KU LEUVEN

FACULTY OF SCIENCE

Developing, Validating and Comparing Prediction Models for Pregnancies of Unknown Location

Nina MORGENSTERN

Supervisor: Prof. Ben Van Calster
Department of Development and Regeneration
(Woman and Child)

Master thesis submitted in fulfillment
of the requirements for the degree in
Master of Science in Statistics and Data Science

Academic year 2021-2022

Preface

The thesis focuses on developing risk prediction models for Pregnancies of Unknown location to detect potentially dangerous ectopic pregnancies. The goal of developing such risk prediction models is to provide useful decision support tools for implementation in clinical practice. I had the chance to work on this relevant topic and hope that my results will contribute to further improvements in this field. Over the last year I gained knowledge in the fields of machine learning, risk modeling and obstetrics.

First, I want to thank my my supervisor Ben Van Calster for guiding me through the process and advising me on the subject matters. I also want to thank my friends Daria, Nusret and Mario for going over my thesis and providing me with valuable feedback. Besides, I want to thank my family and my partner for supporting me during this time.

Summary

A pregnancy of unknown location (PUL) refers to a situation where the outcome of a pregnancy test is positive, but no intra- or extra-uterine pregnancy is detected on a transvaginal ultrasound (TVUS). During the follow-up of patients with PULs, three different outcomes can be identified. A PUL is an intermediary classification and might resolve into a normal (intrauterine) pregnancy (IUP) or a failed PUL (FPUL). The third possible outcome of a PUL is an ectopic pregnancy (EP), posing a high risk for complications. The detection of EP is of vital importance, since it is the main cause of maternal mortality in the first trimester. Women who are experiencing a PUL often go through a lengthy follow-up to verify the location and viability of the pregnancy. This intensive follow-up is necessary for an EP, but not for the low risk FPUL and IUP.

Given the data of a patient experiencing a PUL, risk prediction models aim to predict the probability that the PUL is an EP. The predicted probability indicates the risk of the patient having an EP and can be used for classification. 2894 PUL patients from 8 hospitals in the United Kingdom are considered to develop and validate binary (EP vs. rest) and multi-class (EP vs. IUP vs. FPUL) risk prediction models. The dataset includes missing values and has been imputed several times (multiply imputed data). Nine different algorithms are applied for binary and multi-class risk prediction models. The algorithms include logistic regression, logistic regression with transformations and interaction terms, Ridge regression, Firth logistic regression, Classification and Regression Trees, Random Forest, Extreme Gradient Boosting, Support Vector Machines and Neural Networks.

Internal-external cross-validation (IECV) is used to validate the models. The performance of the models is assessed in terms of discrimination, calibration and clinical utility. On average, multinomial models gave more accurate risk predictions and therefore were better at classifying EP than the binary models. Overall, logistic regression with the appropriate transformations seems to outperform all other approaches. This indicates that more flexible machine learning algorithms do not add a relative merit compared to the classical regression approaches for this use case.

Table of Contents

List of Abbreviations

AUC	Area under the Receiver operating characteristic curve
BhCG	Beta human chorionic gonadotropin
CART	Classification and Regression Trees
cEP	c-statistic for EP
CI	Confidence interval
cIF	c-statstic for IUP vs. FPUL
EP	Ectopic pregnancy
FP	False positive
FPUL	Failing pregnancy of unknown location
hCG	Human chorionic gonadotropin
IECV	Internal-external cross-validation
IU/L	International units per litre
IUP	Intrauterine pregnancy
LFU	Lost in follow-up
LR	Logistic Regression
MAR	Missing at Random
ML	Machine Learning
nmol/L	Nanomols per liter
PDI	Polytomous Discrimination Index
PI	Prediction interval
PPUL	Persistent pregnancy of unknown location
PUL	Pregnancy of unknown location
RBF	Radial Basis Function
RF	Random Forest
ROC	Receiver operating characteristic
SGD	Stochastic Gradient Descent
SVM	Support Vector Machines
TP	True positive
TVUS	Transvaginal ultrasound
XGBoost	Extreme Gradient Boosting

List of Symbols

a	Activation function
α	Complexity parameter
β	Coefficient
b	Bias term
C	Cost function
e	Euler's constant
ϵ	Sampling error
η	Learning rate
f	Mapping function
γ	Minimum required loss reduction
K	Kernel function
λ	Regularization parameter
$\ell(\beta)$	Log-likelihood function
$l(y_i, \hat{y}_i)$	Loss for observation $i = 1, ..., n$
L	Lagrange multiplier
μ	Common effect
ν_k	Variance of the sampling error for a study k
Ω	Complexity of a tree
ϕ	Non-linear transformation function
$\mathbb{R}$	Set of real numbers
$R_\alpha(T)$	Cost-complexity measure of a given tree T
$\sigma(z)$	Sigmoid activation function
$s(\beta)$	Score equation for a coefficient p
τ^2	Between-centre variance
θ_k	True mean for a study k
w	Weight
ξ	Slack variable
x_i	Input for observation $i = 1, ..., n$
y_i	Output for observation $i = 1, ..., n$
$\hat{y}_i$	Predictions for observation $i = 1, ..., n$
ζ_k	Difference between common effect the true mean for a study k

List of Figures

List of Tables

Introduction

1.1 Pregnancies of Unknown Location

A pregnancy of unknown location (PUL) refers to a situation where the outcome of a pregnancy test is positive, but no intra- or extra-uterine pregnancy appears on a transvaginal ultrasound (TVUS). Among women attending early pregnancy units, the reported rates of PUL vary between 5% and 42% [6]. During the follow-up of patients with PULs, four different outcomes can be identified. A PUL is an intermediary classification and might resolve into a normal (intrauterine) pregnancy (IUP). In this case the ultrasound was performed early and the intrauterine pregnancy is not yet recognized on the TVUS. The second and most common outcome is failed PUL (FPUL), meaning that a spontaneous outcome of gestation occurs with negative human chorionic gonadotropin (hCG) and the exact location of gestation is never identified [39]. FPUL is a failing pregnancy which usually does not require further treatment. hCG is a hormone which the body produces during pregnancy. Both IUP and FPUL are PULs with low risk for complications. The third form of PUL is ectopic pregnancy (EP). An EP describes the situation when a fertilized egg implants itself outside of the womb, usually in one of the fallopian tubes posing a high risk for complications. An EP can cause the fallopian tube to bust open leading to life-threatening internal bleeding and maternal death. EP is a medical emergency and should be treated right away. The fourth form of PUL is a persistent PUL (PPUL). In this rare case, the hCG does not decline spontaneously, but an abnormal increase or plateau of hCG occurs and the ultrasound does not show an intrauterine or ectopic gestation. This form is also considered high risk, as it most often turns out to be EP not visualized using ultrasound [60].

The detection of the high risk forms of PUL is of vital importance, since ectopic pregnancy is the main cause of maternal mortality in the first trimester. Despite the decline in maternal mortality in recent decades, ectopic pregnancy remains the cause for around 4.9% of all maternal deaths in developed countries [49]. An ectopic pregnancy ends in pregnancy loss. Patients with EP are treated using medication, laparoscopic surgery or abdominal surgery [28].

1.2 Risk Prediction Models

A risk prediction model is a mathematical equation that combines a number of characteristics to make predictions on the likelihood of a diagnosis of a disease. It is usually performed using multi-variable models that aim to provide reliable predictions in new patients [70]. Risk prediction models can be valuable assets to physicians and health policy-makers in order to aid decision making in clinical practice. The models provide an explicit, empirical approach to estimate probabilities of a disease and relate to the evidence-based medicine movement [46]. They calculate individualized predictions of the absolute risk for the possible outcomes. Many different statistical techniques are available such as logistic regression, linear regression or machine learning. Most often in medical literature, a binary outcome is predicted where the model gives the risk of an outcome as a probability between 0 and 1. Typically, logistic regression techniques are applied [26]. However, techniques for multiple outcomes also exist.

1.3 Related Research

Women who are experiencing a PUL often go through a lengthy follow-up to verify the location and viability of the pregnancy. The management of PULs in clinical practice differs, since standardized protocols are lacking. Oftentimes multiple blood tests measuring hCG and progesterone levels are performed and patients must undergo numerous TVUS. This intensive follow-up is necessary for an EP, but not for the low risk FPUL and IUP. FPUL and IUP represent at least 70% of occurring PULs [33]. For this group an earlier discharge or reduced follow-up can be applied, which would decrease the clinical workload and improve the patient's convenience [32].

In an early attempt to rationalize resources such that women with low-risk PULs avoid unnecessary blood tests and additional hospital visits, classification models were developed and tested by Condous et al. [15]. The logistic regression model (M1) based on hCG ratio outperformed other approaches. However, the study was based on a small sample size of 185 PULs for training and 196 cases for the testing.

In another attempt nine classification models were developed by Van Calster et al. [55]. In their study, the researchers compared prediction models using different probabilistic multi-class methods based on logistic regression, multi-layer perceptrons, least squares support vector machines and kernel logistic regression. Here too, logistic regression models showed a great performance. The developed models are able to predict the outcome of PULs well and therefore provide a useful decision support tools

to implement in clinical practice [55].

The logistic regression model M4 was developed in an attempt to improve on the performance of the previously published M1 model [16]. It uses the ratio of serum hCG level at 48h with the level at presentation, the logarithmic average of the two hCG levels and its quadratic effects to estimate the risk of a FPUL, IUP and EP. On the basis of these risks, pregnancies are classified with the aim to achieve a high detection rate (sensitivity) for the outcome. Even though the M4 model is superior to the M1 model when comparing AUCs for predicting EPs, in real terms the M4 did not result into classifying considerably more EPs correctly [16]. Due to the small sample size of this study (15 cases of EP), the implications are rather limited.

In an interventional clinical trial on 362 women, the M4 model reported sensitivity of 86% for FPUL, 86% for IUP and 73% for EP, meaning that 27% of patients with an EP would be chosen for early discharge or reduced follow-up [32, 58]. In a later multi-centre diagnostic accuracy study of 1962 patients by Van Calster et al. [58] the M4 model was externally validated and classified 88% of ectopic pregnancies as high risk based on a risk threshold of 5%.

In a recent approach by Van Calster et al. [60] the M6P and M6NP models are developed to improve the M4 model. These models are based on multinomial logistic regression. They include the predictors hCG level at presentation, the 2-day hCG ratio and the progesterone level at presentation. Progesterone is an optional predictor. The M6P model includes progesterone, while M6NP excludes this predictor [60]. The models show high sensitivity and classify PULs efficiently. The multi-center validation study by Christodoulou et al. [13] suggests that M6P is the best prediction model for PULs and that even without progesterone as a predictor, M6NP is a clear improvement over M4. Additionally, to further rationalize the management of PULs the researchers developed a two-step triage strategy (2ST), which incorporates the M6P model. M6P and 2ST showed the best clinical utility and good overall calibration [13].

1.4 Machine Learning in Medicine

Machine learning (ML) is a sub-field of artificial intelligence, which studies artificial learning behavior and automates analytical model building for data analysis. It is based on the idea that systems can learn from data, find patterns and make decisions with minimal human intervention. The ability to learn describes the capacity of a system to improve its own performance solving certain problems after receiving additional information about the problem [3]. There are discussions around definitions and

clear demarcations of regression and ML techniques as described by Breiman [8]. A meaningful definition that will be used in this work is that ML concentrates on models which directly and automatically learn from data. Conversely, regression models are based on theory and assumptions, and benefit from human intervention and subject knowledge for model specification [37]. ML techniques have been applied successfully in diverse fields ranging from computer vision, pattern recognition, traffic prediction, finance and computational biology to biomedical and medical applications as well as countless other areas [20]. Artificial neural networks, support vector machines, and random forests fall under the hood of ML techniques. ML techniques have become increasingly popular in medical research as an alternative approach for prediction and classification problems. The growing availability of large data sets such as electronic health records opened up opportunities for the use of ML models in medicine, since ML models typically require large datasets [12].

ML models are data-driven and do not rely on a pre-specified structure (formula) unlike standard regression techniques. This means that ML can produce highly flexible models and utilize advanced mathematical techniques to build complex models [12]. Since ML models learn their structure directly from the data, non-linearity and interaction effects are captured automatically and do not have to be explicitly specified as in standard regression techniques which can lead to an increased predictive accuracy and performance benefits [24]. Due to the data driven approach, ML models can identify specific trends and patterns that researchers might miss when specifying a standard regression model.

However, ML also faces major criticism in the science community for the lack of new knowledge and understanding arising from their use. This is due to the fact that ML models are often treated as black-boxes, since they are too complex for humans to understand [44]. Traditional statistical modeling is based on the belief that the purpose of science is to crack open black boxes in order to better understand the underlying processes. ML arose from a group of young computer scientists, engineers, and statisticians who were looking for a practical solution to a problem. They prioritize accuracy and precision over interpretability [8]. Yet, accountability and interpretability are of key importance in medical research. Ethical issues and feedback loops of ML models are serious problems when using ML in medical settings [51]. In addition, complex ML models are prone to overfit if built without a proper validation procedure. The failure of the model to generalize on unseen data can lead to nonsensical predictions with potentially dangerous consequences in medical settings.

Lastly, research suggests that ML models require more data than e.g. logistic

regression (LR). ML models are "data hungry" in a sense that they need a higher number of events per variable to attain a stable AUC than LR [63]. Besides, in a systematic review comparing ML and LR for clinical prediction models, Christodoulou et al. [12] finds that there is no evidence that ML techniques display a superior performance over LR, suggesting that the claims of improved performance in many studies are due to poor methodology. Researchers, investigating the discrimination and calibration performance of LR and ML algorithms on predicting the outcome of traumatic brain injury, found similar results [27]. The authors emphasized the lack of calibration evaluation and the focus on model AUC [27].

1.5 Research Questions

The thesis is focused on developing risk prediction models using standard regression models, penalized regression models and machine learning algorithms. Given a PUL, the models aim to predict the occurrence of EP, IUP or FPUL and provide an indication of the corresponding risks by providing a probability. A large number of patients from multiple hospitals in the United Kingdom are considered to develop and validate binary and multi-class risk prediction models. The research questions are as follows:

- How do different types of algorithms compare in terms of model performance?
- Do flexible machine learning algorithms add relative merit?
- Do multinomial models lead to more accurate risk predictions for ectopic pregnancies than binary models?

Methodology

2.1 Data Collection

The data was collected between January 2015 and January 2017 at eight UK hospitals: Chelsea and Westminster, Hillingdon, North Middlesex, Queen Charlotte's and Chelsea, Royal Surrey, St Mary's, West Middlesex and Wexham Park. The version of the provided dataset is from May 2019. The study recruited patients classified as having a PUL after their first transvaginal ultrasound and who were suitable for outpatient management. Patients were excluded if their pregnancy did not meet the criteria for PUL, they were hemodynamically unstable or could not safely be managed as an outpatient due to complications such as moderate to severe pelvic pain or hemoperitoneum on transvaginal ultrasound [13].

Women with an initial hCG of less than 25 IU/L were excluded from prediction modeling because this is the level at which most pregnancy tests will return a negative result [13]. The study enrolled 3272 women, however 6 of them satisfied an exclusion criteria, and another 367 had an initial hCG level of less than 25 IU/L. Patients under the age of 16 are also excluded due to additional ethical restrictions (5 cases). As a result, the dataset used in this thesis contains 2894 PULs. In Table 1 the data per center and corresponding distribution of PUL outcomes are shown.

Data cleaning was performed by biostatisticians and ultrasound examiners. The process included checking for inconsistencies and outliers, and sending inquiries to centres to retrieve missing information or to correct errors [13]. Patients were followed up until the PUL outcome was known. The PUL outcome was divided into four categories: FPUL, IUP, EP, and PPUL. EP and PPUL were integrated into a single group, as done in many previous studies, so that the models predict the probability of three outcomes: FPUL, IUP, and EP/PPUL.

Table 1: Sample size per center

Center	N	FPUL	IUP	EP/PPUL	LFU
St. Mary's	507	194 (38%)	176 (35%)	72 (14%)	65 (13%)
Hillingdon	472	229 (49%)	136 (29%)	44 (9%)	63 (13%)
Chelsea and Westminster	470	224 (48%)	163 (35%)	54 (11%)	29 (6%)
Queen Charlotte's and Chelsea	443	176 (40%)	167 (32%)	56 (13%)	44 (10%)
Wexham Park	334	129 (39%)	108 (32%)	44 (13%)	53 (16%)
West Middlesex	297	208 (70%)	63 (21%)	16 (5%)	10 (3%)
Royal Surrey	186	83 (45%)	60 (32%)	21 (11%)	22 (12%)
North Middlesex	185	92 (50%)	56 (30%)	27 (15%)	10 (5%)
All centers	2894	1335 (46%)	929 (32%)	334 (12%)	296 (10%)

FPUL, failed PUL; IUP, intra-uterine pregnancy; EP, ectopic pregnancy; PPUL, persisting PUL; LFU, lost to follow-up.

2.2 Missing Data

Missing values were detected for the initial progesterone, the second beta human chorionic gonadotropin (BhCG) level, pain score, vaginal bleeding, history of EP and the final PUL outcome, as some patients were lost to the follow up before a PUL outcome could be determined [13]. Frequency and distribution of missing values is visualized in Figure 1. The missing data has been assumed to be 'missing at random' (MAR), meaning that the missingness can be fully accounted for by the observed variables in the dataset. The dataset has been imputed several times (multiply imputed data) by Christodoulou et al. [13]. The R package `mice` was used to create multiple imputations for multivariate missing data by chained equations [54, 41]. The method is based on fully conditional specification, in which each incomplete variable is imputed by a separate model [54] . Christodoulou et al. [13] generated 100 imputed datasets with this method, out of which 10 will be used in this thesis. The final dataset used for analysis has therefore 28,940 entries. For the regression models, the coefficients are combined across the imputed data sets using Rubin's rule for pooling parameter estimates.

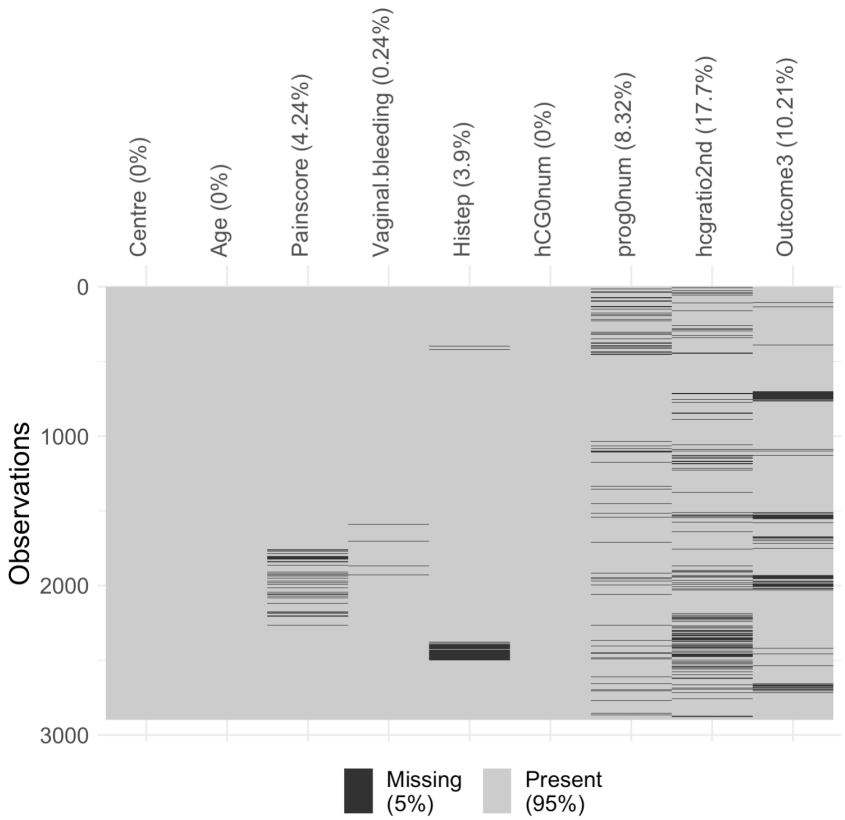


Figure 1: Missing data in original dataset

2.3 Variable Selection

The selected predictors for model building can be classified in two categories: three biomarkers and four clinical variables. The biomarker variables include the initial hCG level, which is a continuous variable measured in IU/L. The initial progesterone level is a continuous variable measured in nmol/L. Additionally, the hCG ratio variable is selected, which corresponds to the ratio between the BhCG at 48 hours after first presentation and the BhCG at first presentation [13]. The clinical variables include the maternal age in years along with the presence of vaginal bleeding as an ordinal variable with 5 levels. Moreover, a binary variable indicating if the patient has a history of ectopic pregnancies and an ordinal pain score with 11 levels are selected. Transformation and interactions between the selected variables will be used in the modeling stage.

2.4 Algorithms

Nine different algorithms are fitted for binary and multi-class risk prediction models. The algorithms are logistic regression, logistic regression with transformations and

interaction terms, Ridge regression, Firth logistic regression, Classification and Regression Trees (CART), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machines (SVM) and Neural networks (NN). The algorithms are supervised, which means that they require labeled data. The learning objective is to optimize the parameters for minimizing the error [30].

Given a training set of the form

$$\{(x_i, y_i)\}_{i=1,...,n},\ x_i \in \mathbb{R}\ and\ y_i \in \mathbb{R}, \tag{1}$$

where x_i are the inputs and y_i are the outputs. The mapping function

$$f : X \to Y \tag{2}$$

is learned from the training set and will be used to generate risk predictions for the different outcomes on unseen data.

2.4.1 Logistic Regression

Logistic regression is a classical and widely applied technique that uses the logistic function to model a binary dependent variable. Extensions exist to model multinomial problems. The logistic function is given by [1]

$$logit[P(Y = 1|\mathbf{X})] = \eta(\mathbf{X}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p, \tag{3}$$

where $logit[P(Y = 1|\mathbf{X})]$ represents the log odds of the outcome $Y = 1$, β_0 is the intercept and β_p are the coefficients of the corresponding predictors x_p. The function is estimated using the maximum likelihood estimation. For a sample $y_1, ..., y_n$ with design matrix $\mathbf{X}$ the log-likelihood equation to maximize is given by [1]

$$\ell(\beta|\mathbf{X}) = \sum_{i=1}^{N} [y_i \log(p_i(\beta)) + (n_i - y_i) \log(1 - p_i(\beta))]. \tag{4}$$

Next, the log-likelihood is differentiated and the resulting score equations are set to zero to find the parameters. To get probabilities between 0 and 1 for the outcomes, the following logistic function is used [1]

$$P(Y = 1|\mathbf{X}) = \frac{e^{\eta(\mathbf{X})}}{1 + e^{\eta(\mathbf{X})}}. \tag{5}$$

One advantage of logistic regression is the interpretability of its coefficients. The coefficient β_p associated with predictor X_p is the expected change in log odds of having the outcome per unit change in X_p while holding all other predictors constant. Increasing X_p by one unit multiplies the odds of having the outcome by e^{β_p}. In addition, logistic regression makes no assumptions about distributions of classes in the feature space [2]. It is easy to implement and efficient to train. The major disadvantage of logistic regression is the assumption of additivity and linearity between the logit of the outcome and the included terms, as it constructs linear decision boundaries [53, 14]. The terms may consist of transformations and interactions of the input variables. For the first model that will be considered, a simplistic approach will be taken by ignoring non-linear and non-additive effects.

2.4.2 Logistic Regression with transformations and interactions

The drawback of logistic regression is that interactions have to be added manually and are not extracted automatically from the data. Adding interaction terms and using variable transformations can potentially improve performance of the logistic regression model as more complex relationships are captured. The transformations and interaction terms have to be formally specified in the modeling stage and are based on the suggestion of previous studies. Using this prior knowledge leads to a more sensible approach compared to the LR model without non-linear and non-additive effects. The considered transformations include logarithmic transformations for the biomarkers as it has been suggested by previous research to be beneficial for modeling [55, 60]. For the maternal age a quadratic term will be considered, which was proposed by Van Calster et al. [55]. For interaction terms, the interaction between hCG ratio and progesterone level will be reviewed as this parameter was included in the M6P model [60].

2.4.3 Ridge Logistic Regression

Ridge regression is a regularization technique that is used to prevent overfitting and dealing with multicollinearity by regularizing the estimated coefficients. It is sometimes called L2 regularization as it adds an L2 penalty when maximizing the log-likelihood function. The L2 penalty is equal to the square of the magnitude of the coefficients. All coefficients are shrunk by the same factor (λ). The factor is used to control the penalty term. Ridge regression reduces the complexity of a model but not the number of variables because it never leads to a zero coefficients unlike L1 (Lasso)

regularization [9]. The log-likelihood function becomes

$$\ell_R(\beta) = \ell(\beta) - \lambda \sum_{p=1}^{p} \beta_p^2. \tag{6}$$

When λ = 0 the solution will be the ordinary maximum likelihood estimation, whereas when $\lambda \to \infty$ the parameters tend to zero [9]. λ will be chosen on the basis of hyper-parameter tuning (see section 2.5).

2.4.4 Firth Logistic Regression

Firth logistic regression's primary idea is to offer a more effective score function by adding a term that counteracts the first-order term from the asymptotic expansion of the bias of the maximum likelihood estimation [22]. This term will go to zero as the sample size increases [22, 68]. This method is attractive because it reduces bias for small sample sizes while also producing finite and consistent estimates, even in case of (complete) separation [68]. The penalized log-likelihood function becomes

$$\ell^*(\beta) = \ell(\beta) + \frac{1}{2} log\ |\mathbf{I}(\beta)|, \tag{7}$$

where $\mathbf{I}(\beta)$ is the Fisher information matrix evaluated at β [1]. The penalized log-likelihood is related to the basing estimation on modified score equations

$$s^*(\beta_p) = s(\beta_p) + \frac{1}{2} tr\left[\mathbf{I}^{-1}(\beta) \frac{\partial \ell(\beta)}{\partial \beta_p}\right] = 0 \tag{8}$$

where $s(\beta_p)$ are the score equations of logistic regression [1].

