Supplementary Material

Table des matières

Supplementary methods

Models

A L2-regularization term was added to the linear regression model to prevent overfitting given the limited amount of data available. This model does not capture non-linear interaction between variables but provide a simple score for each feature, measuring its contribution to the final prediction. On the other hand, ensemble tree-based models such as random forest or gradient boosted trees can unravel nonlinear interactions between the variables. The Random Forest model aggregates multiple decision trees grown on bootstrap subsamples of the training set, while the gradient boosted trees methods successively build decision trees, learning from the mistakes of the previous ones. The gradient boosted trees methods often achieve the best results on structured data as available in our study.

Variable selection

A variable selection process was defined to select the most important features and help to train the models. An automatic backward feature selection (17) was used. It firstly considers all the available features to compute the metrics (AUROC for the prediction of the therapeutic response and MSE for the prediction of the ΔDAS28) of a random forest model using a 5 folds cross-validation on the training data. Then, it successively removes features to recompute the metrics and eliminates the feature with the least impact on model's performances. This feature selection stops when all remaining features significantly contribute to the metrics (i.e., removing one more feature would strongly deteriorate the model's performances, a level assessed with the elbow method). The variable selection was performed for each of the drug class and all the models tested used the same variables.

To deal with missing data (details in Supplementary Table 1) we compared multiple methods on the training set and selected the one that yielded the best results for each drug. The compared methods were mean imputation, median imputation, k-nearest-neighbours-based imputation (18), and MICE (Multiple Imputation by Chained Equations) (19). The Python library Scikit-Learn (20) was used to implement the regression model and the random forest model as well as the KNN and MICE imputation methods. Both XGBoost and CatBoost libraries were compared for the gradient boosted trees.

Statistical methods

For all scores, 95% confidence intervals were computed using bootstrap with 100 resamples. Models were assessed on 100 datasets sampled with replacements from the original dataset and the 95% confidence interval from the resampling scores was outputted. To compare cross-validated models and test if a statistical difference existed between the AUC, we computed p-value through a Wilcoxon test between model results.

Supplementary results

Evaluation of ΔDAS28 prediction

When comparing the four models on all TNFi, ridge regression performed significantly better than the other models ($p < 0.0001$, detail in Supplementary Figure 2 A). When etanercept alone was considered, ridge regression performed better ($p < 0.0001$, detail in Supplementary Figure 2 B) than random forest and XGBoost but the improvement compared to CatBoost was not significant. When monoclonal antibodies were considered alone, ridge regression yielded better results than the other models ($p < 0.0001$, detail in Supplementary Figure 2 C). Overall, the ridge regression model had the best performances suggesting limited non-linear interaction effects between the variables.

For each drug, the model that performed the best on the training set was then evaluated on the validation database (ABIRISK), the results are presented in Supplementary Table 3 with their 95% confidence intervals. The Mean Average Error (MAE) was computed on the validation set as it is easier to interpret. Overall, our models predicted the ΔDAS28 after treatment initiation with an error around 1.1 points of DAS28. This error should be compared to the 0.6 threshold, that is the minimum variation of the DAS28 that is clinically relevant.

Interpretation of the ΔDAS28 prediction

Bouget V, et al. RMD Open 2022; 8:e002442. doi: 10.1136/rmdopen-2022-002442

For the prediction of the ΔDAS28, the model with the best performances on the validation dataset is the ridge regression model. To interpret this model, we plotted its coefficients for each class of drugs. The coefficients values are proportional to their impact on the output and can be compared as the features were normalized prior training (Supplementary Figure 3).

Supplementary Tables

Supplementary Table 1. Missing values statistics.

