Supplementary Online Content

The consistency of a variety of machine learning and statistical models in predicting clinical risks of individual patients: A Longitudinal cohort study using cardiovascular disease as exemplar

eTable 1. Description of the machine learning and statistical models included in this study and the key parameters

eTable 2.1 - 4.2 Model performance comparison

- **eTable 2.1.** Performance indicators of machine learning and statistical models in overall cohort
- **eTable 2.2.** Performance indicators of machine learning and statistical models in overall cohort with logistic caret model as reference model
- **eTable 3.1.** Performance indicators of machine learning and statistical models in cohort without censoring with QRISK3 model as reference model
- **eTable 3.2.** Performance indicators of machine learning and statistical models in cohort without censoring with logistic caret model as reference model
- **eTable 4.1.** More performance indicators of machine learning and statistical models
- **eTable 4.2.** More performance indicators of machine learning and statistical models in cohort without censoring

eTable 5.1 – 5.4 Comparison of individual risk predictions

- **eTable 5.1**. Comparison of individual risk predictions of machine learning and statistical models in overall cohort and cohort without censoring
- **eTable 5.2**. Comparison of individual risk predictions of machine learning and statistical models in overall cohort (with as reference the risk predictions of the QRISK3)
- **eTable 5.3**. Comparison of individual risk predictions of machine learning and statistical models in overall cohort (with as reference the risk predictions of the logistic Caret model)
- **eTable 5.4**. Comparison of the individual risk predictions of machine learning and statistical models in cohort without censoring (with as reference the risk predictions of the QRISK3 model)
- **eTable 5.5**: Comparison of the individual risk predictions of machine learning and statistical models in cohort without censoring (with as reference the risk predictions of the Logistic Caret model)

eTable 6. Spearman correlations of machine learning models and statistical models in risk groups (logistic Caret predicted risk between 7%~8%)

eTable 7. Reclassification of individual risk predictions of machine learning and statistical models with 10% as threshold

eTable 8. Reclassification of individual risk predictions of Caret neural network models with different hyperparameters

eTable 9. Inconsistency of individual risk prediction between machine learning models derived from overall cohort and cohort without censoring

eTable 10. Performance indicators of machine learning and statistical models developed in South and validated in North England

eTable 11. Performance indicators of machine learning and statistical models with lower number of predictors

eFigure 1. Flow chart of sample splitting and model fitting process

eFigure 2.1 – 2.2 Comparison of individual risk predictions

- **eFigure 2.1**: Distribution of individual risk predictions with machine learning and statistical models in overall cohort for patients with predicted CVD risks of 7%~8% in the logistic Caret model

  X axis: predicted CVD risk

  Y axis: relative frequency (estimated density value)

- **eFigure 2.2**: Distribution of individual risk predictions with machine learning and statistical models in cohort without censoring

  a. For patients with predicted CVD risks of 9.5%~10.5% in QRISK3
  b. For patients with predicted CVD risks of 7%~8% in the logistic Caret model

  X axis: predicted CVD risk

  Y axis: relative frequency (estimated density value)

eFigure 3.1 - 4.6 Calibration of models

- **eFigure 3.1** Calibration slope of machine learning models and statistical models in overall cohort in binary framework (Observed events did not consider censoring)

- **eFigure 3.2.** Calibration slope of machine learning models and statistical models in cohort without censoring
  a. Survival framework
  b. Binary framework

- **eFigure 4.1**. Calibration plots in machine learning models of Caret in overall cohort and cohort without censoring

- **eFigure 4.2.** Calibration plots in statistical logistic models in overall cohort and cohort without censoring

- **eFigure 4.3.** Calibration plots in Cox proportional hazard models in overall cohort and cohort without censoring

- **eFigure 4.4.** Calibration plots in parametric survival models in overall cohort and cohort without censoring

- **eFigure 4.5.** Calibration plots in machine learning models of Sklearn in overall cohort and cohort without censoring

- **eFigure 4.6**. Calibration plots in machine learning models of h2o in overall cohort and cohort without censoring

eFigure 5.1 – 5.2: Inconsistency of individual risk predictions in overall models

- **eFigure 5.1.** 95% range of individual risk predictions with machine learning and statistical models stratified by deciles of predicted risks with Caret logistic model in cohort without censoring

- **eFigure 5.2.** 95% range of individual risk predictions with machine learning and statistical models stratified by deciles of predicted risks with Caret logistic model

a. Overall cohort
b. Cohort without censoring

eFigure 6.1 – 6.9: Bland–Altman analysis compare two models

X axis: Average = (model A + model B) / 2

Y axis: Difference = model B – model A

- **eFigure 6.1 Logistic model fitted with machine learning framework (Caret) comparing to Logistic model fitted statistically**
- **eFigure 6.2 QRISK3 comparing to Logistic model Caret**
- **eFigure 6.3 QRISK3 comparing to Random forest Caret**
- **eFigure 6.4 QRISK3 comparing to Parametric Weibull model**
- **eFigure 6.5 QRISK3 comparing to Local Cox model**
- **eFigure 6.6 Local Cox model comparing to Parametric Weibull model**
- **eFigure 6.7 Gradient Boosting Classifier (GBC) sklearn comparing to Random forest h2o**
- **eFigure 6.8 Parametric Gaussian model comparing to Neural network h2o**
- **eFigure 6.9 Logistic model Caret comparing to Random forest Sklearn**

eFigure 7: Inconsistency of individual risk predictions with machine learning and statistical models with Fieller's 95% confidence interval (each dot corresponds to an individual prediction; a random sample of these is displayed with red line enclosing 95% of the observations)

a. Overall cohort
b. Cohort without censoring

X axis: QRISK3 predicted CVD risk

Y axis: predicted CVD risk by other models

eFigure 8 – 10 sensitivity analysis of inconsistency of individual risk predictions

- **eFigure 8**. 95% range of individual risk predictions with Caret neural network models with different grid searched best hyperparameters stratified by deciles of predicted risks with models with the most frequent selected hyperparameters
- **eFigure 9**. Distribution of individual risk predictions with machine learning and statistical models developed in practices from South and tested in practices from North England
- **eFigure 10**. Distribution of individual risk predictions with machine learning and statistical models developed with predictors of age and sex plus 1/3, 1/2, 2/3 of all predictors

eFigure 11. Distribution of age among removed patients due to censoring (death patients excluded)

eTable 1 describes the 19 model families used in main study and selected hyperparameters in grid search process.

eTable 2.1 - 4.2 Model performance comparison

Model performance were calculated using a threshold of 7.5% for all models. The threshold was selected according to ACC/AHA Guideline on the Assessment of Cardiovascular Risk and used in other machine learning studies

eTable 2.1-2.2 shows the model performance among all models with QRISK3 or logistic model Caret as reference. Most models had similar model performance.

eTable 3.1-3.2 shows the model performance of machine learning and statistical models in the cohort without censoring. Though all models generally had a lower C-statistic than the models from the cohort with censoring (Table 2/eTable 2.1 in the main manuscript), the performance of these models was comparable.

eTable 4.1-4.2 shows more model performance measures including F1 score, balanced accuracy, negative predictive value (NPV) and specificity with threshold as 7.5% in binary framework in overall cohort and cohort without censoring. In general, all models had a few slightly better measures than others but also had a few slightly worse measures than other models. This was because these measures are a trade-off and being influenced by the selected threshold (i.e. a different threshold say 10% rather than 7.5% would change values of these measures).

eTable 5.1 – 5.5 Comparison of individual risk predictions

eTable 5.1-5.5 shows the inconsistencies of range of individual CVD risk predictions for different strata of predictions with the QRISK3 or logistic Caret model as reference in the overall cohort and cohort without censoring. eTable 5.1 shows the overall inconsistencies. Logistic models and machine learning models which ignore censoring substantially underestimated patient risks (eTable 5.2), predictions for same individual patients varied substantially (eTable 5.3). Removing patients with censoring makes models overestimated patients risk compared to QRISK3 (eTable 5.4) and it did not change the magnitude of inconsistency of individual risk prediction (eTable 5.5).

eTable 6 shows the low correlation between individual risk prediction among different machine learning models. The results were consistent with the eFigure 2.1: machine learning models with similar model performance predicted the same patients differently.

eTable 7 is a similar reclassification table as Table 3 in the main manuscript except using 10% rather than 7.5% threshold. Similar reclassification was found as in the main study. Of the 735,474 patients with a CVD risk ≤10%, 10% were reclassified when using another model. Of the 180,005 patients with a CVD risk >10%, 62.9% were reclassified when using another model.

eTable 8 shows the reclassification effects of choosing different hyperparameters for the same machine learning model family on individual risk prediction. Neural network Caret with hyperparameters of size (number of neurons) and decay (parameter to control the overfitting) was used as an exemplar. From 100 best models grid searched from random samples, there were 17 groups of best selected hyperparameters. Using the average risk predictions from the most frequent group as reference (in this case the biggest model group has 17 neural network models with size=3 and decay=3.5 from grid search process), risk predictions of the same patients from the same model with different best hyperparameters were compared in **eFigure 8**. The reclassification **eTable 8** shows that among 129,991 patients over threshold 7.5%, 11% of them would be reclassified if a different best hyperparameter was chosen. However, in the main study, the variation of individual risk predictions within the same model family was decreased by model ensembling with soft voting (averaging). This additional analysis shows the inconsistency of individual risk prediction among machine learning models could be worse considering variation of individual risk prediction among the same model family, and current approach to find the best hyperparameters is data-driven and in lack of a principal way to determine what hyperparameters were more proper before fitting the model.

eTable 9 used the machine learning models from the cohort without censoring to calculate risk for the full cohort (the cohort with censoring), and then the risk prediction of the same model of the same risk group of patients was compared. Even within the same model, the risk predictions for the same patients did not agree to each other. Machine learning models derived from a cohort without censoring predicted larger range of risk than the same machine learning models derived from a cohort with censoring on the same patients. The same applied to the fitted statistical models, as both models were fitted on a biased cohort, i.e. patients censoring were artificially removed. This indicates that models developed from a censored removed cohort (this has been done in several machine learning papers) should not be used in a cohort with censoring, as ignoring censoring introduces bias (mis-calibration) to individual risk prediction.

eTable 10 shows the model performance of models developed from practices from South England and validated in practices from North England. It also shows similar inconsistency of individual risk prediction was found as in main manuscript (**eFigure**

9 comparing to Figure 1 in main manuscript). As expected, models have similar model performance in the North and South England, which is consistent to the main study. However, models developed from practices from South generally have lower model performance in practices from North English practices compared to practices from South England in either machine learning models or statistical models. Previous study using the random effects model showed that there was practice variability among UK practices with an effect on individual risk prediction [1]. In this study, this sensitive analysis showed that machine learning models did not automatically capture this variability in coding.

eTable 11 compares the model performance of models with age and sex plus 1/3, 1/2 and 2/3 randomly sampled predictors. Except random forest caret, all models have similar model performance among each other as in the main study and similar inconsistency of individual risk prediction was found as in main manuscript (**eFigure 10**). Random forest Caret model was underfitted with 1/3 predictors as the final model grid searched the best parameter mtry (number of predictors used to grow branches) as the same number of available predictors. As expected, model performance of random forest Caret improved with the increase of number of predictors. Random forest Caret model has the similar model performance as other models with full predictors indicate enough predictors were considered in the study.

eFigure 1 visualises the workflow of sample splitting and model fitting process.

eFigure 2.1 – 2.2 Comparison of individual risk predictions

eFigure 2.1 is a similar graph as Figure 1 in the main manuscript except it uses logistic Caret model as reference and assessing patients with CVD predicted risk of 7%~8%. Similar finding is that models with similar model performance predicts the same individual patients differently. **eFigure 2.2** plotted similar distribution as Figure1 and eFigure 2.1 in cohort without censoring. Multiple models fitted from the cohort without censoring substantially overestimated patients' risk compared to QRISK3. The removal of censored patients changed the magnitude but not the variability of individual CVD risk predictions.

eFigure 3.1 - 4.6 Calibration of models

Figure 2 (in main manuscript) **and eFigure 3.1 -3.2** shows the calibration slopes of all machine learning models and statistical models in overall cohort and cohort without censoring considering survival or binary framework. It shows that both models were well calibrated in their own framework (i.e. Survival models in survival framework and logistic models and machine learning models in binary framework). However, Logistic models and machine learning models ignoring censoring were mis-calibrated (Figure 2) in survival framework (i.e. observed events considering effects of censoring). It appears these models were well

calibrated in both framework in cohort without censoring, this is because survival framework and binary framework were similar once patients with censoring were removed. However, artificially removing patients with censoring makes the cohort non-representative as censoring often occurs over time. **eFigure 2.2a** has shown that logistic models and machine learning models developed from cohort without censoring over-estimated patient risks in overall cohort compared to QRISK3.

eFigure 4.1 – 4.6 shows the calibration plots of all machine learning models and statistical models. Machine learning models have good calibration in a binary framework (i.e. treating the patients with censoring as non-events) irrespective of the cohort with censoring or cohort without censoring. However, the calibration figures of the cohort with censoring showed that machine learning models have poor calibration in the survival framework (i.e. considering the effects of censoring). Once censoring was removed, the calibration of machine learning models improved (shown in calibration plot from cohort without out censoring). This suggests that although machine learning models which ignore patient-censoring can have good calibration in a binary framework, it was poor calibrated in survival framework. Cox models including QRISK3 and Framingham have very good calibration in a survival framework but very poor calibration in a binary framework as they considered censoring with time-to-events outcome. However, censoring is very common with long-term risks and should not be ignored.

eFigure 5.1 – 5.2: Inconsistency of individual risk predictions in overall models

eFigure 5.1 – 5.2 is a similar plot as Figure 3 in main manuscript except eFigure 5.1 results from cohort without censoring and eFigure 5.2 uses logistic Caret model as reference rather than QRISK3. It shows similar finding that inconsistency of individual risk prediction among models were mainly in higher risk group patients.

eFigure 6.1 – 6.9: Bland–Altman analysis compare two models

eFigure 6.1 – 6.7 shows more pairs of model comparison in Bland-Altman analysis. Overall, it shows the similar inconsistency of individual risk prediction between two statistical comparable models in the full spectrum. Differences within model family (e.g. comparison of two machine learning models) are generally smaller than between model family (e.g. comparison of survival models and machine learning models) but still influence treatment decision of patients. **eFigure 6.1** compares logistic model fitted from different framework and could be used as a reference to show what would it look like if two models agree to each other, i.e. the differences of risk prediction of 95% patients would be near 0. **eFigure 6.2 – 6.3 and eFigure 6.8** shows similar inconsistency of individual risk prediction between model family as Figure 4 in main manuscript. **eFgiure 6.4 –**

6.7 shows though the inconsistency of individual risk prediction within the model family is smaller than between model family, it still has substantial influence for treatment decision of patients. The differences between QRISK3 and Local fitted Cox model might be due to the variation of data being used such as different calendar time (2015 comparing to 2018) between Qresearch and CPRD or variation of data quality such as missing value and different coding. The differences between Cox model and Parametric survival model might be due to the later one additionally assumes the survival time follows a known parametric distribution. The differences between machine learning models might be due to different architectures and selected best hyper-parameters. Still, all these results show similar finding that treatment decision for patients is arbitrary based on which model is used eventually in practice as statistical comparable models predict different risks for the same patients.

eFigure 7 shows the variation of individual risk predictions between QRISK3 and to those generated by the other models with Fieller's confidence interval. Similar to previous findings presented above, predictions for these same patients varied substantially between models.

eFigure 8 – 10 sensitivity analysis of inconsistency of individual risk predictions

The interpretation of **eFigure 8** could be found in the interpretation of **eTable8**

The interpretation of **eFigure 9** could be found in the interpretation of **eTable10**

The interpretation of **eFigure 10** could be found in the interpretation of **eTable11**

eFigure 11 shows that among patients who were censored (death excluded), younger patients were the majority. This indicates the reason that average age in the cohort without censoring is higher than the cohort with censoring is the effects of younger patients transferred out from practices that compensated the effects of older patients who died during the follow-up.

eTable 1: Description of the machine learning and statistical models included in this study and the key parameters

	Package description	Model description	Key parameters selected by analyst for grid search
Caret			
Logistic	Classification And REgression Training (Caret) is a R package which has a series of functions to create predictive models in a structural and organized way. It contains functions which can be used to split data, pre-process data, select predictors, tune model and resample [2]	Logistic model is a type of generalised linear model with a binary variable as outcome [3]	None
Random forest		Random forest is an ensemble machine learning model which combines the predictions from multiple decision-trees where each tree grows from an independent sample of predictors [4]	**mtry:** number of randomly selected predictors as candidates at each split **ntree:** number of trees
Neural network		Neural network is a machine learning model whose model-structure mimics the structure of animal brain using hidden layers and neurons in those hidden layer [5]	**size:** number of units in hidden layer (neural network in Caret only fits one hidden layer) **decay:** a regularization parameter to control over-fitting (higher decay means less chance of over-fitting)
Statistical logistic model			
Logistic model	Standard statistical way to fit logistic model with glm() function from basic R library [6]	Logistic model fitted with standard statistical approach	None
Cox proportional hazards model			
QRISK3	Effects of predictors on hazard ratio are assumed to be multiplicative. Cox models take into account censoring [7]	QRISK3 was derived from a UK cohort [8]	None
Framingham		Framingham model was derived from a US cohort [9]	None
Local Cox model		Re-fitted Cox model using the same training cohorts and validation cohorts as machine learning models.	None
Parametric survival model			
Weibull distribution	Parametric survival models are alternatives of Cox model in survival analysis. It assumes survival time follows a known parametric distribution (e.g. Weibull distribution). Parametric survival models also take into account censoring naturally [10].	Assume survival time follows Weibull distribution	None
Gaussian distribution		Assume survival time follows Gaussian distribution	None
Logistic distribution		Assume survival time follows Logistic distribution	None
Sklearn			
Logistic	Scikit-learn (Sklearn) is a free machine learning library written in Python. It supports different machine learning algorithms including classification, regression and clustering tasks [11]	Using the same mathematical algorithm as Caret but written by different computer language (Python rather than R)	**penalty:** L1 Lasso regression or L2 Ridge regression (penalty term adds to loss function to increase model-generalizability) **C**: Inverse of regularization strength to control over-fitting (smaller value means stronger regularization and more-generalizability)
Random forest		Using the same mathematical algorithm as Caret but memory-optimisation and language advantage allow python version to fit more trees than Caret version.	**n_estimators**: number of trees **max_features**: number of predictors to consider when searching for the best split

	Package description	Model description	Key parameters selected by analyst for grid search
Neural network		Using the same mathematical algorithm as Caret but additional options provided by Sklearn to further control the structure and fitting process of neural network	**hidden_layer_sizes**: control number of hidden layers and number of neurons in each hidden layer **activation:** activation function for the hidden layer calculation. **solver:** different methods to optimise weights (beta) estimation
Gradient boosting classifier		Gradient boosting is a machine learning boosting method to train model by adding new predictor to the residual error of previous predictor rather than using all predictors at once [12]	**n_estimators:** the number of boosting stages to perform (larger means better performance but higher risk of overfitting) **learning_rate**: shrinks the contribution of each tree (a trade-off to n_estimators to control overfitting) **max_features**: number of predictors to consider when searching for the best split
extra-trees		Extra-trees model is similar to random forest except that random forest grows decision-trees by searching for the best splitting while Extra-trees uses random split for each decision-tree [12]	**n_estimators:** number of trees **max_features**: number of predictors to consider when searching for the best split
h2o			
Logistic	h2o is a Java-based machine learning library which has been implanted to both of R and python. Its main strength consists of memory allocation and the ability to distributed and paralleled machine learning process (which accelerates the model creating process) [13]	Using the same mathematical algorithm as Caret and Sklearn but being optimised for better memory allocation and parallelizing	None
Random forest		Using the same mathematical algorithm as Caret and Sklearn, but further memory-optimization and parallelizing allow to fit even more trees than Sklearn.	**max_depth:** maximum tree depth **mtries**: number of randomly selected predictors as candidates at each split **ntrees:** number of trees
Neural network		Using the same mathematical algorithm as Caret and Sklearn, but parallelising makes it possible to fit neural networks with large number of hidden layers with a more complex structure (deep learning)	**hidden:** control number of hidden layers and number of neurons in each hidden layer
autoML		AutoML is an automatic machine learning model training algorithm provided by h2o, which choose a best model among several candidate machine learning models such as gradient boosting machine, deep neural net and extremely randomised forest [14]	None

eTable 2.1: Performance indicators of machine learning and statistical models in overall cohort