2.4.5 Classification and Regression Trees

Decision tree learning is one of the most popular methods in machine learning and data mining. Decision trees combine good predictive accuracy with a high interpretability and efficient learning [3]. Additionally, decisions trees are generally considered insensitive to outliers. The goal is to divide (split) the data into homogeneous groups by using a splitting criteria. An illustration of a decision tree is given in Figure 2.

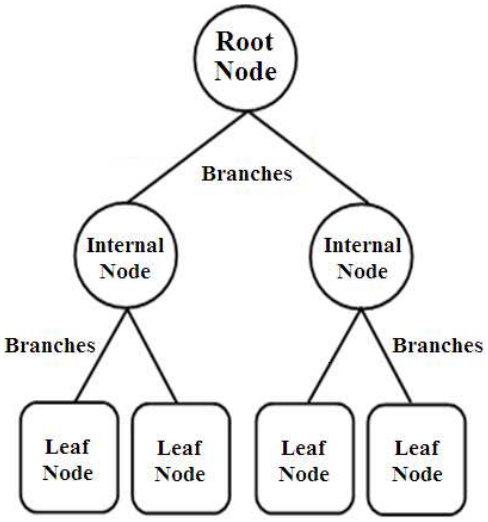


Figure 2: Structure of a decision tree [47]

The algorithm of the R package `rpart` works by splitting the dataset recursively, which means that the arising subsets from previous splits are further split until a termination criterion is reached [41, 50]. Each split is carried out on the basis of a predictor variable that results into the largest possible reduction of heterogeneity of the outcome variable [3]. To determine the best split for categorical outcomes, rpart offers the Gini-index (default) or entropy as choices. The Gini-index lies between values 0 and 1, where 0 expresses the purity of classification, meaning all the elements belong one class and 1 indicates a random distribution of elements across the classes. When determining the best split in a decision tree, the Gini-index is calculated for all possible splits. The split with the lowest Gini-index is carried out since it indicates the biggest reduction in class-heterogeneity of the node leaf. For a set of items with K classes and p_k being the fraction of items labeled with class $k \in 1, 2, ..., K$ Gini-index and the entropy are given by [3]

$$Gini = \sum_{k=1}^{K} p_k(1 - p_k) \tag{9}$$

$$Entropy = -\sum_{k=1}^{K} p_k \log p_k. \tag{10}$$

The algorithm is terminated if the subsets reach a minimal size or no improvements can be made. In the second stage, the complexity of the tree is reduced to prevent overfitting by using cross-validation to trim back the full tree. This is called cost-complexity pruning, where α is the complexity parameter, a hyperparameter which requires tuning. The complexity parameter is needed to define the cost-complexity measure $R_\alpha(T)$

of a given tree T

$$R_\alpha(T) = R(T) + \alpha|T| \tag{11}$$

where $|T|$ is the number of terminal nodes in T and $R(T)$ is traditionally defined as the total miss-classification rate of the terminal nodes. The cost-complexity measure is optimized to carry out the pruning procedure. The predictions from the trained and pruned tree are given by a matrix whose columns give the probabilities of belonging to the distinct classes for each observation [50].

2.4.6 Random Forest

Random Forest is an ensemble technique that is based on bootstrap aggregating. The idea of bootstrap aggregating is to create several variants of the original training data set by drawing randomly n elements from the training data set without replacement. Each variant might lead to a different model. This works especially well with decisions trees as they are considered unstable learners [3]. The mapping function f is given by collection of tree classifiers $h_1, ..., h_j$ where j is the number of trees [72]. The trees are grown with recursive binary splitting, the same way as classification trees. In addition, randomness is introduced into the procedure by randomly sampling only a subset of all variables at each split. The best split on a certain node is chosen from the sampled subset only. This increases the variability among the induced decision trees. Conventionally, the trees stay unpruned to ensure low bias. Next, the trees are averaged which leads to a lower variance. A single tree predicts the class of an observation by applying the tree's splitting rules. The observation is allocated into a leaf node and assigned to the class which the majority of observations in the leaf node are corresponding to. The predictions of all trees are combined by majority voting [7, 72]

$$f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, I(y = h_j(x)). \tag{12}$$

An instance is classified into the class for which most trees voted for and the class probabilities are given by the proportion of trees that voted for a certain class. The number of trees to grow is a hyperparameter that can be tuned, but generally can be set to a large, computationally feasible number. The size of the randomly sampled subset of variables is another hyperparameter which can be tuned.

The advantage of the random forest algorithm is that it runs efficiently on large volumes of data and it can handle thousands of input variables. It gives an unbiased estimate of the generalization error as the forest building progresses [7]. Due to the tree-based structure, interactions do not have to be specified but are included implicitly.

The main disadvantage is the loss of interpretability as hundreds or even thousands of trees are grown which creates a black-box. However, interpretability techniques like partial dependence and individual conditional expectation plots for random forest exist.

2.4.7 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a decision tree ensemble technique based on boosting. Here, trees are added sequentially to the ensemble and fit to correct the prediction errors made by prior models. Gradient boosting employs gradient descent algorithm to minimize errors in sequential models. XGBoost is an optimized version, which uses parallel processing, tree pruning and regularization [10]. XGBoost has become a very popular algorithm due to its high predictive performance, often outperforming other ML algorithms in competitions [3]. The downsides are that many hyperparameters need to be tuned and the loss of interpretability make XGBoost a black-box model.

A version of the algorithm is implemented in the `xgboost` package in R [41, 11]. The prediction value at step t is written as $\hat{y}_i^{(t)}$. The objective function Q used during training is defined as

$$Q = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i), \tag{13}$$

and consists of two parts: training loss l and regularization term Ω [10]. The training is performed in an additive manner: fix what is learned and add a new tree at each iteration. In the general case, the Taylor expansion of the loss function up to the second order is taken [10]. The loss function is therefore given by

$$\sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t), \tag{14}$$

where g_i and h_i are defined as

$$\begin{aligned} g_i &= \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}), \\ h_i &= \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}). \end{aligned} \tag{15}$$

The complexity of the tree (regularization term) is given by

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2, \tag{16}$$

where w is the vector scores on leaves, T is the number of leaves, λ is a L2 regu-

larization term on the weights and γ specifies the minimum loss reduction required to make a split [10] . The formulation of the objective function can be further compressed by defining $I_j = \{i|q(x_i) = j\}$ as the set of indices of data points assigned to the j-th leaf, $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$ [10]. This leads to the objective function being defined as

$$Q^{(t)} = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T. \tag{17}$$

Next, the optimal weight for leaf j and the best objective reduction can be calculated as [10]

$$\begin{aligned} w_j^* &= -\frac{G_j}{H_j + \lambda}, \\ Q^* &= -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T. \end{aligned} \tag{18}$$

The best objective reduction is a measure of how good the tree structure is [10]. Since it is infeasible to enumerate over all possible trees to pick the best one, one level of the tree at a time is optimized by using the score gains [10]

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma. \tag{19}$$

The gain can be decomposed as the score on the new left leaf, the score on the new right leaf, the score on the original leaf and the regularization on the additional leaf. The predictions produced by XGBoost are probabilities for each observation to belong to a class.

2.4.8 Support Vector Machines

The Support Vector Machine (SVM) is a linear model that can be used to solve classification and regression tasks. It can solve both linear and nonlinear problems and is useful for a wide range of applications. The method divides the data into classes by drawing a hyperplane. The points closest to the hyperplane are called support vectors [3]. The distance between the points and a hyperplane is the margin. The logic behind SVM is to maximize the margin in order to find the optimal hyperplane [17]. As a result, SVM seeks to create a decision boundary with as much separation between the two classes as possible. This approach is illustrated in Figure 3. SVM has the advantages that it performs well in high dimensions and when the data is linearly separable. Disadvantages include slow processing of large datasets and poor performance when classes are overlapping [67].

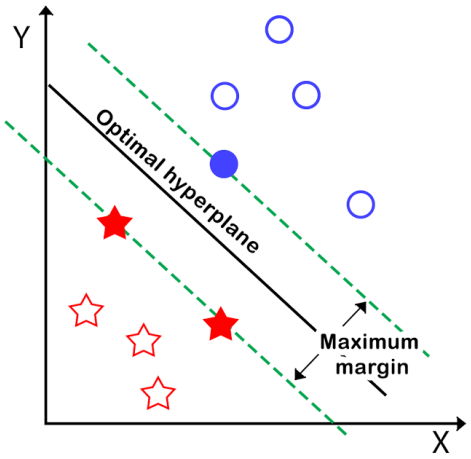


Figure 3: Support Vector Machine [18]

To find the optimal decision boundary the term $\frac{1}{2}||\mathbf{w}||^2$ is minimized under the constraint of $y_i(\mathbf{w}^T \cdot \mathbf{x_i} + b) \geq 1$, where $\mathbf{w}$ is the vector of weights and b is the bias term. This results into a Lagrange multiplier equation which needs to be optimized [52]

$$L = \frac{1}{2}||w||^2 - \sum_i \alpha_i(y_i(\mathbf{w}^T \cdot \mathbf{x_i} + b) - 1). \tag{20}$$

Solving the above Lagrangian optimization lead to w, b and α parameters that uniquely determine the maximum margin solution. However, in the case that the data is not linearly separable, a so called "soft margin" hyperplane is used, where C is an added regularization parameter which controls the importance of the slack variable ξ. C is a hyperparameter, which requires tuning. The slack variable ξ_i determines to which extent the constraint of being on the right side of the margin can be violated for every observation [52]. Besides, when the data is not linearly separable, a mapping function $\phi(x)$ is used to transform the data and map it to a space of higher dimension, where a linear decision boundary can be constructed [3]. The kernel trick can be applied, which converts the dot product of a support vector to the dot product of the mapping function, making the construction into higher dimensions unnecessary. Any function K for which a transformation ϕ exists such that $K(x_i, x_j) = \phi(x_i)\phi(x_j)$ can be called a kernel function [3]. Examples of kernel functions are linear, polynomial, sigmoid or radial basis function (RBF) kernels. The kernel function has to be chosen in advance, or can also be considered a hyperparameter to be tuned. In this thesis, a

Gaussian kernel will be used. The term to be minimized becomes [52]

$$\begin{aligned} \min_{b,w,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to } \xi_i \geq 0, y_i(w^T \cdot \phi(x_i) + b) \geq 1 - \xi_i \; \forall_i. \end{aligned} \tag{21}$$

Hence, the Lagrangian (Wolfe) dual objective function is obtained

$$\begin{aligned} L_D = \sum_{i}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j), \\ \text{subject to } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^{N}\alpha_i y_i = 0. \end{aligned} \tag{22}$$

Together with the Karush-Kuhn-Tucker conditions, the equations uniquely characterize the solution to the dual problem [52]. Since SVMs do not output probabilities, probability calibration methods have to be used to get the class probabilities. Platt scaling is the most common method which produces probability estimates for the form [4]

$$P(y = 1 \mid x) = \frac{1}{1 + \exp(Af(x) + B)}, \tag{23}$$

which corresponds to a logistic transformation of the classifier scores $f(x)$, where A and B are two scalar parameters that are learned by the algorithm. Extensions for multi-class problems exist. For this purpose, binary SVM will be performed for each pair of outcomes and the resulting probabilities will be combined with pairwise coupling.

2.4.9 Neural Networks

Artificial neural networks are a general class of nonlinear models which are composed of a number of neurons organized in input, hidden and output layers. The layers are connected to the following layers by weights (Figure 4). It has been proven that neural networks such as multi-layer perceptrons are universal approximators [3]. They are able to approximate any continuous nonlinear function arbitrarily well on a given interval. Neural networks typically have may inputs and outputs, which makes them attractive for modeling multi-variable system. Critical design issues when building a neural network are the number of neurons, layers, training algorithms, activation functions and dealing with generalization, noise and local minima solutions [3].

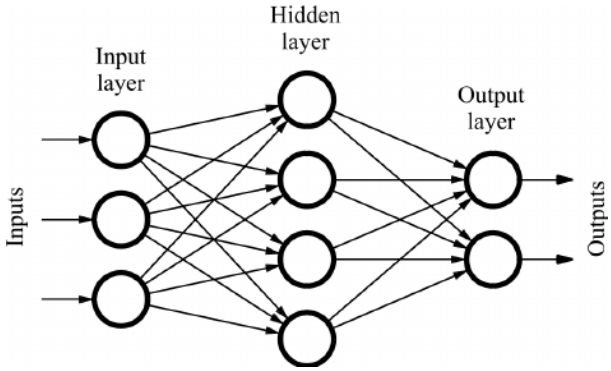


Figure 4: Multi-layer perceptron with one hidden layer [29]

Neural networks are tremendously successful at predictive tasks with large volumes of data. They are flexible and can be applied to structured and unstructured data. Complex relationships and interactions are modeled automatically and require less feature engineering than traditional models. However, neural networks require large amounts of training data and are computationally expensive and time consuming to train. They are often referred to as black-boxes, since the model does not explain the influence of the independent variables on the outcomes. In addition, overfitting and generalization need to be carefully evaluated. The weights of a neural network are initialized to small random numbers. The input to a neuron in a fully connected hidden layer is given by $z = \sum_{i=1}^{m} w_i x_i + b$, where m is the number of features in the input matrix X. Next, each neuron has an activation function such as the sigmoid function [31]

$$a = \sigma(z) = \frac{1}{1 + e^{-z}}. \tag{24}$$

For binary or multi-class classification problems the output layer consists of a sigmoid or softmax activation function, which takes the transformed hidden layer outputs as inputs to predict the class. Different loss functions for optimization exist. For classification task a common loss function is the cross-entropy, which is equivalent to the log-likelihood [52]

$$L_{CE} = -\sum_{i=1}^{\mathrm{N}_c} y_i \log(p_i), \tag{25}$$

where y_i is the true label, p_i is the probability for the i^th class and N_c is the number of classes. An algorithm such as gradient descent has to be selected for training. The cost function corresponds to the average loss over the entire training dataset. Given the results of the cost function for the first initialization of the network (forward pass), the optimization procedure begins by using the backpropagation algorithm to update the weights and biases of the network (backwards pass). The partial derivatives of the cost function are obtained by applying the chain rule and allow to compute the

relationship between components of the neural network and the cost function. The partial derivative for the weights of the output layer are given by [3]

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}}, \tag{26}$$

where C is the cost function, L is the last layer, a is the activation function. Each partial derivative is saved in a gradient vector. The gradient is computed in mini-batches by sub-sampling the observations into batches and averaging the output for each weight and bias. After each mini-batch, the weights and bias terms are updated by a certain amount in each layer [3]

$$w^{(l)} = w^{(l)} - \eta \cdot \frac{\partial C}{\partial w^{(l)}}, \tag{27}$$

where η corresponds to the learning rate. This procedure repeated for every weight and bias in the network. The number of times the forward and backwards pass are completed is given by the chosen number of epochs. To prevent overfitting, early stopping, weight decay and dropout are common measures to take. To avoid the problem of sub-optimal local minima solutions, the stochastic gradient descent algorithm (SGD) can be used for training [52].

2.5 Hyperparameter Tuning

The algorithms presented in section 2.4 utilize hyperparameters to build flexible methods and to optimize the performance of the procedures. A hyperparameter in machine learning is a parameter whose value is used to influence the learning process of an algorithm. Hyperparameters have to be set before the start of the training process and cannot be learned directly from the training data [40]. They require tuning as no standard analytical formula can be used to derive them. R methods provide default values. However, the hyperparameters should be tailored to the dataset used, since they can influence the predictive performance of the model [69]. The validation of hyperparameters will be performed using 10-fold cross-validation for each imputed dataset. Optimizing hyperparameters can be carried out using different search strategies such as grid search, random search or the more complex Bayesian optimization. Tuning is generally computationally expensive and becomes more expensive if the search space increases [69]. Choosing adequate hyperparameters is crucial because of their influence on convergence and model performance.

For ridge regression, the hyperparameter λ requires tuning. For SVM, σ of the Gaussian kernel and the regularization parameter need tuning. For random forest, the number of variables randomly sampled as candidates at each split and the minimal

node size will be tuned. For CART complexity parameter will be tuned. For the neural network the number of hidden neurons and the decay will be tuned. XGBoost will be tuned with the learning rate η, which controls the weighting of new trees added to the model and the booster method, which sets the type of learner (either a tree or a linear function).

2.6 Evaluating the Performance

2.6.1 Discrimination

The performance of the models will be validated in terms of discrimination, calibration and clinical utility. Discrimination refers to the model's capacity to distinguish between patients who have an EP and those who do not. For binary outcomes (EP vs. rest) the area under the receiver operating curve (AUC) can be used for model evaluation. The receiver operating curve (ROC) is a graph illustrating sensitivity (true positive rate) against specificity (false positive rate) for several cutoffs for the probability of an outcome. The AUC is threshold independent and can be interpreted as the probability that a randomly selected patient with an EP is given a higher probability of having EP by the model than a randomly selected patient without an EP [46]. The AUC can lay between 0 and 1, where 0.5 is an uninformative model and 1 is a perfect model. An AUC < 0.5 indicates that the classifier is systematically predicting the wrong class, which can be reversed by changing the class labeling. Next to the AUC, sensitivity, specificity, positive predicted value and negative predicted value will be assessed. For binary outcomes, the AUC is equal to the c-index which is a rank order statistic for predictions against true outcome. It is insensitive to errors in calibration such as differences in average outcome [46].

For the multi-class prediction of EP vs. IUP vs. FPUL, several extensions to the c-index exist to measure the discriminative ability of a model. The polytomous discrimination index (PDI) is a set approach which can be interpreted as correctly identifying a patient from a randomly selected category within a set of patients [57]. It ranges from 0 to 1 where $1/k$ is a uninformative model, with k as the number of categories. It is an elegant measure to assess the overall discriminative ability because it is more strongly influenced by simultaneous discrimination between all categories than by partial discrimination [56]. If the PDI suggests that the model has relevant discriminative ability, pairwise c-indexes for each pair of outcome categories will be evaluated. For pairwise c-indexes the conditional-risk method will be used, which is consistent with the analytical approach of the multinomial logistic regression used to

develop polytomous risk models [57].

2.6.2 Calibration

Calibration evaluates how correctly the model's predictions match the overall observed event rates [38]. Calibration is an important metric to evaluate because it gives an indication whether the risk estimates are accurate. The reliability of risk estimates is important for clinical decision making and informing patients about their risks. In practice calibration is rarely assessed as a systematic review by Christodoulou et al. [12] showed. This is problematic since poor calibration can make predictions misleading [62]. Calibration has been referred to as the 'Achilles heel' of predictive analytics [12].

A weak calibrated model will, on average, not over- or underestimate risk and does not give overly extreme/modest risk estimates. It can be assessed by the calibration intercept and slope. The calibration slope can be estimated by using the logistic recalibration framework which fits the model $logit(Y) = a + b_L \cdot L$, where L is the linear combination of predictors. If the calibration slope $b_L < 1$ for new data from the same population, the model is overfitting whereas $b_L > 1$ the model is underfitting [61]. In general, $b_L < 1$ means that the predicted risks are too extreme, $b_L > 1$ means that the predicted risks are too modest. The calibration intercept is obtained by fitting the model $logit(Y) = a + L$ where the slope b_L is set to 1. The predicted risks are on average underestimated if $a|b_L = 1 > 0$, and overestimated if $a|b_L = 1 < 0$ [61].

A calibration plot can be used to assess calibration of a model. It shows the risk predictions on the x-axis, and the observed proportions on the y-axis. The observed proportions per level of predicted risk cannot be directly observed, but are estimated. One way this can be done is by fitting the logistic model $logit(Y) = a + b_L \cdot L$. A different way is to use flexible, nonlinear calibration curve by fitting $logit(Y) = a + f(L)$. Here, f can be a continuous function of the linear predictor L, such as loess or spline transformations [61]. Perfect predictions should be at the 45° line [46]. Deviations from the 45° line can indicate over- or underestimation of risks.

2.6.3 Net Benefit

Traditional metrics like sensitivity, specificity, and AUC are statistical abstractions that are not directly informative about clinical value [65]. Medical decisions often involve trade-offs like obsolete tests for IUP/FPUL patients versus not diagnosing patients with an EP correctly. The Net Benefit is a decision analytic metric, which puts benefits and detriments on the same scale, by specifying a clinical judgement of the

relative values for the benefits (detection of EP) and detriments (unnecessary tests) associated with the prediction models [65]. For a model with a chosen risk threshold t the Net Benefit is calculated as

$$\text{Net Benefit} = \frac{TP}{n} - \frac{FP}{n} \cdot \frac{t}{1-t}, \tag{28}$$

where TP is the number of true positives (patients having a EP and being correctly classified) and FP is the number of false positives (patients not having an EP, but classified as having an EP). The risk threshold t indicates the relative importance of true positives. The lower t, the more important is the detection of true positives relative to the gain in false positives. Calculating the Net Benefit for a range of reasonable risk thresholds and plotting them in a graph leads to a decision curve. In this case this will be done for for EP/PPUL based on risk thresholds ranging from 3% to 10%. The Net Benefit is important for assessing if using a model to make medical choices is more beneficial than harmful. This is why reporting the Net Benefit is advisable to create models that contribute to making better clinical decisions.

2.7 Model validation

Validating a model on the exact same dataset it is built on leads to overly optimistic and potentially biased performance estimates. Independent assessments of the models performance is crucial. When no external validation data is available, methods exist to mimic external validation. Examples are sample splitting, k-fold cross-validation, and bootstrapping. These methods try to obtain an independent assessment of the model, even though data is coming from exactly the same population the model was built on (internal validation). Alternatively, applying the model on a new dataset (e.g. from another hospital, or another time period) assesses performance on a different population (external validation). The multi-center nature of the available dataset allows the combination of both into internal-external cross-validation (IECV).

2.7.1 Internal-external cross-validation

In a study by Takada et al. [48], the use of IECV is recommended in large clustered datasets to assess the generalizability of prediction models during their development and to to identify whether complex modelling strategies may offer advantages. In contrast to the classical internal validation methods, IECV allows for non-random hold-out samples with patients from different populations to evaluate performance of the model [48].

Considering that the data is collected from multiple centers, external validation can be mimicked using IECV. Here, instead of using random folds to calculate model performance, one center at a time is left out, corresponding to 8-folds. The procedure works as follows: A model is fit on each of the 10 imputed datasets except for the data from one center. The 10 resulting models are applied to the left-out-center data in each imputed dataset, leading to 10 predictions for the estimated risk of each observation. These estimated risks are averaged within each imputed dataset. Next, the performance metrics discussed in the previous sections are calculated for each imputed dataset on the left-out center data. The performance metrics are combined across the imputed datasets using Rubin's rules for pooling parameter estimates [43]. The process is visualized in Figure 20 in the appendix. The procedure is repeated eight times, each time leaving out a distinct center, leading to center-specific performance estimates.

2.7.2 Meta Analysis

Since the dataset comes from eight different hospitals in the UK, meta analysis is performed to obtain overall performance measures of the models by combining the center specific performance estimates resulting from the IECV. Meta analysis techniques are used to combine centre-specific results to account for the clustering by centers [13]. Meta analysis is performed for discrimination and calibration metrics. For the Net Benefit, meta analysis is not performed. There are two types of meta analysis: fixed-effects and random-effects. For the purpose of this thesis, random-effects meta analysis will be applied. Random-effects meta analysis assumes that the observed estimates of the independent variables can vary across studies because of real differences in the true effect likely due to differences between hospital-specific populations, as well as random sampling variability [5]. This is more reasonable than the fixed effects model, since it does not assume that the true effect sizes are the same across studies. With random-effects we acknowledge that the study populations may not come from a homogeneous group, but are drawn from $\theta_k \sim N(\mu, \tau^2)$, where τ^2 represents the heterogeneity between studies [42]. The overall effects are calculated as $\hat{\theta}_k = \mu + \zeta_k + \epsilon_k$, where μ is the mean of the distribution. ζ_k is the difference between μ and the true mean θ_k for a study k with $\zeta_k \sim N(0, \tau^2)$. ϵ_i denotes the sampling error for θ_i with $\epsilon_i \sim N(0, \nu_k)$. The random-effects weights for each study are calculated as [5]

$$W_k = \frac{1}{\nu_k + \tau^2}. \tag{29}$$

The pooled effect sizes can be computed as a weighted means [5]

$$M = \frac{\sum_{k=1}^{K} W_k \theta_k}{\sum_{k=1}^{K} W_k}. \tag{30}$$

2.8 Statistical Analysis

The risk prediction models are developed from 2894 PUL patients. Due to missing values, 100 imputed datasets were generated by Christodoulou et al. [13] leading to a final dataset of 289,400 observations. For this thesis, 10 imputed datasets will be used. The apparent model performance is the model performance on the very same dataset the model is built. It is obtained as follows: a model is fitted each of the 10 imputed datasets, leading to 10 models. The models are applied again to their corresponding imputed dataset. Using the predictions, the performance estimates are obtained for each imputed dataset and the estimates are combined using Rubin's rules to get the apparent performance estimates. This procedure is visualized in Figure A in the appendix. Next, internal-external cross-validation procedure is applied to ensure proper validation of the model. The performance metrics of binary and multi-class models are plotted using forest plots, calibration plots and decision diagrams.

In the first part, nine binary models are developed for predicting the risk of a PUL being an ectopic pregnancy. For this EP and PPUL were grouped together, as PPUL is most likely an ectopic pregnancy not detected by ultrasound. IUP and FPUL were combined since they have a low risk for complications. The apparent performance of a model is assessed using AUC. Next, internal-external cross-validation is used to evaluate discrimination, calibration and Net Benefit of the model. Discrimination was assessed by computing the AUC on each left-out center for all imputed datasets and combined by applying Rubin's rule to get the specific center results. The calibration intercept and slope and calculated by applying Rubin's rules across the imputed datasets. The calibration intercept and slope values are used to create logistic calibration curves for each center. To get the overall performance metrics, the center specific results are combined using random-effects meta analysis. Furthermore, the classification performance for EP/PPUL using a risk threshold of 5% on the estimated risk of EP/PPUL is evaluated. For the Net Benefit, meta analysis is not performed.