Feature's name	Missing values in ESPOIR	Missing values in ABIRISK
Age	0%	0%
Sex	0%	0%
Weight	1%	3%
Height	0%	3%
Body Mass Index	1%	3%
Autoimmunity Family history	0%	2%
Ever smokers	0%	0%
Current smokers	0%	0%
Smoking cumulative dose - pack-year	0%	43%
Past pregnancy (among sex=female)	0%	26%
DAS28	1%	0%
CRP	1%	2%
Erythrocyte sedimentation rate	0%	3%
Creatininemia	2%	3%
AST	1%	4%
ALT	1%	2%
White blood	1%	2%
Neutrophils	2%	2%
Lymphocytes	1%	2%
Presence of Anti-Citrullinated Protein Antibody	0%	3%
Presence of Rheumatoid factor IgM	0%	3%

Supplementary Table 2. Result of the variable selection process for the ΔDAS28 prediction.

All TNFi	**Etanercept**	**Monoclonal antibodies TNFi**
Sex	Sex	DAS28
DAS28	DAS28	Ever Smoked
BMI	BMI	Creatinine
Ever smoked	Weight	AST
ESR	ALT	ALT
RF IgM		
Family history		
Past pregnancy		

Supplementary Table 3. Performances of the best models for the ΔDAS28 prediction. The best model is selected on the training set (ESPOIR) and the replication is assessed on the validation set (ABIRISK).

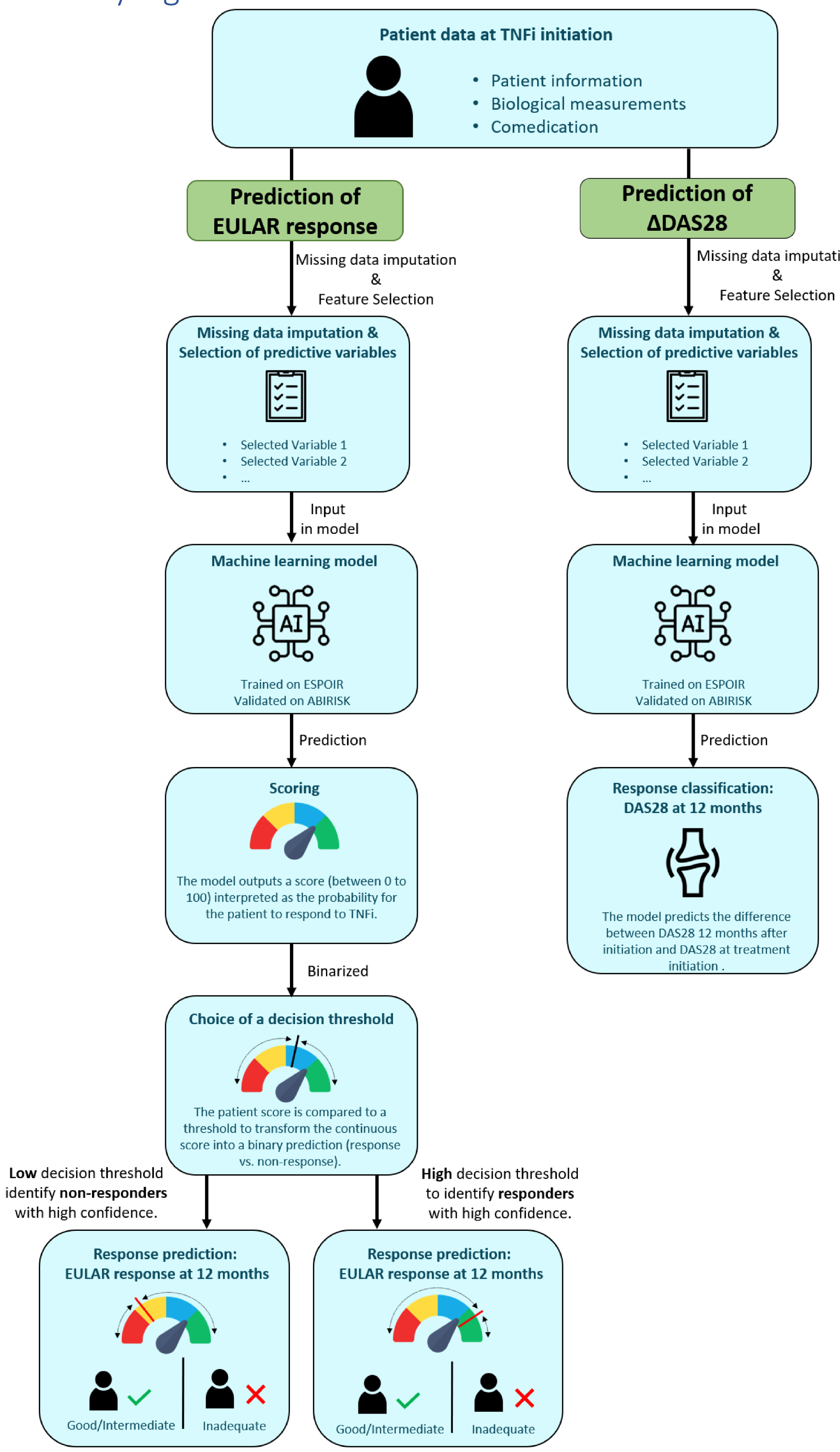
DRUG	BEST MODEL	MEAN SQUARE ERROR IN DAS28 POINTS (TRAIN)	MEAN SQUARE ERROR IN DAS28 POINTS (VALIDATION)	MEAN AVERAGE ERROR IN DAS28 POINTS (VALIDATION)
Overall TNFi	Ridge regression	1.7 (1.6 – 1.8)	Not evaluated since worse than drug-class-specific models on the training set	Not evaluated since worse than drug-class-specific models on the training set
Etanercept	CatBoost	1.5 (1.5 – 1.6)	1.8 (1.4 – 2.4)	1.1 (0.9 - 1.3)
Monoclonal anti-TNF antibodies	Ridge regression	2.1 (2 - 2.1)	2 (1.4 - 2.9)	1.2 (1.0 - 1.4)