	Model performance* (95% range #)				Average absolute change of model performance (95% range)
	C-statistic (2.5% ~ 97.5%) #	**Brier score (2.5% ~ 97.5%) #**	**Recall (Sensitivity) (2.5% ~ 97.5%) #**	**Precision (PPV) (2.5% ~ 97.5%) #**	**C-statistic (2.5% ~ 97.5%) #**
Logistic (Caret)	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.615 (0.609, 0.620)	0.163 (0.162, 0.164)	+0.00% (-0.03%, 0.04%)
Random forest (Caret)	0.869 (0.867, 0.869)	0.028 (0.028, 0.028)	0.656 (0.620, 0.675)	0.144 (0.139, 0.153)	-1.20% (-1.33%, -1.10%)
Neural network (Caret)	0.878 (0.867, 0.880)	0.028 (0.027, 0.028)	0.670 (0.642, 0.687)	0.148 (0.141, 0.153)	-0.15% (-1.35%, 0.06%)
Statistic logistic model	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.614 (0.607, 0.620)	0.163 (0.162, 0.164)	+0.01% (-0.02%, 0.04%)
QRISK3	0.879	0.031	0.834	0.107	Reference model
Framingham	0.865	0.031	0.892	0.085	-1.66% (-1.66%, -1.66%)
Local Cox model	0.877 (0.877, 0.878)	0.032 (0.031, 0.032)	0.810 (0.804, 0.816)	0.112 (0.110, 0.113)	-0.22% (-0.28%, -0.17%)
Parametric survival model (Weibull)	0.877 (0.876, 0.877)	0.031 (0.031, 0.032)	0.810 (0.804, 0.815)	0.111 (0.110, 0.113)	-0.29% (-0.35%, -0.24%)
Parametric survival model (Gaussian)	0.876 (0.876, 0.877)	0.031 (0.030, 0.031)	0.834 (0.830, 0.839)	0.104 (0.103, 0.105)	-0.33% (-0.39%, -0.29%)
Parametric survival model (Logistic)	0.876 (0.875, 0.876)	0.031 (0.031, 0.032)	0.796 (0.791, 0.802)	0.114 (0.113, 0.115)	-0.36% (-0.43%, -0.31%)
Logistic (Sklearn)	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.615 (0.609, 0.620)	0.163 (0.161, 0.164)	0.00% (-0.05%, 0.03%)
Random forest (Sklearn)	0.872 (0.871, 0.873)	0.028 (0.028, 0.028)	0.670 (0.661, 0.679)	0.142 (0.140, 0.144)	-0.80% (-0.89%, -0.71%)
Neural network (Sklearn)	0.872 (0.832, 0.879)	0.028 (0.028, 0.029)	0.556 (0.174, 0.692)	0.163 (0.137, 0.224)	-0.85% (-5.39%, -0.03%)
Gradient boosting classifier (Sklearn)	0.878 (0.877, 0.878)	0.028 (0.028, 0.028)	0.642 (0.623, 0.657)	0.154 (0.150, 0.157)	-0.17% (-0.29%, -0.08%)
extra-trees (Sklearn)	0.863 (0.861, 0.864)	0.028 (0.028, 0.029)	0.639 (0.628, 0.650)	0.139 (0.136, 0.141)	-1.89% (-2.05%, -1.76%)
Logistic (h2o)	0.879 (0.878, 0.879)	0.028 (0.028, 0.028)	0.615 (0.608, 0.621)	0.162 (0.161, 0.164)	-0.06% (-0.10%, -0.02%)
Random forest (h2o)	0.877 (0.877, 0.878)	0.028 (0.028, 0.028)	0.646 (0.631, 0.659)	0.152 (0.149, 0.154)	-0.22% (-0.29%, -0.17%)
Neural network (h2o)	0.875 (0.870, 0.879)	0.028 (0.028, 0.031)	0.552 (0.163, 0.780)	0.169 (0.118, 0.238)	-0.45% (-1.09%, -0.04%)
autoML (h2o)	0.879 (0.879, 0.880)	0.028 (0.028, 0.028)	0.616 (0.605, 0.642)	0.162 (0.157, 0.164)	-0.01% (-0.07%, 0.06%)

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

 # 95% range (2.5% ~ 97.5%) of model performance was derived from 100 random samples.

eTable 2.2 : Performance indicators of machine learning and statistical models in overall cohort with logistic caret model as reference model

	Model performance* (95% range #)				Average absolute change of model performance (95% range)
	C-statistic (2.5% ~ 97.5%) #	**Brier score (2.5% ~ 97.5%) #**	**Recall (Sensitivity) (2.5% ~ 97.5%) #**	**Precision (PPV) (2.5% ~ 97.5%) #**	**C-statistic (2.5% ~ 97.5%) #**
Logistic (Caret)	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.615 (0.609, 0.620)	0.163 (0.162, 0.164)	Reference model
Random forest (Caret)	0.869 (0.867, 0.869)	0.028 (0.028, 0.028)	0.656 (0.620, 0.675)	0.144 (0.139, 0.153)	-1.21% (-1.35%, -1.10%)
Neural network (Caret)	0.878 (0.867, 0.880)	0.028 (0.027, 0.028)	0.670 (0.642, 0.687)	0.148 (0.141, 0.153)	-0.16% (-1.34%, 0.05%)
Statistic logistic model	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.614 (0.607, 0.620)	0.163 (0.162, 0.164)	+0.00% (0.00%, 0.00%)
QRISK3	0.879	0.031	0.834	0.107	0.00% (-0.04%, 0.03%)
Framingham	0.865	0.031	0.892	0.085	-1.66% (-1.69%, -1.63%)
Local Cox model	0.877 (0.877, 0.878)	0.032 (0.031, 0.032)	0.810 (0.804, 0.816)	0.112 (0.110, 0.113)	-0.22% (-0.26%, -0.18%)
Parametric survival model (Weibull)	0.877 (0.876, 0.877)	0.031 (0.031, 0.032)	0.810 (0.804, 0.815)	0.111 (0.110, 0.113)	-0.30% (-0.34%, -0.26%)
Parametric survival model (Gaussian)	0.876 (0.876, 0.877)	0.031 (0.030, 0.031)	0.834 (0.830, 0.839)	0.104 (0.103, 0.105)	-0.34% (-0.38%, -0.30%)
Parametric survival model (Logistic)	0.876 (0.875, 0.876)	0.031 (0.031, 0.032)	0.796 (0.791, 0.802)	0.114 (0.113, 0.115)	-0.37% (-0.41%, -0.33%)
Logistic (Sklearn)	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.615 (0.609, 0.620)	0.163 (0.161, 0.164)	-0.01% (-0.04%, 0.00%)
Random forest (Sklearn)	0.872 (0.871, 0.873)	0.028 (0.028, 0.028)	0.670 (0.661, 0.679)	0.142 (0.140, 0.144)	-0.81% (-0.91%, -0.71%)
Neural network (Sklearn)	0.872 (0.832, 0.879)	0.028 (0.028, 0.029)	0.556 (0.174, 0.692)	0.163 (0.137, 0.224)	-0.85% (-5.41%, -0.06%)
Gradient boosting classifier (Sklearn)	0.878 (0.877, 0.878)	0.028 (0.028, 0.028)	0.642 (0.623, 0.657)	0.154 (0.150, 0.157)	-0.17% (-0.28%, -0.09%)
extra-trees (Sklearn)	0.863 (0.861, 0.864)	0.028 (0.028, 0.029)	0.639 (0.628, 0.650)	0.139 (0.136, 0.141)	-1.89% (-2.05%, -1.77%)
Logistic (h2o)	0.879 (0.878, 0.879)	0.028 (0.028, 0.028)	0.615 (0.608, 0.621)	0.162 (0.161, 0.164)	-0.06% (-0.09%, -0.04%)
Random forest (h2o)	0.877 (0.877, 0.878)	0.028 (0.028, 0.028)	0.646 (0.631, 0.659)	0.152 (0.149, 0.154)	-0.23% (-0.29%, -0.16%)
Neural network (h2o)	0.875 (0.870, 0.879)	0.028 (0.028, 0.031)	0.552 (0.163, 0.780)	0.169 (0.118, 0.238)	-0.45% (-1.10%, -0.05%)
autoML (h2o)	0.879 (0.879, 0.880)	0.028 (0.028, 0.028)	0.616 (0.605, 0.642)	0.162 (0.157, 0.164)	-0.02% (-0.05%, 0.03%)

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

 # 95% range (2.5% ~ 97.5%) of model performance was derived from 100 random samples.

eTable 3.1: Performance indicators of machine learning and statistical models in cohort without censoring with QRISK3 as reference model

	Model performance*				Average absolute change of model performance
	C-statistic	**Brier score**	**Recall (Sensitivity)**	**Precision (PPV)**	**C-statistic**
Logistic (Caret)	0.851	0.125	0.957	0.346	+0.49%
Random forest (Caret)	0.849	0.126	0.926	0.384	+0.34%
Neural network (Caret)	0.849	0.126	0.953	0.354	+0.25%
Statistical logistic model	0.851	0.125	0.957	0.346	+0.49%
QRISK3	0.847	0.150	0.844	0.455	Reference
Framingham	0.815	0.161	0.899	0.385	-3.74%
Local Cox model	0.850	0.126	0.968	0.330	+0.39%
Parametric survival model (Weibull)	0.849	0.128	0.955	0.347	+0.25%
Parametric survival model (Gaussian)	0.848	0.130	0.932	0.379	+0.23%
Parametric survival model (Logistic)	0.848	0.129	0.925	0.386	+0.20%
Logistic (Sklearn)	0.851	0.125	0.957	0.346	+0.49%
Random forest (Sklearn)	0.849	0.126	0.957	0.346	+0.30%
Neural network (Sklearn)	0.852	0.125	0.965	0.336	+0.63%
Gradient boosting classifier (Sklearn)	0.853	0.124	0.953	0.354	+0.74%
extra-trees (Sklearn)	0.845	0.127	0.954	0.345	-0.17%
Logistic (h2o)	0.849	0.126	0.957	0.343	+0.28%
Random forest (h2o)	0.851	0.125	0.960	0.344	+0.52%
Neural network (h2o)	0.852	0.126	0.927	0.386	+0.65%
autoML (h2o)	0.853	0.125	0.952	0.356	+0.71%

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

eTable 3.2: Performance indicators of machine learning and statistical models in cohort without censoring with logistic caret model as reference model

	Model performance*				Average absolute change of model performance
	C-statistic	**Brier score**	**Recall (Sensitivity)**	**Precision (PPV)**	**C-statistic**
Logistic (Caret)	0.851	0.125	0.957	0.346	Reference
Random forest (Caret)	0.849	0.126	0.926	0.384	-0.15%
Neural network (Caret)	0.849	0.126	0.953	0.354	-0.24%
Statistical logistic model	0.851	0.125	0.957	0.346	0.00%
QRISK3	0.847	0.150	0.844	0.455	-0.48%
Framingham	0.815	0.161	0.899	0.385	-4.21%
Local Cox model	0.850	0.126	0.968	0.330	-0.10%
Parametric survival model (Weibull)	0.849	0.128	0.955	0.347	-0.24%
Parametric survival model (Gaussian)	0.848	0.130	0.932	0.379	-0.25%
Parametric survival model (Logistic)	0.848	0.129	0.925	0.386	-0.28%
Logistic (Sklearn)	0.851	0.125	0.957	0.346	+0.00%
Random forest (Sklearn)	0.849	0.126	0.957	0.346	-0.18%
Neural network (Sklearn)	0.852	0.125	0.965	0.336	+0.15%
Gradient boosting classifier (Sklearn)	0.853	0.124	0.953	0.354	+0.25%
extra-trees (Sklearn)	0.845	0.127	0.954	0.345	-0.66%
Logistic (h2o)	0.849	0.126	0.957	0.343	-0.21%
Random forest (h2o)	0.851	0.125	0.960	0.344	+0.03%
Neural network (h2o)	0.852	0.126	0.927	0.386	+0.16%
autoML (h2o)	0.853	0.125	0.952	0.356	+0.22%

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

eTable 4.1: More performance indicators of machine learning and statistical models

	Model performance* (95% range #)			
	F1 score (2.5% ~ 97.5%) #	**Balanced accuracy (2.5% ~ 97.5%) #**	**NPV (2.5% ~ 97.5%) #**	**Specificity (2.5% ~ 97.5%) #**
Logistic (Caret)	0.258 (0.256, 0.259)	0.756 (0.754, 0.758)	0.986 (0.986, 0.986)	0.897 (0.895, 0.899)
Random forest (Caret)	0.236 (0.230, 0.245)	0.765 (0.754, 0.770)	0.987 (0.986, 0.988)	0.873 (0.864, 0.888)
Neural network (Caret)	0.242 (0.234, 0.248)	0.772 (0.752, 0.777)	0.988 (0.987, 0.988)	0.874 (0.864, 0.885)
Statistical logistic model	0.258 (0.256, 0.259)	0.756 (0.754, 0.758)	0.986 (0.986, 0.986)	0.898 (0.896, 0.900)
QRISK3	0.190	0.804	0.993	0.775
Framingham	0.155	0.790	0.995	0.688
Local Cox model	0.197 (0.194, 0.199)	0.800 (0.799, 0.801)	0.992 (0.992, 0.992)	0.791 (0.786, 0.796)
Parametric survival model (Weibull)	0.196 (0.194, 0.198)	0.800 (0.799, 0.800)	0.992 (0.992, 0.992)	0.789 (0.785, 0.794)
Parametric survival model (Gaussian)	0.185 (0.183, 0.187)	0.800 (0.800, 0.801)	0.993 (0.993, 0.993)	0.766 (0.762, 0.771)
Parametric survival model (Logistic)	0.199 (0.197, 0.201)	0.797 (0.796, 0.798)	0.992 (0.992, 0.992)	0.798 (0.795, 0.802)
Logistic (Sklearn)	0.258 (0.256, 0.259)	0.756 (0.754, 0.758)	0.986 (0.986, 0.986)	0.897 (0.895, 0.899)
Random forest (Sklearn)	0.235 (0.232, 0.237)	0.769 (0.766, 0.772)	0.988 (0.988, 0.988)	0.869 (0.864, 0.872)
Neural network (Sklearn)	0.240 (0.191, 0.272)	0.728 (0.576, 0.777)	0.984 (0.973, 0.988)	0.901 (0.858, 0.979)
Gradient boosting classifier (Sklearn)	0.248 (0.244, 0.251)	0.763 (0.757, 0.768)	0.987 (0.986, 0.987)	0.885 (0.880, 0.890)
extra-trees (Sklearn)	0.228 (0.225, 0.231)	0.755 (0.752, 0.758)	0.987 (0.986, 0.987)	0.871 (0.867, 0.875)
Logistic (h2o)	0.257 (0.255, 0.258)	0.756 (0.753, 0.758)	0.986 (0.986, 0.986)	0.897 (0.895, 0.899)
Random forest (h2o)	0.246 (0.243, 0.248)	0.764 (0.760, 0.768)	0.987 (0.987, 0.988)	0.883 (0.878, 0.887)
Neural network (h2o)	0.246 (0.172, 0.273)	0.728 (0.573, 0.795)	0.984 (0.973, 0.991)	0.904 (0.811, 0.983)
autoML (h2o)	0.257 (0.252, 0.259)	0.756 (0.753, 0.765)	0.986 (0.986, 0.987)	0.897 (0.888, 0.900)

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

 # 95% range (2.5% ~ 97.5%) of model performance was derived from 100 random samples.

eTable 4.2: More performance indicators of machine learning and statistical models in cohort without censoring

	Model performance* (95% range #)			
	F1 score (2.5% ~ 97.5%) #	**Balanced accuracy (2.5% ~ 97.5%) #**	**NPV (2.5% ~ 97.5%) #**	**Specificity (2.5% ~ 97.5%) #**
Logistic (Caret)	0.508	0.702	0.972	0.447
Random forest (Caret)	0.543	0.736	0.960	0.547
Neural network (Caret)	0.516	0.711	0.970	0.470
Statistical logistic model	0.509	0.703	0.971	0.449
QRISK3	0.592	0.768	0.936	0.692
Framingham	0.539	0.730	0.948	0.561
Local Cox model	0.492	0.683	0.976	0.399
Parametric survival model (Weibull)	0.510	0.704	0.971	0.452
Parametric survival model (Gaussian)	0.539	0.733	0.962	0.534
Parametric survival model (Logistic)	0.545	0.738	0.960	0.552
Logistic (Sklearn)	0.508	0.702	0.972	0.447
Random forest (Sklearn)	0.508	0.702	0.971	0.448
Neural network (Sklearn)	0.498	0.691	0.975	0.417
Gradient boosting classifier (Sklearn)	0.516	0.711	0.971	0.468
extra-trees (Sklearn)	0.507	0.701	0.969	0.447
Logistic (h2o)	0.505	0.698	0.971	0.439
Random forest (h2o)	0.506	0.700	0.973	0.441
Neural network (h2o)	0.545	0.739	0.961	0.551
autoML (h2o)	0.518	0.713	0.970	0.475

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

 # 95% range (2.5% ~ 97.5%) of model performance was derived from 100 random samples.

eTable 5.1: Comparison of individual risk predictions of machine learning and statistical models in overall cohort and cohort without censoring

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to reference model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Overall cohort									
QRISK3 as reference model									
Soft voting *	0.3~3.5	2.7~5.2	3.1~6.0	3.5~6.8	3.9~7.5	4.3~8.3	4.6~9.0	5.1~9.8	7.2~36.4
All models #	0.1~4.9	1.5~10.5	1.8~11.6	2.0~12.7	2.4~13.6	2.6~14.7	2.9~15.7	3.2~16.7	5.0~44.8
Logistic model (Caret) as reference model									
Soft voting	0.3~7.8	8.2~12.3	9.4~13.9	10.5~15.4	11.7~16.8	12.8~18.2	14.0~19.3	15.0~20.5	17.1~41.6
All models	0.1~9.7	5.4~20.1	6.2~22.1	7.0~24.0	7.8~26.1	8.5~28.1	9.2~29.9	9.9~31.7	12.6~53.7
Cohort without censoring									
QRISK3 as reference model									
Soft voting	1.4~16.4	12.5~25.2	14.3~28.5	15.9~30.7	17.1~33.7	18.8~36.6	20.4~38.9	21.8~41.2	28.4~80.7
All models	0.6~18.1	8.4~29.5	9.5~33.4	10.5~36.0	11.4~39.4	12.3~42.4	13.2~45.2	13.9~47.4	19.3~85.9
Logistic model (Caret) as reference model									
Soft voting	1.2~5.3	4.7~7.7	5.4~8.9	6.2~9.8	7.1~10.9	7.8~11.9	8.5~13.5	9.4~14.3	11.9~76.2
All models	0.2~6.3	1.6~9.2	2.0~10.9	2.3~12.2	2.7~14.1	3.1~15.3	3.4~17.0	3.8~18.2	8.4~82.0

*** 95% range of individual risk prediction from soft voting (averaging) of all models except the reference model**

 # 95% range of individual risk prediction from all models except the reference model

eTable 5.2: Comparison of individual risk predictions of machine learning and statistical models in overall cohort (with as reference the risk predictions of the QRISK3)

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from QRISK3 model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Caret									
Logistic	0.1~2.3	1.4~3.7	1.6~4.3	1.8~4.9	2.1~5.5	2.3~6.1	2.5~6.7	2.7~7.4	4.2~35.7
Random forest	0.0~2.9	1.7~6.3	2.0~7.3	2.4~8.1	2.8~9.0	3.1~9.9	3.5~10.9	3.9~11.5	5.5~30.4
Neural network	0.1~2.5	1.4~4.4	1.7~5.2	1.9~6.1	2.3~6.9	2.5~7.6	2.8~8.4	3.2~9.2	5.3~24.6
Statistical logistic model									
Statistical logistic model	0.1~2.3	1.4~3.7	1.6~4.3	1.8~4.9	2.1~5.5	2.3~6.1	2.5~6.6	2.7~7.3	4.2~35.5
Cox model									
QRISK3	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Framingham	0.0~10.0	4.5~15.3	5.0~16.7	5.3~18.1	5.9~19.4	6.2~20.8	6.2~22.2	6.7~23.2	8.2~49.7
Local Cox model	0.4~5.4	4.1~8.2	4.7~9.2	5.2~10.4	5.7~11.3	6.2~12.4	6.7~13.7	7.2~14.7	10.3~61.1
Parametric survival model									
Parametric survival model (Weibull)	1.0~5.5	4.2~8.3	4.7~9.4	5.2~10.7	5.7~11.7	6.2~12.8	6.7~14.0	7.2~15.1	10.1~59.2
Parametric survival model (Gaussian)	1.0~6.2	4.7~9.7	5.3~11.1	6.0~12.4	6.6~13.6	7.2~14.9	7.7~16.1	8.3~17.3	11.4~49.4
Parametric survival model (Logistic)	1.0~5.1	3.8~8.2	4.3~9.5	4.8~10.8	5.3~12.0	5.9~13.3	6.3~14.7	6.9~15.9	9.8~57.5
Sklearn									
Logistic	0.1~2.3	1.4~3.7	1.6~4.3	1.8~4.9	2.1~5.5	2.3~6.1	2.5~6.7	2.7~7.3	4.3~35.5
Random forest	0.0~3.1	1.8~6.4	2.1~7.4	2.5~8.2	2.9~9.0	3.4~9.9	3.6~11.0	4.1~11.6	5.8~30.6
Neural network	0.1~2.4	1.2~4.8	1.4~5.6	1.6~6.4	1.9~7.1	2.1~7.6	2.3~8.3	2.5~8.9	4.1~23.9
Gradient boosting classifier	0.1~2.5	1.4~4.4	1.7~5.2	2.0~5.8	2.3~6.7	2.7~7.3	2.9~8.3	3.3~9.2	5.2~30.5
extra-trees	0.0~3.2	1.6~6.2	2.0~7.1	2.3~7.9	2.6~8.7	3.0~9.5	3.2~10.5	3.7~11.2	5.6~29.4
h2o									
Logistic	0.1~2.4	1.4~3.8	1.6~4.4	1.8~5.0	2.0~5.6	2.3~6.1	2.4~6.7	2.7~7.4	4.3~35.0