In the second part, nine multi-class models are developed with three possible outcomes: EP/PPUL vs. IUP vs. FPUL. Apparent model performance was assessed using the PDI, c-statistic for EP vs. other and the the c-statistic for FPUL vs. IUP using

the conditional risk method. Next, internal-external cross-validation is used for evaluating discrimination and calibration. The PDI as well as pairwise AUC for each pair of the three outcome categories were computed using the conditional risk technique to assess the models' ability to differentiate between the different PUL outcomes. Additionally, the calibration of the risk of EP was evaluated by computing the calibration intercept and slope across the imputed datasets and combining them with Rubin's rules to get center specific calibration curves. Meta analysis is carried out to get overall performance metrics. The overall calibration slope and intercept was used to create the overall calibration curves. The classification performance for EP/PPUL using a risk threshold of 5% on the estimated risk of EP/PPUL is evaluated. Here too, meta analysis is not performed for the Net Benefit.

2.9 Software

For the analysis the open-source software R version 4.1.1 is used [41]. For model fitting, additional packages are needed such as `glmnet`, `nnet`, `brglm2`, `rpart`, `caret` and `VGAM` [45, 64, 34, 50, 35, 71]. To calculate the AUC of the models, the package `auRoc` is utilized [21]. For the PDI, the `mcca` package is used [23]. For calculating results across imputed datasets with Rubin's rule, the `mitools` package is needed [36]. The meta analyses were performed using the `metafor` package [66]. Forest plots are created using the `forestplot` package [25]. Many additional packages were needed for data wrangling and visualization, which can be found in the code attached to this thesis.

Results

3.1 Descriptive Statistics

The dataset includes 334 cases of EP. The most common outcome is FPUL with 1335 cases. IUP outcome corresponds to 929 cases. 296 patients were lost in the follow-up. The median age among the patients is similar across the outcomes (see Table 2). The majority of patients do not have a history of EP. Vaginal bleeding is common across categories, but the majority of patients with the IUP outcome report no bleeding. The reported pain score from 0 to 10 shows a clear peak around 0 for all outcomes. The mean initial hCG is lowest for EP followed by FPUL. The IUP and LFU outcomes have a higher mean initial hCG and a larger variation than EP and FPUL. The mean initial progesterone is lowest for the FPUL and highest for the IUP. Similarly, the mean ratio hCG is lowest for FPUL and highest for IUP. Figure 5 displays a scatter plot of initial progesterone and logarithmic initial hCG colored by outcome category. It is visible that for the available measurements, FPUL outcomes seem to have a lower progesterone level, whereas IUP shows higher progesterone level. EP outcomes are somewhat between. In Figure 6 the logarithmic ratio hCG and the logarithmic initial hCG by outcome category are displayed. Separation between IUP, EP and FPUL is visible. The LFU cases seem to be randomly scattered in the plot.

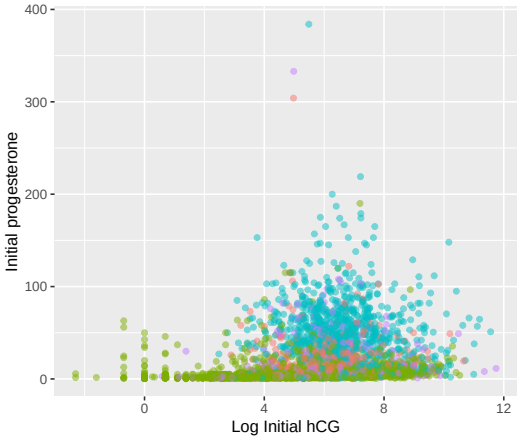


Figure 5: Initial progesterone vs. log initial hCG by outcome categories

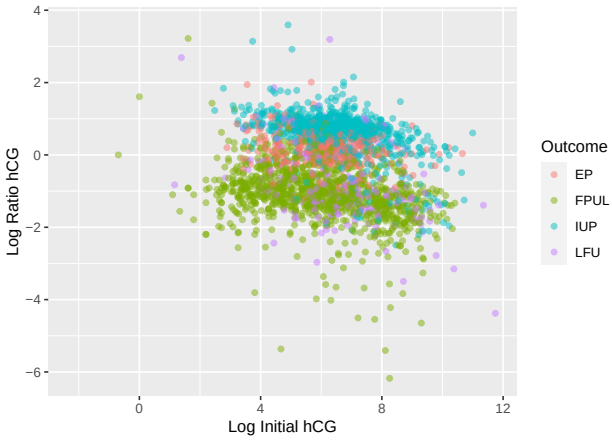


Figure 6: Log ratio hCG vs. log initial hCG by outcome categories

Table 2: Descriptive statistics per outcome category

Variable	EP/PPUL (n=334)	FPUL (n=1335)	IUP (n=929)	LFU (n=296)
Median Age	32	32	31	31
History of EP				
yes	46 (13.7%)	41 (3.1%)	87 (9.4%)	7 (2.4%)
no	280 (83.8%)	1213 (90.9%)	822(88.6%)	284 (95.6%)
NA	8 (2.4%)	80 (6%)	19 (2%)	5 (1.7%)
Vaginal Bleeding				
0	73 (21.2%)	87 (6.5%)	528 (56.9%)	45 (15.2%)
1	152 (45.5%)	316 (23.7%)	256 (27.6%)	77 (26%)
2	79(23.6%)	408 (30.6%)	69 (7.4%)	68 (23%)
3	18 (5.4%)	270 (20.2%)	30 (3.2%)	45 (15.2%)
4	12(3.6%)	250 (18.7%)	41 (4.4%)	61 (20.6%)
NA	-	3 (0.2%)	4 (0.4%)	-
Pain score				
0	126 (37.8%)	392 (29.4%)	281 (30.3%)	77 (26%)
1	21 (6.3%)	116 (8.7%)	45 (4.8%)	14 (4.7%)
2	48 (14.4%)	169 (12.7%)	94 (10.1%)	31 (10.5%)
3	41 (12.2%)	185 (13.9%)	132 (14.2%)	49 (16.6%)
4	35 (10.5%)	145 (10.9%)	116 (12.5%)	31 (10.5%)
5	19 (5.7%)	122 (9.1%)	87 (9.4%)	28 (9.5%)
6	12 (3.6%)	56 (4.2%)	41 (4.4%)	21 (7.1%)
7	7 (2.1%)	28 (2.1%)	33 (3.6%)	10 (3.4%)
8	9 (2.7%)	35 (2.6%)	31 (3.3%)	10 (3.4%)
9	4 (1.2%)	20 (1.5%)	16 (1.7%)	3 (1%)
10	-	14 (1%)	10 (1.1%)	5 (1.7%)
NA	12 (3.6%)	52 (3.9%)	42 (4.5%)	17 (5.7%)
Mean Initial hCG (std)	1309 (3431)	1622 (3454)	2465 (6906)	2589 (9517)
Mean Initial progesterone (std)	24 (25)	8 (13)	53 (35)	20 (31)
Mean hCG ratio (std)	1.34 (0.73)	0.44 (0.35)	2.15 (1.79)	0.87 (1.87)

3.2 Binary Models

In this section the results of the binary models (EP vs. other) are discussed in terms of the chosen evaluation measures: discrimination, calibration and Net Benefit.

3.2.1 Discrimination

The apparent performance of the nine different models is displayed in the forest plot in Figure 7. The LR model has by far the lowest AUC score (0.69, 95% CI 0.65-0.72), followed by the SVM model (0.87, 95% CI 0.84-0.90). The highest AUC is

obtained by the RF model with a perfect AUC score (1, 95% CI 0.99-1), followed by the NN (0.94, 95% CI 0.89-0.97). The other models are showing quite similar AUC scores of around 0.9, indicating an excellent discriminative ability.

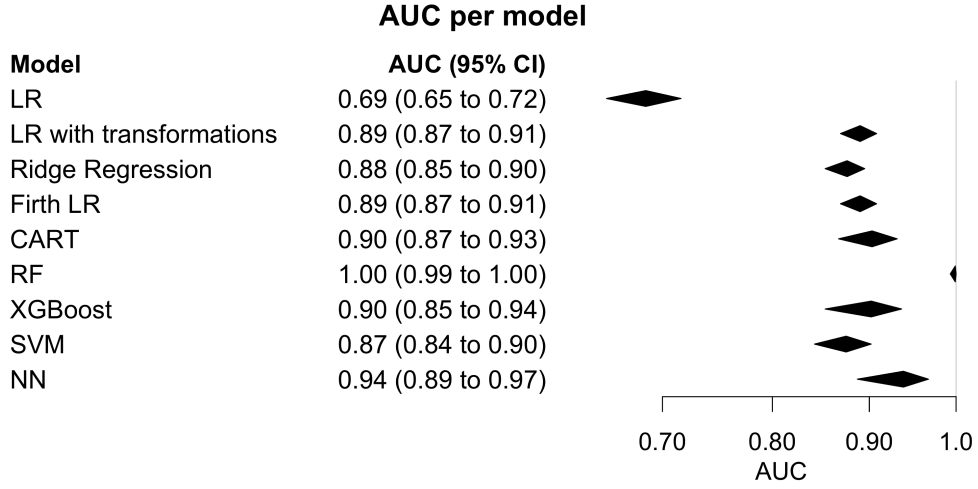


Figure 7: Apparent Model performance of different models

Next, the IECV is used to build and validate the models. The performance of the models using IECV is shown in Figure 8. Again the LR is again performing worst in terms of AUC (0.68, 95% CI 0.65-0.71). The highest AUC scores are obtained by the LR with transformations (0.89, 95% CI 0.87-0.91) and Firth LR (0.89, 95% CI 0.87-0.91), followed by RF (0.88, 95% CI 0.85-0.9). The NN IECV performance (0.83, 95% CI 0.8-0.86) is notably worse than the apparent performance. Similarly, the RF model was overfitting the data for the apparent performance. The AUC dropped from a perfect AUC of 1 (apparent performance) to 0.88 (IECV) which shows the importance of model validation. CART, XGBoost, SVM and NN are showing similar AUC scores between 0.82 to 0.84. LR with transformations, Ridge LR, and Firth LR are showing the same IECV AUC score as for the apparent AUC score. The center specific AUC scores for all nine models are displayed in Figures 21 - 29 in the appendix.

It is to note that for all models except LR, the confidence interval (CI) equals the prediction interval (PI). The same is observed in the multinomial setting. In general, the prediction interval is wider than the confidence interval due to the extra uncertainty it takes into account. The relatively small number of centers (eight), the low variation among the centers' AUC scores (except LR) and performing the logit transformation of the AUC scores prior to the meta analysis lead to τ^2 which is rounded to zero. τ^2 is defined as the estimated variance of underlying true effects across studies. The τ^2 equal to zero leads to PI=CI for all models except LR. To overcome this a Bayesian

approach to meta analysis can be favorable [19].

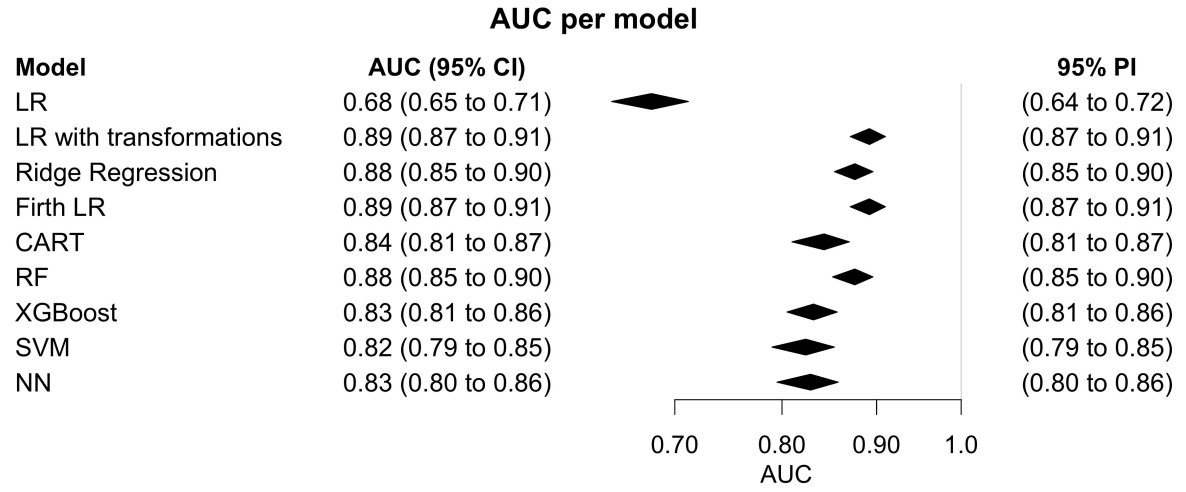


Figure 8: IECV Model performance of different models

In Table 3 the classification performance of the models based on a risk threshold of 5% is presented. The LR and SVM models are classifying almost every patient as high risk, leading to a high sensitivity but a low specificity. The highest combination of sensitivity and specificity is given by the LR with transformations and the RF model. The XGBoost model has the highest specificity, but the lowest sensitivity among the models. In our case the detection of EP is of vital importance, which means a high sensitivity beneficial. The NPV of LR with transformations, Ridge LR, Firth LR, and RF is 0.99 indicating that out of 100 negative predictions, there is 1 false negative. In turn, the PPV is highest for XGBoost with 0.33, meaning that out of 100 positive predictions, 33 are true positives. Overall, in terms of discrimination, LR with transformations, Firth LR and RF seem to outperform the other approaches.

Table 3: Classification Performance of binary models with 5% threshold

Model	% of high risk PUL	Sensitivity	Specificity	PPV	NPV
LR	0.91	0.98	0.08	0.13	0.96
LR with transformations	0.49	0.96	0.57	0.24	0.99
Ridge LR	0.70	0.99	0.35	0.17	0.99
Firth LR	0.49	0.96	0.58	0.25	0.99
CART	0.41	0.88	0.66	0.26	0.97
RF	0.48	0.95	0.59	0.25	0.99
XGBoost	0.26	0.68	0.80	0.33	0.95
SVM	0.97	0.99	0.01	0.13	0.94
NN	0.39	0.84	0.68	0.26	0.96

3.2.2 Calibration

Calibration was not assessed for the apparent model performance, since the comparison of mean predictions to observed outcomes on the same data the model is built on leads to a overly optimistic or even perfect calibration curve [46]. Calibration of the models for the IECV is assessed using logistic calibration curves. In Figure 9 the logistic calibration curves of the nine models are given, where the grey curve indicates the ideal model calibration curve. By comparing the calibration curves of the models with the ideal calibration curve, it is visible that not all models are well calibrated. To validate this, in the right corner of the Figure 9, the calibration slopes and intercepts for the models are given. The 95% confidence intervals for calibration intercepts include 0 in all models, not indicating over- or underestimating of the average risks. The calibration slopes however, show that the predicted risks for LR with transformations (1.17, 95% CI 1.09-1.25), Ridge LR (2.09, 95% CI 1.96-2.22), Firth LR (1.19, 95% CI 1.11-1.27), and SVM (1.63, 95% CI 1.08-2.19) are significantly too weak. Significantly too extreme predicted risks are detected for CART (0.74, 95% CI 0.7-0.78), XGBoost (0.31, 95% CI 0.3-0.32) and NN (0.45, 95% CI 0.4-0.51). Solely LR and RF have a calibration slope of 1.17 (95% CI 0.86-1.41) and 0.97 (95% CI 0.88-1.07) not indicating over- or underfitting of the predicted risks. In the Figures 30- 38 in the appendix, center specific calibration curves for the different models are given. Additionally, in Figure 39 in the appendix violin plots of the predicted risks for the different models are shown.

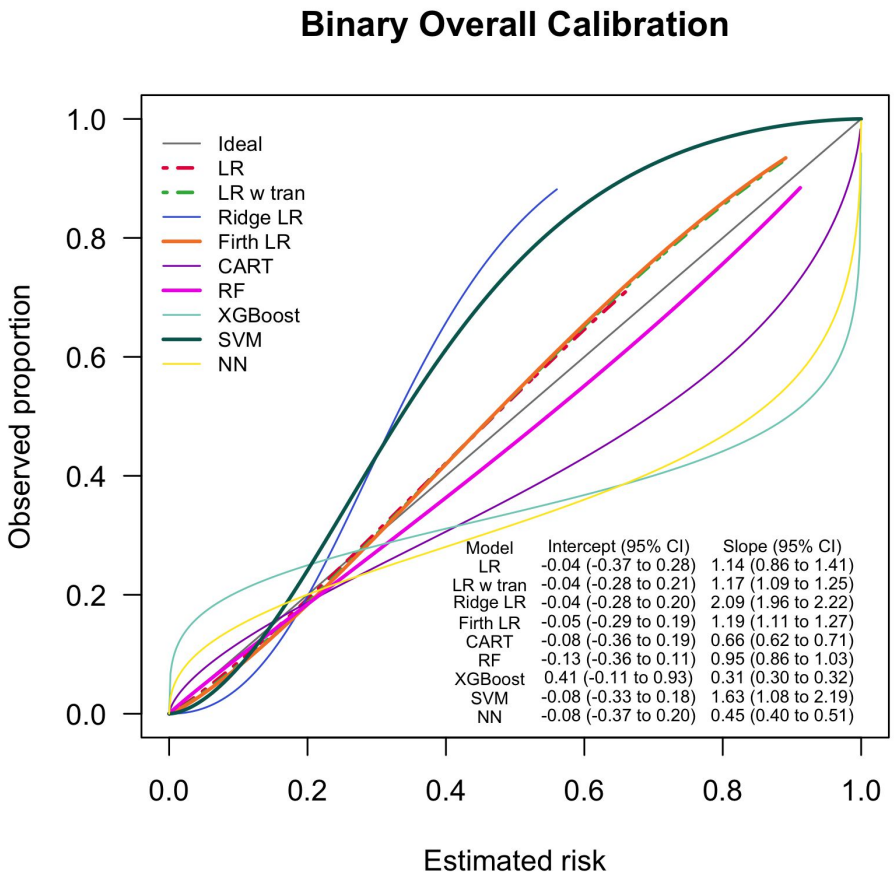


Figure 9: Calibration of different models

3.2.3 Net Benefit

The Net Benefits for different risk thresholds ranging from 3% to 10% are compared for all models in Figure 10. The black line indicates the Net Benefit if all patients are considered high risk (treat all), the grey line if no patient is considered high risk (treat none). The Net Benefit of the Firth LR, LR with transformations and RF are highest, with almost completely overlapping curves, always situated above the treat all curve. Ridge LR and CART have a slightly lower Net Benefit as the previous, while still above the treat all curve. The treat all curve is crossing with the LR, NN, XGBoost and SVM Net Benefit curves. The SVM Net Benefit curve is almost fully overlapping with the treat all curve. This indicates that for some risk threshold the models have the same Net Benefit as treating all patients, which translates to no benefit of using the model.

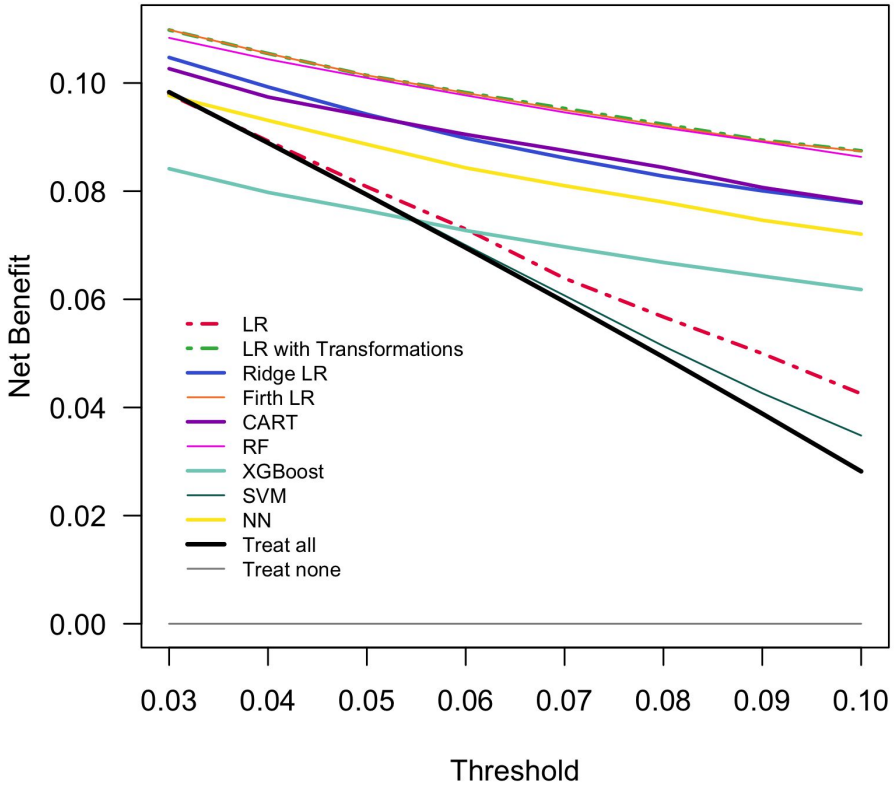


Figure 10: Net Benefit of Models

3.3 Multinomial Models

In this section the results of the multinomial models (EP vs. IUP vs. FPUL) are discussed. Again, discrimination, calibration and the Net Benefit are assessed.

3.3.1 Discrimination

In the forest-plot in Figure 11 the apparent model performance in terms of c-statistic for EP (cEP) for the nine different multinomial models is assessed. For this, the probabilities of FPUL and IUP are added and the AUC for the EP vs. IUP+FPUL is calculated. One can see that the apparent cEP is lowest for the LR (0.87, 95% CI 0.85-0.89). The NN (1, 95% CI 1.00-1.00) and RF (1.00, 95% CI 0.96-1.00) are performing highest in terms of cEP with a perfect score of 1. The other models show a similar score of around 0.9, indicating an excellent discriminative ability in the apparent approach for classifying EP vs. IUP+FPUL.

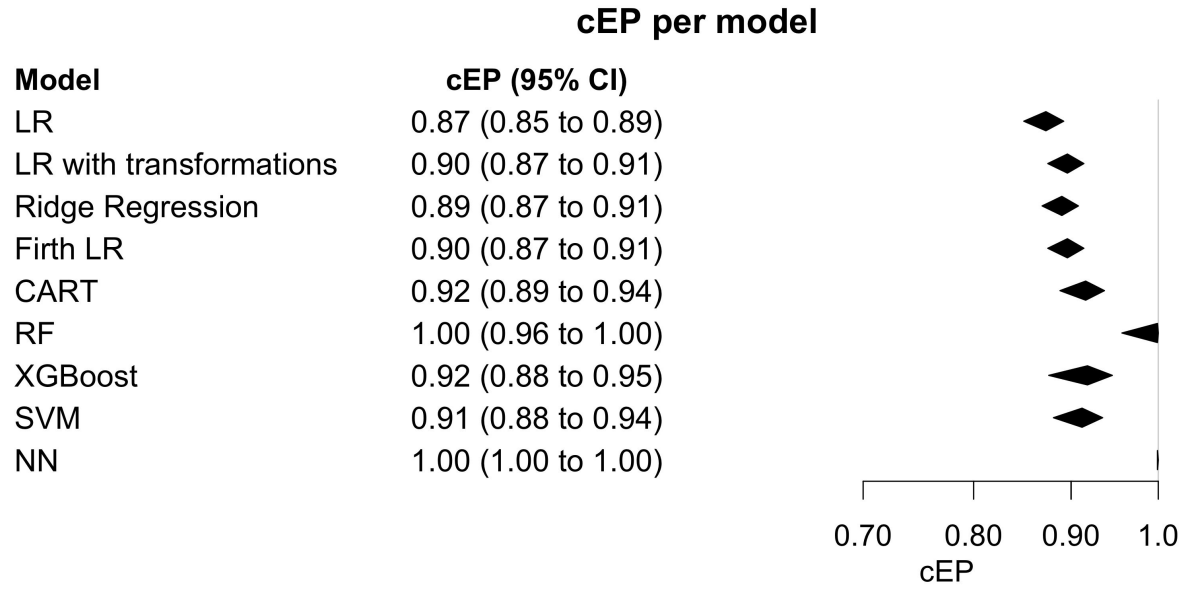


Figure 11: apparent cEP for different models

In the forest-plot in Figure 12 the apparent c-statistic for IUP vs. FPUL (cIF) for the different models is displayed. It is calculated using the conditional risk method. Overall, all models show a high cIF, indicating an excellent discrimination between IUP and FPUL. NN and RF have a perfect cIF (1, 95% CI 1.00-1.00). LR (0.97, 95% CI 0.96-0.98) and Ridge LR (0.97, 95% CI 0.97-0.98) have the lowest cIF among the models.

