Grant Agreement no. 777167

BOUNCE

Predicting Effective Adaptation to Breast Cancer to Help Women to BOUNCE Back

Research and Innovation Action
SC1-PM-17-2017: *Personalised computer models and in-silico systems for well-being*

Deliverable: D4.4b Accuracy Evaluation, Finalization and Quality Assurance of the BOUNCE In Silico Resilience Trajectory Predictor (Stage 2)

Due date of deliverable: (31-03-2022)
Actual submission date: (28-04-2022))

Start date of Project: 01 November 2017 Duration: 54 months (extended duration)

Responsible WP: ICCS

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777167

Dissemination level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Service	
RE	Restricted to a group specified by the consortium (including the Commission Services)	x
CO	Confidential, only for members of the consortium (excluding the Commission Services)	

0. Document Info

0.1. Author

Author	Company	E-mail
Georgios Stamatakos	ICCS	gestam@central.ntua.gr
Eleni Kolokotroni	ICCS	ekolok@mail.ntua.gr
Panagiotis Simos	FORTH	akis.simos@gmail.com
Georgios Manikis	FORTH	gmanikis@gmail.com
Vangelis Karademas	FORTH	karademas@uoc.gr
Konstantina Kourou	FORTH	konstadina.kourou@gmail.com
Eugenia Mylona	FORTH	mylona.eugenia@gmail.com
Mikko Nuutinen	NHG	mikko.nuutinen@nhg.fi

0.2. Documents history

Document version #	Date	Change
V0.1	06-01-2022	Starting version, template
V0.2	28-01-2022	Definition of ToC
V0.3	08-04-2022	First complete draft
V0.4	12-04-2022	Integrated version (send to WP members)
V0.5	17-04-2022	Updated version (send PCP)
V0.6	21-04-2022	Updated version (send to project internal reviewers)
Sign off	28-04-2022	Signed off version (for approval to PMT members)
V1.0	28-04-2022	Approved Version to be submitted to EU

0.3. Document data

Keywords	breast cancer, resilience, predictive model, accuracy, evaluation, quality assurance
Editor Address data	Name: Georgios S. Stamatakos Partner: ICCS Address: Iroon Polytechniou 9, Zografos, ICCS, National Technical University of Athens Phone: +30 210 772 2287 Fax: - E-mail: gestam@central.ntua.gr
Delivery date	28/ 04/2022

Table of Contents

NOTE: The character F before a number refers to work done by FORTH whereas the character I before a number refers to work done by ICCS.

2. Introduction

This document presents the accuracy evaluation of various models related to the resilience trajectory predictor that have been developed by both modelling partners of BOUNCE i.e. FORTH (Foundation for Research and Technology Hellas) and ICCS (Institute of Communication and Computer Systems). The in silico resilience trajectory predictors have been developed by exploiting the data generated by the BOUNCE prospective pilot study using artificial intelligence and statistical methods. More precisely, the document focuses on supervised models predicting Month12 (M12) mental health and quality of life or QoL (FORTH), the generalizability of risk prediction models across clinical sites (FORTH), classifiers for detecting depression and QoL 18 month trajectories (ICCS) including a classifier for detecting patients with poor and good depression trajectory (ICCS), a classifier detecting patients with low decreasing QoL trajectory (ICCS) and a classifier for detecting patients with high increasing QoL trajectory (ICCS). An outline of the finalized versions of the model implementation within the BOUNCE decision support tool / platform is also provided. A summary of quality assurance considerations with cross-references to other relevant deliverables complements the current document.

3. Accuracy Evaluation of the BOUNCE In Silico Resilience Trajectory Predictor

3.1 FORTH MODELS

3.1.1 Supervised Models predicting M12 Mental Health and QoL: Generalizability of risk prediction models across clinical sites (FORTH)

In this version of deliverable D4.4 we extended the assessment of the generalizability of the supervised learning algorithms and associated pipeline across the four clinical sites. Specifically, we focused on the models that address the most challenging clinical question, namely identification of patients who initially report minimal mental health symptoms and/or good QoL and later experience a significant decline in psychological well-being (Models F2 and F4 in D4.3b). In this framework, for each one of the two models (one for mental health and one for QoL) we considered each cohort separately as the testing/validation set of the training model. The training phase was implemented each time on three out of four datasets keeping the last cohort for testing purposes regarding the different combinations of the four available cohorts. This approach therefore validates the proposed ML-based pipeline when new unseen patient data are considered for predicting M12 outcomes given the patient characteristics at baseline (M0) and 3-months post diagnosis.

3.1.1.1 Model Description

The Random Forest (RF) estimator was applied on a total of 532 (Model F2) and 528 patients with sufficient M12 data (Models F4). Model generalizability across study sites is demonstrated here for the prediction of mental health outcomes at M12 (stable good vs deteriorating mental health groups including HADS and EORTC global QoL ratings at M0 and M3). The number of features to retain for model testing was limited to 15 given the smaller size of the test set.

Model performance metrics

To evaluate the performance of the classification models based on the ensemble-based methodology, 5 measures were calculated: balanced accuracy, F1 score, sensitivity, specificity, and AUC. Additionally, the concordance of the external validation schemes on the most important predictors was assessed.

3.1.1.2. Results

3.1.1.2.1 Model F2 (Mental Health)

Model F2 Performance

As shown in Table F1, model performance was fairly constant across sites with mean accuracy at 78.8%, sensitivity at 80.5%, specificity at 77.8% and AUC=.79. These results are promising, especially in light of the small size of the deteriorated groups included in the test data set each time (ranging from 8 to 17 participants).

Table F1. Model F2 performance predicting M12 mental health deterioration in cross validation tests.

Training data set	Test data set	Stable good Group (n)	Deterio-rated Group (n)	Accu-racy	Sensi-tivity	Speci-ficity	F1	AUC
IEO, HUS, HUJI	Champ	89	12	84	85	84	61	84
Champ, HUS, HUJI	IEO	53	17	73	88	66	70	77
Champ, IEO, HUJI	HUS	162	8	81	89	81	34	85
Champ, IEO, HUS	HUJI	71	10	77	60	80	43	70

Concordance of predictors across sites

Table F2 lists the predictors of M12 overall Mental Health (Model F2) across the four cross-validation schemes (each tested with data from a single site that was not used in the training phase). All predictor variables that emerged as important features for classification across the four sites were among the highest-ranking features identified by Model F2 in the total sample. Of those, four variables were common to the four cross-validation models (i.e., featured among the 15 highest ranking features in models tested on data from each of the four clinical

sites; according to impurity-based feature importances -RF-based meta-estimator). As expected, variables which consistently featured among the highest-ranking predictors of deterioration in M12 Mental Health across sites were those that contributed most to model output (illustrated in the SHAP Figures F1-F4[1]). Inspection of Figures F1-F4 reveals that high levels of mental health symptoms, negative affect, worrying thoughts and feelings of helplessness at M3, combined with low levels of potential protective factors (such as optimism, social and emotional support, mindfulness, and sense of illness manageability) consistently predicted mental health decline 9-12 months later.

Five additional variables featured among the highest-ranking features in the tests performed on data from only 2 clinical sites (high levels of negative affectivity at M0, poor body image at M3, and low levels of positive emotional regulation strategies, mindfulness, and emotional support predicting poor mental health at M12).

These figures reveal notable inconsistencies across sites on two biological markers that ranked highly in Model F2 conducted on the total sample (NLR, platelet count). Thus, the direction of each variable's impact on model performance varied across sites (e.g., positive for NLR in the model predicting mental health for IEO patients and negative in the model predicting mental health deterioration for HUS patients).

[1] An expanded set of the 30-highest ranking variables are included in the SHAP figures in order to allow comparison between high- and low-ranking features as well as between clinical sites on lower-ranking features. As a result of including additional variables in the overall classification model some variables may not be included in the figures (i.e., NLR in Fig. F1, Optimism in Fig. F4).

Table F2. Predictors of M12 Mental Health across cross-validation schemes.

	Champ	IEO	HUS	HUJI
HADS Anxiety (M3)	√	√	√	√
HADS Depression (M3)	√	√	√	√
Negative Affectivity (PANAS, M3)	√	√	√	√
Anxious Preoccupation (MAC, M3)	√	√	√	√
HADS Anxiety (M0)	√	√	√	
Social Support (M3)	√	√	√	
Optimism (LOT, M0)	√		√	√
Manageability (SOC, M0)	√		√	√
Helplessness (MAC, M3)	√	√		√
NLR (M0)	√		√	√
Negative Affectivity (PANAS, M0)		√		√
Positive Emotion Regulation (CERQ, M0)			√	√
Mindfulness (MAAS, M0)		√	√	
Emotional Support (M3)		√	√	
Body Image (BR-23, M3))		√	√	
Thrombocyte count (M0)	√			√
HADS Depression (M0)				√
Coping with Cancer (CBI, M0)				√
Global QoL (M3)				√
Meaningfulness (SOC, M0)		√		
Trait Resilience (CDRISC, M0)	√			
Avoidance (MAC, M3)	√			
Community Cohesion (FARE, M3)		√		
PTGI (M3)	√			
Side Effects (BR23, M3)		√		
Arm Symptoms (BR23, M0)				

