martinsen, “Prediction of geometry deviations in additive manufactured parts: comparison of linear regression with machine learning algorithms,” *Journal of Intelligent Manufacturing*, Apr. 2020, doi: 10.1007/s10845-020-01567-0.

[18] N. Sapountzoglou, J. Lago, and B. Raison, “Fault diagnosis in low voltage smart distribution grids using gradient boosting trees,” *Electric Power Systems Research*, vol. 182, p. 106254, May 2020, doi: 10.1016/j.epsr.2020.106254.

[19] R. Bartzatt, “Determination of Dermal Permeability Coefficient (Kp) by Utilizing Multiple Descriptors in Artificial Neural Network Analysis and Multiple Regression Analysis,” *Journal of Scientific Research and Reports*, vol. 3, no. 22, pp. 2884–2899, Jan. 2014, doi: 10.9734/jsrr/2014/13125.

[20] S. N. Wood, *Generalized Additive Models: An Introduction with R*. CRC Press, 2017.

[21] T. Hastie and R. Tibshirani, “Generalized Additive Models,” *Statistical Science*, vol. 1, no. 3, pp. 297–310, Aug. 1986, doi: 10.1214/ss/1177013604.

[22] A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, Jul. 1959, doi: 10.1147/rd.33.0210.

[23] T. van der Ploeg, P. C. Austin, and E. W. Steyerberg, “Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints,” *BMC Medical Research Methodology*, vol. 14, no. 1, Dec. 2014, doi: 10.1186/1471-2288-14-137.

[24] T. O. Ayodele, *Machine Learning Overview*. IntechOpen, 2010. Accessed: Apr. 26, 2022. [Online]. Available: https://www.intechopen.com/chapters/10683

[25] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/a:1010933404324.

[26] M. Mijwil, A. Esen, and A. Alsaadi. “Overview of Neural Networks,” 2019.

[27] E. Çomak, “Destek Vektör Makineleri Çoklu Sınıf Problemleri İçin Çözüm Önerileri,” Available: https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=wxbpdaC5s6csBwBsKQHTrw&no=4EpPet-PBm-Mw12OijFV7w (accessed Apr. 26, 2022).

[28] S. Malik, “XGBoost: A Deep Dive Into Boosting - DZone AI,” Available: https://dzone.com/articles/xgboost-a-deep-dive-into-boosting

[29] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, 2016, doi: 10.1145/2939672.2939785.

[30] S. Zhu and F. Zhu, “Cycling comfort evaluation with instrumented probe bicycle,” *Transportation Research Part A: Policy and Practice*, vol. 129, pp. 217–231, Nov. 2019, doi: 10.1016/j.tra.2019.08.009.

[31] H. Mo, H. Sun, J. Liu, and S. Wei, “Developing window behavior models for residential buildings using XGBoost algorithm,” *Energy and Buildings*, vol. 205, Dec. 2019, Accessed: Apr. 26, 2022. [Online]. Available: https://discovery.ucl.ac.uk/id/eprint/10085493/

[32] A. Członkowska *et al.*, “High early case fatality after ischaemic stroke in Poland: exploration of possible explanations in the International Stroke Trial,” *Journal of the Neurological Sciences*, vol. 202, no. 1–2, pp. 53–57, Oct. 2002, doi: 10.1016/s0022-510x(02)00203-4.

[33] C. Wang, C. Deng, and S. Wang, “Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost,” *Pattern Recognition Letters*, Jun. 2020, doi: 10.1016/j.patrec.2020.05.035.

[34] E. W. Steyerberg *et al.*, “Assessing the Performance of Prediction Models,” *Epidemiology*, vol. 21, no. 1, pp. 128–138, Jan. 2010, doi: 10.1097/ede.0b013e3181c30fb2.

[35] E. W. Steyerberg and Y. Vergouwe, “Towards better clinical prediction models: seven steps for development and an ABCD for validation,” *European Heart Journal*, vol. 35, no. 29, pp. 1925–1931, Jun. 2014, doi: 10.1093/eurheartj/ehu207.

[36] P. Galdi and R. Tagliaferri, “Data Mining: Accuracy and Error Measures for Classification and Prediction,” *Encyclopedia of Bioinformatics and Computational Biology*, pp. 431–436, 2019, doi: 10.1016/B978-0-12-809633-8.20474-3.

[37] M. van Smeden *et al.*, “No rationale for 1 variable per 10 events criterion for binary logistic regression analysis,” *BMC Medical Research Methodology*, vol. 16, no. 1, Nov. 2016, doi: 10.1186/s12874-016-0267-3.

[38] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, “Calibration: the Achilles heel of predictive analytics,” *BMC Medicine*, vol. 17, no. 1, Dec. 2019, doi: 10.1186/s12916-019-1466-7.

[39] M. van Smeden *et al.*, “Sample size for binary logistic prediction models: Beyond events per variable criteria,” *Statistical Methods in Medical Research*, vol. 28, no. 8, pp. 2455–2474, Jul. 2018, doi: 10.1177/0962280218784726.

[40] A. K. Ma and G. Liu, “Machine Learning for Predicting Delayed Onset Trauma following Ischemic Stroke,” Available: https://www.semanticscholar.org/paper/Machine-Learning-for-Predicting-Delayed-Onset-Ma-Liu/6323faf68ca03e7f39392d15a7aceaa38f177ea4 (accessed Apr. 26, 2022).

[41] P. C. Austin and E. W. Steyerberg, “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*, vol. 26, no. 2, pp. 796–808, Nov. 2014, doi: 10.1177/0962280214558972.

[42] H. Chen and D. Boning, “Machine Learning Approaches for IC Manufacturing Yield Enhancement,” *Machine Learning in VLSI Computer-Aided Design*, 2019, doi: 10.1007/978-3-030-04666-8_6.

[43] S. K. Kiangala and Z. Wang, "An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment," *Machine Learning with Applications*, vol. 4, p. 100024, Jun. 2021, doi: 10.1016/j.mlwa.2021.100024.

[44] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319:1317.

[45] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18.

[46] Steyerberg EW, Harrell FE Jr, Borsboom GJ et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001; 54: 774–781

[47] Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016 Jan;69:245-7. doi: 10.1016/j.jclinepi.2015.04.005.

[48] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6):515-524.

A
R PROGRAMMING CODES

 # Libraries

```
library(ggrepel)
library(gridExtra)
library(ggpubr)
library(xgboost)
library(rpart)
library(nnet)
library(caTools)
```

```
library(foreign)
library(randomForest)
library(kernlab)
library(RCurl)
library(tidyverse)
library(caret)
library(glmnet)
library(mgcv)
library(curl)
library(dplyr)
library(stringr)
library(maps)
library(ggplot2)
library(data.table)
library(rms)
library(CalibrationCurves)
library(gbm)
library(DesignLibrary)
library(data.table)
library(mlr)
library(writexl)
```

 # Downloading data

```
tezdatasi=fread('https://datashare.ed.ac.uk/bitstream/handle/10283/124/IST_corrected.csv?sequence=5&isAllowed=y')
```

 # Choosing the variables

```
veri<-tezdatasi %>% select(AGE,SEX,RDELAY,RCONSC,RATRIAL,RVISINF,RSBP,RXASP,DDIAGHA,DALIVE,COUNTRY,CNTRYNUM)
```

 # Removing NA (also unknown values)

```
veri$DDIAGHA <- ifelse(veri$DDIAGHA=="U", "",veri$DDIAGHA)
veri$DALIVE <- ifelse(veri$DALIVE=="U", "",veri$DALIVE)
```

```
veri <- veri %>% mutate_all(na_if,"")
veri=na.omit(veri)
```

 # Making data binary

```
veri$SEX <- ifelse(veri$SEX=="F", 1, 0)
veri$RCONSC=ifelse(veri$RCONSC=="F", 1, 0)
veri$RATRIAL=ifelse(veri$RATRIAL=="Y", 1, 0)
veri$RVISINF=ifelse(veri$RVISINF=="Y", 1, 0)
veri$RXASP=ifelse(veri$RXASP=="Y", 1, 0)
veri$DDIAGHA=ifelse(veri$DDIAGHA=="Y", 1, 0)
veri$DALIVE=ifelse(veri$DALIVE=="N", 1, 0)
```

 # Creating map data (named as “harita”)

```
group_by_veri <-   veri%>%  group_by(COUNTRY) %>%
  summarise(unique_customer = n_distinct(COUNTRY),
            number_of_entries = n(),      unique_product = n_distinct(COUNTRY)  )
group_by_veri<-group_by_veri%>%arrange(desc(number_of_entries))
harita <- data.frame(Ulke = group_by_veri$COUNTRY, Sayi = group_by_veri$number_of_entries)
```

 # Plotting the map

```
WorldData <- map_data('world') %>% filter(region != "Antarctica") %>% fortify
df <- data.frame(region=c("UK","Italy","Switzerland","Poland","Netherlands", "Sweden","Australia","Argentina","Norway", "Spain", "Czech Republic" ,"New Zealand" , "Portugal" ,"Belgium" ,"Turkey",  "Austria" ,"India" ,"Greece" ,"Singapore", "USA" , "Canada", "Hong Kong" , "Israel", "Hungary", "Slovakia", "Finland", "Brazil" ,"South Africa", "Chile", "Ireland",  "Slovenia", "Denmark" ,"Sri Lanka" , "Romania" ,"Japan" ,"France"), value=harita$Sayi, stringsAsFactors=T)
p <- ggplot() +
  geom_map(data = WorldData, map = WorldData,
           aes(x = long, y = lat, group = group, map_id=region),
```

```
        fill = "white", colour = "#7f7f7f", size=0.5) +
  geom_map(data = df, map=WorldData,
        aes(fill=value, map_id=region),
        colour="#7f7f7f", size=0.5) +
  coord_map("rectangular", lat0=0, xlim=c(-180,180), ylim=c(-60, 90)) +
  scale_fill_continuous(low="thistle2", high="darkred", guide="colorbar") +
  scale_y_continuous(breaks=c()) +
  scale_x_continuous(breaks=c()) +
  labs(fill="Patients", title="Countries Including the Patients in International Stroke Trial
Data", x="", y="") +   theme_bw() +
  theme(axis.text.x = element_text(size = 12, color = "black"),
       axis.text.y = element_text(size = 12, color = "black"),
       panel.grid.major = element_line(color = gray(0.5), linetype = "dashed", size = 0.5),
       panel.background = element_rect(fill = "lightsteelblue1"))
p        # plots the map
```

 # Creating new data to create models

```
veri2<-(veri[,1:10]) # country and country numbers were extracted
ukveri<-rbind(veri[veri$COUNTRY=="UK"],veri[veri$COUNTRY=="EIRE"])
ukveri<-ukveri[,1:10]          # UK data creation
nonukveri<-rbind(veri[veri$COUNTRY=="ITAL"],veri[veri$COUNTRY=="FINL"],
                 veri[veri$COUNTRY=="CZEC"],veri[veri$COUNTRY=="HUNG"],
                 veri[veri$COUNTRY=="PORT"],veri[veri$COUNTRY=="NETH"],
                veri[veri$COUNTRY=="SWIT"],veri[veri$COUNTRY=="AUST"],
                veri[veri$COUNTRY=="SLOV"],veri[veri$COUNTRY=="SPAI"],
                veri[veri$COUNTRY=="NORW"],veri[veri$COUNTRY=="SWED"],
                veri[veri$COUNTRY=="BELG"],veri[veri$COUNTRY=="GREE"],
```

```
            veri[veri$COUNTRY=="POLA"],veri[veri$COUNTRY=="TURK"],
            veri[veri$COUNTRY=="SLOK"],veri[veri$COUNTRY=="DENM"],
            veri[veri$COUNTRY=="FRAN"],veri[veri$COUNTRY=="ROMA"])
        # non UK data creation
nonukveri<-nonukveri[,1:10]
noneuveri<-rbind(veri[veri$COUNTRY=="AUSL"],veri[veri$COUNTRY=="USA"],
            veri[veri$COUNTRY=="NEW"],veri[veri$COUNTRY=="CHIL"],
            veri[veri$COUNTRY=="SOUT"],veri[veri$COUNTRY=="ISRA"],
            veri[veri$COUNTRY=="HONG"],veri[veri$COUNTRY=="CANA"],
            veri[veri$COUNTRY=="BRAS"],veri[veri$COUNTRY=="ARGE"],
            veri[veri$COUNTRY=="INDI"],
            veri[veri$COUNTRY=="SRI"],veri[veri$COUNTRY=="SING"],
            veri[veri$COUNTRY=="JAPA"])
# non EU data creation
noneuveri<-noneuveri[,1:10]
```

 # Sample Creation

```
# This process is applied for following sample sizes : 50, 100, 150, 200, 250, 300, 350,
#400, 450, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000. To put it short, just
#first and last ones are shared here.
# set.seed(4) is applied throughout this study. Both for samples and models.
set.seed(4)
sample_rows <- sample(1:nrow(ukveri), 50)
ukveri50 <- ukveri[sample_rows, ]
sample_rows <- sample(1:nrow(ukveri), 3000)
ukveri3000 <- ukveri[sample_rows, ]
```

 # Model Creation (models are created for all sub-samples of ukveri)

logmodel50=lrm(DALIVE~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA+RVISINF+SEX,data=ukveri50,x=T,y=T) *# logistic regression*

logmodel50rcs=lrm(DALIVE~rcs(AGE,5)+rcs(RDELAY,5)+RCONSC+RATRIAL+rcs(RSBP,5)+RXASP+DDIAGHA+RVISINF+SEX,data=ukveri50,x=T,y=T) *#logistic #regression with RCS*

gammodel150=gam(DALIVE~s(AGE,k=9)+s(RDELAY,k=9)+s(RSBP,k=9)+RATRIAL+RCONSC+RXASP+DDIAGHA+RVISINF+SEX,data=ukveri150,family="binomial") *# generalized additive model # Only GAM does not include sample sizes 50 and 100.*

set.seed(4)

rf50=randomForest(as.factor(DALIVE)~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA+RVISINF+SEX,data=ukveri50) *# Random Forest*

gbm50=gbm(DALIVE~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA+RVISINF+SEX, data=ukveri50, distribution="bernoulli", bag.fraction=0.5, train.fraction=1.0, cv.folds=0, keep.data=TRUE) *# GBM*

set.seed(4)

#NN

nn50=nnet(as.factor(DALIVE)~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA+RVISINF+SEX, data = ukveri50,size=10,maxit=750,decay=5e-2)

 # Creating Other Models and Calibration Plots

 # Here is where predicted probabilities are validated against binary events.

XGB

set.seed(4)

trains = ukveri50

tests = ukveri

train_x = data.matrix(trains[1:nrow(trains),1:9])

train_y = trains[1:nrow(trains),10]

test_x = data.matrix(tests[1:nrow(tests),1:9])

test_y = tests[1:nrow(tests),10]

```
xgb_train = xgb.DMatrix(data = train_x, label = train_y$DALIVE)

xgb_test = xgb.DMatrix(data = test_x, label = test_y$DALIVE)

watchlist = list(train=xgb_train, test=xgb_test)

model = xgb.train(data = xgb_train, watchlist=watchlist,max.depth = 3, nrounds = 70)

modelnew=as.list(model)

mintestrmse<-which.min(modelnew$evaluation_log$test_rmse)

final = xgboost(data = xgb_train, max.depth = 3, nrounds = mintestrmse, verbose = 0, objective = "binary:logistic")

predxgb <- predict(final, newdata = as.matrix(test_x))

valxgb_uk_50 <- val.prob.ci.2(predxgb,y=tests$DALIVE,pl=T,main="XGB n=50, uk internal")

valxgb_uk_50

# As a different example, ridge model with 800 sample size will be internally validated:

x=model.matrix(DALIVE~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA+RVISINF+SEX, ukveri800)[,-1]

y <- ukveri800$DALIVE

set.seed(4)

cv <- cv.glmnet(x, y, alpha = 0,family="binomial")

model <- glmnet(x, y, alpha = 0, lambda = cv$lambda.min,family="binomial")

x.test=model.matrix(DALIVE~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA+RVISINF+SEX, ukveri)[,-1]

predictions <- model %>% predict(x.test) %>% as.vector()

val.ridge800uk=val.prob.ci.2(y=ukveri$DALIVE,logit=predictions,pl=T,main="RIDGE uk n=800 ")

# As a different example, lasso model with 800 sample size will be externally validated in #non-UK data:

x=model.matrix(DALIVE~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA+RVISINF+SEX, ukveri800)[,-1]
```

```
y <- ukveri800$DALIVE

set.seed(4)

cv <- cv.glmnet(x, y, alpha = 1,family="binomial")

model <- glmnet(x, y, alpha = 1, lambda = cv$lambda.min,family="binomial")

x.test=model.matrix(DALIVE~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA+RVISINF+SEX, nonukveri)[,-1]

predictions <- model %>% predict(x.test) %>% as.vector()

val.lasso800nonuk=val.prob.ci.2(y=nonukveri$DALIVE,logit=predictions,pl=T,main=" LASSO non uk, n=800 ")

# As a different example, elastic net model with 800 sample size will be externally validated in #non-EU data:

x=model.matrix(DALIVE~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA+RVISINF+SEX, ukveri800)[,-1]

y <- ukveri800$DALIVE

set.seed(4)

cv <- cv.glmnet(x, y, alpha = 0.5,family="binomial")

model <- glmnet(x, y, alpha = 0.5, lambda = cv$lambda.min,family="binomial")

x.test=model.matrix(DALIVE~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA+RVISINF+SEX, noneuveri)[,-1]

predictions <- model %>% predict(x.test) %>% as.vector()

val.elastic800noneu=val.prob.ci.2(y=noneuveri$DALIVE,logit=predictions,

pl=T,main="ELASTIC_NET non eu , n=800 ")

# Random Forest

pred.valid.rf50 <- predict(rf50,newdata=ukveri,type="prob")[,2]

val.rf50uk=val.prob.ci.2(pred.valid.rf50,y=ukveri$DALIVE,pl=T,main="n=50, UK",CL.BT = F)

# GBM

pred.valid.gbm50 <- predict(gbm50,newdata=ukveri,type="response")
```

```
val.gbm50uk <- val.prob.ci.2(pred.valid.gbm50,y=ukveri$DALIVE,pl=T
,main="GBM n=50, uk internal")
# Logistic Regression with RCS
rcsplot50uk=val.prob.ci.2(y=ukveri$DALIVE,
logit=predict(logmodel50rcs,ukveri),pl=T,main="RCS n=50,  internal UK")
# Logistic Regression without RCS
plot50uk<-val.prob.ci.2(y=ukveri$DALIVE,
logit=predict(logmodel50,ukveri),pl=T,main="LOGISTIC n=50,  internal UK")
# GAM
plotgammodel150uk<-val.prob.ci.2(y=ukveri$DALIVE,
predict(gammodel150,ukveri,type="response"),pl=T,
main="GAM n=150,  internal UK")
# NN
plotnn50uk<-val.prob.ci.2(y=ukveri$DALIVE,predict(nn50,ukveri),pl=T,
main="NN n=50, uk internal")
rbf <- rbfdot(sigma=0.1)
# SVM
svm50<-
ksvm(DALIVE~AGE+RDELAY+RCONSC+RATRIAL+RSBP+RXASP+DDIAGHA
+RVISINF+SEX,   data=ukveri50, kernel=rbf,C=10,type="C-bsvc",prob.model=TRUE)
predsvm=predict(   svm50,   ukveri50,   type="probabilities")
val.svm50uk <- val.prob.ci.2(predsvm[,2],y=ukveri50$DALIVE, main="SVM n=50, uk
internal")
```

 # Plotting The Results

#After results from calibration plots are written into different excel files, they are #imported into R. Here in the following example, only Brier score "machine learning" #results for internal validation are plotted.

```
brierlerinternalukml<-
ggplot(nninternaluk,aes(x=epv,y=brier))+geom_line(aes(color='nn '),size=1) +
    geom_line(data=svminternaluk,aes(color='svm '),size=1) +
    geom_line(data=rfinternaluk,aes(color='rf '),size=1) +
    geom_line(data=gbminternaluk,aes(color='gbm '),size=1) +
    geom_line(data=xgbinternaluk,aes(color='xgb '),size=1)
brierlerinternalukml<-  brierlerinternalukml+ggtitle("BRIER SCORE : UK (Machine Learning)")+  theme(plot.title = element_text(hjust = 0.5)) + labs(x = "event per df")
```

B
TABLES AND FIGURES OF RESULTS

Table B.1 Results of logistic regression without RCS

Sample info	N	50	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	3.3	6.6	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53.0	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	-	0.18	0.19	0.21	0.22	0.20	0.22	0.23	0.21	0.21	0.20	0.22	0.22	0.22	0.21	0.21	0.22	0.22	0.22
	AUC	-	0.78	0.76	0.70	0.69	0.72	0.66	0.66	0.73	0.68	0.72	0.68	0.67	0.68	0.68	0.69	0.69	0.69	0.70
	R-squared	-	0.33	0.27	0.19	0.16	0.21	0.10	0.11	0.21	0.12	0.19	0.14	0.12	0.14	0.13	0.15	0.14	0.14	0.15
	intercept	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	slope	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
External validation non-UK	Brier	0.30	0.28	0.27	0.27	0.25	0.26	0.27	0.26	0.26	0.26	0.26	0.26	0.25	0.25	0.26	0.26	0.26	0.26	0.25
	AUC	0.61	0.62	0.63	0.60	0.61	0.63	0.61	0.59	0.62	0.62	0.63	0.62	0.62	0.62	0.62	0.63	0.62	0.63	0.62
	R-squared	0.05	0.05	0.07	0.04	0.05	0.07	0.05	0.03	0.06	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.06	0.07	0.07
	intercept	-0.91	-0.98	-0.77	-0.64	-0.34	-0.62	-0.78	-0.45	-0.58	-0.70	-0.71	-0.66	-0.62	-0.58	-0.72	-0.64	-0.59	-0.63	-0.56
	slope	0.23	0.25	0.38	0.39	0.57	0.49	0.64	0.47	0.51	0.68	0.52	0.64	0.71	0.66	0.68	0.63	0.61	0.64	0.61
External validation non-EU	Brier	0.28	0.30	0.25	0.26	0.23	0.25	0.28	0.25	0.25	0.26	0.26	0.26	0.26	0.25	0.27	0.25	0.25	0.25	0.25
	AUC	0.65	0.64	0.66	0.65	0.68	0.67	0.66	0.64	0.67	0.68	0.69	0.68	0.69	0.67	0.68	0.68	0.68	0.68	0.69
	R-squared	0.08	0.06	0.10	0.08	0.11	0.11	0.09	0.08	0.11	0.11	0.12	0.12	0.12	0.11	0.11	0.12	0.11	0.11	0.12
	intercept	-1.33	-1.70	-1.17	-1.13	-0.84	-1.09	-1.26	-1.02	-1.08	-1.19	-1.27	-1.19	-1.17	-1.12	-1.25	-1.11	-1.08	-1.14	-1.11
	slope	0.30	0.26	0.48	0.62	0.90	0.63	0.85	0.80	0.70	0.96	0.73	0.89	1.02	0.87	0.90	0.90	0.86	0.87	0.88

Table B.2 Results of logistic regression with RCS

Sample info	N	50	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	3.3	6.6	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	-	0.15	0.18	0.20	0.21	0.20	0.21	0.22	0.20	0.21	0.20	0.21	0.22	0.22	0.21	0.21	0.21	0.21	0.22
	AUC	-	0.85	0.81	0.74	0.74	0.74	0.69	0.69	0.74	0.70	0.73	0.69	0.68	0.69	0.69	0.70	0.69	0.70	0.70
	R-squared	-	0.46	0.34	0.24	0.23	0.23	0.14	0.14	0.23	0.15	0.21	0.15	0.13	0.15	0.15	0.15	0.15	0.15	0.15
	intercept	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	slope	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
External validation non-UK	Brier	0.38	0.32	0.29	0.28	0.26	0.27	0.28	0.27	0.26	0.27	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.25
	AUC	0.56	0.60	0.61	0.58	0.60	0.62	0.60	0.57	0.61	0.60	0.62	0.61	0.62	0.61	0.62	0.62	0.62	0.62	0.62
	R-squared	0.01	0.04	0.05	0.03	0.04	0.05	0.04	0.02	0.06	0.04	0.07	0.05	0.06	0.05	0.06	0.06	0.06	0.06	0.06
	intercept	-1.04	-1.13	-0.94	-0.66	-0.46	-0.62	-0.86	-0.47	-0.59	-0.69	-0.72	-0.66	-0.60	-0.61	-0.71	-0.66	-0.59	-0.64	-0.56
	slope	0.04	0.13	0.24	0.26	0.38	0.38	0.46	0.30	0.43	0.49	0.49	0.54	0.63	0.54	0.59	0.60	0.57	0.59	0.59
External validation non-EU	Brier	0.36	0.32	0.28	0.27	0.25	0.25	0.29	0.26	0.25	0.27	0.26	0.26	0.26	0.26	0.27	0.25	0.25	0.26	0.25
	AUC	0.60	0.63	0.65	0.62	0.64	0.66	0.64	0.62	0.66	0.67	0.68	0.67	0.68	0.66	0.67	0.67	0.67	0.67	0.68
	R-squared	0.03	0.05	0.07	0.06	0.07	0.10	0.07	0.05	0.10	0.10	0.12	0.10	0.12	0.10	0.11	0.11	0.11	0.11	0.12
	intercept	-2.32	-2.03	-1.33	-1.20	-1.02	-1.08	-1.36	-1.04	-1.08	-1.24	-1.27	-1.19	-1.18	-1.17	-1.29	-1.12	-1.08	-1.15	-1.12
	slope	0.07	0.16	0.30	0.39	0.57	0.53	0.62	0.50	0.62	0.80	0.68	0.79	0.95	0.75	0.81	0.83	0.80	0.82	0.86

Table B.3 Results of generalized additive models

Sample info	N	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53.0	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	0.193	0.203	0.204	0.203	0.219	0.220	0.204	0.208	0.204	0.216	0.218	0.215	0.212	0.213	0.215	0.214	0.215
	AUC	0.761	0.721	0.737	0.725	0.666	0.685	0.733	0.695	0.721	0.683	0.672	0.692	0.685	0.693	0.691	0.694	0.696
	R-squared	0.270	0.211	0.236	0.216	0.111	0.134	0.219	0.149	0.194	0.137	0.122	0.151	0.137	0.145	0.143	0.146	0.151
	intercept	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	slope	1.000	1.055	1.121	1.022	1.046	1.080	1.023	1.066	1.000	1.009	1.017	1.030	1.027	1.000	1.003	1.011	1.001
External validation non-UK	Brier	0.269	0.270	0.255	0.258	0.269	0.260	0.257	0.263	0.260	0.259	0.254	0.255	0.260	0.256	0.255	0.256	0.252
	AUC	0.626	0.596	0.601	0.626	0.615	0.580	0.620	0.615	0.627	0.623	0.624	0.615	0.624	0.625	0.621	0.624	0.622
	R-squared	0.07	0.04	0.04	0.07	0.06	0.03	0.06	0.05	0.07	0.06	0.07	0.06	0.07	0.07	0.06	0.07	0.07
	intercept	-0.77	-0.66	-0.40	-0.62	-0.79	-0.45	-0.59	-0.70	-0.71	-0.66	-0.62	-0.59	-0.71	-0.64	-0.59	-0.63	-0.56
	slope	0.38	0.34	0.40	0.47	0.65	0.39	0.50	0.60	0.52	0.63	0.71	0.60	0.66	0.63	0.60	0.63	0.61
External validation non-EU	Brier	0.253	0.261	0.241	0.245	0.277	0.253	0.247	0.266	0.260	0.258	0.256	0.256	0.268	0.250	0.249	0.254	0.248
	AUC	0.664	0.644	0.648	0.670	0.657	0.633	0.673	0.670	0.686	0.678	0.692	0.666	0.678	0.682	0.675	0.678	0.685
	R-squared	0.10	0.07	0.07	0.11	0.09	0.07	0.11	0.10	0.12	0.11	0.13	0.10	0.11	0.12	0.11	0.11	0.12
	intercept	-1.17	-1.17	-0.94	-1.08	-1.28	-1.03	-1.09	-1.22	-1.27	-1.18	-1.17	-1.15	-1.27	-1.11	-1.08	-1.15	-1.11
	slope	0.48	0.55	0.57	0.62	0.86	0.67	0.70	0.87	0.73	0.88	1.03	0.83	0.90	0.90	0.84	0.86	0.88