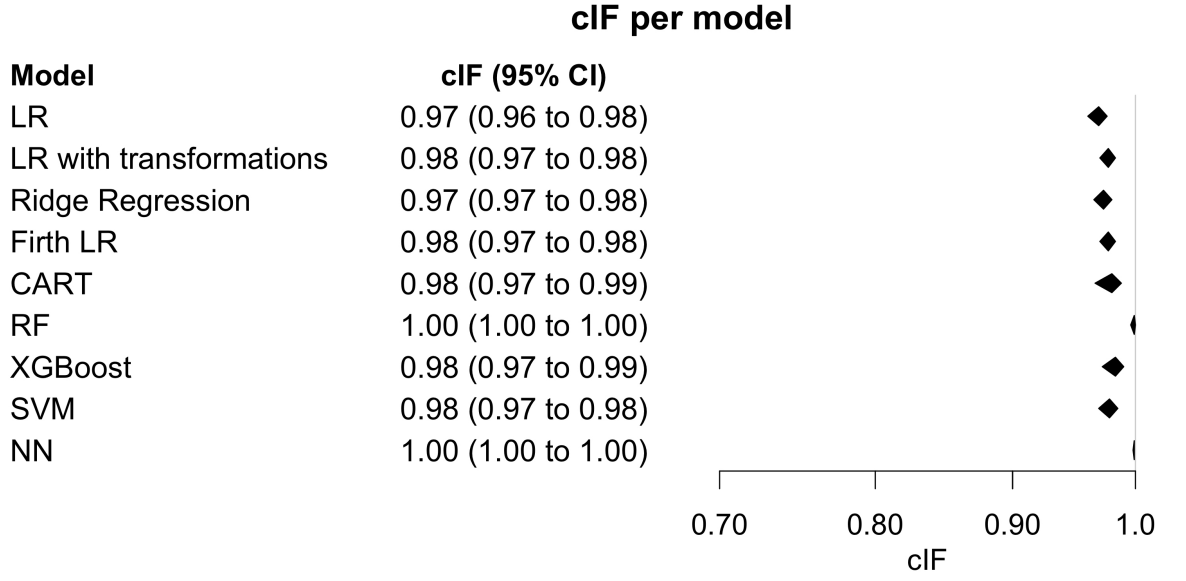


Figure 12: apparent cIF for different models

The apparent PDI scores of the different models can be found in the in Table 4. They can be compared with the PDI scores resulting from the IECV, shown in the same

table. The PDI is lowest for the CART model. It is visible that the ML models show a higher PDI in the apparent set up than in the IECV, indicating overfitting, whereas the "classical" regression models (LR, LR with transformations, Firth LR, Ridge LR) show a consistent performance. Overall, the IECV PDI is highest for the LR with transformations, Firth LR, XGBoost and NN.

Table 4: PDI per model

Model	apparent PDI (95% CI)	IECV PDI (95% CI)
LR	0.83 (0.81-0.85)	0.84 (0.82-0.86)
LR with Transformations	0.86 (0.84-0.88)	0.86 (0.84-0.88)
Ridge LR	0.85 (0.83-0.87)	0.85 (0.83-0.87)
Firth LR	0.86 (0.84-0.88)	0.86 (0.84-0.88)
CART	0.86 (0.82-0.90)	0.80 (0.77-0.82)
RF	0.99 (0.98-0.99)	0.85 (0.83-0.87)
XGBoost	0.88 (0.84-0.92)	0.86 (0.84-0.87)
SVM	0.87 (0.85-0.90)	0.85 (0.83-0.86)
NN	1.00 (1.00-1.00)	0.86 (0.84-0.88)

Using the IECV, the cEP of the different models is displayed in Figure 13. We can see that LR with transformations and Firth LR show the highest cEP, indicating a good discriminative ability of EP (0.9, 95% CI 0.88-0.91). The CART model shows the lowest performance in terms of cEP (0.85, 95% CI 0.82-0.87). Interestingly, the classical regression models show almost identical cEP for IECV and apparent approach as opposed to the ML models. This indicates that the ML models are prone to overfit and do not generate the same results when model validation is performed, whereas the classical approaches generalize better. The center specific cEP per model can be found in the appendix from Figure 40 - Figure 48.

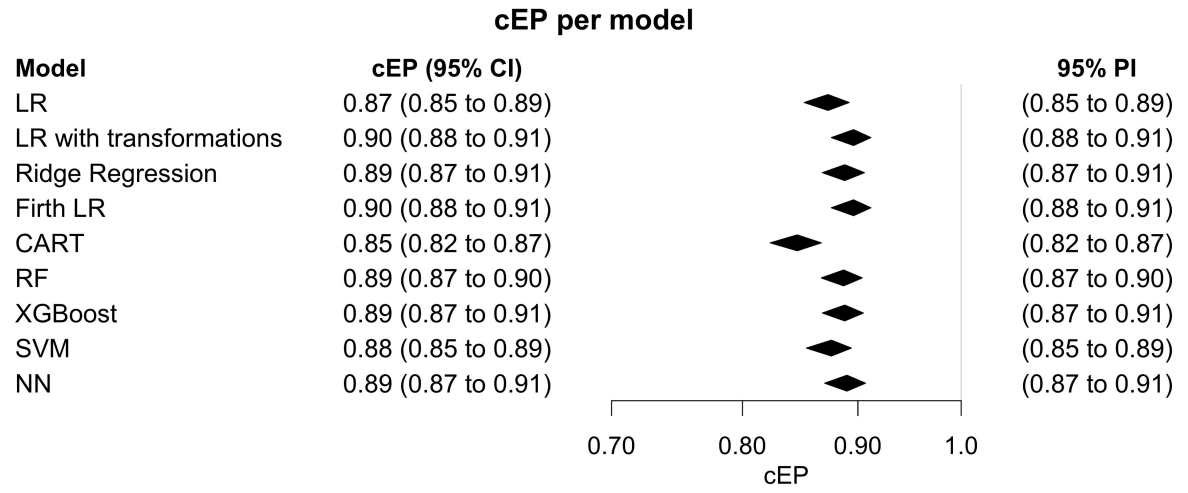


Figure 13: cEP for IECV models

In the forest-plot in Figure 14 the IECV cIF for the different models is displayed. Again, all models show a high cIF, indicating an excellent discrimination between IUP and FPUL. LR with transformations and Firth LR have the highest cIF (0.98, 95% CI 0.96-0.99), CART has the lowest cIF among the models 0.95, 95% CI 0.94-0.97). Comparing the apparent cIF and IECV cIF of the models, the ML models over-promise the cIF performance in the apparent approach, while the classical models cIF remains consistent, indicating a better generalizability of the models.

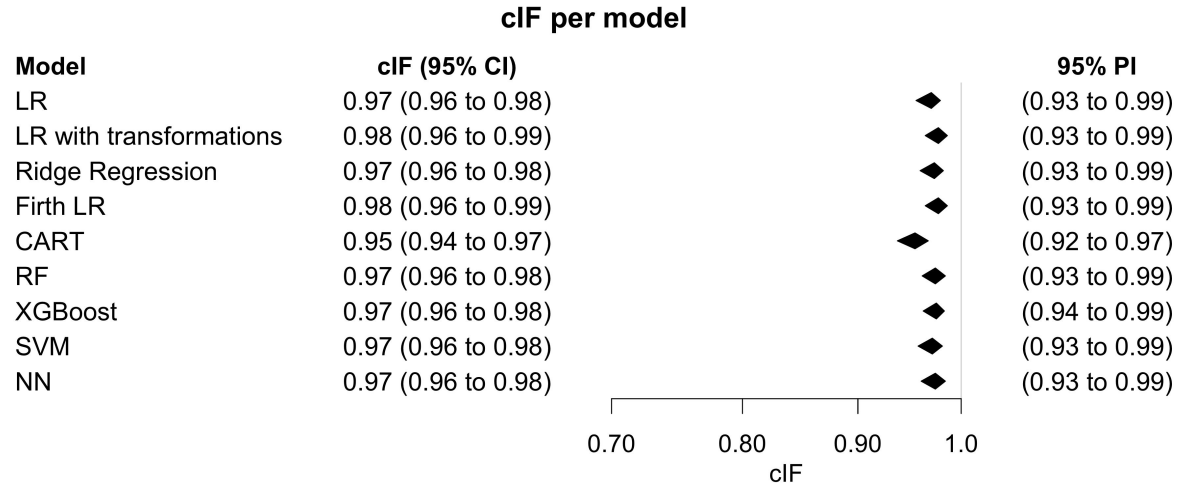


Figure 14: cIF for IECV models

In the Table 5 the classification performance for EP with a risk threshold of 5% is displayed. The LR, SVM, and Ridge LR models are classifying more than half of the patients as high risk, leading to a high sensitivity but a low specificity compared to the

other models. The highest combination of sensitivity and specificity is given by the LR with transformations, Firth LR and NN. The CART model has the highest specificity, but the lowest sensitivity among the models. The NPV all models except for CART is 0.99 indicating that out of 100 negative predictions, there is 1 false negative. In turn, the PPV is highest for CART with 0.45, meaning that out of 100 positive predictions, 45 are true positives. Overall, in terms of discrimination, LR with transformations, Firth LR and NN seem to outperform the other approaches.

Table 5: Classification Performance of multinomial models with 5% threshold

Model	% of high risk PUL	Sensitivity	Specificity	PPV	NPV
LR	0.69	0.98	0.36	0.18	0.99
LR with transformations	0.47	0.97	0.60	0.26	0.99
Ridge LR	0.71	0.99	0.33	0.17	0.99
Firth LR	0.47	0.97	0.60	0.26	0.99
CART	0.45	0.92	0.63	0.45	0.98
RF	0.51	0.97	0.55	0.24	0.99
XGBoost	0.46	0.95	0.61	0.26	0.99
SVM	0.56	0.96	0.50	0.22	0.99
NN	0.47	0.97	0.61	0.26	0.99

3.3.2 Calibration

In Figure 15 the logistic calibration curves in terms of EP of the nine multinomial models are given. The 95% confidence intervals for calibration intercepts include 0 in all models, not indicating over- or underestimating of the average risks. However, the calibration slopes show that the predicted risks for CART (0.66, 95% CI 0.59-0.72), is significantly too extreme. Significantly too weak predicted risks are detected for LR (1.71, 95% CI 1.62-1.79), LR with transformations(1.13, 95% CI 1.04-1.22), Ridge LR (2.23, 95% CI 2.08-2.38), Firth LR (1.13, 95% CI 1.04-1.22), RF (1.2, 95% CI 1.12-1.29), SVM (1.28, 95% CI 1.22-1.1.34) and NN (1.11, 95% CI 1.09-1.13). Solely XGBoost has a calibration slope of 1.06 (95% CI 1.0-1.12) not indicating over- or underfitting of the predicted risks. In the Figures 49- 57 in the appendix, center specific calibration curves for the different models are given. Additionally, in Figure 58 violin plots of the predicted risks for EP of the different models are shown.

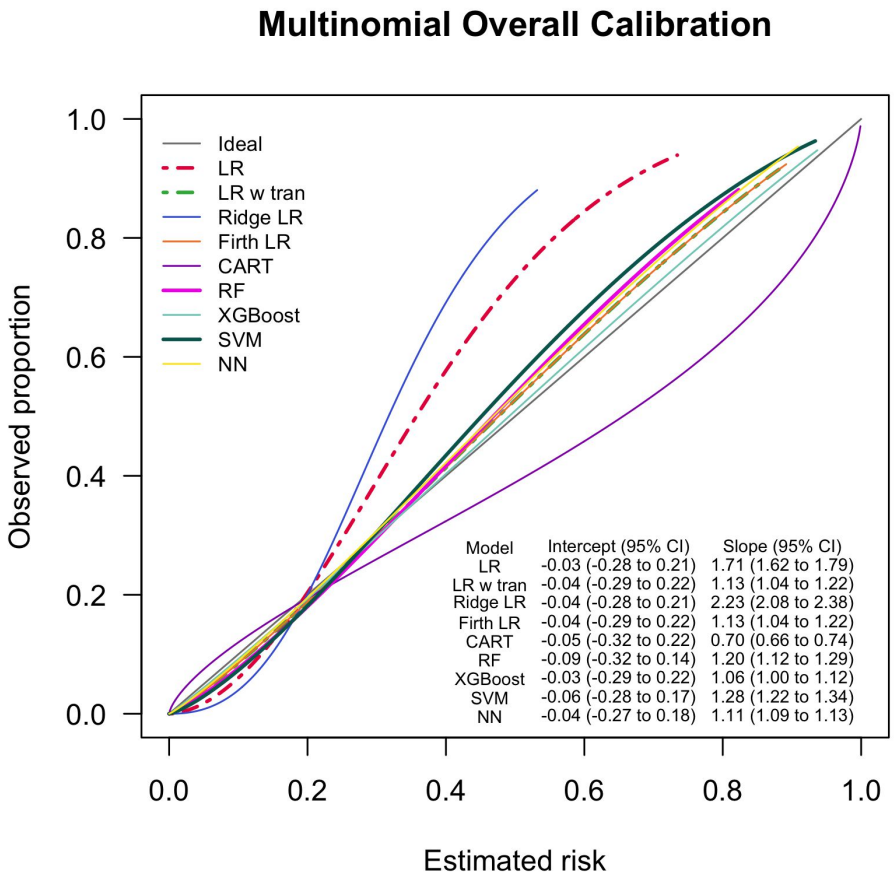
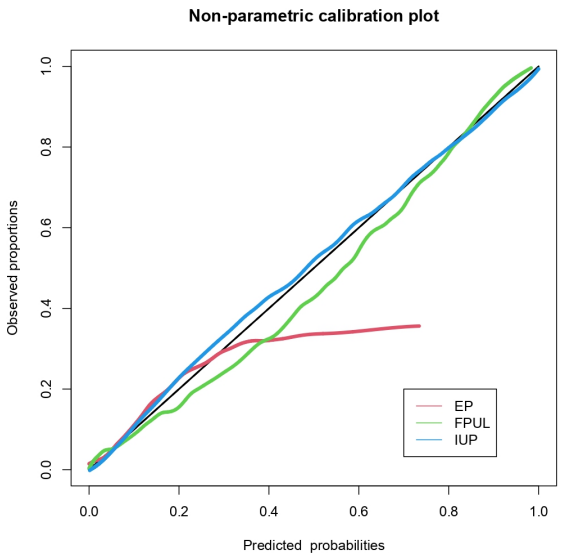
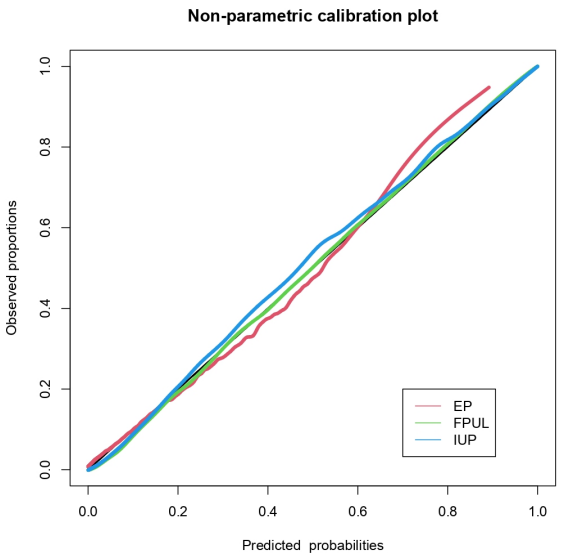


Figure 15: Multinomial models calibration

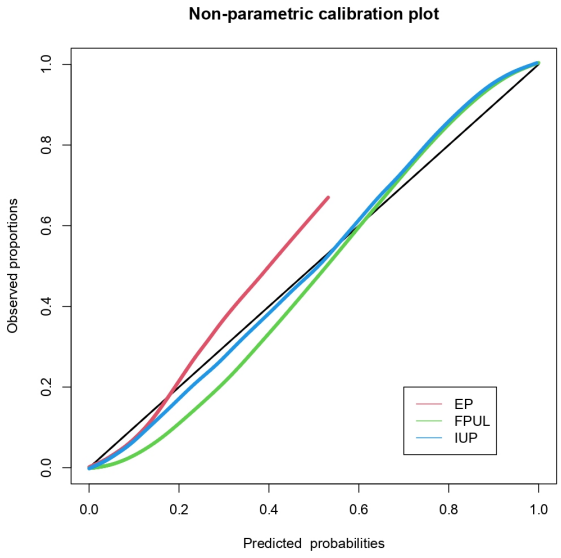
In Figure 16 and Figure 17 multinomial flexible calibration plots using a spline smoother are shown for the different models. All models seem to be fairly well calibrated with the exception of LR, Ridge LR and CART. These models seem to deviate heavily from the optimal 90° line when estimating the probability of EP, which is in line with the observations from the logistic calibration plot in Figure 15.



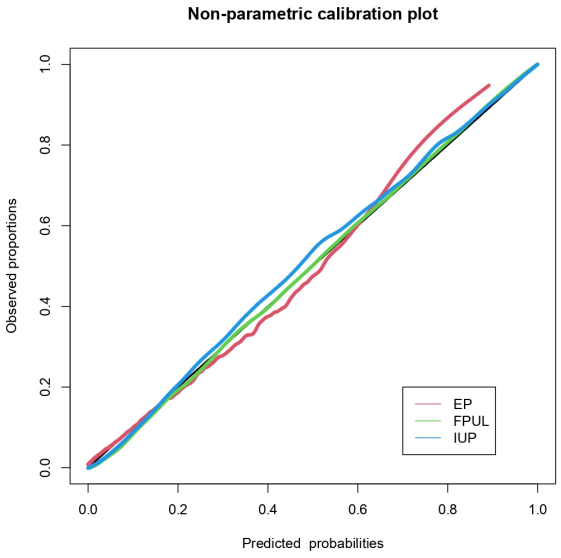
(a) LR



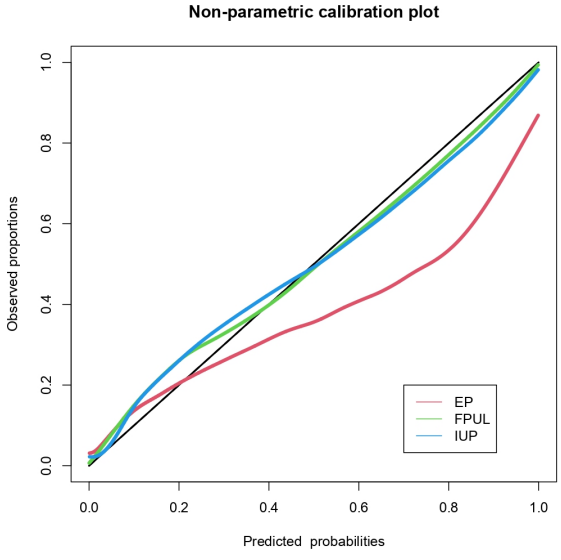
(b) LR with Transformations



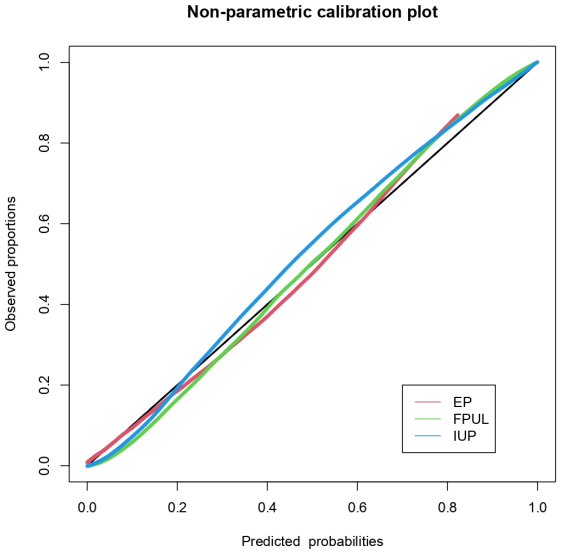
(c) Ridge LR



(d) Firth LR

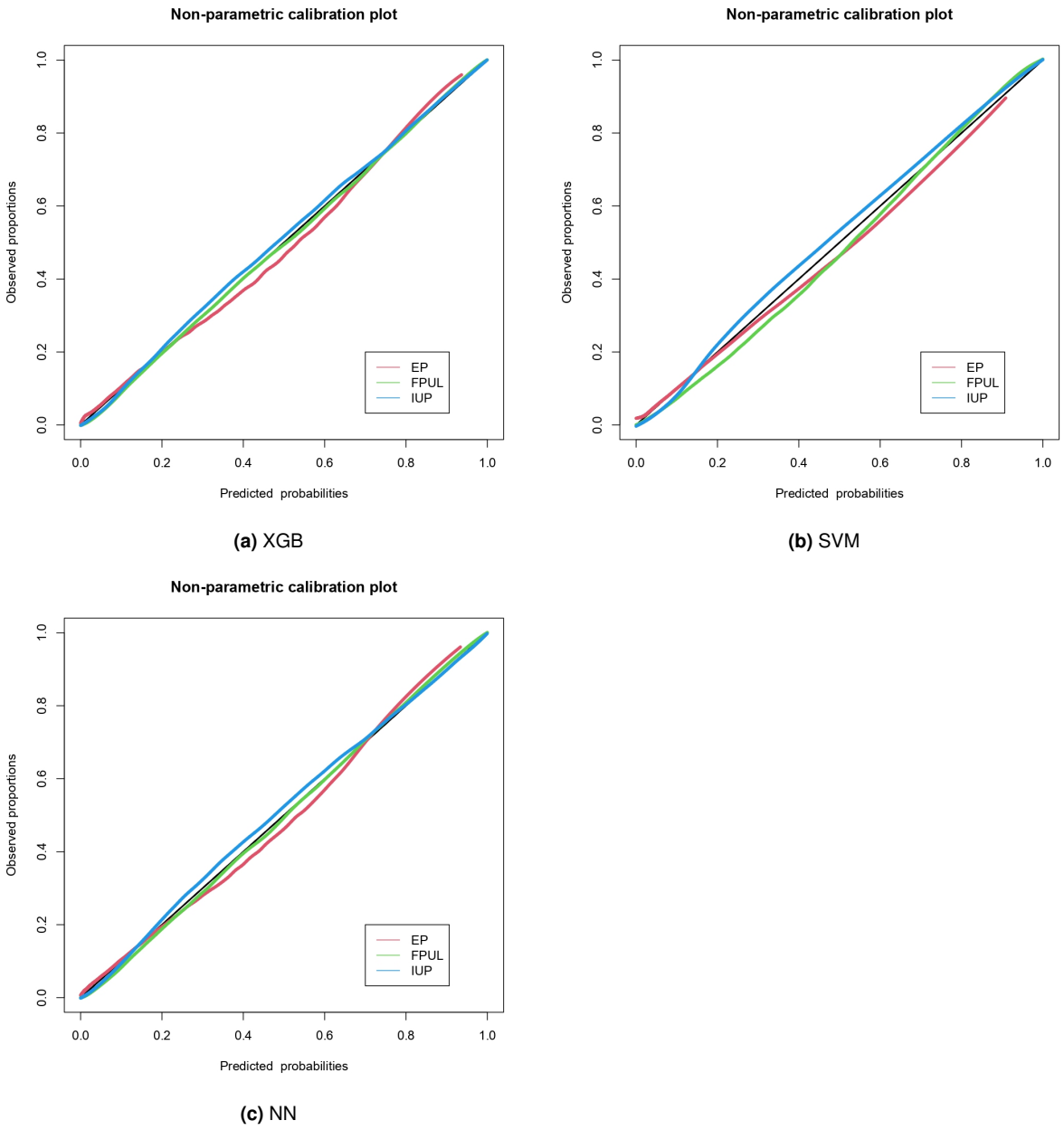


(e) CART



(f) RF

Figure 16: Multinomial flexible calibration curves



(a) XGB

(b) SVM

(c) NN

Figure 17: Multinomial flexible calibration curves

3.3.3 Net Benefit

The Net Benefit curves of the different models with risk thresholds ranging from 3% to 10% are compared in Figure 18. The Net Benefit of the LR with transformations, Firth LR and NN are highest, with almost completely overlapping curves. RF and XGBoost have a slightly lower Net Benefit curves as the previous. For smaller cut-offs, LR has the lowest Net Benefit, while for higher cut-offs CART has the lowest Net Benefit among the models. The treat all curve not crossing with any Net Benefit curve, indicating that using any model with the specified risk cut-offs always results in a higher Net Benefit than treating all patients.

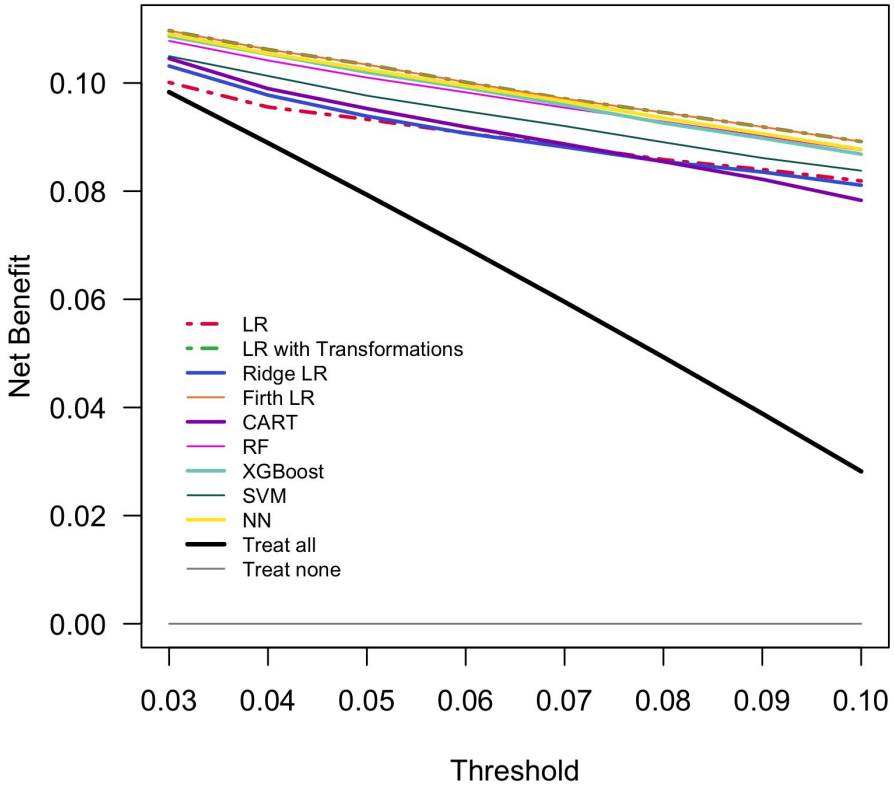


Figure 18: Multinomial models Net Benefit

Discussion

4.1 Findings

This study developed, validated and compared different risk prediction models for PULs using IECV. In the binary setting, LR with transformations and Firth LR performed best in terms of discrimination followed closely by RF. The LR was least effective in discriminating between the classes. For calibration of the risk estimates, solely LR and RF showed no indication of poor calibration. The Net Benefit of using the binary models in practice is highest for LR with transformations, Firth LR and RF. LR with transformations and Firth LR are behaving similar because of the large sample size which decreases the bias-reduction term of Firth LR to almost zero, making it equivalent to LR with transformations. Overall, for the binary models LR with transformations is the best choice when considering the results of discrimination, calibration and Net Benefit. The discriminative performance of the model varied among centers but remained on a high level. The center specific calibration curves show a reasonable calibration of the model. RF performed good as well, but is less interpretable and computationally heavier than LR with transformations.

