

Supplementary Material.

Title: Accessible Model Predicts Response in Hormone Receptor Positive HER2 Negative Breast Cancer Receiving Neoadjuvant Chemotherapy

Authors:

Luca Mastrantoni, Giovanna Garufi, Giulia Giordano, Noemi Maliziola, Elena Di Monte, Giorgia Arcuri, Valentina Frescura, Angelachiara Rotondi, Armando Orlandi, Luisa Carbognin, Antonella Palazzo, Federica Miglietta, Letizia Pontolillo, Alessandra Fabi, Lorenzo Gerratana, Sergio Pannunzio, Ida Paris, Sara Pilotto, Fabio Marazzi, Antonio Franco, Gianluca Franceschini, Maria Vittoria Dieci, Roberta Mazzeo, Fabio Puglisi, Valentina Guarneri, Michele Milella, Giovanni Scambia, Diana Giannarelli, Giampaolo Tortora, Emilio Bria.

Other Sources:

Interactive WebApp: <https://coralaine-whxxbg8wwtnntoxbxqjhu.streamlit.app>.

Summary

Supplementary Table 1. Patients' characteristics between the training and internal validation cohorts.....	3
Supplementary Table 2. Patients' characteristics between the internal and the external validation cohort.....	4
Supplementary Table 3. Patients' characteristics among the three cohorts.....	5
Supplementary Table 4. Exploratory univariable logistic regression for pCR in the train cohort.	6
Supplementary Table 5. Variance inflation factor (VIF).	7
Supplementary Fig. 1 Feature selection for pCR prediction.....	8
Supplementary Fig. 2 Summary of model performance comparison.	10
Supplementary Fig. 3 Hyperparameters optimization for the best GLM model trained with SGD (SGD Elastic Net).	12
Supplementary Fig. 4 Performance of the best GLM model trained with SGD (SGD Elastic Net).	13
Supplementary Fig. 5 Performance of the best models in the external validation dataset.	14
Supplementary Fig. 6 Partial dependence plots comparison for the SGD Elastic Net model.....	15
Supplementary Fig. 7 Patient levels predictions of the SGD Elastic Net model.	16
Supplementary Fig. 8 Association between pCR and disease-free survival in the internal validation cohort.....	17
Supplementary Fig. 9 Association between pCR and overall survival in the internal validation cohort.....	18
Supplementary Table 6. TRIPOD-AI Checklist.....	20

		Train	Internal Validation	P-Value
n		463	109	
Menopausal_Status, n (%)	Postmenopausal	191 (41.3)	48 (44.0)	0.673*
	Premenopausal	272 (58.7)	61 (56.0)	
Age, mean (SD)		49.8 (10.3)	49.6 (11.4)	0.821**
cT, n (%)	T1	44 (9.5)	7 (6.4)	0.021*
	T2	279 (60.3)	54 (49.5)	
	T3	90 (19.4)	26 (23.9)	
	T4	50 (10.8)	22 (20.2)	
cN, n (%)	N0	73 (15.8)	22 (20.2)	0.404*
	N1	229 (49.5)	46 (42.2)	
	N2	131 (28.3)	31 (28.4)	
	N3	30 (6.5)	10 (9.2)	
Histology, n (%)	Ductal	356 (76.9)	83 (76.1)	0.625*
	Lobular	59 (12.7)	17 (15.6)	
	Other	48 (10.4)	9 (8.3)	
G, n (%)	G1_2	217 (46.9)	46 (42.2)	0.440*
	G3	246 (53.1)	63 (57.8)	
ER, mean (SD)		78.7 (27.2)	79.9 (25.2)	0.654**
PR, mean (SD)		47.8 (37.1)	48.2 (37.2)	0.918**
Ki67, mean (SD)		41.1 (21.8)	42.4 (20.2)	0.571**
HER2, n (%)	Low	284 (61.3)	58 (53.2)	0.147*
	Zero	179 (38.7)	51 (46.8)	
pCR, n (%)	No	391 (84.4)	94 (86.2)	0.749*
	Yes	72 (15.6)	15 (13.8)	
ypT, n (%)	T0	97 (21.0)	19 (17.4)	0.882*
	T1	229 (49.5)	56 (51.4)	
	T2	100 (21.6)	24 (22.0)	
	T3	26 (5.6)	6 (5.5)	
	T4	11 (2.4)	4 (3.7)	
ypN, n (%)	N0	160 (34.6)	39 (35.8)	0.761*
	N1	192 (41.5)	43 (39.4)	
	N2	80 (17.3)	22 (20.2)	
	N3	31 (6.7)	5 (4.6)	

Supplementary Table 1. Patients' characteristics between the training and internal validation cohorts.

An anthracycline and taxane-based regimen was used in nearly all patients (99%), mainly using a sequential schedule (76%).

* Chi-squared test. ** Two sample t-test.

		Internal	External Validation	P-Value
n		572	151	
Menopausal_Status, n (%)	Postmenopausal	239 (41.8)	48 (31.8)	0.032*
	Premenopausal	333 (58.2)	103 (68.2)	
Age, mean (SD)		49.8 (10.5)	48.7 (9.1)	0.231**
cT, n (%)	T1	51 (8.9)	8 (5.3)	0.380*
	T2	333 (58.2)	86 (57.0)	
	T3	116 (20.3)	33 (21.9)	
	T4	72 (12.6)	24 (15.9)	
cN, n (%)	N0	95 (16.6)	21 (13.9)	<0.001*
	N1	275 (48.1)	112 (74.2)	
	N2	162 (28.3)	13 (8.6)	
	N3	40 (7.0)	5 (3.3)	
Histology, n (%)	Ductal	439 (76.7)	77 (51.0)	<0.001*
	Lobular	76 (13.3)	53 (35.1)	
	Other	57 (10.0)	21 (13.9)	
G, n (%)	G1_2	263 (46.0)	81 (53.6)	0.113*
	G3	309 (54.0)	70 (46.4)	
ER, mean (SD)		78.9 (26.8)	86.7 (8.9)	<0.001**
PR, mean (SD)		47.9 (37.1)	59.7 (33.1)	<0.001**
Ki67, mean (SD)		41.4 (21.5)	28.2 (15.1)	<0.001**
HER2, n (%)	Low	342 (59.8)	102 (67.5)	0.099*
	Zero	230 (40.2)	49 (32.5)	
pCR, n (%)	No	485 (84.8)	138 (91.4)	0.050*
	Yes	87 (15.2)	13 (8.6)	

Supplementary Table 2. Patients' characteristics between the internal and the external validation cohort.

Chemotherapy administration was similar: in the external validation set 97% of patients received an anthracycline and taxane-based regimen and a sequential schedule was used in 76% of patients.

* Chi-squared test. ** Two sample t-test.