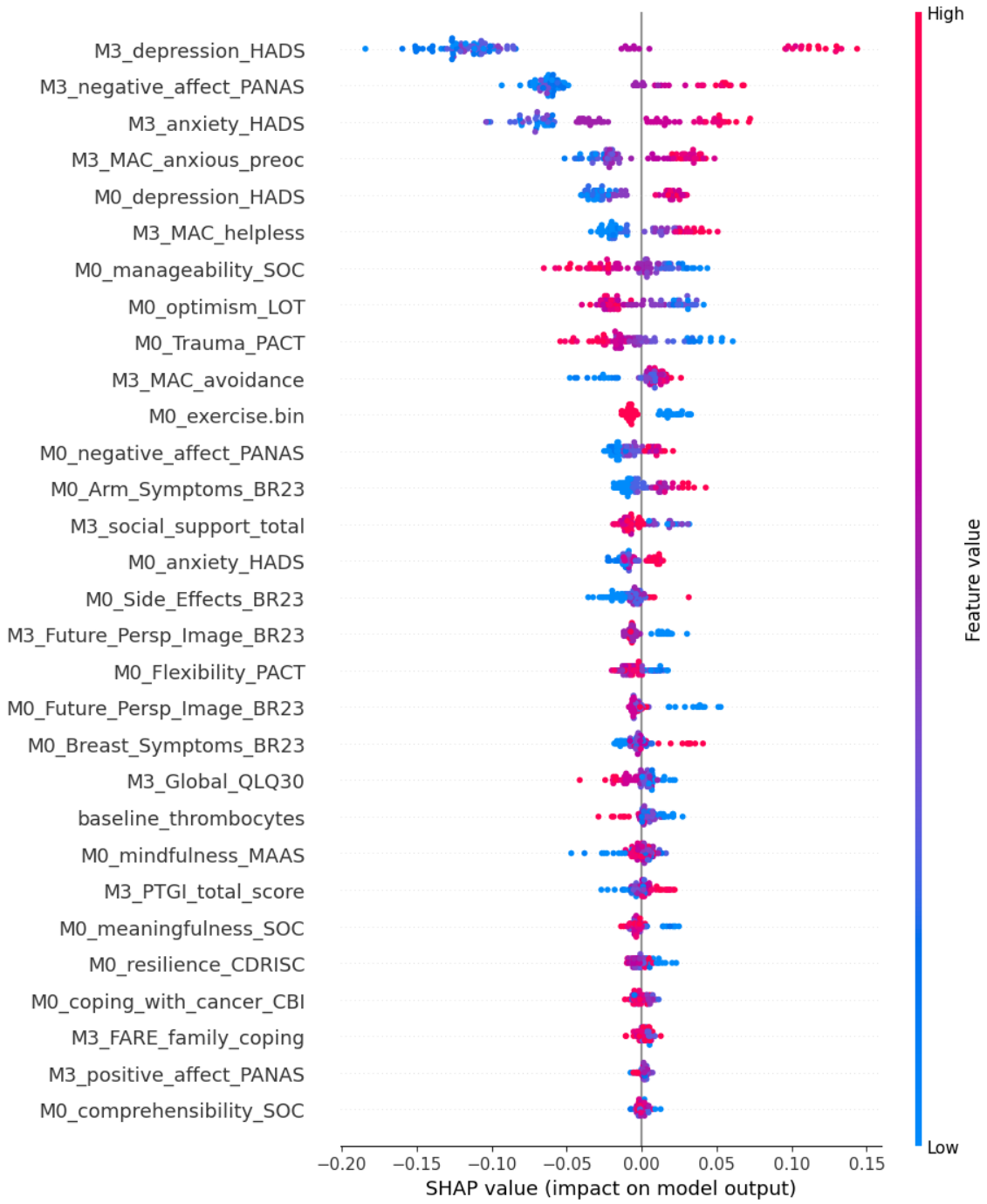

Figure F1. Shapley additive explanations (SHAP) summary plot generated using data from CHAMP in the testing phase of Model F2 (predicting overall Mental Health at M12 from M0 and M3 variables including QoL and mental health indices).

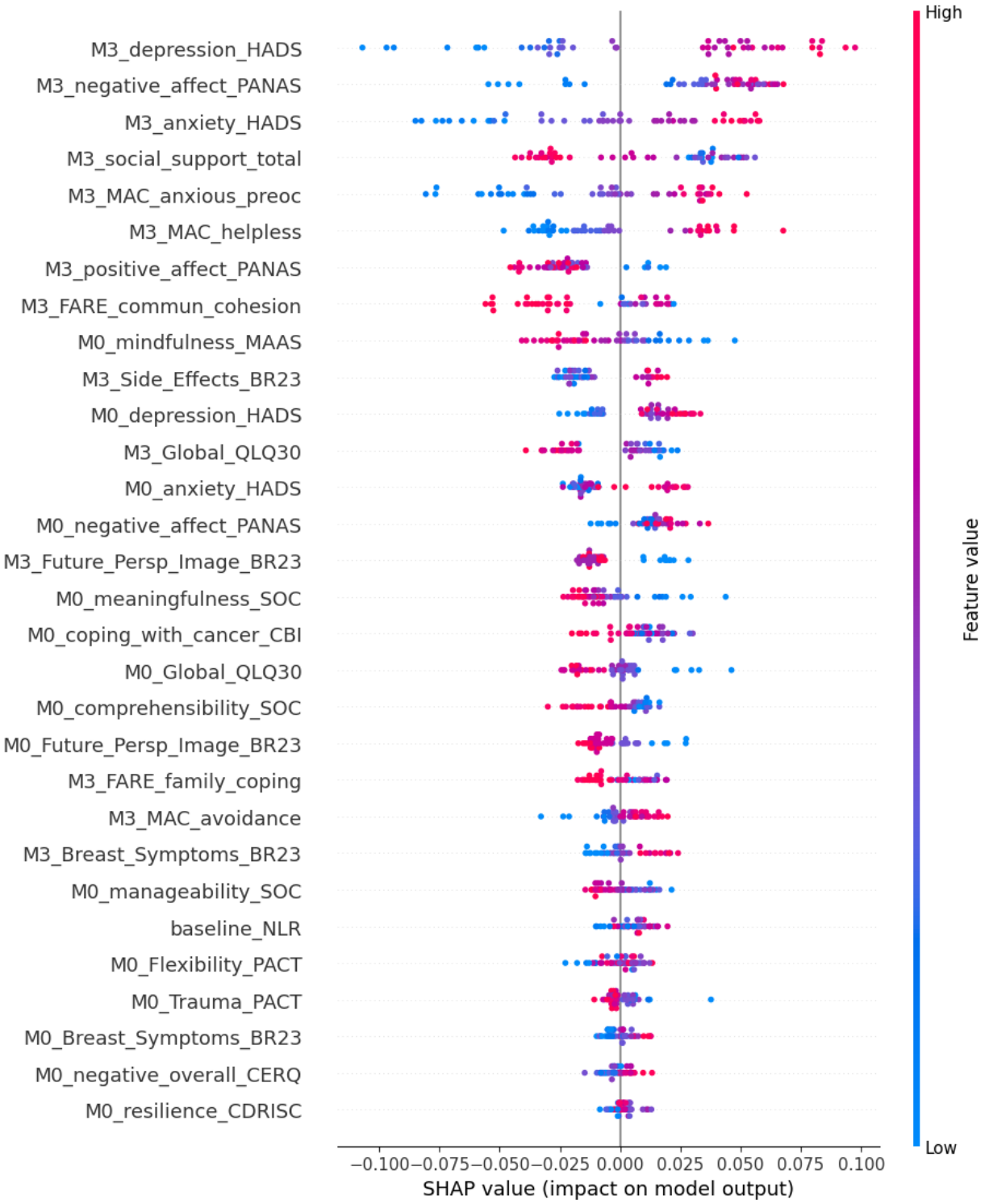

Figure F2. Shapley additive explanations (SHAP) summary plot generated using data from IEO in the testing phase of Model F2 (predicting overall Mental Health at M12 from M0 and M3 variables including QoL and mental health indices).

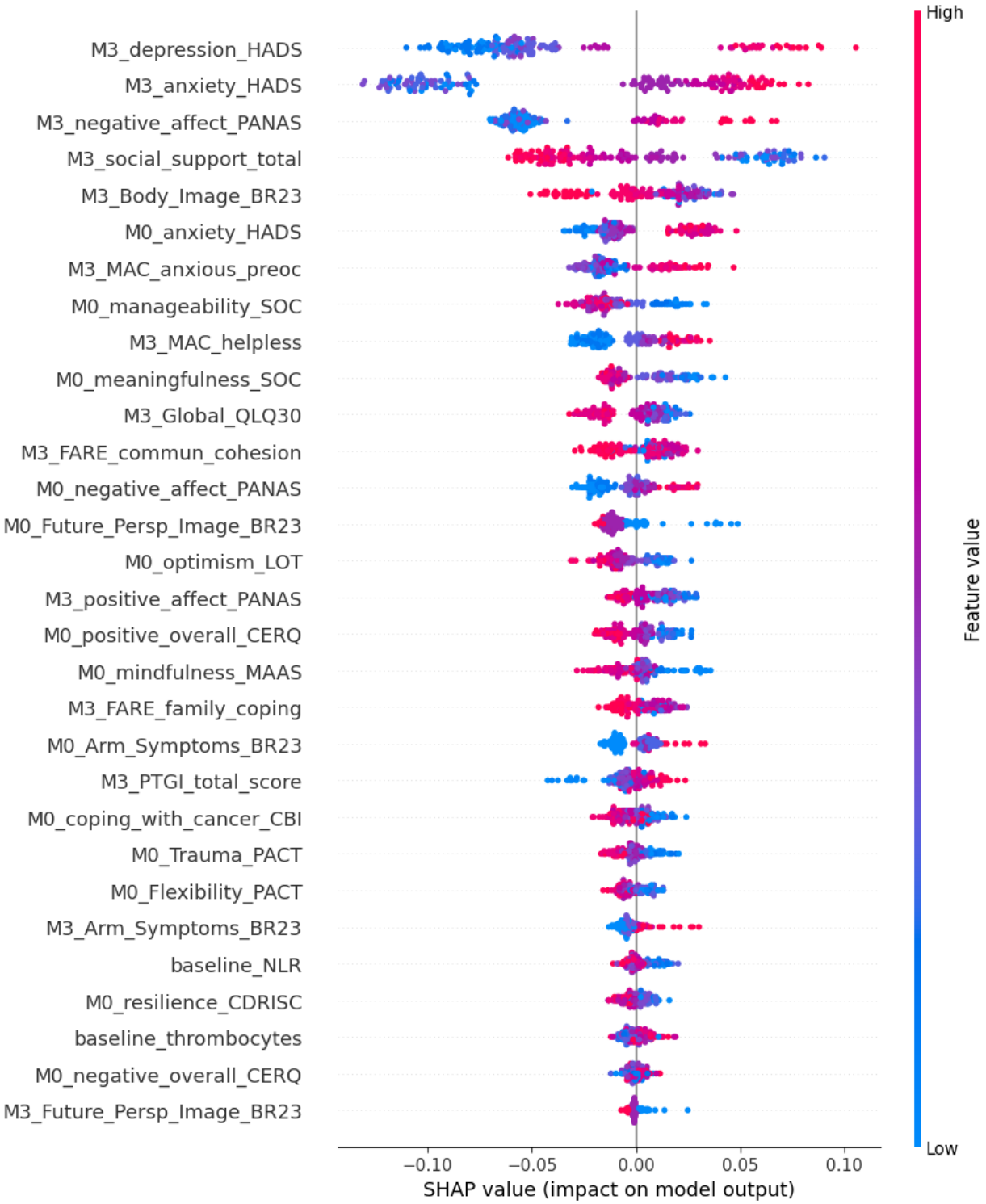

Figure F3. Shapley additive explanations (SHAP) summary plot generated using data from HUS in the testing phase of Model F2 (predicting overall Mental Health at M12 from M0 and M3 variables including QoL and mental health indices).