In the multinomial setting, the highest PDI was obtained by LR with transformations, Firth LR, XGB and NN. For cEP, LR with transformations, Firth LR perform best, closely followed by RF and NN. The cIF is high for all models, but best for LR with transformations and Firth LR. In terms of calibration, only the XGBoost model did not indicate poor calibration of risk estimates. The CART model performed worst in the multinomial setting in terms of PDI, cEP, cIF and calibration. This model seems not to be suitable for building risk prediction models for PULs. The Net Benefit is highest for LR with transformations and Firth LR. Overall, for the multinomial case LR with transformations seems to be best suitable for predicting the risks of PULs. Again, the performance varies among the centers but resides on a high level. The center specific calibration curves and multinomial flexible calibration curves of the LR with transformations are indicating reasonable calibration of risk estimates.

In comparison, the multinomial models performed better or equal to the binary models in terms of discrimination. The calibration curves show a better calibration of risk estimates for the multinomial models compared to the binary models. Besides, the Net Benefit of the multinomial models is equal or higher to the binary models. All these results indicate that applying multinomial risk prediction models for predicting the outcome of PULs are more suitable than binary ones. The risk estimates for EP coming from multinomial models are on average more accurate than from binary models. As a consequence, using multinomial LR with transformations produces the best performing model when considering discrimination, calibration and Net Benefit.

4.2 Strengths and Limitations

The strengths of this study lie in the large data set used to develop, validate and compare the models. The large sample size in combination with observations coming from different populations (hospitals) should improve the generalizability of the models for application on other populations within the UK. Furthermore, the data was collected using a standardized protocol, which ensures data quality and reduces bias. Prior data cleaning was performed to check inconsistencies, retrieve missing data and correct errors. The variables included to build the models were selected based on previous research, and not in a data-driven approach. Lastly, the models were evaluated using discrimination, calibration and Net Benefit, which allowed an in-depth comparison of model performance.

However, a limitation of this study is that the observations are solely coming from the UK. The results might not be generalizable to other countries with different populations and healthcare systems. Additionally, external validation was only mimicked using IECV. It is strongly advised to externally validate the recommended models to further ensure generalizability. Another limitation of the study arises from application of multiple imputation due to missing data issues. Imputation should generally be done separately for training and testing data. In this study however, imputation was done for the whole data set without splitting the data first, which might lead to data leakage from training to testing set, possibly causing overoptimistic performance estimates. Lastly, for the models including hyperparameters, the search was performed using grid and random search techniques. Advanced search techniques such as Bayesian Optimization exist, which could further improve model performance.

4.3 Comparison with Other Studies

The M1 model, developed by Condous et al. [15] is a logistic regression model based on hCG ratio had a sensitivity of 91.7%, a specificity of 84.2%, a positive predictive value of 27.5% and a negative predictive value of 99.4%. Notably, the M1 model was an early attempt and built on a small dataset of 185 PUL cases. The recommended model built in this study (multinomial logistic regression model with transformations) outperforms the M1 model on the reported metrics; is based on a larger dataset; and additionally evaluates calibration and Net Benefit.

The M4 model developed by Condous et al. [16] is a multinomial logistic regression model which includes the hCG ratio and transformations of hCG ratio. M4 was developed to improve the M1 model and achieved an AUC of 0.9 which is equal cEP of the model recommended in this thesis. However, the M4 models implications are rather limited due to the small sample size. In a later study by Van Calster et al. [58] following the developed M4 model, the reported sensitivity for classifying EP was 88%. In comparison, the recommended model of this thesis achieved a sensitivity of 97% based on a risk threshold of 5%.

In a study by Van Calster et al. [55] nine different classification models for PUL were built and compared. Here, binary LR models using pairwise coupling performed best. The evaluation was done solely looking at discrimination metrics and did not take calibration and Net Benefit into account. The sample size was lower compared to the sample size used in this thesis.

The M6P model developed by Van Calster et al. [59] is based on progesterone, hCG and hCG ratio. The model is a multinomial logistic regression model that achieved 95% sensitivity and NPV of 99% and cEP of 0.9. In the external validation study by Christodoulou et al. [13] the M6P cEP was 0.89, the sensitivity 96% and the specificity 65% The recommended model in this thesis resulted into an cEP of 0.9, sensitivity of 97%, specificity of 60% and a NPV of 99%, meaning that the recommended model has a sightly higher sensitivity than the M6P model while having a lower specificity.

Overall, the results of this study are in line with previous efforts because they indicate that multinomial logistic regression shows great performance and that logistic regression is the preferred technique for developing risk prediction models for PULs.

4.4 Implications for Practice

This study demonstrated that a multinomial logistic regression model with the appropriate transformations performs best when comparing PUL risk prediction models based on discrimination, calibration and Net Benefit. The model discriminated efficiently between EP, IUP and FPUL. Especially important when classifying PULs is a high sensitivity. The high sensitivity of the model in combination with moderate specificity shows that the model can be a useful decision support tool in clinical practice. In terms of calibration, the risk estimates from the model were relatively calibrated, indicating accurate risk estimates which can be used to inform patients about their estimated risks. The Net Benefit of the model demonstrated that the use of the model to support medical decisions in practice is beneficial. Furthermore, this study implies that ML models do not add a relative merit over logistic regression techniques. However, the recommended model needs to be externally validated to confirm the results.

Conclusion

This thesis focused on developing, validating and comparing risk prediction models for PULs using binary and multinomial risk prediction models. The algorithms compared were logistic regression, logistic regression with transformations, Ridge regression, Firth logistic regression, Classification and Regression Trees, Random Forest, Extreme Gradient Boosting, Support Vector Machines and Neural Networks. The results presented in the previous chapters indicate that using data-driven, flexible machine learning algorithms do not add relative merit over standard regression techniques. In addition, using multinomial models rather than binary models led to more accurate risk predictions for ectopic pregnancies. As a result, multinomial logistic regression with transformations is the recommended risk prediction model for the use in clinical practice when predicting the outcome of PULs. Compared to the ML approaches, this model is less computationally expensive, well interpretable and explainable and performs well in terms of discrimination, calibration and Net Benefit.

Bibliography

[1] A. A. Abad. Lecture notes generalized linear models, February 2020.

[2] A. Agresti. *An introduction to categorical data analysis*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, third edition. edition, 2019. ISBN 9781119405269.

[3] H. Blockeel. *Machine learning and inductive inference*. Acco, Leuven, 2018. ISBN 9789033482977.

[4] B. Boeken. On the appropriateness of platt scaling in classifier calibration. *Information systems (Oxford)*, 95:101641, 2021. ISSN 0306-4379.

[5] M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, 1(2):97–111, 2010. ISSN 1759-2879.

[6] G. Boyraz and G. Bozdağ. Pregnancy of unknown location. *Journal of the Turkish German Gynecological Association*, 14:104–108, 06 2013. doi: $10.5152/\mathrm{jtgga}.2013.74317$.

[7] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. ISSN 0885-6125.

[8] L. Breiman. Statistical modeling: The two cultures. *Statistical science*, 16(3): 199–215, 2001. ISSN 0883-4237.

[9] S. Cessie and J. C. Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):191–201, 1992. ISSN 0035-9254.

[10] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining*, volume 13-17- of *KDD '16*, pages 785–794. ACM, 2016. ISBN 1450342329.

[11] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li. *xgboost: Extreme*

Gradient Boosting, 2021. URL https://CRAN.R-project.org/package=xgboost. R package version 1.4.1.1.

[12] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110:12–22, 2019. ISSN 0895-4356.

[13] E. Christodoulou, S. Bobdiwala, C. Kyriacou, J. Farren, N. Mitchell-Jones, F. Ayim, B. Chohan, O. Abughazza, B. Guruwadahyarhalli, M. Al-Memar, S. Guha, V. Vathanan, D. Gould, C. Stalder, L. Wynants, D. Timmerman, T. Bourne, and B. Van Calster. External validation of models to predict the outcome of pregnancies of unknown location: a multicentre cohort study. *BJOG : an international journal of obstetrics and gynaecology*, 128(3):552–562, 2021. ISSN 1470-0328.

[14] I. Cohen and M. Goldszmidt. Properties and benefits of calibrated classifiers. In *Knowledge Discovery in Databases: PKDD 2004*, volume 3202 of *Lecture Notes in Computer Science*, pages 125–136, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 9783540231080.

[15] G. Condous, E. Okaro, A. Khalid, D. Timmerman, C. Lu, Y. Zhou, S. Van Huffel, and T. Bourne. The use of a new logistic regression model for predicting the outcome of pregnancies of unknown location. *Human Reproduction*, 19(8):1900–1910, 08 2004. ISSN 0268-1161. doi: 10.1093/humrep/deh341. URL https://doi.org/10.1093/humrep/deh341.

[16] G. Condous, B. Van Calster, E. Kirk, Z. Haider, D. Timmerman, S. Van Huffel, and T. Bourne. Prediction of ectopic pregnancy in women with a pregnancy of unknown location. *Ultrasound in obstetrics & gynecology*, 29(6):680–687, 2007. ISSN 0960-7692.

[17] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

[18] C. Dabakoglu. What is support vector machine (svm)? https://medium.com/@cdabakoglu/what-is-support-vector-machine-svm-fd0e9e39514f, Dec 2018. Accessed: 2021-09-27.

[19] T. P. Debray, J. A. Damen, R. D. Riley, K. Snell, J. B. Reitsma, L. Hooft, G. S. Collins, and K. G. Moons. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Statistical Methods in Medical*

Research, 28(9):2768–2786, 2019. doi: 10.1177/0962280218785504. URL https://doi.org/10.1177/0962280218785504. PMID: 30032705.

[20] I. El Naqa and M. J. Murphy. *What Is Machine Learning?*, pages 3–11. Springer International Publishing, Cham, 2015. ISBN 978-3-319-18305-3. doi: 10.1007/978-3-319-18305-3_1. URL https://doi.org/10.1007/978-3-319-18305-3_1.

[21] D. Feng, D. Manevski, and M. P. Perme. *auRoc: Various Methods to Estimate the AUC*, 2020. URL https://CRAN.R-project.org/package=auRoc. R package version 0.2-1.

[22] D. FIRTH. Bias reduction of maximum-likelihood-estimates. *Biometrika*, 80(1): 27–38, 1993. ISSN 0006-3444.

[23] M. Gao and J. Li. *mcca: Multi-Category Classification Accuracy*, 2019. URL https://CRAN.R-project.org/package=mcca. R package version 0.7.0.

[24] B. A. Goldstein, A. M. Navar, and R. E. Carter. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European Heart Journal*, 38(23):1805–1814, 07 2016. ISSN 0195-668X. doi: 10.1093/eurheartj/ehw302. URL https://doi.org/10.1093/eurheartj/ehw302.

[25] M. Gordon and T. Lumley. *forestplot: Advanced Forest Plot Using 'grid' Graphics*, 2021. URL https://CRAN.R-project.org/package=forestplot. R package version 2.0.1.

[26] S. W. Grant, G. S. Collins, and S. A. M. Nashef. Statistical Primer: developing and validating a risk prediction model. *European Journal of Cardio-Thoracic Surgery*, 54(2):203–208, 05 2018. ISSN 1010-7940. doi: 10.1093/ejcts/ezy180. URL https://doi.org/10.1093/ejcts/ezy180.

[27] B. Gravesteijn, D. Nieboer, A. Ercole, C. Åkerlund, N. Andelic, L. Aneassen, A. Antoni, G. Audibert, M. Azzolini, A. Belli, L. Beretta, M. Blaabjerg, A. Brazinova, C. Brorsson, A. Buki, M. Bullinger, A. Caccioppola, M. Calvi, M. Carbonara, G. Chevallard, J. P. Coles, A. Covic, N. Curry, V. De Keyser, F. Della Corte, A. Dixit, J. Guy-Loup Dulière, P. Esser, V. Feigin, Kelly Foks, P. Gagliardo, G. Gao, P. George, J. Golubovic, P. Gomez, F. Grossi, J. Haagsma, L. Horton, J. Huijben, B. Jacobs, M. Ji-yao Jiang, K. M. Jones, A. G. Kolias, S. Laureys, F. Lecky, R. Lefering, L. Levi, R. Lightfoot, H. Lingsma, A. Castaño-León, M. Majdan, G. Manley, C. Martino, L. Murray, A. Negru, V. F. Newcombe, M. Oresic, F. Ortolano, A. Piippo-Karjalainen, M. Pirinen, S. Polinder, J. Posti, M. Rambadagalla,

R. Real, S. Ripatti, S. Rocka, O. Roise, G. Rosenthal, R. Rossaint, M. Rusnák, J. Sahuquillo, O. Sakowitz, N. Schäfer, G. Schoonman, E. Schwendenwein, C. Sewalt, T. Skandsen, P. Smielewski, W. Stewart, R. Takala, B. Ao, A. Theadom, C. M. Tolias, T. Trapani, C. Tudora, P. Vajkoczy, E. Valeinis, J. van Dijck, T. van Essen, C. M. van Heugten, D. Van Praag, A. Vanhaudenhuyse, A. Vargiolu, E. Vega, A. Vik, V. Volovici, D. Voormolen, P. Vulekovic, L. Wilson, S. Winzeck, and S. Wolf. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *Journal of clinical epidemiology*, 122: 95–107, 2020. ISSN 0895-4356.

[28] K. W. Hoover, G. Tao, and C. K. Kent. Trends in the diagnosis and treatment of ectopic pregnancy in the united states. *Obstetrics & Gynecology*, 115(3), 2010.

[29] D. Inc. What is a neural network?, Aug 2020. URL `https://databricks.com/de/glossary/neural-network`.

[30] T. Jo. *Machine learning foundations: supervised, unsupervised, and advanced learning*. Springer, Cham, Switzerland, 1st ed. 2021. edition, 2021. ISBN 3030659003.

[31] K. Johnson. *Applied Predictive Modeling*. Springer New York : Imprint: Springer, New York, NY, 1st ed. 2013. edition, 2013. ISBN 1461468493.

[32] E. Kirk, G. Condous, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Rationalizing the follow-up of pregnancies of unknown location. *Human reproduction (Oxford)*, 22(6):1744–1750, 2007. ISSN 0268-1161.

[33] E. Kirk, G. Condous, and T. Bourne. Pregnancies of unknown location. *Best practice & research. Clinical obstetrics & gynaecology*, 23(4):493–499, 2009. ISSN 1521-6934.

[34] I. Kosmidis, E. C. Kenne Pagui, and N. Sartori. Mean and median bias reduction in generalized linear models. *Statistics and Computing*, 30:43–59, 2020. URL `https://doi.org/10.1007/s11222-019-09860-6`.

[35] M. Kuhn. *caret: Classification and Regression Training*, 2021. URL `https://CRAN.R-project.org/package=caret`. R package version 6.0-90.

[36] T. Lumley. *mitools: Tools for Multiple Imputation of Missing Data*, 2019. URL `https://CRAN.R-project.org/package=mitools`. R package version 2.4.

[37] T. M. Mitchell. *Machine learning*. McGraw-Hill series in computer science. McGraw-Hill, Boston, 1997. ISBN 0070428077.

[38] M. J. Pencina and S. D'Agostino, Ralph B. Evaluating Discrimination of Risk Prediction Models: The C Statistic. *JAMA*, 314(10):1063–1064, 09 2015. ISSN 0098-7484. doi: $10.1001/\mathrm{jama.2015.11082}$. URL `https://doi.org/10.1001/jama.2015.11082`.

[39] P. P. Pereira, C. Fábio Roberto, G. Úrsula Trovato, and F. Rossana Pulcineli Vieira. Pregnancy of unknown location. *Clinics (Sao Paulo, Brazil)*, 74:e1111, 10 2019. doi: $10.6061/\mathrm{clinics}/2019/\mathrm{e1111}$.

[40] P. Probst, B. Bischl, and A.-L. Boulesteix. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20, 2018.

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

[42] R. D. Riley, J. P. T. Higgins, and J. J. Deeks. Interpretation of random effects meta-analyses. *BMJ*, 342(7804):c221–967, 2011. ISSN 0959-8138.

[43] D. B. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley, New York (N.Y.), 1987. ISBN 047108705X.

[44] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques. Recent advances and applications of machine learning in solid-state materials science. *npj computational materials*, 5(1):1–36, 2019. ISSN 2057-3960.

[45] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011. URL `https://www.jstatsoft.org/v39/i05/`.

[46] E. W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for biology and health. Springer International Publishing AG, Cham, 2019. ISBN 3030163989.

[47] J. Sá, A. Almeida, B. Pereira da Rocha, M. Mota, J. R. De Souza, and L. Dentel. Lightning forecast using data mining techniques on hourly evolution of the convective available potential energy, 03 2016.

[48] T. Takada, S. Nijman, S. Denaxas, K. I. Snell, A. Uijl, T.-L. Nguyen, F. W. Asselbergs, and T. P. Debray. Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. *Jour-*

nal of Clinical Epidemiology, 137:83–91, 2021. ISSN 0895-4356. doi: https://doi.org/10.1016/j.jclinepi.2021.03.025. URL https://www.sciencedirect.com/science/article/pii/S0895435621001074.

[49] K. Takeuchi, E. A. R. Filho, D. S. Santana, J. G. Cecatti, M. L. Costa, S. M. Haddad, M. A. Parpinelli, M. H. Sousa, R. S. Camargo, R. C. Pacagnella, F. G. Surita, and J. L. Pinto e Silva. Awareness about a life-threatening condition: Ectopic pregnancy in a network for surveillance of severe maternal morbidity in brazil. *BioMed Research International*, 2014:965724, 2014. doi: 10.1155/2014/965724. URL https://doi.org/10.1155/2014/965724.

[50] T. Therneau, B. Atkinson, and B. Ripley. *rpart: Recursive partitioning for classification, regression and survival trees.*, 2019. URL https://cran.pau.edu.tr/web/packages/rpart/rpart.pdf. R package version4.1-15.

[51] R. Thomas. Medicine's machine learning problem, Apr 2021. URL https://bostonreview.net/science-nature/rachel-thomas-medicines-machine-learning-problem.

[52] R. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York : Imprint: Springer, New York, NY, 2nd ed. 2009. edition, 2009. ISBN 9780387848587.

[53] J. V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231, 1996. ISSN 0895-4356. doi: https://doi.org/10.1016/S0895-4356(96)00002-9. URL https://www.sciencedirect.com/science/article/pii/S0895435696000029.

[54] S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. URL https://www.jstatsoft.org/v45/i03/.

[55] B. Van Calster, G. Condous, E. Kirk, T. Bourne, D. Timmerman, and S. Van Huffel. An application of methods for the probabilistic three-class classification of pregnancies of unknown location. *Artificial intelligence in medicine*, 46(2):139–154, 2008. ISSN 0933-3657.

[56] B. Van Calster, V. Van Belle, Y. Vergouwe, D. Timmerman, S. Van Huffel, and E. W. Steyerberg. Extending the c-statistic to nominal polytomous outcomes: the

polytomous discrimination index. *Statistics in medicine*, 31(23):2610–2626, 2012. ISSN 0277-6715.

[57] B. Van Calster, Y. Vergouwe, C. W. N. Looman, V. Van Belle, D. Timmerman, and E. W. Steyerberg. Assessing the discriminative ability of risk models for more than two outcome categories. *European journal of epidemiology*, 27(10):761–770, 2012. ISSN 0393-2990.

[58] B. Van Calster, Y. Abdallah, S. Guha, E. Kirk, K. Van Hoorde, G. Condous, J. Preisler, W. Hoo, C. Stalder, C. Bottomley, D. Timmerman, and T. Bourne. Rationalizing the management of pregnancies of unknown location: temporal and external validation of a risk prediction model on 1962 pregnancies. *Human Reproduction*, 28(3):609–616, 01 2013. ISSN 0268-1161. doi: 10.1093/humrep/des440. URL https://doi.org/10.1093/humrep/des440.

[59] B. Van Calster, Y. Abdallah, S. Guha, E. Kirk, K. Van Hoorde, G. Condous, J. Preisler, W. Hoo, C. Stalder, C. Bottomley, D. Timmerman, and T. Bourne. Rationalizing the management of pregnancies of unknown location: temporal and external validation of a risk prediction model on 1962 pregnancies. *Human reproduction (Oxford)*, 28(3):609–616, 2013. ISSN 0268-1161.

[60] B. Van Calster, S. Bobdiwala, S. Guha, K. Van Hoorde, M. Al-Memar, R. Harvey, J. Farren, E. Kirk, G. Condous, S. Sur, C. Stalder, D. Timmerman, and T. Bourne. Managing pregnancy of unknown location based on initial serum progesterone and serial serum hcg levels: development and validation of a two-step triage protocol. *Ultrasound in obstetrics & gynecology*, 48(5):642–649, 2016. ISSN 0960-7692.

[61] 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*, 74:167–176, 2016. ISSN 0895-4356.

[62] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, E. W. Steyerberg, P. Bossuyt, G. S. Collins, P. Macaskill, D. J. McLernon, K. G. M. Moons, E. W. Steyerberg, A. J. Vickers, O. behalf of Topic Group 'Evaluating diagnostic tests, and prediction models'of the STRATOS initiative. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019. doi: 10.1186/s12916-019-1466-7. URL https://doi.org/10.1186/s12916-019-1466-7.

[63] 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*, 14(1):137–137, 2014. ISSN 1471-2288.

[64] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `https://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

[65] A. J. Vickers, B. Van Calster, and E. W. Steyerberg. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, 352, 2016. doi: $10.1136/\mathrm{bmj.i6}$. URL `https://www.bmj.com/content/352/bmj.i6`.

[66] W. Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010. URL `https://doi.org/10.18637/jss.v036.i03`.

[67] H. Wang and D. Hu. Comparison of svm and ls-svm for regression. In *2005 International Conference on Neural Networks and Brain*, volume 1, pages 279–283, 2005. doi: $10.1109/\mathrm{ICNNB.2005.1614615}$.

[68] X. Wang. Firth logistic regression for rare variant association tests. *Frontiers in genetics*, 5:187–187, 2014. ISSN 1664-8021.

[69] H. J. P. Weerts, A. C. Mueller, and J. Vanschoren. Importance of tuning hyperparameters of machine learning algorithms, 2020.

[70] L. Wynants, G. Collins, and B. Van Calster. Key steps and common pitfalls in developing and validating risk models. *BJOG : an international journal of obstetrics and gynaecology*, 124(3):423–432, 2017. ISSN 1470-0328.

[71] T. W. Yee. The VGAM package for negative binomial regression. *Australian and New Zealand Journal of Statistics*, 61, 2020.

[72] C. Zhang and Y. Ma. *Ensemble Machine Learning: Methods and Applications*. Springer New York : Imprint: Springer, New York, NY, 1st ed. 2012. edition, 2012. ISBN 1280794429.