		Train	Internal Validation	External Validation	P-Value	Mapping
n		463	109	151		
Menopausal_Status, n (%)	Postmenopausal	191 (41.3)	48 (44.0)	48 (31.8)	0.072*	0
	Premenopausal	272 (58.7)	61 (56.0)	103 (68.2)		1
Age, mean (SD)		49.8 (10.3)	49.6 (11.4)	48.7 (9.1)	0.527**	
cT, n (%)	T1	44 (9.5)	7 (6.4)	8 (5.3)	0.051*	0
	T2	279 (60.3)	54 (49.5)	86 (57.0)		1
	T3	90 (19.4)	26 (23.9)	33 (21.9)		2
	T4	50 (10.8)	22 (20.2)	24 (15.9)		3
cN, n (%)	N0	73 (15.8)	22 (20.2)	21 (13.9)	<0.001*	0
	N1	229 (49.5)	46 (42.2)	112 (74.2)		1
	N2	131 (28.3)	31 (28.4)	13 (8.6)		2
	N3	30 (6.5)	10 (9.2)	5 (3.3)		3
Histology, n (%)	Ductal	356 (76.9)	83 (76.1)	77 (51.0)	<0.001*	0
	Lobular	59 (12.7)	17 (15.6)	53 (35.1)		1
	Other	48 (10.4)	9 (8.3)	21 (13.9)		2
G, n (%)	G1_2	217 (46.9)	46 (42.2)	81 (53.6)	0.167*	0
	G3	246 (53.1)	63 (57.8)	70 (46.4)		1
ER, mean (SD)		78.7 (27.2)	79.9 (25.2)	86.7 (8.9)	0.002**	
PR, mean (SD)		47.8 (37.1)	48.2 (37.2)	59.7 (33.1)	0.002**	
Ki67, mean (SD)		41.1 (21.8)	42.4 (20.2)	28.2 (15.1)	<0.001**	
HER2, n (%)	Low	284 (61.3)	58 (53.2)	102 (67.5)	0.064*	0
	Zero	179 (38.7)	51 (46.8)	49 (32.5)		1
pCR, n (%)	No	391 (84.4)	94 (86.2)	138 (91.4)	0.100*	0
	Yes	72 (15.6)	15 (13.8)	13 (8.6)		1
ypT, n (%)	T0	97 (21.0)	19 (17.4)		0.882	0
	T1	229 (49.5)	56 (51.4)			1
	T2	100 (21.6)	24 (22.0)			2
	T3	26 (5.6)	6 (5.5)			3
	T4	11 (2.4)	4 (3.7)			4
ypN, n (%)	N0	160 (34.6)	39 (35.8)		0.761	0
	N1	192 (41.5)	43 (39.4)			1
	N2	80 (17.3)	22 (20.2)			2
	N3	31 (6.7)	5 (4.6)			3

Supplementary Table 3. Patients' characteristics among the three cohorts.

The mapping for ordinal encoding is shown in the column "mapping".

* Chi-squared test. ** One-way ANOVA

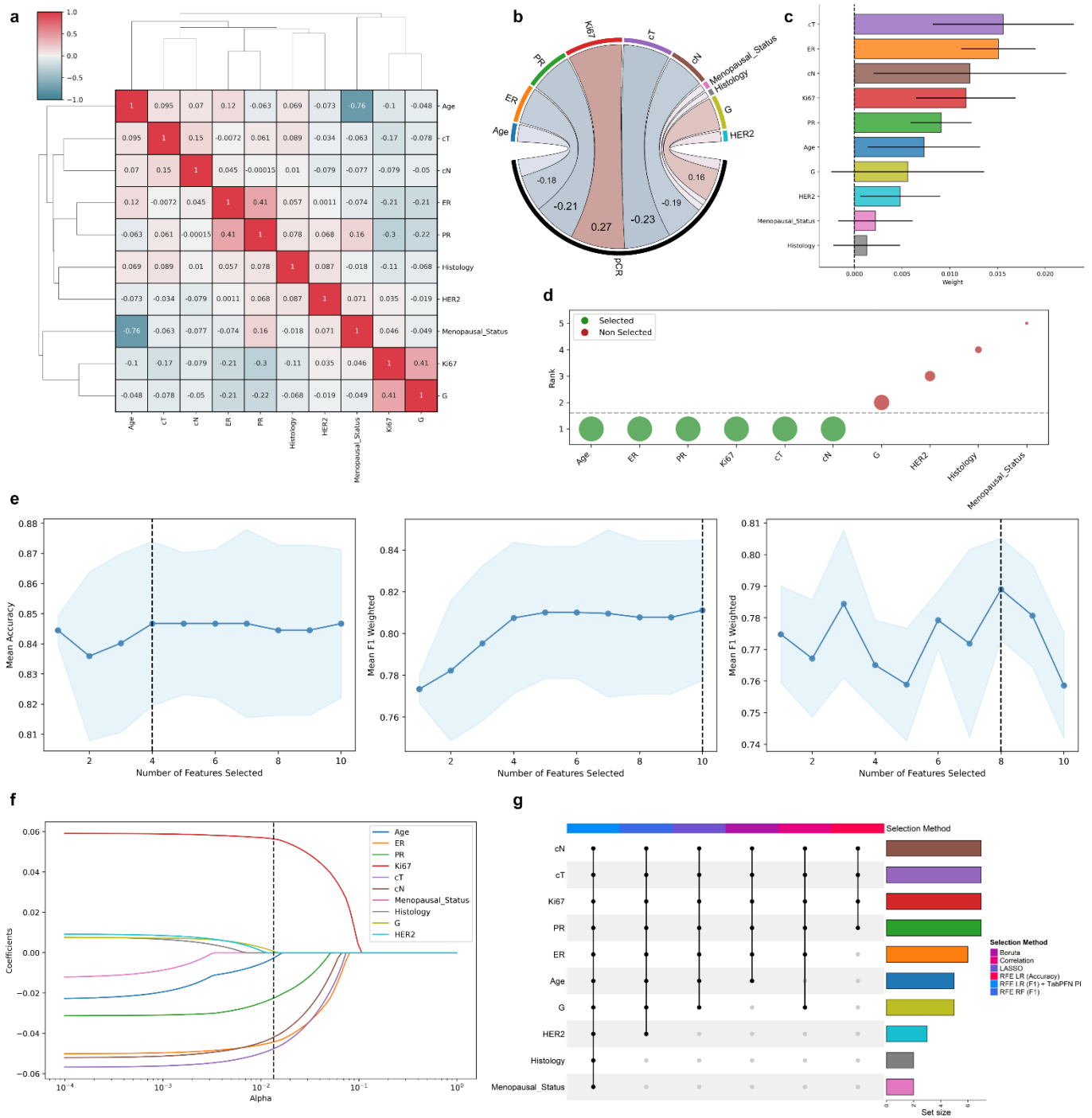
		pCR No	pCR Yes	P-Value
n		391	72	
Menopausal_Status, n (%)	Postmenopausal	164 (41.9)	27 (37.5)	0.482
	Premenopausal	227 (58.1)	45 (62.5)	
Age, mean (SD)		50.2 (10.3)	47.7 (10.5)	0.062
cT, n (%)	T1	29 (7.4)	15 (20.8)	<0.001
	T2	229 (58.6)	50 (69.4)	
	T3	85 (21.7)	5 (6.9)	
	T4	48 (12.3)	2 (2.8)	
cN, n (%)	N0	53 (13.6)	20 (27.8)	<0.001
	N1	190 (48.6)	39 (54.2)	
	N2	118 (30.2)	13 (18.1)	
	N3	30 (7.7)		
Histology, n (%)	Ductal	298 (76.2)	58 (80.6)	0.829
	Lobular	54 (13.8)	5 (6.9)	
	Other	39 (10.0)	9 (12.5)	
G, n (%)	G1_2	197 (50.4)	20 (27.8)	<0.001
	G3	194 (49.6)	52 (72.2)	
ER, mean (SD)		81.5 (23.7)	63.5 (37.9)	<0.001
PR, mean (SD)		51.0 (36.9)	30.5 (33.5)	<0.001
Ki67, mean (SD)		38.5 (20.5)	55.5 (23.0)	<0.001
HER2, n (%)	Low	244 (62.4)	40 (55.6)	0.274
	Zero	147 (37.6)	32 (44.4)	