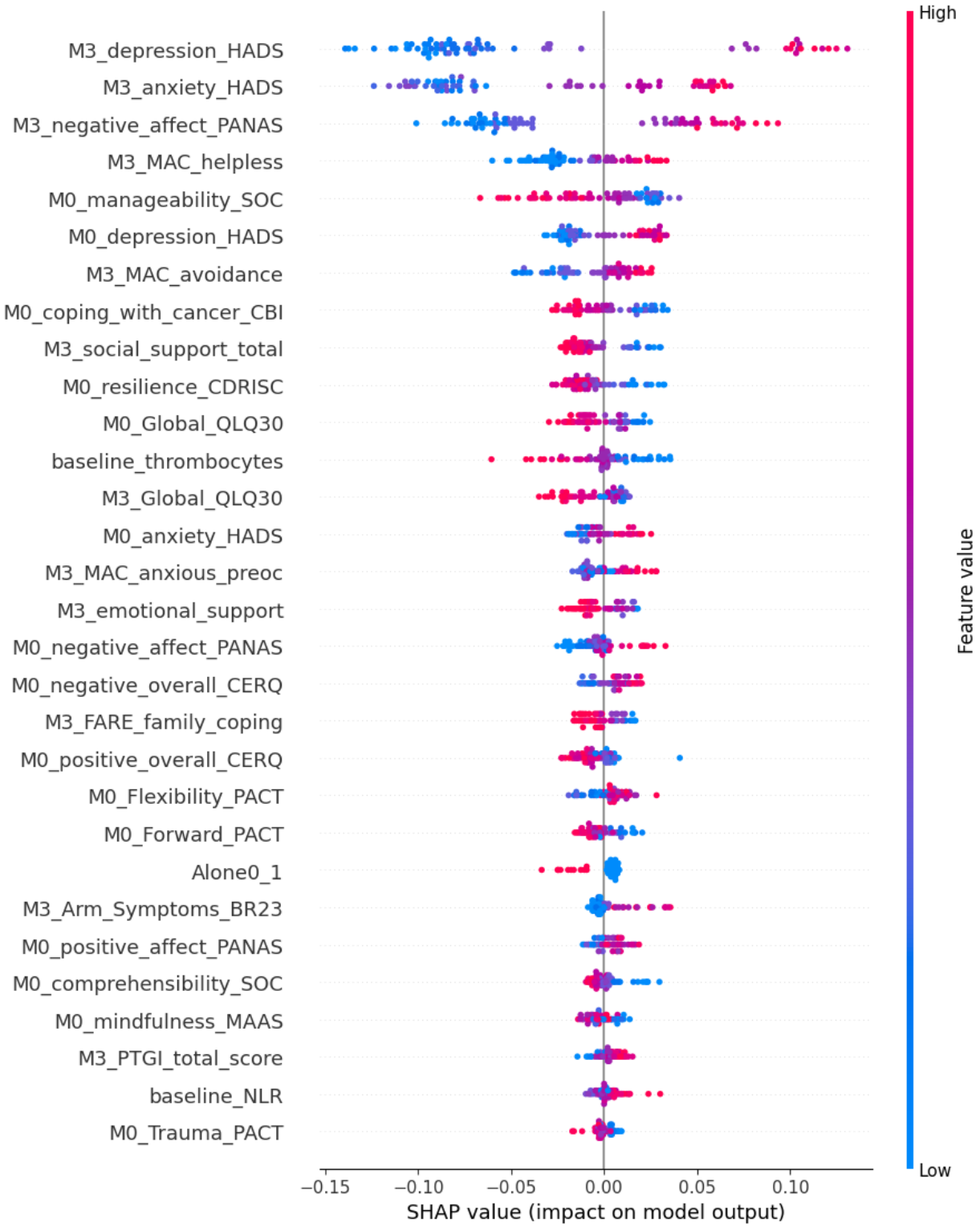

Figure F4. Shapley additive explanations (SHAP) summary plot generated using data from HUJI in the testing phase of Model F2 (predicting overall Mental Health at M12 from M0 and M3 variables including QoL and mental health indices).

3.1.1.2.2 Model F4 (Global QoL)

Model F4 Performance

As shown in Table F3, model performance was somewhat less consistent across sites, as compared to Model F2, with mean accuracy at 72.3%, sensitivity at 72.3%, specificity at 72.3% and AUC=.73. Overall accuracy in predicting decline in global QoL in the total sample was also considerably lower than accuracy in predicting decline in overall mental health (see D4.3b). Despite these limitations, these results are promising, especially in light of the small size of the deteriorated groups included in the test data set each time (ranging from 9 to 14 participants).

Table F3. Model F4 performance predicting M12 global QoL deterioration in cross validation tests.

Training data set	Test data set	Stable good Group (n)	Deterio-rated Group (n)	Accu-racy	Sensi-tivity	Speci-ficity	F1	AUC
IEO, HUS, HUJI	Champ	70	12	88	75	93	77	84
Champ, HUS, HUJI	IEO	56	9	48	91	37	42	64
Champ, IEO, HUJI	HUS	114	10	72	50	76	31	63
Champ, IEO, HUS	HUJI	60	14	81	75	83	69	79

Concordance of predictors across sites

Table F4 lists the predictors of Global QoL decline at M12 (Model F4) across the four cross-validation schemes (each tested with data from a single site that was not used in the training phase). All predictor variables that emerged as important features for classification across the four sites were among the highest-ranking features identified by Model F4 in the total sample. Of those, five variables were common to the four cross-validation models (i.e., featured among the 15 highest ranking features in models tested on data from each of the four clinical sites). Symptoms of depression and treatment side effects experienced at M3, combined with low ratings of global QoL (M3), illness coping strategies and mindfulness (M0), predicted a decline in global QoL at M12. Five additional variables were among the highest-ranking in 3/4 cross-validation schemes. Thus, decline in global QoL was also consistently predicted by high levels of worrying thoughts, and poor body image at M3, whereas a sense of illness manageability, availability of positive emotion regulation strategies and a sense of personal growth in response to the illness appeared to exert a protective role against such deterioration.

Table F4. Predictors of M12 Global QoL across cross-validation schemes.

	Champ	IEO	HUS	HUJI
Coping with Cancer (CBI, M0)	√	√	√	√
Global QoL (M3)	√	√	√	√
HADS Depression (M3)	√	√	√	√
Side Effects (BR23, M3)	√	√	√	√
Mindfulness (MAAS, M0)	√	√	√	√
Anxious Preoccupation (MAC, M3)	√	√	√	
Body Image (BR-23, M3))	√	√	√	
Manageability (SOC, M0)		√	√	√
Positive Emotion Regulation (CERQ, M0)		√	√	√
PTGI (M3)	√	√		√
Negative Affectivity (PANAS, M3)	√	√		
Positive Affectivity (PANAS, M0)		√	√	
Forward (PACT, M0)		√		√
Family Coping (FARE, M3)	√		√	
Thrombocyte count (M0)	√	√		
HADS Anxiety (M3)	√			
HADS Anxiety (M0)				√
Negative Affectivity (PANAS, M0)		√		
Global QoL (M0)				√
Emotional Support (M3)	√			
Social Support (M3)	√			
Flexibility (PACT)			√	
Trait Resilience (CDRISC, M0)				√
Avoidance (MAC, M3)				√
Community Cohesion (FARE, M3)	√			√
Trauma (PACT)				√
NLR (M0)			√	
Chemotherapy			√	

Generalizability of predictor profiles across clinical sites were explored in more detail through global interpretation analyses SHAPs (Figures F5-F8).

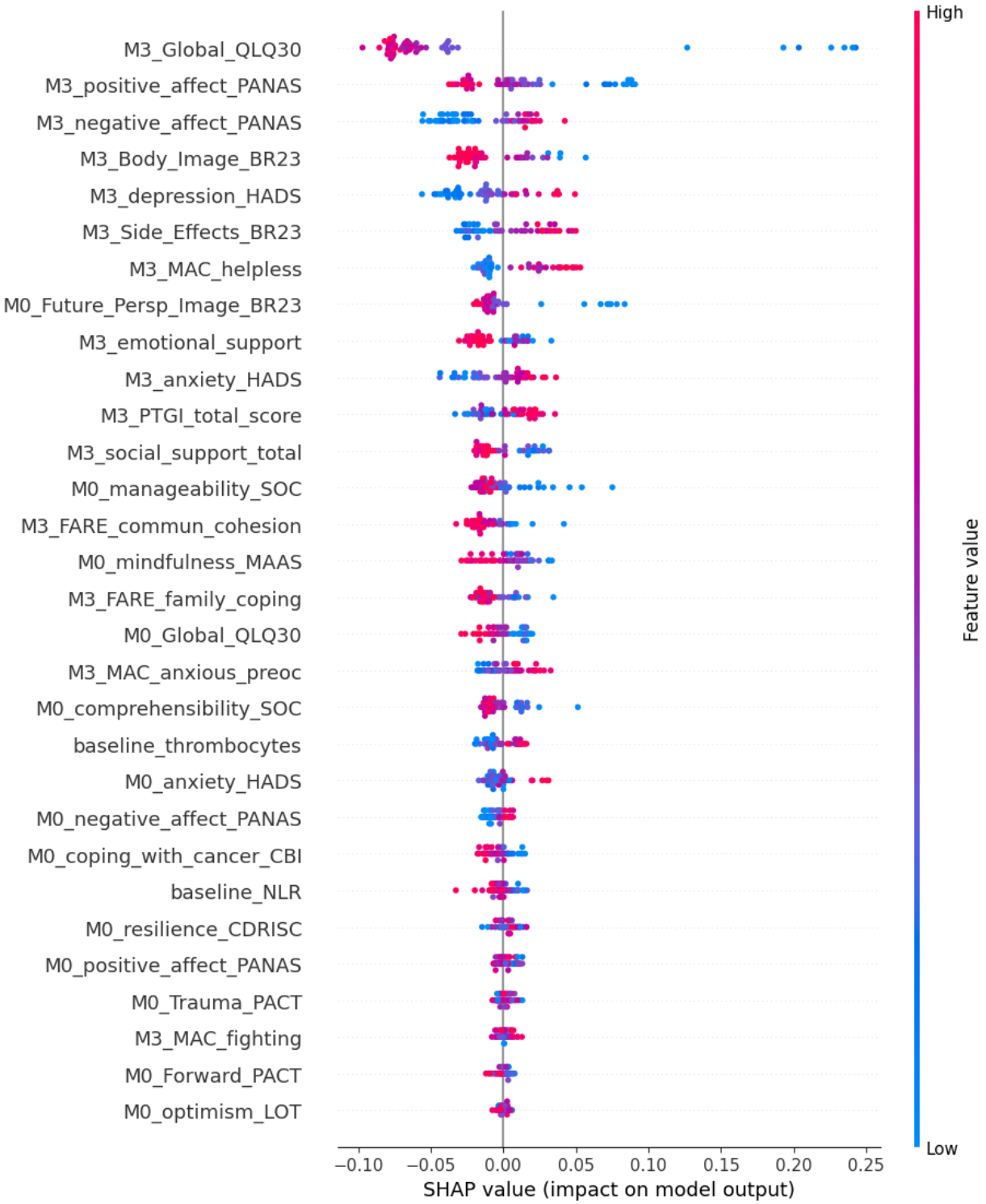

Figure F5. Shapley additive explanations (SHAP) summary plot generated using data from CHAMP in the testing phase of Model F4 (predicting Global QoL at M12 from M0 and M3 variables including QoL and mental health indices).