Appendices

Additional Tables and Figures

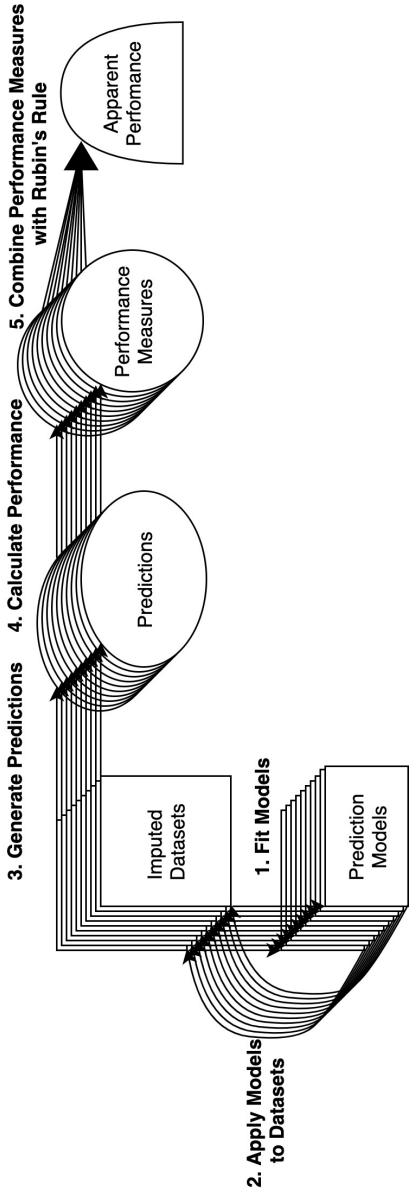


Figure 19: Process to obtain the apparent Model Performance

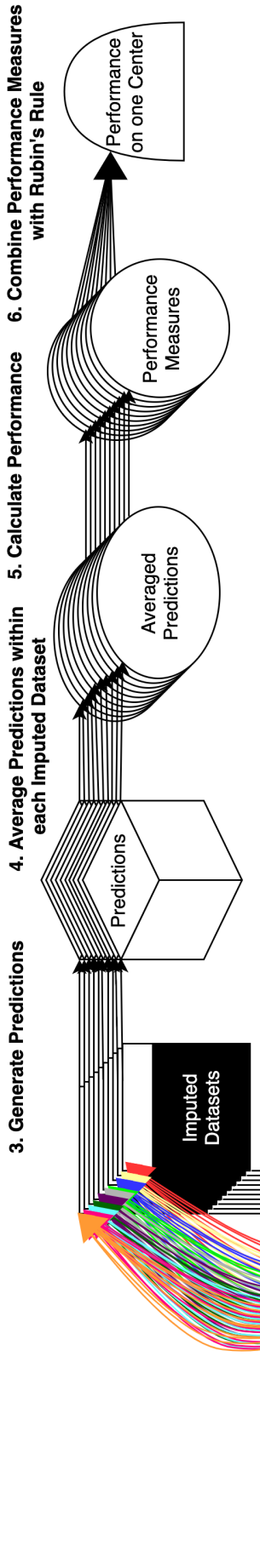


Figure 20: Process to obtain the Model Performance with IECV

Binary Models - Discrimination

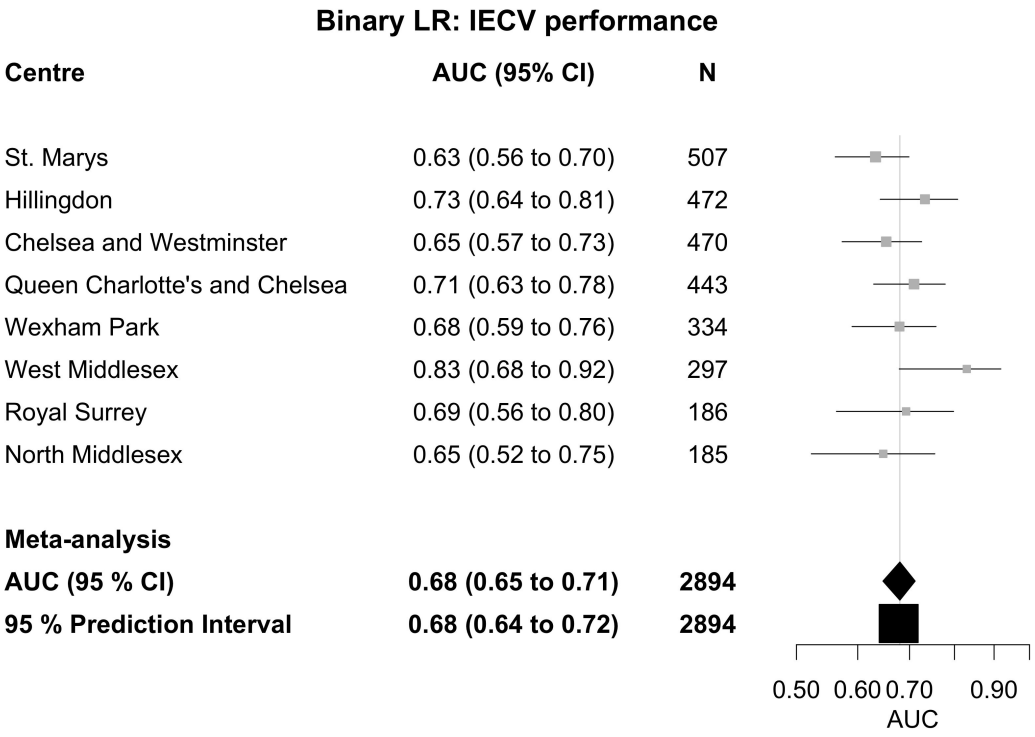


Figure 21: AUC per center for the binary LR model

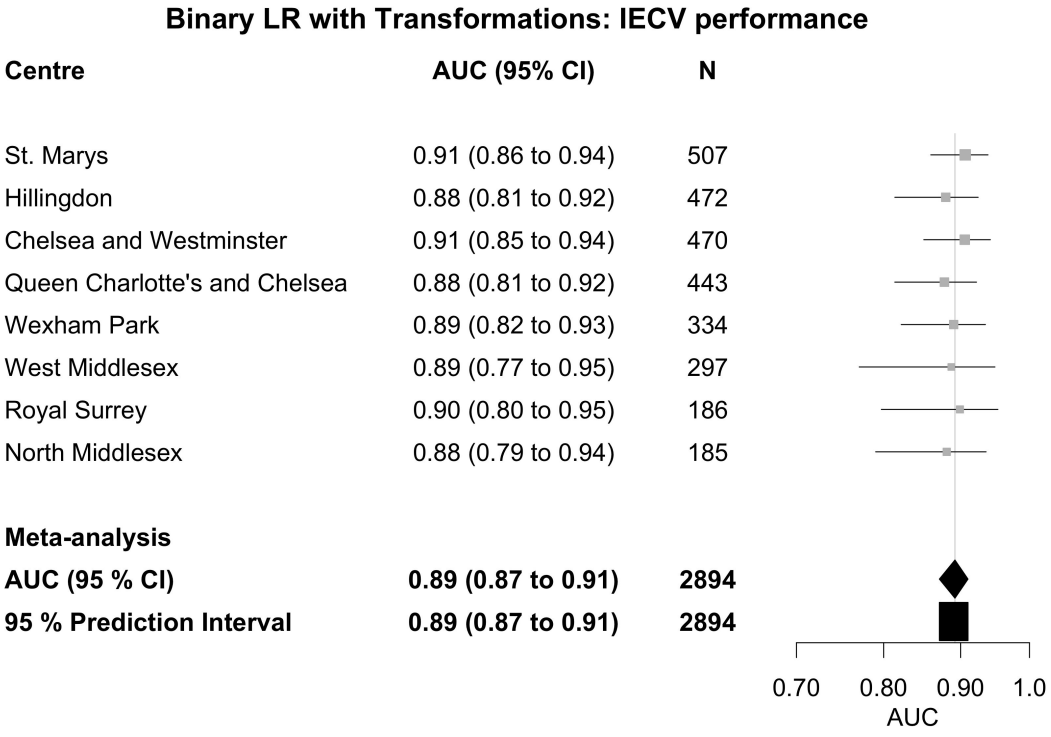


Figure 22: AUC per center for the binary LR model with Transformations

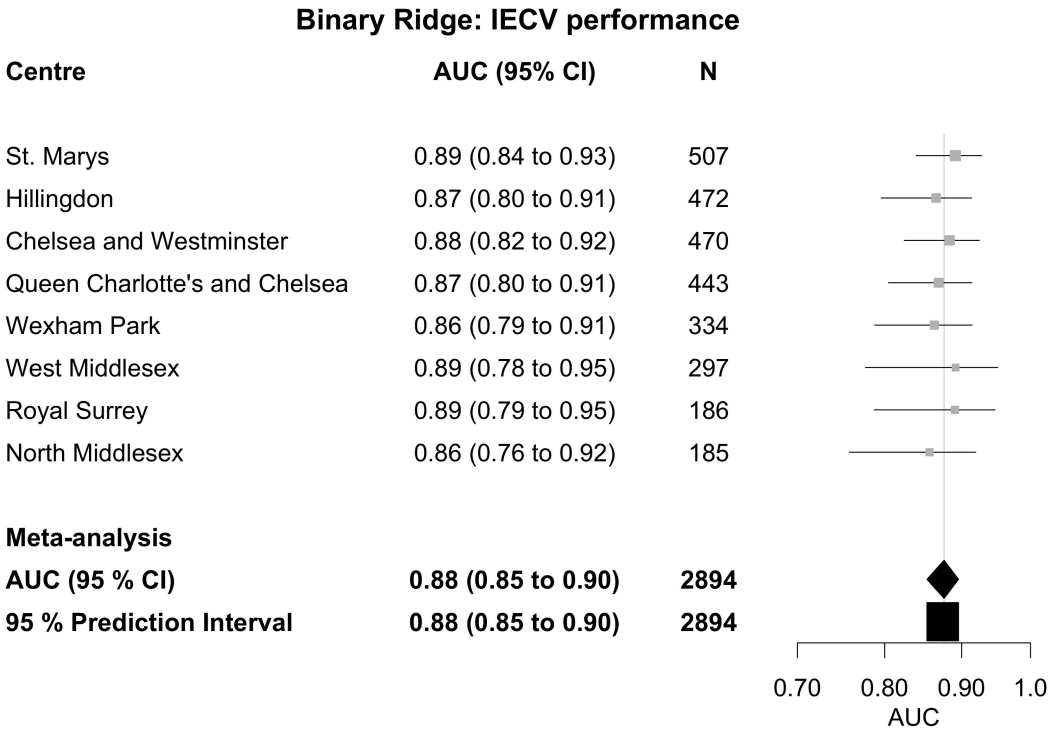


Figure 23: AUC per center for the binary Ridge Regression model

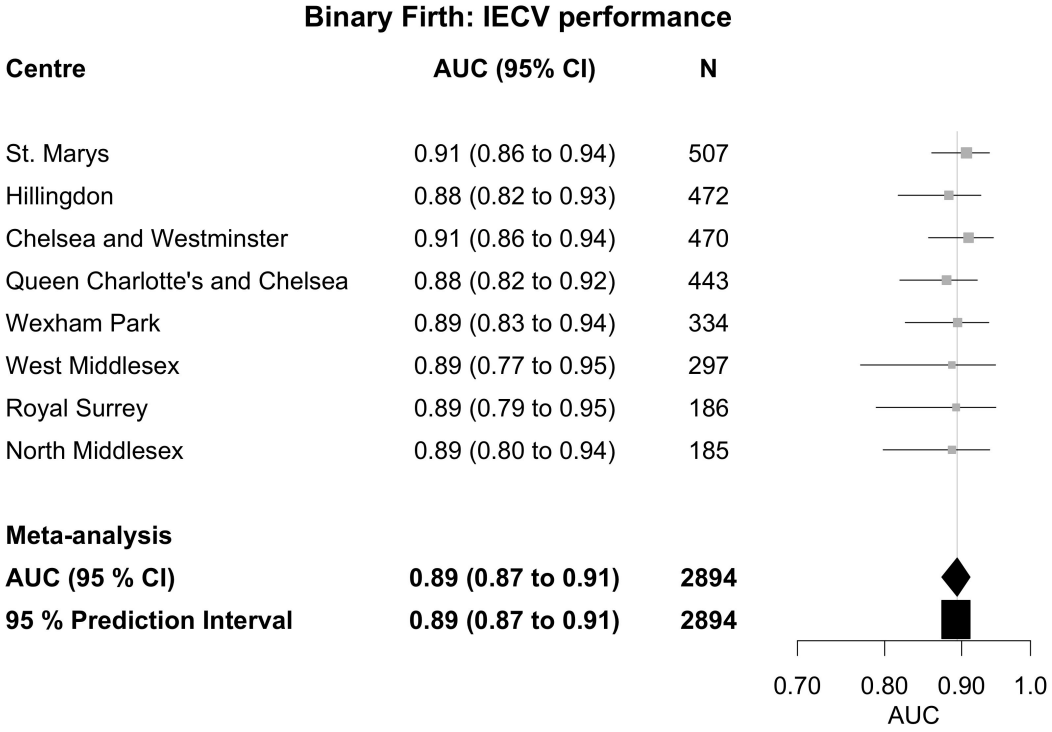


Figure 24: AUC per center for the binary Firth LR model

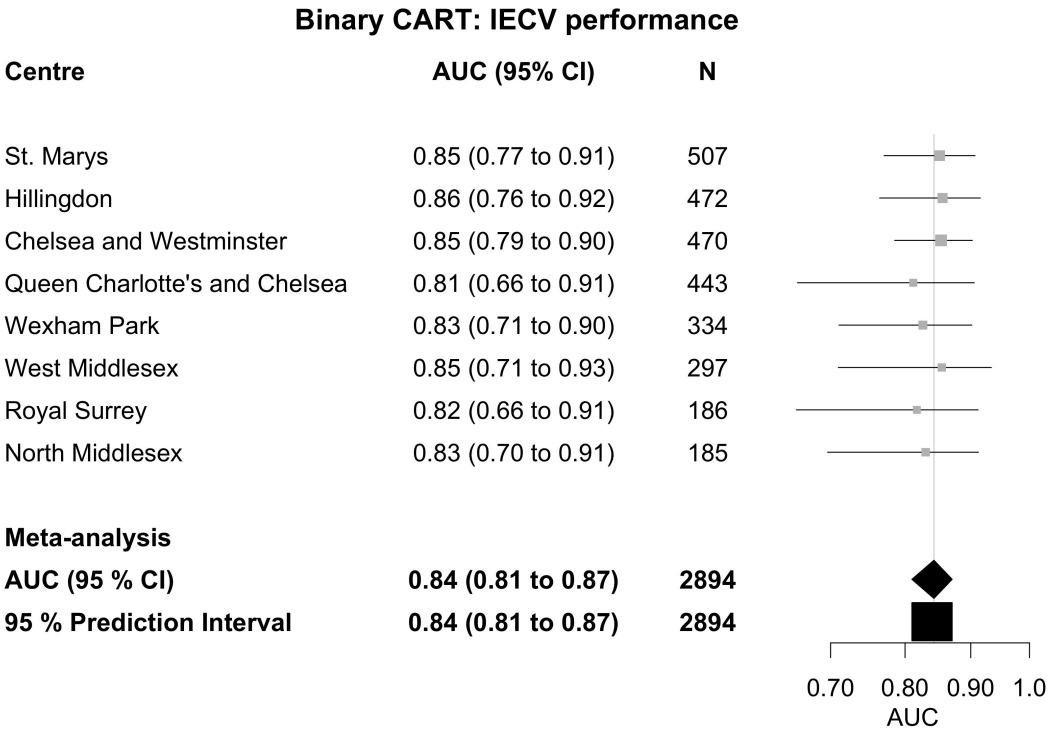


Figure 25: AUC per center for the binary CART model

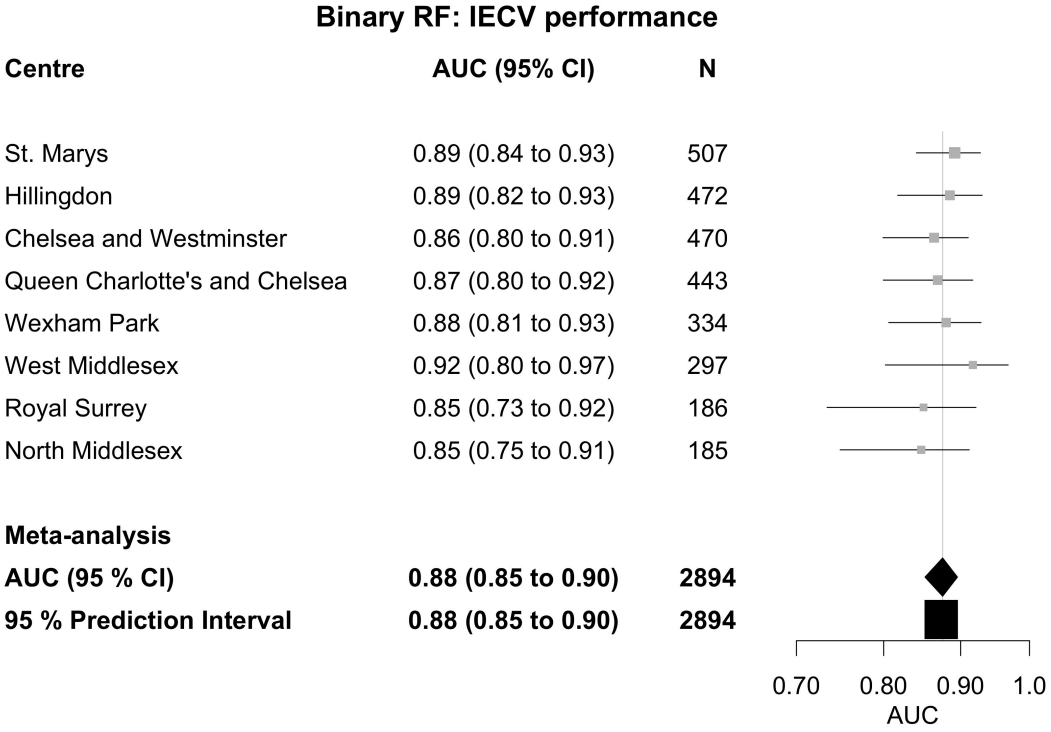


Figure 26: AUC per center for the binary RF model

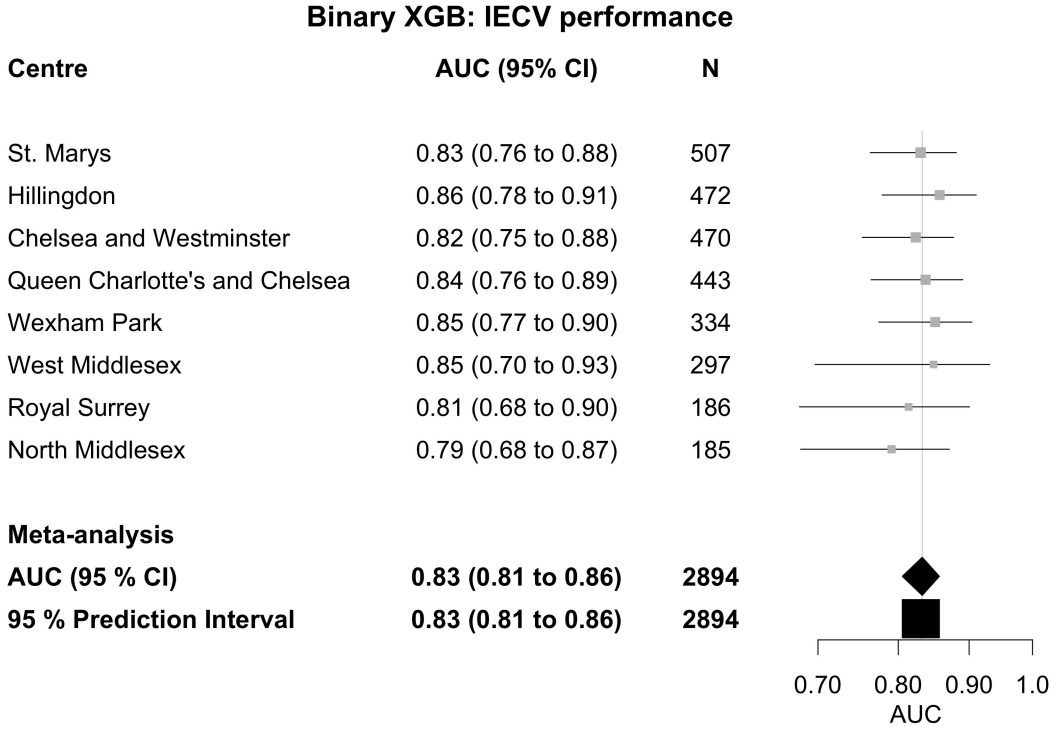


Figure 27: AUC per center for the binary XGB model

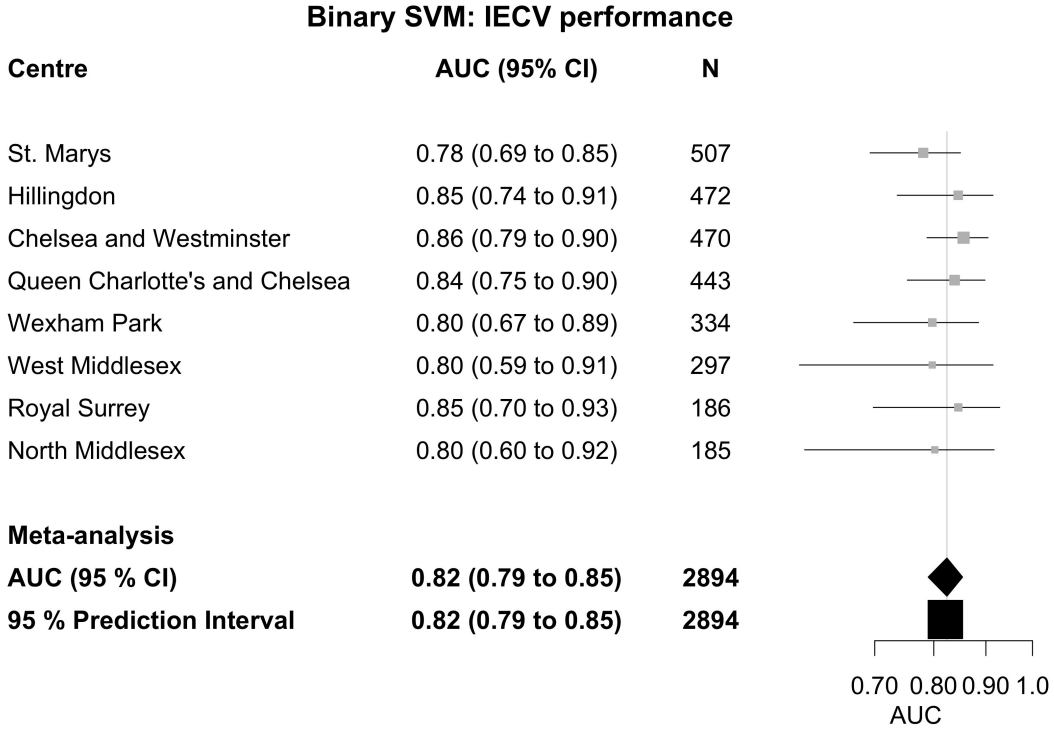


Figure 28: AUC per center for the binary SVM model

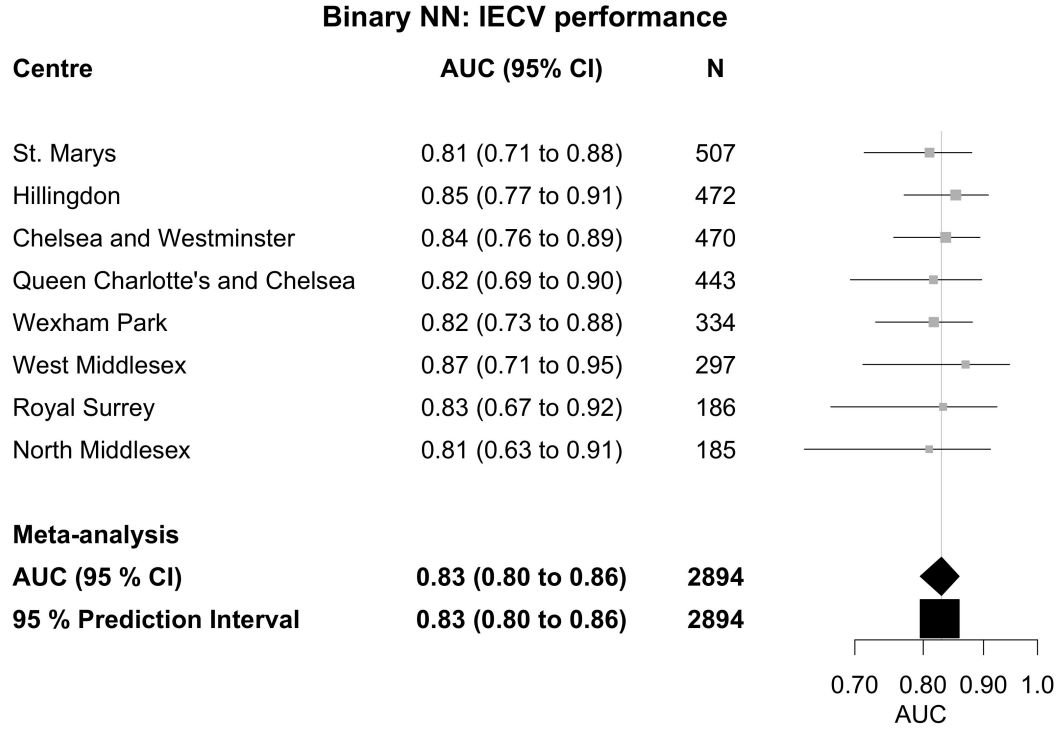


Figure 29: AUC per center for the binary NN model

Binary Models - Calibration

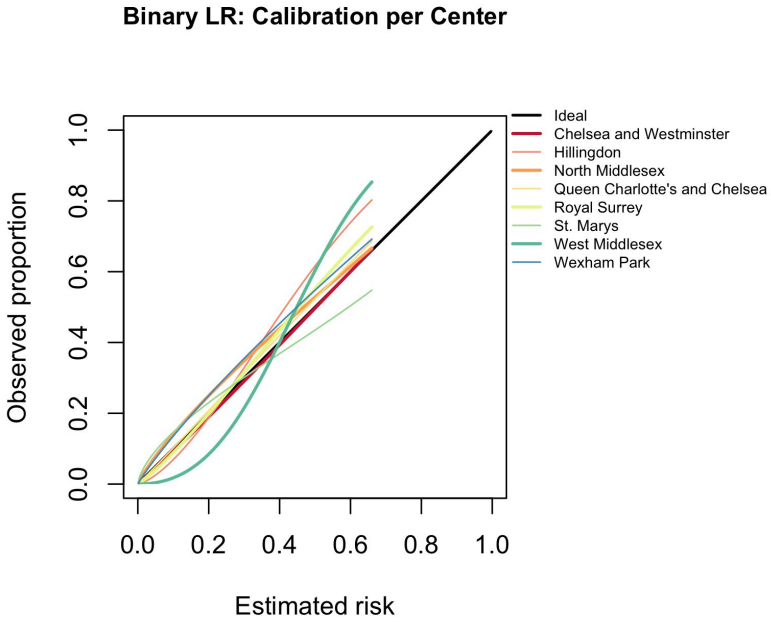


Figure 30: Calibration per center for the binary LR model

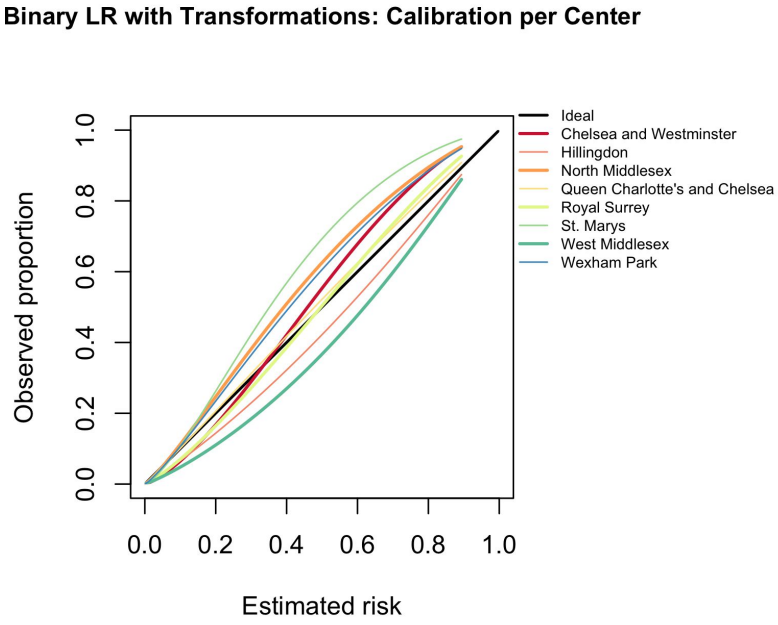


Figure 31: Calibration per center for the binary LR model with Transformations

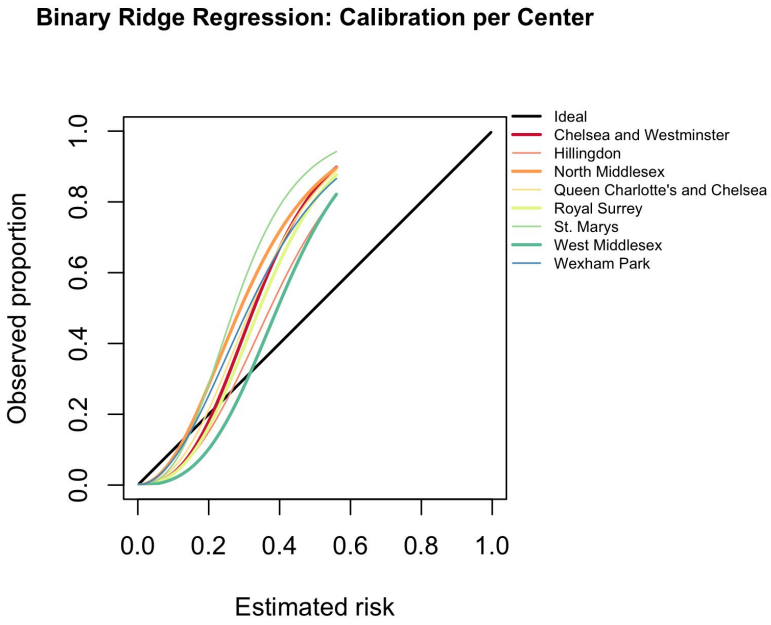


Figure 32: Calibration per center for the binary Ridge Regression model

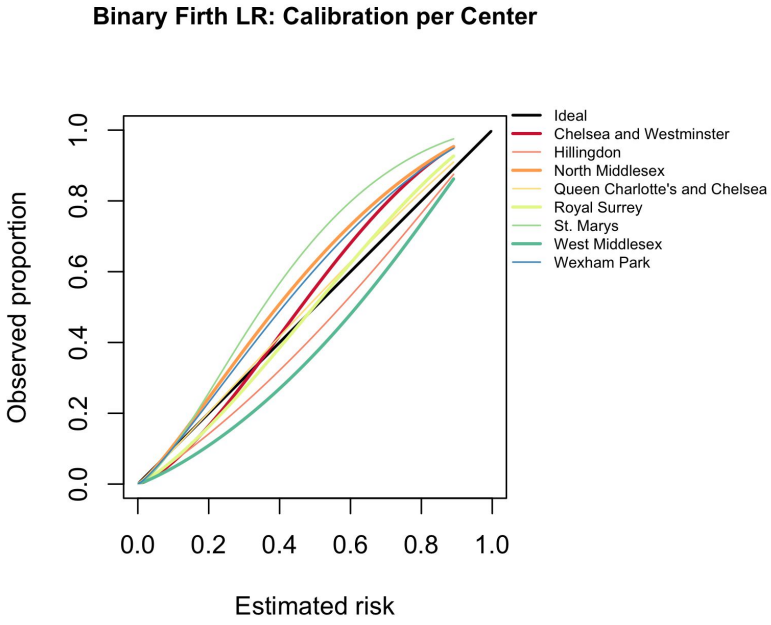


Figure 33: Calibration per center for the binary Firth LR model

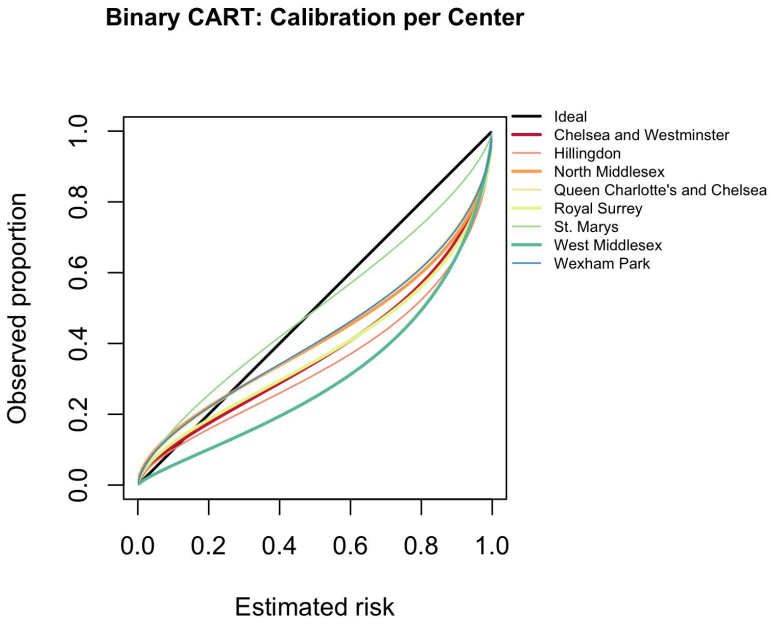