Table B.4 Results of lasso

Sample info	N	50	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	3.3	6.6	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	-	0.20	0.20	0.21	0.22	0.20	0.22	0.23	0.21	0.22	0.20	0.22	0.22	0.22	0.21	0.21	0.22	0.22	0.22
	AUC	-	0.75	0.76	0.70	0.69	0.71	0.66	0.66	0.73	0.67	0.72	0.68	0.67	0.68	0.68	0.69	0.69	0.69	0.70
	R-squared	-	0.29	0.26	0.18	0.16	0.20	0.10	0.10	0.20	0.12	0.19	0.13	0.11	0.13	0.13	0.14	0.14	0.14	0.15
	intercept	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	slope	-	2.04	1.36	1.35	1.32	1.17	1.25	1.26	1.27	1.30	1.05	1.35	1.24	1.19	1.12	1.11	1.04	1.04	1.02
External validation non-UK	Brier	0.27	0.27	0.26	0.26	0.24	0.25	0.27	0.25	0.25	0.26	0.26	0.26	0.26	0.25	0.26	0.26	0.26	0.26	0.25
	AUC	0.58	0.62	0.63	0.60	0.62	0.63	0.61	0.60	0.62	0.62	0.63	0.62	0.63	0.62	0.62	0.63	0.62	0.63	0.62
	R-squared	0.02	0.07	0.07	0.04	0.06	0.07	0.05	0.04	0.06	0.06	0.07	0.06	0.07	0.07	0.07	0.07	0.06	0.07	0.07
	intercept	-0.68	-0.82	-0.70	-0.63	-0.38	-0.61	-0.73	-0.49	-0.51	-0.71	-0.70	-0.64	-0.65	-0.57	-0.71	-0.64	-0.59	-0.63	-0.56
	slope	0.64	0.62	0.54	0.52	0.76	0.58	0.80	0.73	0.68	0.87	0.55	0.87	0.91	0.81	0.76	0.70	0.63	0.66	0.62
External validation non-EU	Brier	0.27	0.29	0.25	0.26	0.23	0.25	0.28	0.25	0.24	0.27	0.26	0.26	0.26	0.25	0.27	0.25	0.25	0.25	0.25
	AUC	0.64	0.63	0.68	0.65	0.68	0.68	0.66	0.66	0.68	0.68	0.69	0.68	0.69	0.68	0.68	0.68	0.68	0.68	0.69
	R-squared	0.07	0.08	0.11	0.08	0.10	0.11	0.09	0.09	0.11	0.11	0.12	0.12	0.12	0.12	0.12	0.12	0.11	0.11	0.12
	intercept	-1.15	-1.43	-1.15	-1.12	-0.87	-1.08	-1.24	-1.05	-1.02	-1.21	-1.26	-1.15	-1.16	-1.10	-1.24	-1.12	-1.08	-1.13	-1.11
	slope	1.09	0.67	0.71	0.85	1.20	0.77	1.14	1.14	0.96	1.27	0.76	1.25	1.24	1.09	1.03	0.99	0.89	0.91	0.90

Table B.5 Results of ridge

Sample info	N	50	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	3.3	6.6	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	-	0.19	0.20	0.21	0.22	0.21	0.22	0.23	0.21	0.22	0.21	0.22	0.22	0.22	0.21	0.21	0.22	0.22	0.22
	AUC	-	0.78	0.76	0.70	0.69	0.72	0.66	0.66	0.73	0.68	0.72	0.68	0.67	0.68	0.68	0.69	0.69	0.69	0.70
	R-squared	-	0.32	0.27	0.18	0.16	0.21	0.10	0.11	0.21	0.12	0.19	0.13	0.12	0.14	0.13	0.15	0.14	0.14	0.15
	intercept	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	slope	-	1.76	1.63	1.53	1.43	1.33	1.45	1.32	1.19	1.43	1.23	1.26	1.22	1.16	1.13	1.10	1.06	1.07	1.06
External validation non-UK	Brier	0.27	0.27	0.26	0.26	0.24	0.25	0.27	0.25	0.25	0.26	0.26	0.26	0.25	0.25	0.26	0.26	0.25	0.26	0.25
	AUC	0.61	0.63	0.62	0.60	0.61	0.63	0.62	0.59	0.62	0.62	0.63	0.62	0.62	0.62	0.63	0.63	0.62	0.63	0.62
	R-squared	0.05	0.07	0.07	0.05	0.05	0.07	0.06	0.04	0.06	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.06	0.07	0.07
	intercept	-0.71	-0.82	-0.70	-0.61	-0.39	-0.62	-0.73	-0.48	-0.57	-0.71	-0.69	-0.65	-0.63	-0.58	-0.72	-0.64	-0.59	-0.63	-0.56
	slope	0.76	0.60	0.62	0.63	0.82	0.66	0.96	0.67	0.61	0.98	0.64	0.81	0.87	0.77	0.77	0.70	0.64	0.69	0.65
External validation non-EU	Brier	0.27	0.28	0.25	0.26	0.23	0.25	0.28	0.25	0.25	0.27	0.26	0.26	0.26	0.25	0.27	0.25	0.25	0.25	0.25
	AUC	0.65	0.66	0.67	0.66	0.68	0.68	0.66	0.65	0.67	0.68	0.69	0.68	0.69	0.68	0.68	0.68	0.68	0.68	0.69
	R-squared	0.08	0.09	0.10	0.09	0.11	0.11	0.10	0.08	0.11	0.12	0.12	0.12	0.13	0.11	0.11	0.12	0.11	0.11	0.12
	intercept	-1.22	-1.38	-1.12	-1.11	-0.89	-1.08	-1.23	-1.03	-1.06	-1.20	-1.23	-1.17	-1.16	-1.11	-1.24	-1.11	-1.08	-1.13	-1.10
	slope	1.03	0.72	0.80	1.02	1.29	0.85	1.30	1.11	0.84	1.41	0.90	1.14	1.26	1.02	1.03	1.00	0.91	0.93	0.94

Table B.6 Results of elastic net

Sample info	N	50	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	3.3	6.6	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	-	0.19	0.20	0.21	0.22	0.20	0.22	0.23	0.21	0.22	0.20	0.22	0.22	0.22	0.21	0.21	0.22	0.22	0.22
	AUC	-	0.78	0.76	0.70	0.69	0.72	0.66	0.66	0.73	0.67	0.72	0.68	0.67	0.68	0.68	0.69	0.69	0.69	0.70
	R-squared	-	0.32	0.27	0.18	0.16	0.21	0.10	0.10	0.21	0.12	0.19	0.13	0.12	0.13	0.13	0.14	0.14	0.14	0.15
	intercept	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	slope	-	1.71	1.43	1.43	1.37	1.20	1.31	1.28	1.18	1.38	1.09	1.40	1.25	1.20	1.15	1.09	1.05	1.05	1.02
External validation non-UK	Brier	0.27	0.27	0.26	0.26	0.24	0.25	0.27	0.25	0.25	0.26	0.26	0.26	0.26	0.25	0.26	0.26	0.25	0.26	0.25
	AUC	0.60	0.62	0.63	0.60	0.62	0.63	0.61	0.60	0.62	0.62	0.63	0.62	0.63	0.62	0.62	0.63	0.62	0.63	0.62
	R-squared	0.03	0.07	0.07	0.04	0.06	0.07	0.05	0.04	0.06	0.06	0.07	0.06	0.07	0.07	0.07	0.07	0.06	0.07	0.07
	intercept	-0.70	-0.85	-0.70	-0.63	-0.38	-0.61	-0.72	-0.49	-0.54	-0.71	-0.70	-0.64	-0.64	-0.57	-0.71	-0.64	-0.59	-0.63	-0.56
	slope	0.87	0.53	0.56	0.56	0.79	0.59	0.85	0.73	0.62	0.93	0.57	0.90	0.91	0.81	0.77	0.69	0.63	0.67	0.63
External validation non-EU	Brier	0.28	0.29	0.25	0.26	0.23	0.25	0.28	0.25	0.24	0.27	0.26	0.26	0.26	0.25	0.27	0.25	0.25	0.25	0.25
	AUC	0.66	0.65	0.67	0.65	0.68	0.68	0.66	0.66	0.68	0.68	0.69	0.68	0.69	0.68	0.68	0.68	0.68	0.68	0.69
	R-squared	0.08	0.08	0.10	0.08	0.10	0.11	0.10	0.09	0.11	0.11	0.12	0.12	0.12	0.12	0.12	0.12	0.11	0.11	0.12
	intercept	-1.20	-1.45	-1.14	-1.12	-0.87	-1.08	-1.24	-1.04	-1.03	-1.21	-1.25	-1.16	-1.16	-1.10	-1.24	-1.12	-1.08	-1.13	-1.11
	slope	1.42	0.59	0.73	0.93	1.24	0.79	1.20	1.15	0.86	1.35	0.79	1.29	1.25	1.10	1.05	0.98	0.90	0.92	0.90

Table B.7 Results of neural network

Sample info	N	50	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	3.3	6.6	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	-	0.06	0.09	0.13	0.16	0.11	0.14	0.16	0.14	0.16	0.16	0.17	0.18	0.19	0.19	0.19	0.20	0.20	0.21
	AUC	-	0.98	0.94	0.89	0.84	0.92	0.87	0.85	0.87	0.83	0.83	0.81	0.79	0.78	0.76	0.77	0.73	0.73	0.72
	R-squared	-	0.87	0.74	0.58	0.46	0.65	0.55	0.47	0.54	0.44	0.41	0.38	0.32	0.32	0.29	0.30	0.22	0.21	0.21
	intercept	-	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	slope	-	2.92	1.63	1.40	1.30	1.44	1.56	1.44	1.31	1.32	1.15	1.15	1.13	1.08	1.07	1.08	1.04	1.03	1.02
External validation non-UK	Brier	0.38	0.35	0.33	0.33	0.30	0.33	0.32	0.30	0.30	0.31	0.29	0.30	0.29	0.29	0.29	0.28	0.27	0.27	0.26
	AUC	0.55	0.60	0.58	0.57	0.56	0.58	0.54	0.56	0.58	0.57	0.59	0.58	0.58	0.58	0.59	0.59	0.61	0.61	0.61
	R-squared	0.01	0.04	0.02	0.02	0.01	0.03	0.01	0.02	0.02	0.02	0.03	0.02	0.02	0.03	0.03	0.03	0.03	0.04	0.05
	intercept	-1.68	-1.60	-0.99	-0.95	-0.51	-0.88	-0.88	-0.58	-0.78	-0.81	-0.80	-0.82	-0.69	-0.69	-0.81	-0.71	-0.62	-0.68	-0.59
	slope	0.06	0.14	0.12	0.10	0.12	0.12	0.10	0.13	0.14	0.14	0.21	0.17	0.22	0.22	0.24	0.21	0.32	0.41	0.45
External validation non-EU	Brier	0.41	0.37	0.32	0.33	0.29	0.31	0.33	0.30	0.30	0.31	0.30	0.30	0.30	0.29	0.31	0.27	0.26	0.27	0.26
	AUC	0.59	0.60	0.60	0.58	0.58	0.64	0.58	0.61	0.61	0.63	0.62	0.62	0.61	0.63	0.61	0.64	0.65	0.66	0.67
	R-squared	0.03	0.04	0.03	0.02	0.02	0.07	0.02	0.04	0.04	0.05	0.04	0.06	0.04	0.06	0.04	0.06	0.07	0.09	0.10
	intercept	-2.98	-2.32	-1.60	-1.66	-1.14	-1.65	-1.65	-1.37	-1.50	-1.53	-1.46	-1.50	-1.36	-1.37	-1.52	-1.29	-1.15	-1.24	-1.17
	slope	0.11	0.14	0.13	0.14	0.16	0.20	0.16	0.20	0.19	0.24	0.24	0.30	0.28	0.31	0.26	0.33	0.49	0.57	0.64

Table B.8 Results of support vector machine

Sample info	N	50	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	3.3	6.6	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53.0	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	-	0.22	0.19	0.18	0.22	0.19	0.22	0.23	0.21	0.21	0.18	0.21	0.22	0.20	0.20	0.20	0.20	0.20	0.21
	AUC	-	0.95	0.93	0.95	0.91	0.90	0.87	0.85	0.89	0.87	0.87	0.84	0.81	0.83	0.81	0.79	0.79	0.77	0.76
	R-squared	-	0.72	0.74	0.72	0.59	0.58	0.47	0.45	0.54	0.48	0.48	0.41	0.34	0.40	0.37	0.29	0.28	0.26	0.24
	intercept	-	0.01	-0.02	-0.02	0.01	0.00	-0.03	0.00	-0.01	0.00	-0.02	-0.01	0.00	-0.02	-0.01	-0.02	-0.02	-0.01	-0.01
	slope	-	39.51	10.43	7.82	12.78	6.04	9.63	22.15	6.79	11.21	3.28	5.26	7.26	4.35	4.33	2.95	2.24	1.90	2.06
External validation non-UK	Brier	0.27	0.28	0.27	0.28	0.26	0.27	0.28	0.26	0.26	0.28	0.27	0.27	0.27	0.27	0.28	0.27	0.26	0.27	0.26
	AUC	0.57	0.59	0.56	0.53	0.56	0.57	0.55	0.55	0.57	0.56	0.57	0.57	0.57	0.57	0.56	0.59	0.59	0.59	0.59
	R-squared	0.02	0.03	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03
	intercept	-0.68	-0.76	-0.68	-0.72	-0.49	-0.65	-0.75	-0.57	-0.56	-0.76	-0.67	-0.70	-0.67	-0.65	-0.76	-0.68	-0.64	-0.66	-0.60
	slope	5.79	4.94	0.77	0.34	1.29	0.64	0.94	2.35	0.83	1.32	0.51	0.84	1.27	0.68	0.68	0.76	0.58	0.54	0.65
External validation non-EU	Brier	0.30	0.31	0.29	0.30	0.27	0.28	0.31	0.28	0.27	0.30	0.28	0.29	0.29	0.29	0.30	0.28	0.27	0.27	0.27
	AUC	0.60	0.60	0.58	0.56	0.60	0.61	0.55	0.60	0.61	0.63	0.62	0.61	0.62	0.61	0.62	0.65	0.64	0.65	0.65
	R-squared	0.04	0.03	0.03	0.02	0.03	0.05	0.01	0.03	0.04	0.06	0.06	0.04	0.05	0.04	0.05	0.07	0.06	0.07	0.08
	intercept	-1.24	-1.32	-1.19	-1.28	-1.05	-1.19	-1.30	-1.13	-1.12	-1.31	-1.24	-1.24	-1.22	-1.22	-1.30	-1.21	-1.16	-1.18	-1.15
	slope	8.34	5.18	1.00	0.61	2.05	1.15	0.98	4.67	1.34	2.89	0.89	1.41	2.18	1.11	1.20	1.27	0.91	0.93	1.15

Table B.9 Results of random forest

Sample info	N	50	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	3.3	6.6	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53.0	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	-	0.07	0.07	0.08	0.08	0.09	0.09	0.09	0.09	0.09	0.11	0.11	0.12	0.12	0.12	0.13	0.15	0.15	0.16
	AUC	-	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.98	0.98	0.98	0.97	0.97	0.95	0.93	0.89	0.89	0.87
	R-squared	-	0.95	0.98	0.95	0.95	0.94	0.95	0.90	0.85	0.86	0.83	0.85	0.83	0.79	0.75	0.67	0.55	0.55	0.50
	intercept	-	0.08	0.05	-0.10	0.04	-0.19	-0.11	-0.11	-0.02	-0.19	-0.26	-0.28	-0.23	-0.28	-0.36	-0.33	-0.34	-0.40	-0.37
	slope	-	6.33	16.19	6.95	8.66	7.44	9.38	5.54	4.26	4.16	4.32	4.72	5.18	3.82	3.07	2.38	1.53	1.53	1.38
External validation non-UK	Brier	0.28	0.27	0.26	0.28	0.25	0.28	0.29	0.28	0.27	0.28	0.28	0.29	0.28	0.29	0.30	0.29	0.30	0.30	0.29
	AUC	0.58	0.61	0.61	0.57	0.59	0.60	0.58	0.56	0.61	0.58	0.61	0.59	0.60	0.59	0.59	0.60	0.60	0.60	0.61
	R-squared	0.02	0.05	0.05	0.02	0.03	0.05	0.02	0.01	0.05	0.03	0.04	0.03	0.04	0.03	0.03	0.04	0.04	0.04	0.05
	intercept	-0.72	-0.75	-0.61	-0.67	-0.32	-0.79	-0.88	-0.61	-0.62	-0.84	-0.91	-0.94	-0.82	-0.90	-1.04	-0.99	-0.95	-1.05	-0.93
	slope	0.31	0.43	0.41	0.26	0.38	0.39	0.37	0.23	0.37	0.34	0.38	0.34	0.44	0.34	0.32	0.33	0.28	0.29	0.30
External validation non-EU	Brier	0.28	0.27	0.25	0.27	0.24	0.27	0.30	0.28	0.26	0.29	0.29	0.30	0.29	0.30	0.31	0.29	0.29	0.30	0.29
	AUC	0.63	0.66	0.65	0.64	0.65	0.64	0.61	0.62	0.66	0.65	0.67	0.65	0.65	0.64	0.66	0.66	0.66	0.66	0.68
	R-squared	0.06	0.09	0.08	0.07	0.08	0.08	0.05	0.05	0.09	0.07	0.10	0.08	0.08	0.07	0.09	0.09	0.09	0.10	0.11
	intercept	-1.32	-1.31	-1.06	-1.18	-0.87	-1.29	-1.40	-1.22	-1.14	-1.42	-1.50	-1.47	-1.39	-1.50	-1.62	-1.48	-1.49	-1.60	-1.56
	slope	0.50	0.60	0.56	0.58	0.66	0.56	0.55	0.48	0.57	0.55	0.61	0.59	0.64	0.51	0.57	0.51	0.47	0.47	0.48

Table B.10 Results of GBM

Sample info	N	50	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	3.3	6.6	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	-	0.17	0.17	0.19	0.20	0.19	0.20	0.21	0.20	0.20	0.20	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
	AUC	-	0.84	0.83	0.77	0.76	0.77	0.74	0.73	0.77	0.72	0.75	0.71	0.70	0.71	0.70	0.71	0.70	0.70	0.70
	R-squared	-	0.43	0.41	0.30	0.28	0.27	0.22	0.20	0.27	0.20	0.24	0.18	0.16	0.17	0.16	0.16	0.15	0.15	0.16
	intercept	-	-0.01	-0.07	-0.01	0.00	0.01	0.00	-0.05	-0.03	0.00	-0.03	-0.01	-0.01	0.02	0.00	0.00	0.00	0.02	0.01
	slope	-	1.82	1.50	1.50	1.38	1.39	1.49	1.41	1.21	1.35	1.29	1.23	1.27	1.12	1.21	1.12	1.12	1.13	1.07
External validation non-UK	Brier	0.29	0.27	0.27	0.27	0.25	0.26	0.28	0.26	0.26	0.27	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.25
	AUC	0.54	0.61	0.62	0.59	0.61	0.61	0.60	0.57	0.62	0.60	0.62	0.62	0.61	0.61	0.62	0.62	0.62	0.62	0.62
	R-squared	0.01	0.05	0.06	0.03	0.05	0.05	0.04	0.02	0.06	0.04	0.06	0.06	0.05	0.05	0.06	0.06	0.06	0.06	0.06
	intercept	-0.68	-0.72	-0.81	-0.61	-0.42	-0.63	-0.87	-0.47	-0.58	-0.71	-0.70	-0.68	-0.61	-0.57	-0.71	-0.68	-0.60	-0.63	-0.57
	slope	0.15	0.38	0.44	0.36	0.53	0.43	0.53	0.33	0.49	0.48	0.54	0.65	0.65	0.60	0.62	0.62	0.60	0.64	0.63
External validation non-EU	Brier	0.28	0.27	0.26	0.26	0.24	0.25	0.29	0.26	0.25	0.27	0.26	0.26	0.26	0.26	0.27	0.26	0.25	0.25	0.25
	AUC	0.58	0.65	0.67	0.64	0.66	0.66	0.64	0.63	0.67	0.65	0.68	0.67	0.68	0.66	0.67	0.67	0.67	0.68	0.68
	R-squared	0.02	0.08	0.09	0.07	0.09	0.10	0.07	0.06	0.11	0.09	0.12	0.11	0.11	0.10	0.11	0.11	0.11	0.11	0.12
	intercept	-1.13	-1.30	-1.23	-1.14	-0.97	-1.08	-1.39	-1.05	-1.07	-1.24	-1.22	-1.19	-1.18	-1.11	-1.26	-1.14	-1.09	-1.14	-1.12
	slope	0.28	0.52	0.59	0.61	0.86	0.62	0.74	0.63	0.72	0.79	0.79	0.94	1.03	0.85	0.88	0.86	0.88	0.92	0.95

Table B.11 Results of XGB

Sample info	N	50	100	150	200	250	300	350	400	450	500	600	700	800	900	1000	1500	2000	2500	3000
	EPV	3.3	6.6	9.9	13.2	16.6	19.9	23.2	26.5	29.8	33.1	39.7	46.4	53	59.6	66.2	99.3	132.4	165.6	198.7
Internal validation UK	Brier	-	0.04	0.05	0.05	0.07	0.09	0.09	0.10	0.11	0.12	0.12	0.13	0.15	0.15	0.16	0.17	0.18	0.18	0.19
	AUC	-	1.00	1.00	1.00	0.99	0.98	0.98	0.96	0.95	0.94	0.93	0.92	0.89	0.88	0.86	0.84	0.81	0.80	0.77
	R-squared	-	0.98	0.94	0.93	0.89	0.83	0.84	0.78	0.73	0.71	0.69	0.63	0.57	0.53	0.49	0.44	0.37	0.33	0.30
	intercept	-	0.01	-0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	slope	-	9.35	3.90	3.91	3.57	3.37	3.65	2.82	2.59	2.80	2.28	2.27	2.52	2.10	1.94	1.82	1.62	1.49	1.41
External validation non-UK	Brier	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
	AUC	0.52	0.59	0.61	0.55	0.58	0.60	0.57	0.55	0.59	0.60	0.60	0.61	0.61	0.59	0.59	0.61	0.61	0.61	0.62
	R-squared	0.00	0.04	0.06	0.01	0.03	0.04	0.02	0.02	0.04	0.04	0.05	0.05	0.05	0.04	0.04	0.05	0.05	0.05	0.06
	intercept	-0.31	-0.39	-0.42	-0.37	-0.34	-0.37	-0.39	-0.39	-0.32	-0.39	-0.38	-0.37	-0.37	-0.35	-0.37	-0.36	-0.35	-0.36	-0.40
	slope	0.29	1.47	1.18	1.04	0.95	1.71	1.39	0.66	1.39	2.06	1.87	2.14	2.44	1.87	1.81	1.91	2.03	2.12	1.34
External validation non-EU	Brier	0.25	0.25	0.25	0.25	0.24	0.25	0.26	0.25	0.24	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.24
	AUC	0.58	0.62	0.63	0.58	0.61	0.63	0.60	0.59	0.66	0.65	0.63	0.66	0.65	0.64	0.65	0.66	0.66	0.66	0.67
	R-squared	0.02	0.06	0.08	0.02	0.05	0.08	0.03	0.03	0.09	0.08	0.08	0.09	0.09	0.09	0.09	0.10	0.09	0.10	0.11
	intercept	-0.84	-0.93	-0.94	-0.88	-0.87	-0.90	-0.95	-0.91	-0.90	-0.93	-0.93	-0.91	-0.91	-0.90	-0.92	-0.90	-0.89	-0.90	-0.94
	slope	0.91	1.72	1.35	1.37	1.52	2.42	1.96	1.34	1.37	2.71	2.42	2.88	3.36	2.64	2.84	2.76	2.87	2.83	1.94