Supplementary Table 4. Best model, F1-score and probability cut-off on the training set when optimizing for F1-score instead of AUROC. Optimizing either the AUROC or the F1-score little change the F1-score and the decision cut-off. We did not conduct the analysis on the test set to avoid overfitting,

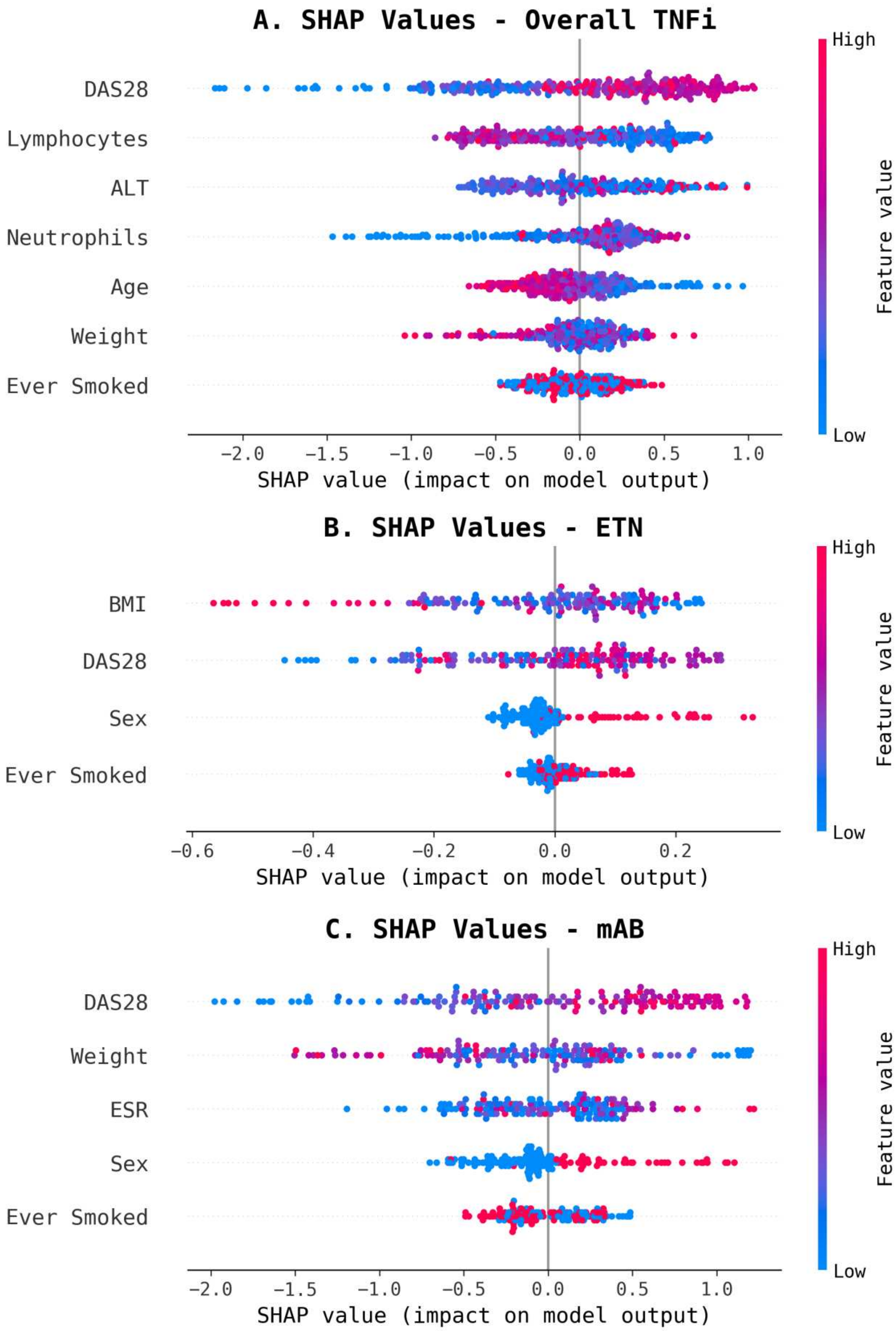
Drug	AUROC Optimization			F1-score optimization		
	Best model	F1-score	Decision cut-off for high confidence in non-response	Best model	F1-score	Decision cut-off