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from QRISK3 model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Random forest	0.1~2.8	1.8~5.1	2.0~6.0	2.4~6.7	2.7~7.3	3.1~8.2	3.4~9.1	3.8~9.6	5.4~28.9
Neural network	0.1~2.2	1.2~4.0	1.4~4.7	1.6~5.4	1.9~6.1	2.1~6.7	2.2~7.5	2.4~8.2	4.0~29.2
autoML	0.1~2.3	1.4~3.8	1.6~4.3	1.8~4.9	2.1~5.5	2.3~6.1	2.5~6.7	2.8~7.4	4.3~35.0
Overall									
Soft voting *	0.3~3.5	2.7~5.2	3.1~6.0	3.5~6.8	3.9~7.5	4.3~8.3	4.6~9.0	5.1~9.8	7.2~36.4
All models #	0.1~4.9	1.5~10.5	1.8~11.6	2.0~12.7	2.4~13.6	2.6~14.7	2.9~15.7	3.2~16.7	5.0~44.8

*** 95% range of individual risk prediction from soft voting (averaging) of all models except the reference model**

 # 95% range of individual risk prediction from all models except the reference model

eTable 5.3: Comparison of individual risk predictions of machine learning and statistical models in overall cohort (with as reference the risk predictions of the logistic Caret model)

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from logistic Caret model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Caret									
Logistic	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Random forest	0.0~6.7	4.9~14.5	5.8~16.1	6.6~17.7	7.4~19.1	8.2~20.0	9.1~20.9	10.0~21.8	12.5~33.6
Neural network	0.1~6.3	6.7~9.9	7.8~11.4	8.8~12.6	9.8~13.9	10.6~14.9	11.5~15.7	12.2~16.6	13.8~26.2
Statistical logistic model									
Statistical logistic model	0.1~5.0	6.0~6.9	7.0~7.9	8.0~8.9	9.0~9.9	10.0~10.9	11.0~11.9	12.0~12.9	13.2~40.9
Cox model									
QRISK3	0.1~12.5	10.9~23.4	12.4~26.2	13.9~28.6	15.4~31.3	16.7~33.5	18.2~35.2	19.5~36.6	23.7~59.6
Framingham	0.1~17.1	7.1~28.3	7.3~30.3	7.6~32.0	7.6~34.5	8.1~35.7	7.9~36.6	8.5~38.5	10.5~57.7
Local Cox model	0.4~11.0	10.5~19.0	11.9~21.7	13.4~24.0	14.8~26.9	16.1~29.5	17.5~31.5	18.9~33.3	23.1~69.0
Parametric survival model									
Parametric survival model (Weibull)	1.0~11.1	10.3~19.0	11.7~21.8	13.1~24.6	14.3~27.6	15.5~30.0	16.8~32.5	18.2~34.4	22.2~66.6
Parametric survival model (Gaussian)	1.0~12.6	11.3~21.2	12.8~23.8	14.1~26.0	15.4~28.4	16.7~30.4	17.9~32.4	19.0~33.6	22.6~54.0
Parametric survival model (Logistic)	1.0~11.0	9.7~20.8	11.2~23.9	12.6~27.0	13.9~30.3	15.2~33.0	16.5~35.4	17.9~37.1	22.3~63.1
Sklearn									
Logistic	0.1~5.1	6.0~7.0	7.0~8.0	8.0~9.0	9.0~10.0	10.0~11.0	11.0~12.0	12.0~13.0	13.2~40.9
Random forest	0.0~6.9	5.2~14.4	6.1~15.9	7.0~17.7	7.8~19.1	8.6~19.8	9.5~20.7	10.5~21.8	13.0~33.6
Neural network	0.1~5.3	4.5~9.4	5.1~10.5	5.8~11.4	6.4~12.3	7.1~13.1	7.7~13.9	8.3~14.7	10.2~26.6
Gradient boosting classifier	0.1~6.0	5.3~10.7	6.2~12.2	7.1~13.5	8.0~14.9	8.8~16.3	9.5~17.4	10.4~18.8	12.5~34.3
extra-trees	0.0~6.9	5.0~13.5	5.8~14.8	6.6~16.2	7.4~17.5	8.1~18.5	9.1~19.6	9.8~20.4	12.5~32.4
h2o									
Logistic	0.1~5.1	5.9~7.2	6.8~8.4	7.7~9.5	8.7~10.6	9.6~11.6	10.5~12.8	11.5~13.9	13.2~40.0

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from logistic Caret model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Random forest	0.1~5.9	5.5~12.2	6.3~13.7	7.0~15.4	7.8~16.7	8.6~17.5	9.4~18.6	10.4~19.6	12.8~32.1
Neural network	0.1~5.0	4.7~8.6	5.4~9.8	6.3~10.9	7.0~12.0	7.8~13.0	8.6~14.0	9.4~14.9	11.6~33.0
autoML	0.1~5.1	6.0~7.1	7.0~8.1	8.0~9.1	9.0~10.2	10.0~11.2	11.0~12.2	11.9~13.2	13.3~40.1
Overall									
Soft voting *	0.3~7.8	8.2~12.3	9.4~13.9	10.5~15.4	11.7~16.8	12.8~18.2	14.0~19.3	15.0~20.5	17.1~41.6
All models #	0.1~9.7	5.4~20.1	6.2~22.1	7.0~24.0	7.8~26.1	8.5~28.1	9.2~29.9	9.9~31.7	12.6~53.7

*** 95% range of individual risk prediction from soft voting (averaging) of all models except the reference model**

 # 95% range of individual risk prediction from all models except the reference model

eTable 5.4: Comparison of the individual risk predictions of machine learning and statistical models in cohort without censoring (with as reference the risk predictions of the QRISK3 model)

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from QRISK3 model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Caret									
Logistic	1.2~17.8	12.4~26.3	14.4~30.5	15.6~32.7	17.3~35.7	18.9~38.8	20.3~42.3	21.7~44.8	30.1~88.6
Random forest	0.2~17.9	6.5~31.4	7.9~35.8	9.5~37.9	10.7~40.4	12.6~44.7	13.5~46.9	14.9~48.9	24.1~85.8
Neural network	2.8~18.0	11.3~29.5	13.4~34.8	15.4~37.9	16.9~41.7	19.0~43.8	21.6~48.0	23.0~50.5	31.8~75.1
Statistical logistic model									
Statistical logistic model	1.2~17.7	12.3~26.2	14.3~30.4	15.5~32.5	17.2~35.6	18.7~38.7	20.2~42.2	21.6~44.7	30.0~88.6
Cox model									
QRISK3	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Framingham	0.1~11.1	4.8~15.7	5.4~16.9	5.9~18.8	5.9~19.9	6.6~21.4	7.2~22.1	7.2~23.9	9.0~53.6
Local Cox model	1.8~15.6	10.7~22.1	12.2~25.2	13.1~27.1	14.3~29.8	15.6~32.5	16.4~35.1	17.3~37.4	24.2~90.2
Parametric survival model									
Parametric survival model (Weibull)	1.6~15.5	10.6~22.1	11.9~25.3	12.7~27.0	13.7~29.7	14.4~32.1	15.6~34.4	16.3~37.3	22.5~87.8
Parametric survival model (Gaussian)	1.0~13.8	8.1~21.4	9.5~24.3	10.3~26.2	11.3~28.9	12.5~30.8	13.6~33.2	14.4~36.4	20.4~76.5
Parametric survival model (Logistic)	1.0~12.7	7.6~20.1	9.0~23.4	9.6~25.1	10.6~28.1	11.7~30.1	12.6~32.8	13.5~36.5	19.4~80.3
Sklearn									
Logistic	1.2~17.8	12.4~26.3	14.4~30.5	15.6~32.7	17.3~35.7	18.8~38.8	20.3~42.3	21.7~44.8	30.1~88.6
Random forest	0.4~22.9	10.5~35.8	12.3~41.0	14.2~42.4	15.4~45.3	17.2~48.0	19.0~49.8	21.0~51.7	29.7~83.4
Neural network	0.3~18.8	12.5~30.1	14.5~34.4	16.1~36.8	17.6~41.0	19.6~43.9	21.6~47.1	22.5~50.1	30.8~84.1
Gradient boosting classifier	1.0~19.2	12.4~31.6	14.4~35.7	15.9~38.3	17.5~42.2	19.0~47.2	21.0~48.2	22.6~53.1	30.1~87.2
extra-trees	0.3~24.0	9.1~38.0	10.8~42.4	12.7~45.8	13.6~48.7	15.0~51.2	16.5~53.6	17.7~55.5	26.8~86.3
h2o									
Logistic	1.3~17.9	12.2~25.6	14.3~29.1	15.6~31.5	16.9~34.6	18.0~37.6	20.3~40.4	21.3~42.8	30.2~87.2

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from QRISK3 model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Random forest	1.6~19.5	13.4~30.6	15.1~33.8	16.8~35.8	18.1~39.3	19.5~42.0	21.5~45.0	22.7~46.7	30.7~83.1
Neural network	0.5~20.4	13.4~32.3	15.9~37.2	17.4~39.6	19.1~44.3	21.1~47.4	23.3~48.8	24.6~52.2	33.7~87.2
autoML	5.3~13.6	10.2~23.0	11.4~27.6	12.3~30.0	13.6~34.4	14.8~36.5	16.0~40.0	17.3~43.8	23.8~87.7
Overall									
Soft voting *	1.4~16.4	12.5~25.2	14.3~28.5	15.9~30.7	17.1~33.7	18.8~36.6	20.4~38.9	21.8~41.2	28.4~80.7
All models #	0.6~18.1	8.4~29.5	9.5~33.4	10.5~36.0	11.4~39.4	12.3~42.4	13.2~45.2	13.9~47.4	19.3~85.9

*** 95% range of individual risk prediction from soft voting (averaging) of all models except the reference model**

 # 95% range of individual risk prediction from all models except the reference model

eTable 5.5: Comparison of the individual risk predictions of machine learning and statistical models in cohort without censoring (with as reference the risk predictions of the Caret Logistic model)

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from Caret Logistic model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Caret									
Logistic	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Random forest	0.0~7.1	1.2~12.0	1.6~14.4	2.1~15.1	2.4~17.4	2.7~18.6	3.2~20.9	3.9~21.8	7.7~83.6
Neural network	2.8~6.0	4.7~7.8	5.2~9.1	5.8~10.4	6.4~11.9	7.1~13.1	7.8~14.9	8.5~16.1	11.9~74.2
Statistical logistic model									
Statistical logistic model	1.0~5.8	6.0~6.9	7.0~7.9	8.0~8.9	9.0~9.9	10.0~10.9	11.0~11.9	12.0~12.9	13.7~85.7
Cox model									
QRISK3	0.1~1.8	0.9~2.8	1.2~3.5	1.4~4.1	1.5~4.6	1.9~5.2	2.1~5.9	2.4~6.4	4.2~52.5
Framingham	0.1~5.1	1.1~8.0	1.5~9.2	1.7~10.4	2.0~11.3	2.2~12.6	2.6~13.6	2.7~14.3	5.1~47.8
Local Cox model	1.5~6.2	6.0~7.3	6.8~8.2	7.6~9.0	8.3~9.9	9.0~10.7	9.7~11.5	10.4~12.3	12.4~86.3
Parametric survival model									
Parametric survival model (Weibull)	1.3~6.2	5.8~7.4	6.5~8.2	7.3~9.2	7.9~10.1	8.6~11.0	9.2~12.0	9.8~12.8	11.8~83.0
Parametric survival model (Gaussian)	1.0~3.8	3.5~5.2	4.1~6.1	4.8~7.1	5.4~8.0	6.1~9.1	6.8~10.1	7.3~11.0	9.3~72.2
Parametric survival model (Logistic)	1.0~3.9	3.6~5.0	4.2~5.8	4.8~6.7	5.3~7.5	5.9~8.4	6.5~9.3	7.0~10.1	8.7~76.4
Sklearn									
Logistic	1.0~5.8	6.0~7.0	7.0~8.0	8.0~9.0	9.0~10.0	10.0~11.0	11.0~12.0	12.0~13.0	13.7~85.7
Random forest	0.3~9.4	2.4~15.5	3.0~18.5	3.7~19.1	4.3~21.8	4.9~22.5	5.6~24.9	6.3~26.3	11.8~80.8
Neural network	0.3~4.8	4.0~7.1	5.0~8.6	6.0~9.8	7.1~11.6	8.2~12.6	9.2~14.4	10.4~15.5	13.4~81.3
Gradient boosting classifier	0.9~5.8	3.6~9.5	4.5~10.7	5.3~12.8	6.1~14.9	7.0~15.5	7.8~18.4	8.7~19.6	12.7~84.1
extra-trees	0.2~10.2	2.1~16.5	2.6~18.6	3.1~20.1	3.8~22.6	4.4~23.6	4.9~27.2	5.5~28.2	11.2~82.9
h2o									
Logistic	1.1~6.1	5.5~7.4	6.4~8.4	7.4~9.5	8.3~10.6	8.2~11.6	10.0~12.6	10.8~13.7	13.9~84.4

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from Caret Logistic model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Random forest	1.5~7.1	4.7~11.9	5.5~13.7	6.3~14.1	7.2~16.8	7.7~17.5	8.5~20.1	9.2~21.4	13.1~79.7
Neural network	0.4~6.1	4.6~8.8	5.9~10.6	7.0~11.7	8.2~13.4	9.2~14.5	10.4~16.1	11.3~17.6	14.6~84.0
autoML	5.3~6.8	6.2~8.0	6.5~9.3	6.8~10.1	7.2~11.3	7.5~12.1	7.8~13.7	8.3~14.5	10.3~85.3
Overall									
Soft voting *	1.2~5.3	4.7~7.7	5.4~8.9	6.2~9.8	7.1~10.9	7.8~11.9	8.5~13.5	9.4~14.3	11.9~76.2
All models #	0.2~6.3	1.6~9.2	2.0~10.9	2.3~12.2	2.7~14.1	3.1~15.3	3.4~17.0	3.8~18.2	8.4~82.0

*** 95% range of individual risk prediction from soft voting (averaging) of all models except the reference model**

 # 95% range of individual risk prediction from all models except the reference model

eTable 6: SPEARMAN correlations of Machine learning models and statistical models in risk groups with logistic (Caret) predicted 7%~8%

	SPEARMAN Correlation																		
	Caret			Statistical model							Sklearn					H2o			
	Logit *	RF	NN	Logit	QRISK3	Framingham	Cox	Parametric Weibull	Parametric Gaussian	Parametric Logistic	Logit	RF	NN	GBC	extra-trees	Logit	RF	NN	auto ML
Caret																			
Logistic	1.00	0.11	0.37	1.00	0.15	0.05	0.21	0.19	0.17	0.16	0.98	0.11	0.18	0.20	0.12	0.60	0.15	0.26	0.85
Random forest	0.11	1.00	0.16	0.11	0.44	-0.06	0.35	0.38	0.38	0.38	0.13	0.99	0.15	0.48	0.65	0.16	0.90	0.02	0.26
Neural network	0.37	0.16	1.00	0.38	0.30	0.09	0.32	0.15	0.17	0.13	0.42	0.15	0.67	0.38	0.38	0.14	0.21	0.71	0.36
Statistical logistic model																			
Statistical logistic model	1.00	0.11	0.38	1.00	0.15	0.06	0.21	0.19	0.17	0.16	0.98	0.11	0.18	0.20	0.12	0.60	0.14	0.26	0.85
Cox model																			
QRISK3	0.15	0.44	0.30	0.15	1.00	0.32	0.60	0.50	0.48	0.49	0.17	0.43	0.14	0.37	0.35	0.20	0.43	0.11	0.23
Framingham	0.05	-0.06	0.09	0.06	0.32	1.00	-0.04	-0.32	-0.34	-0.36	0.01	-0.05	-0.30	0.21	-0.02	0.03	-0.23	-0.06	0.06
Local Cox model	0.21	0.35	0.32	0.21	0.60	-0.04	1.00	0.85	0.79	0.80	0.28	0.33	0.32	0.13	0.25	0.30	0.33	0.15	0.25
Parametric survival model																			
Parametric survival model (Weibull)	0.19	0.38	0.15	0.19	0.50	-0.32	0.85	1.00	0.97	0.99	0.26	0.36	0.23	0.04	0.22	0.32	0.44	0.03	0.26
Parametric survival model (Gaussian)	0.17	0.38	0.17	0.17	0.48	-0.34	0.79	0.97	1.00	0.99	0.23	0.36	0.27	0.03	0.24	0.24	0.44	0.10	0.23
Parametric survival model (Logistic)	0.16	0.38	0.13	0.16	0.49	-0.36	0.80	0.99	0.99	1.00	0.23	0.36	0.25	0.03	0.22	0.28	0.45	0.06	0.24
Sklearn																			
Logistic	0.98	0.13	0.42	0.98	0.17	0.01	0.28	0.26	0.23	0.23	1.00	0.13	0.27	0.21	0.14	0.67	0.18	0.29	0.88

	SPEARMAN Correlation																		
	Caret			Statistical model							Sklearn					H2o			
	Logit *	RF	NN	Logit	QRISK3	Framingham	Cox	Parametric Weibull	Parametric Gaussian	Parametric Logistic	Logit	RF	NN	GBC	extra-trees	Logit	RF	NN	auto ML
Random forest	0.11	0.99	0.15	0.11	0.43	-0.05	0.33	0.36	0.36	0.36	0.13	1.00	0.12	0.50	0.68	0.17	0.89	-0.00	0.27
Neural network	0.18	0.15	0.67	0.18	0.14	-0.30	0.32	0.23	0.27	0.25	0.27	0.12	1.00	0.19	0.24	0.17	0.21	0.68	0.22
Gradient boosting classifier	0.20	0.48	0.38	0.20	0.37	0.21	0.13	0.04	0.03	0.03	0.21	0.50	0.19	1.00	0.45	0.14	0.52	0.26	0.36
extra-trees	0.12	0.65	0.38	0.12	0.35	-0.02	0.25	0.22	0.24	0.22	0.14	0.68	0.24	0.45	1.00	0.05	0.65	0.27	0.21
h2o																			
Logistic	0.60	0.16	0.14	0.60	0.20	0.03	0.30	0.32	0.24	0.28	0.67	0.17	0.17	0.14	0.05	1.00	0.23	0.12	0.86
Random forest	0.15	0.90	0.21	0.14	0.43	-0.23	0.33	0.44	0.44	0.45	0.18	0.89	0.21	0.52	0.65	0.23	1.00	0.12	0.34
Neural network	0.26	0.02	0.71	0.26	0.11	-0.06	0.15	0.03	0.10	0.06	0.29	-0.00	0.68	0.26	0.27	0.12	0.12	1.00	0.28
autoML	0.85	0.26	0.36	0.85	0.23	0.06	0.25	0.26	0.23	0.24	0.88	0.27	0.22	0.36	0.21	0.86	0.34	0.28	1.00

*** Abbreviation: Logit - Logistic model, RF - Random forest, NN - Neural network, Cox - Cox proportional hazard model, GBC - Gradient boosting classifier**

eTable 7: Reclassification of individual risk predictions of machine learning and statistical models with 10% as threshold

	Reclassification in overall testing cohort	
	Reclassified*	**Not reclassified**
Overall cohort		
QRISK3 as reference model		
Below or equal to the threshold (≤10%)	73871 (10.0%)	661603 (90.0%)
Above the threshold (>10%)	113260 (62.9%)	66745 (37.1%)
Logistic model (Caret) as reference model		
Below or equal to the threshold (≤10%)	170983 (20.5%)	661603 (79.5%)
Above the threshold (>10%)	16148 (19.5%)	66745 (80.5%)
Cohort without censoring		
QRISK3 as reference model		
Below or equal to the threshold (≤10%)	34691 (49.1%)	35891 (50.9%)
Above the threshold (>10%)	2269 (5.5%)	39017 (94.5%)
Logistic model (Caret) as reference model		
Below or equal to the threshold (≤10%)	6872 (16.1%)	35891 (83.9%)
Above the threshold (>10%)	30088 (43.5%)	39017 (56.5%)