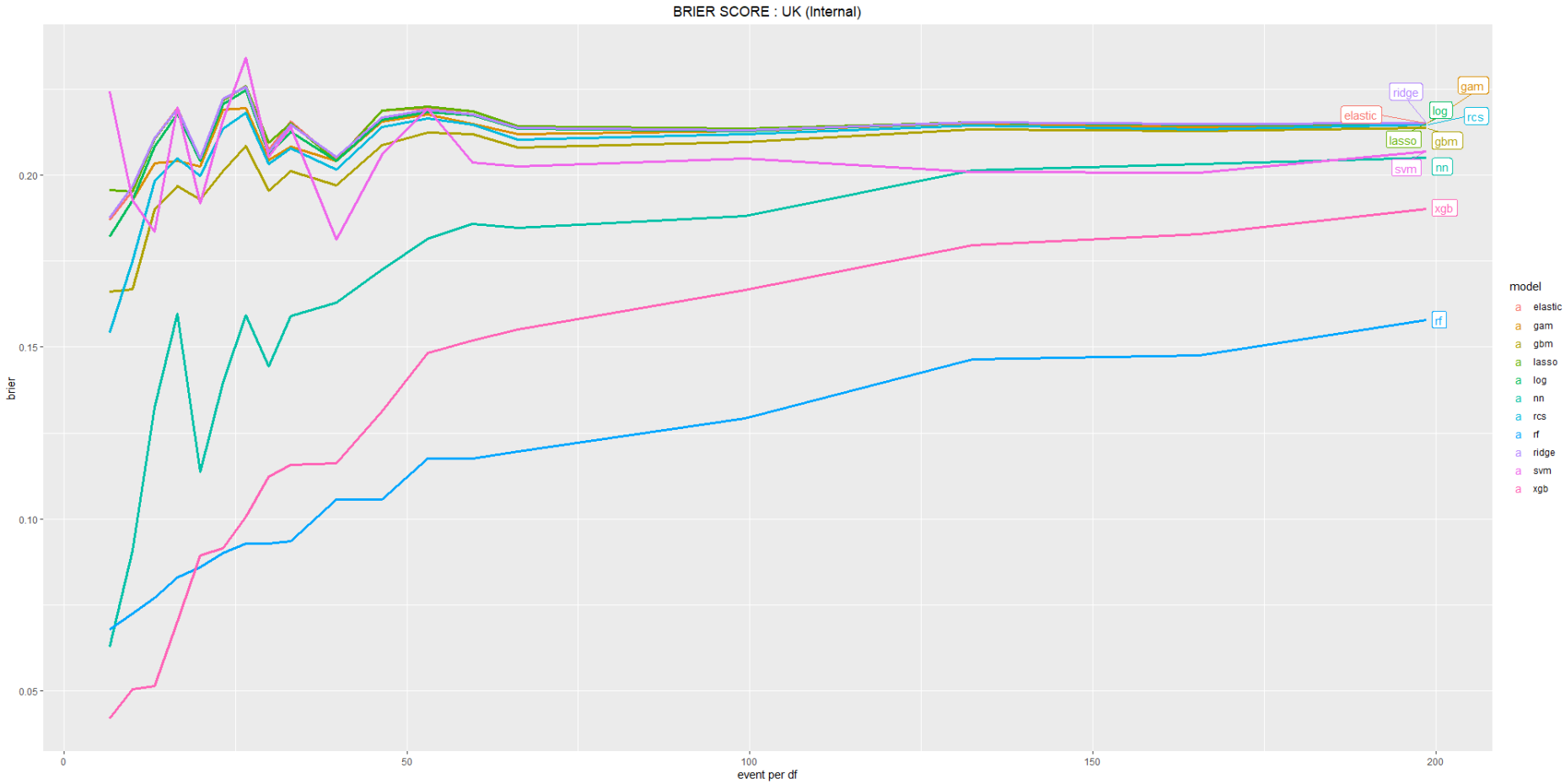


Figure B.1 Brier score results of all models, UK data

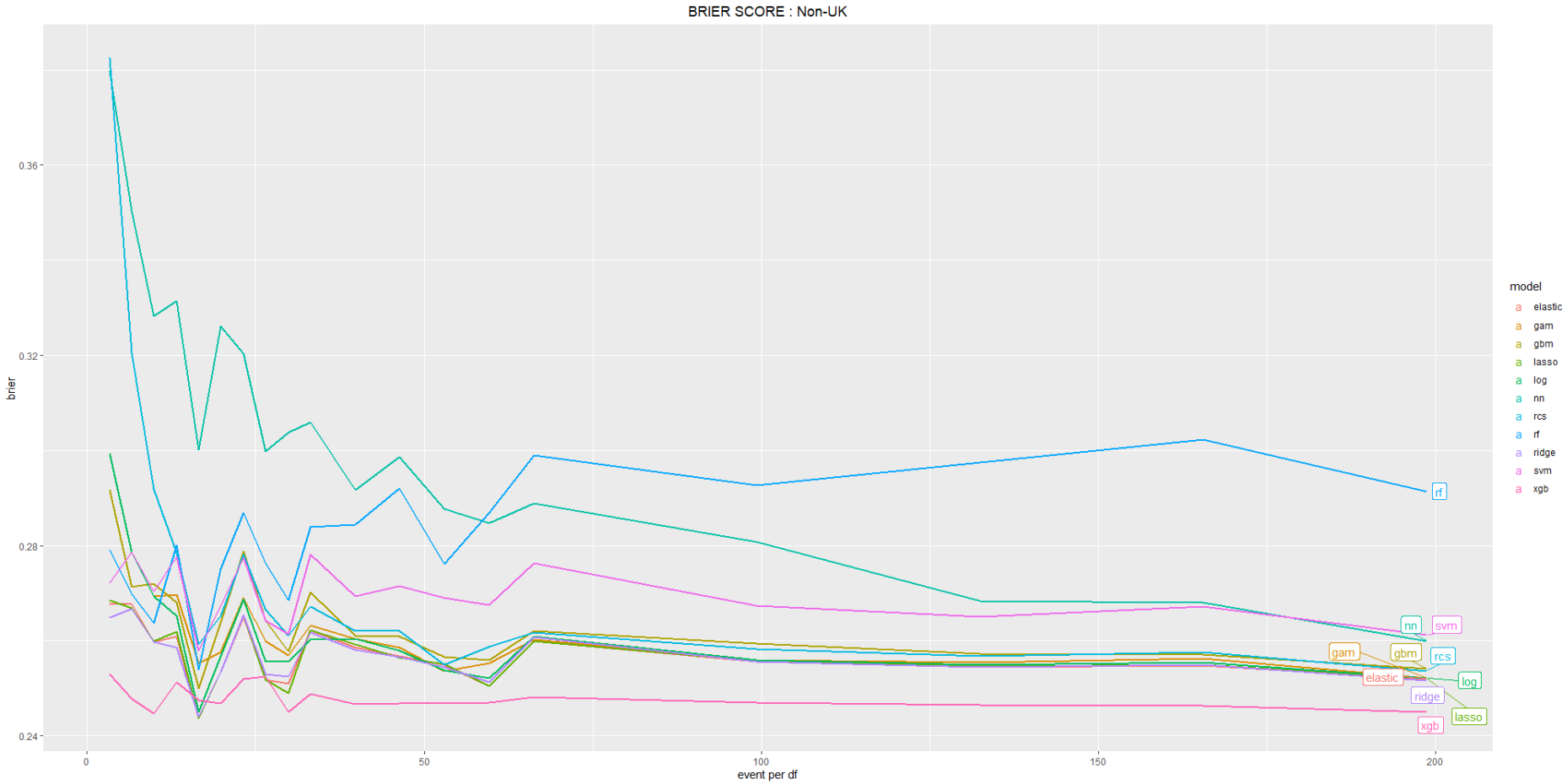


Figure B.2 Brier score results of all models, non-UK data

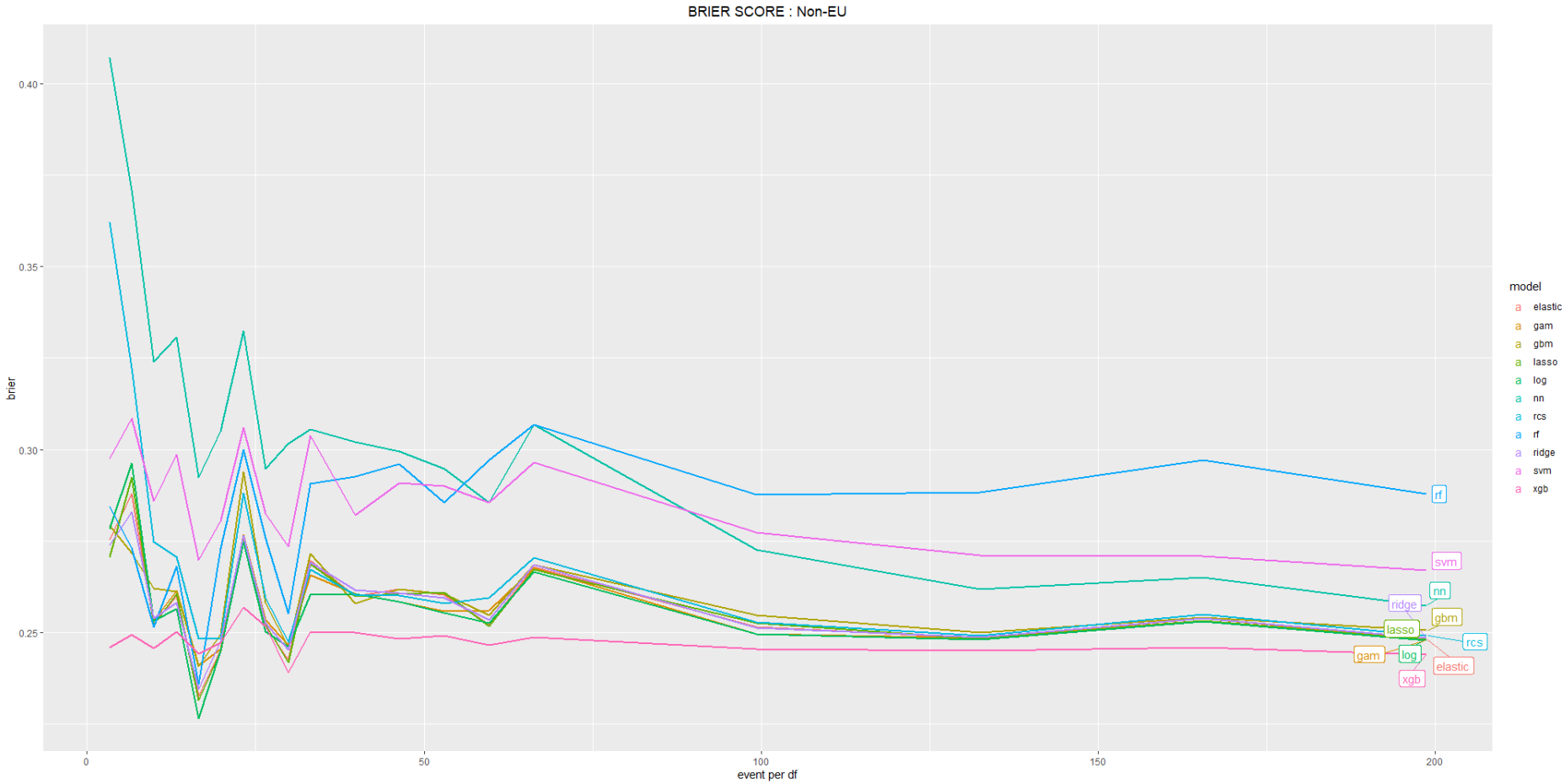


Figure B.3 Brier score results of all models, non-EU data

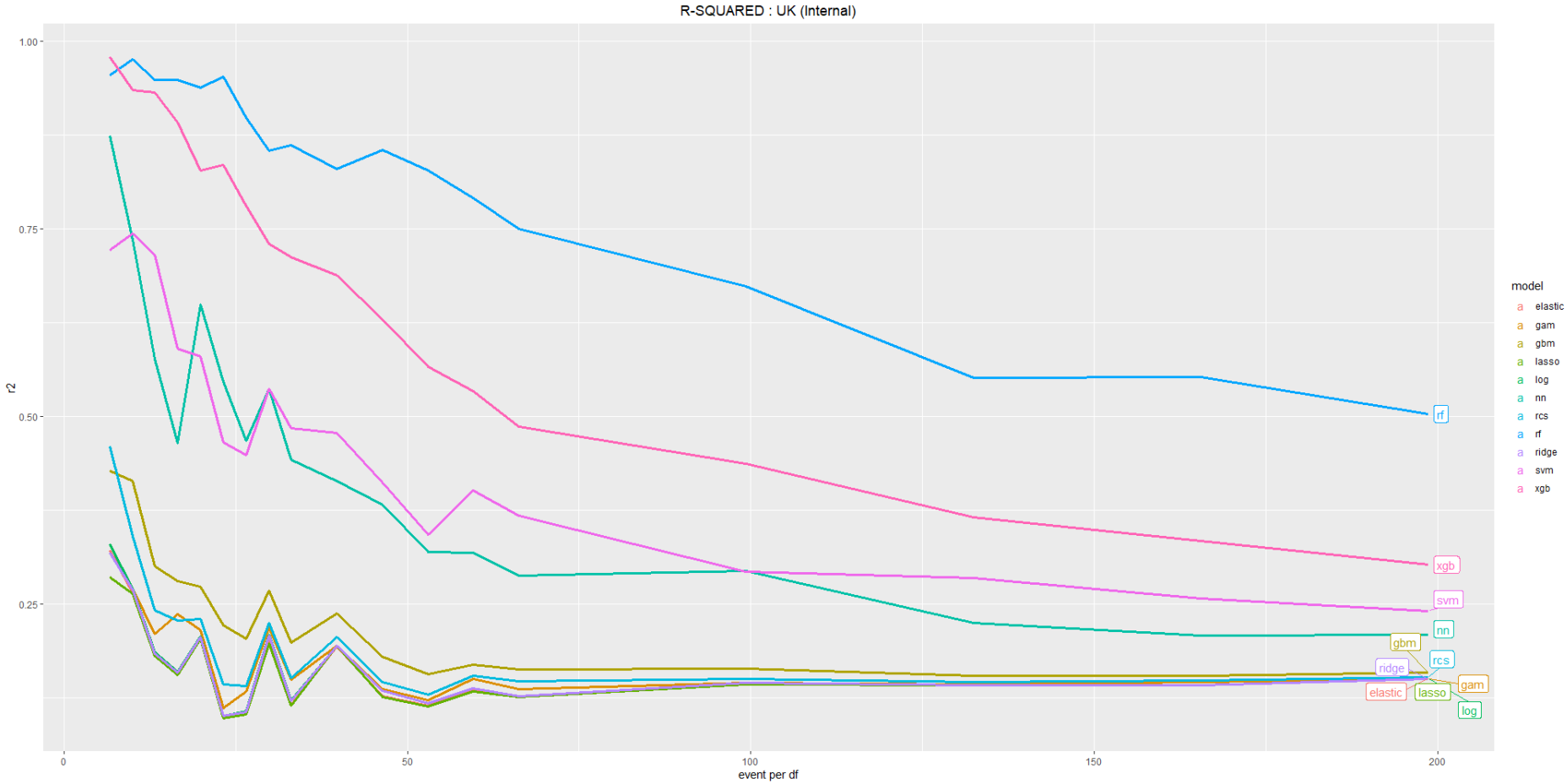


Figure B.4 R-squared results of all models, UK data

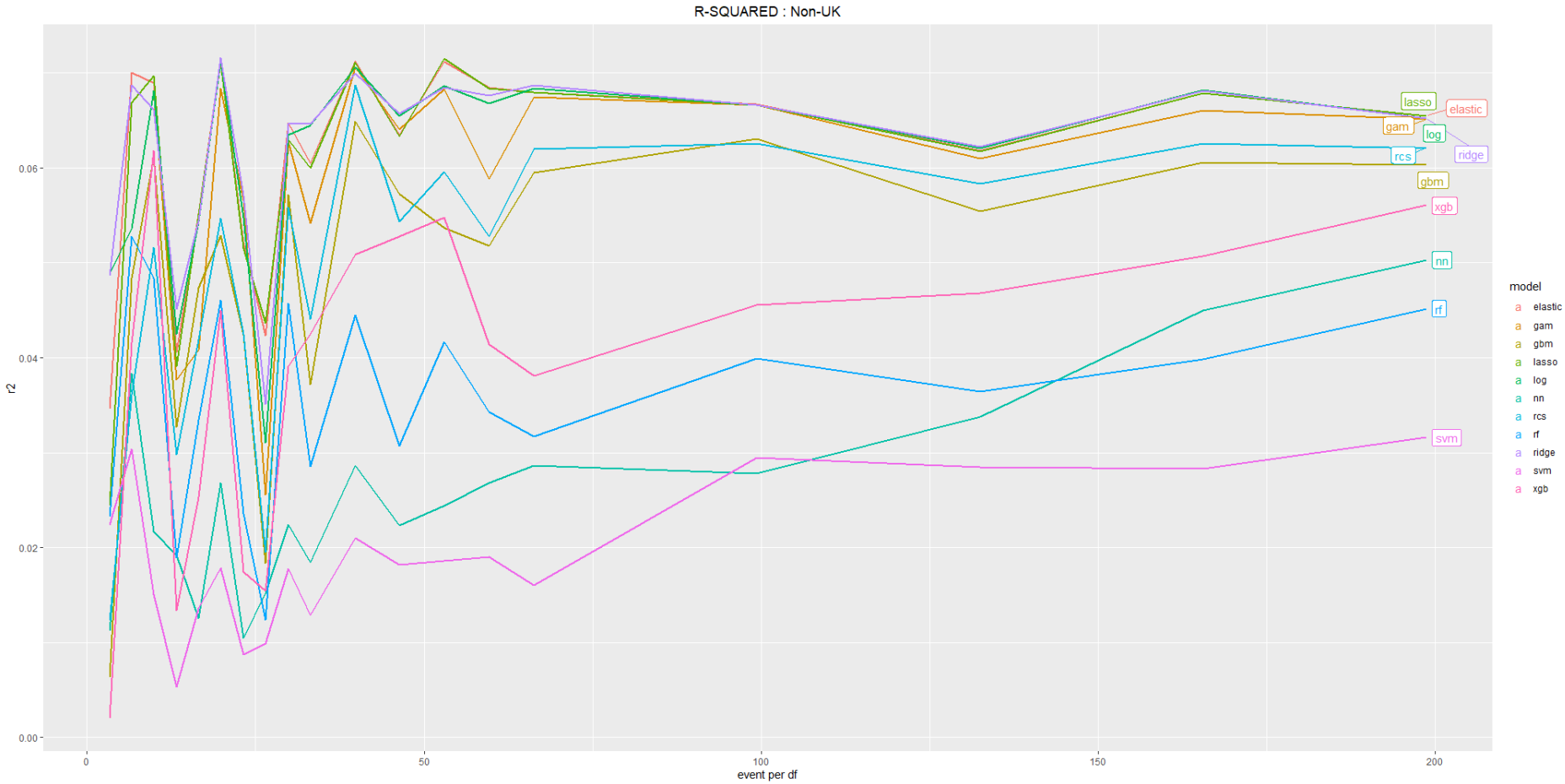


Figure B.5 R-squared results of all models, non-UK data

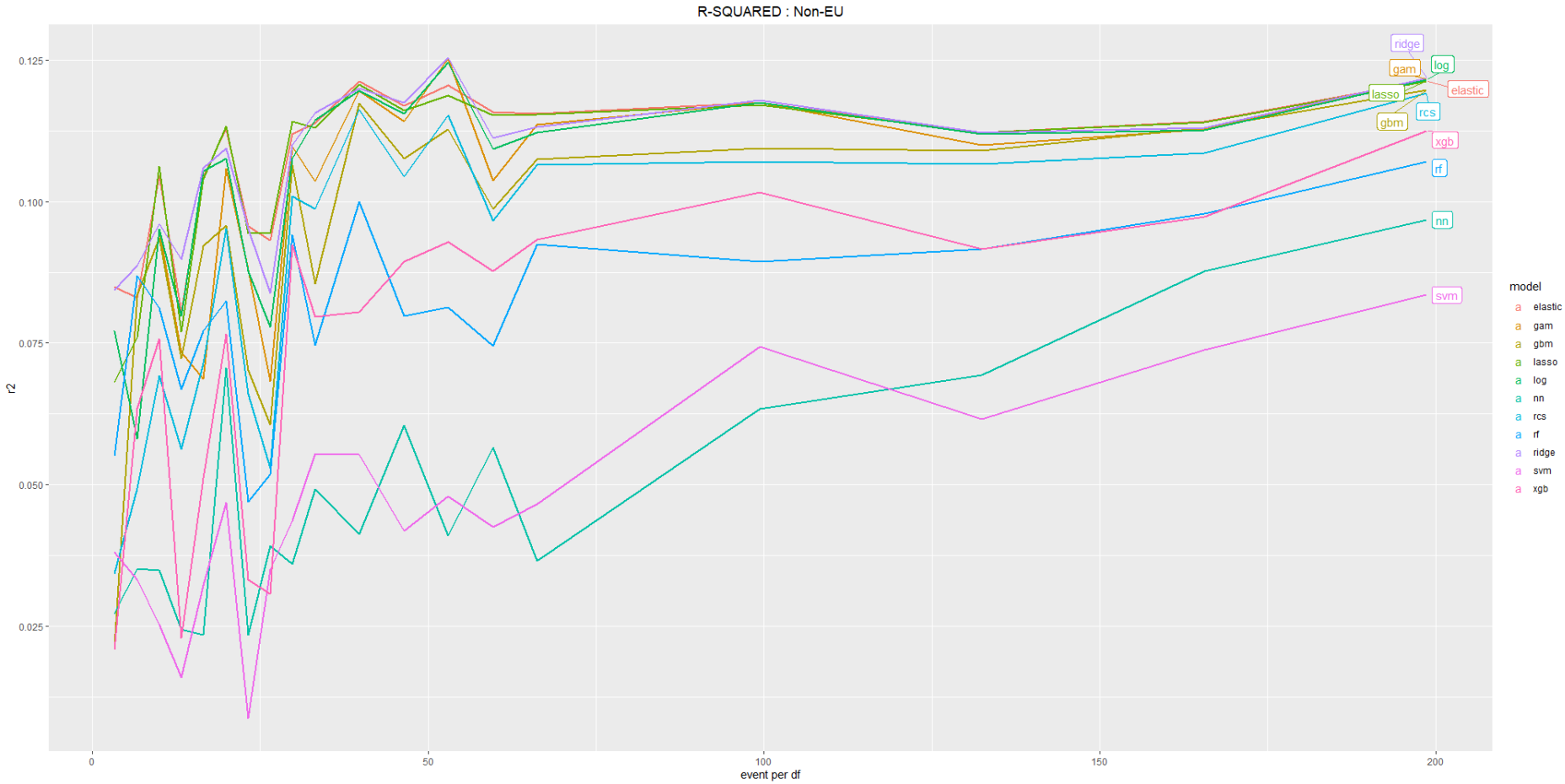


Figure B.6 R-squared results of all models, non-EU data

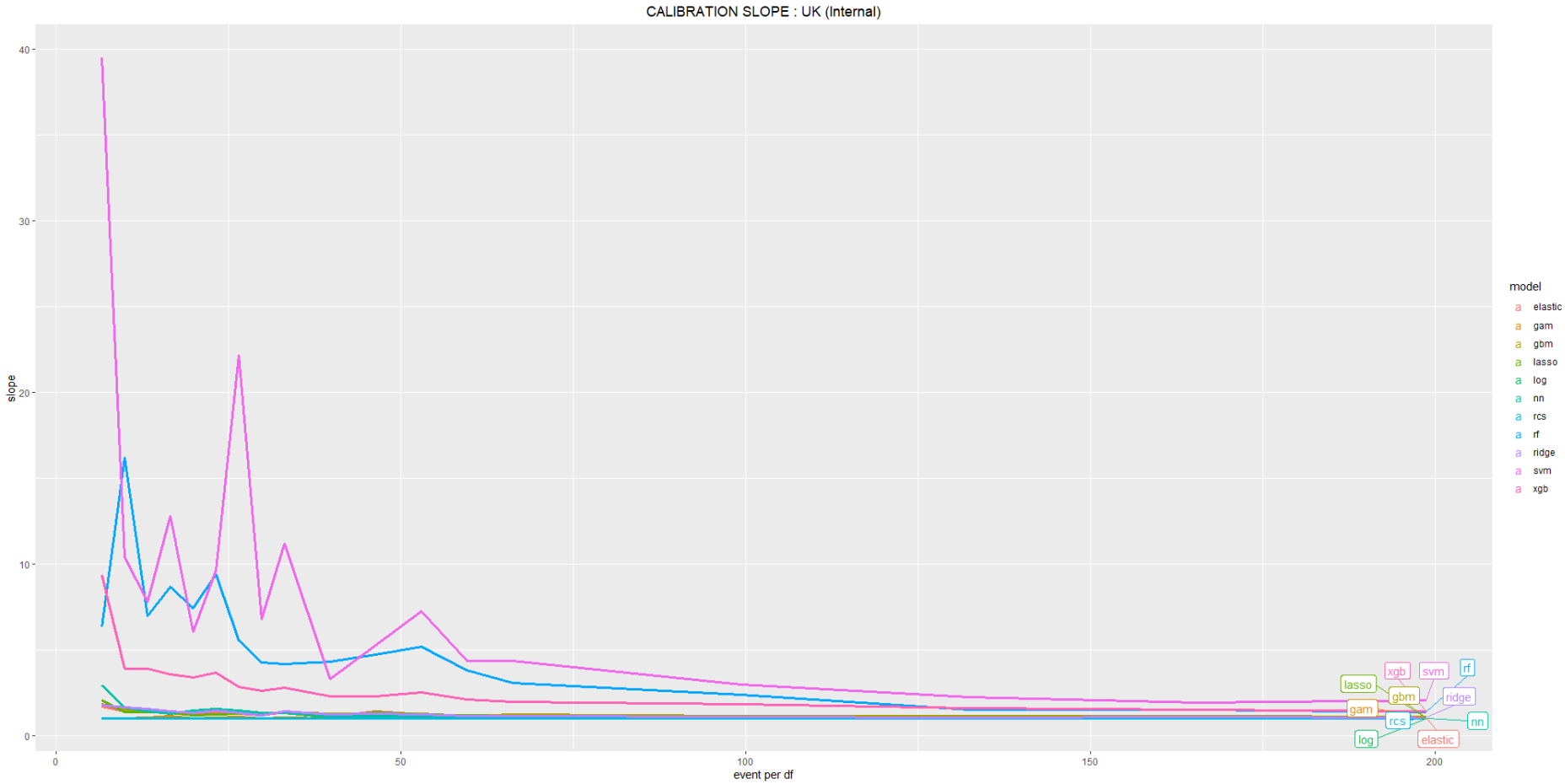


Figure B.7 Calibration slopes of all models, UK data

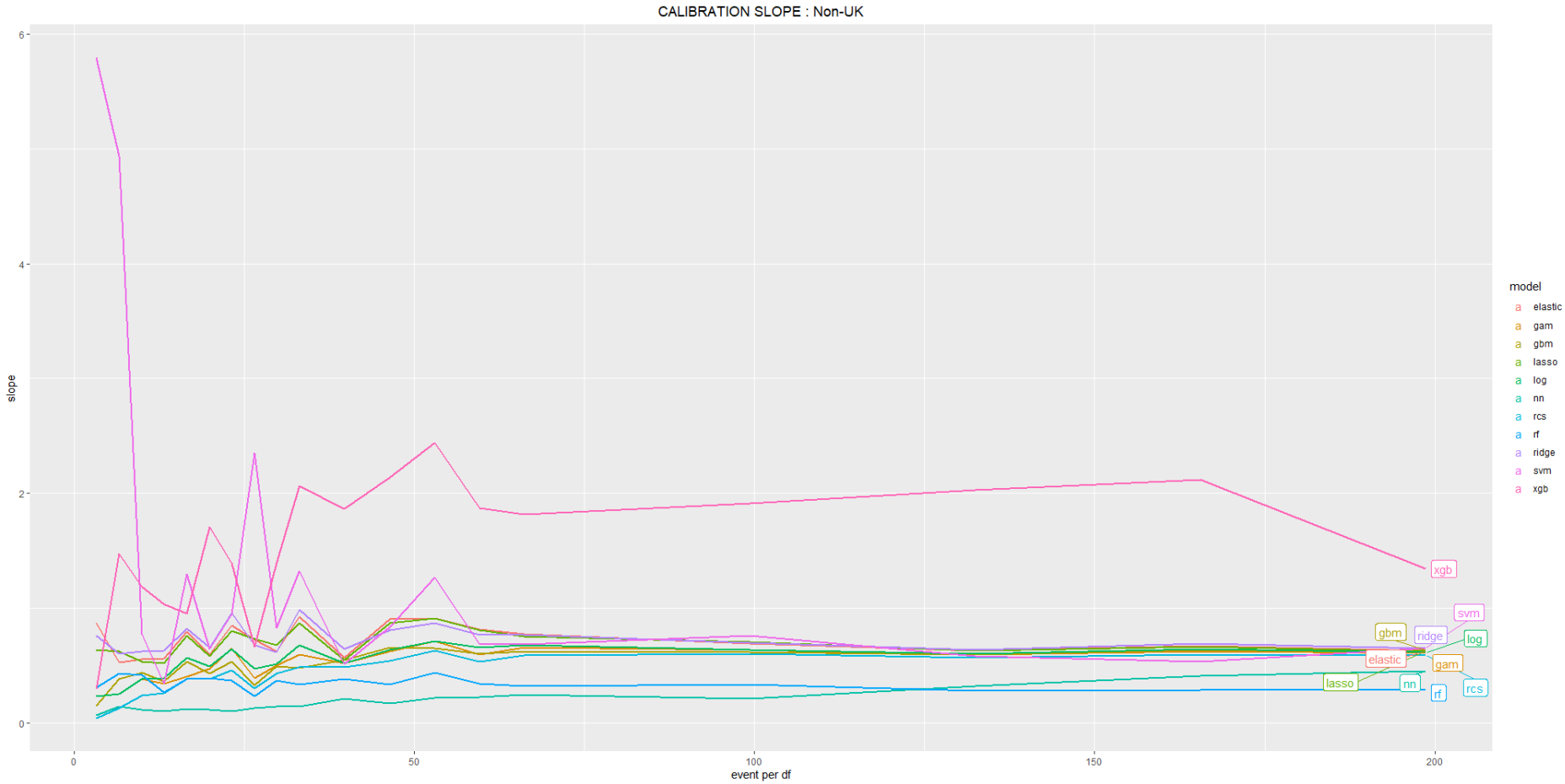


Figure B.8 Calibration slopes of all models, non-UK data

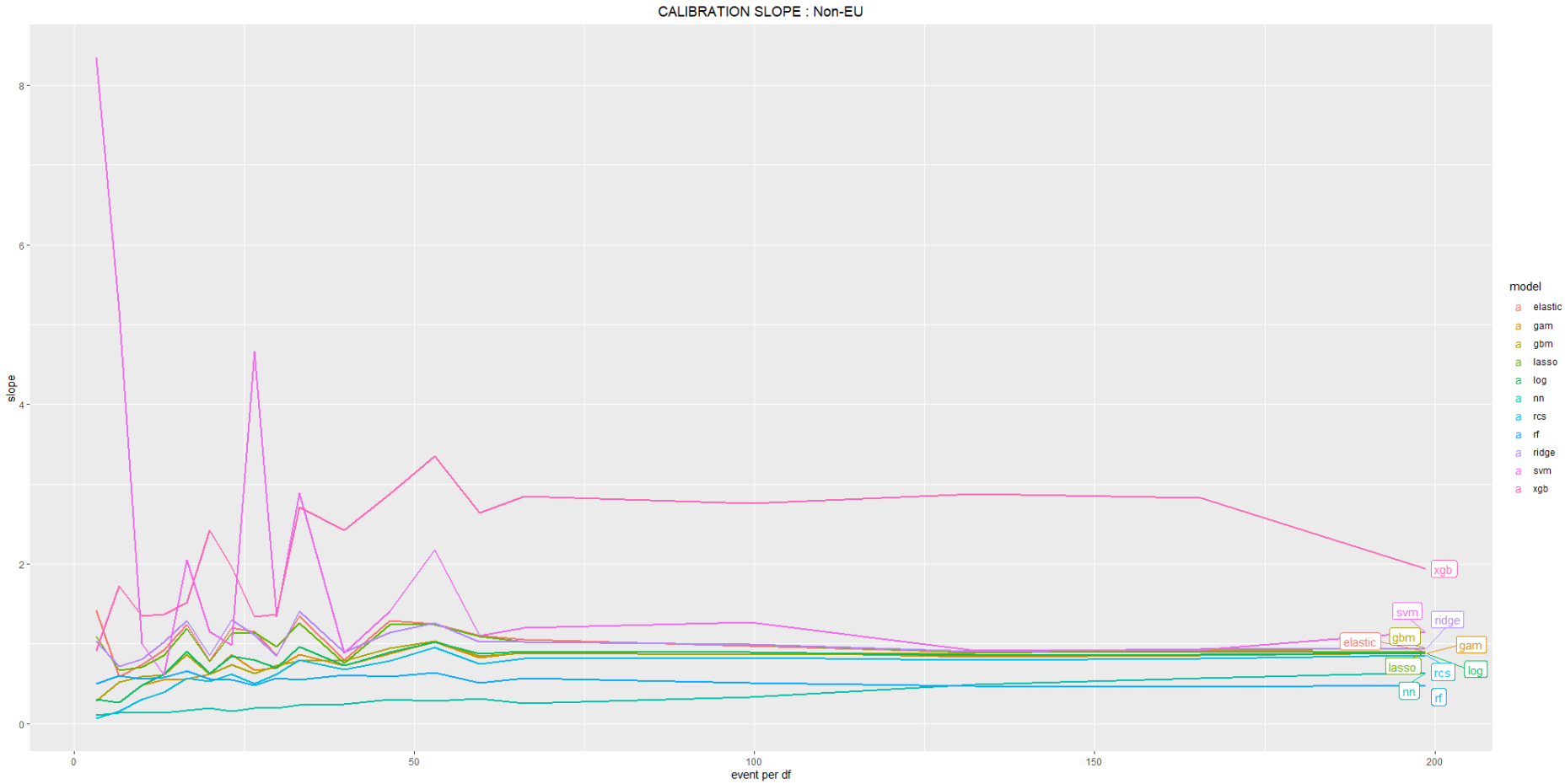


Figure B.9 Calibration slopes of all models, non-EU data