Supplementary Table 4. Exploratory univariable logistic regression for pCR in the train cohort.
Results are reported after encoding. Significant values are reported in bold (at a significance level of 0.05).

a		b	
Feature	VIF	Feature	VIF
Age	2.94	Age	1.05
ER	1.36	ER	1.36
PR	1.33	PR	1.30
Ki67	1.34	Ki67	1.33
cT	1.07	cT	1.07
cN	1.04	cN	1.04
Menopausal Status	2.31	G	1.23
Histology	1.02		
G	1.24		
HER2	1.03		

Supplementary Table 5. Variance inflation factor (VIF).

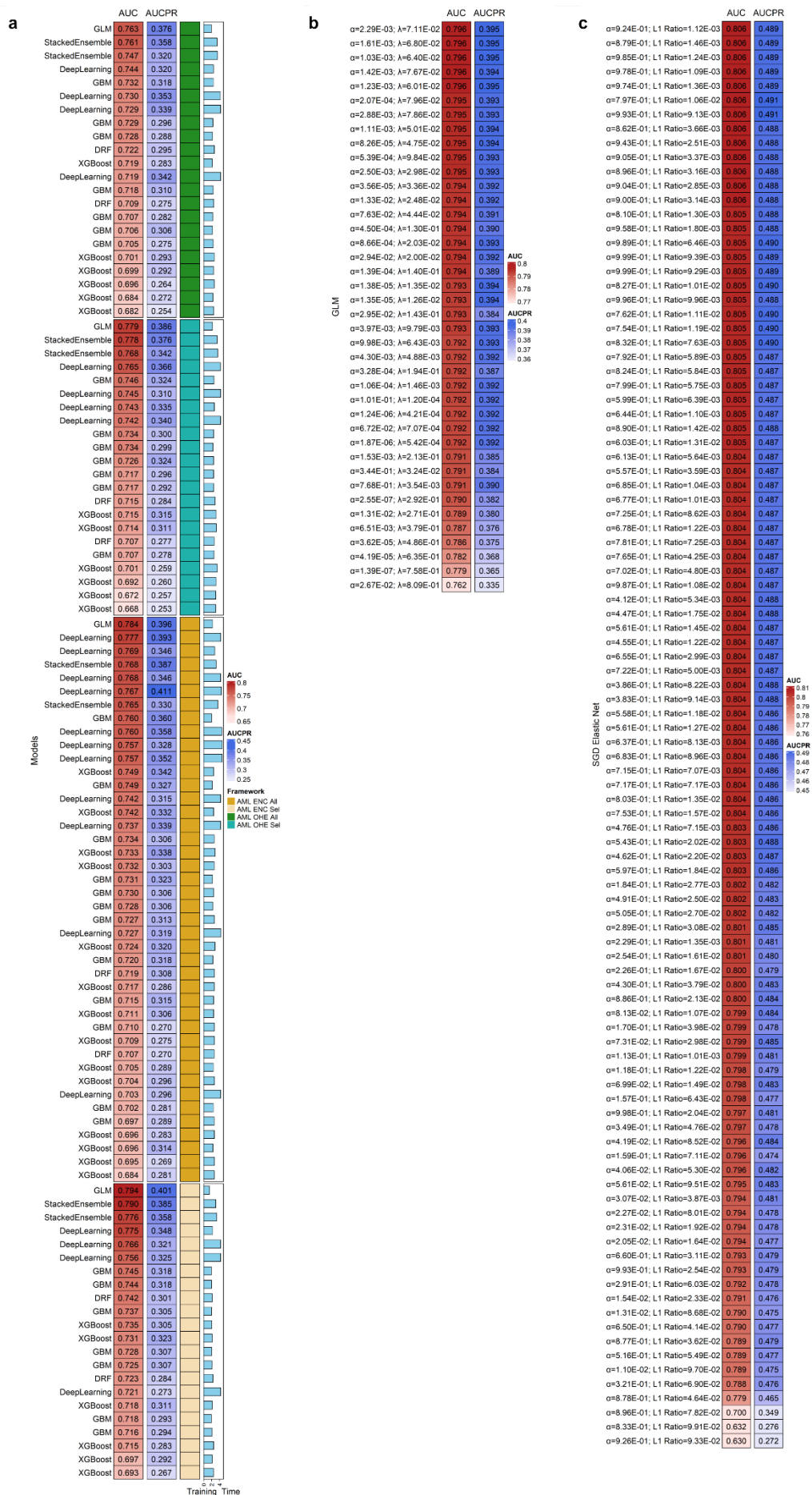
- a. VIF including all features. The highest multicollinearity was observed for age and menopausal status.
- b. VIF after feature selection. A reduction in the VIF was observed, with all features having a VIF lower than one.



Supplementary Fig. 1| Feature selection for pCR prediction.
Legend on next page.

Supplementary Fig. 1| Feature selection for pCR prediction.

- a. Spearman correlation coefficient between clinicopathological features in the training cohort. All variables were pre-processed using ordinal encoding. The nearest point algorithm (single method) with Euclidean distance was used for hierarchical clustering.
- b. Spearman correlation coefficient between clinicopathological features and pCR in the training cohort. Variables with $r > 0.10$ or $r < -0.10$ were selected.
- c. TabPFN permutation importance. Permutation importance estimator measures performance decrease when a feature is not available. Weights are given to indicate the importance, with relative standard errors providing a measure of variability. Using TabPFN permutation importance, all features showed a decrease in model performance upon permutation, with different weights. 15 iterations were used.
- d. Feature ranking and selection using Boruta algorithm. Boruta is a feature selection method that identifies and ranks the most relevant features in a dataset. It works by comparing the importance of original features with that of random shadow features using a random forest algorithm, iteratively removing less important ones to retain only the most significant. 100 iterations were run and features with rank 1 were selected. Applying the Boruta algorithm age, ER, PR, Ki67, cT and cN were selected as rank one features, while G – whereas formally excluded – was the only rank two feature.
- e. Recursive feature elimination with cross-validation. Three models were evaluated: RFE for logistic regression with score accuracy (left), logistic regression with score weighted-F1 (center) and a random forest with score weighted-f1 (right). The three recursive feature elimination (RFE) algorithms implemented selected respectively 4, 10 and 8 features. As expected, there was no increase in accuracy from 4 to 7 seven features, due to imbalance in the outcome frequency.
- f. LASSO regression. 10-fold stratified cross-validation were used to selected the best alpha. Features with non-zero coefficients were selected. The best regularization parameter alpha was identified using cross-validation and non-zero features were Ki67, grade, age, PR, cN, ER and cT.
- g. Upset plot for different feature selection methods. Connected points represent a variable selected by the corresponding model. The bar plot (right) indicates how many times a feature was selected. In the end, seven features were selected for pCR: cN, cT, Ki67, PR, ER, age and grading.