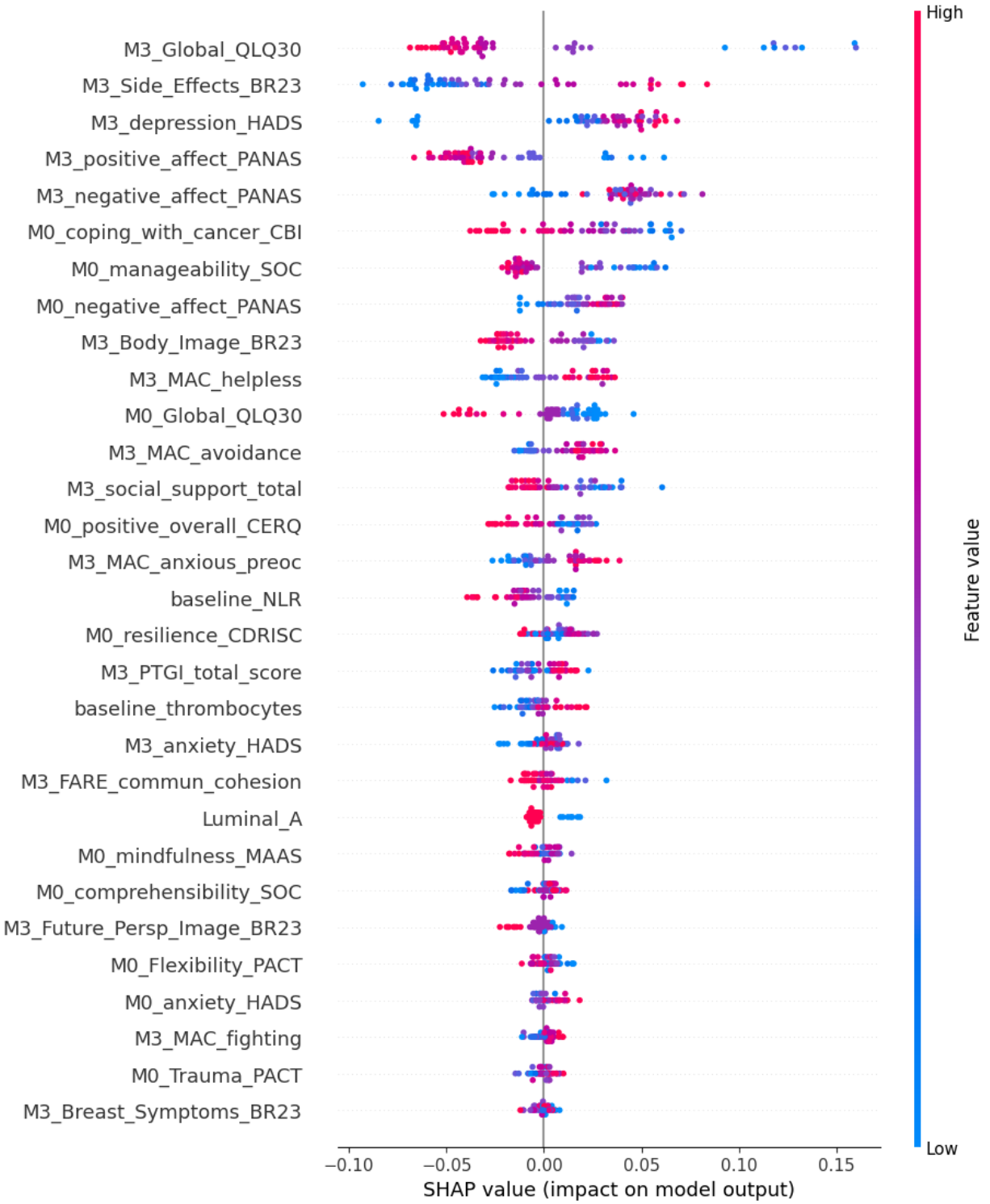

Figure F6. Shapley additive explanations (SHAP) summary plot generated using data from IEO in the testing phase of Model F4 (predicting Global QoL at M12 from M0 and M3 variables including QoL and mental health indices).

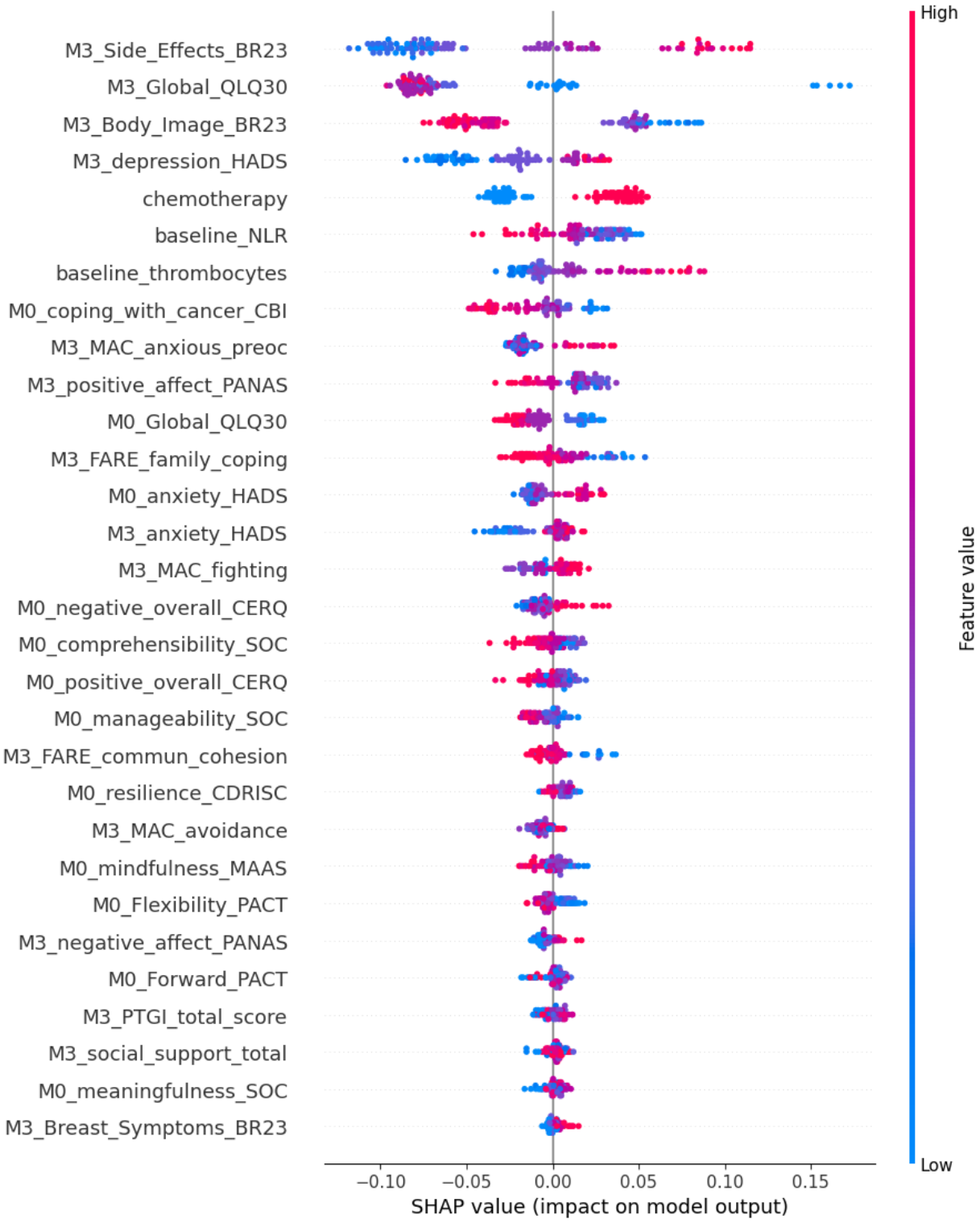

Figure F7. Shapley additive explanations (SHAP) summary plot generated using data from HUS in the testing phase of Model F4 (predicting Global QoL at M12 from M0 and M3 variables including QoL and mental health indices).

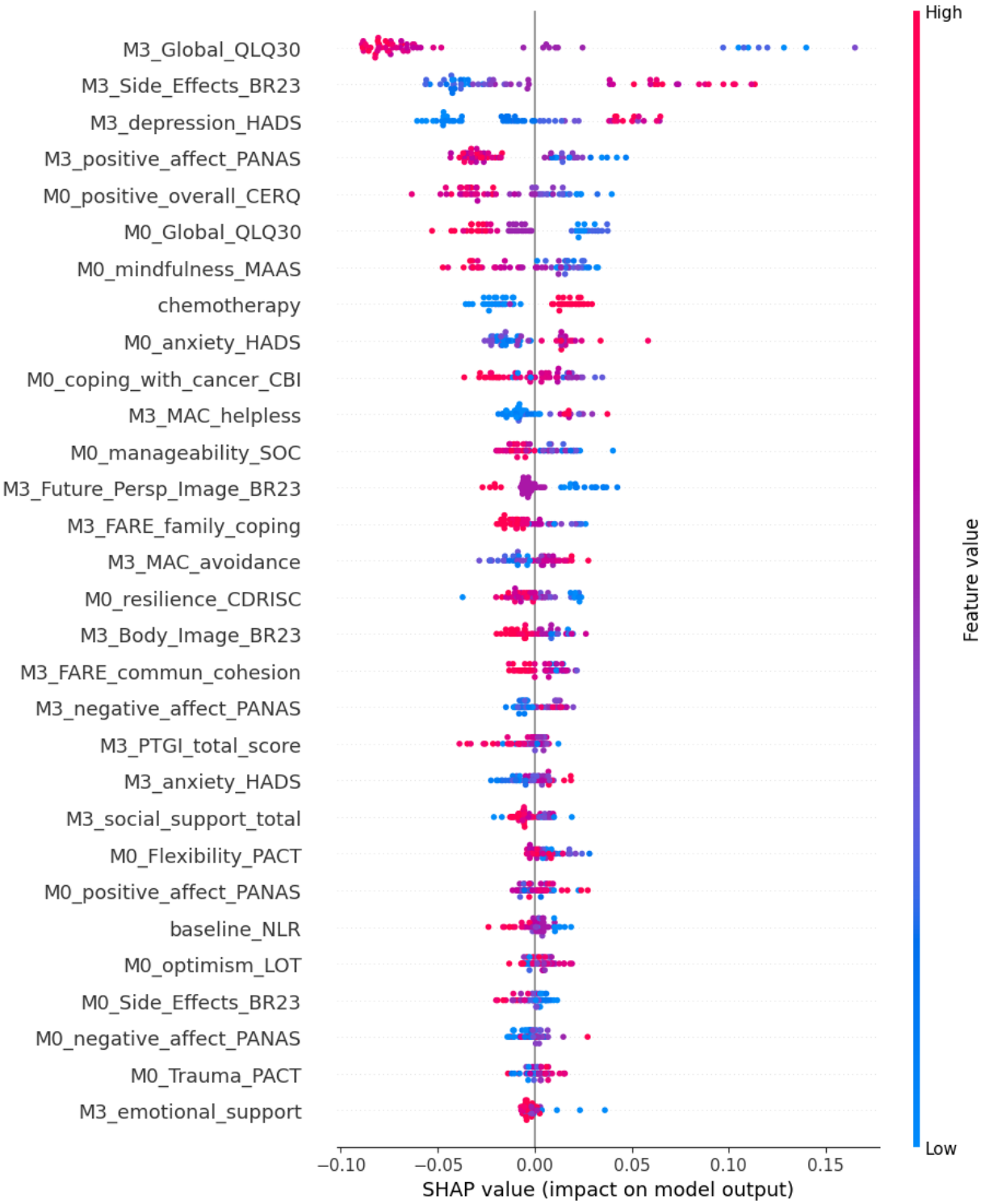

Figure F8. Shapley additive explanations (SHAP) summary plot generated using data from HUJI in the testing phase of Model F4 (predicting Global QoL at M12 from M0 and M3 variables including QoL and mental health indices).

In sum, these results demonstrate a fair level of concordance across clinical samples. However, results in terms of the highest-ranking predictor variables should be considered with caution given the small, and highly unbalanced sample size of the test set in each cross-validation runs.