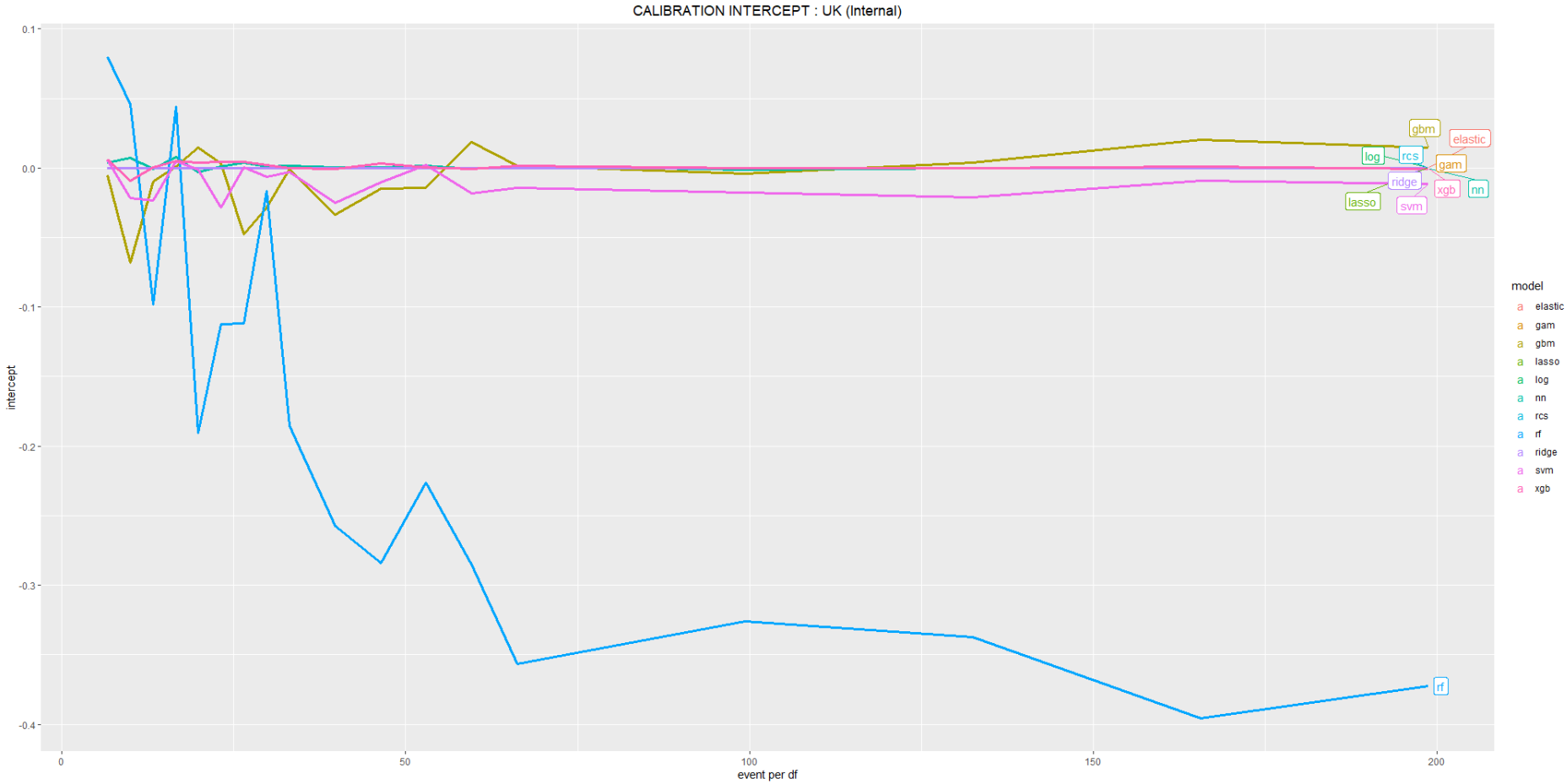


Figure B.10 Calibration intercepts of all models, UK data

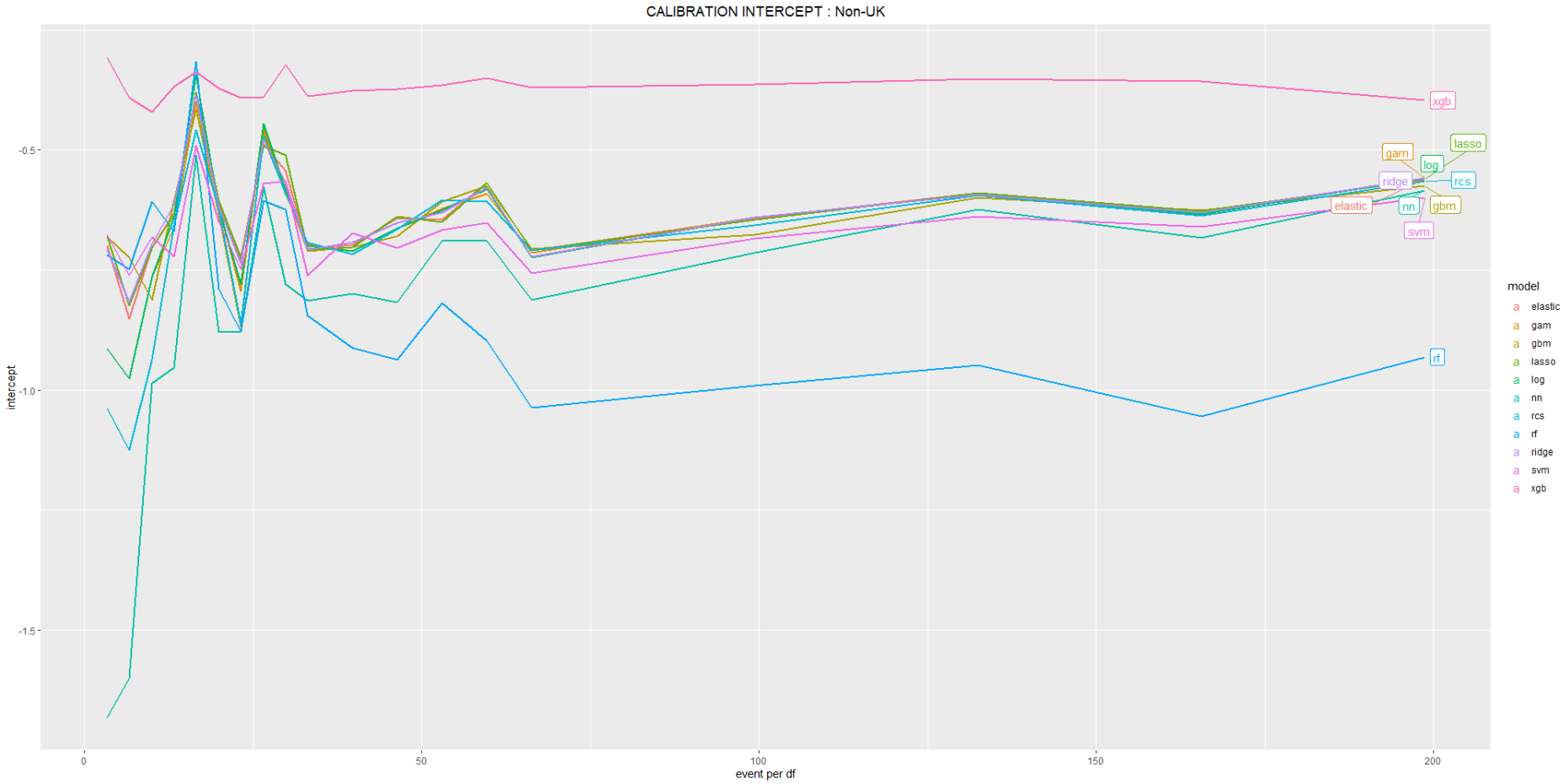


Figure B.11 Calibration intercepts of all models, non-UK data

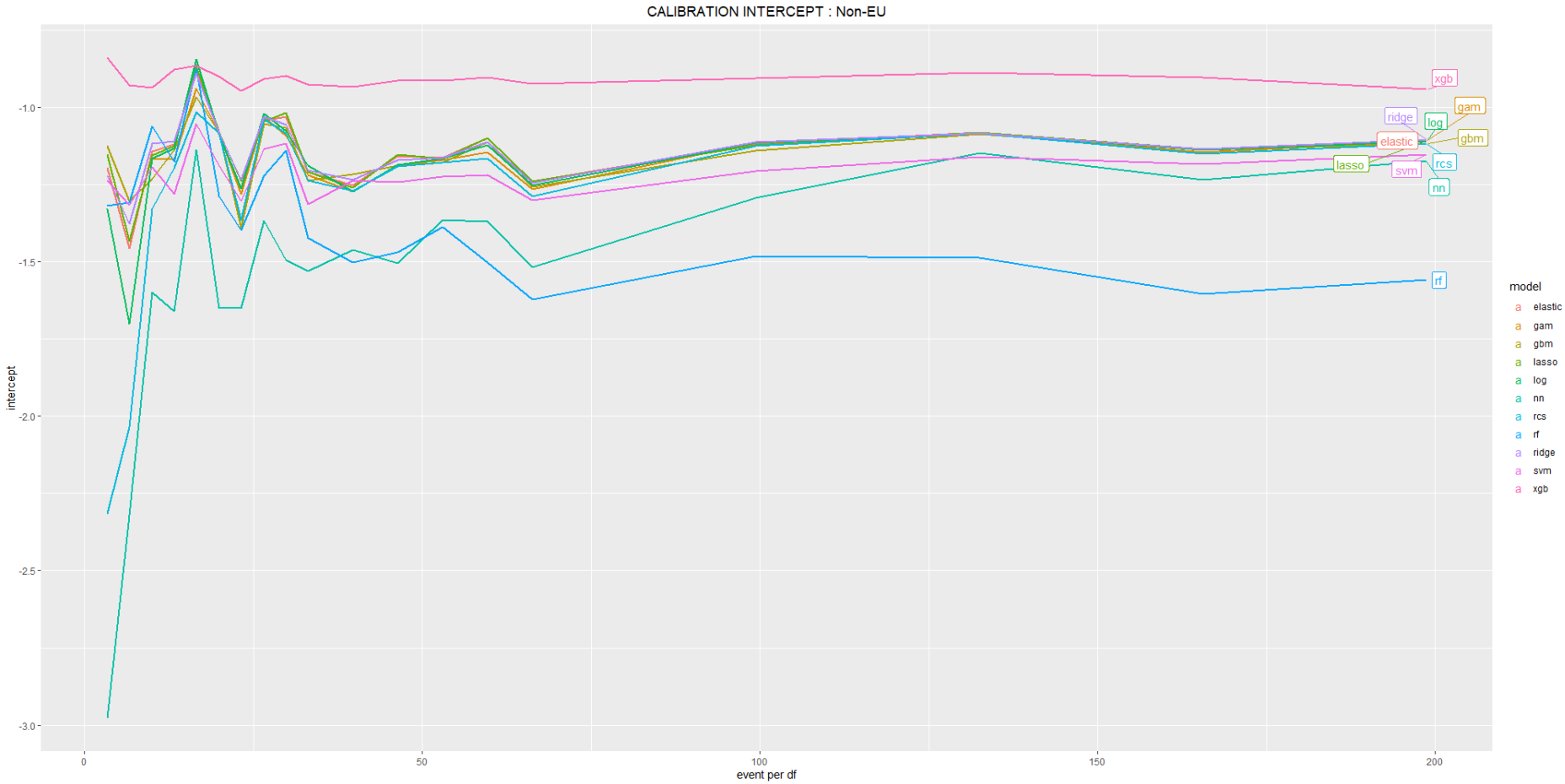


Figure B.12 Calibration intercepts of all models, non-EU data

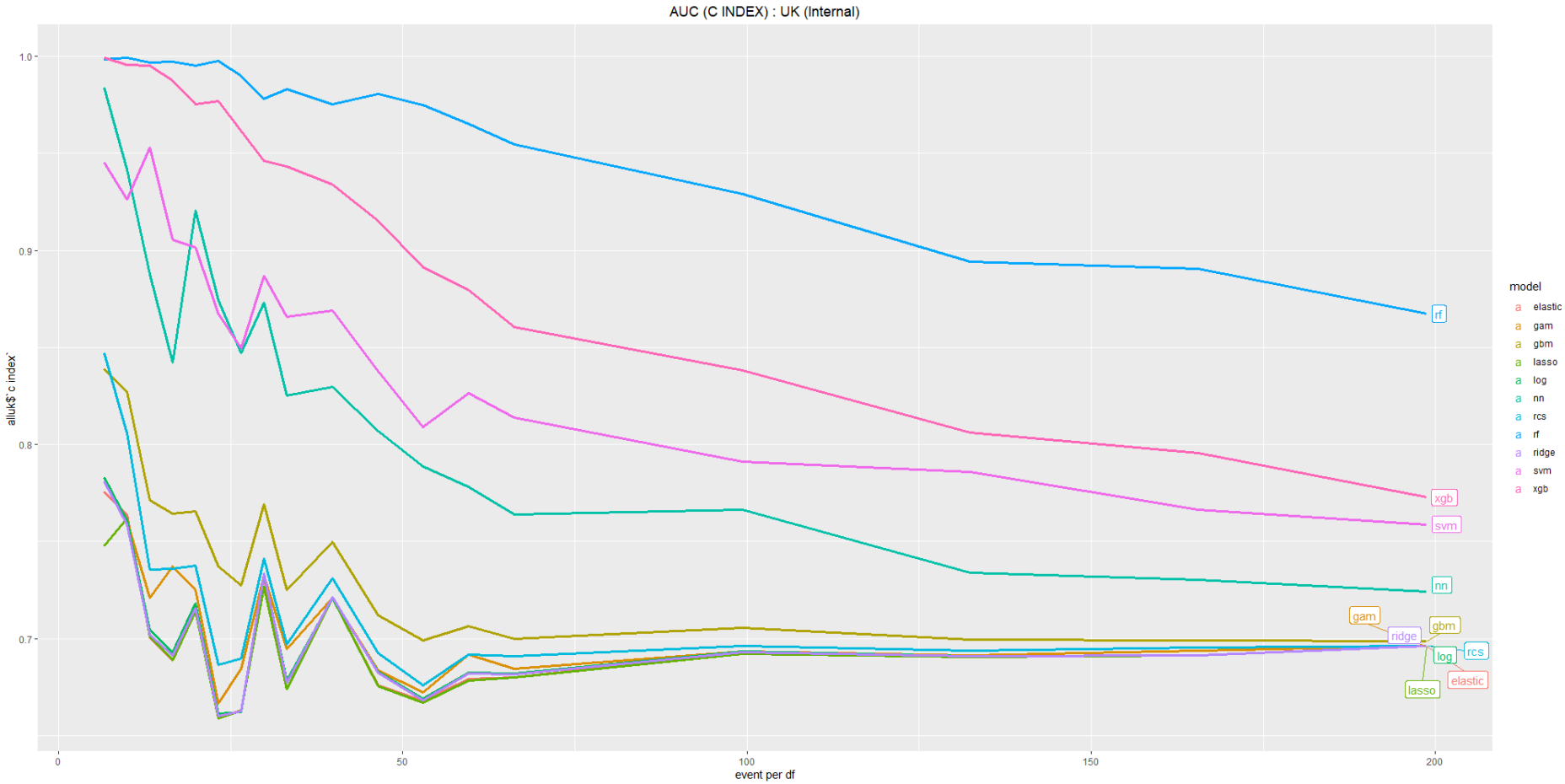


Figure B.13 AUC (C-index) of all models, UK data

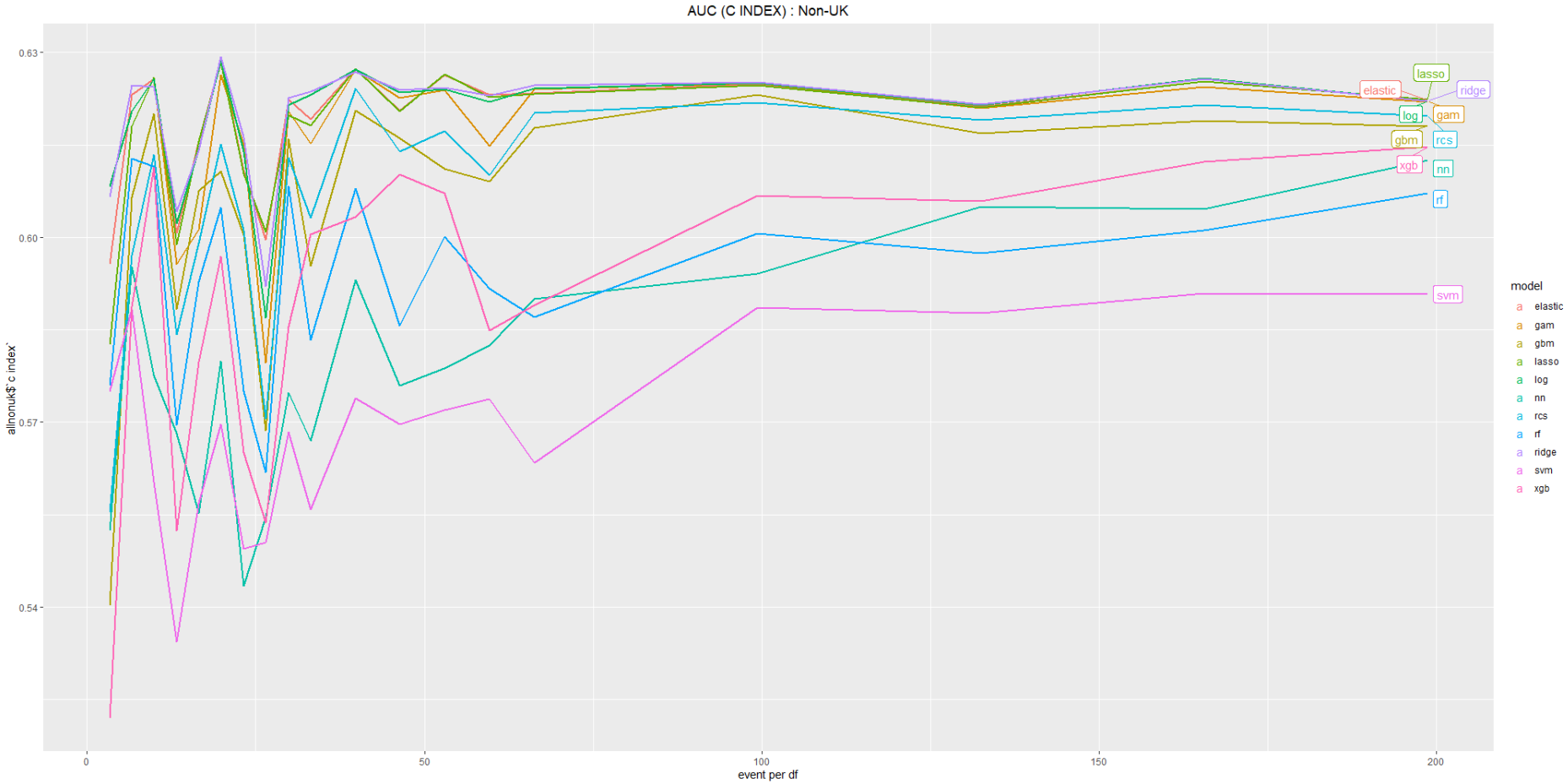


Figure B.14 AUC (C-index) of all models, non-UK data

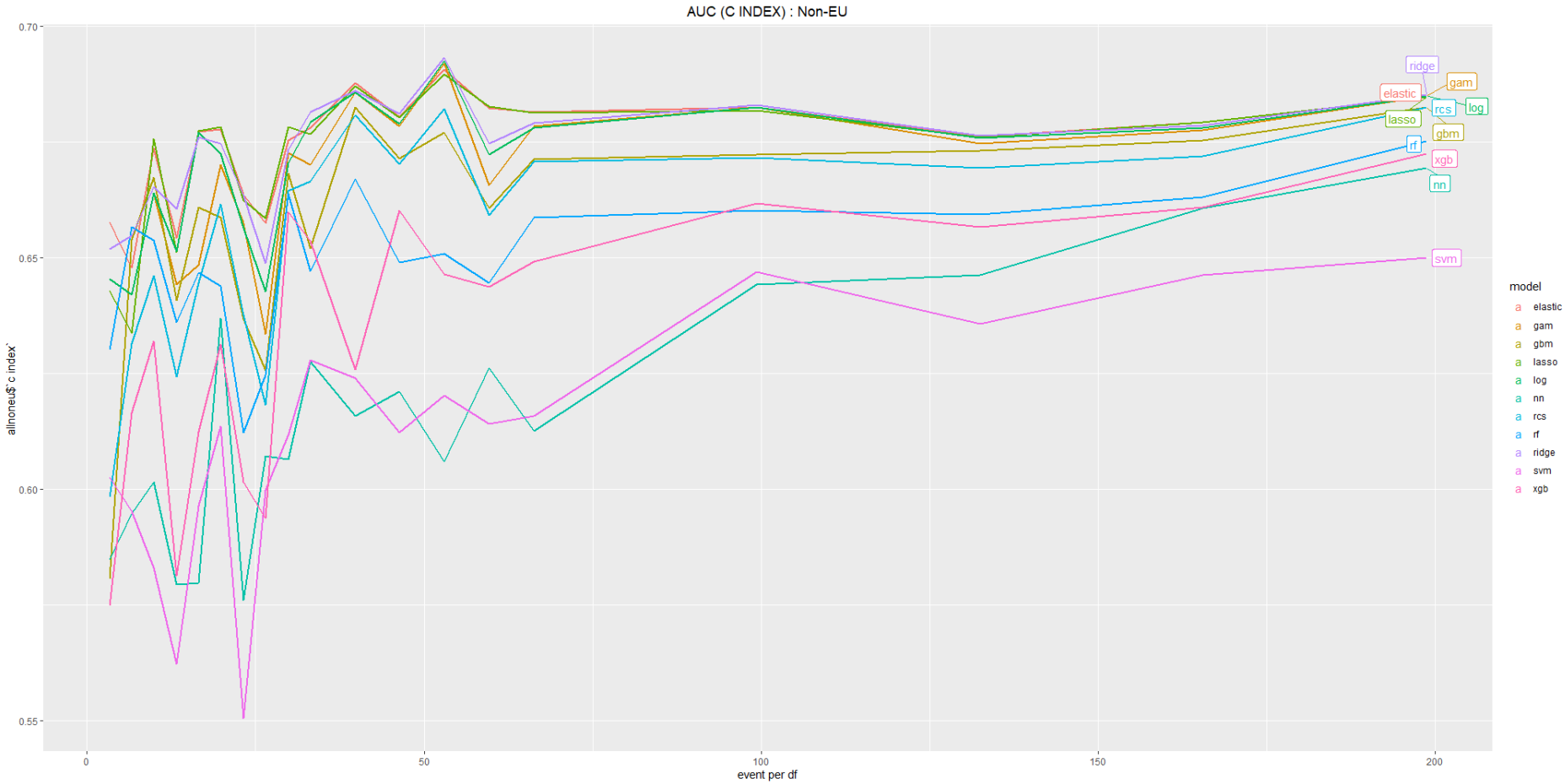


Figure B.15 AUC (C-index) of all models, non-EU data

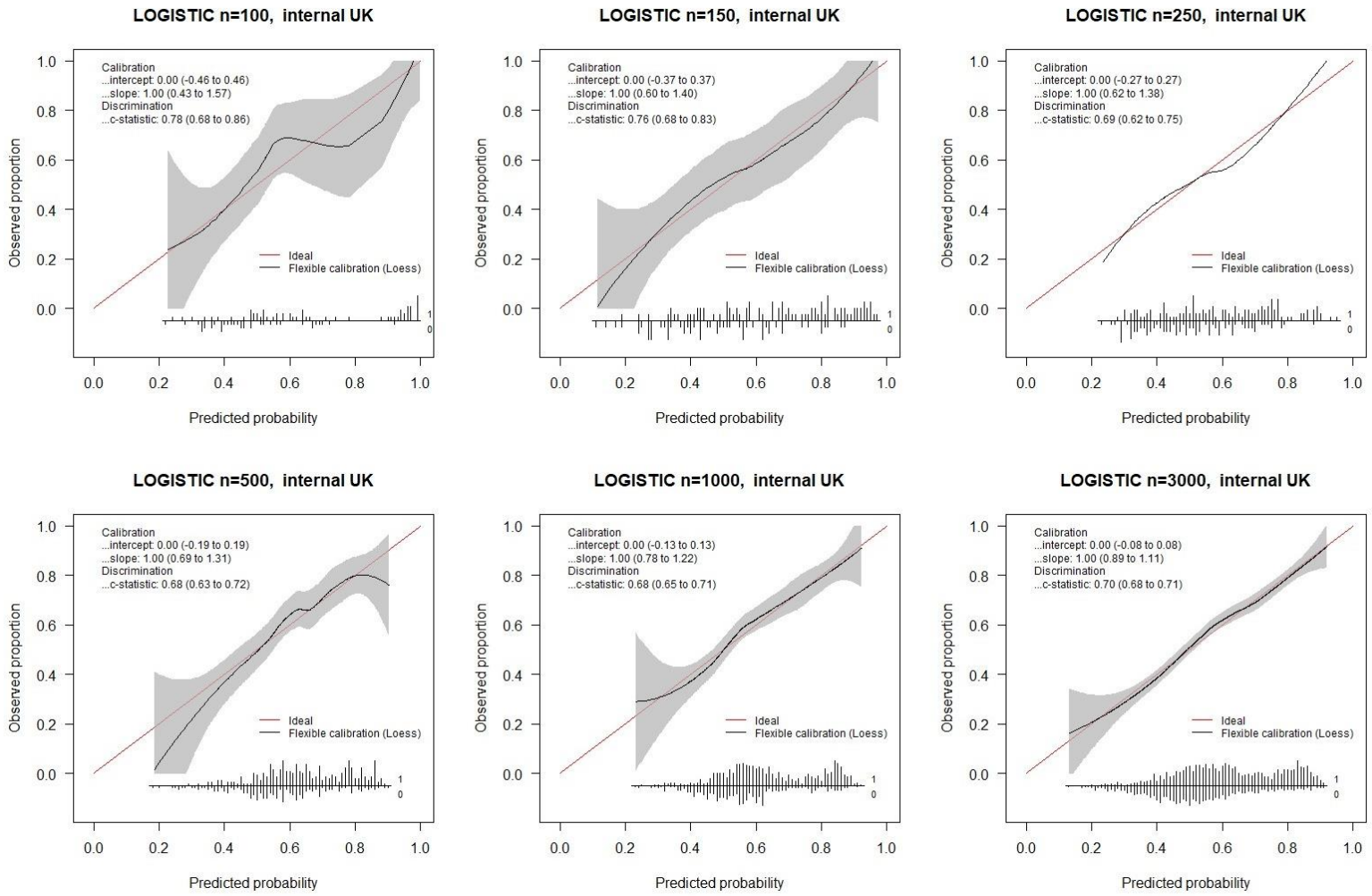


Figure B.16 Calibration plots of logistic regression models with 6 different sizes, UK data

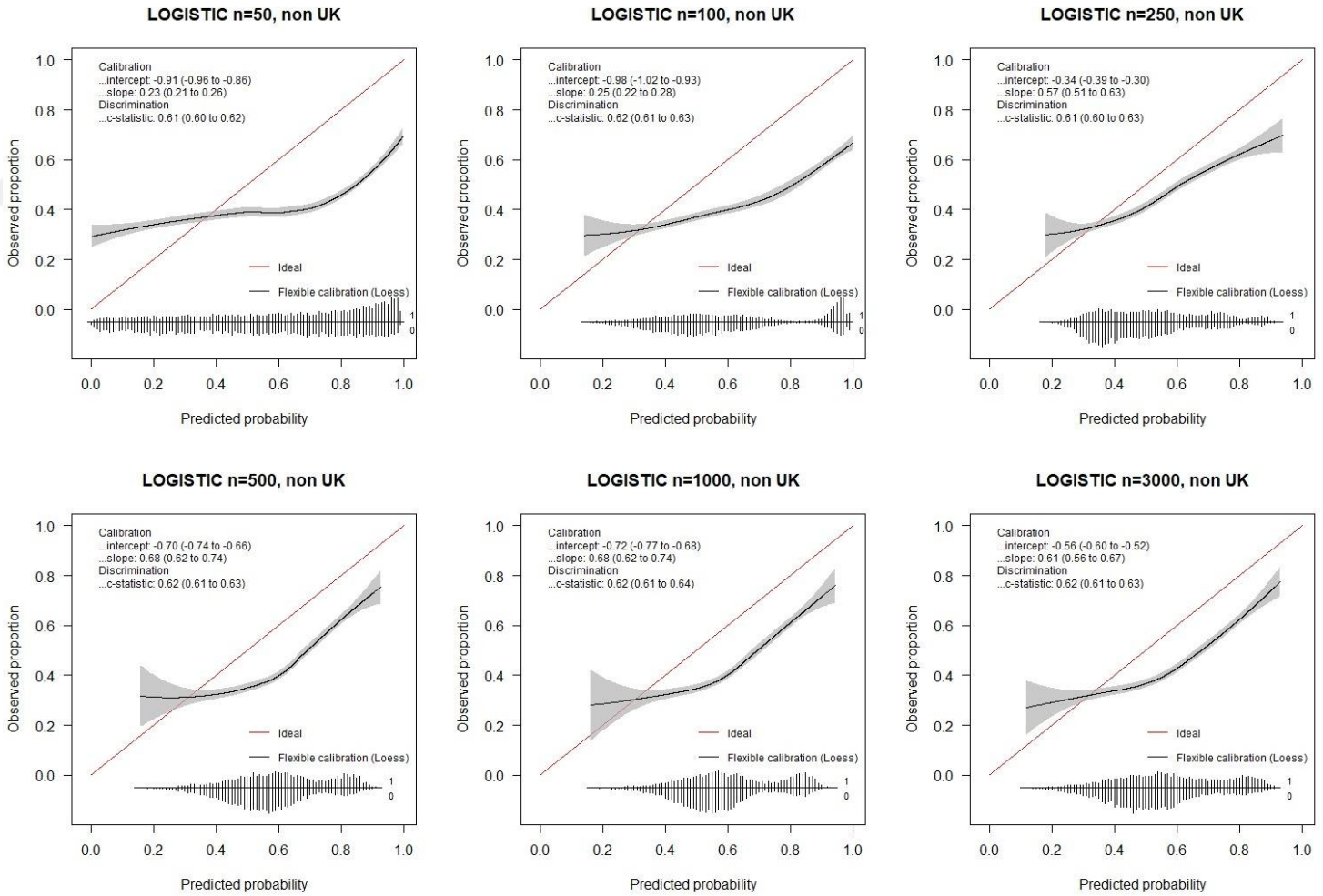


Figure B.17 Calibration plots of logistic regression models with 6 different sizes, non-UK data