*** For patients who are below or equal to the threshold, they are re-classified if they have prediction above the threshold in any model.**
For patients who are above the threshold, they are re-classified if they have prediction below or equalt to the threshold in any model.

eTable 8: Reclassification of individual risk predictions of Caret neural network models with different hyperparameters

	Reclassification in overall testing cohort	
	Reclassified*	**Not reclassified**
Overall cohort		
Models with the most frequent selected hyperparameters as reference model		
Below or equal to the threshold (<=7.5%)	12016 (1.5%)	773472 (98.5%)
Above the threshold (>7.5%)	14987 (11.5%)	115004 (88.5%)

*** For patients who are below or equal to the threshold, they are re-classified if they have prediction above the threshold in any model.**
For patients who are above the threshold, they are re-classified if they have prediction below or equalt to the threshold in any model.

eTable 9: Inconsistency of individual risk prediction between machine learning models derived from overall cohort and cohort without censoring

	Range of individual risk predictions (2.5^{th}~97.5^{th}) for the same group of patients *								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Caret									
Logistic	0.8~31.9	21.6~61.2	24.2~65.5	27.2~69.4	29.5~73.7	31.8~75.4	34.1~77.4	36.2~79.6	44.2~90.2
Random forest	0.2~26.8	10.5~52.5	11.6~57.9	13.5~61.7	15.3~65.3	16.8~69.5	18.8~73.7	19.4~76.8	29.1~87.9
Neural network	0.8~29.7	19.7~54.4	22.7~57.5	25.3~60.7	28.0~64.3	31.0~68.0	33.3~70.2	35.7~72.8	42.2~76.2
Cox model									
QRISK3	0.1~5.4	6.0~7.0	7.0~8.0	8.0~9.0	9.0~10.0	10.0~11.0	11.0~12.0	12.0~13.0	13.3~54.0
Framingham	0.0~5.7	6.0~7.0	7.0~8.0	8.0~9.0	9.0~10.0	10.0~11.0	11.0~12.0	12.0~13.0	13.2~48.2
Local Cox model	2.5~28.0	24.1~42.9	27.1~47.7	29.9~52.2	32.7~56.2	35.4~59.7	37.8~62.9	40.4~66.3	48.4~99.2
Sklearn									
Logistic	0.8~31.8	21.6~61.1	24.3~65.0	27.3~69.9	29.6~73.2	32.0~75.3	34.1~77.5	36.3~79.1	44.3~90.2
Random forest	0.4~32.1	12.7~56.8	13.7~60.4	15.1~64.8	16.1~68.8	18.4~72.1	19.0~74.7	19.5~77.5	27.1~87.2
Neural network	1.0~35.5	20.9~59.8	23.7~63.5	26.8~67.4	30.3~70.5	33.9~73.5	37.4~75.0	40.8~76.3	49.1~83.0
Gradient boosting classifier	0.8~33.2	17.7~62.1	19.8~65.9	22.0~71.5	23.4~76.9	25.2~81.0	27.9~83.3	29.7~85.2	38.1~89.2
extra-trees	0.1~33.0	6.8~64.9	7.6~71.2	6.0~74.2	6.5~80.1	5.1~85.1	10.1~86.6	11.2~88.5	15.0~97.1
h2o									
Logistic	0.8~30.9	21.9~55.3	24.6~59.8	27.4~63.4	29.8~66.7	32.0~70.9	34.2~72.4	36.4~74.2	44.4~88.4
Random forest	1.4~30.5	18.2~51.0	20.1~54.8	22.3~57.8	24.4~61.1	26.4~65.4	28.6~67.9	31.1~71.4	39.2~84.9
Neural network	0.1~31.8	19.6~69.9	22.5~75.2	25.4~80.0	28.1~82.5	31.1~84.7	33.2~86.8	36.4~88.0	48.4~93.5
autoML	5.0~29.2	17.5~63.5	19.9~69.1	22.6~76.3	24.9~80.7	27.2~81.8	29.9~83.9	31.3~85.7	40.4~91.3

*** 95% range of individual risk prediction of the same risk-group patients predicted by model derived from cohort without censoring comparing to the same model dervived from overall cohort (risk-group displayed in the second line of the table title)**

eTable 10: Performance indicators of machine learning and Cox models developed in South and validated in North

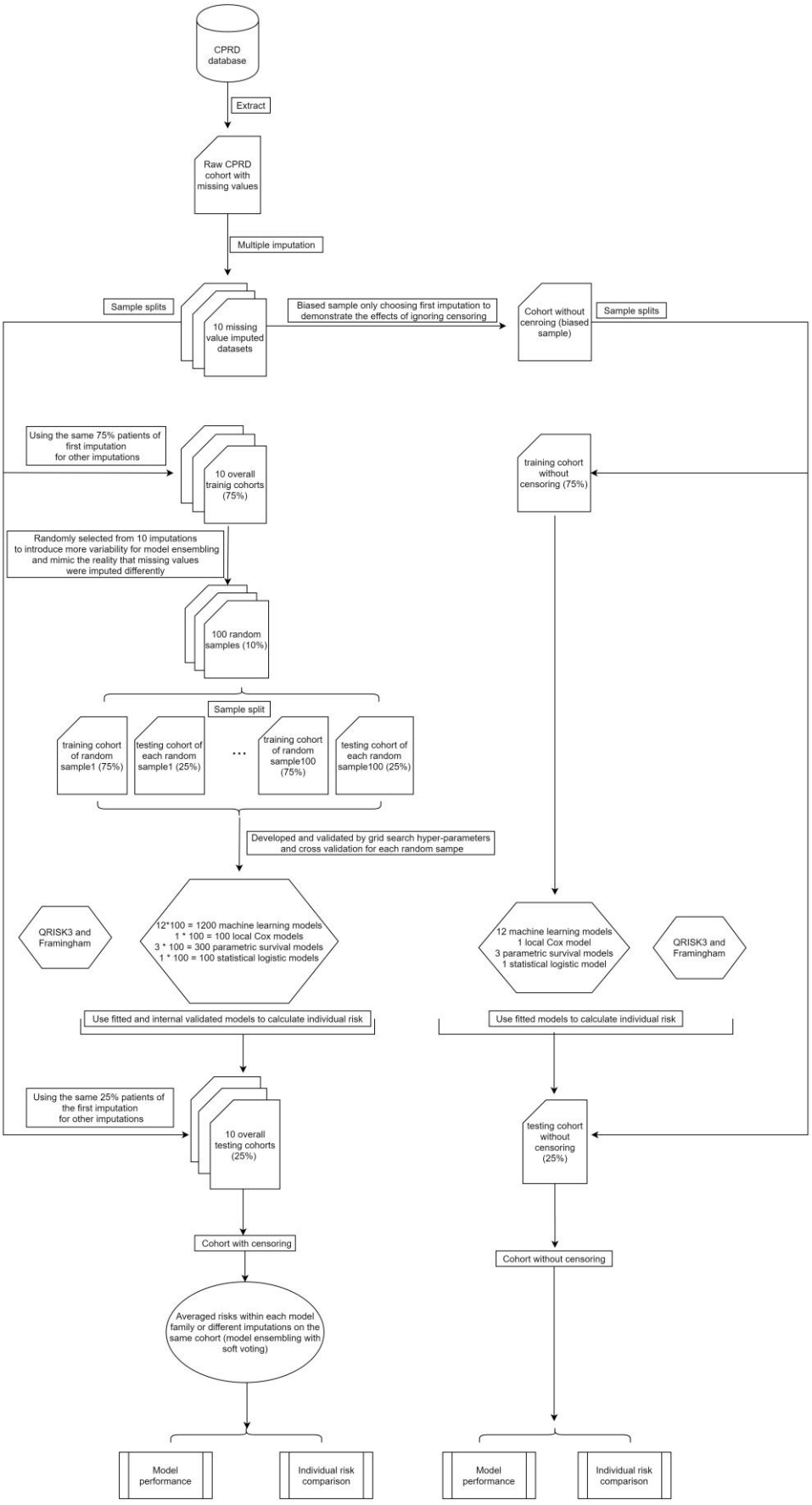
	Model performance*				Average absolute change of model performance
	C-statistic	Brier score	Recall (Sensitivity)	Precision (PPV)	C-statistic
North#					
Logistic (Caret)	0.871	0.032	0.575	0.179	Reference
Neural network (Caret)	0.871	0.032	0.631	0.167	-0.02%
Local Cox model	0.869	0.036	0.798	0.124	-0.21%
South$					
Logistic (Caret)	0.877	0.028	0.607	0.164	Reference
Neural network (Caret)	0.877	0.028	0.659	0.151	+0.01%
Local Cox model	0.875	0.031	0.803	0.112	-0.21%

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**
 # Testing cohort only including practices from North of UK which was different from development cohort (i.e. practices from south)
$ Testing cohort only including practices from South of UK which was similar to development cohort

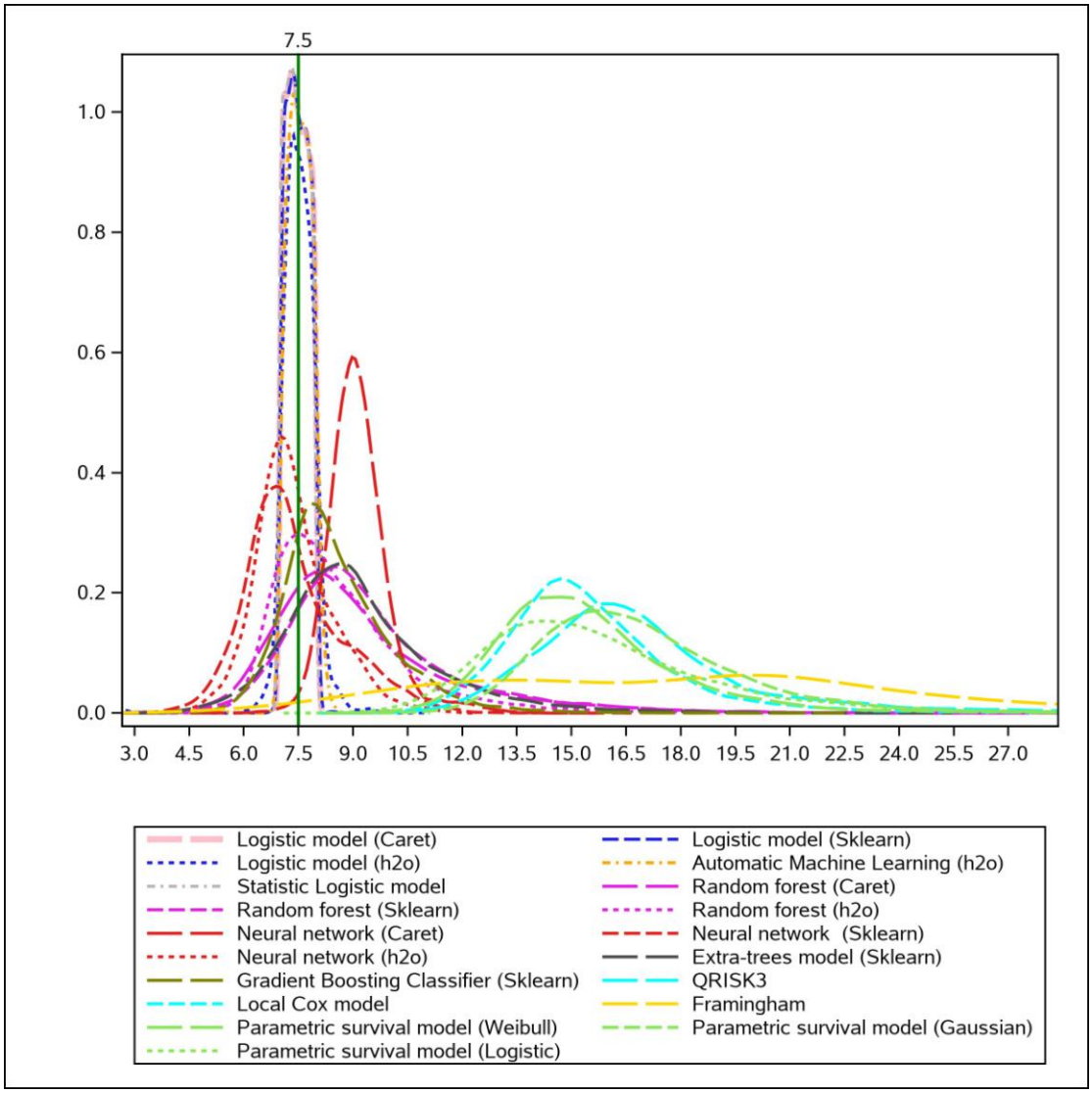
eTable 11: Performance indicators of machine learning and Cox models with lower number of predictors

	Model performance*				Average absolute change of model performance
	C-statistic	**Brier score**	**Recall (Sensitivity)**	**Precision (PPV)**	**C-statistic**
Using the same 1/3 random predictors #					
Logistic (Caret)	0.870	0.028	0.591	0.157	Reference
Random forest (Caret)	0.705	0.036	0.302	0.125	-18.95%
Neural network (Caret)	0.870	0.028	0.655	0.142	+0.01%
Local Cox model	0.869	0.032	0.801	0.108	-0.08%
Using the same 1/2 random predictors					
Logistic (Caret)	0.875	0.028	0.602	0.160	Reference
Random forest (Caret)	0.832	0.029	0.594	0.132	-4.96%
Neural network (Caret)	0.876	0.028	0.669	0.145	+0.03%
Local Cox model	0.875	0.031	0.809	0.110	-0.07%
Using the same 2/3 random predictors					
Logistic (Caret)	0.878	0.028	0.610	0.162	Reference
Random forest (Caret)	0.858	0.028	0.621	0.143	-2.27%
Neural network (Caret)	0.878	0.028	0.665	0.149	+0.02%
Local Cox model	0.876	0.031	0.810	0.111	-0.22%

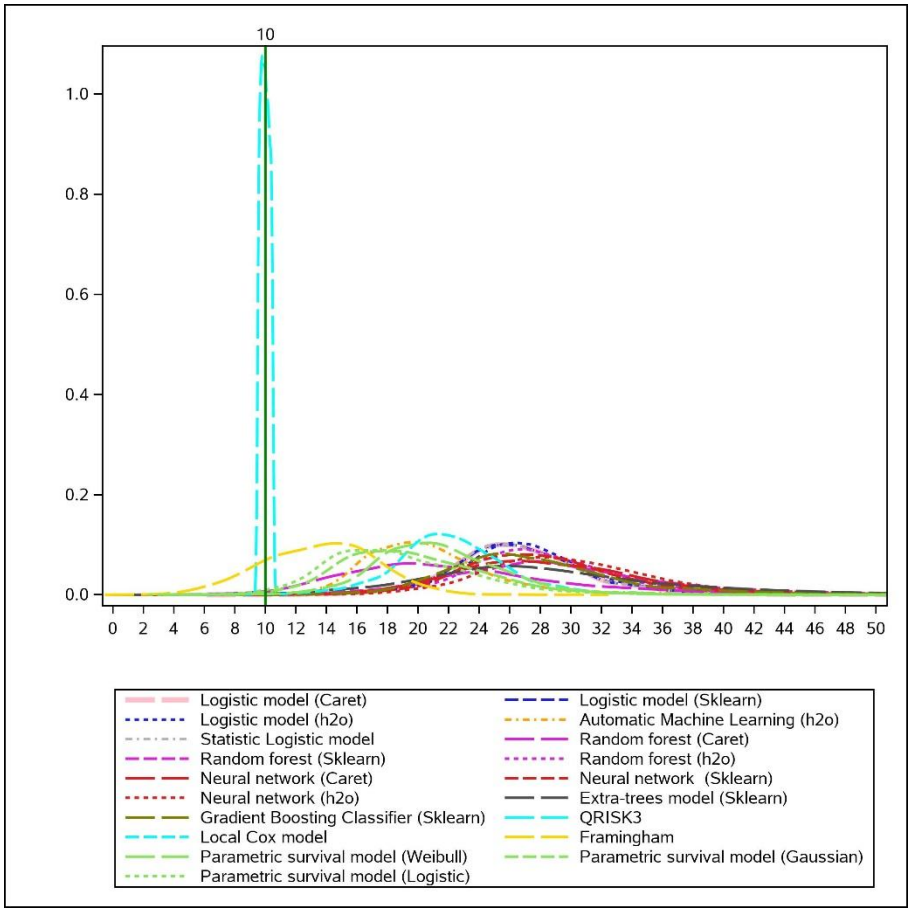
*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**
 # Age and gender were always included as predictors in all scenarios.



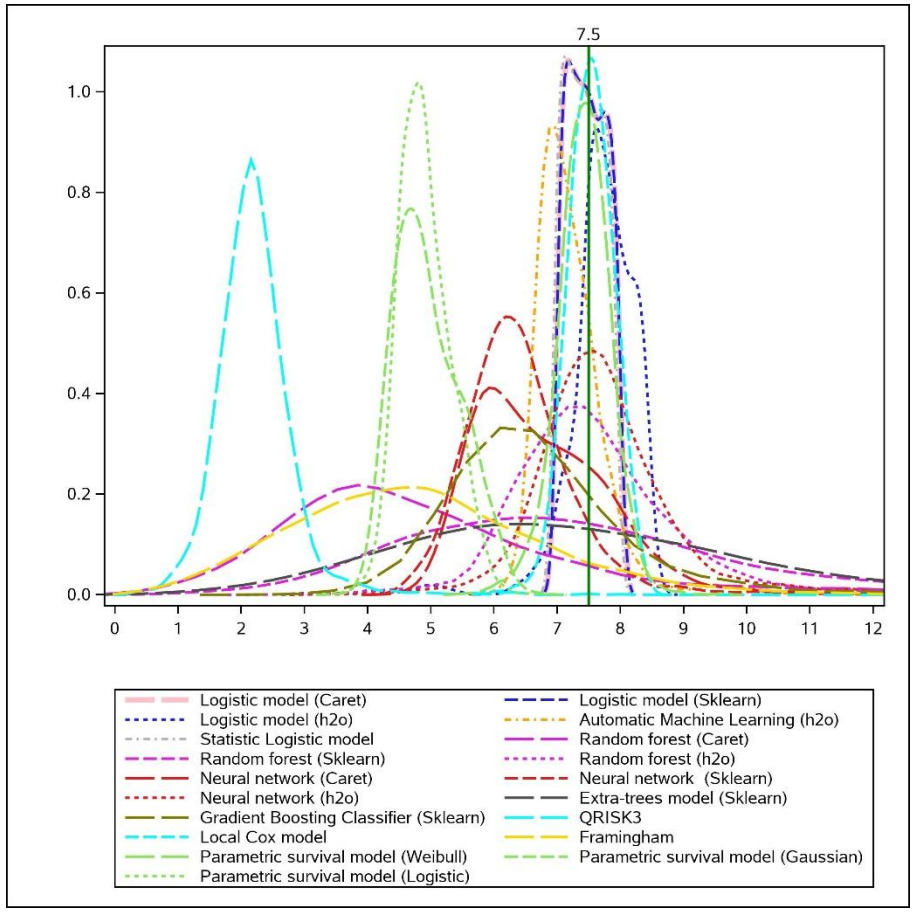
eFigure 1. Workflow of sample splitting and model fitting process



eFigure 2.1: Distribution of individual risk predictions with machine learning and statistical models in overall cohort for patients with predicted CVD risks of 7%~8% in the logistic Caret model