Figure 34: Calibration per center for the binary CART model

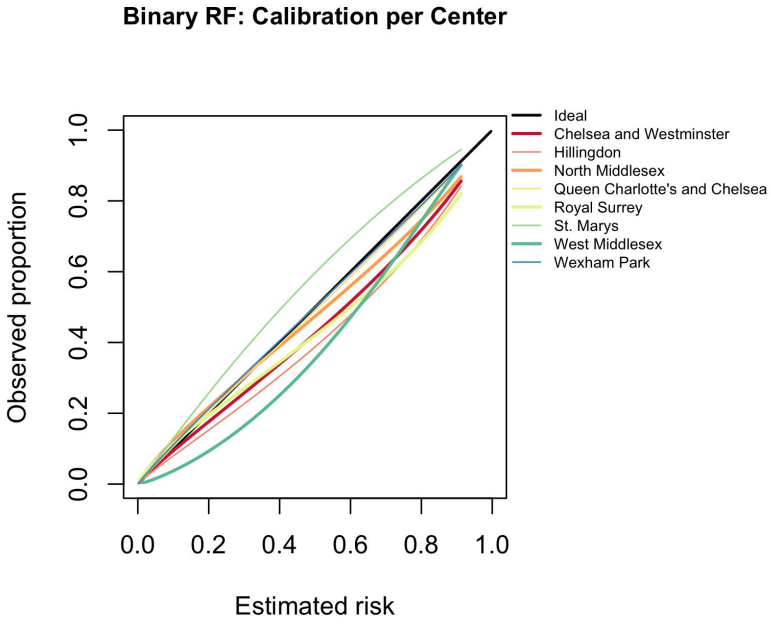


Figure 35: Calibration per center for the binary RF model

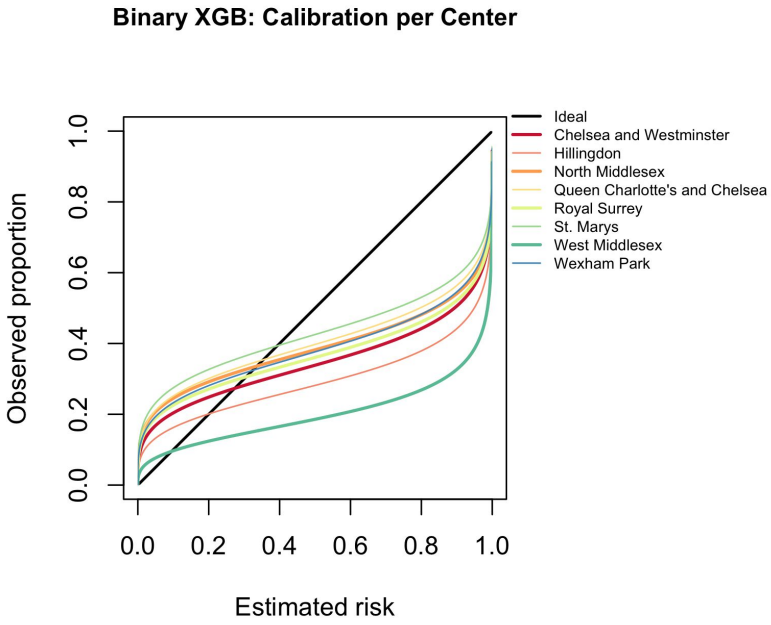


Figure 36: Calibration per center for the binary XGB model

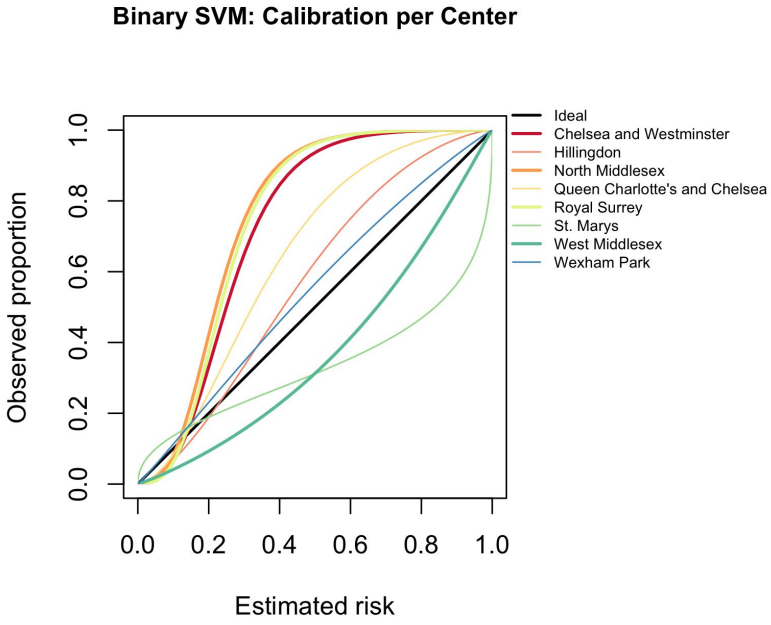


Figure 37: Calibration per center for the binary SVM model

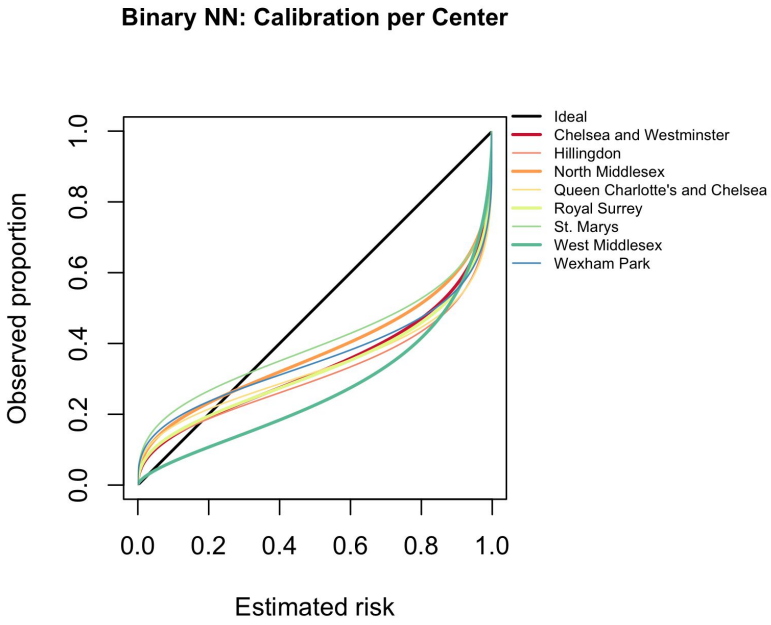


Figure 38: Calibration per center for the binary NN model

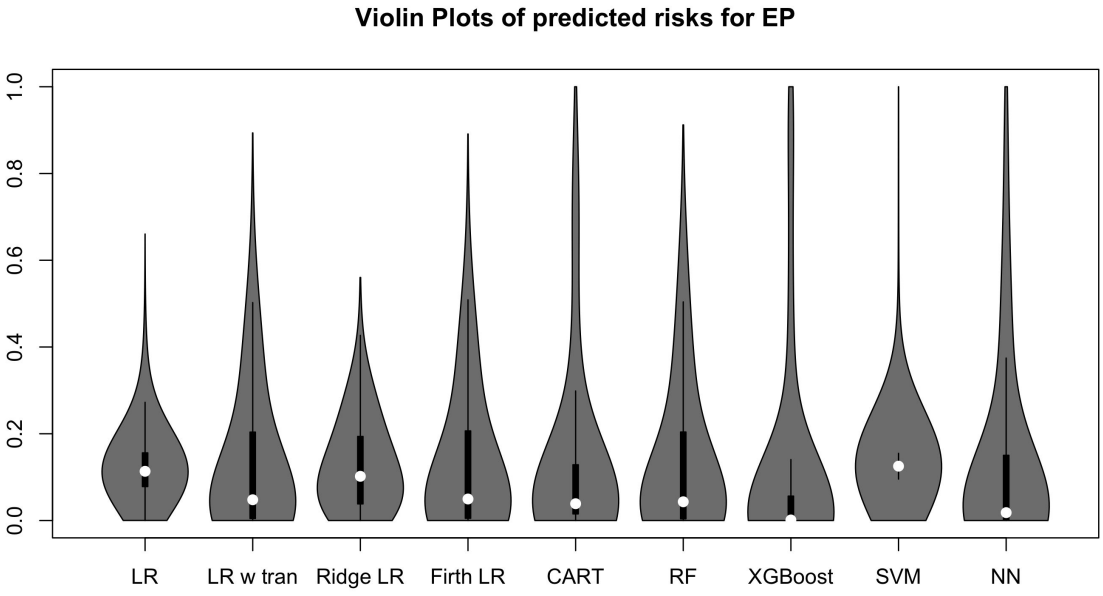


Figure 39: Violin plot of predicted risks for binary models

Multinomial Models - Discrimination

Multinomial LR: IECV cEP

Centre	AUC (95% CI)	N
St. Marys	0.87 (0.81 to 0.91)	507
Hillingdon	0.88 (0.82 to 0.93)	472
Chelsea and Westminster	0.87 (0.81 to 0.92)	470
Queen Charlotte's and Chelsea	0.86 (0.80 to 0.91)	443
Wexham Park	0.87 (0.80 to 0.92)	334
West Middlesex	0.92 (0.81 to 0.97)	297
Royal Surrey	0.87 (0.77 to 0.94)	186
North Middlesex	0.86 (0.77 to 0.92)	185
Meta-analysis		
cEP (95 % CI)	**0.87 (0.85 to 0.89)**	**2894**
95 % Prediction Interval	**0.87 (0.85 to 0.89)**	**2894**

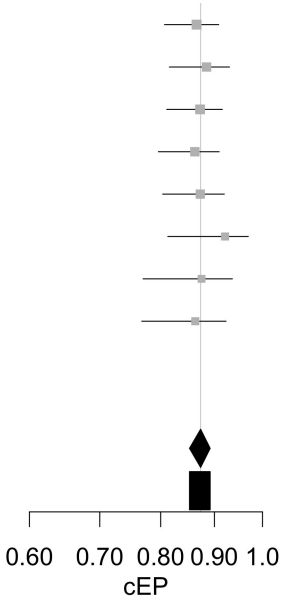


Figure 40: cEP per center for the multinomial LR model

Multinomial LR with Transformations: IECV cEP

Centre	AUC (95% CI)	N
St. Marys	0.92 (0.88 to 0.95)	507
Hillingdon	0.89 (0.82 to 0.93)	472
Chelsea and Westminster	0.90 (0.85 to 0.94)	470
Queen Charlotte's and Chelsea	0.88 (0.82 to 0.93)	443
Wexham Park	0.90 (0.83 to 0.94)	334
West Middlesex	0.89 (0.77 to 0.95)	297
Royal Surrey	0.90 (0.80 to 0.95)	186
North Middlesex	0.87 (0.78 to 0.93)	185
Meta-analysis		
cEP (95 % CI)	**0.90 (0.88 to 0.91)**	**2894**
95 % Prediction Interval	**0.90 (0.88 to 0.91)**	**2894**