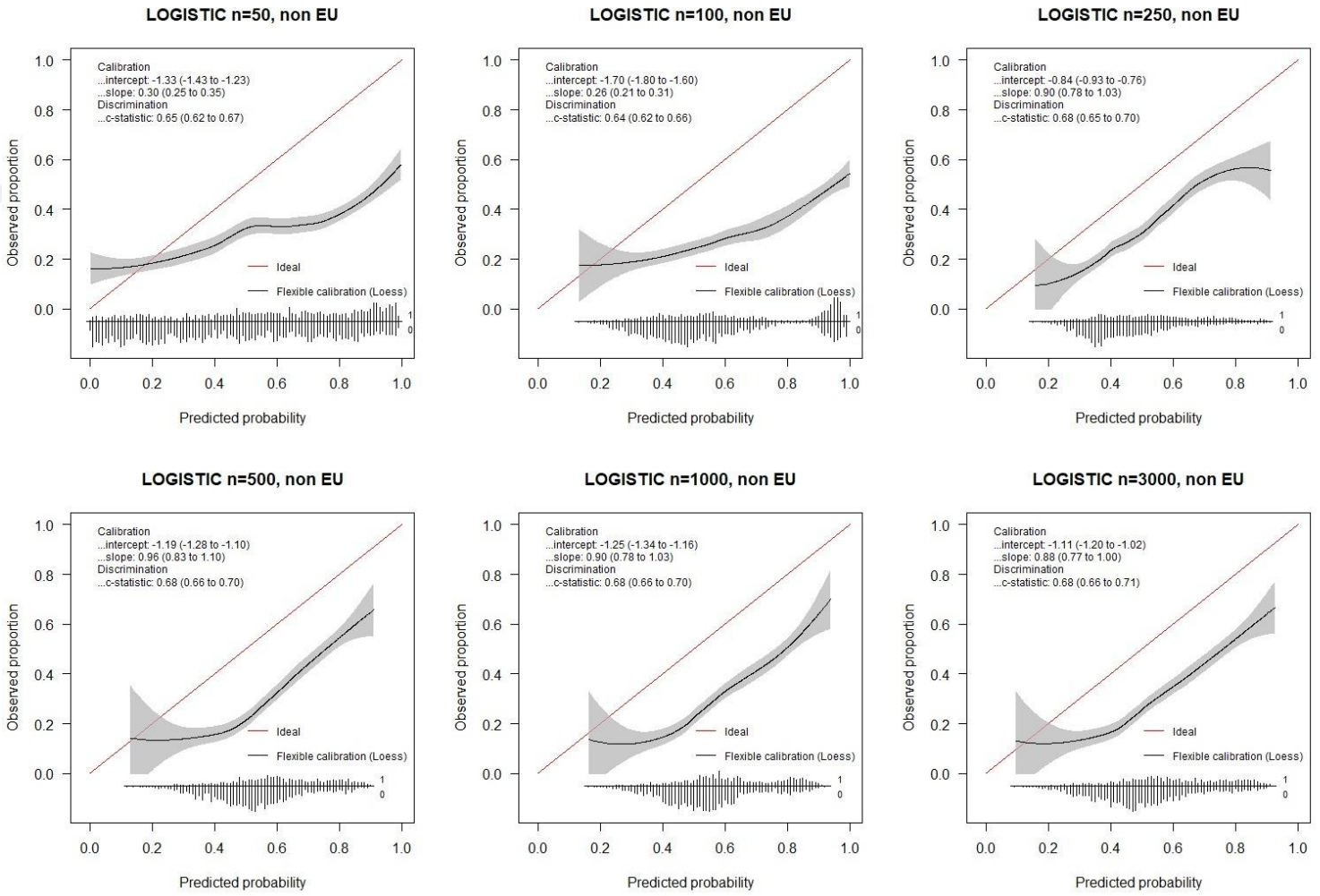


Figure B.18 Calibration plots of logistic regression models with 6 different sizes, non-EU data

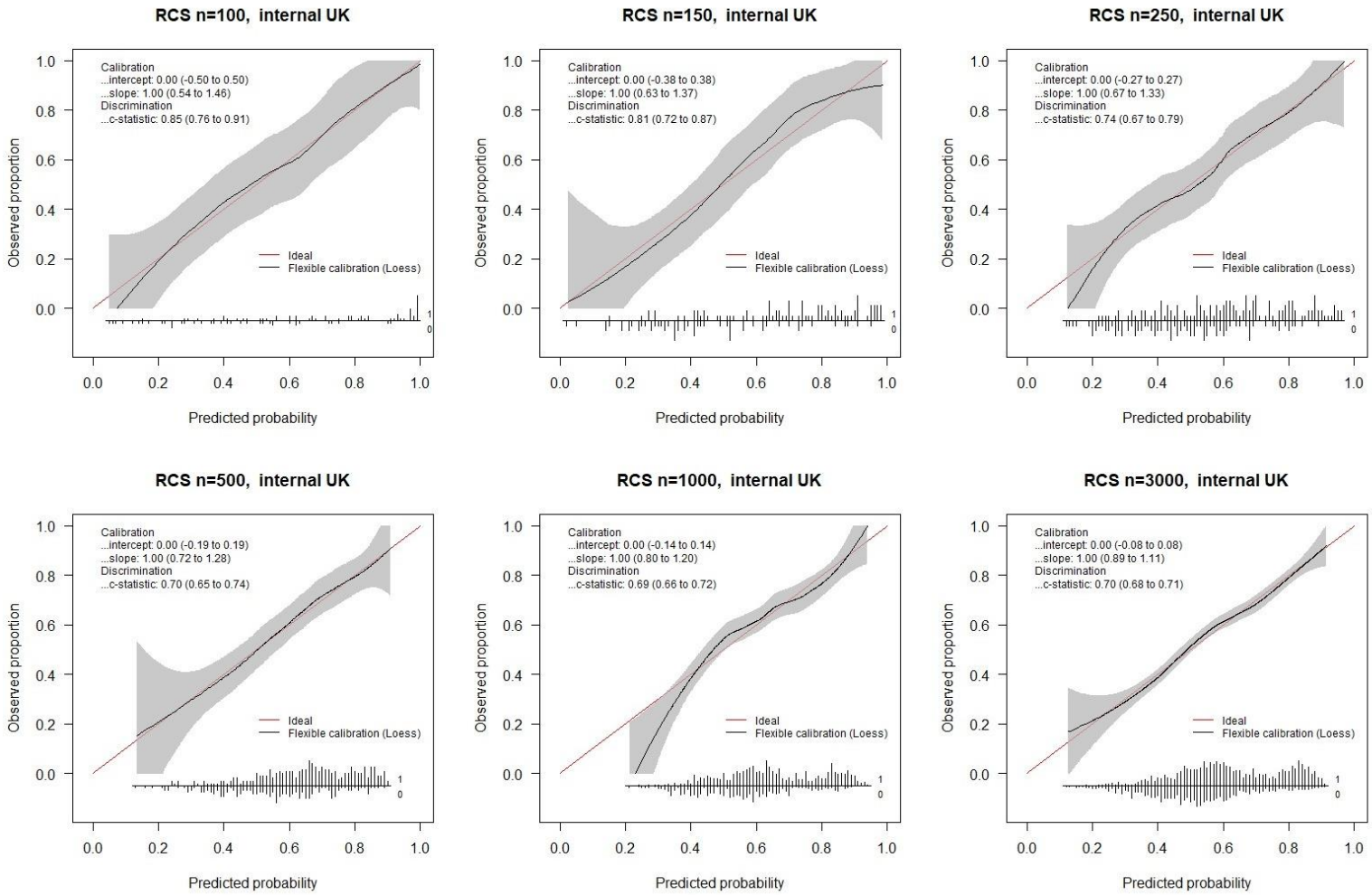


Figure B.19 Calibration plots of logistic regression models (RCS) with 6 different sizes, UK data

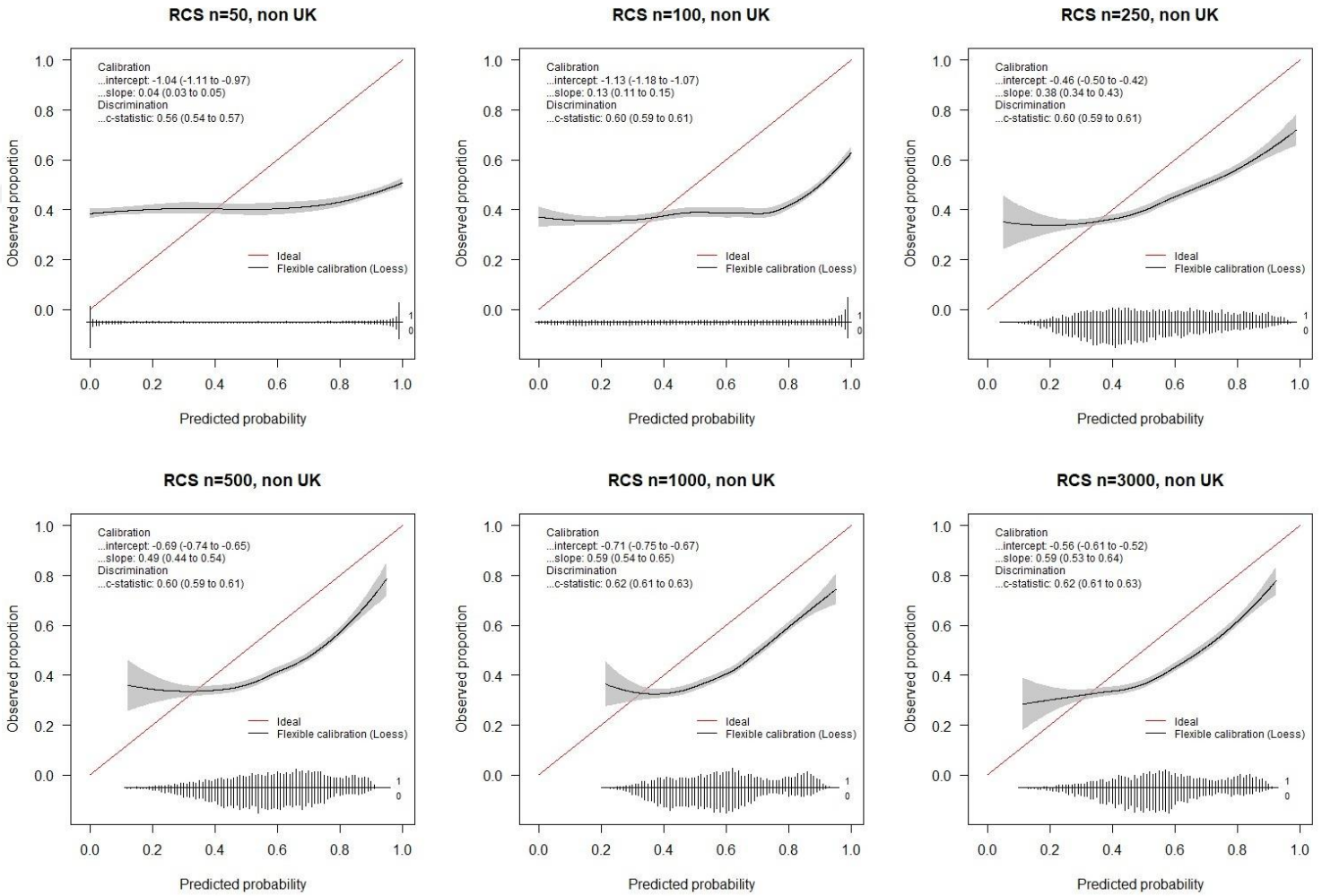


Figure B.20 Calibration plots of logistic regression models (RCS) with 6 different sizes, non-UK data

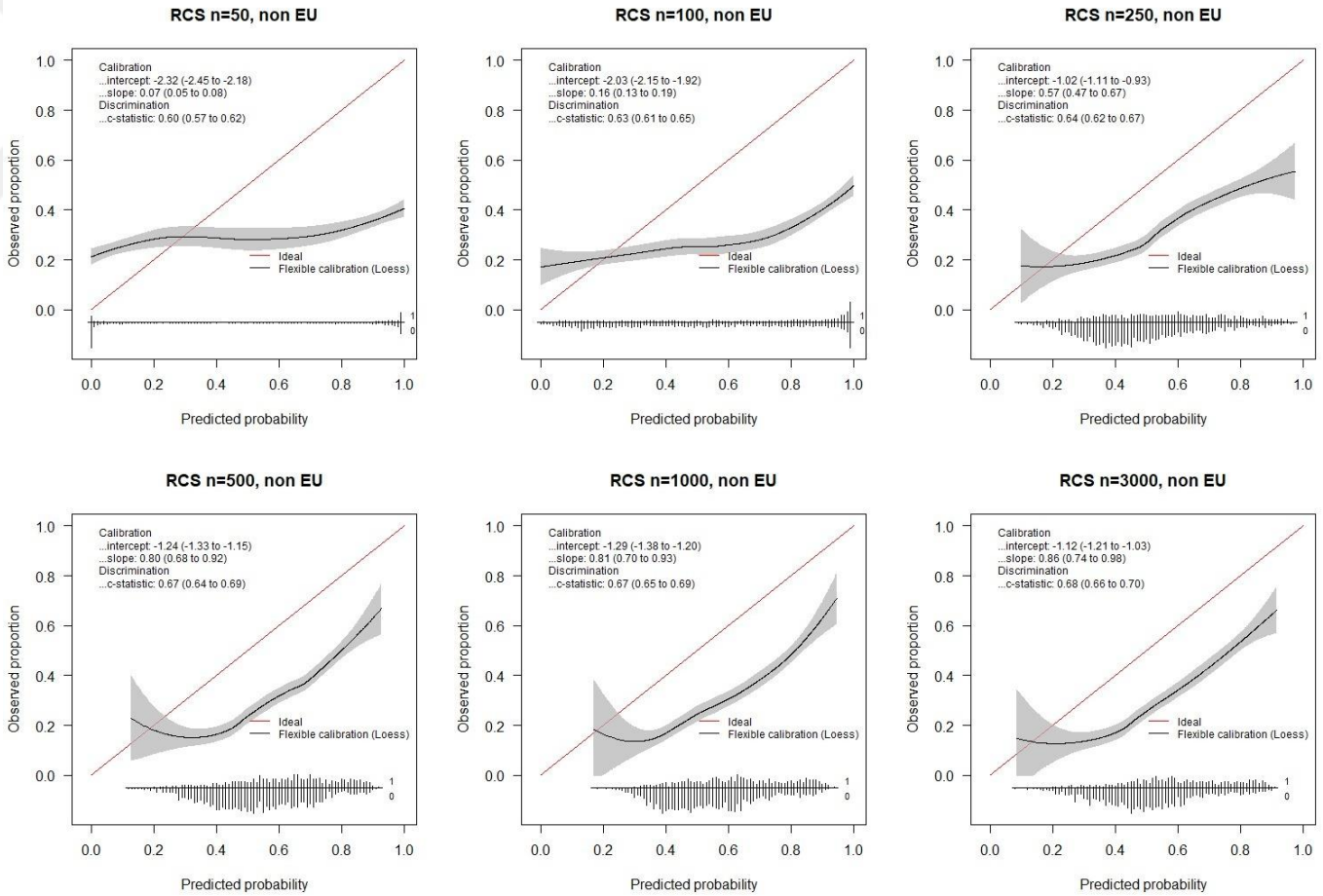


Figure B.21 Calibration plots of logistic regression models (RCS) with 6 different sizes, non-EU data

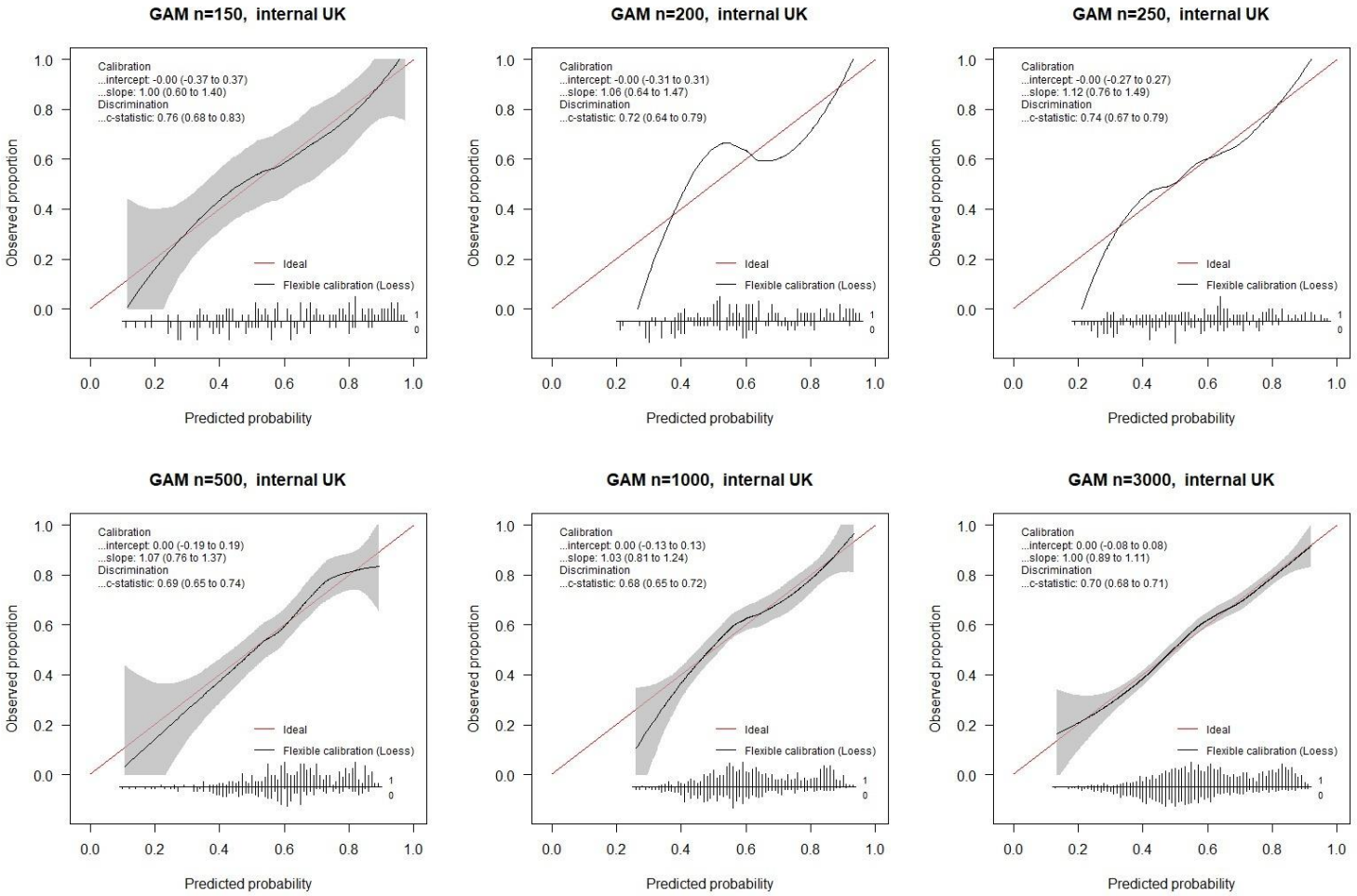


Figure B.22 Calibration plots of generalized additive models with 6 different sizes, UK data

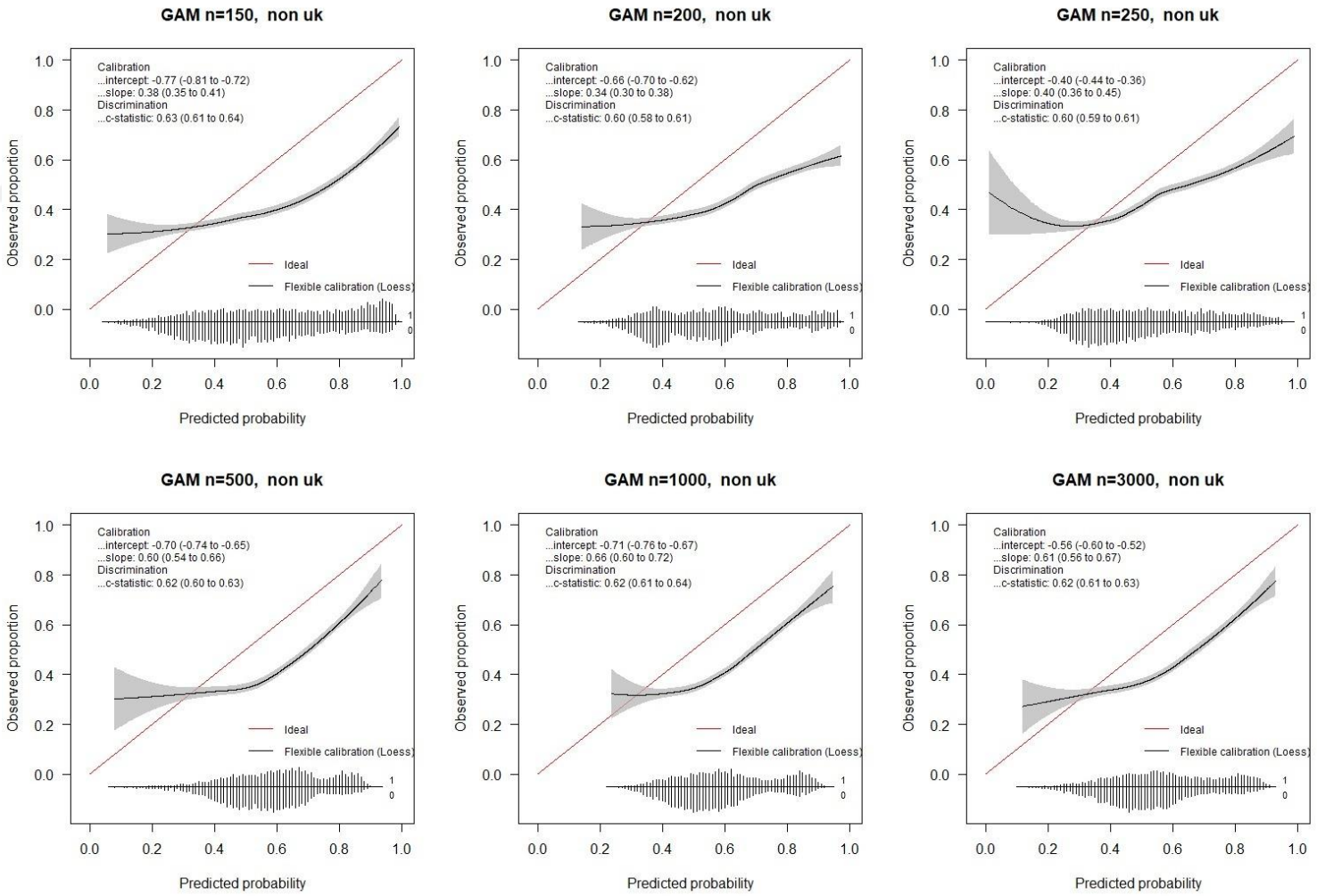


Figure B.23 Calibration plots of generalized additive models with 6 different sizes, non-UK data

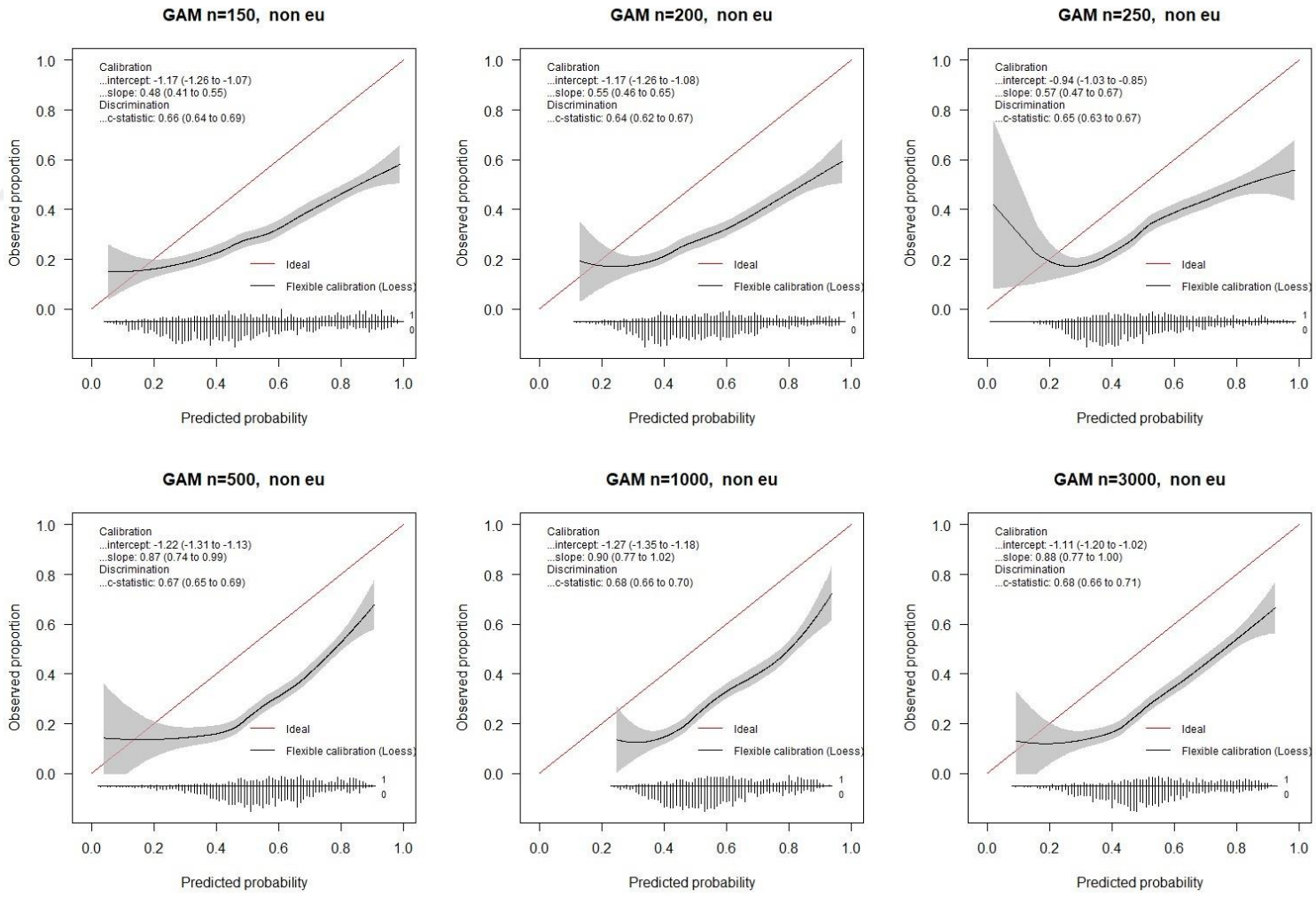


Figure B.24 Calibration plots of generalized additive models with 6 different sizes, non-EU data

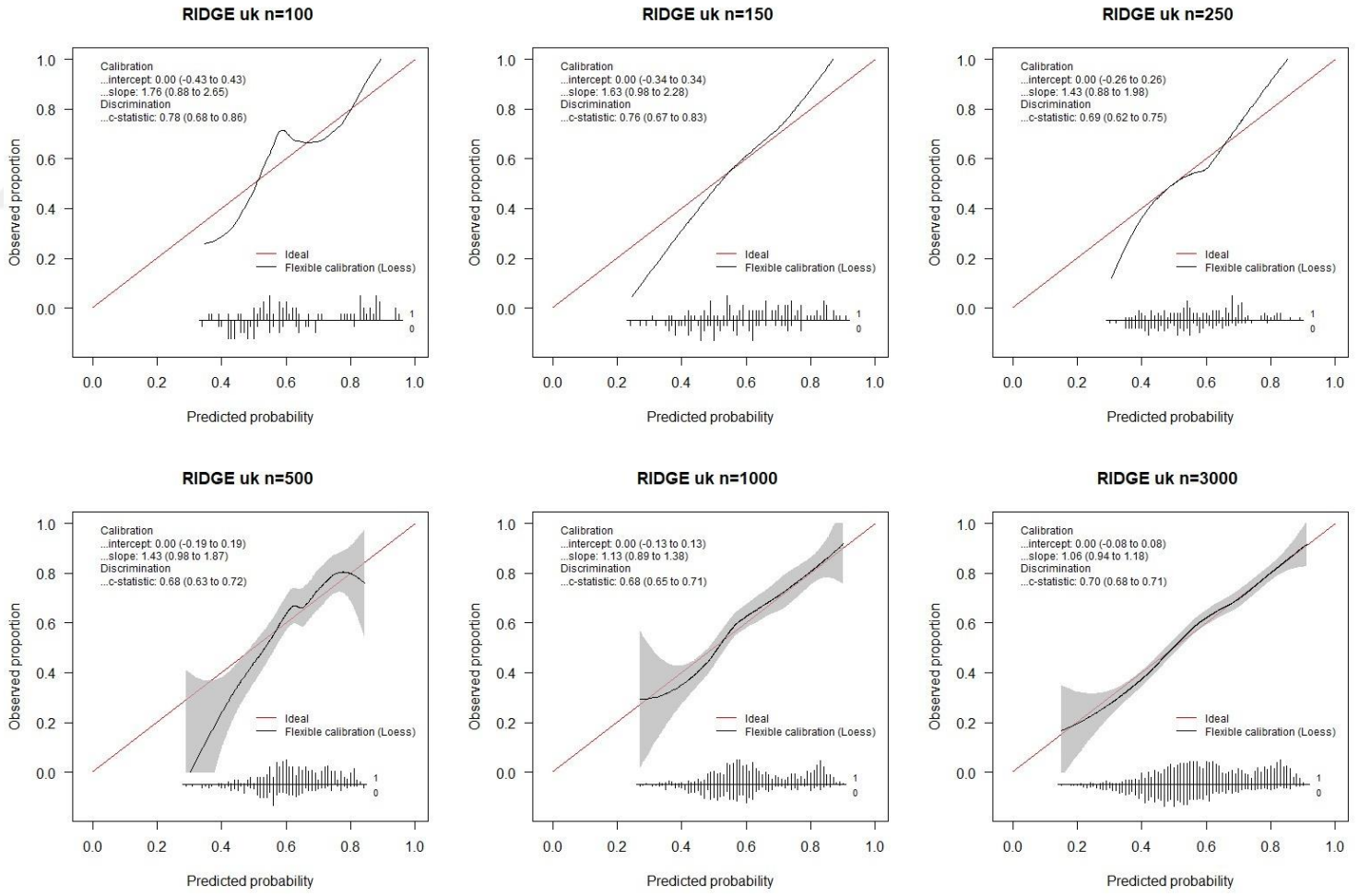


Figure B.25 Calibration plots of ridge regression models with 6 different sizes, UK data