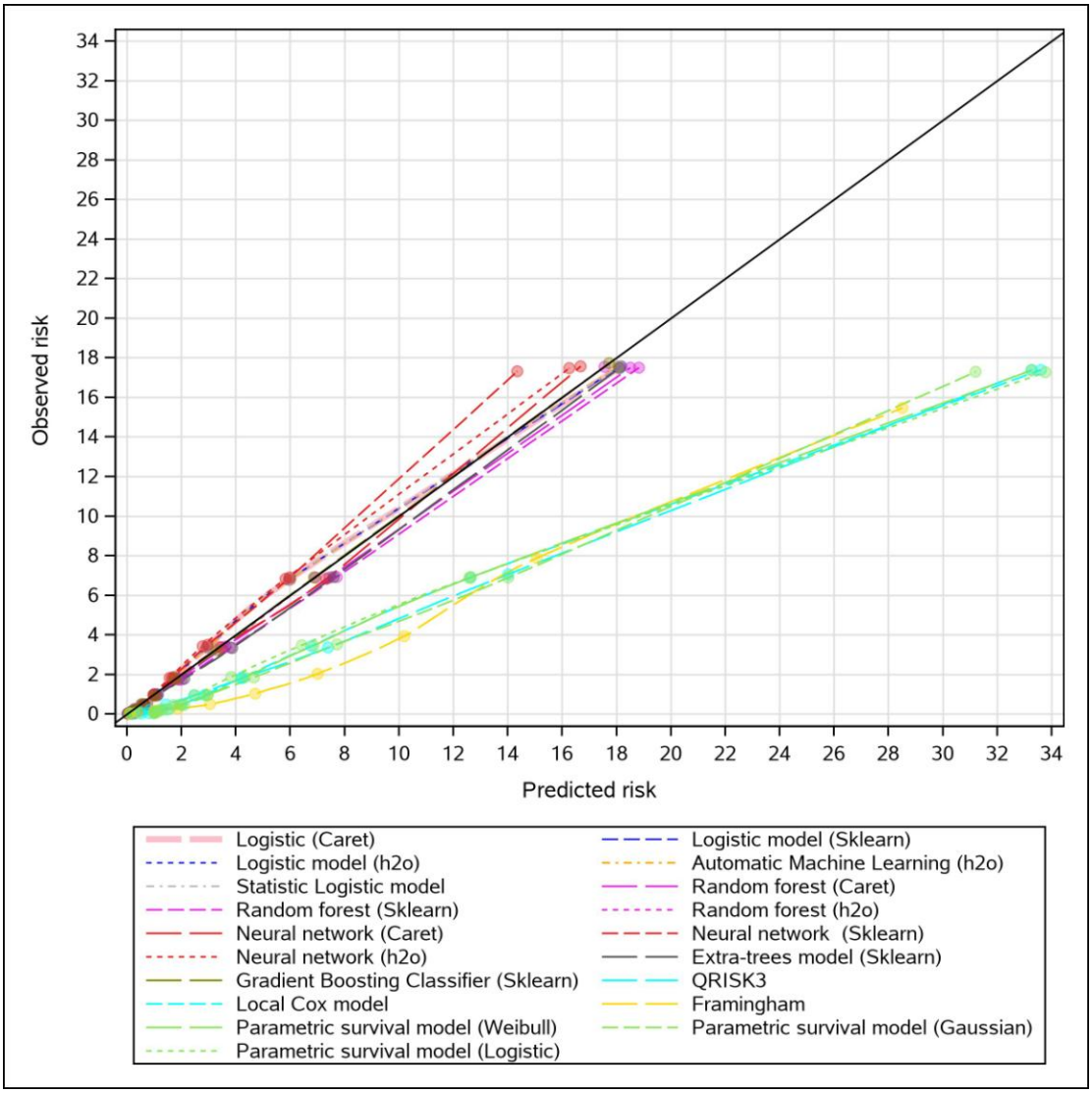


eFigure2.2a



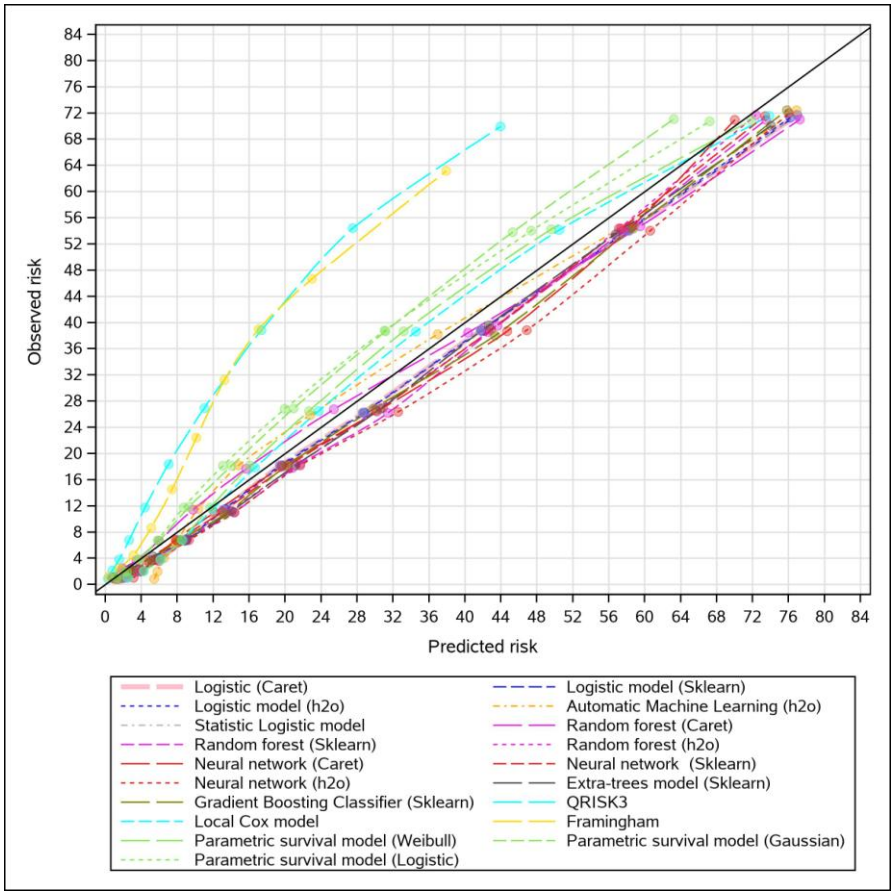
eFigure2.2b

eFigure 2.2: Distribution of individual risk predictions with machine learning and statistical models in cohort without censoring

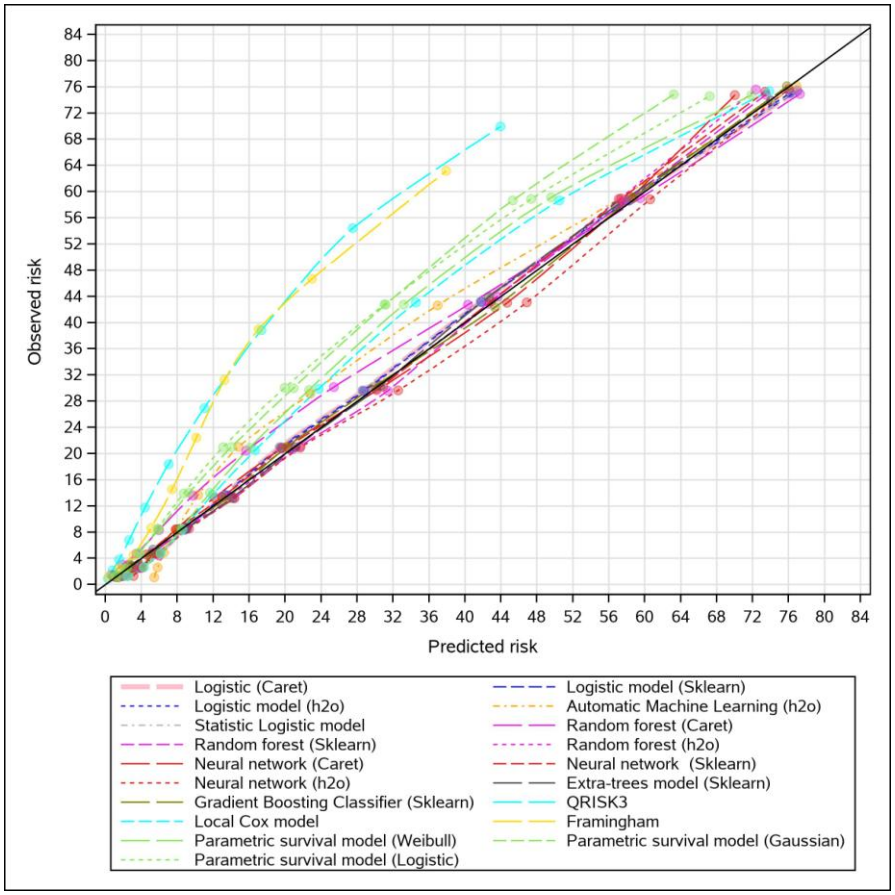


eFigure3.1

eFigure 3.1 Calibration slope of machine learning models and statistical models in overall cohort in binary framework (Observed events did not consider censoring)

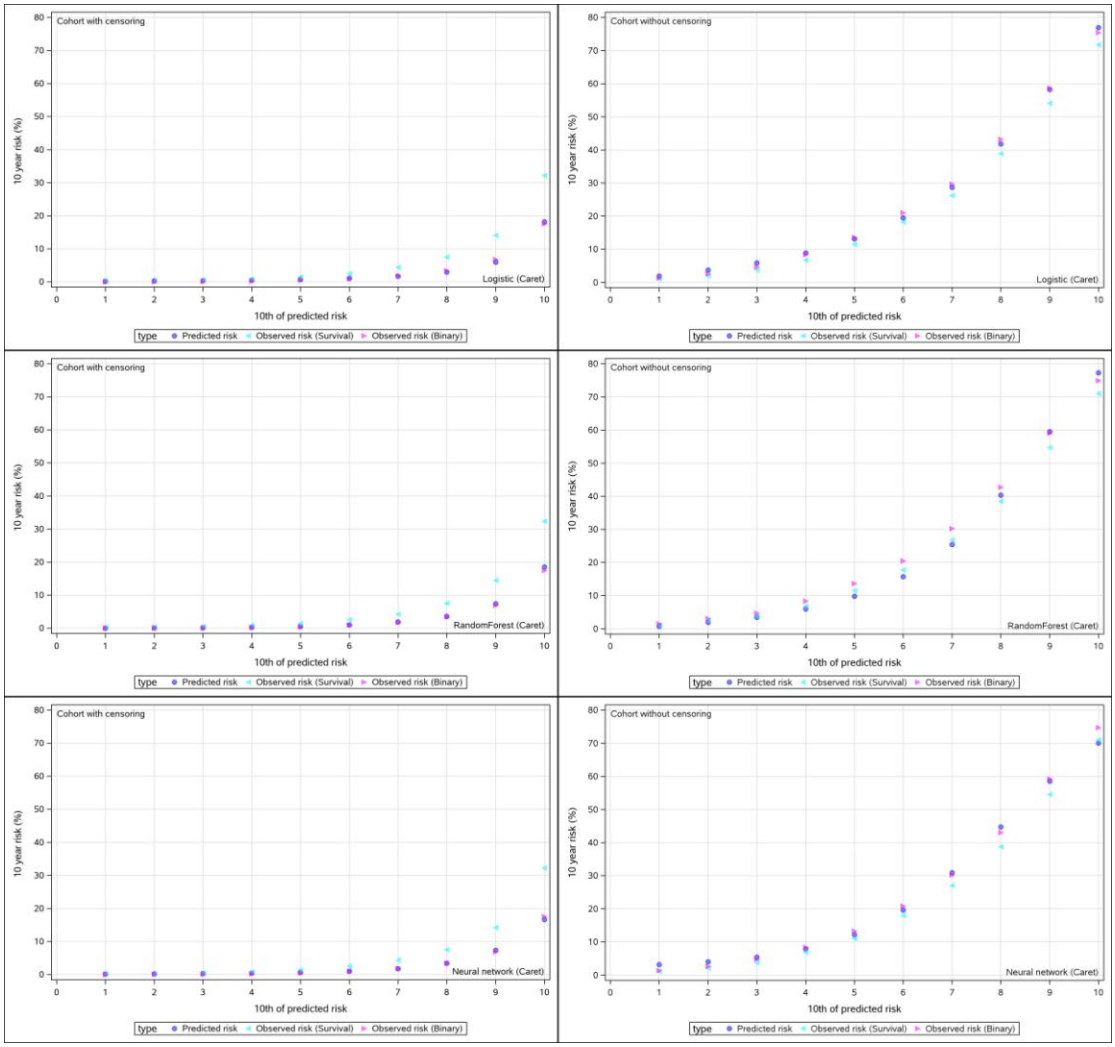


eFigure3.2a

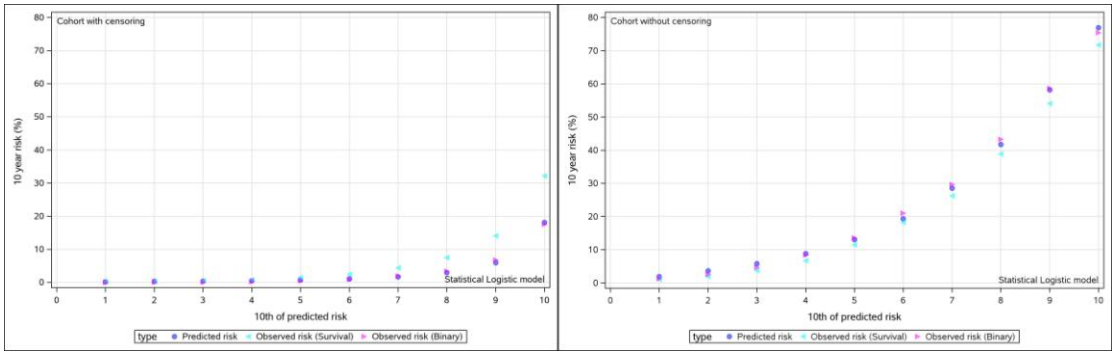


eFigure3.2b

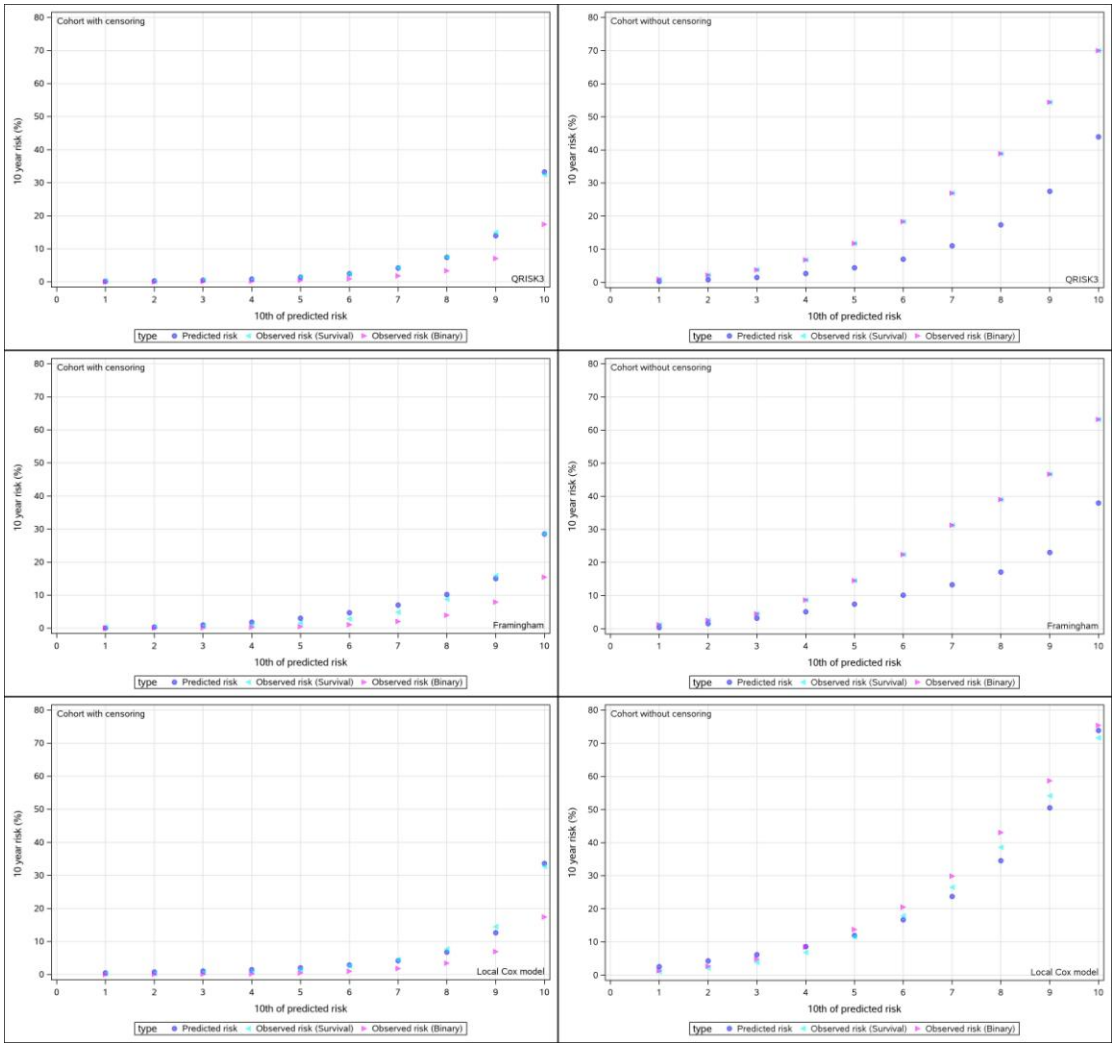
eFigure 3.2. Calibration slope of machine learning models and statistical models in cohort without censoring



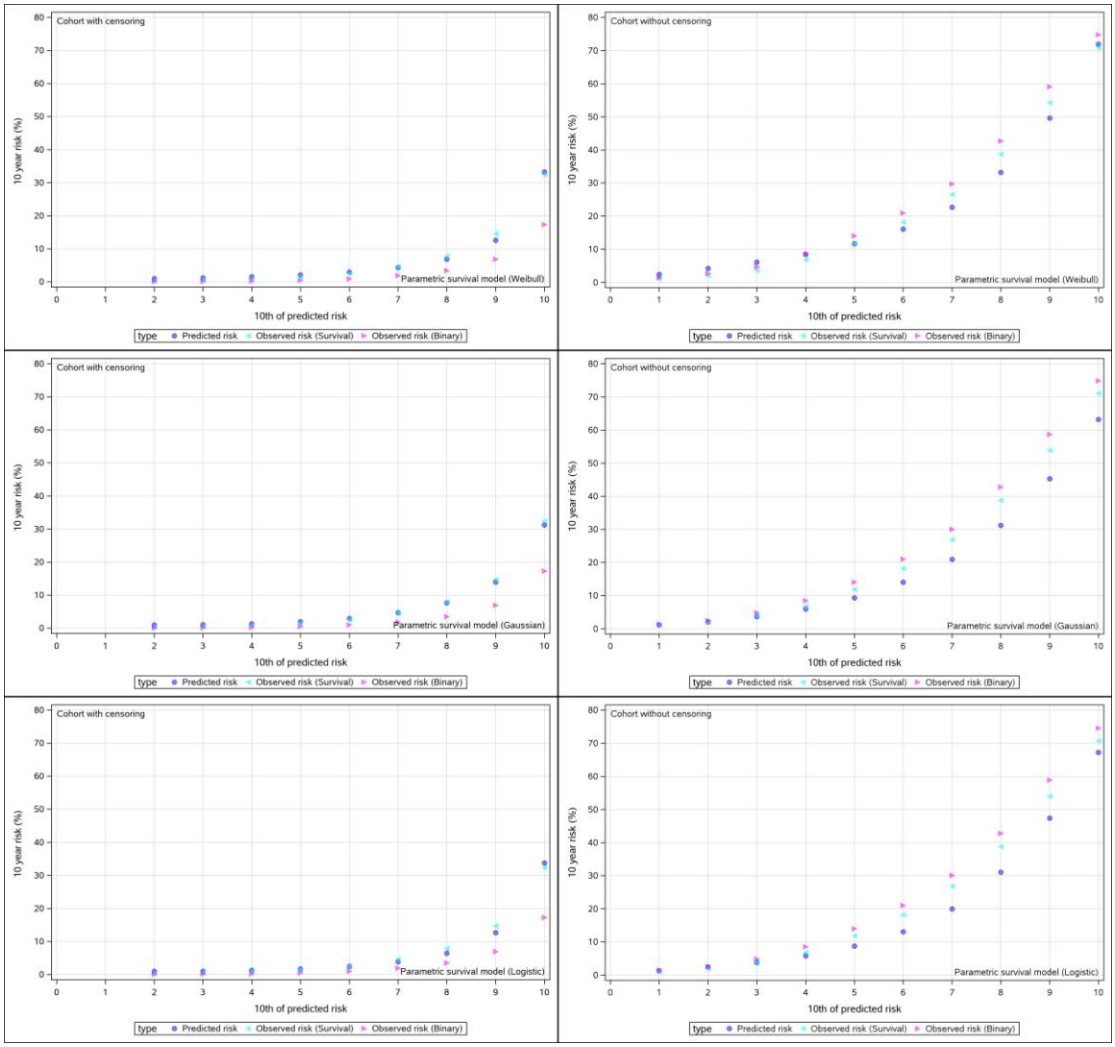
eFigure 4.1. Calibration plots in machine learning models of Caret in overall cohort and cohort without censoring