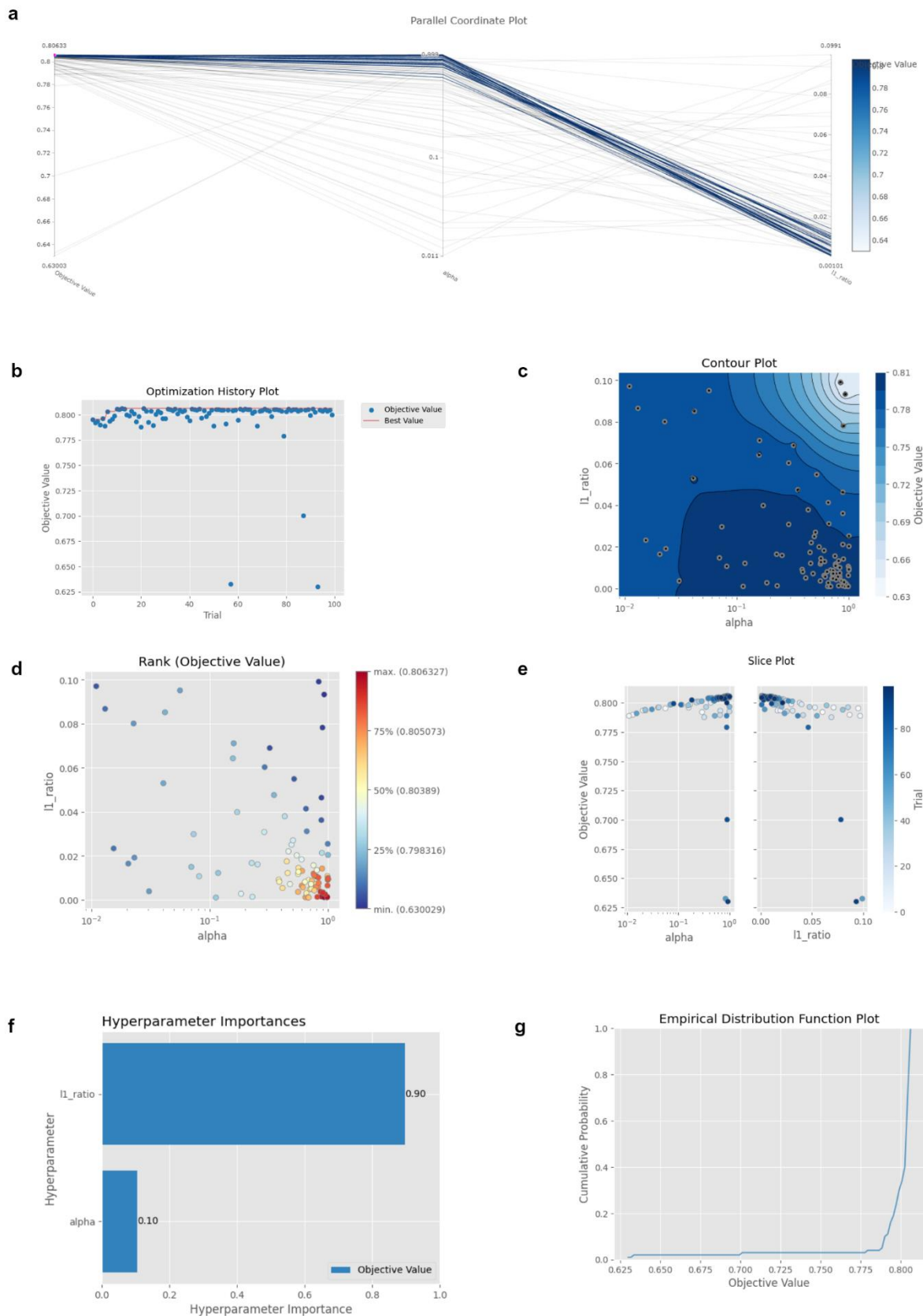


Supplementary Fig. 2| Summary of model performance comparison.
Legend on next page.

Supplementary Fig. 2| Summary of model performance comparison.

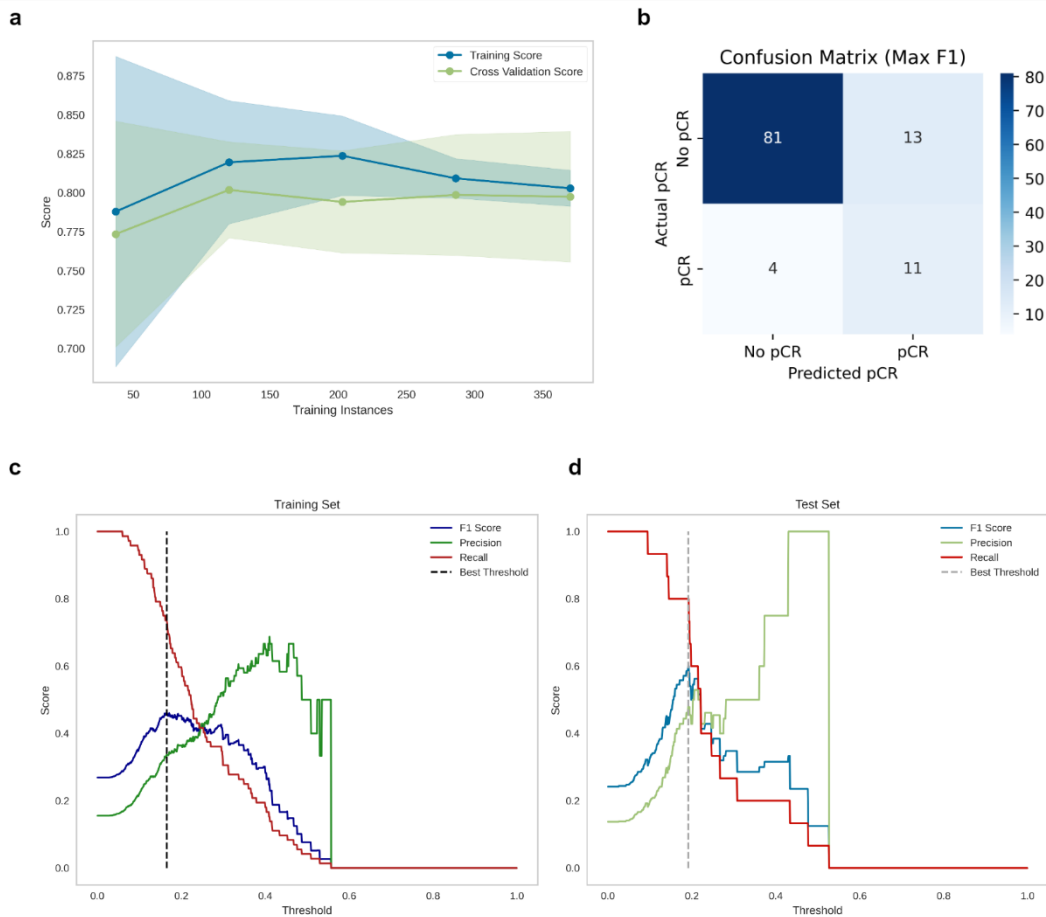
All models were compared using 5-fold cross-validation, storing the AUROC and the AUC-PR for each set of hyperparameters tested.