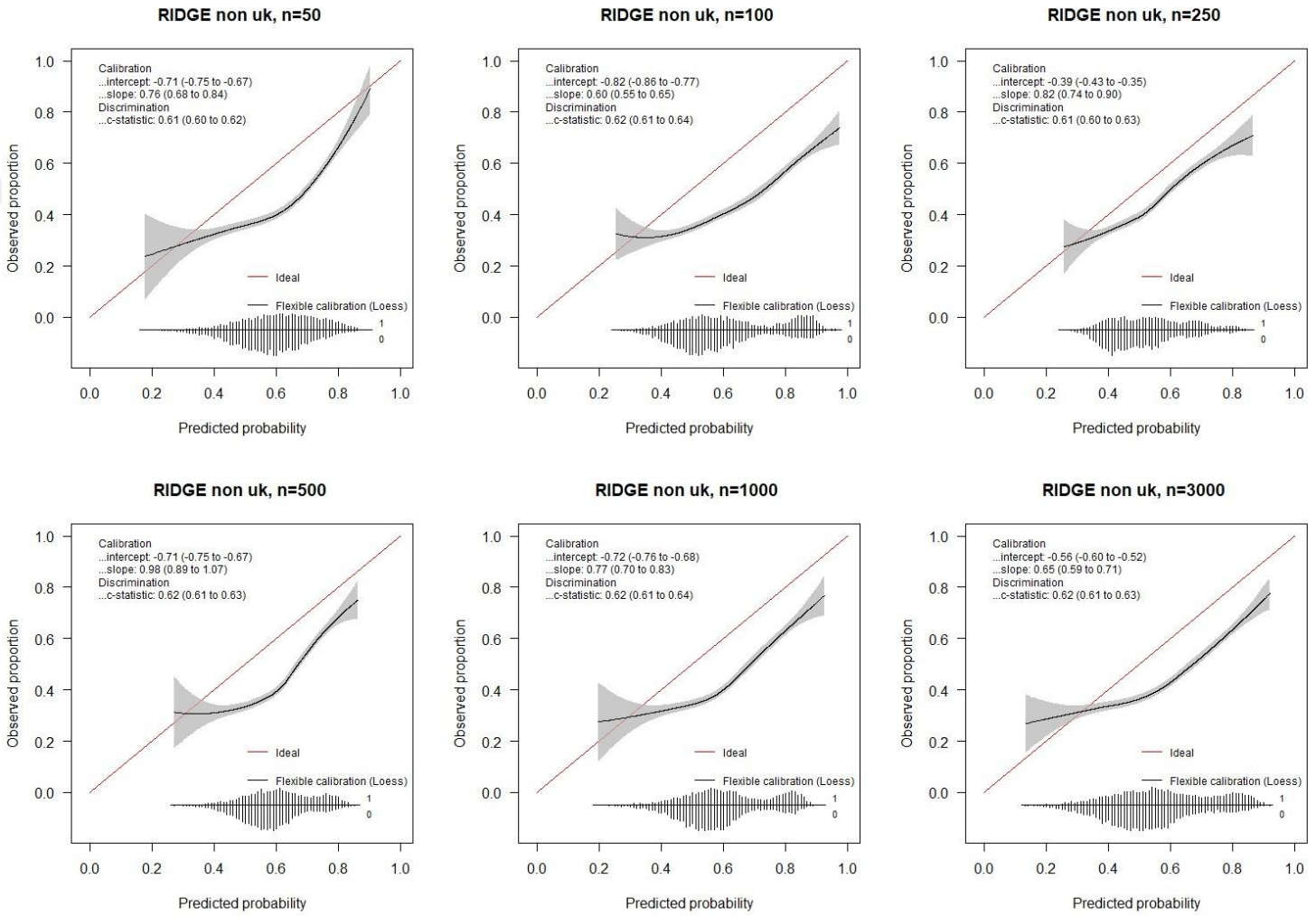


Figure B.26 Calibration plots of ridge regression models with 6 different sizes, non-UK data

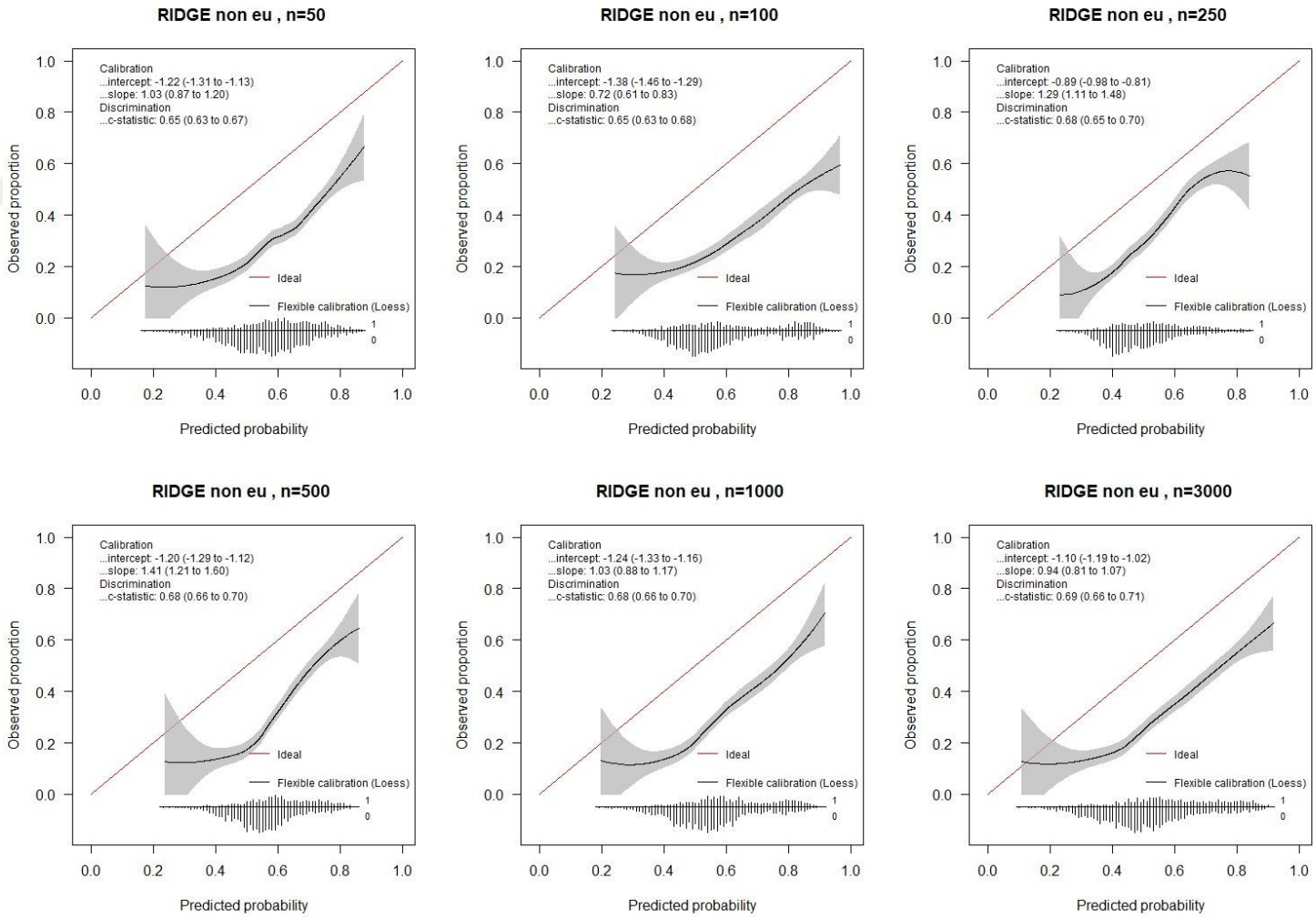


Figure B.27 Calibration plots of ridge regression models with 6 different sizes, non-EU data

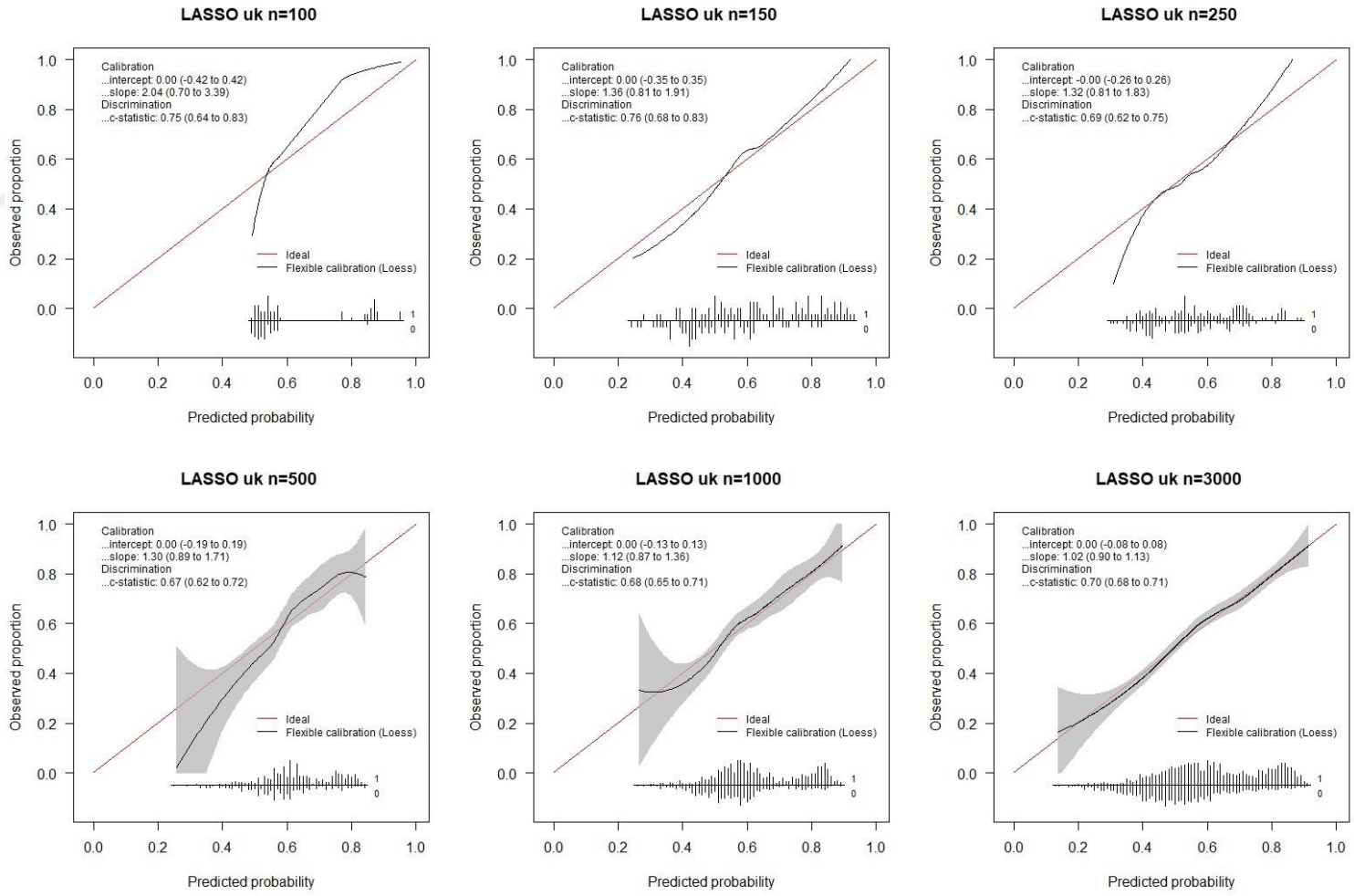


Figure B.28 Calibration plots of lasso regression models with 6 different sizes, UK data

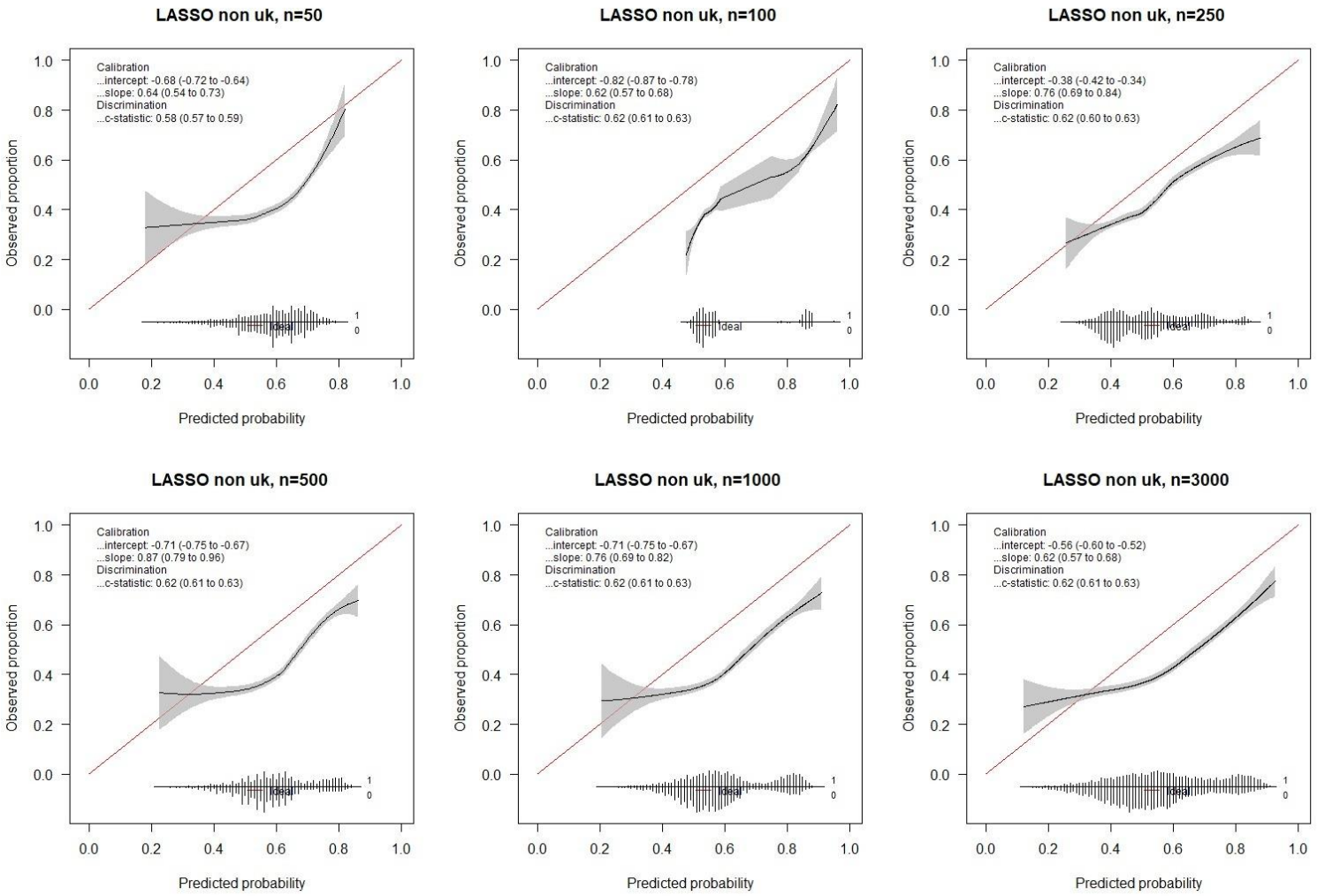


Figure B.29 Calibration plots of lasso regression models with 6 different sizes, non-UK data

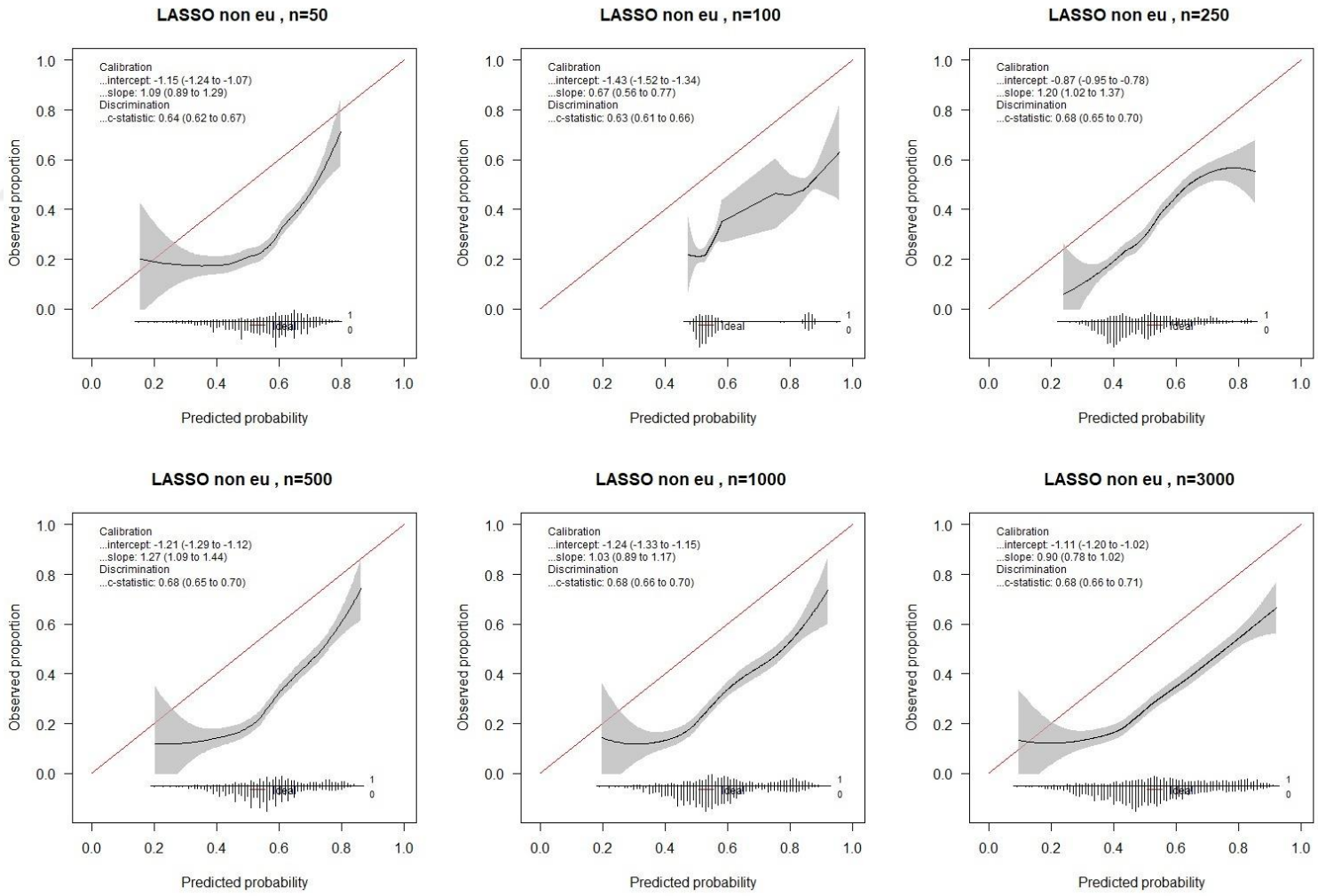


Figure B.30 Calibration plots of lasso regression models with 6 different sizes, non-EU data

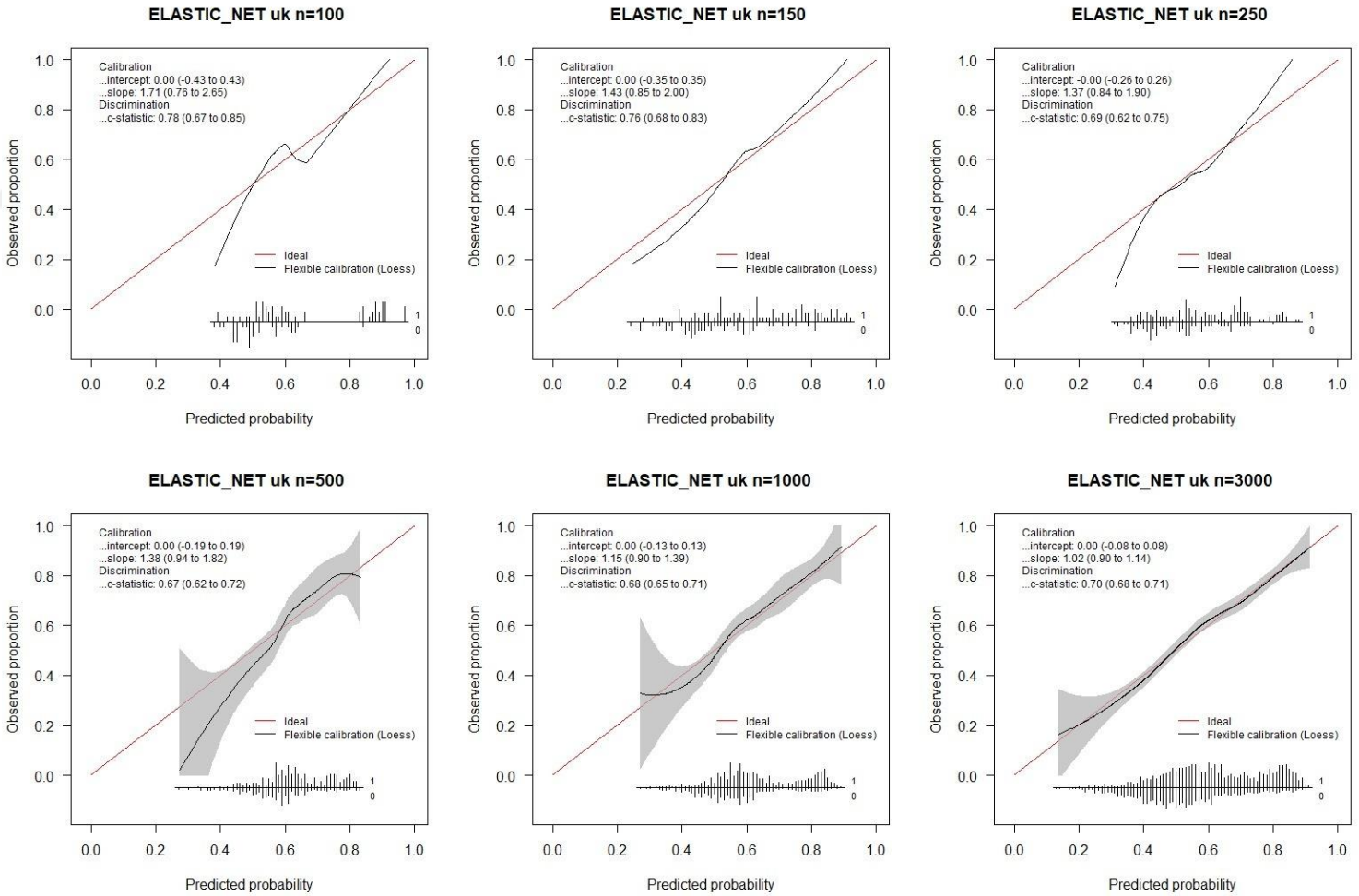


Figure B.31 Calibration plots of elastic net regression models with 6 different sizes, UK data

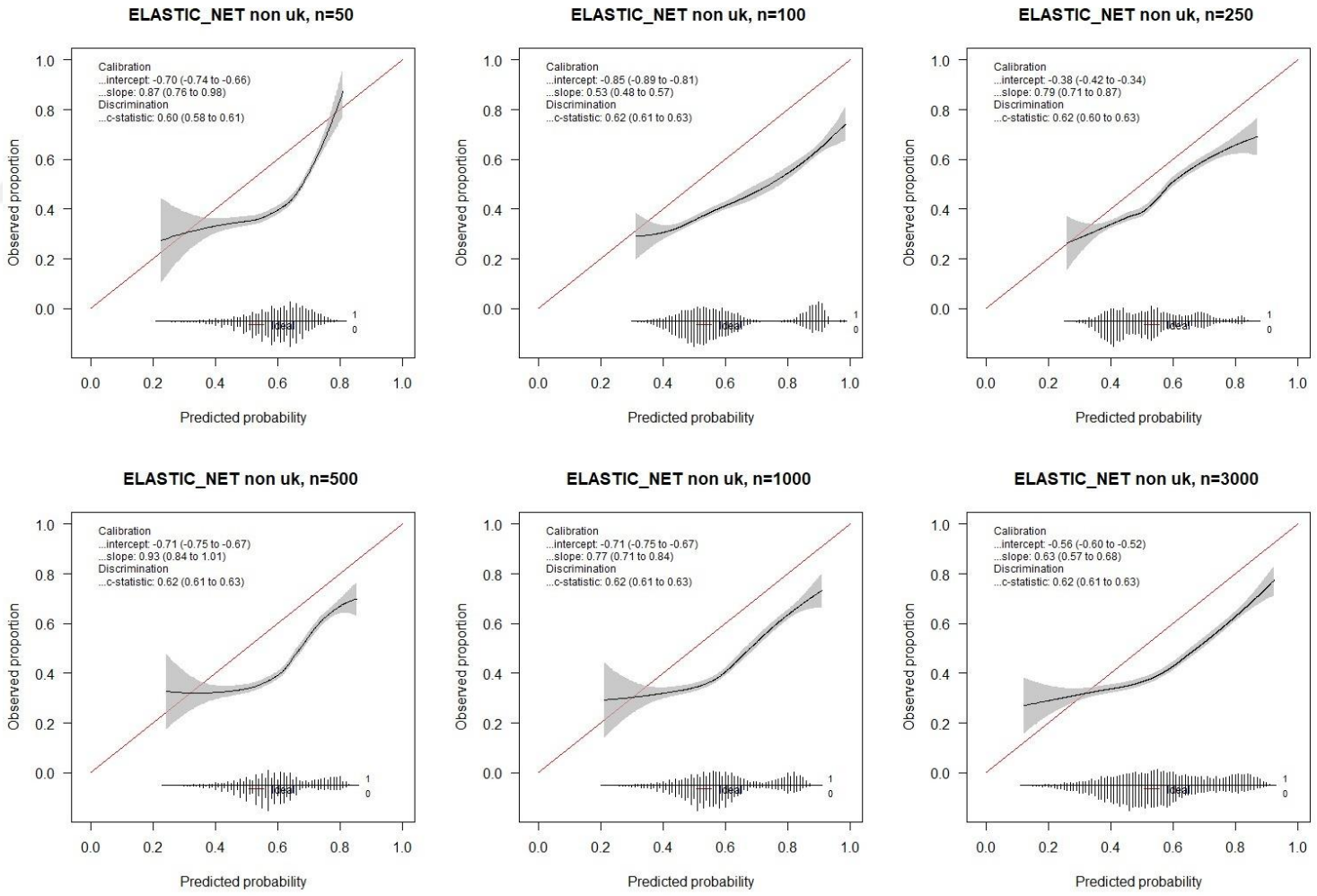


Figure B.32 Calibration plots of elastic net regression models with 6 different sizes, non-UK data

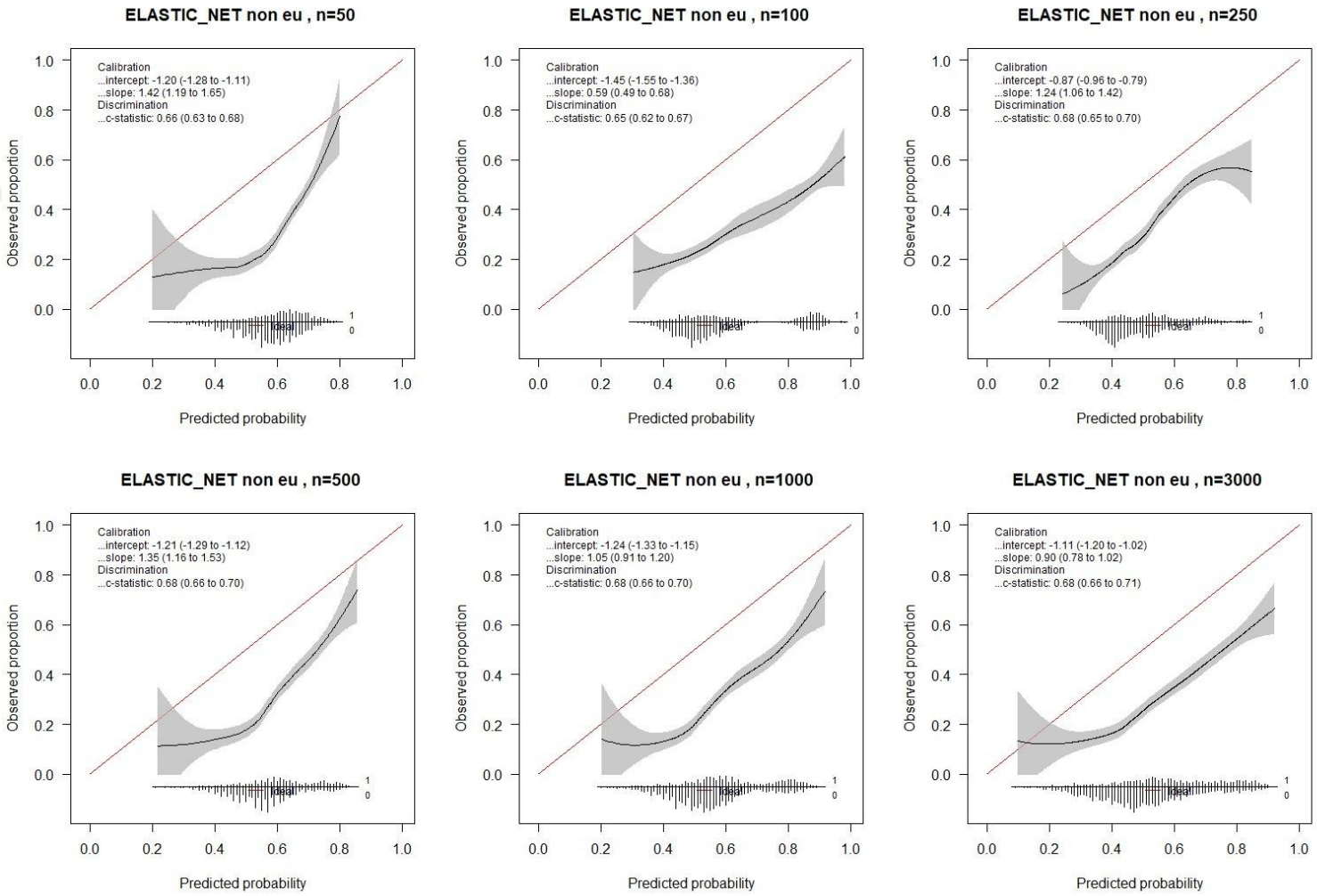


Figure B.33 Calibration plots of elastic net regression models with 6 different sizes, non-EU data