All TNFi	Catboost	78% (75%-79%)	0.32	Catboost	78% (75%-79%)	0.31
Etanercept	Random Forest	81% (78%-84%)	0.31	XGBoost	83% (79% - 85%)	0.21
Monoclonal TNFi antibodies	Catboost	74% (71%-77%)	0.22	Catboost	74% (71%-77%)	0.24

Bouget V, *et al. RMD Open* 2022; 8:e002442. doi: 10.1136/rmdopen-2022-002442

Supplementary Figures

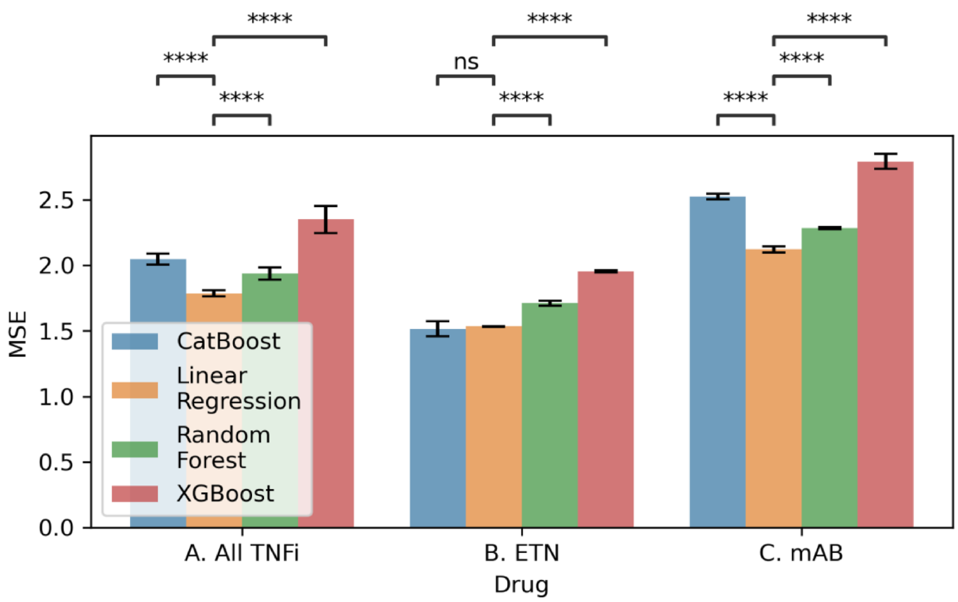


Supplementary Figure 1. Presentation of the different blocks to predict the therapeutic response.