3.2 *ICCS MODELS*

3.2.1 Classifiers for detecting depression and QoL 18-months trajectories

3.2.1.1 Classifier for detecting patients with poor and good depression trajectory

3.2.1.1.1 Model Description

The binary classifier aims to detect which patients belong to the 'poor' or 'good' mean trajectory during the 18 month period from baseline. The two classes considered emerge from the grouping of the four depression clusters identified during trajectory analysis (D4.3) (Fig. I1). The 'good' trajectory class comprise the decreasing and low depression trajectory groups and the 'poor' trajectory class comprise the increasing and high depression trajectory groups. The features considered were the ones selected based on recursive feature elimination (see D4.3b and Table I4 below).

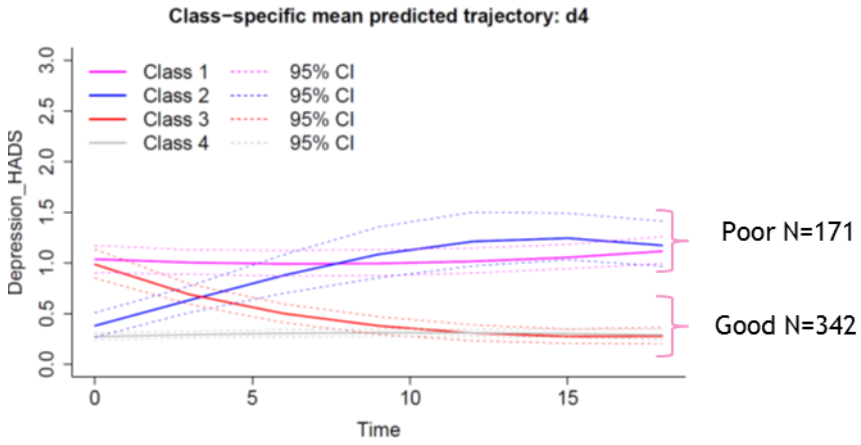

Figure I1: The classes considered

3.2.1.1.2 Nested Cross Validation

Classifier's performance is evaluated by means of nested cross validation (cv) (Table I1). The inner loop of nested cv is responsible for model selection/hyperparameter tuning (validation set), while the outer loop is for error estimation (test set). A stratified four-fold resampling is used for the inner and outer loop. Model parameters are tuned using grid search againstROC (area under the ROC curve) metric with 4-fold cross validation. Class imbalance was handled using downsampling, performed inside the inner cv resampling. Inclusion of M3 variables improves the performance of the classifier. However, the ability of the model to distinguish

between the two classes is good in both cases (mean ROC 0.84 and 0.87 at M0 and M0&M3 respectively) (Fig. I2).

Table I1. Model performance using nested cross validation. Abbreviations: AUC: Area under the receiver operating characteristic curve, sens: sensitivity, spec: specificity, BACC: balanced accuracy, PPV: positive predictive value and NPV: negative predictive value

Predictors	Sens	Spec	AUC	BACC	F1	PPV	NPV
M0 (10 Selected)	0.81±0.06	0.74±0.06	0.84±0.02	0.78±0.01	0.70±0.01	0.62±0.04	0.89±0.02
m0 & m3 (10 SELECTED)	0.82±0.07	0.78±0.08	0.87±0.03	0.80±0.04	0.73±0.04	0.66±0.07	0.90±0.03

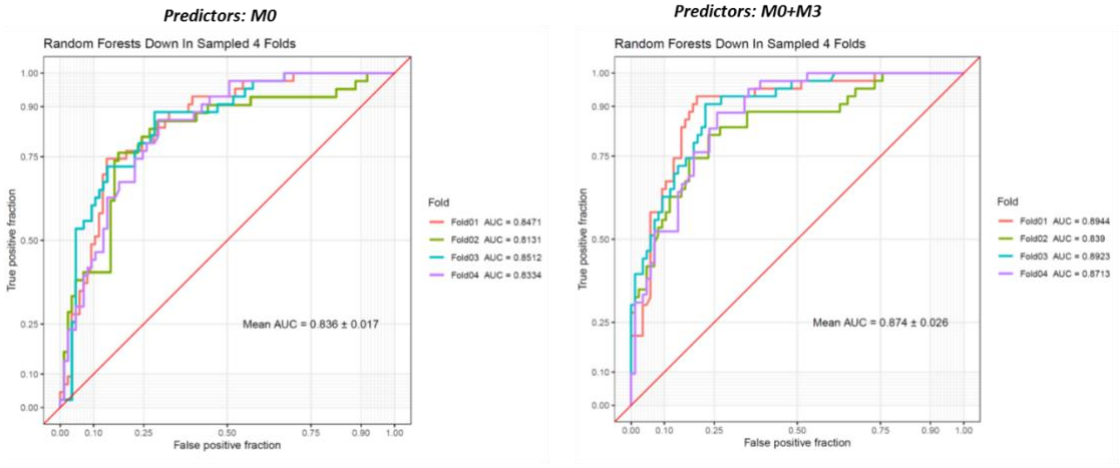

Figure I2: ROC curves for each outer fold of the nested cross validation. Model with 10 selected predictors at M0. Model with 10 selected predictors at M0 and M3.

3.2.1.1.3 Leave-one-hospital out cross-validation: Generalizability across clinical sites

The dataset is split into test and training set based on data origin. Each time, one of the four clinical subsets (i.e. CHAMP, IEO, HUS and HUJI) comprise the test set and the rest three the training set. Such a sampling is not stratified, meaning that the ratio between the target classes is not the same in each fold as it is in the whole dataset. Inner cross validation is as previously described. The distribution of the trajectory classes between the clinical sites is reported in Table I2. The majority of patients in each clinical site is assigned to the 'good' trajectory class, with the exception of IEO where the majority of patients are assigned to the 'poor' trajectory class. Moreover, in the 'good' trajectory class, IEO is under-represented, whereas almost half of the patients come from HUS.

Table I2. *Class distribution* (in *percentage) in each clinical site*

	CHAMP	IEO	HUS	HUJI	
Good	16%	7%	33%	11%	67%
Poor	9%	11%	6%	8%	33%
	24%	18%	39%	19%	

Performance metrics

Table I3 reports various performance evaluation metrics when considering predictors from M0 only or from both M0&M3. The features considered in each case were the ones selected based on recursive feature elimination (see D4.3b and Table I3). Performance when only M0 predictors are considered is acceptable with ROC>0.8 in all cases. Mean balanced accuracy, sensitivity and specificity are 73%, 77% and 69% respectively. Performance increases when M0 and M3 predictors are considered (ROC values are >0.85). Mean balanced accuracy, sensitivity and specificity are 78%, 81% and 75% respectively.

Table I3. Model performance using leave one-hospital-out cross validation. Abbreviations: ROC: Area under the receiver operating characteristic curve, sens: sensitivity, spec: specificity, BACC: balanced accuracy, PPV: positive predictive value and NPV: negative predictive value

PREDICTORS	TRAIN SITES	TEST SITE	GOOD TRAJECTORY CLASS IN TEST SET (N)	POOR TRAJECTORY CLASS IN TEST SET (N)	SENS	SPEC	ROC	BACC	F1	PPV	NPV
M0	IEO, HUS, HUJI	CHAMP	80	45	0.73	0.83	0.83	0.78	0.72	0.70	0.85
	CHAMP, HUS, HUJI	IEO	38	55	0.84	0.53	0.80	0.68	0.77	0.72	0.69
	CHAMP,IEO, HUJI	HUS	167	31	0.74	0.78	0.84	0.76	0.51	0.38	0.94
	CHAMP, IEO, HUS	HUJI	57	40	0.78	0.65	0.81	0.71	0.68	0.61	0.80
M0&M3	IEO, HUS, HUJI	CHAMP	80	45	0.76	0.81	0.85	0.78	0.72	0.69	0.86
	CHAMP, HUS, HUJI	IEO	38	55	0.91	0.61	0.88	0.76	0.83	0.77	0.82
	CHAMP,IEO, HUJI	HUS	167	31	0.81	0.82	0.89	0.81	0.58	0.45	0.96
	CHAMP, IEO, HUS	HUJI	57	40	0.78	0.77	0.88	0.77	0.74	0.70	0.83

Selected Features

As described in D4.3b, feature selection was performed based on recursive feature elimination (RFE). RFE was repeated a number of times (N=10) to evaluate the variability of the selected features. The most frequently selected features were finally chosen (features in bold at first column of Tables I4 & I5). The procedure was also applied in each leave-one-hospital out cross-validation scheme to assess the consistency of selected features across clinical sites. The most (tick with no brackets) and less (tick in brackets) frequently selected features in six repetitions of RFE at each validation scheme are listed in Tables I4 and I5.

At M0, the vast majority of the finally selected features were also among the most frequently selected in all four cross validation schemes. These are anxiety HADS (M0), depression HADS

(M0), coping with cancer CBI (M0), manageability SOC (M0), optimism LOT (M0) and negative affect PANAS (M0). Fear of recurrence FCRI (M0) appeared in all cross validation schemes but was less frequently selected in the scheme that did not contain CHAMP dataset. Two of the selected variables, catastrophizing CERQ (M0) and resilience CDRISC (M0), were selected in three out four validation schemes (the ones that did not contain CHAMP and IEO respectively). Self-efficacy (M0) was frequently selected in two validation schemes (the ones without CHAMP and HUS) and less frequently selected in one (without HUJI). The features pain C30 (M0) and meaningfulness SOC (M0) were frequently selected in three out of four validation schemes. However, because they appeared in less than 50% of the repetitions of RFE procedure when the whole dataset was considered they were not selected for the final models.

There are similar observations when M0&M3 predictors are considered. Eight out of ten features selected based on RFE procedure were also consistently selected when RFE procedure was applied in each cross validation scheme. These features are Anxiety HADS (M3), Depression HADS (M0), Depression HADS (M3), Negative affect PANAS (M3), manageability SOC (M0), optimism LOT (M0), anxious preoccupation MAC (M3), coping with cancer CBI (M0). The features Negative affect PANAS (M0) and resilience CDRISC (M0) were selected in three out of four cross validation schemes.

Overall, a high consistency among the clinical sites is observed. The finally selected features are among the most-frequently selected features across the four cross validation schemes considered. Features that were less frequently selected in each cross validation scheme or were selected in one or two schemes were not among the finally selected features. Concluding, the selected features generalize across all clinical sites.

Table I4: Selected features of six repetitions of recursive feature elimination (RFE) across leave-one-hospital-out cross-validation schemes. Predictors from M0 only. Features in bold are the ones selected based on ten repetitions of recursive feature elimination procedure on whole dataset and are considered in the final models. Tick in brackets correspond to features that are less frequently selected (<50%) among N repetitions of RFE.