- a. H2O AutoML algorithms. From upper to lower: OHE with all features, OHE with the 7 selected features, ordinal encoding with all features and ordinal encoding with the 7 selected features. The training time is expressed as base 10 logarithm of the total training time. The models trained by H2O include: XGBoost, Generalized Linear Models (GLM), Random Forest (RF), Extremely Randomized Forest (XRT), Gradient Boosting Machine (GBM) and Deep Neural Networks (DNN).
- b. H2O GLM algorithms with respective alphas and lambdas.
- c. SGD Elastic Net with Platt's calibration. Since the calibration step was included in the pipeline for the optimization process, the train set was further splitted in a test dataset and a calibration dataset. Therefore, this model was indeed trained on less patients compared to the others model, without a loss in performance. The best-performing model achieved a cross-validated AUROC of 0.81.



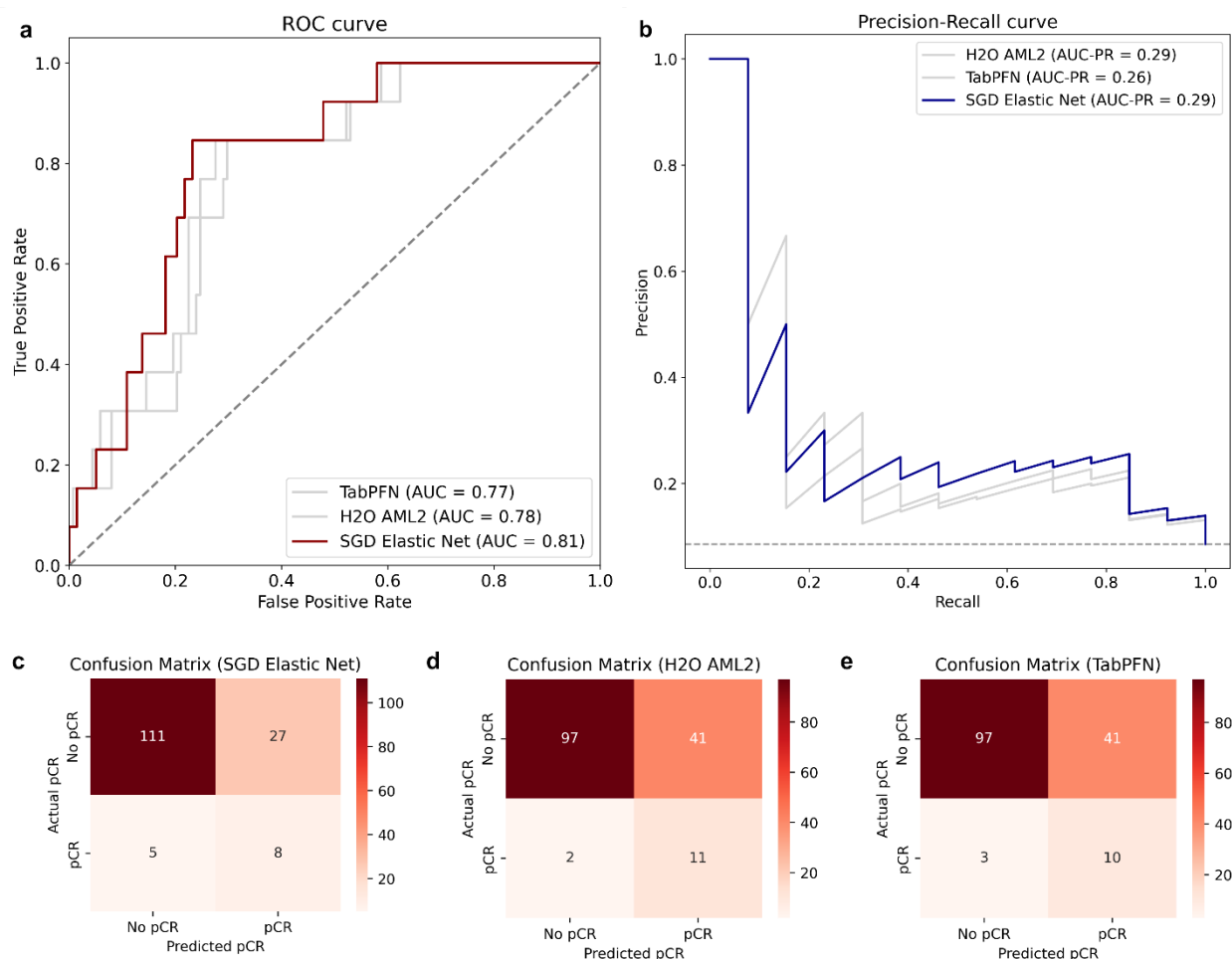
Supplementary Fig. 3| Hyperparameters optimization for the best GLM model trained with SGD (SGD Elastic Net).

- Parallel coordinates plot. The plot shows the relationship between hyperparameters tuned and their respective outcomes.
- Optimization history plot. The plot shows the progression of optimization over iterations, highlighting convergence at an AUROC of 0.81. Converge of the estimate provide reliability for the performance.
- Contour plot. The optimal regions for alpha and L1 ratio are highlighted in dark blue.
- Rank plot. The plot displays the ranking of hyperparameter sets based on their performance.
- Slice plot. The plot provides a view of model performance with variations in one hyperparameter, holding others constant.
- Hyperparameter importance. The barplot indicates the significance of each hyperparameter in influencing model performance. The L1 ratio was found to be the most important hyperparameter.
- Empirical distribution function plot. The plot shows the cumulative distribution of model performance metrics across different hyperparameter settings.



Supplementary Fig. 4| Performance of the best GLM model trained with SGD (SGD Elastic Net).