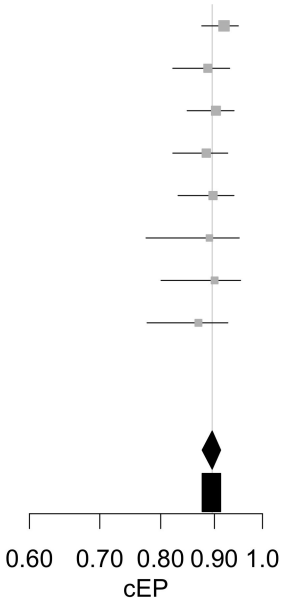


Figure 41: cEP per center for the multinomial LR with transformations

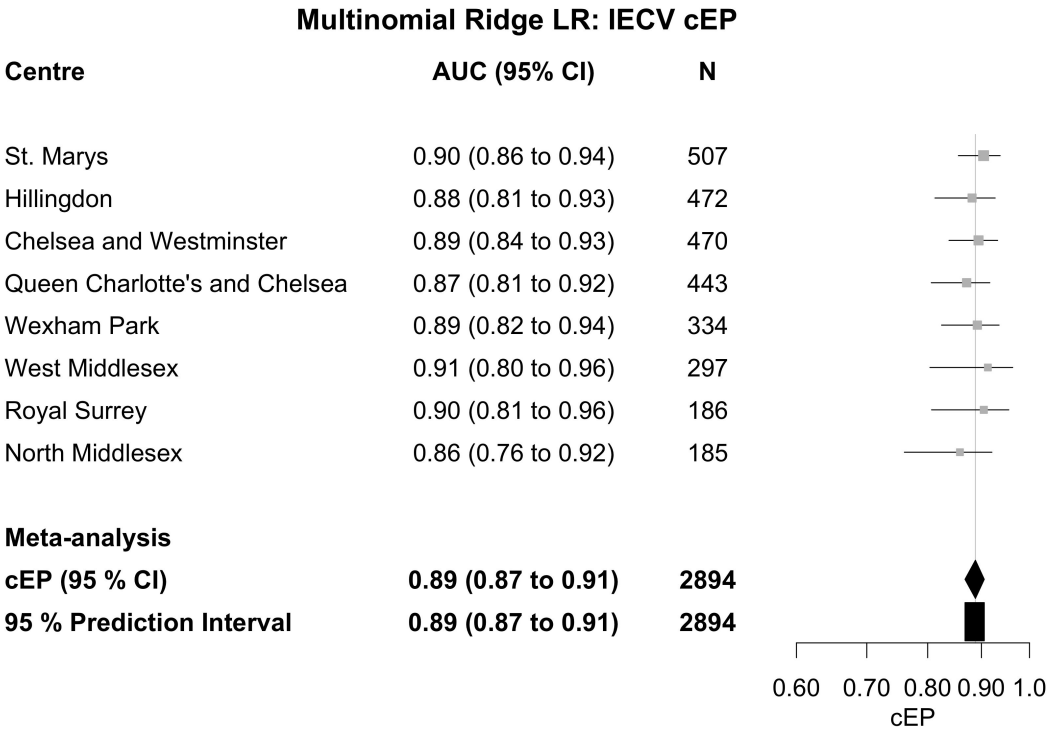


Figure 42: cEP per center for the multinomial Ridge LR model

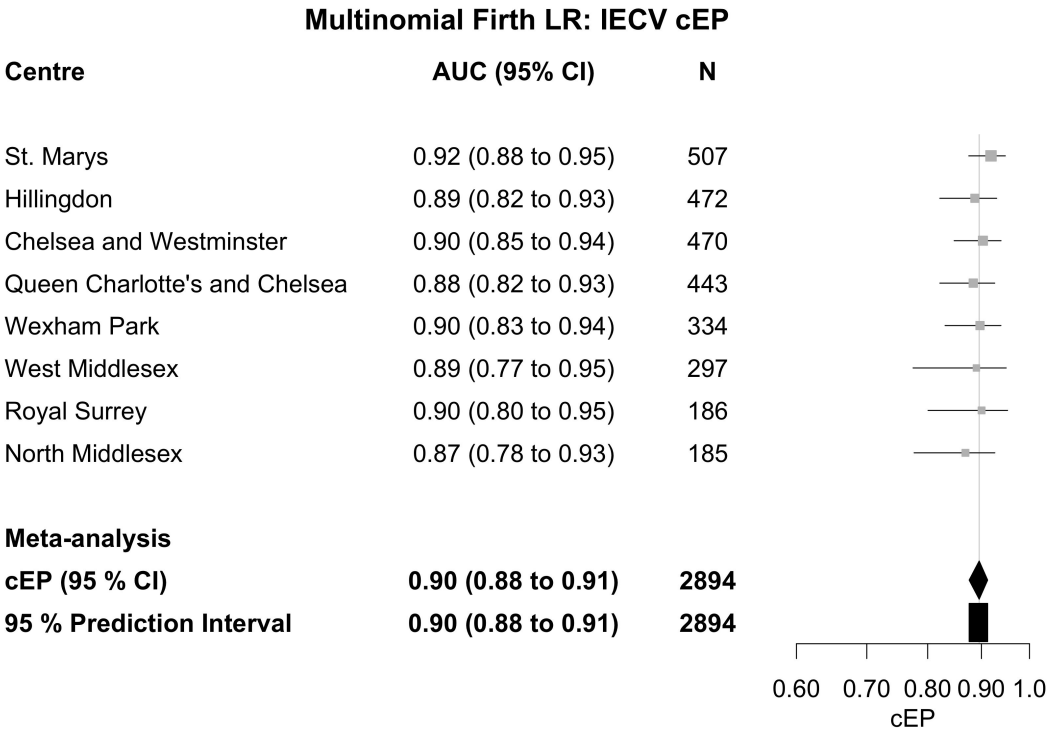


Figure 43: cEP per center for the multinomial Firth LR model

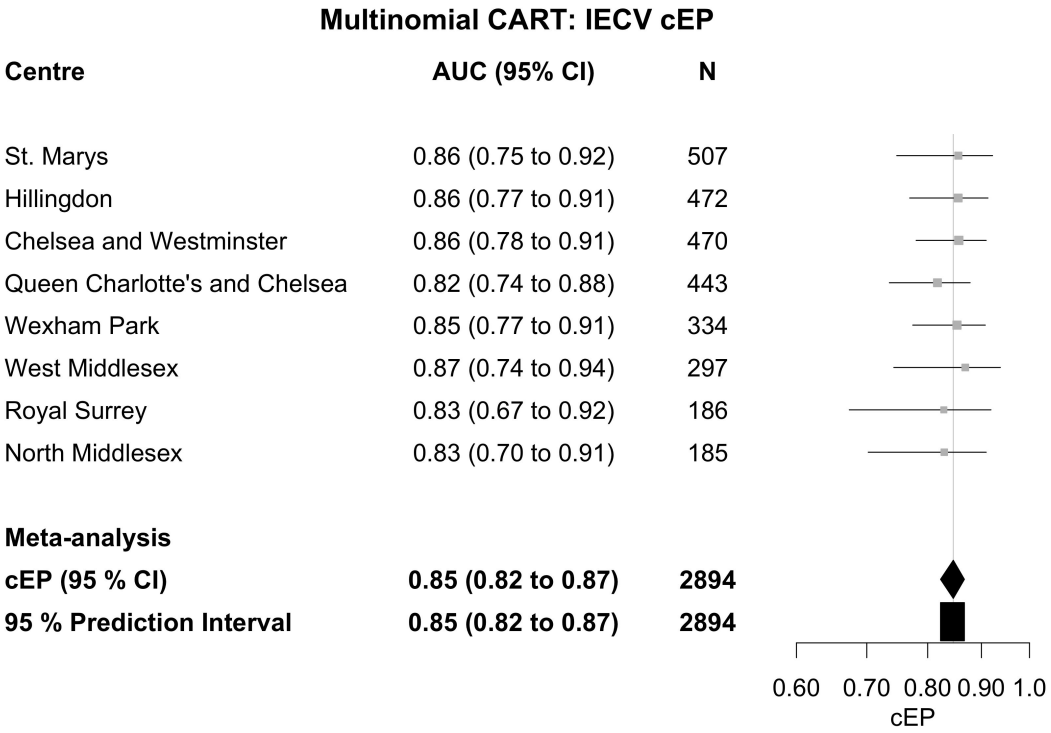


Figure 44: cEP per center for the multinomial CART

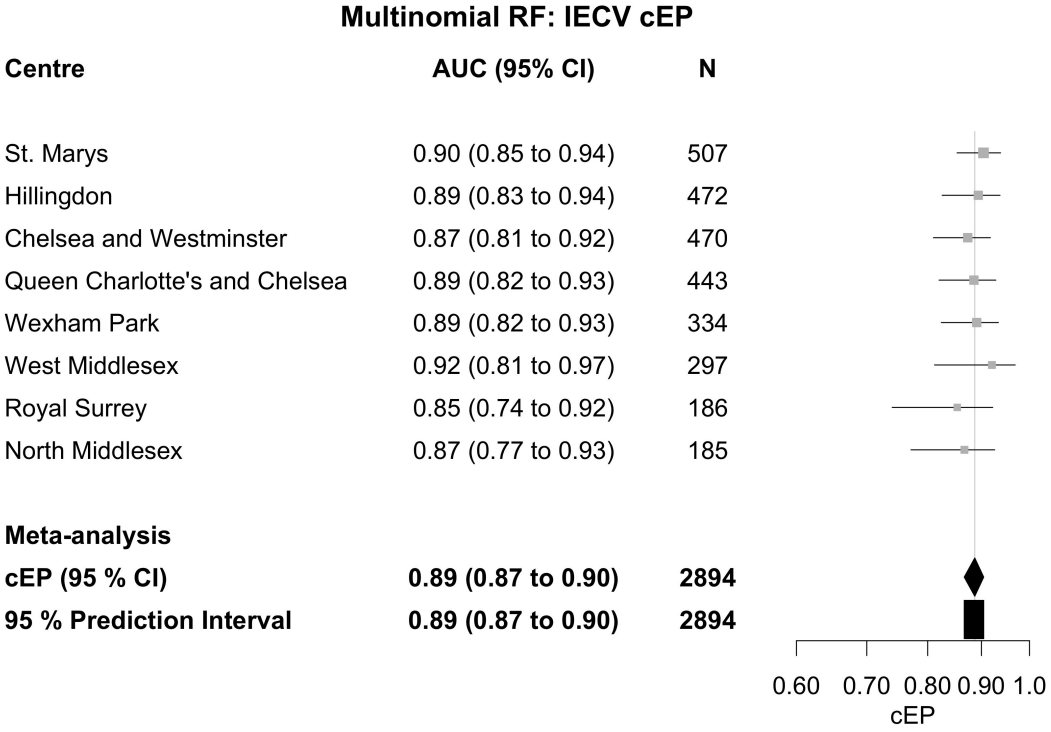


Figure 45: cEP per center for the multinomial RF

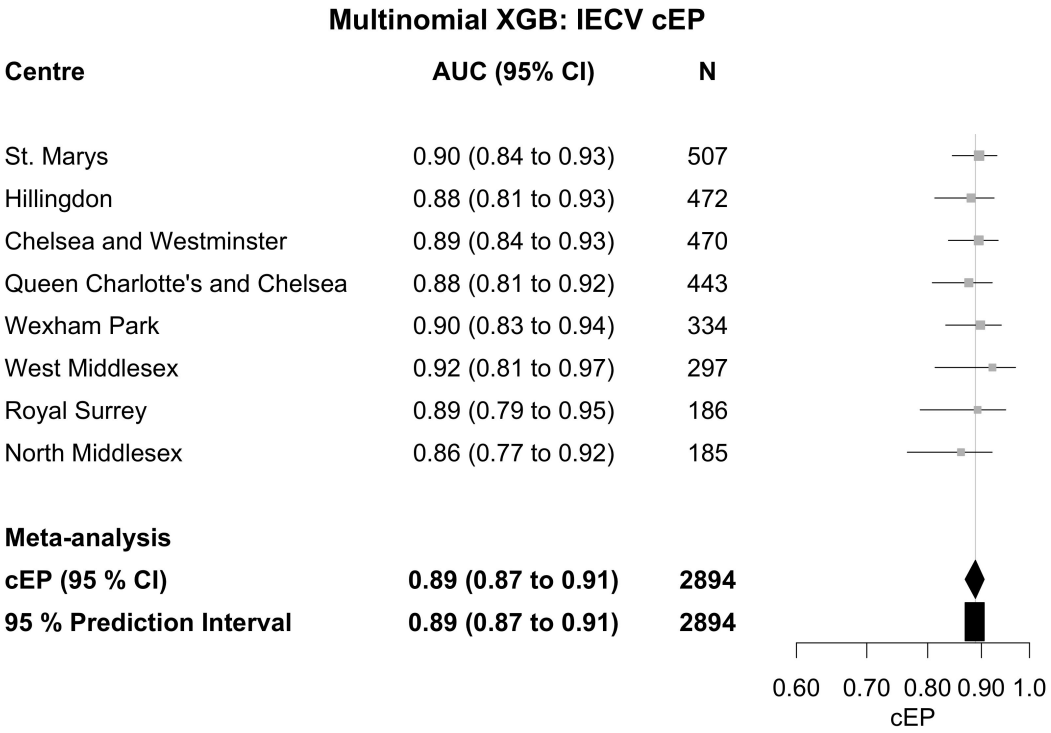


Figure 46: cEP per center for the multinomial XGB

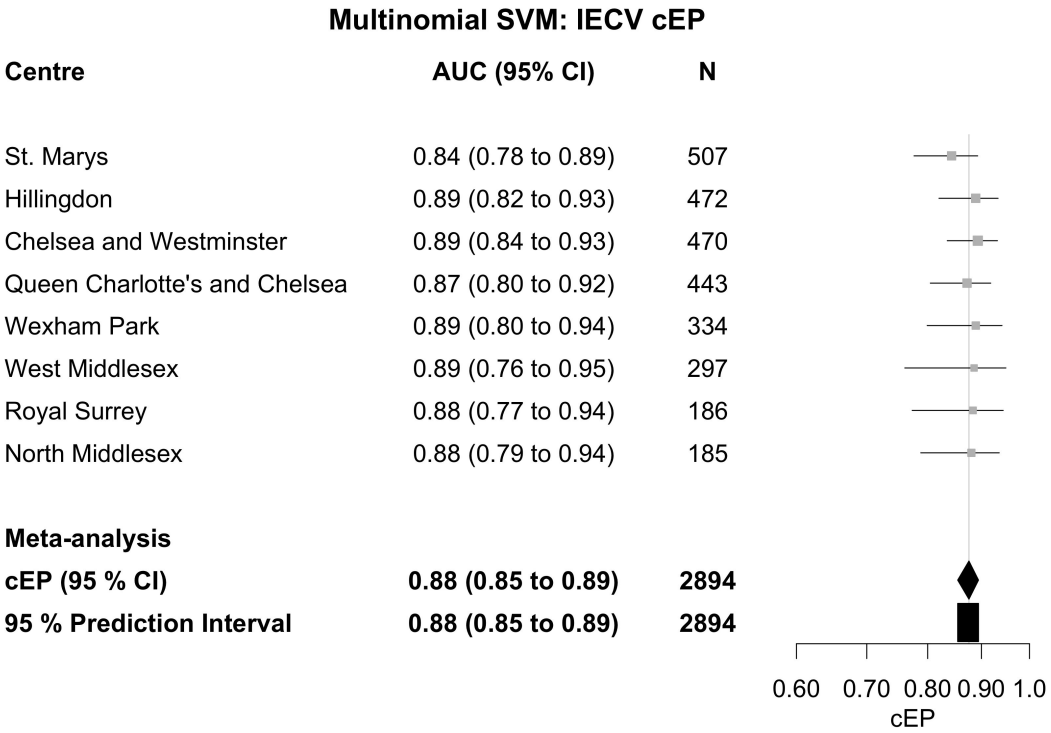


Figure 47: cEP per center for the multinomial SVM

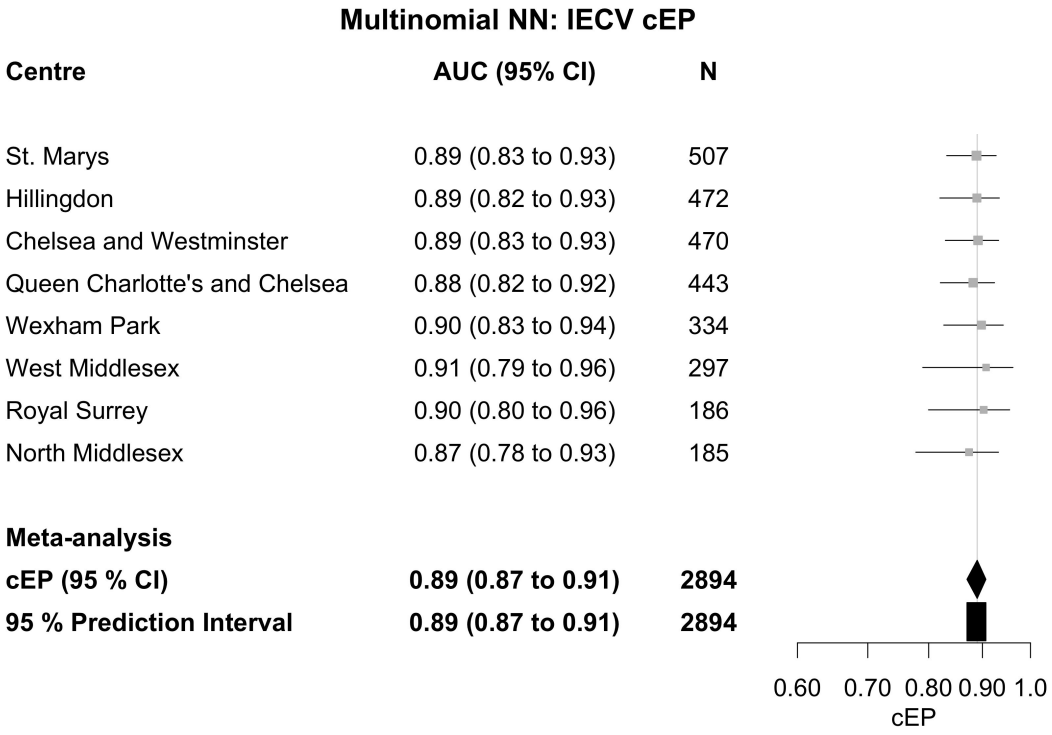


Figure 48: cEP per center for the multinomial NN

Multinomial Models - Calibration

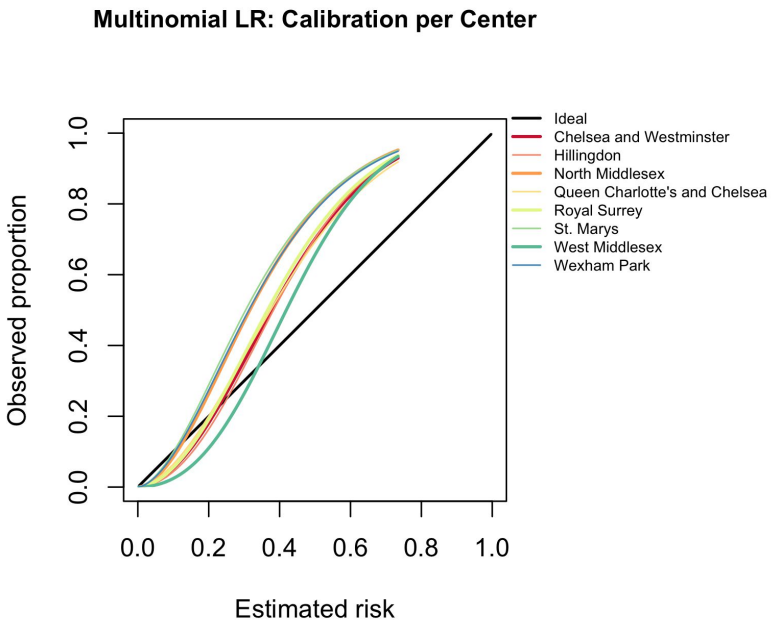


Figure 49: Calibration per center for the multinomial LR model

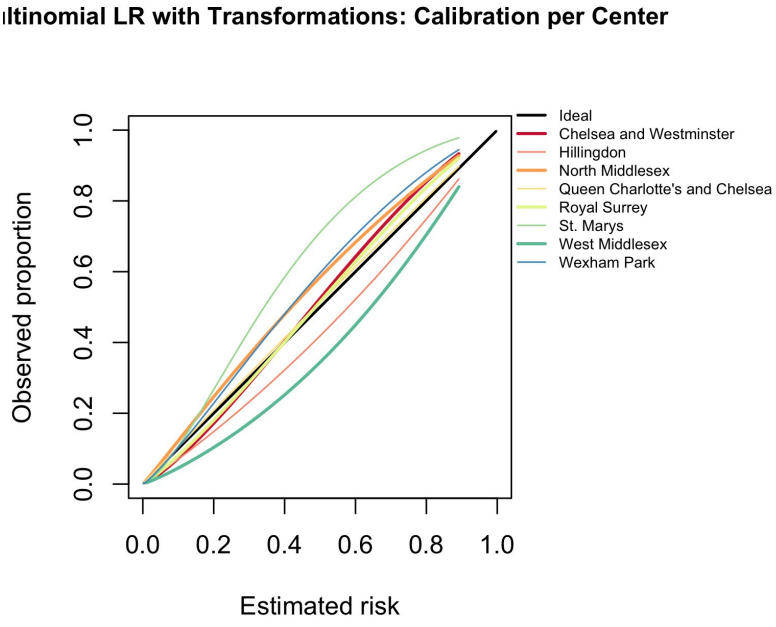


Figure 50: Calibration per center for the multinomial LR model with transformations

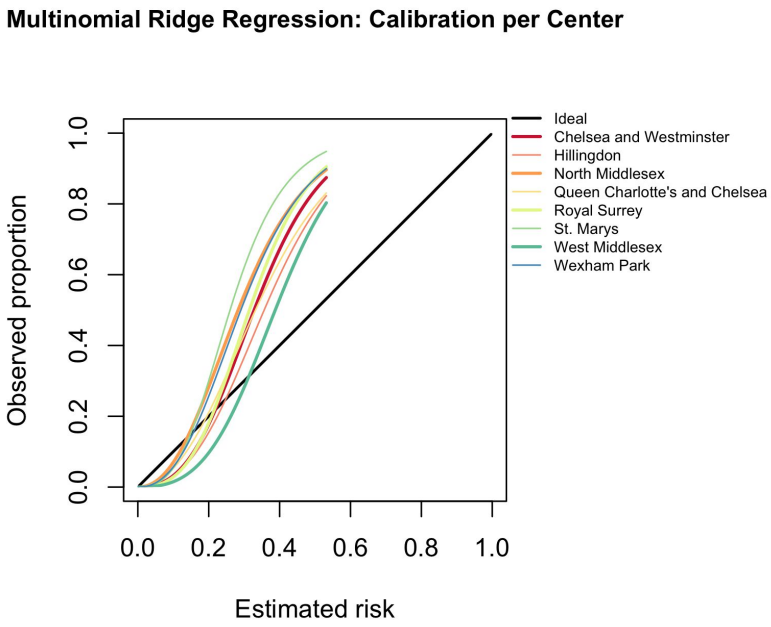


Figure 51: Calibration per center for the multinomial Ridge LR

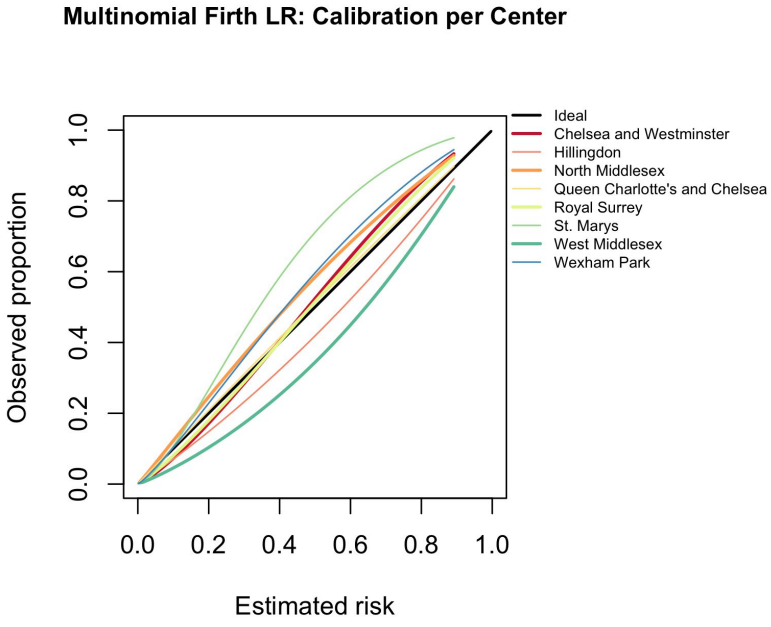


Figure 52: Calibration per center for the multinomial Firth LR

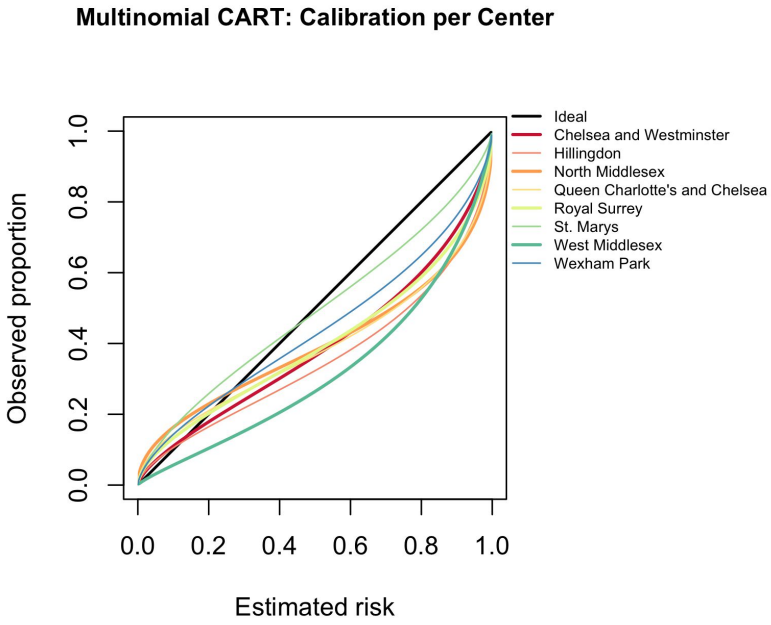


Figure 53: Calibration per center for the multinomial CART

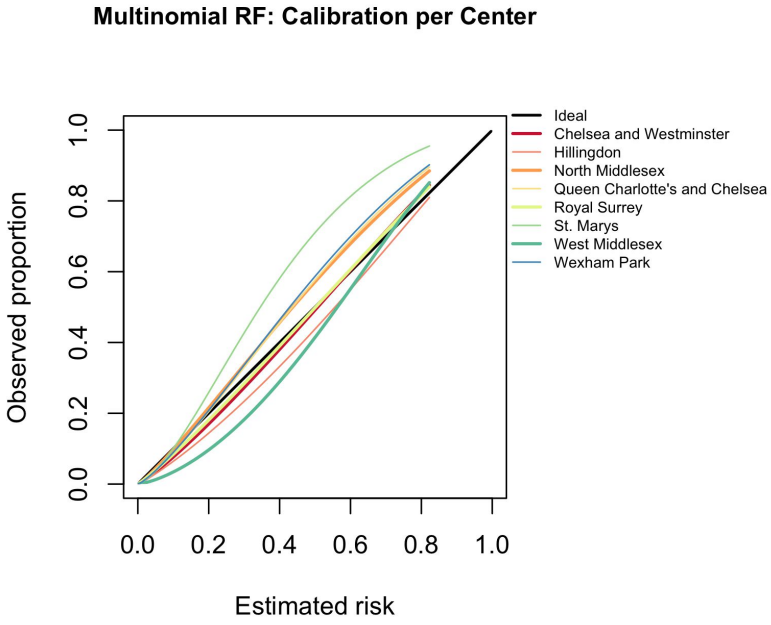


Figure 54: Calibration per center for the multinomial RF

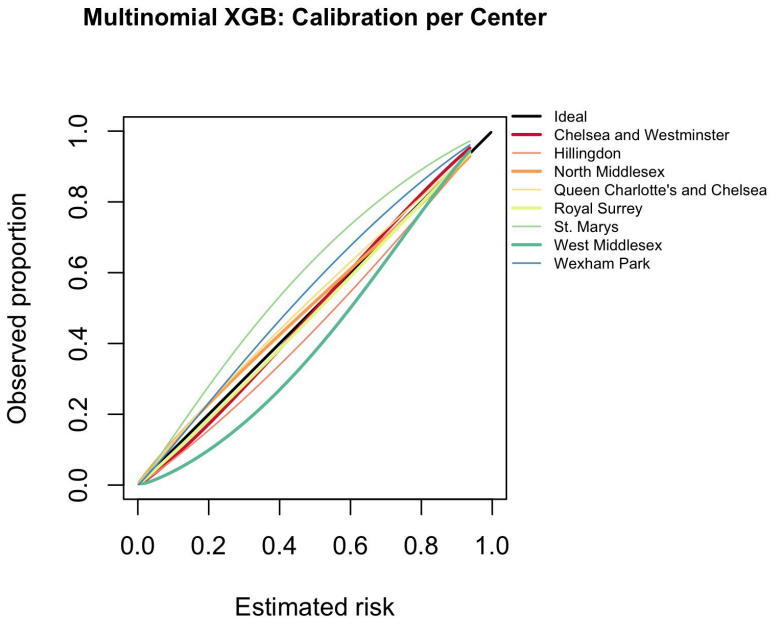


Figure 55: Calibration per center for the multinomial XGB

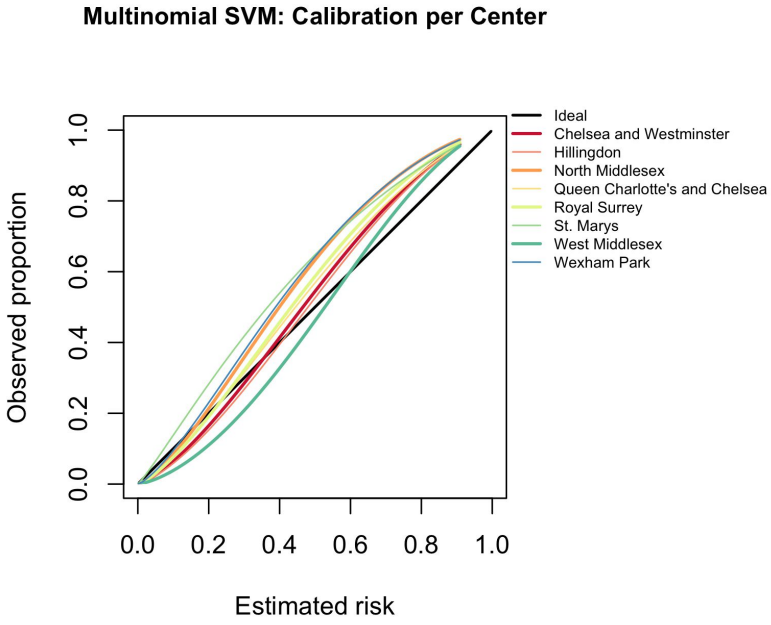


Figure 56: Calibration per center for the multinomial SVM

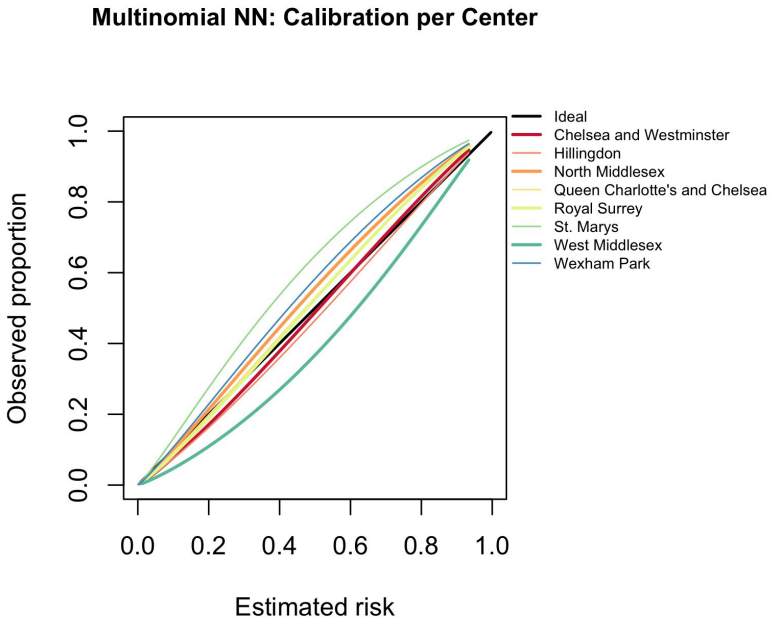


Figure 57: Calibration per center for the multinomial NN

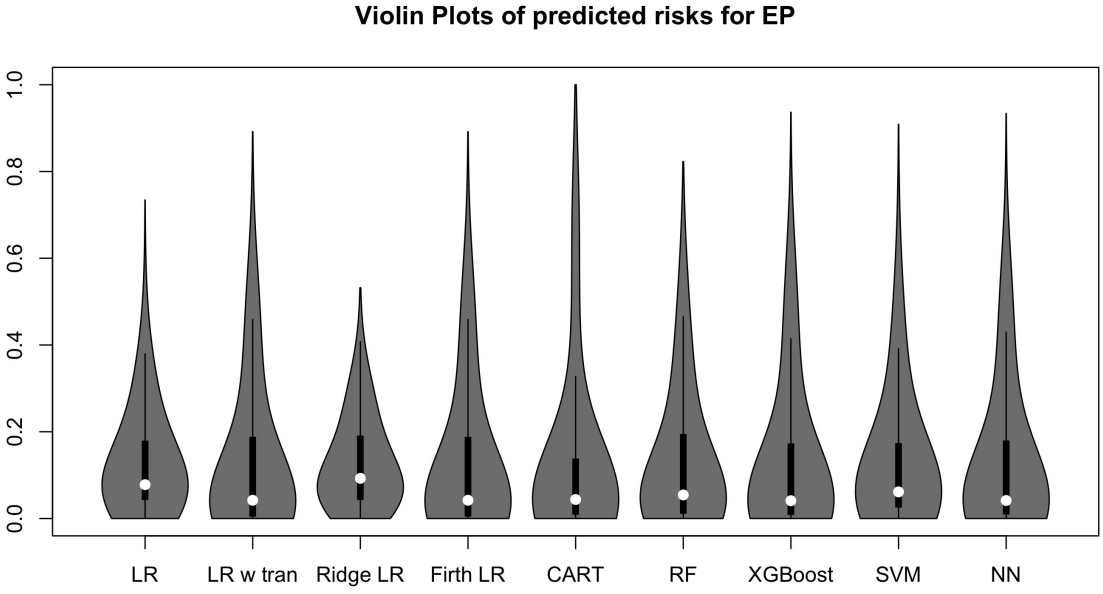


Figure 58: Violin plot of predicted risks for multinomial models

Code

The code for the analysis can be found in this GitHub repository.

Leuven Statistics Research Centre (LStat)
Celestijnenlaan 200 B
3001 HEVERLEE, BELGIË
tel. +32 16 32 88 75
https://lstat.kuleuven.be/