	Training set				
	All clinical sites	**IEO, HUS & HUJI**	**CHAMP, HUS & HUJI**	**CHAMP, IEO & HUJI**	**CHAMP, IEO & HUS**
Anxiety_HADS.0	√	√	√	√	√
Depression_HADS.0	√	√	√	√	√
M0_catastrophizing_CERQ	√		√	√	√
Negative_affect_PANAS.0	√	√	√	√	√
M0_coping_with_cancer_CBI	√	√	√	√	√
M0_manageability_SOC	√	√	√	√	√
M0_optimism_LOT	√	√	√	√	√
M0_resilience_CDRISC	√	√		√	√
M0_fear_of_recur_FCRI	√	(√)	√	√	√
general_se_1_item.0	√	√		√	(√)
Pain_QLQ30.0	(√)	√	√	√	(√)
M0_meaningfulness_SOC	(√)	√	√	√	
Soc_Fun_QLQ30.0		√	(√)		(√)
education		√		√	(√)
M0_negative_overall_CERQ			√		
perceived_suppport_1_item.0				√	
Positive_affect_PANAS.0				√	
M0_Forward_PACT				√	(√)
Arm_Symptoms_BR23.0				√	(√)
Global_QLQ30.0		(√)			√
m0_cardio_exercise_min		(√)			√
Side_Effects_BR23.0		(√)			
chemo0_type		(√)			
m0_drinking_EK		(√)			
m0_employment_status		(√)			
m0_exercise_012		(√)			(√)
Role_Fun_QLQ30.0		(√)			(√)
M0_mindfulness_MAAS			(√)		(√)
Fatigue_QLQ30.0				(√)	
m0_BMI				(√)	
m0_employment_status				(√)	
M0_positive_overall_CERQ				(√)	
m0_sick_leave_days					(√)
chemo0_type					(√)
M0_pos_refus_CERQ					(√)
Sex_Funct_BR23.0					(√)

Table I5: Selected features of six repetitions of recursive feature elimination (RFE) across leave-one-hospital-out cross-validation schemes. Predictors from M0 & M3. Features in bold are the ones selected based on ten repetitions of recursive feature elimination procedure on whole dataset and are considered in the final models. Tick in brackets correspond to features that are less frequently selected (<50%) among N repetitions of RFE.

	All clinical sites	**IEO, HUS & HUJI**	**CHAMP, HUS & HUJI**	**CHAMP, IEO & HUJI**	**CHAMP, IEO & HUS**
Anxiety_HADS.3	√	√	√	√	√
Depression_HADS.0	√	√	√	√	√
Depression_HADS.3	√	√	√	√	√
Negative_affect_PANAS.0	√	√		√	√
Negative_affect_PANAS.3	√	√	√	√	√
M0_manageability_SOC	√	√	√	√	√
M0_optimism_LOT	√	√	√	√	√
M0_resilience_CDRISC	√	√		√	√
M3_MAC_anxious_preoc	√	√	√	√	√
M0_coping_with_cancer_CBI	√	√	√	√	√
M3_MAC_helpless	(√)			(√)	√
M3_MAC_avoidance	(√)				
Anxiety_HADS.0	(√)		(√)	√	√
M3_PTGI_spiritual_change		(√)			
general_se_1_item.0		(√)		√	
single_item_cope4.3		(√)			
Pain_QLQ30.0			√	√	
Pain_QLQ30.3			√	(√)	
Global_QLQ30.3			√	(√)	
Positive_affect_PANAS.3			(√)	√	
M0_fear_of_recur_FCRI			(√)	√	
Fatigue_QLQ30.3			(√)	(√)	
Future_Persp_Image_BR23.3			(√)		
M3_mMOS_emotional_support				√	
M3_mMOS_social_support_total				√	
M0_catastrophizing_CERQ				√	√
M0_Forward_PACT				(√)	
Body_Image_BR23.3				(√)	
education_3				(√)	
M0_meaningfulness_SOC				(√)	
M3_mMOS_instumental_support				(√)	
perceived_suppport_1_item.0				(√)	
Positive_affect_PANAS.0				(√)	
Sex_Funct_BR23.3				(√)	
Side_Effects_BR23.3				(√)	

3.2.1.2 Classifier for detecting patients with low decreasing QoL trajectory

3.2.1.2.1 Model Description

The binary classifier aims to detect which patients belong to the low decreasing QoL trajectory (positive class) during the 18 month period from baseline. The negative class emerges from the grouping of the high increasing and moderate QoL trajectory clusters identified during trajectory analysis (D4.3b), as depicted in Fig. I3.

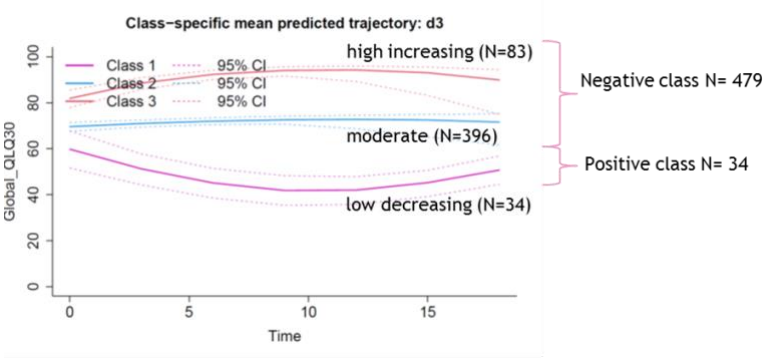

Figure I3: The considered classes

3.2.1.2.2 Nested Cross Validation

Classifier's performance is evaluated by means of nested cross validation (cv) as previously described. The features considered were the ones selected based on recursive feature elimination (see D4.3b and Table I9). The ability of the model to identify low decreasing trajectory is good (mean ROC 0.87) (Table I6, Fig. I4). Mean balanced accuracy, sensitivity and specificity are 82%, 85% and 78% respectively.

Table I6. Model performance using nested cross validation. Abbreviations: ROC: Area under the receiver operating characteristic curve, sens: sensitivity, spec: specificity, BACC: balanced accuracy, PPV: positive predictive value and NPV: negative predictive value

Predictors	Sens	Spec	ROC	BACC	F1	PPV	NPV
M0 & M3 (14 sELECTED)	0.85±0.07	0.78±0.05	0.87±0.06	0.82±0.05	0.36±0.08	0.23±0.06	0.99±0.01

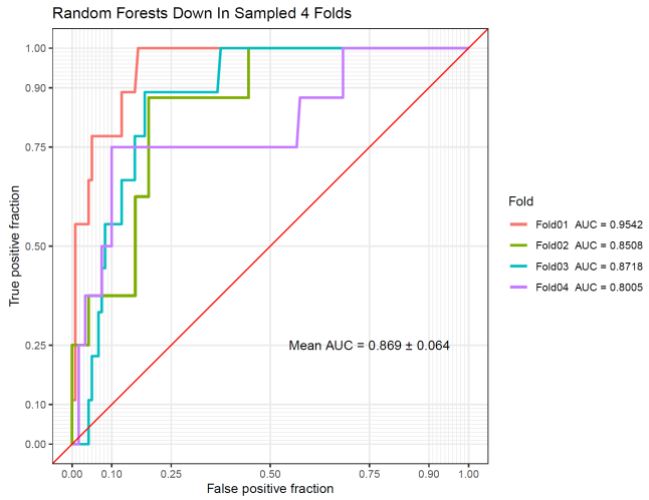

Figure I4: ROC curves for each outer fold of the nested cross validation. Model with 14 selected predictors at M0 and M3.

3.2.1.2.3 Leave-one-hospital out cross-validation: Generalizability across clinical sites

As previously described, the dataset is split into test and training set based on data origin. Each time, one of the four clinical subsets (i.e. CHAMP, IEO, HUS and HUJI) comprise the test set and the rest three the training set. The distribution of the trajectory classes between the clinical sites is reported in Table I7. The vast majority of patients in each clinical site is assigned to the moderate/low trajectory class. Relatively less patients from HUS are assigned to the low decreasing class.

Table I7. *Class distribution* (in *percentage) in each clinical site*

	CHAMP	IEO	HUS	HUJI	
low decreasing	2%	1%	1%	2%	7%
moderate/high	22%	17%	37%	17%	93%
	24%	18%	39%	19%	

Performance metrics

Table I8 reports various performance evaluation metrics when considering predictors from both M0&M3. The features considered were the ones selected based on recursive feature elimination (see D4.3b and Table I9). A small deviation across sites is evident. Despite the small size of the positive class (low decreasing trajectory), the ROC values are >0.85 in all cross validation schemes. Mean balanced accuracy, sensitivity and specificity are 81%, 82% and 80% respectively.

Table I8. Model performance using leave one-hospital-out cross validation. Abbreviations: ROC: Area under the receiver operating characteristic curve, sens: sensitivity, spec: specificity, BACC: balanced accuracy, PPV: positive predictive value and NPV: negative predictive value

PREDICTORS	TRAIN SITES	TEST SITE	NEGATIVE TRAJECTORY CLASS IN TEST SET (N)	LOW DECREASING TRAJECTORY CLASS IN TEST SET (N)	SENS	SPEC	ROC	BACC	F1	PPV	NPV
M0&M3	IEO, HUS, HUJI	CHAMP	115	10	0.70	0.83	0.85	0.76	0.38	0.26	0.97
	CHAMP, HUS, HUJI	IEO	86	7	1.00	0.86	0.95	0.93	0.54	0.37	1.00
	CHAMP,IEO, HUJI	HUS	192	6	0.67	0.81	0.85	0.74	0.17	0.10	0.99
	CHAMP, IEO, HUS	HUJI	86	11	0.91	0.71	0.89	0.81	0.43	0.29	0.98

Selected Features

Eight out of fourteen features selected based on RFE procedure were also consistently selected when RFE procedure was applied in each cross validation scheme. These features are cognitive function C30 (M3), physical function C30 (M3), role function C30 (M3), social function C30 (M3), systemic therapy side effects BR23 (M3), global QoL C30 (M3), pain C30 (M3) and

depression HADS (M3). The features treatment control beliefs (M3), fatigue C30 (M3), coping with cancer CBI (M0) and anxiety (M3) were frequently selected in three out of four validation schemes. Selected in three schemes, but, overall, less frequently, were dyspnoea symptoms C30 (M3) and manageability SOC (M0).

Summarizing, the finally selected features are among the most-frequently selected features across the four cross validation schemes considered. In other words, the selected features generalize across all clinical sites.

Table I9: Selected features of six repetitions of recursive feature elimination (RFE) across leave-one-hospital-out cross-validation schemes. Predictors from M0 & M3. Features in bold are the ones selected based on ten repetitions of recursive feature elimination procedure on whole dataset and are considered in the final models. Tick in brackets correspond to features that are less frequently selected (<50%) among N repetitions of RFE.

	Training set				
	All clinical sites	IEO, HUS & HUJI	CHAMP, HUS & HUJI	CHAMP, IEO & HUJI	CHAMP, IEO & HUS
Cogn_Fun_QLQ30.3	√	√	√	√	√
Depression_HADS.3	√	√	√	√	√
Global_QLQ30.3	√	√	√	√	√
Pain_QLQ30.3	√	√	√	√	√
Phys_Fun_QLQ30.3	√	√	√	√	√
Role_Fun_QLQ30.3	√	√	√	√	√
Side_Effects_BR23.3	√	√	√	√	√
Soc_Fun_QLQ30.3	√	√	√	√	√
bipq2.3	√	(√)	√	√	√
Fatigue_QLQ30.3	√	√	√	√	(√)
M0_coping_with_cancer_CBI	√		√	√	√
Anxiety_HADS.3	√	√	(√)	√	√
Dyspnoea_QLQ30.3	√	√		(√)	(√)
M0_manageability_SOC	√		(√)	(√)	√
M3_MAC_helpless		√		(√)	(√)
baseline_pt		√			
Negative_affect_PANAS.3		√			
Depression_HADS.0	(√)		(√)	(√)	
M0_fear_of_recur_FCRI	(√)			√	
perceived_suppport_1_item.0	(√)				√
Global_QLQ30.0					√
TamoxifenVSother	(√)	(√)	√		
Arm_Symptoms_BR23.3	(√)		(√)	(√)	(√)

Table I9 (cont.) : Selected features of six repetitions of recursive feature elimination (RFE) across leave-one-hospital-out cross-validation schemes. Predictors from M0 & M3. Features in bold are the ones selected based on ten repetitions of recursive feature elimination procedure on whole dataset and are considered in the final models. Tick in brackets correspond to features that are less frequently selected (<50%) among N repetitions of RFE.