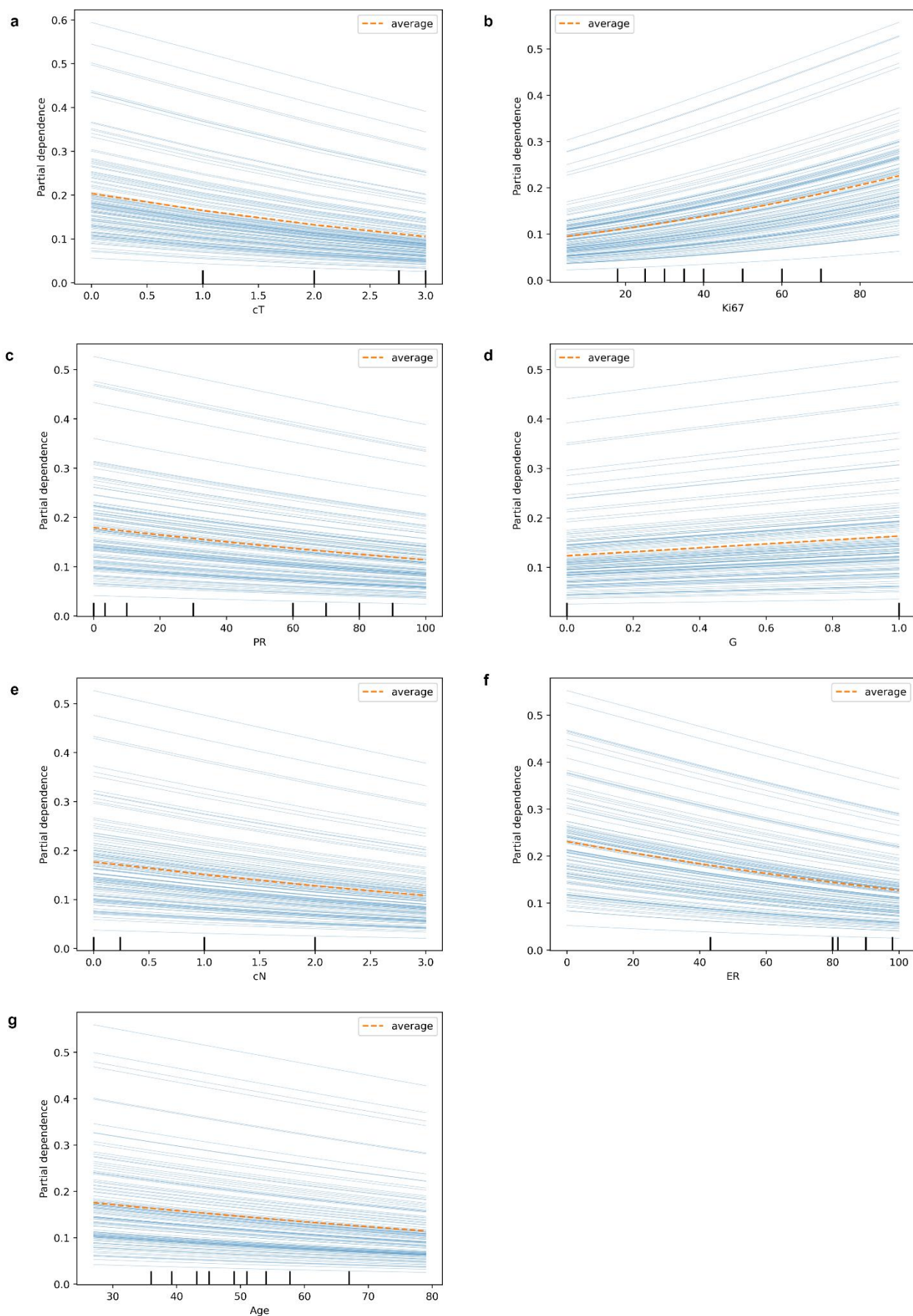
- Model learning curve. Training and calibration curves are reported. While the training error seems to continue decreasing, the cross-validation error reached a plateau, confirming the number of instances on which the model was trained was enough.
- Confusion Matrix. The cut-off of 0.1925 was used. Using this cut-off, precision was 0.46, recall (sensitivity) was 0.80 and specificity 0.86
- Discrimination threshold plot in the training set. The dashed black line corresponds to the cut-off of 0.1925.
- Discrimination threshold plot in the internal validation set. The dashed gray line corresponds to the cut-off of 0.19. The best cut-off was evaluated the train set and in the internal validation set to maximize F1 score, with similar results (cut-offs were always between 0.17 and 0.19). The best F1 score was 0.60 in the internal validation dataset. We considered 0.19 more appropriate and directly applicable in clinical practice and we used it for further evaluation in the external independent set. Using this cut-off, the rate of pCR in patients with positive prediction in the train set is 36% and in the internal validation set is 46%.



Supplementary Fig. 5| Performance of the best models in the external validation dataset.

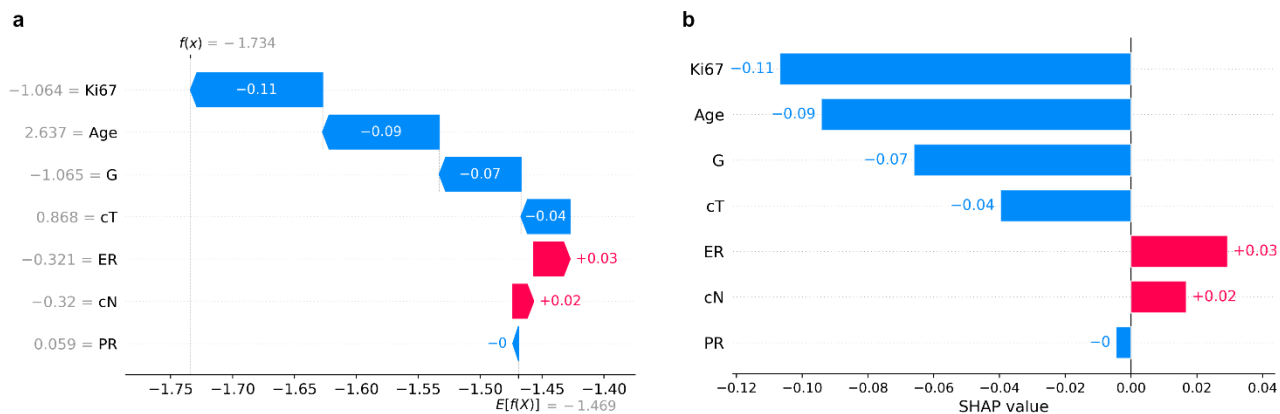
The external validation dataset was used only to validate the performance of the best model (SGD Elastic Net), which was selection on the internal validation set. Therefore, the following comparisons on the external validation set should be considerate exploratory. The three best models are reported: H2O AML2, SGD Elastic Net and TabPFN.

- Comparison of ROC curves. Elastic Net trained with SGD solver and calibrated using Platt's logistic model was the best model. The performance of the other models is presented in gray solid lines. The performance of a random classifier is represented in a gray dashed line.
- Comparison of Precision-Recall Curves. Elastic Net trained with SGD solver and calibrated using Platt's logistic model was the best model. The baseline performance is represented in a gray dashed line.
- Confusion Matrix for SGD Elastic Net. The cut-off of 0.1925 was used. Using this cut-off, precision was 0.23, recall (sensitivity) was 0.62 and specificity 0.80.
- Confusion Matrix for H2O AML2.
- Confusion Matrix for TabPFN.