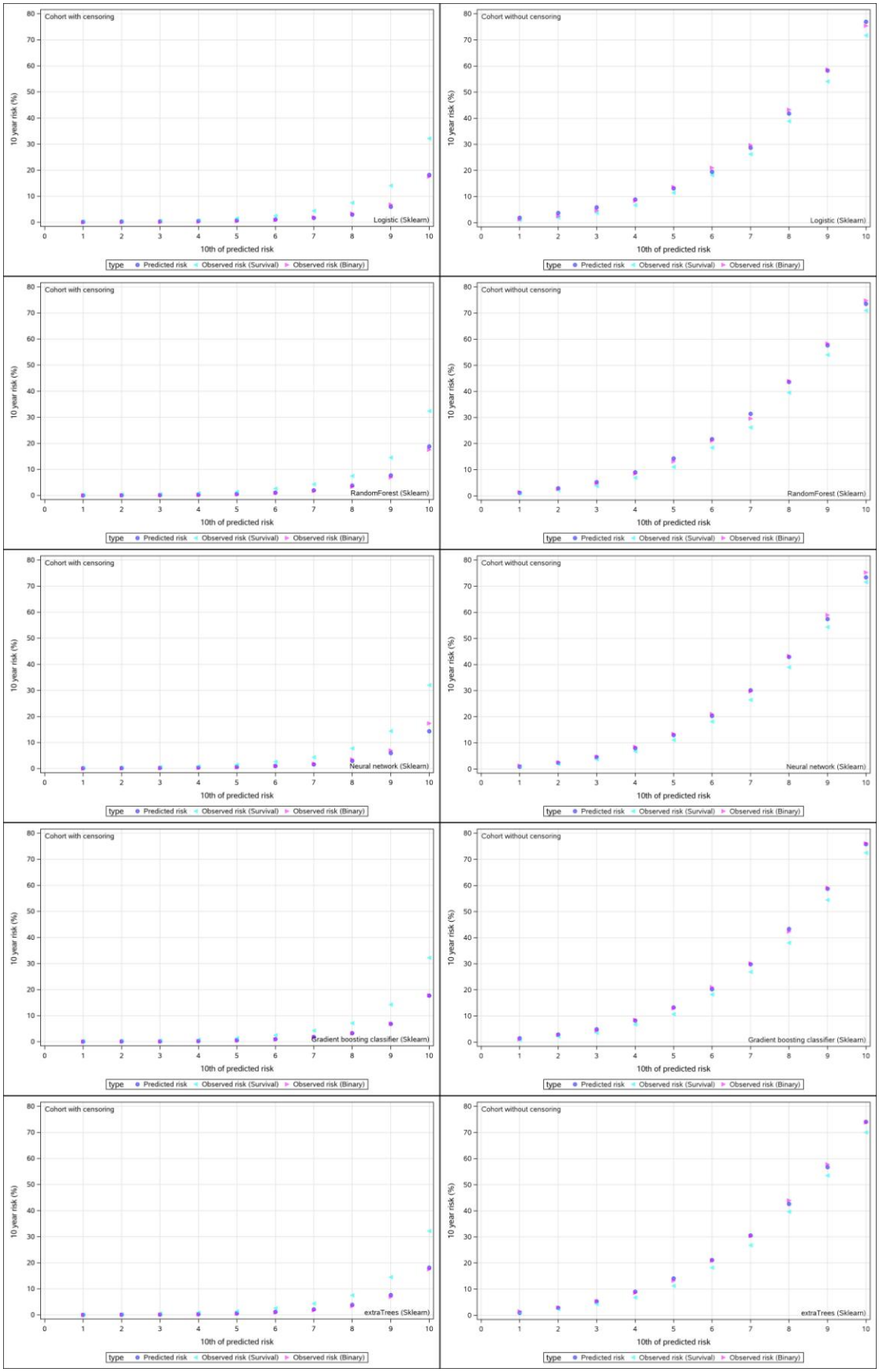
eFigure 4.2. Calibration plots in statistical logistic models in overall cohort and cohort without censoring



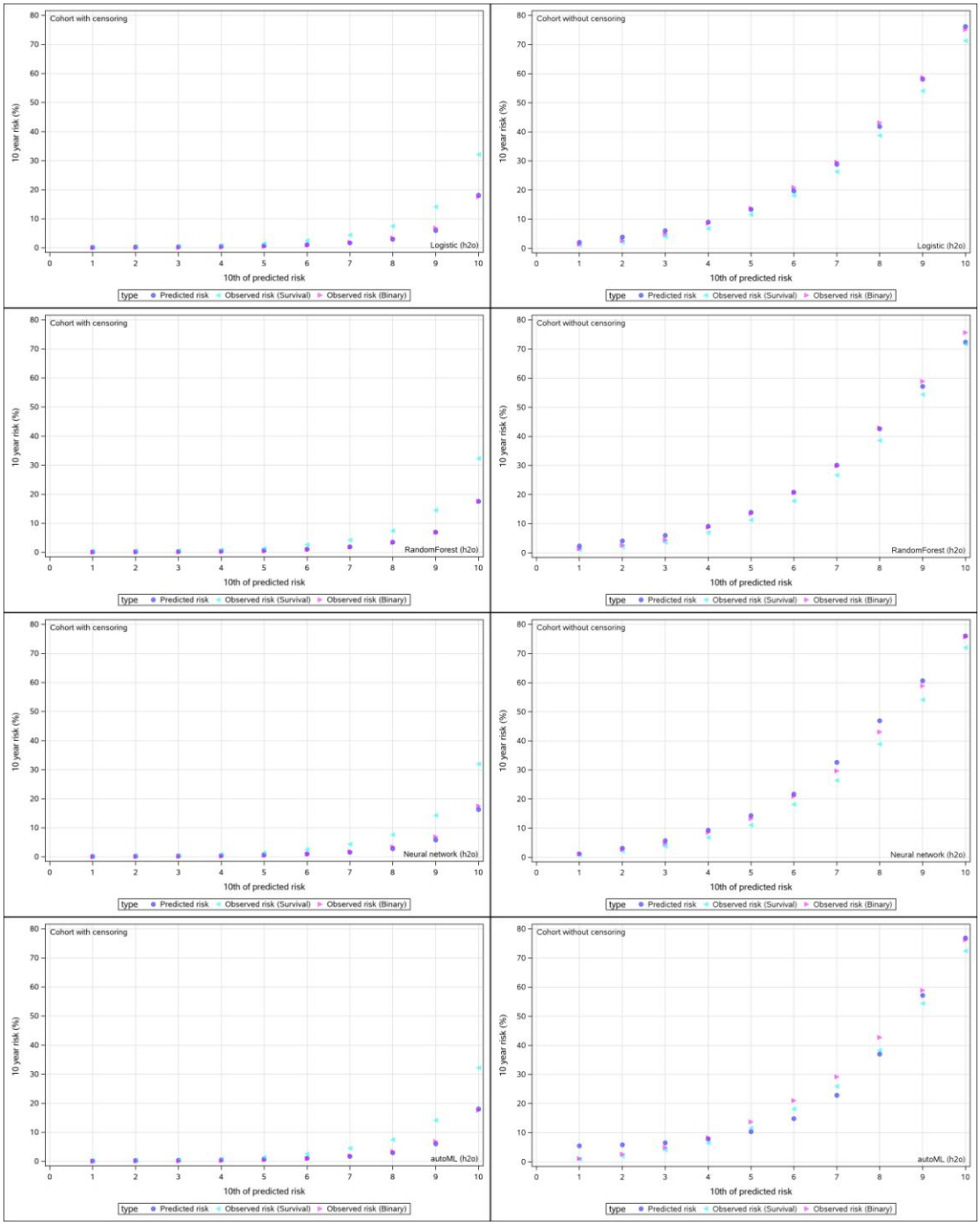
eFigure 4.3. Calibration plots in Cox proportional hazard models in overall cohort and cohort without censoring



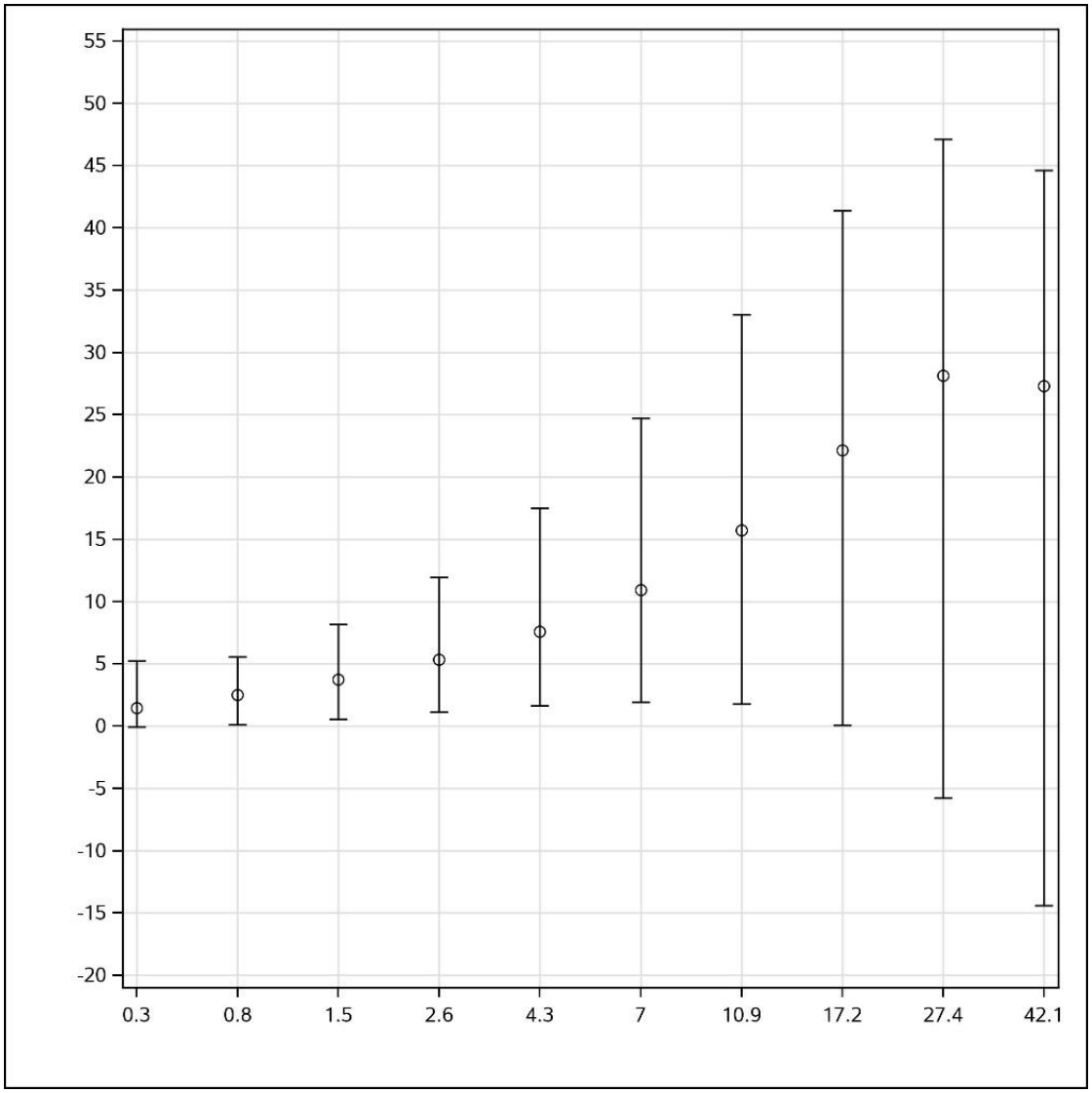
eFigure 4.4. Calibration plots in parametric survival models in overall cohort and cohort without censoring



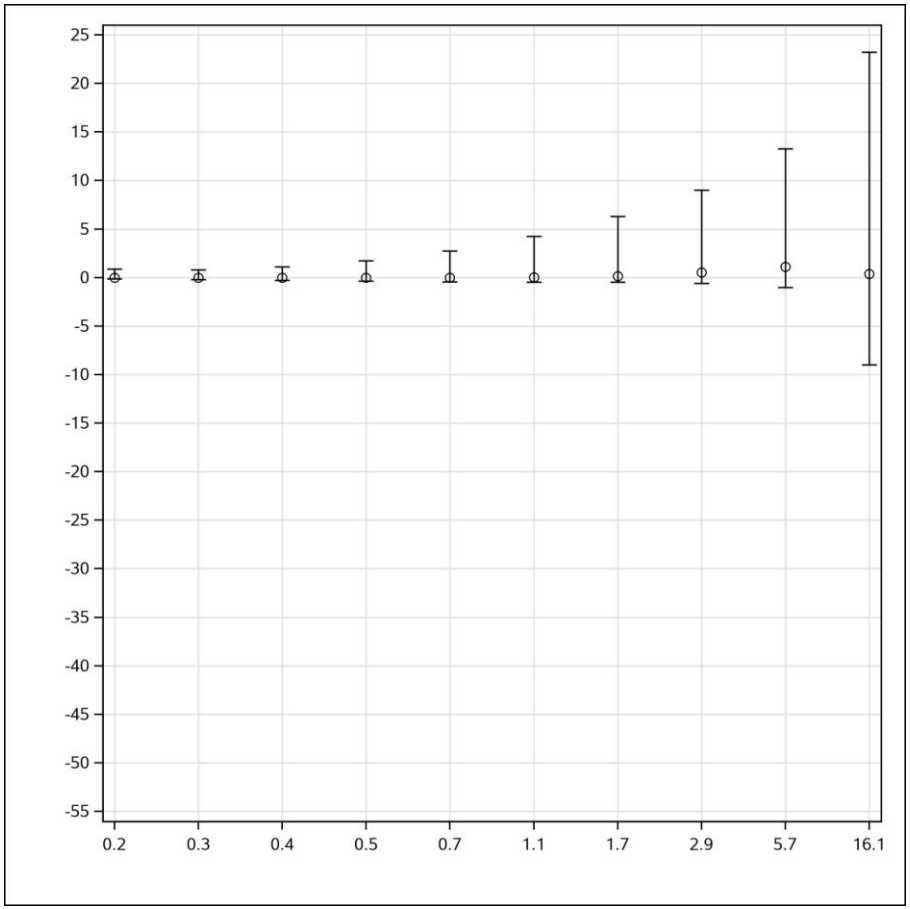
eFigure 4.5. Calibration plots in machine learning models of Sklearn in overall cohort and cohort without censoring



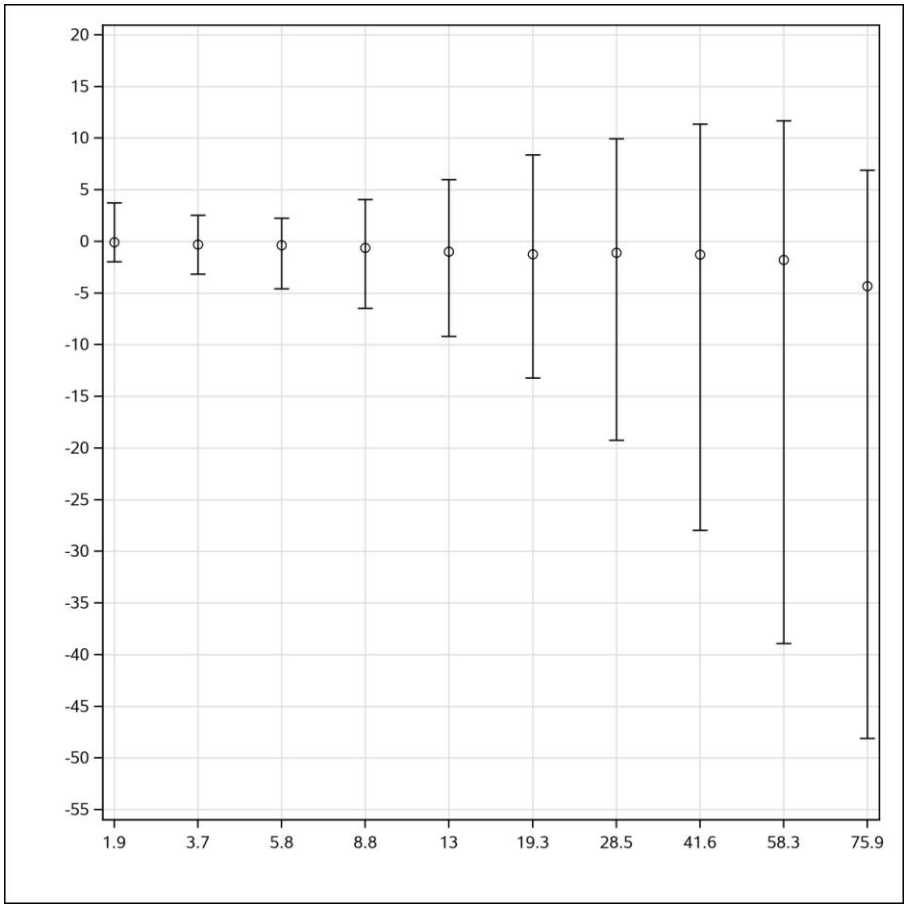
eFigure 4.6. Calibration plots in machine learning models of h2o in overall cohort and cohort without censoring



eFigure 5.1. 95% range of individual risk predictions with machine learning and statistical models stratified by deciles of predicted risks with QRISK3 in cohort without censoring