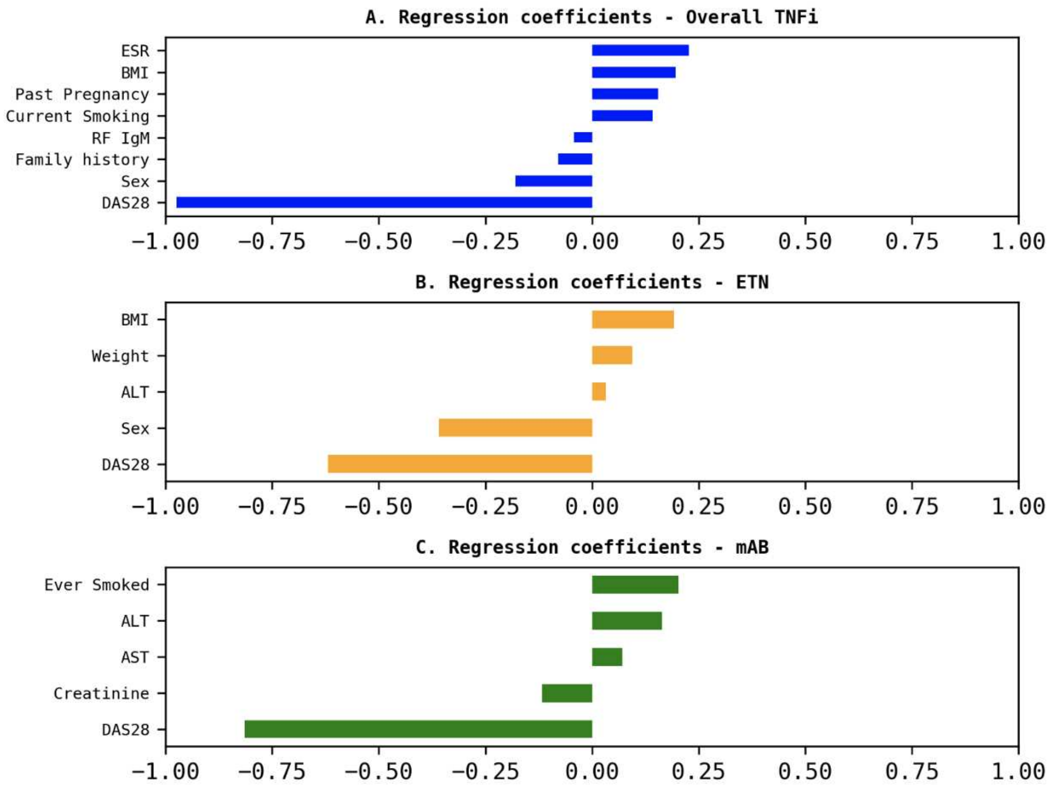

A. SHAP Values - Overall TNFi
DAS28
Lymphocytes
ALT
Neutrophils
Age
Weight
Ever Smoked
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
High
Feature value
Low
B. SHAP Values - ETN
BMI
DAS28
Sex
Ever Smoked
−0.6
−0.4
−0.2
0.0
0.2
SHAP value (impact on model output)
High
Feature value
Low
C. SHAP Values - mAB
DAS28
Weight
ESR
Sex
Ever Smoked
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
High
Feature value
Low

Supplementary Figure 2. Summary plot of the SHAP values of the best models for the prediction of EULAR response to overall TNFi (A) etanercept (B) and monoclonal anti-TNF antibodies (C). The Shapley values are computed on a concatenation of the training and validation sets. Each dot represents a patient's data at treatment initiation. On the x-axis are represented the SHAP value; and on the y-axis, the features are ranked based on their importance (the higher the more important) given by the mean of their absolute Shapley values. The dots are coloured depending on the features' value. Female sex is encoded by 0. Jitter was added to binary variable to facilitate the reading.