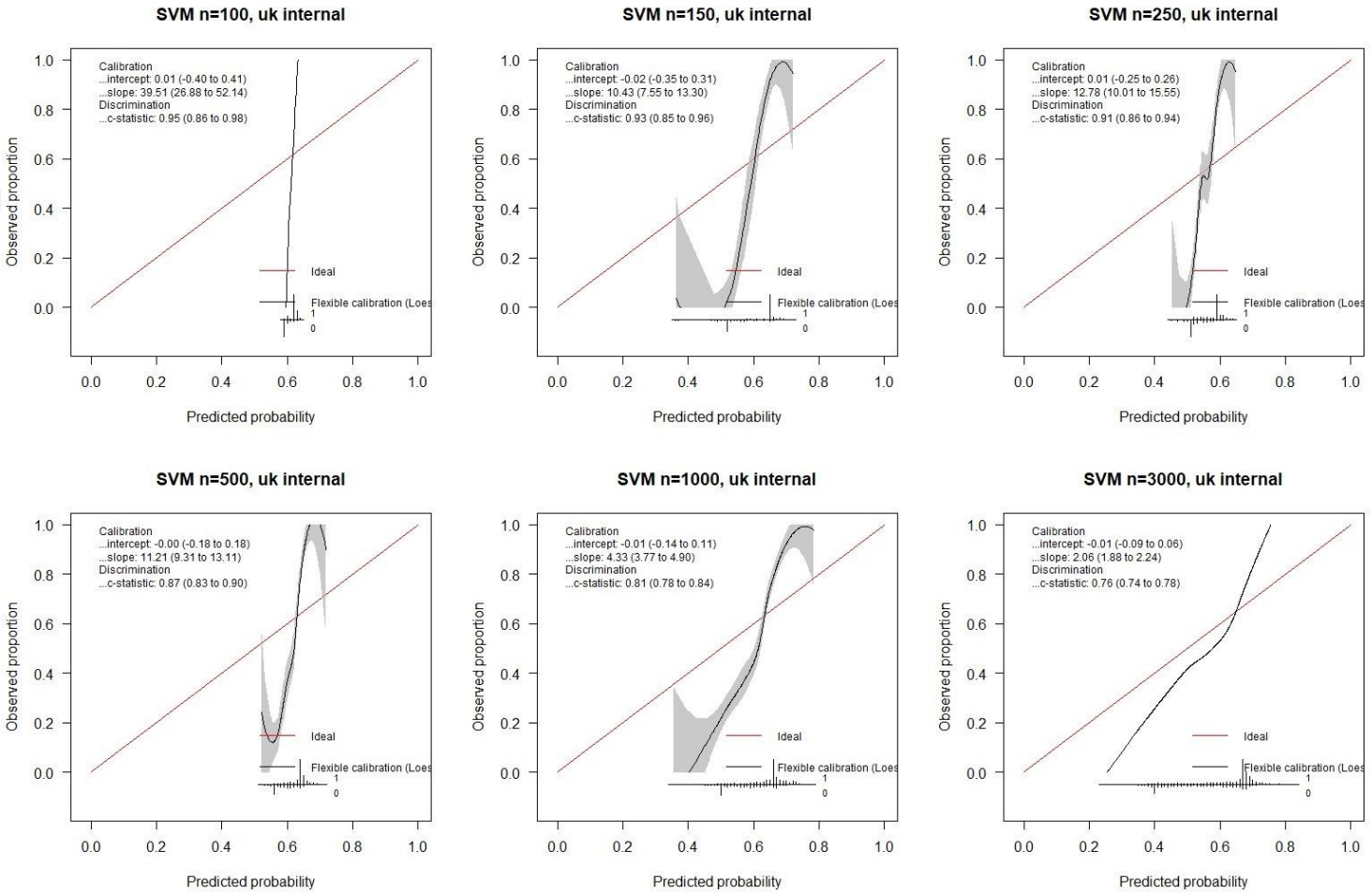


Figure B.34 Calibration plots of SVM models with 6 different sizes, UK data

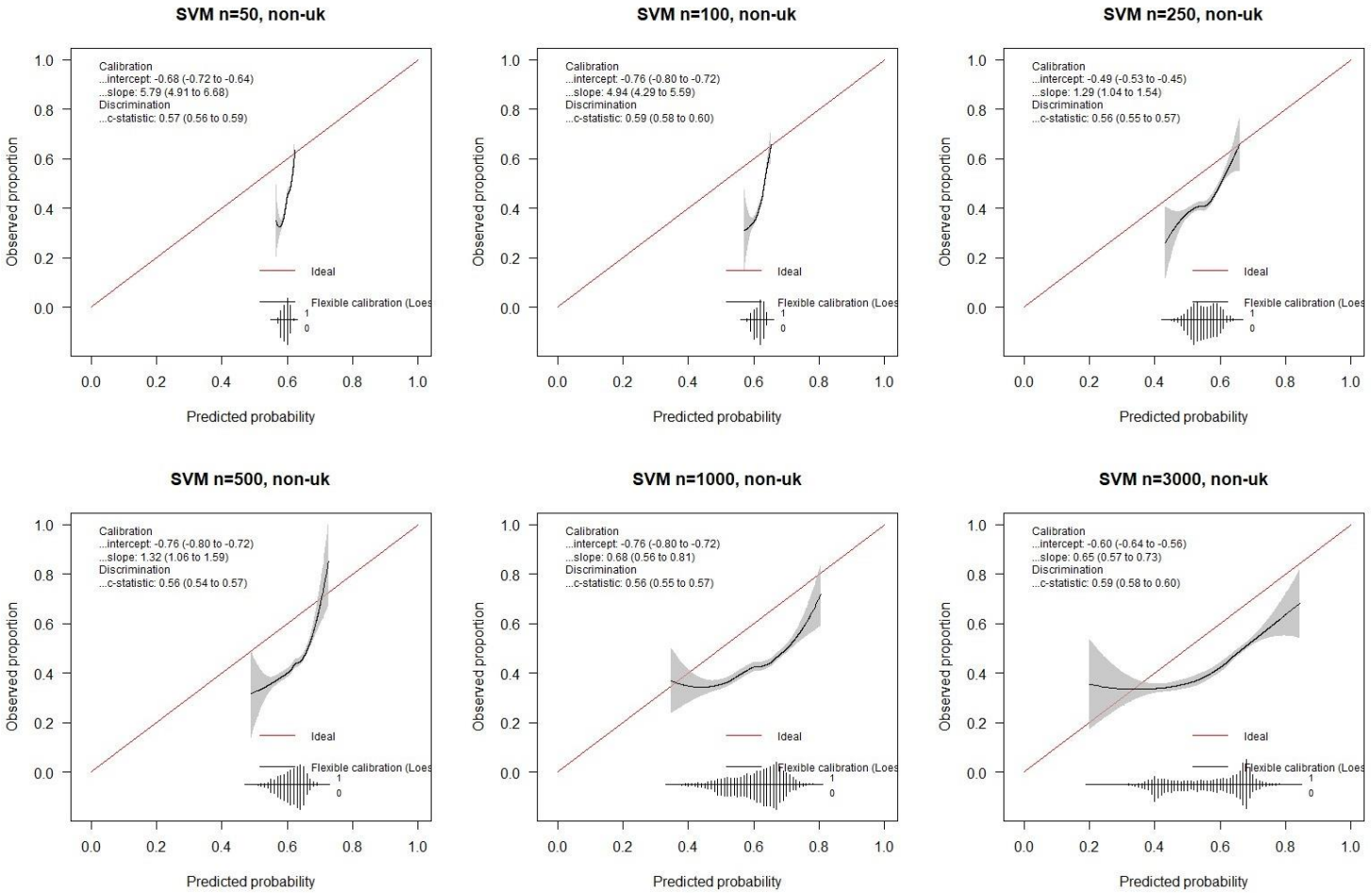


Figure B.35 Calibration plots of SVM models with 6 different sizes, non-UK data

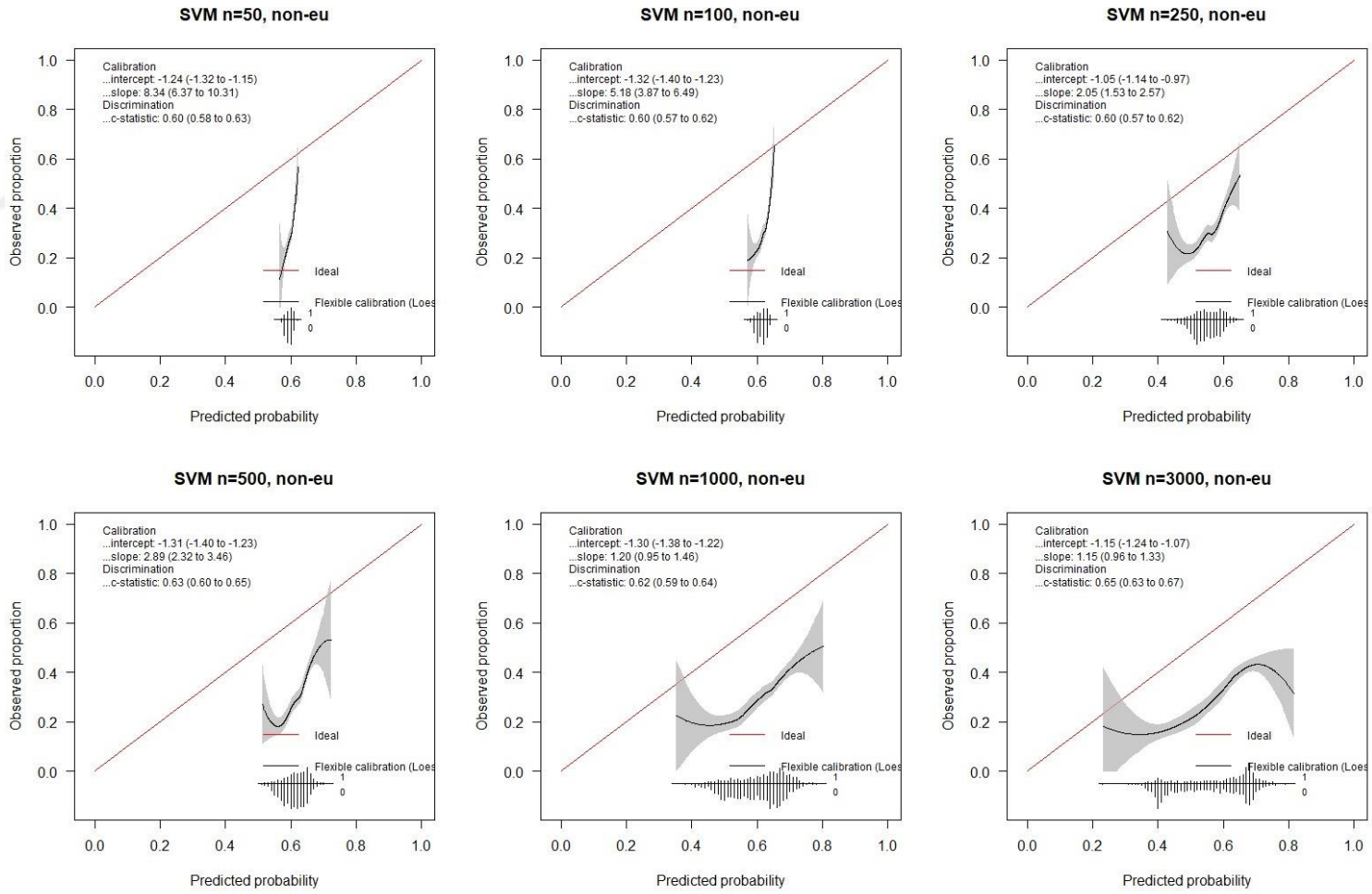


Figure B.36 Calibration plots of SVM models with 6 different sizes, non-EU data

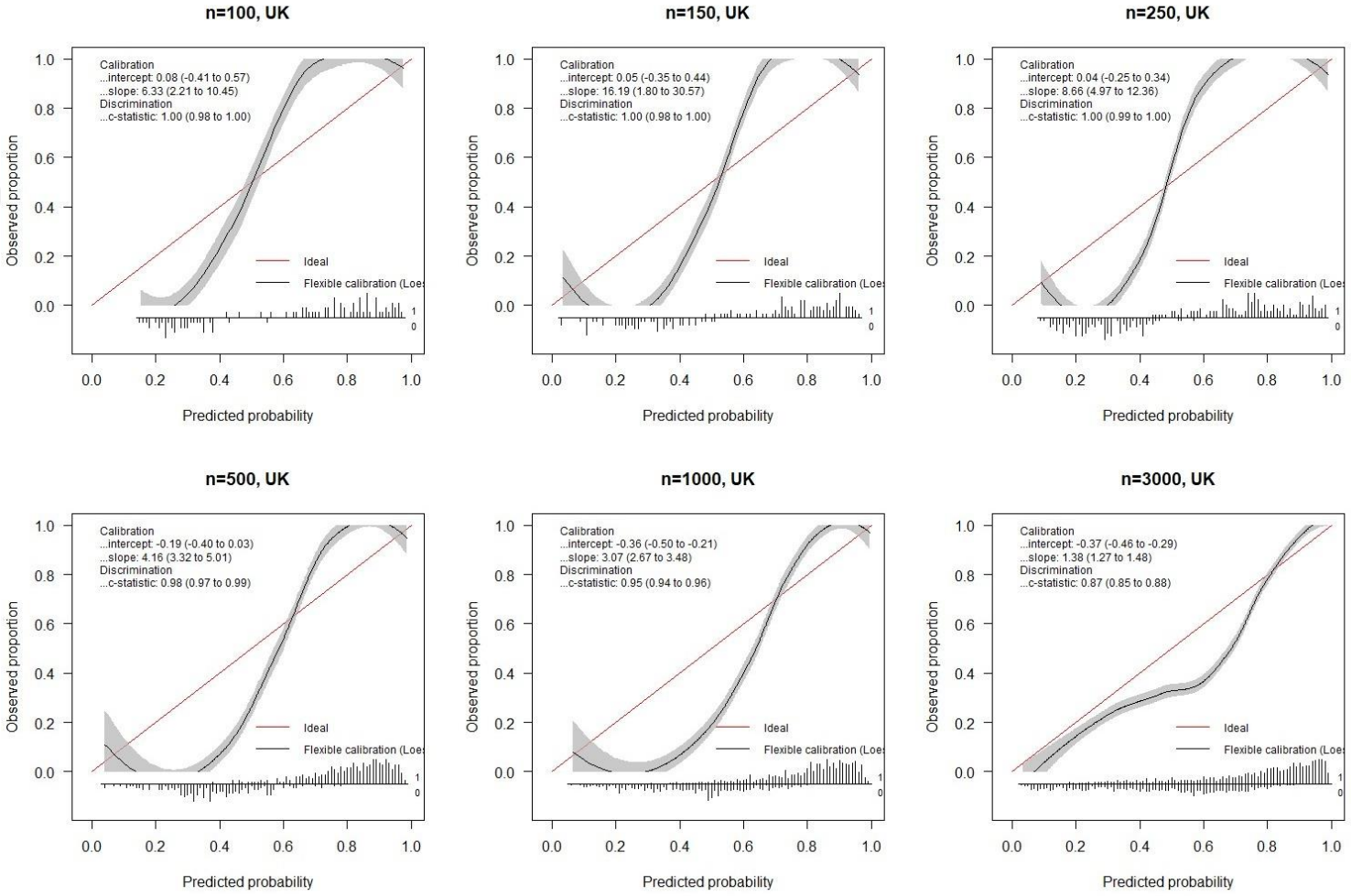


Figure B.37 Calibration plots of random forest models with 6 different sizes, UK data

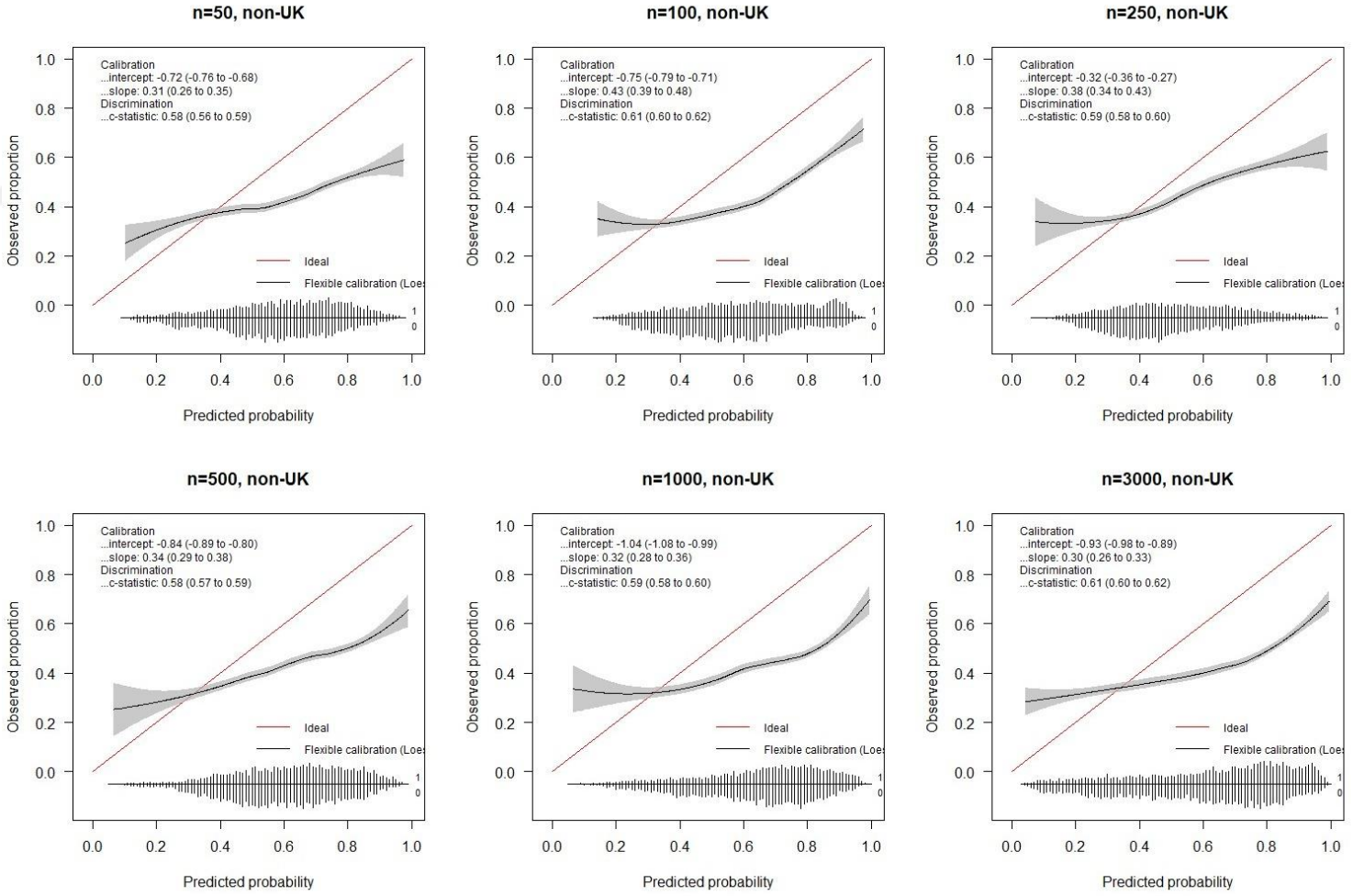


Figure B.38 Calibration plots of random forest models with 6 different sizes, non-UK data

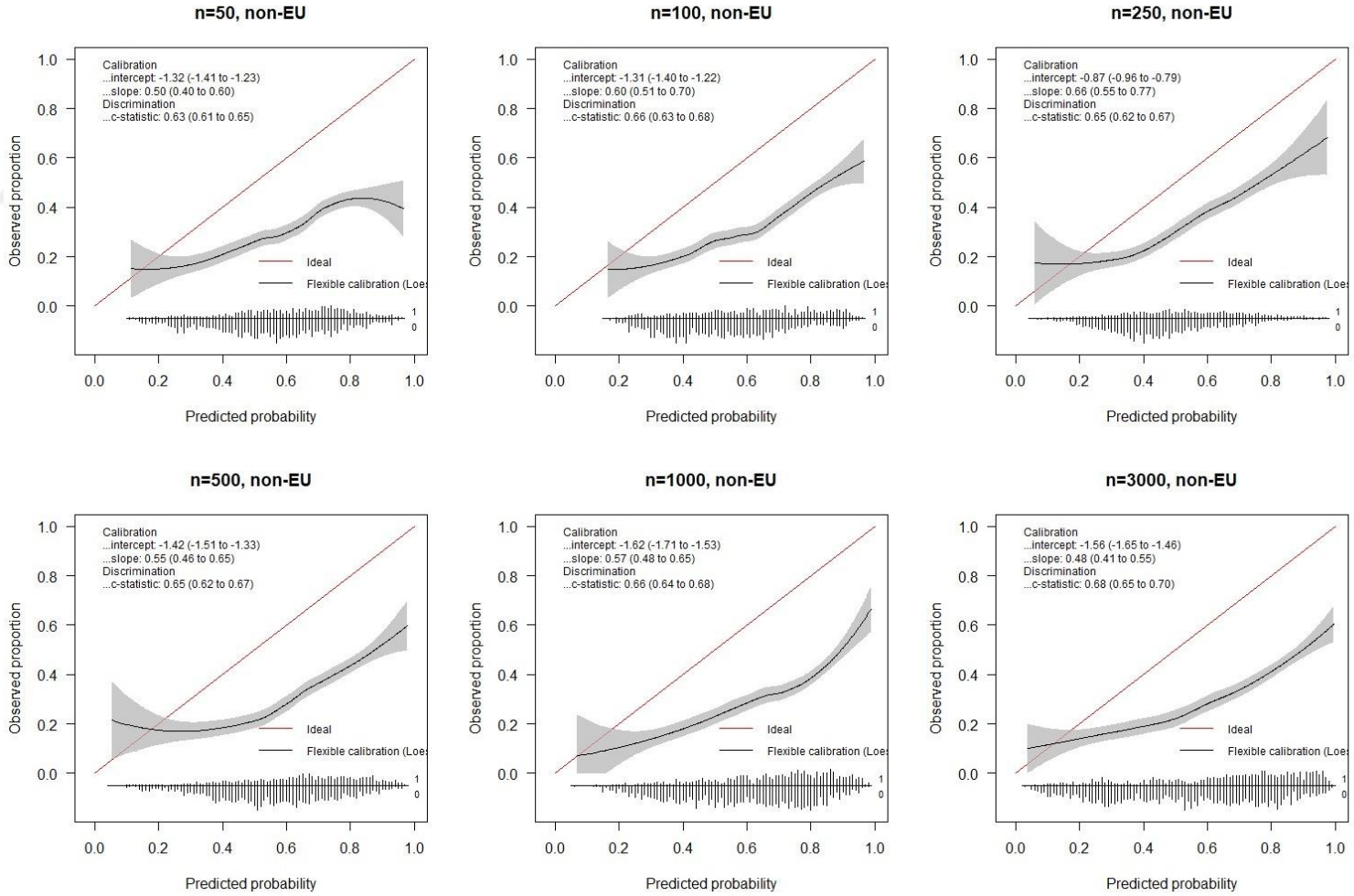


Figure B.39 Calibration plots of random forest models with 6 different sizes, non-EU data

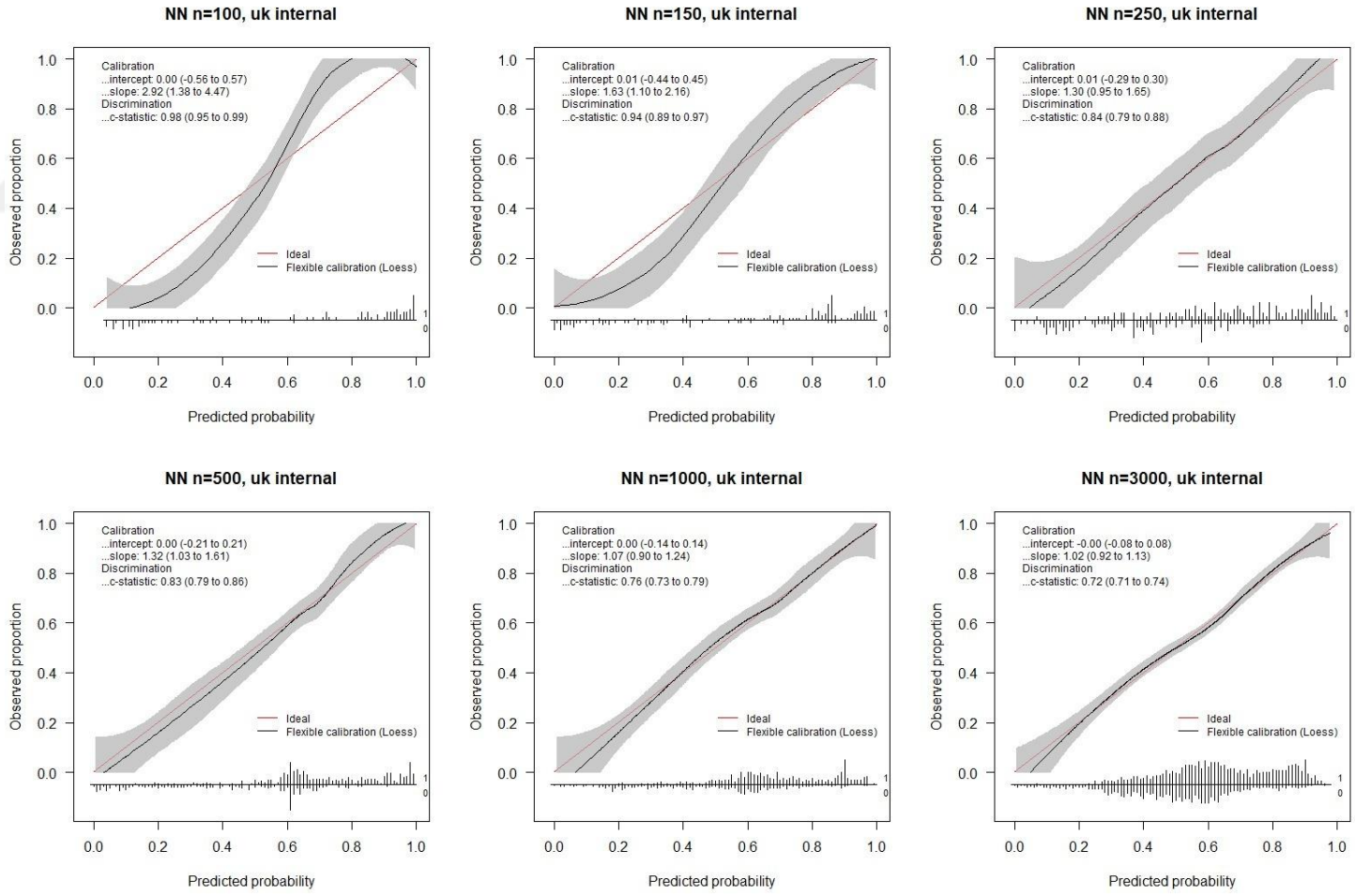


Figure B.40 Calibration plots of neural network models with 6 different sizes, UK data