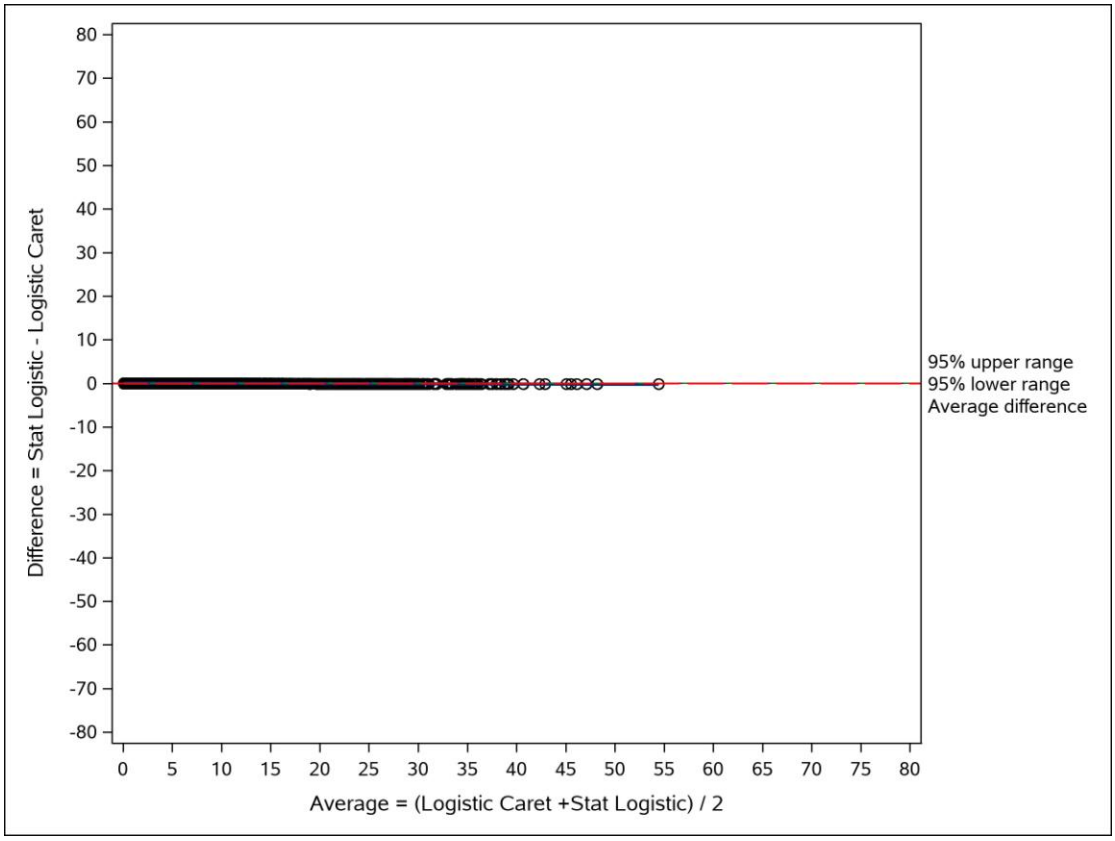


eFigure 5.2a

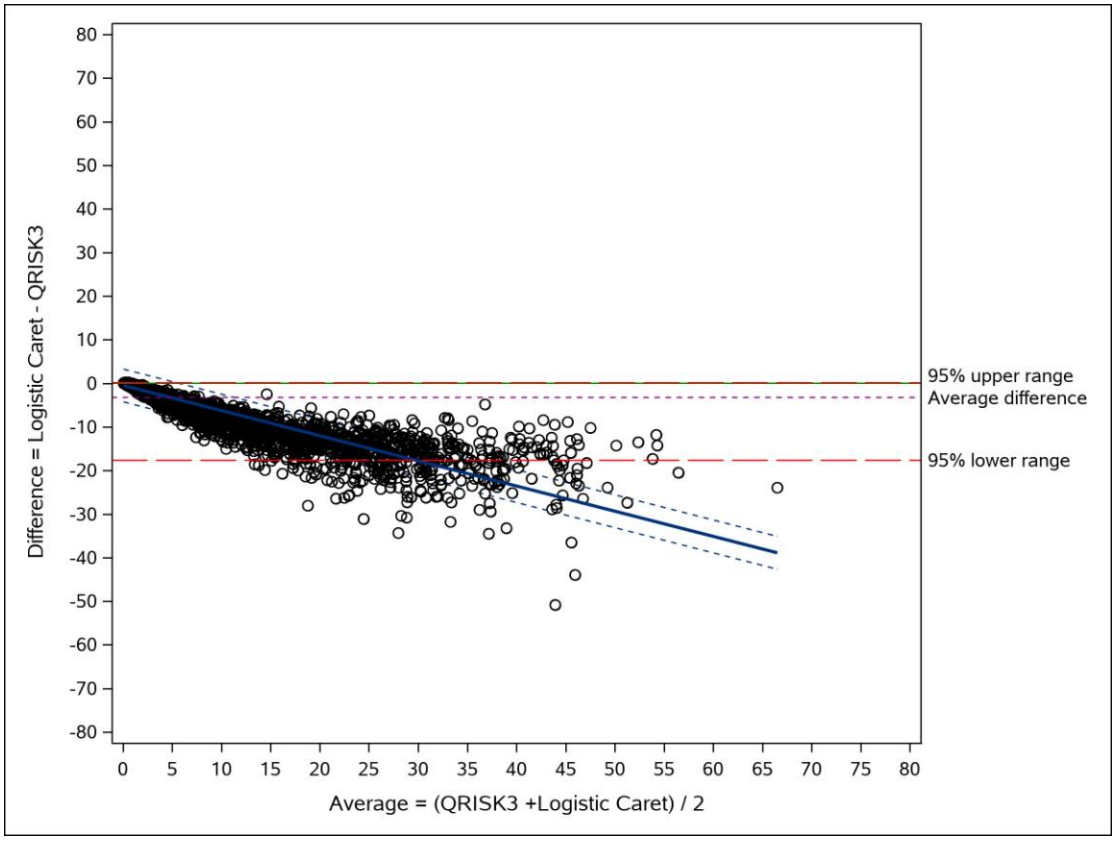


eFigure 5.2b

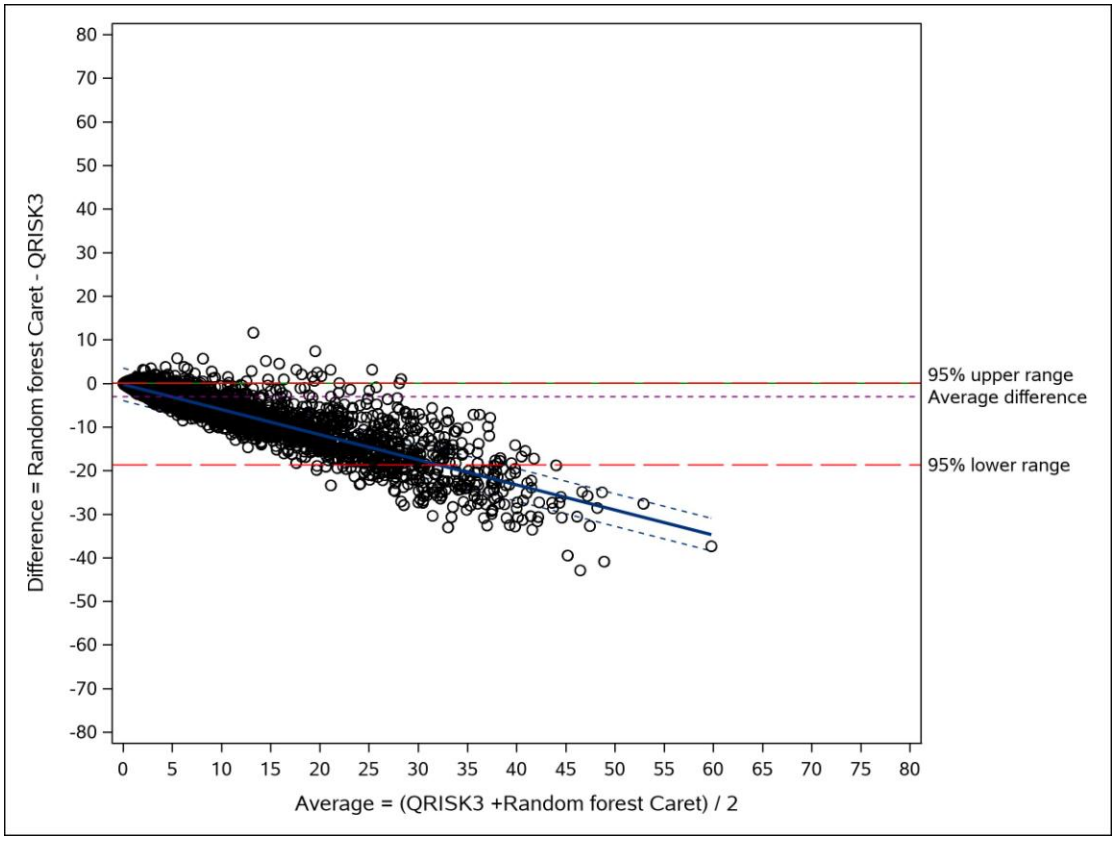
eFigure 5.2. 95% range of individual risk predictions with machine learning and statistical models stratified by deciles of predicted risks with Caret logistic model



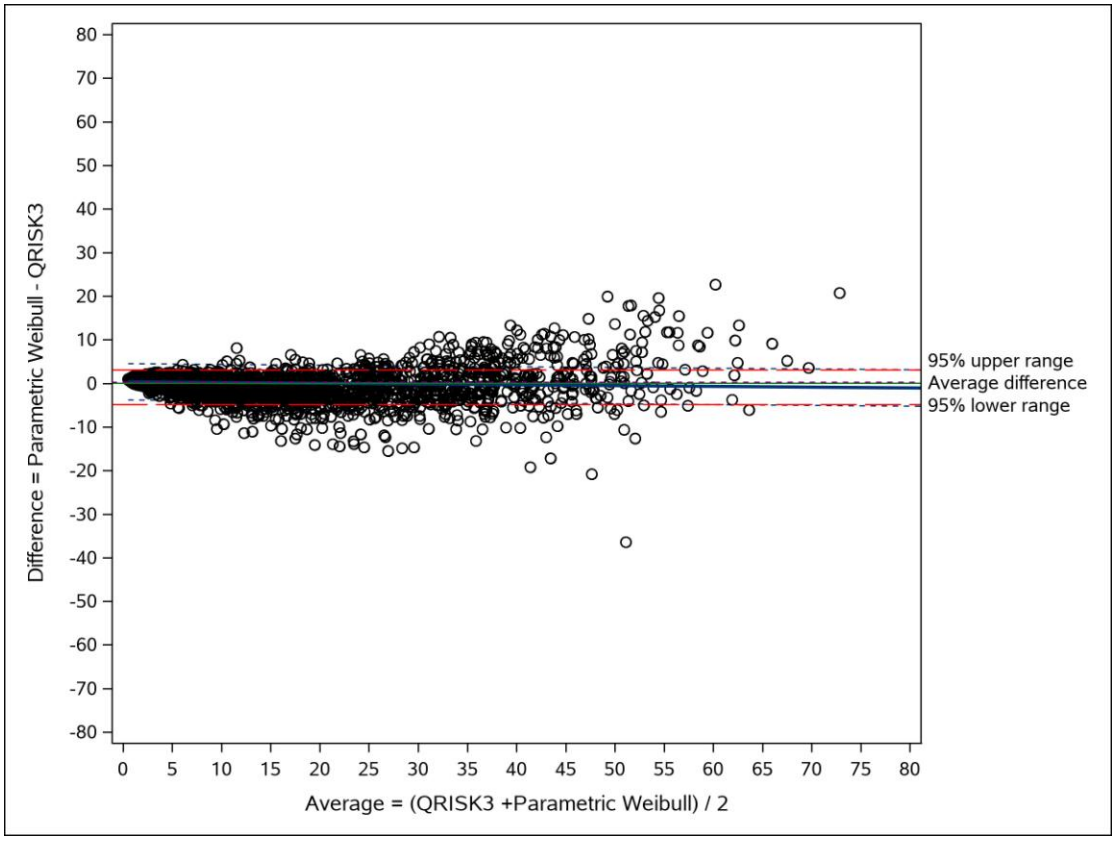
eFigure 6.1 Logistic model fitted with machine learning framework (Caret) comparing to Logistic model fitted statistically



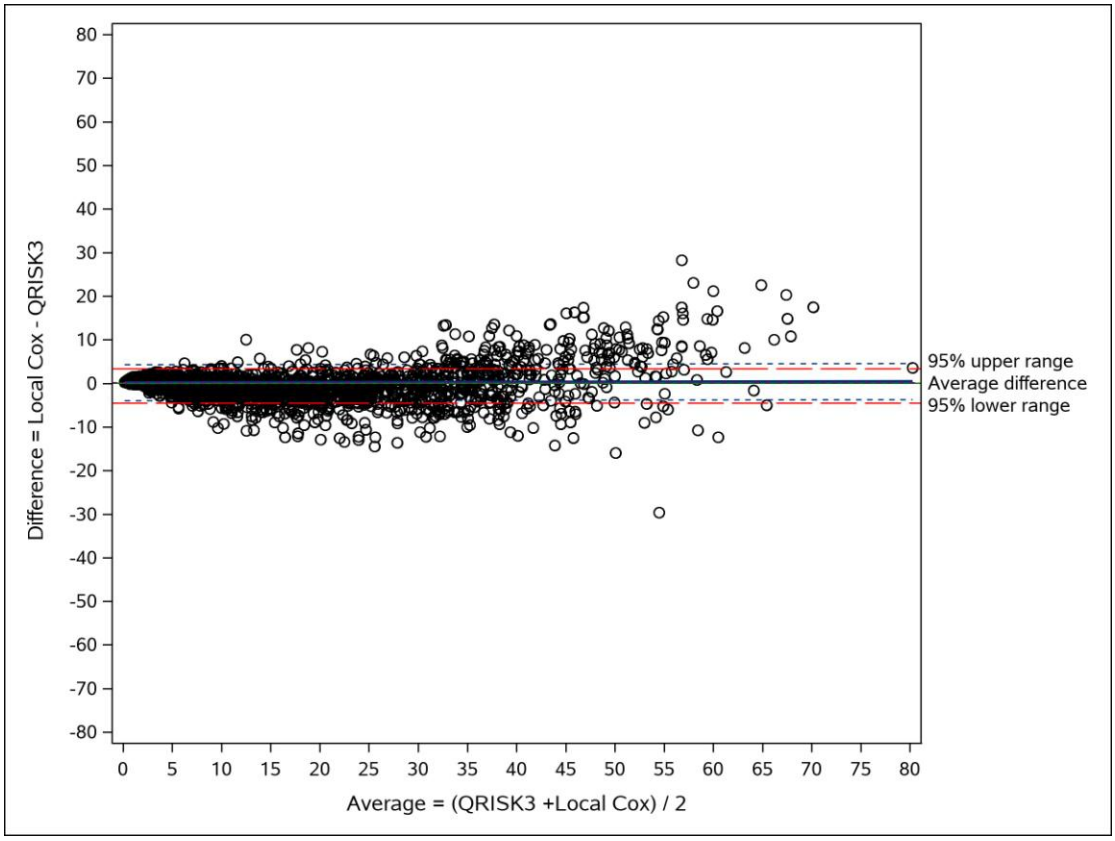
eFigure 6.2 QRISK3 comparing to Logistic model Caret



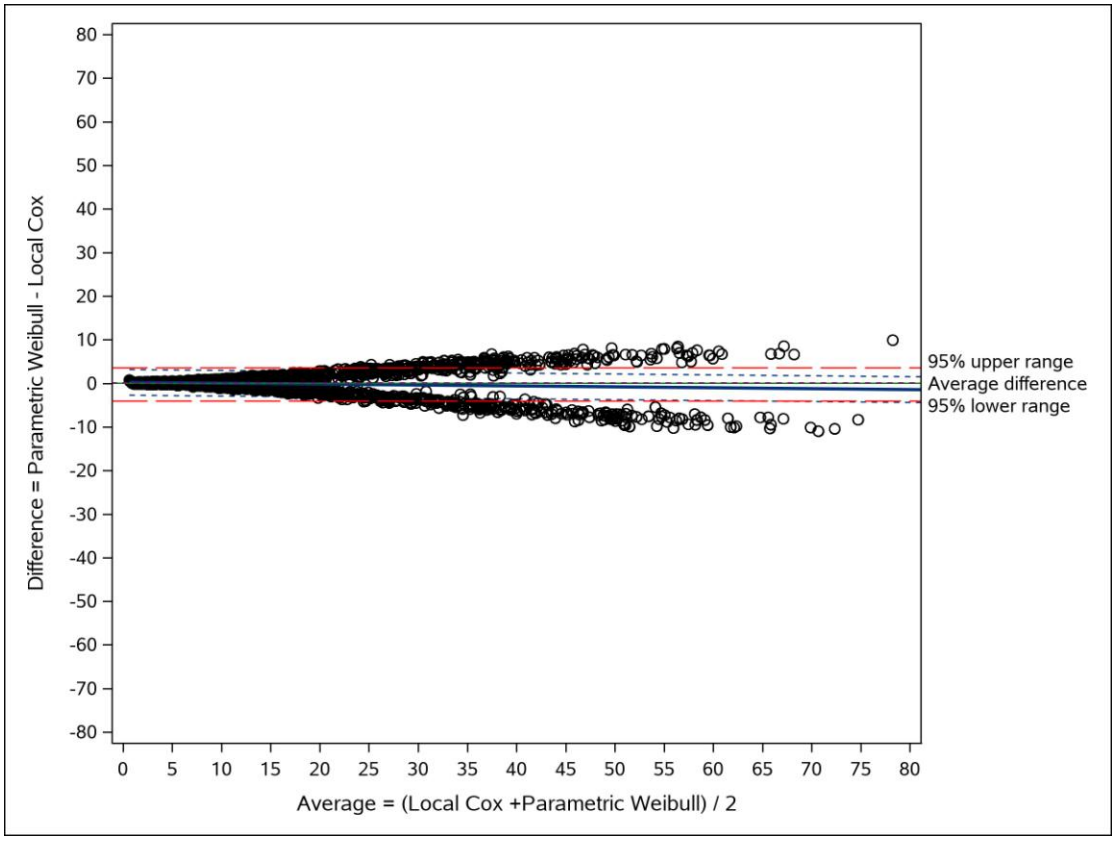
eFigure 6.3 QRISK3 comparing to Random forest Caret



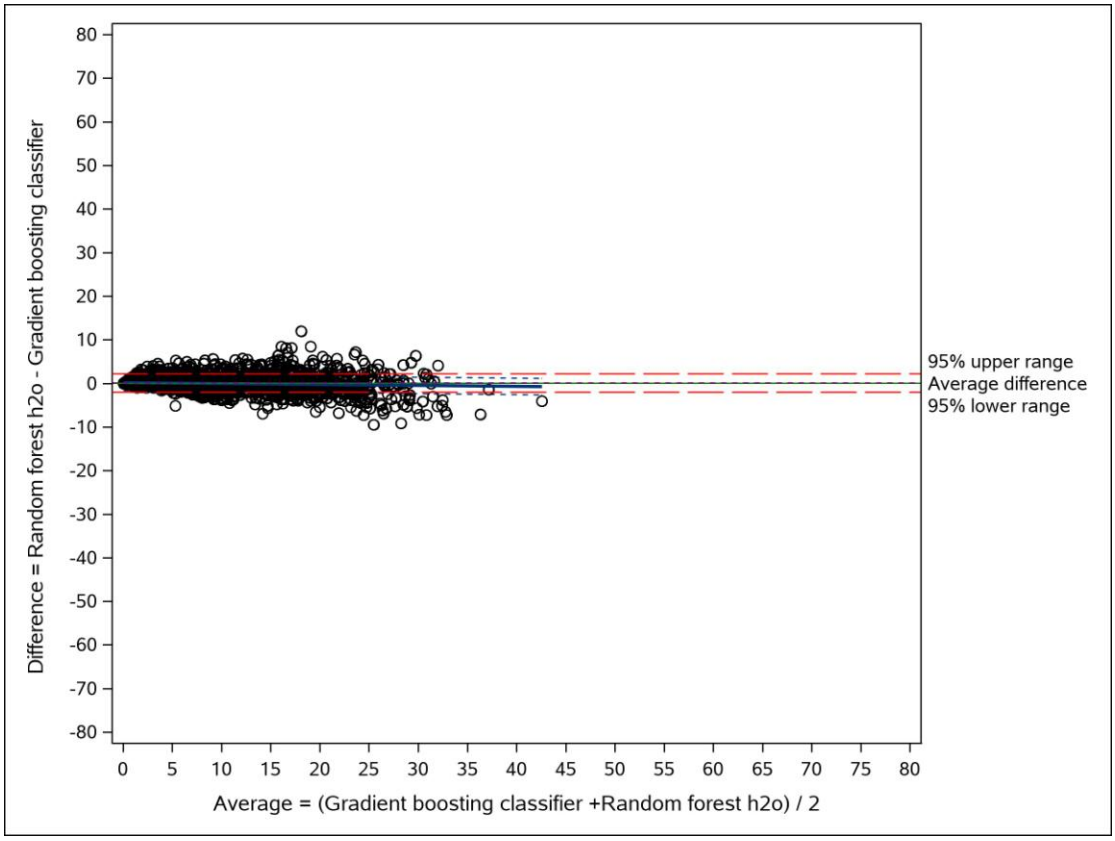
eFigure 6.4 QRISK3 comparing to Parametric Weibull model



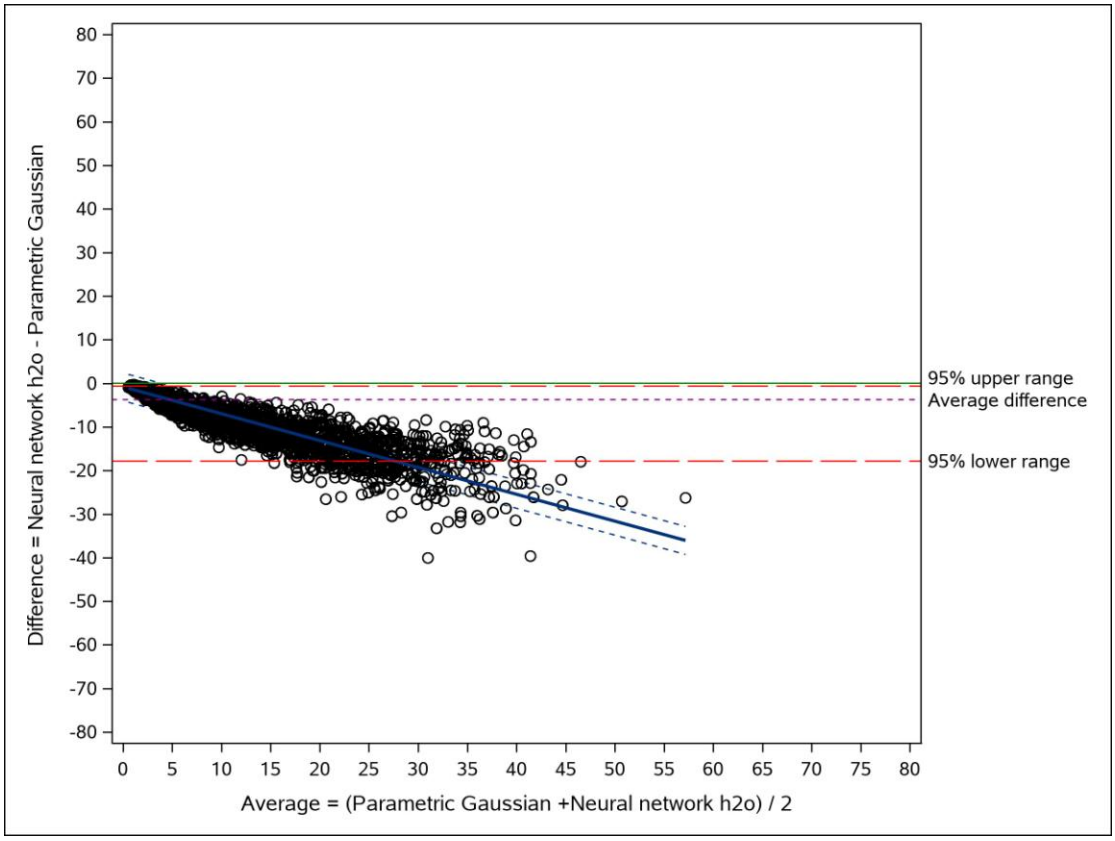
eFigure 6.5 QRISK3 comparing to Local Cox model