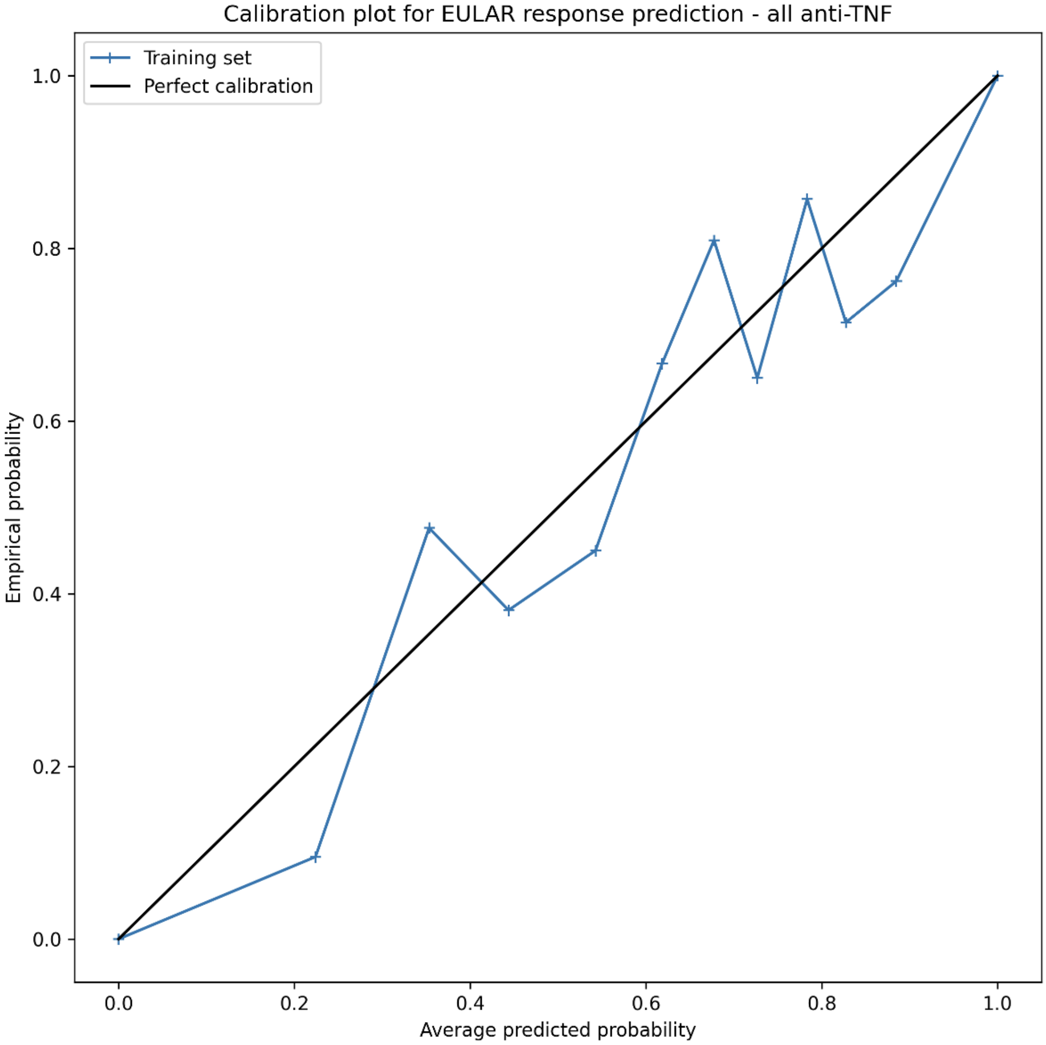
Bouget V, *et al. RMD Open* 2022; 8:e002442. doi: 10.1136/rmdopen-2022-002442