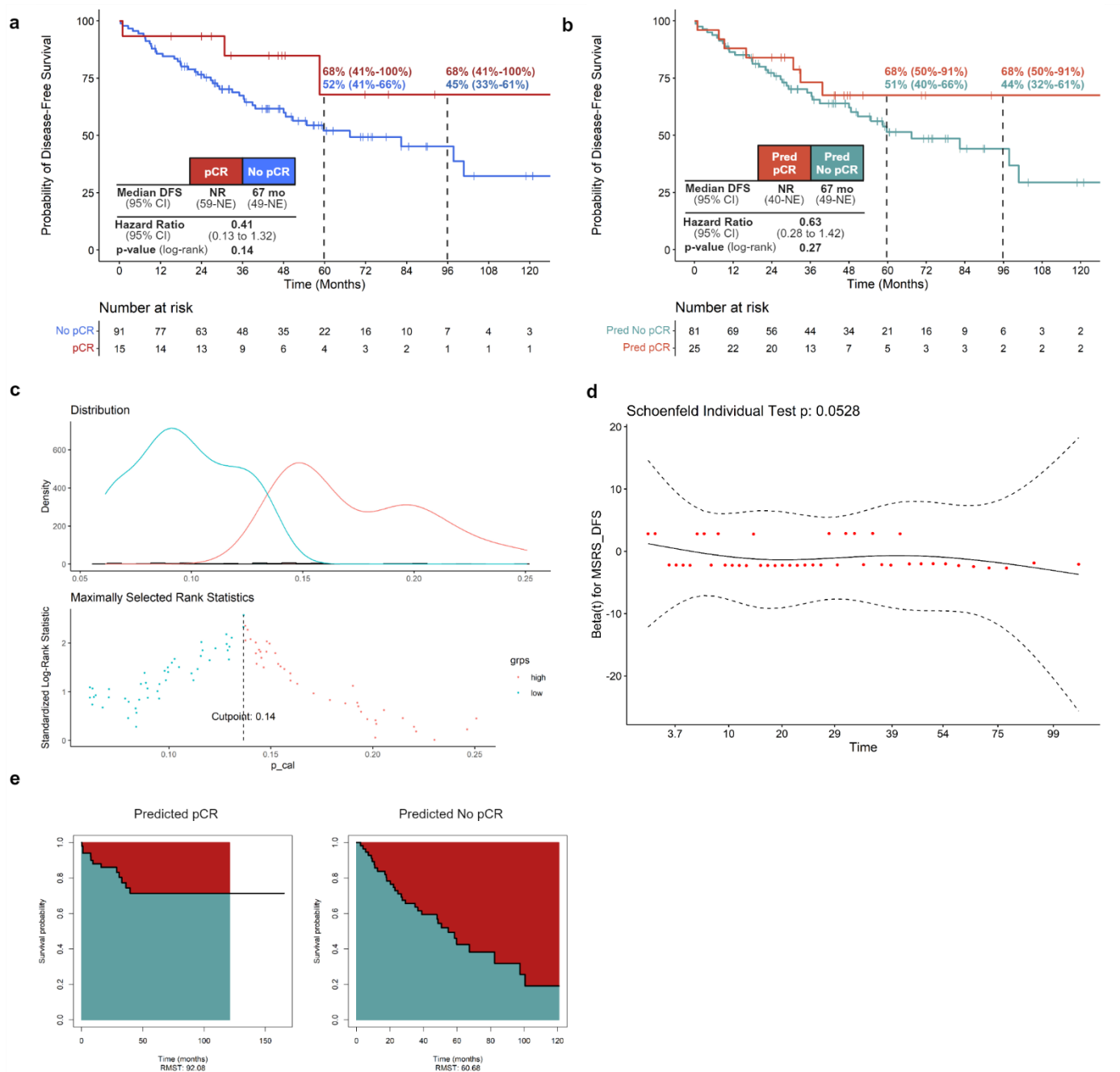
Supplementary Fig. 6| Partial dependence plots comparison for the SGD Elastic Net model.

The framework using ordinal encoding and SGD Elastic Net including the 7 selected features was considered. Ordinal encoding for categorical features was applied in alphabetical order with the following mapping. a-g. PDP for different variables in order of their SHAP value. From left to right and from upper to lower: cT, Ki67, PR, G, cN, ER and Age.



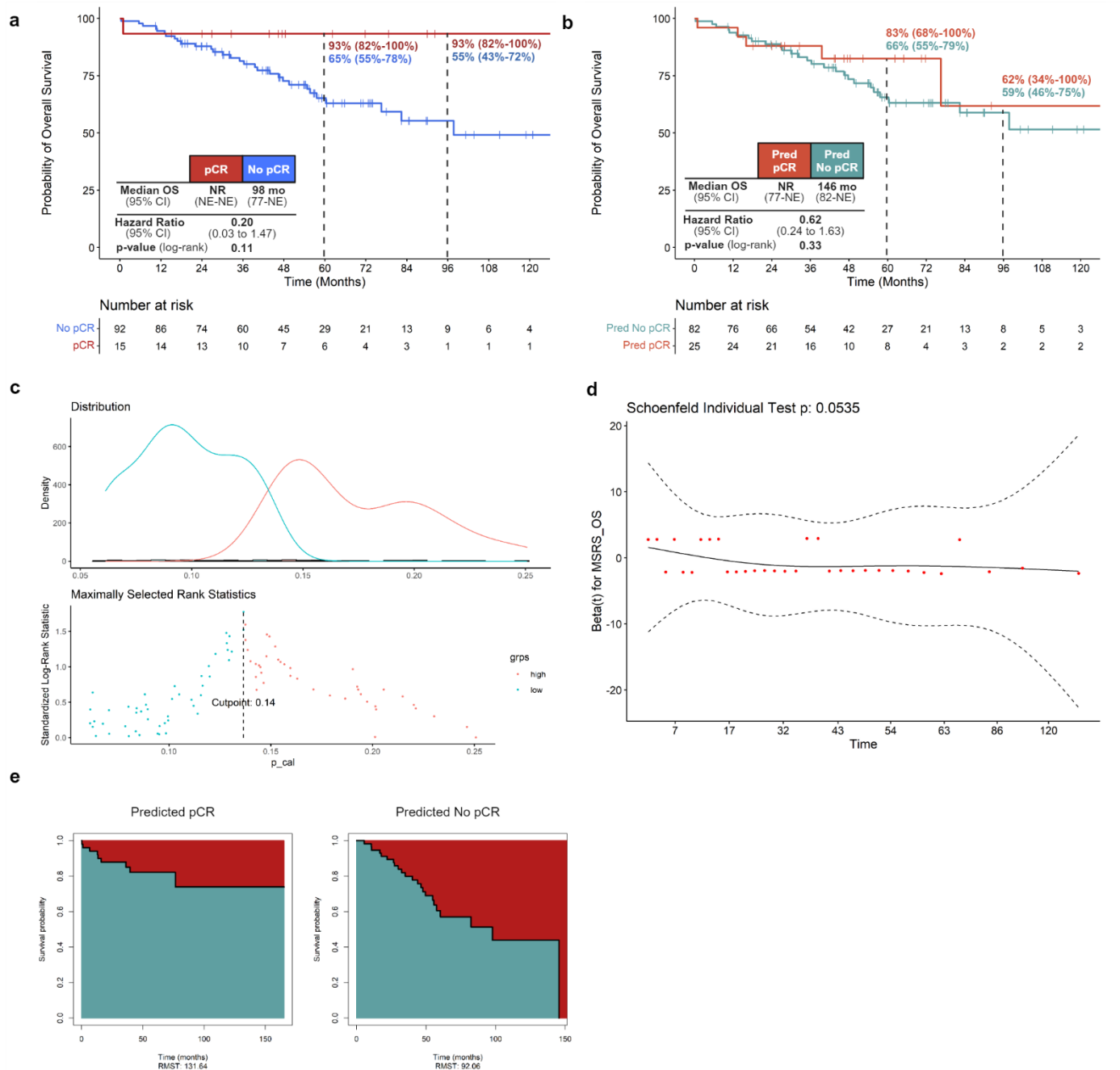
Supplementary Fig. 7| Patient level predictions of the SGD Elastic Net model.