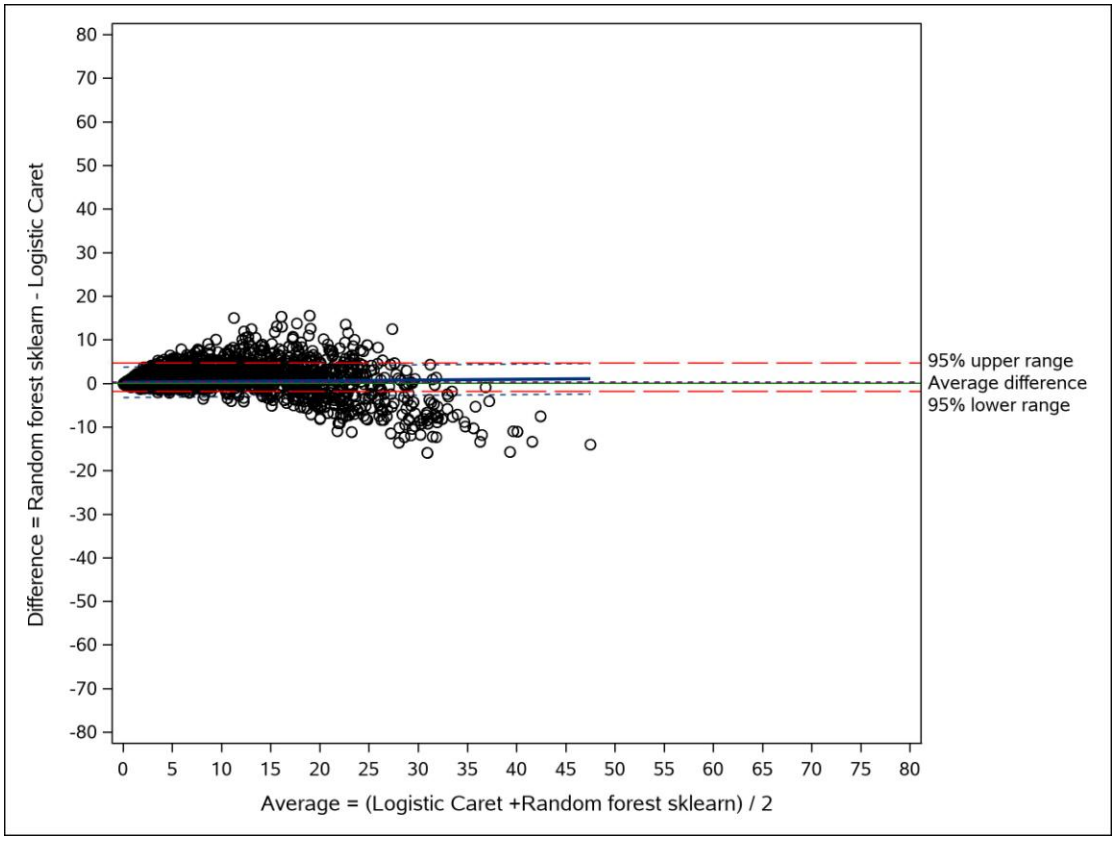
eFigure 6.6 Local Cox model comparing to Parametric Weibull model



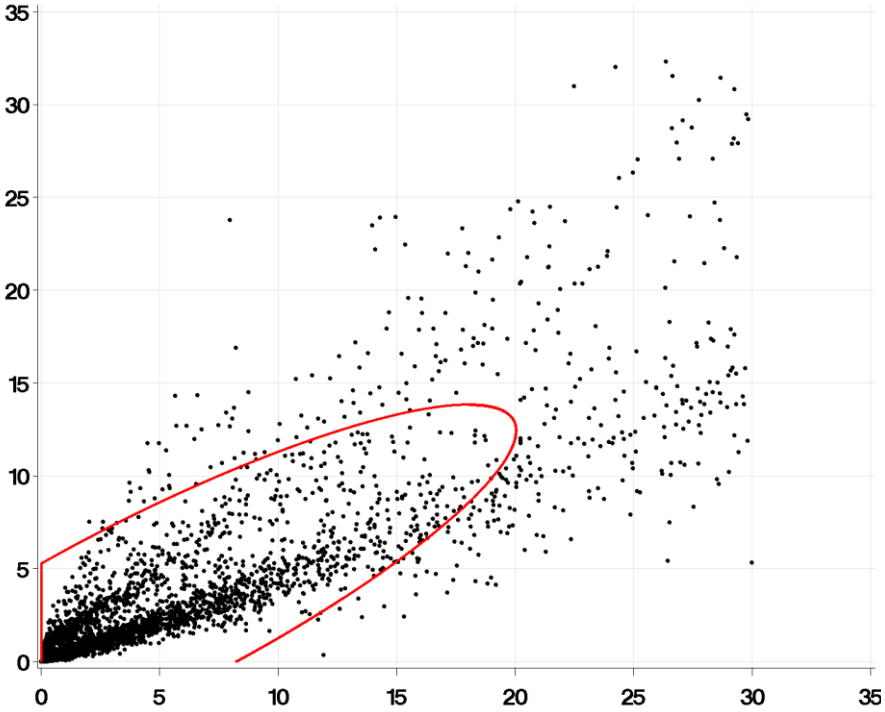
eFigure 6.7 Gradient Boosting Classifier (GBC) sklearn comparing to Random forest h2o



eFigure 6.8 Parametric Gaussian model comparing to Neural network h2o



eFigure 6.9 Logistic model Caret comparing to Random forest Sklearn



eFigure 7a

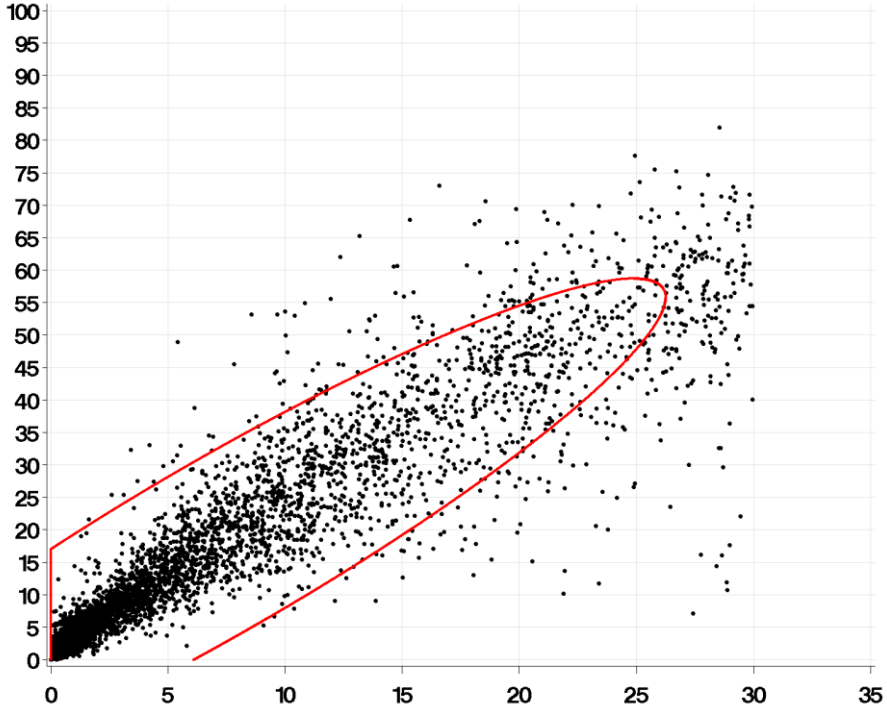
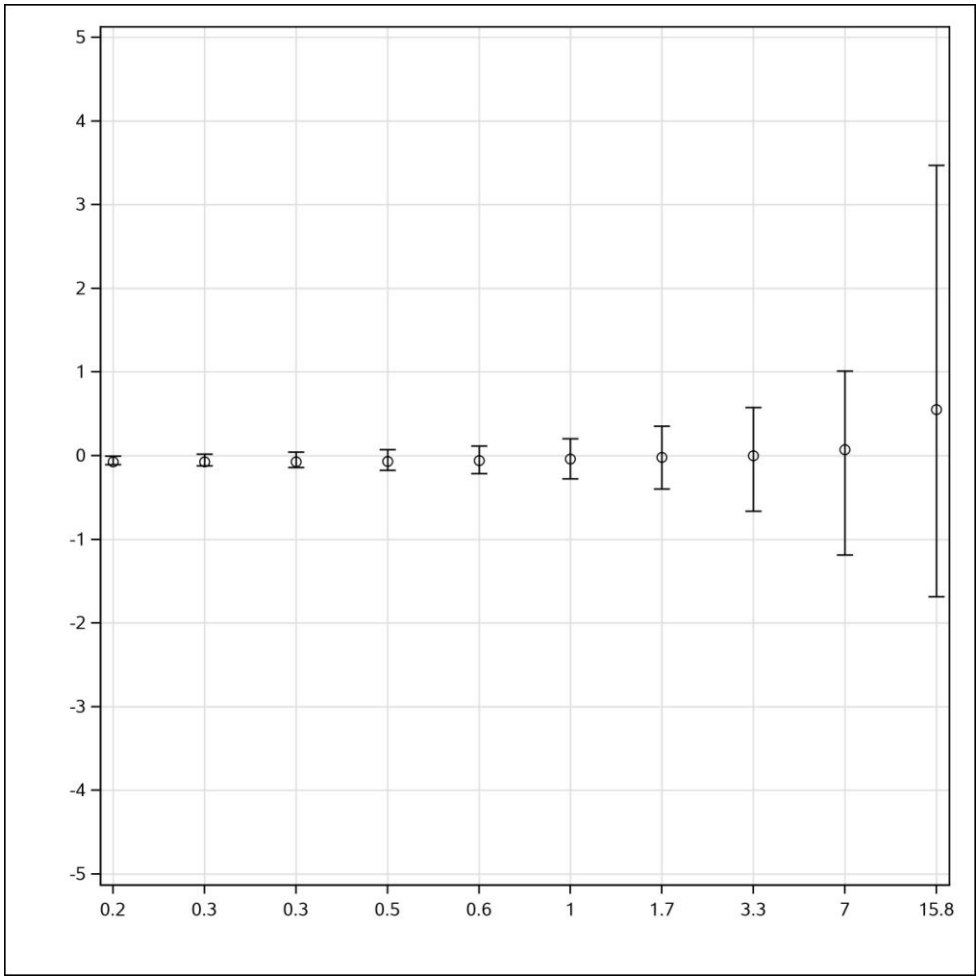
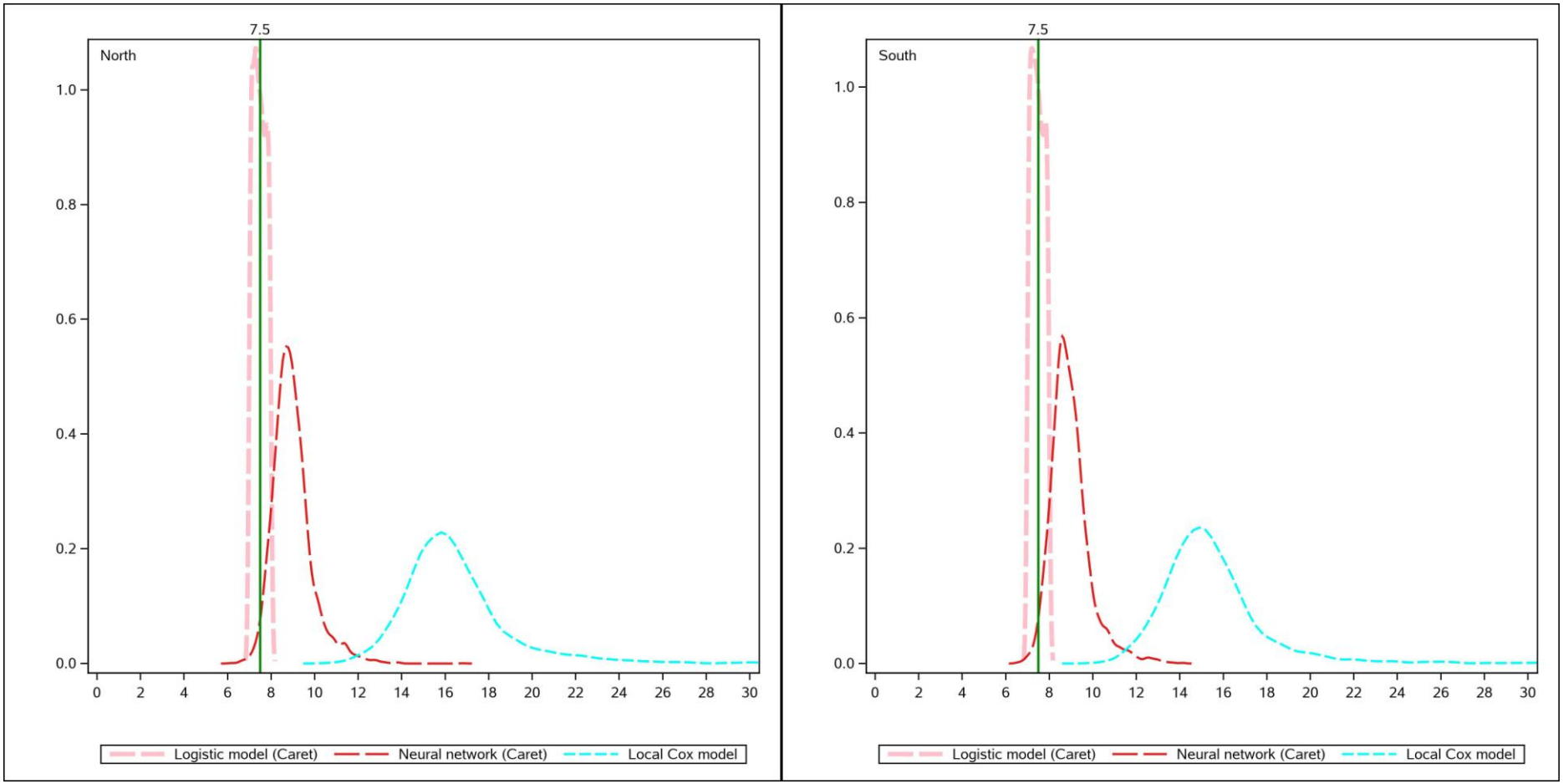


Figure 7b

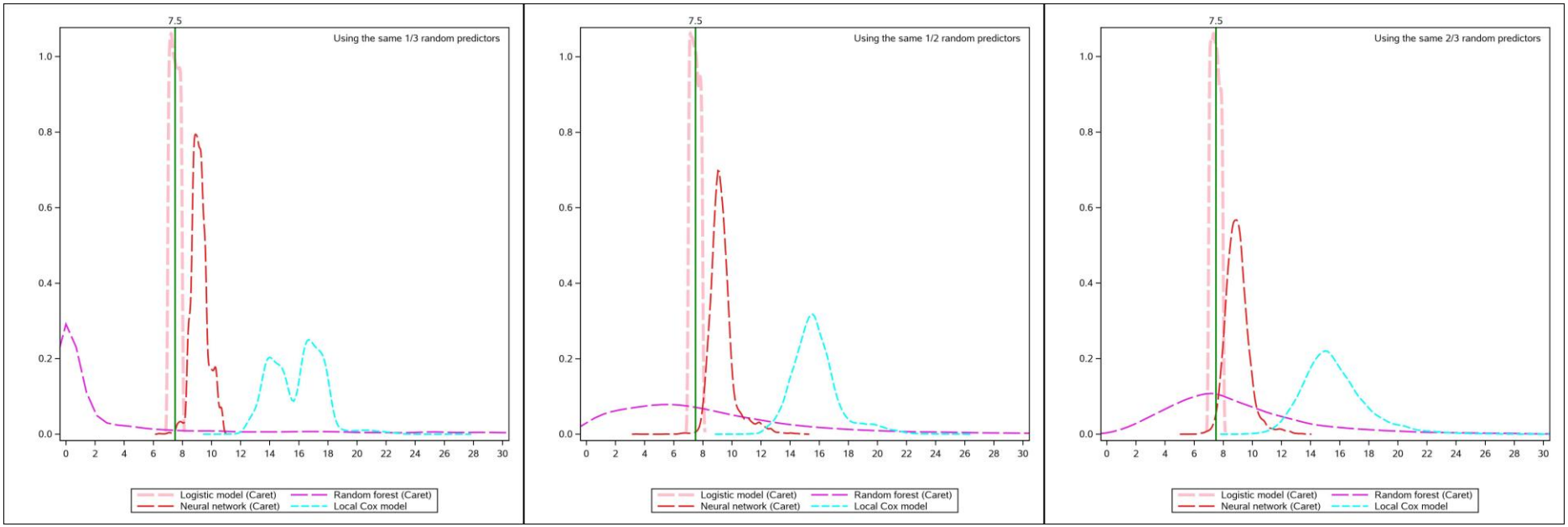
eFigure 7: Inconsistency of individual risk predictions with machine learning and statistical models with Fieller's 95% confidence interval



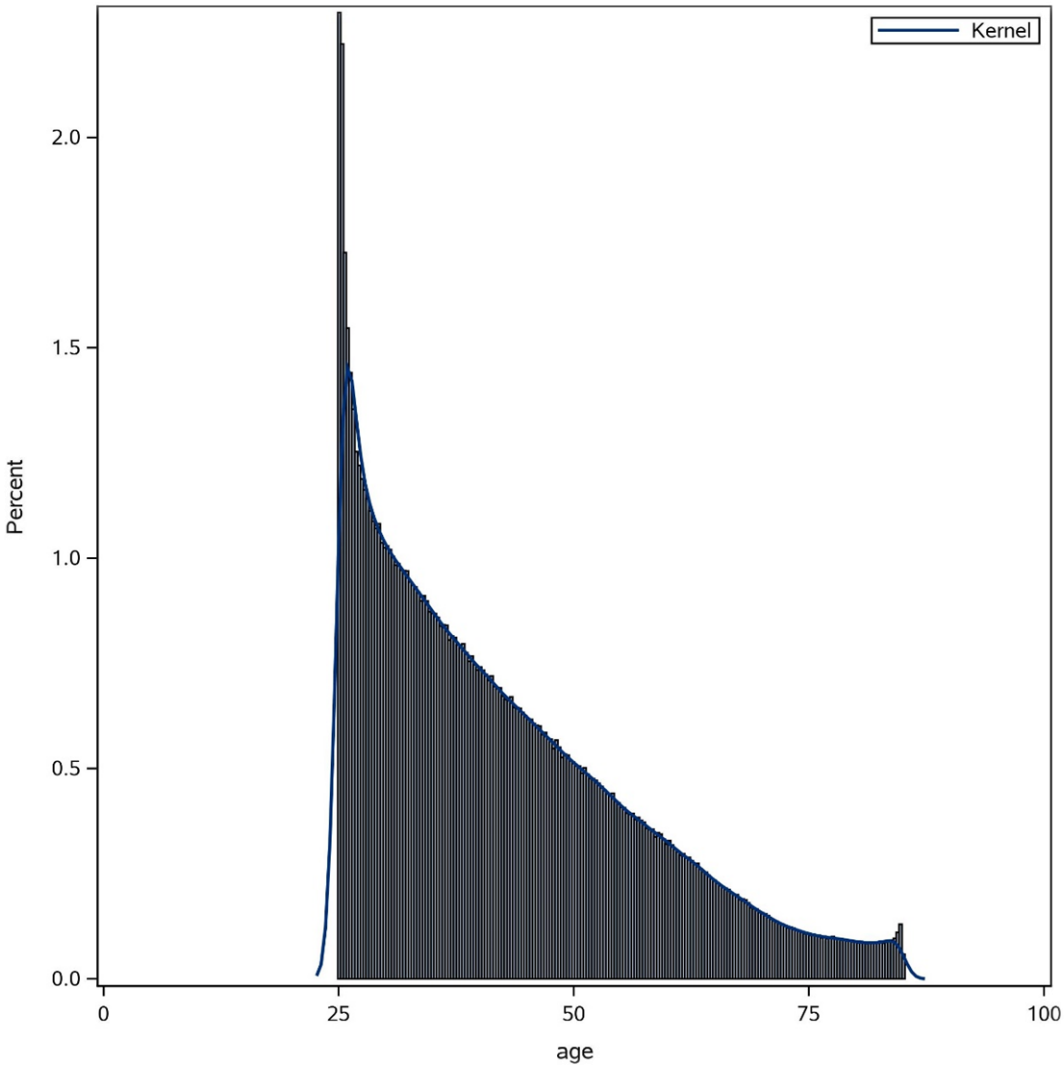
eFigure 8. 95% range of individual risk predictions with Caret neural network models with different grid searched best hyperparameters stratified by deciles of predicted risks with models with the most frequent selected hyperparameters



eFigure 9. Distribution of individual risk predictions with machine learning and statistical models developed in practices from South and tested in practices from North



eFigure 10. Distribution of individual risk predictions with machine learning and statistical models developed with predictors of age and sex plus 1/3, 1/2, 2/3 of all predictors



eFigure 11. Distribution of age among removed patients due to censoring (death patients excluded)