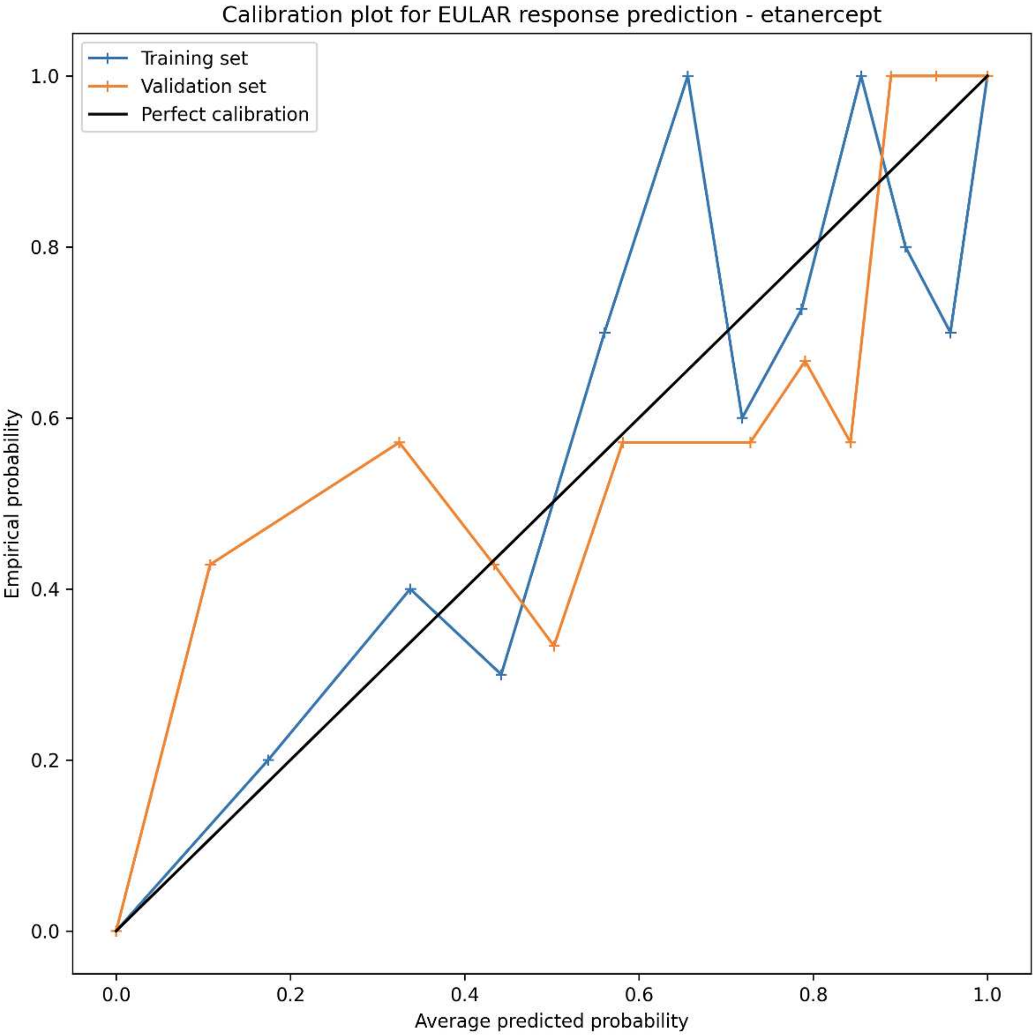
Supplementary Figure 3. Performances of the models predicting the ΔDAS28 on the training set. Cross-validated Mean Squared Error (MSE) of our models for each drug class on training set with the 95% confidence interval. The smallest the MSE, the better the model. Stars legend the p-value ns: 5.00e-02 < p <= 1.00e+00 and ****: p <= 1.00e-04.



Supplementary Figure 4. Coefficients of the ridge regression model for the prediction of the ΔDAS28, with model trained on the training set. Positive (respectively negative) coefficients influence the model towards an increase (resp. decrease) in DAS28. The greater the absolute value of the coefficient, the greater the increase (resp. decrease). Female sex is encoded by 0.