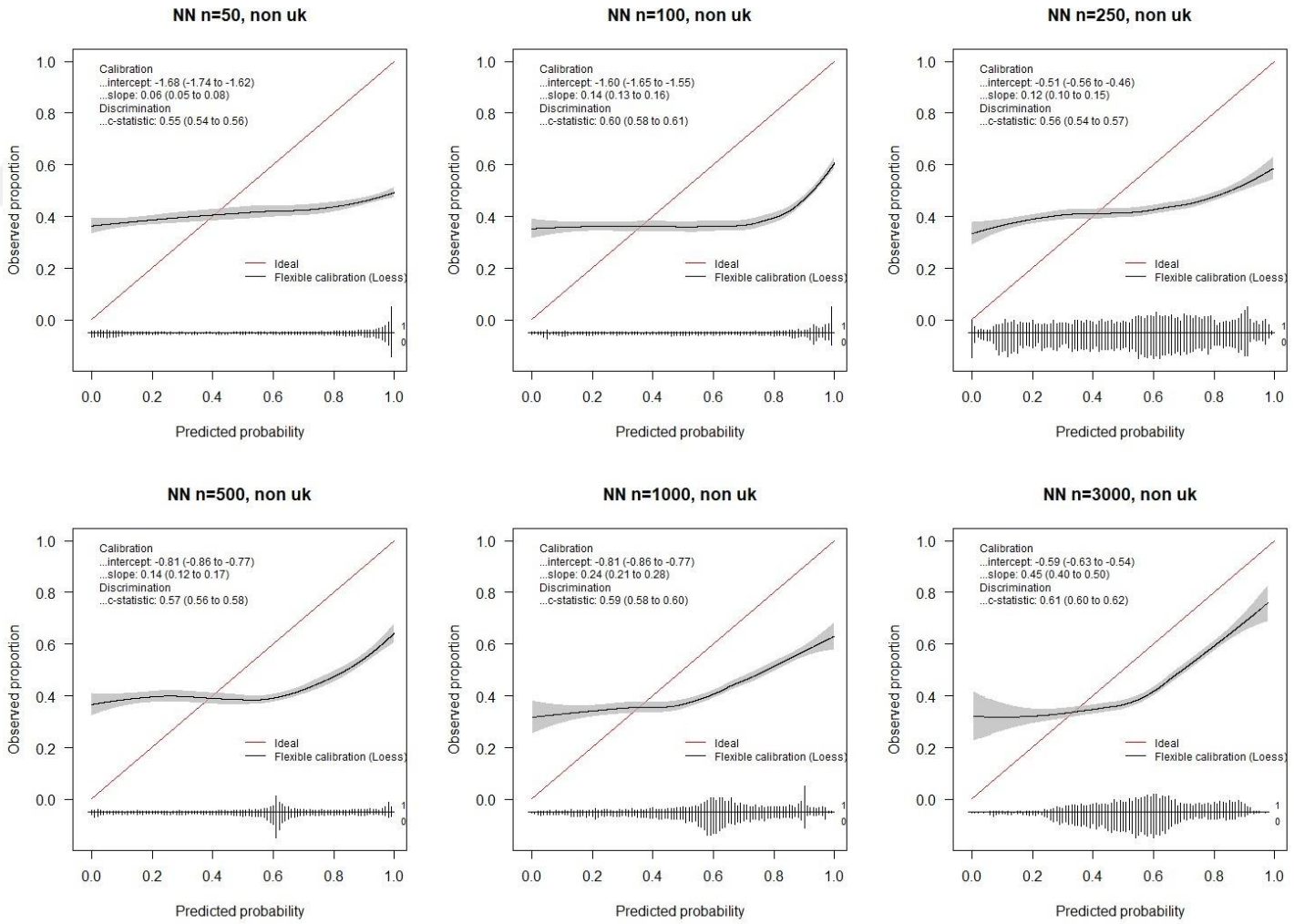


Figure B.41 Calibration plots of neural network models with 6 different sizes, non-UK data

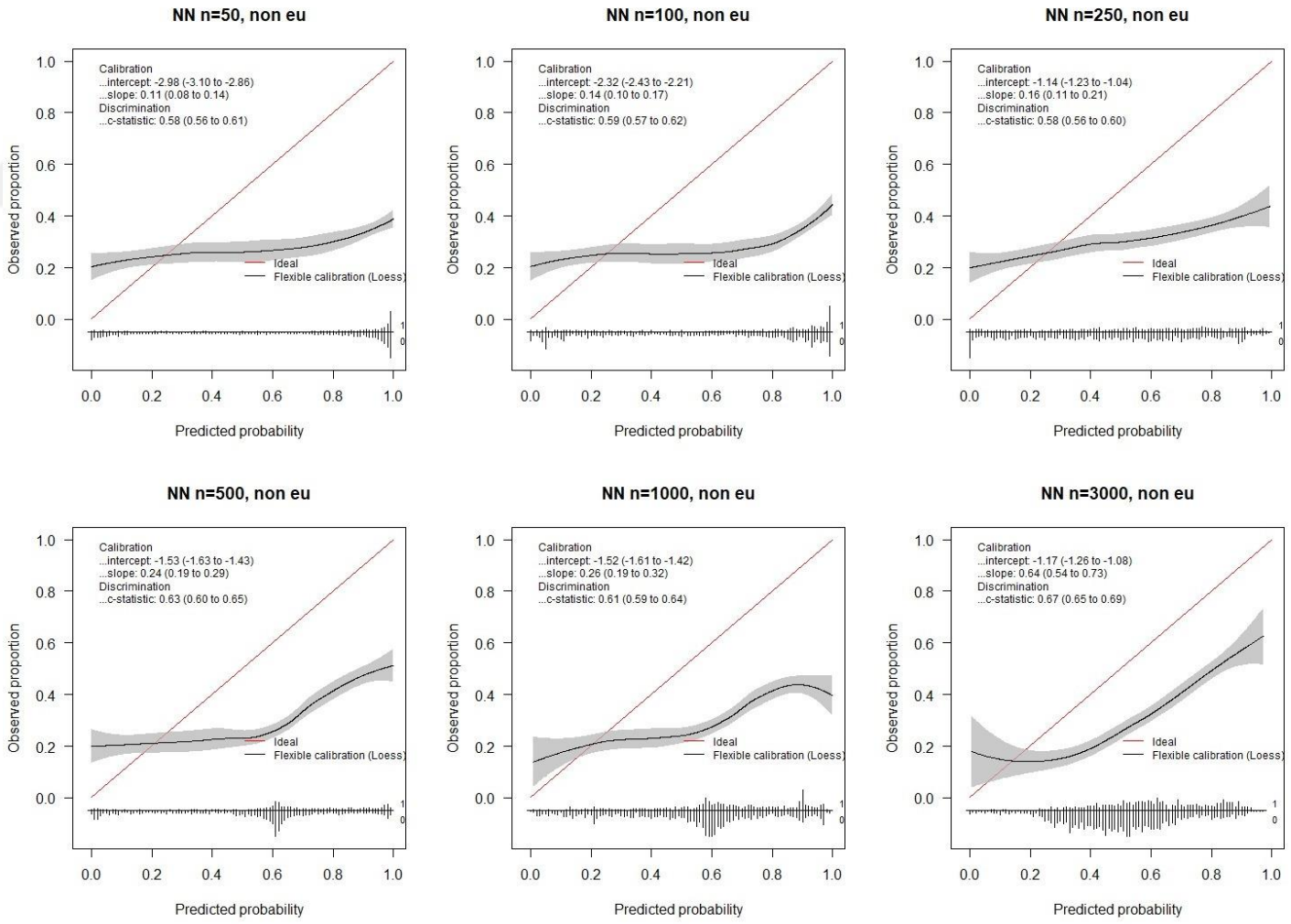


Figure B.42 Calibration plots of neural network models with 6 different sizes, non-EU data

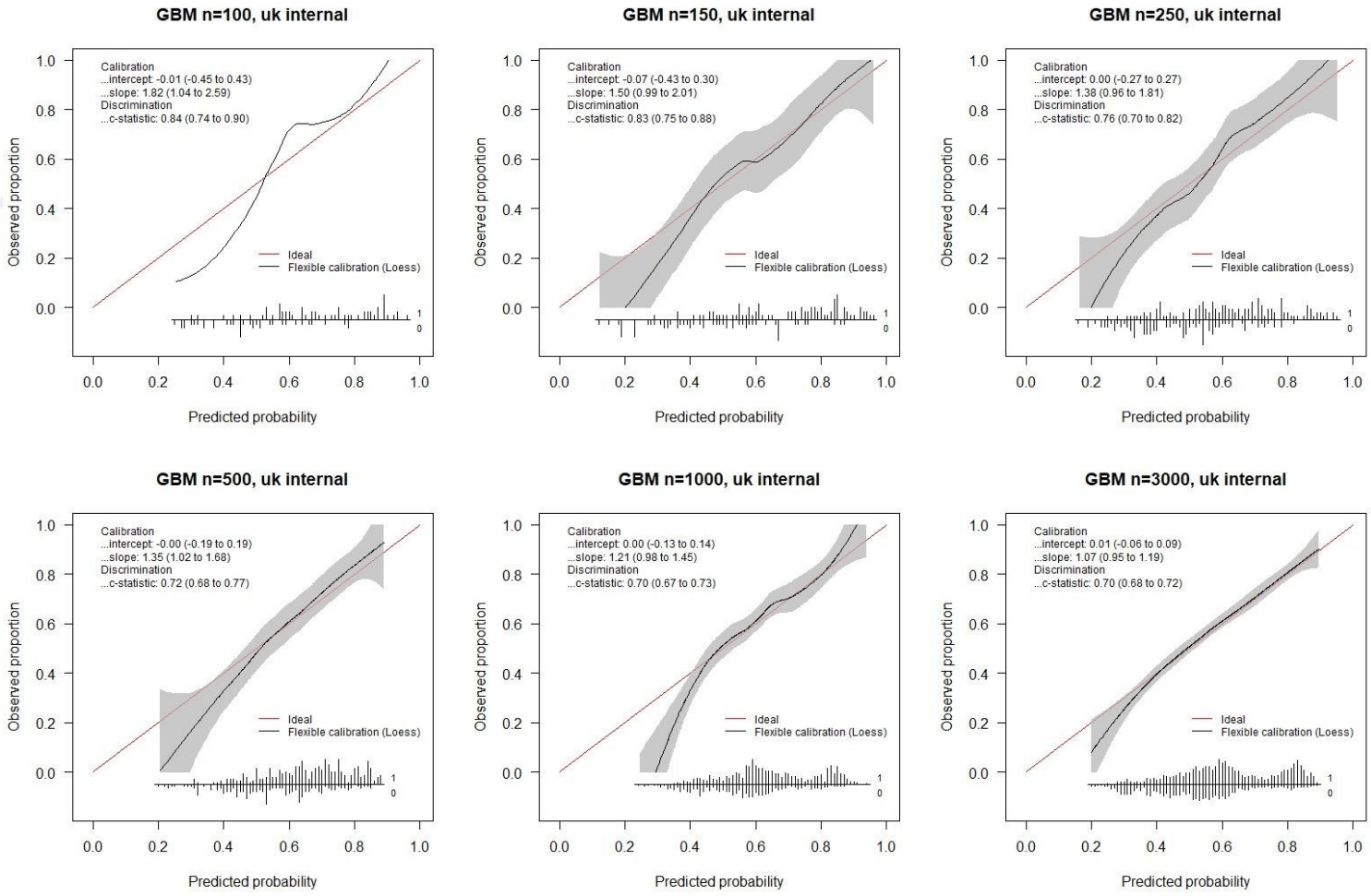


Figure B.43 Calibration plots of GBM models with 6 different sizes, UK data

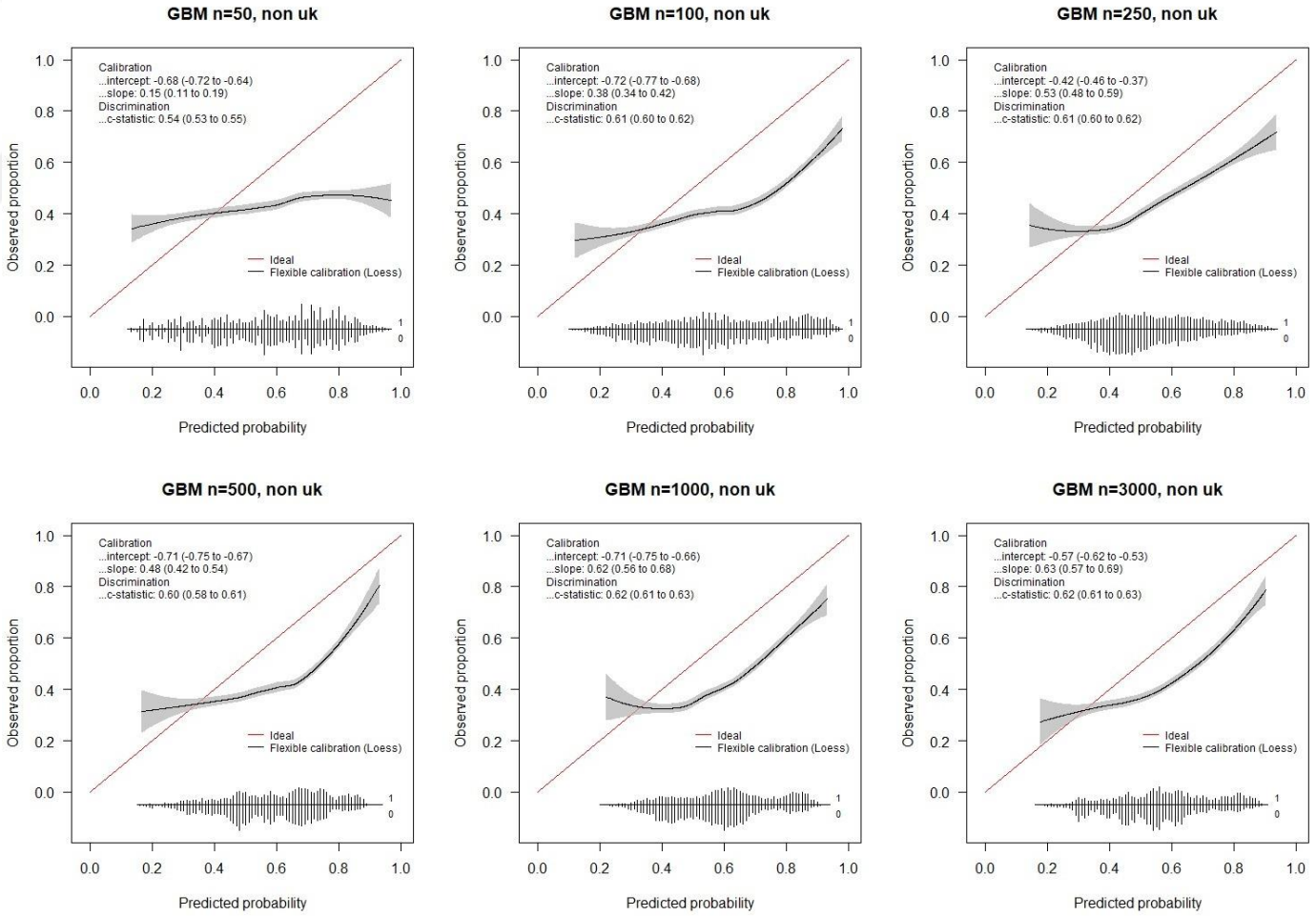


Figure B.44 Calibration plots of GBM models with 6 different sizes, non-UK data

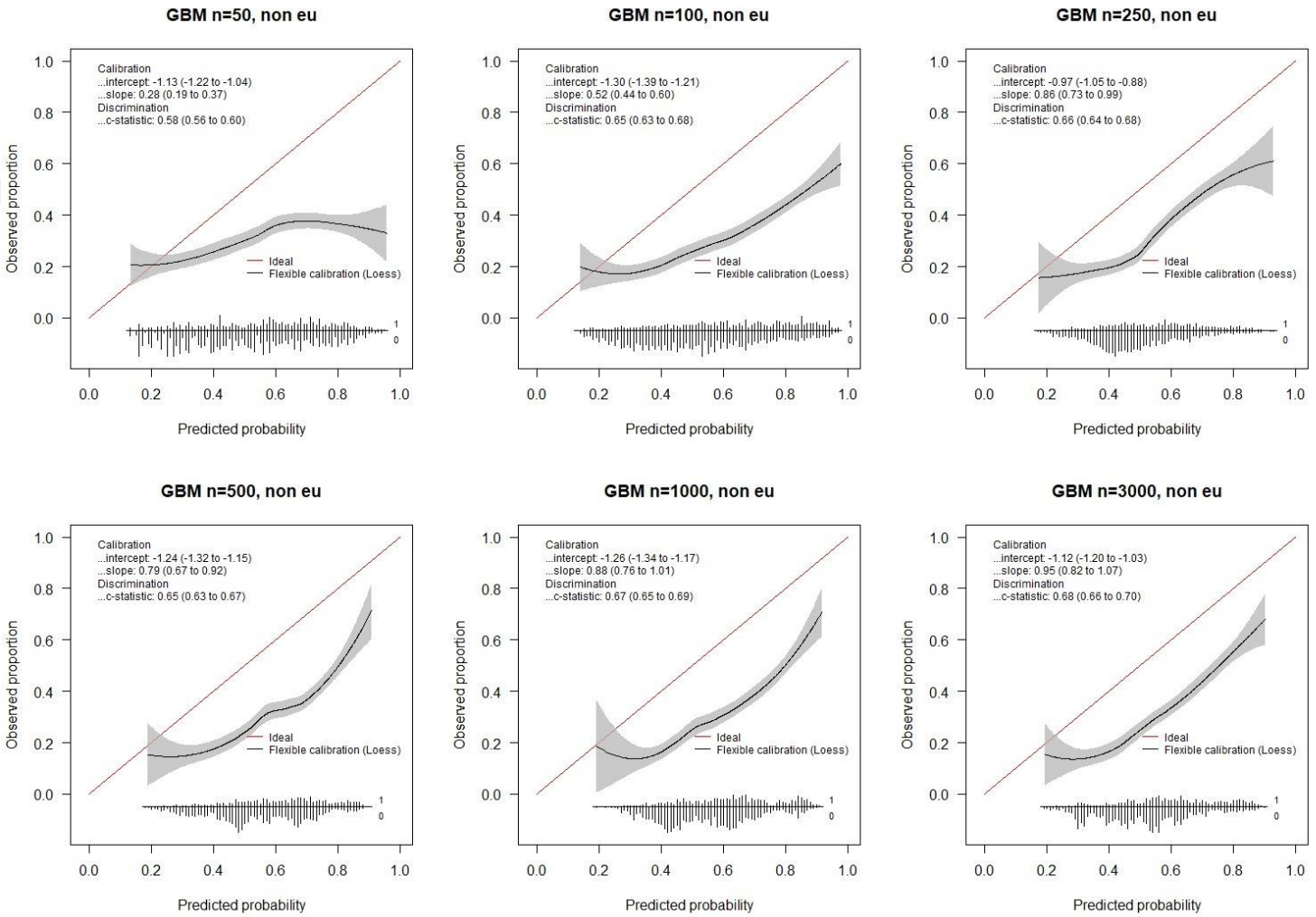


Figure B.45 Calibration plots of GBM models with 6 different sizes, non-EU data

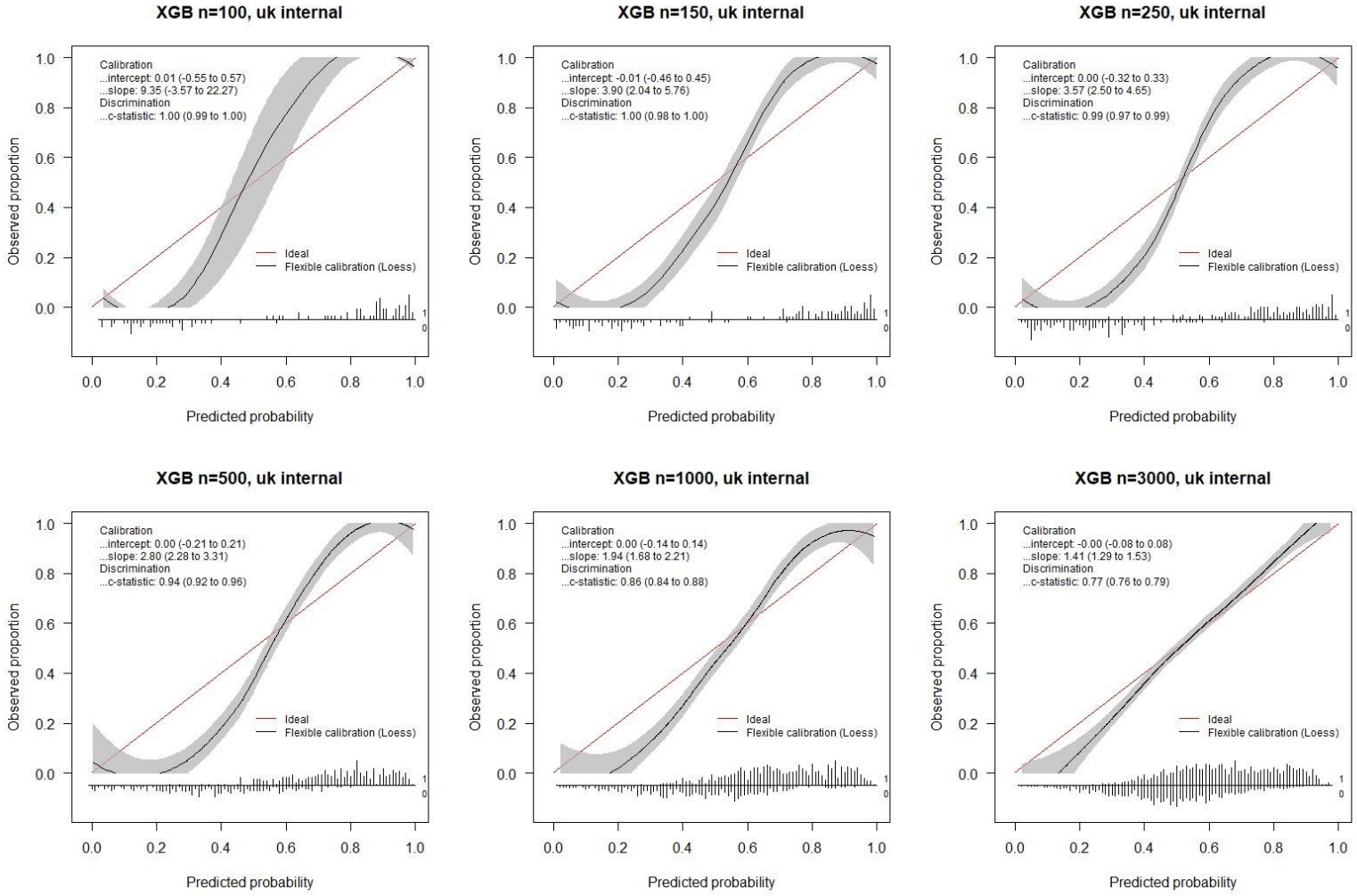


Figure B.46 Calibration plots of XGB models with 6 different sizes, UK data

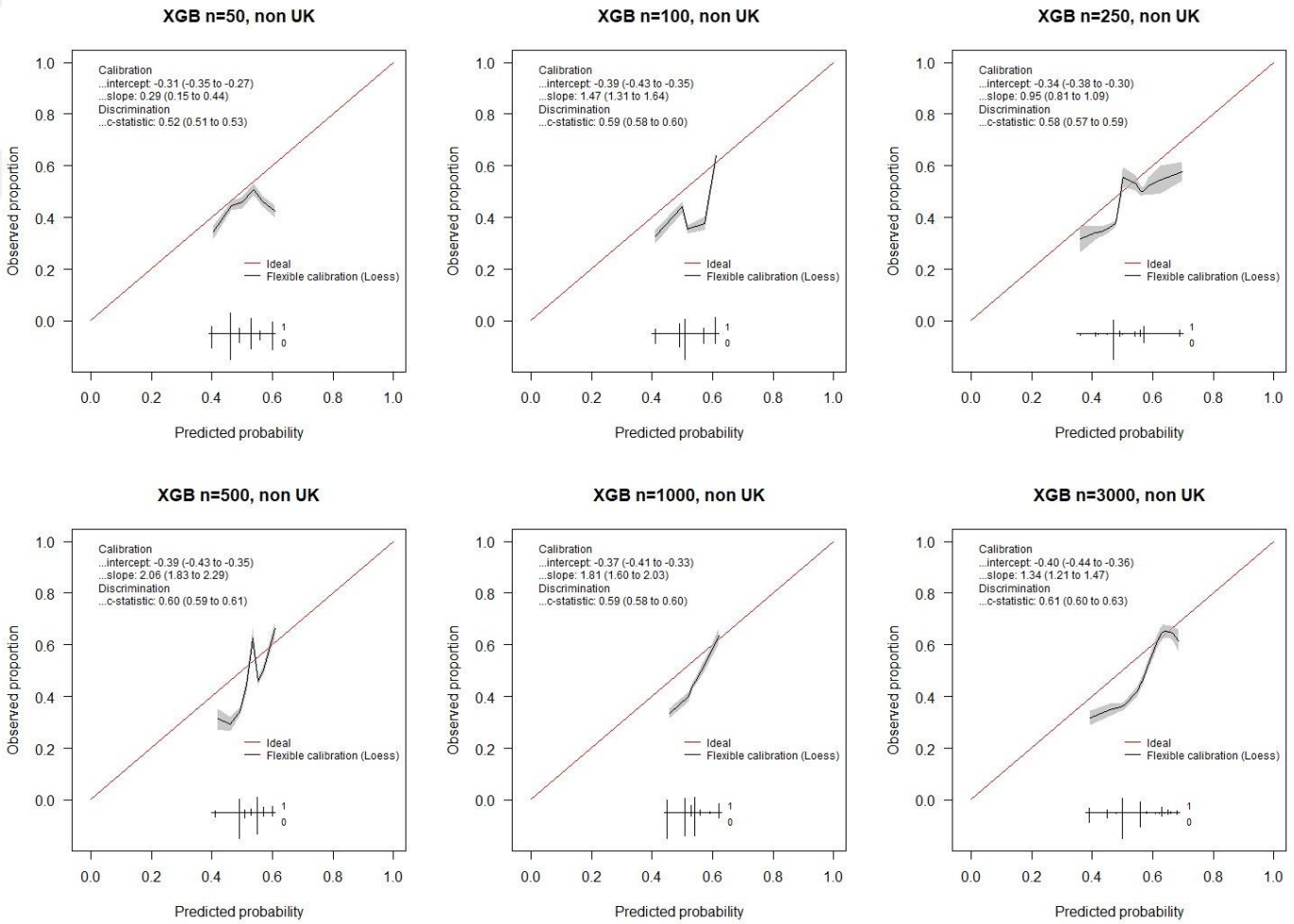


Figure B.47 Calibration plots of XGB models with 6 different sizes, non-UK data

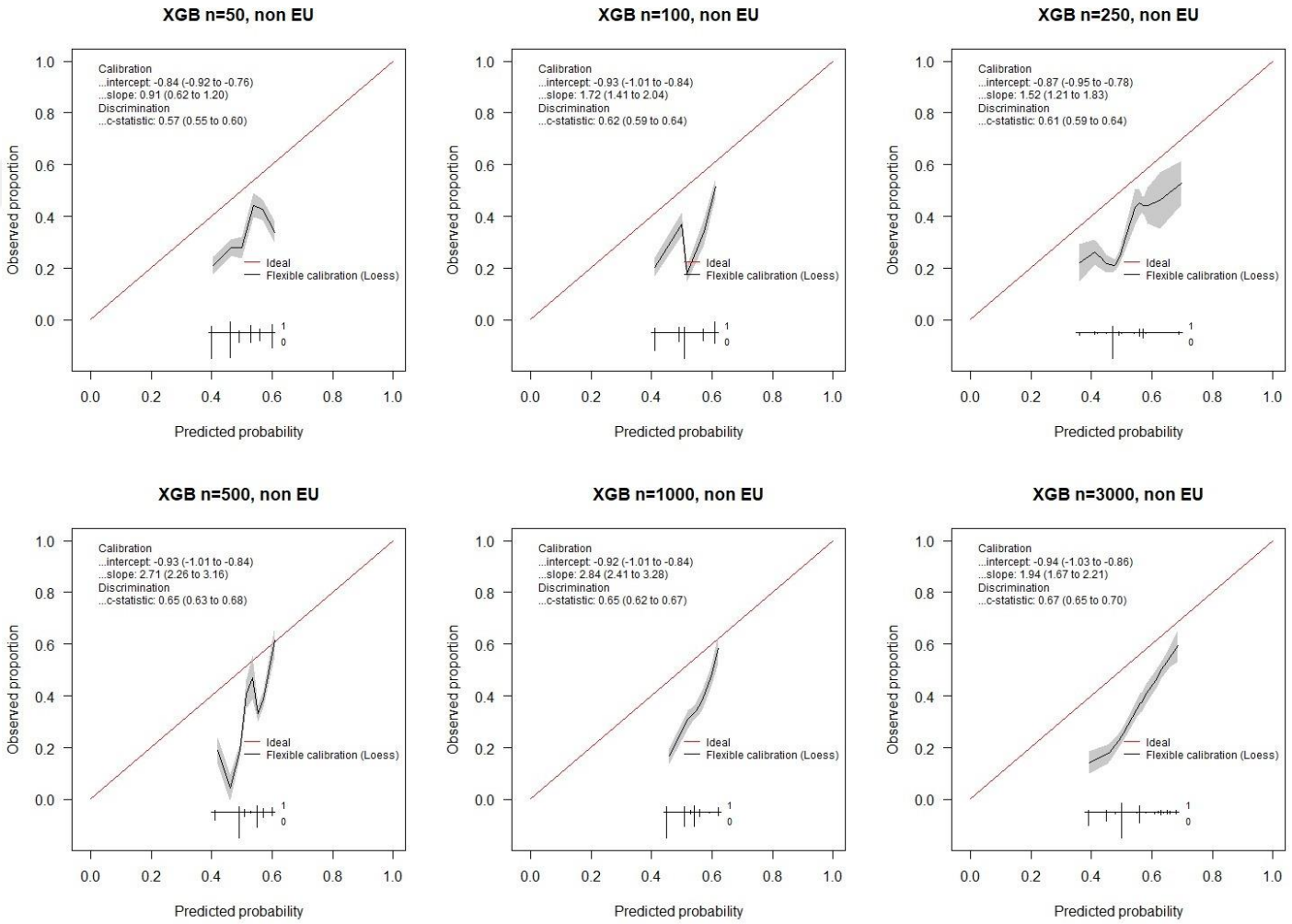


Figure B.48 Calibration plots of XGB models with 6 different sizes, non-EU data

PUBLICATIONS FROM THE THESIS

Conference Papers

1. A. U. Tosun and F. Karaman, "Prediction of Death at 6. Month on International Stroke Trial Dataset with the Comparison of Different Statistical Analyses and Machine Learning Methods," in *4th International April 23 Scientific Studies Congress*, Gaziantep, Turkey, May 10, 2022, pp. 236–248. [Online]. Available: https://www.23april.org