	Training set				
	All clinical sites	IEO, HUS & HUJI	CHAMP, HUS & HUJI	CHAMP, IEO & HUJI	CHAMP, IEO & HUS
general_se_1_item.0	(√)		√		
M3_FARE_commun_cohesion	(√)		√	(√)	√
perceived_suppport_1_item.3	(√)	(√)	(√)		√
Positive_affect_PANAS.3	(√)	(√)	(√)	(√)	
Anxiety_HADS.0	(√)			(√)	
baseline_alt	(√)	(√)	(√)	√	(√)
general_se_1_item.3	(√)				
m0_BMI	(√)	(√)	(√)	(√)	
m0_exercise_012	(√)	(√)	√		
M0_optimism_LOT	(√)				
m0_sick_leave_days	(√)		(√)		
M3_MAC_anxious_preoc	(√)	(√)		√	
Negative_affect_PANAS.0	(√)				
Sex_Funct_BR23.3	(√)		(√)	(√)	(√)
Dyspnoea_QLQ30.0		(√)			
LifeEvents_012.3		(√)	√		
M0_fear_of_recur_FCRI		(√)	(√)		
Insomnia_QLQ30.3			(√)		
baseline_leukocytes			(√)		
Future_Persp_Image_BR23.0			(√)		
general_se_1_item.3			(√)		
Body_Image_BR23.3				(√)	
M0_optimism_LOT				(√)	(√)
m3_do_you_do_any_activities_to_support_your_wellbeing				(√)	(√)
Pain_QLQ30.0				(√)	
Upset_Hair_Image_BR23B.3				(√)	
m0_sick_leave_days					(√)
m3_domestic_help_days					(√)
Negative_affect_PANAS.0					(√)
Pain_QLQ30.0					(√)
bipq1.3					(√)
Constipation_QLQ30.0					(√)
Depression_HADS.0					(√)
M0_acceptance_CERQ					(√)
m0_exercise_012					(√)
M0_other_blame_CERQ					(√)
M0_perspective_CERQ					(√)
M0_pos_reapp_CERQ					(√)
M3_PTGI_appreciation_of_life					(√)
M3_PTGI_new_possibilities					(√)
Nausea_QLQ30.3					(√)
single_item_cope2.3					(√)
single_item_cope4.3					(√)

3.2.1.3 Classifier for detecting patients with high increasing QoL trajectory

3.2.1.3.1 Model Description

The binary classifier aims to detect which patients belong to the high increasing QoL trajectory (positive class) during the 18 month period from baseline. The negative class emerges from the grouping of the low decreasing and moderate QoL trajectory clusters identified during trajectory analysis (D4.3b), as depicted in Fig. I5.

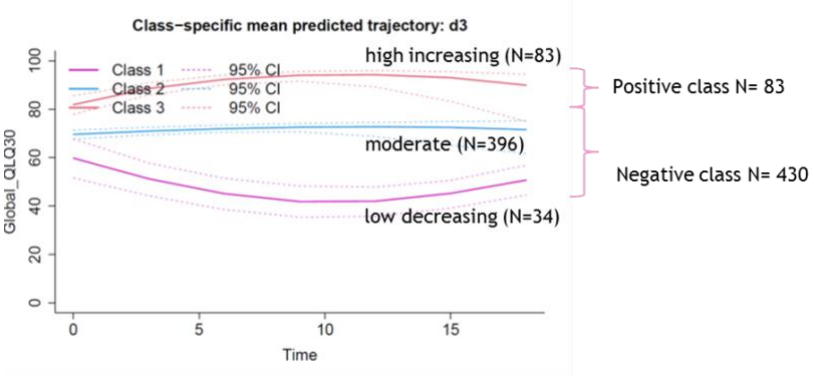

Figure I5: The considered classes

3.2.1.3.2 Nested Cross Validation

Table I10 reports various performance evaluation metrics when considering specific predictors from both M0&M3. The features considered were the ones selected based on recursive feature elimination (see D4.3b and Table I13). Mean ROC is 0.85 (Fig I6). Mean balanced accuracy, sensitivity and specificity are 79%, 81% and 78% respectively (Table I10). Overall performance is good.

Table I10. Model performance using nested cross validation. Abbreviations: ROC: Area under the receiver operating characteristic curve, sens: sensitivity, spec: specificity, BACC: balanced accuracy, PPV: positive predictive value and NPV: negative predictive value

Predictors	Sens	Spec	ROC	BACC	F1	PPV	NPV
m0 & m3 (22 SELECTED)	0.81±0.05	0.78±0.03	0.85±0.05	0.79±0.04	0.56±0.05	0.41±0.04	0.95±0.01

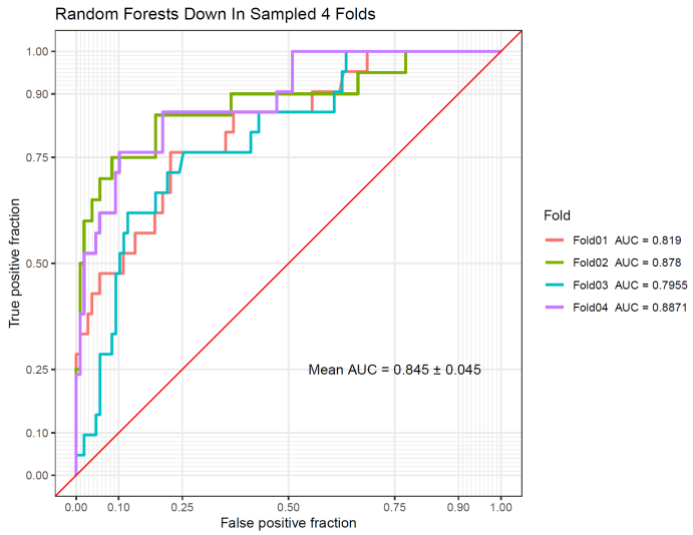

Figure I6: ROC curves for each outer fold of the nested cross validation. Model with 22 selected predictors at M0 and M3.

3.2.1.3.3 Leave-one-hospital out cross-validation: Generalizability across clinical sites

The distribution of the trajectory classes between the clinical sites is reported in Table I11. The majority of patients in each clinical site is assigned to the moderate/low trajectory class. Relatively less patients from IEO are assigned to the high increasing class.

Table I11. *Class distribution* (in *percentage) in each clinical site*

	CHAMP	IEO	HUS	HUJI	
moderate/low	21%	16%	32%	15%	84%
high increasing	4%	2%	7%	4%	16%
	24%	18%	39%	19%	

Performance metrics

Table I12 reports various performance evaluation metrics when considering predictors from both M0&M3. The features considered were the ones selected based on recursive feature elimination (see D4.3b and Table I13). A somewhat higher deviation across sites is evident than in the case of low decreasing trajectory. The ROC values are >0.81 in all cross validation schemes. Mean balanced accuracy, sensitivity and specificity are 77%, 77% and 78% respectively.

Table I12. Model performance using leave one-hospital-out cross validation. Abbreviations: ROC: Area under the receiver operating characteristic curve, sens: sensitivity, spec: specificity, BACC: balanced accuracy, PPV: positive predictive value and NPV: negative predictive value

PREDICTORS	TRAIN SITES	TEST SITE	GOOD TRAJECTORY CLASS IN TEST SET (N)	POOR TRAJECTORY CLASS IN TEST SET (N)	SENS	SPEC	ROC	BACC	F1	PPV	NPV
M0	IEO, HUS, HUJI	CHAMP	107	18	0.89	0.70	0.92	0.79	0.48	0.33	0.97
	CHAMP, HUS, HUJI	IEO	84	9	0.89	0.83	0.92	0.86	0.52	0.36	0.99
	CHAMP,IEO, HUJI	HUS	164	34	0.65	0.82	0.80	0.73	0.51	0.42	0.92
	CHAMP, IEO, HUS	HUJI	75	22	0.64	0.79	0.81	0.71	0.54	0.47	0.88

Selected Features

Eight out of twenty two features selected based on RFE procedure were also consistently selected when RFE procedure was applied in each cross validation scheme. These features are depression HADS (M3), fatigue C30 (M3), self-efficacy (M0), global QoL C30 (M3), role function C30 (M3), communication & cohesion FARE (M3), physical function C30 (M3), future perspective BR23 (M3). The following eleven features were most frequently selected in three validation schemes (some also appeared less frequently in the fourth scheme): cognitive function C30 (M3), social function C30 (M3), positive affect PANAS (M3), personal *control* over the illness (M3), coping with cancer CBI (M0), resilience CDRISC (M0), pain C30 (M3), anxiety (M3), physical function C30 (M3), global QoL C30 (M0) and family coping FARE (M3). Frequently selected in two validation schemes were perceived support (M3), depression HADS (M0) and helpless MAC (M3).

Summarizing, the finally selected features comprise the most-frequently selected features across the four cross validation schemes.

Table I13: Selected features of six repetitions of recursive feature elimination (RFE) across leave-one-hospital-out cross-validation schemes. Predictors from M0 & M3. Features in bold are the ones selected based on ten repetitions of recursive feature elimination procedure on whole dataset and are considered in the final models. Tick in brackets correspond to features that are less frequently selected (<50%) among N repetitions of RFE.

	Training set				
	All clinical sites	**IEO, HUS & HUJI**	**CHAMP, HUS & HUJI**	**CHAMP, IEO & HUJI**	**CHAMP, IEO & HUS**
Depression_HADS.3	√	√	√	√	√
Fatigue_QLQ30.3	√	√	√	√	√
general_se_1_item.0	√	√	√	√	√
Global_QLQ30.3	√	√	√	√	√
Role_Fun_QLQ30.3	√	√	√	√	√
M3_FARE_commun_cohesion	√	√	√	√	√
Phys_Fun_QLQ30.3	√	√	√	√	√
Future_Persp_Image_BR23.3	√	√	√	√	√
Cogn_Fun_QLQ30.3	√	√	√	√	(√)
Soc_Fun_QLQ30.3	√	√	√	√	(√)
Positive_affect_PANAS.3	√	√	√	√	(√)
bipq1.3	√	√	(√)	√	√
M0_coping_with_cancer_CBI	√		√	√	√
M0_resilience_CDRISC	√	√	√		√
Pain_QLQ30.3	√		√	√	√
Anxiety_HADS.3	√	√		√	√
Phys_Fun_QLQ30.0	√	√	√		√
Global_QLQ30.0	√	√	√		√
M3_FARE_family_coping	√	√	√	√	
perceived_suppport_1_item.3	√	√	(√)		√
Depression_HADS.0	√	√			√
M3_MAC_helpless	√	√		√	(√)
Breast_Symptoms_BR23.3		√			
perceived_suppport_1_item.0	(√)	√			(√)
Fatigue_QLQ30.0		√			
M0_meaningfulness_SOC		√			(√)
M3_MAC_fighting		√	(√)		
Side_Effects_BR23.3	(√)		(√)	√	
baseline_ki67	(√)		(√)		(√)
Financial_QLQ30.3	(√)	√			(√)
M0_optimism_LOT	(√)		(√)	√	√
Negative_affect_PANAS.3	(√)	√		√	(√)
Pain_QLQ30.0	(√)	√	√		√
Anxiety_HADS.0	(√)			√	(√)

Table I13 (cont.) : Selected features of six repetitions of recursive feature elimination (RFE) across leave-one-hospital-out cross-validation schemes. Predictors from M0 & M3. Features in bold are the ones selected based on ten repetitions of recursive feature elimination procedure on whole dataset and are considered in the final models. Tick in brackets correspond to features that are less frequently selected (<50%) among N repetitions of RFE.