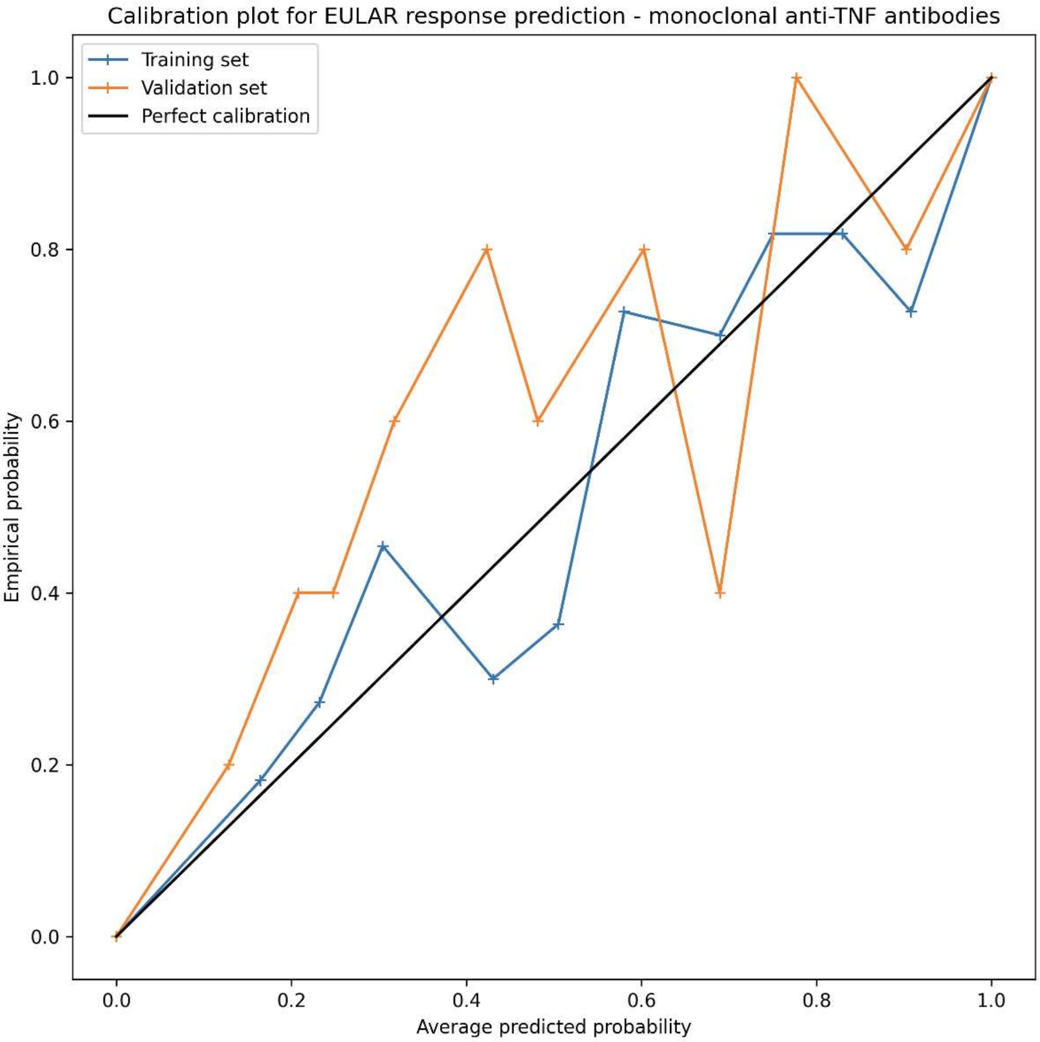


Supplementary Figure 5. Calibration plot for the prediction of the EULAR response for all TNFi computed on the training set. The calibration on the validation set was not evaluated since the model's performances are worse than drug-class-specific models on the training set.



Supplementary Figure 6. Calibration plot for the prediction of the EULAR response for all etanercept computed on both the training and validation sets.

Bouget V, *et al. RMD Open* 2022; 8:e002442. doi: 10.1136/rmdopen-2022-002442



Supplementary Figure 7. Calibration plot for the prediction of the EULAR response for monoclonal antibodies computed on both the training and validation sets.