- Waterfall plot of SHAP predictions for patient #1 in the internal validation dataset. Patient's characteristics were: Premenopausal, Age 77, cT3, cN1, Histology Ductal, G1 or 2, ER 70, PR 50, Ki67 18. The final prediction of the model was pCR No with $f(x) = -1.743$, which was actually correct.
- SHAP barplot for patient #1 in the internal validation dataset. For each feature, the corresponding SHAP value is reported.



Supplementary Fig. 8| Association between pCR and disease-free survival in the internal validation cohort.

- Kaplan-Meier curves of DFS according to pCR in the internal validation cohort (log-rank test: $p=0.14$).
- Kaplan-Meier curves of DFS according to pCR prediction in the internal validation cohort. Predictions were based on the maximum threshold for F1 score (log-rank test: $p=0.27$).
- Maximally selected rank statistics (MSRS) analysis. MSRS was conducted to identify the best pCR probability to maximize log-rank DFS differences between predicted pCR and predicted no-pCR patients in the internal validation cohort. The cut-off was 0.14.
- Schoenfeld residuals and test for proportional hazards assumption. The test is referred to pCR prediction based on MSRS cut-off and DFS. A violation of proportional hazard was graphically suspected and confirmed with the test.
- Exploratory analysis restricted mean survival (RMST). A significant difference in mean DFS between predicted pCR and predicted no-pCR patients was reported.



Supplementary Fig. 9| Association between pCR and overall survival in the internal validation cohort.

- Kaplan-Meier curves of OS according to pCR in the internal validation cohort (log-rank test: $p=0.11$).
- Kaplan-Meier curves of OS according to pCR prediction in the internal validation cohort. Predictions were based on the maximum threshold for F1 scores (log-rank test: $p=0.27$).
- Maximally selected rank statistics (MSRS) analysis. MSRS was conducted to identify the best pCR probability to maximize log-rank OS differences between predicted pCR and predicted no-pCR patients in the internal validation cohort. The cut-off was 0.14.
- Schoenfeld residuals and test for proportional hazards assumption. The test is referred to pCR prediction based on MSRS cut-off and OS. A violation of proportional hazard was graphically suspected and confirmed with the test.
- Exploratory analysis restricted mean survival (RMST). A significant difference in mean OS between predicted pCR and predicted no-pCR patients was reported.

Section/Topic	Item	Development / evaluation ¹	Checklist item	Reported on page
TITLE				
<i>Title</i>	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	1
ABSTRACT				
<i>Abstract</i>	2	D;E	See TRIPOD+AI for Abstracts checklist	2
INTRODUCTION				
<i>Background</i>	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	3
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	3
	3c	D;E	Describe any known health inequalities between sociodemographic groups	NA
<i>Objectives</i>	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	3
METHODS				
<i>Data</i>	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	14
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	14
<i>Participants</i>	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	14
	6b	D;E	Describe the eligibility criteria for study participants	14
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	14
<i>Data preparation</i>	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	16
<i>Outcome</i>	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	15
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	15
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	NA
<i>Predictors</i>	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	16-17
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	14-16
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	14-16
<i>Sample size</i>	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	15
<i>Missing data</i>	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	16
<i>Analytical methods</i>	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	16-16
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	16
	12c	D	Specify the type of model, rationale ² , all model-building steps, including any hyperparameter tuning, and method for internal validation	16-18
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations ³	NA
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	16-18
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	NA
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	16-18
<i>Class imbalance</i>	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	18
<i>Fairness</i>	14	D;E	Describe any approaches that were used to address model fairness and their rationale	NA
<i>Model output</i>	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	17-18

¹ D=items relevant only to the development of a prediction model; E=items relating solely to the evaluation of a prediction model; D;E=items applicable to both the development and evaluation of a prediction model

² Separately for all model building approaches.

³ TRIPOD-Cluster is a checklist of reporting recommendations for studies developing or validating models that explicitly account for clustering or explore heterogeneity in model performance (eg, at different hospitals or centres). Debray et al, BMJ 2023; 380: e071018 [DOI: 10.1136/bmj-2022-071018]

<i>Training versus evaluation</i>	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	4, 14
<i>Ethical approval</i>	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	14
OPEN SCIENCE				
<i>Funding</i>	18a	D;E	Give the source of funding and the role of the funders for the present study	23
<i>Conflicts of interest</i>	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	23
<i>Protocol</i>	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	14
<i>Registration</i>	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	14
<i>Data sharing</i>	18e	D;E	Provide details of the availability of the study data	19
<i>Code sharing</i>	18f	D;E	Provide details of the availability of the analytical code ⁴	19
PATIENT & PUBLIC INVOLVEMENT				
<i>Patient & Public Involvement</i>	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	NA
RESULTS				
<i>Participants</i>	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	4, Fig.1
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	4, Suppl Table 1-3
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	4-5, Suppl Table 1-3
<i>Model development</i>	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	4
<i>Model specification</i>	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary) ⁵	6-7
<i>Model performance</i>	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	6-7, Figure 3
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details ³ .	NA
<i>Model updating</i>	24	E	Report the results from any model updating, including the updated model and subsequent performance	NA
DISCUSSION				
<i>Interpretation</i>	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	9-13
<i>Limitations</i>	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	12-13
<i>Usability of the model in the context of current care</i>	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	NA
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	11
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	12

From: Collins GS, Moons KGM, Dhiman P, et al. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378

Supplementary Table 6. TRIPOD-AI Checklist.

Suppl. indicates the information is available in the supplementary material. NA = Not applicable. Patients' characteristics and distributions of variables are available in tabular form in the supplementary material. Individual model predictions can be accessed through the WebApp.