	Training set				
	All clinical sites	**IEO, HUS & HUJI**	**CHAMP, HUS & HUJI**	**CHAMP, IEO & HUJI**	**CHAMP, IEO & HUS**
general_se_1_item.3	(√)		(√)	√	(√)
baseline_thrombocytes	(√)	√			
chemo0_type	(√)		(√)		
Fatigue_QLQ30.0	(√)				(√)
LifeEvents_012.0	(√)	√		√	
M0_fear_of_recur_FCRI	(√)		(√)	√	
m0_income	(√)				
M0_mindfulness_MAAS	(√)		√	√	(√)
M0_Polarity_PACT	(√)		(√)		
M3_MAC_anxious_preoc	(√)			√	(√)
M3_mMOS_social_support_total	(√)	√			(√)
M3_mMOS_emotional_support		√			√
M3_MAC_avoidance		√			
Insomnia_QLQ30.3		√	(√)		
Arm_Symptoms_BR23.3		√			
Anxiety_HADS.3			(√)		
Positive_affect_PANAS.0			(√)		
cancer_subtypes			(√)		(√)
M0_BMI			(√)		
M0_comprehensibility_SOC					(√)
Soc_Fun_QLQ30.0			(√)		
Negative_affect_PANAS.0				√	
Body_Image_BR23.3				√	

4. Finalization of the BOUNCE In Silico Resilience Trajectory Predictor

4.1 FORTH MODELS

4.1.1 Implementation of prediction models within the Decision Support Tool (FORTH)

Following extensive modeling work on the total sample (D4.3b, first part) and the across-site validation experiments (Section 3.1, present document) the functionalities of the Decision Support Tool (DST) were revised accordingly. The final DST was designed on the following principles:

(i) flexibility toward future use in clinical settings,
(ii) performance accuracy in predicting key aspects of patient well-being,
(iii) robustness in formulating personalized risk profiles of potentially modifiable patient characteristics, and
(iv) directly linking personalized needs assessment with concrete suggestions regarding psychological prevention strategies.

Regarding the first principle, the DST incorporates models addressing eight distinct clinical outcomes:

(a) Prediction of overall mental health status ("good" vs"poor") at M12 post-diagnosis
(b) Prediction of overall mental health status ("good" vs "poor") at M18 post-diagnosis
(c) Prediction of Global QoL status ("good" vs "poor") at M12 post-diagnosis
(d) Prediction of Global QoL status ("good" vs "poor") at M12 post-diagnosis
(e) Prediction of mental health decline ("good" at M0 or M6 → "poor" at M12 vs "good" at M0 or M6 → "good" at M12
(f) Prediction of mental health decline ("good" at M0 or M6 → "poor" at M18 vs "good" at M0 or M6 → "good" at M18
(g) Prediction of QoL decline ("good" at M0 or M6 → "poor" at M12 vs "good" at M0 or M6 → "good" at M12
(h) Prediction of QoL decline ("good" at M0 or M6 → "poor" at M18 vs "good" at M0 or M6 → "good" at M18

Models (a-d) address the need to identify patients at risk of either overall poor mental health (or global QoL, respectively) at a particular end-time. These models were deemed more appropriate in terms of classification performance for patients who report poor mental health (or global QoL, respectively) at the time of diagnosis. Models (e-h) address the need to identify patients at risk of declining mental health (or global QoL, respectively) having displayed adequate classification performance on the subgroup of patients who reported good mental health (or global QoL, respectively) at the time of diagnosis. Added clinical flexibility is afforded by the tool which permits prediction of key end points (overall "poor" vs "good" or "stable good" vs "decline") based on available psychological and life-style measurements (1) at the time of diagnosis and 3 months later, (2) only at 6 months post-diagnosis, or (3) at the time of diagnosis and at 6 months post-diagnosis. In sum, 24 different clinical scenarios can be accommodated by the final version of the DST (see Table F5).

Regarding the second principle, selected models displayed adequate classification accuracy (especially in the total sample). Complementary models (e.g., contrasting "stable poor" vs. "improving" overall mental health (or QoL, respectively) over 12 or 18 months post-diagnosis displayed inferior performance (AUCs<.70).

The third principle was addressed through model agnostic analyses at the local (patient-specific) level which permit identification of specific psychological or lifestyle characteristics, which are deemed to be underdeveloped in a given patient based on available and appropriately timed measurements. Thus, the platform reviews scores from the break-down plot, derived from the selected prediction model for a given patient. Then the platform compares the patient's actual scores on the displayed variables to the population distribution and identifies variables where the patient's score is above the 75%tile for "risk" variables (such as anxious preoccupation) or below the 25%tile for "protective" variables (such as coping). Variables that meet these criteria are "flagged" as potential targets of prevention strategies.

The fourth principle was addressed by composing clinical recommendations each targeting a specific underdeveloped or deficient psychological or lifestyle characteristics of the patient (clinical recommendations are supplied in a separate document). The platform integrates appropriate recommendations for a given patient into a single document in two versions. One (abbreviated) version addressed to clinicians who come in direct contact with the patient but are not trained in administering systematic psychotherapeutic support (such as physicians, nurses, and social workers). An extended version is also available for use by mental health professionals who have some training in psychotherapy interventions.

Table F5. Complementary models included in the DST platform.

	Predicted outcome	**Prediction end-point**	**Predictors**
1	Mental health-overall	M12	All M0 & M3
2	Mental health-overall	M18	All M0 & M3
3	Mental health-deterioration	M12	All M0 & M3
4	Mental health-deterioration	M18	All M0 & M3
5	QoL-overall	M12	All M0 & M3
6	QoL-overall	M18	All M0 & M3
7	QoL-deterioration	M12	All M0 & M3
8	QoL-deterioration	M18	All M0 & M3
9	Mental health-overall	M12	M0 clinical and all M6
10	Mental health-overall	M18	M0 clinical and all M6
11	Mental health-deterioration	M12	M0 clinical and all M6
12	Mental health-deterioration	M18	M0 clinical and all M6
13	QoL-overall	M12	M0 clinical and all M6
14	QoL-overall	M18	M0 clinical and all M6
15	QoL-deterioration	M12	M0 clinical and all M6
16	QoL-deterioration	M18	M0 clinical and all M6
17	Mental health-overall	M12	All M0 & M6
18	Mental health-overall	M18	All M0 & M6
19	Mental health-deterioration	M12	All M0 & M6
20	Mental health-deterioration	M18	All M0 & M6

21	QoL-overall	M12	All M0 & M6
22	QoL-overall	M18	All M0 & M6
23	QoL-deterioration	M12	All M0 & M6
24	QoL-deterioration	M18	All M0 & M6

In sum, the DST user will have several options based on their clinical needs—in terms of both prediction endpoints and capabilities to engage diverse prevention strategies—and also according to the timing of available psychological and life-style data. These features are expected to facilitate the applicability of the DS platform for a wider variety of clinical scenario and settings.

4.2. ICCS MODELS

4.2.1 Additional models within the BOUNCE platform (ICCS)

Table I14 lists additional models uploaded on the BOUNCE platform. They view resilience as a process. Models I1 & I2 aim to identify patients at risk of poor overall depression trajectory over a period of 18 months after baseline. A flexibility in the timing of the predictors is provided. Model I1 can be applied at the time of diagnosis/baseline, whereas model I2 requires psychological measurements at baseline and 3 months later. Model I3 aims to detect patients at risk of poor/low-decreasing QoL trajectory over a period of 18 months after baseline. These patients are characterized by a high mean score of depression throughout the period of interest (D4.3b). It can be applied 3 months following baseline and utilizes measurements from two time points (baseline and 3 months later). Model I4 aims to detect patients of excellent/high-increasing QoL trajectory over a period of 18 months after baseline. These patients are characterized by a very low mean score of depression throughout the period of interest (D4.3b). It can be applied 3 months following baseline and utilizes measurements from two time points (baseline and 3 months later). The model does not identify patients at risk of poor resilience, but, depending on clinical needs or interests, it could be applied complementary to other models e.g. to Model 2 to further distinguish between patients of excellent and adequate resilience in terms of both depression and QoL. All models displayed adequate classification performance based on nested cross validation and leave-one-hospital-out cross validation (see section 3.2.1). These models are not linked to the clinical recommendations due to the agnostic nature of the procedures followed to categorize the patients (i.e. through latent class mixed effects analysis-see D4.3b).

Table I14. Additional models included in the BOUNCE platform.

	Predicted outcome	**Prediction period**	**Predictors**
1	Depression trajectories	M18	Selected M0
2	Depression trajectories	M18	Selected M0 & M3
3	QoL-low decreasing trajectory	M18	Selected M0 & M3

4	QoL-high increasing trajectory	M18	Selected M0 & M3

5. Quality Assurance of the BOUNCE In Silico Resilience Trajectory Predictor

Quality assurance of the approach and models developed in BOUNCE was carried out by analysing the coverage of the reported performance measurements of the FORTH and ICCS models (Section 3 in D7.1) and by conducting external user experiments (Section 5 in D7.1). The aim of analysing the performance measurements was to indicate that all relevant performance and functionality aspects of the developed models have been measured and analysed. The aim of the user experiments was to measure whether the machine learning algorithm incorporated in the clinical decision support system improves clinicians' performance to predict patients' state during treatment process. That is, the user experiments approximated the performance of the methods for a real clinical environment. The results of the both analyses are presented in details in D7.1.

6. External Validation Considerations

Despite systematic efforts to locate and utilize external to BOUNCE pertinent data sets, e.g. from an external to BOUNCE clinical study deployed in IEO (European Institute of Oncology), a strict validation of the models developed through this channel has proven non feasible. This appears to be due to the novelty and the originality of the BOUNCE approach. What could be done in the future in order to externally (in relation to the BOUNCE clinical centres) validate the BOUNCE models, would be to utilize the BOUNCE platform in conjunction with its In Silico Prediction Repository (ISPR) for new patients, predict the relevant trajectories, store the predictions into the ISPR and compare the actual responses and trajectories of the new patients with the respective model predictions.

7. Conclusions

Based on the quantitative data and the graphical visualizations of the performance of the various in silico resilience related trajectory predictors presented in this document, it appears that the model performance is overall good and the models are promising. The results demonstrate a fair level of concordance across clinical samples. However, results in terms of the highest-ranking predictor variables should be considered with caution given the small, and highly unbalanced sample size of the test set in each cross-validation runs. A palette of such finalized models has been uploaded on the BOUNCE Decision Support Tool. Quality assurance aspects have also been addressed in D7.1,