

Using Machine Learning to Predict Platelet Counts: A Methodology Framework

Enayetur Raheem, Limei Cheng, Hong Yang, Jennifer Sheng

Incyte Research Institute, Wilmington, DE, USA

Background

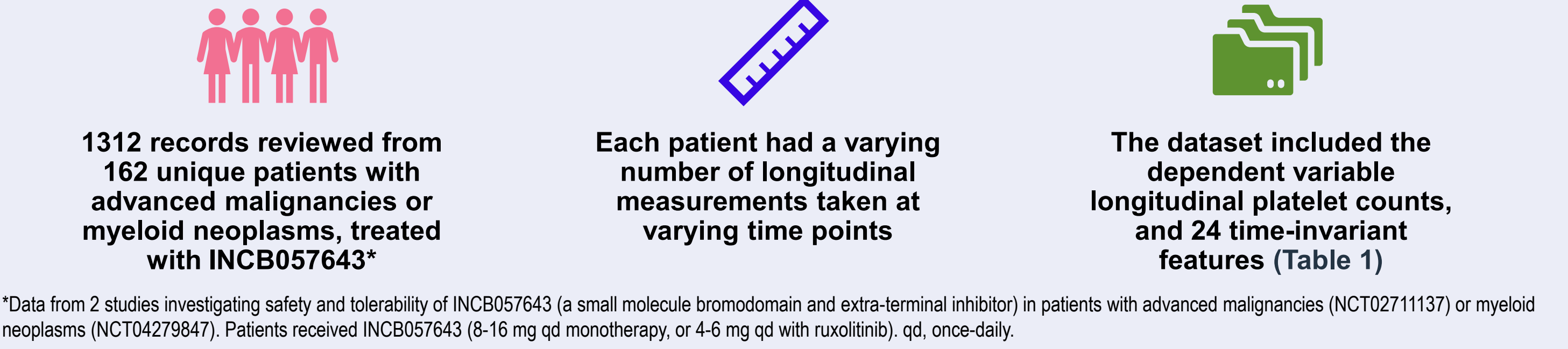
- Conventional exposure-response (ER) models have previously been used for predicting platelet counts during patient treatment and can provide insights into the impact of drug exposure and guide dosing regimens¹⁻³
- However, this approach can be limited by available information and inherent ER modelling assumptions
- Consequently, ER models may not fully reflect platelet signalling pathways and non-linear relationships and/or confounders that exist in the ER dynamic
- An appropriate model should also consider the hierarchical nature of the data
- The advent of machine learning (ML) provides opportunities to advance the potential unknown covariate search and selections⁴⁻⁵ and could fill in some of the gaps in platelet count modelling⁶

Objective

- To explore the use of various ML methods and compare select ML modelling with conventional ER analyses used in platelet count prediction

Methods

Figure 1. Important Remarks About the Data Methods



*Data from 2 studies investigating safety and tolerability of INCB057643 (a small molecule bromodomain and extra-terminal inhibitor) in patients with advanced malignancies (NCT02711137) or myeloid neoplasms (NCT04273647). Patients received INCB057643 (8-16 mg qd monotherapy, or 4-6 mg qd with ruxolitinib), qd, once-daily.

Table 1. Target Outcome and Features/Exposure Variables Investigated

Target/Outcome	Feature/Exposure Variables		
	Pharmacokinetic Variables*	Data Mining Variables	Clinical Variables
Platelet count <75×10 ⁹ /L (grade 2 thrombocytopenia)	Time (since first dose) Cp Cm C _{min} for Cp and Cm, C _{max} for Cp and Cm, C _{avg} for Cp and Cm, Average C _{max} for Cp and Cm	Sex Race Ethnicity Age	BMI Albumin, ALP ALT, AST Bilirubin, creatinine, eGFR (by MDRD equation) Baseline platelet count

*Related to study drug INCB057643.
ALP, alkaline phosphatase; ALT, alanine transaminase; AST, aspartate transferase; BMI, body mass index; C_{avg}, average concentration; Cm, metabolite concentration; C_{max}, maximum observed serum concentration; C_{min}, minimum observed serum concentration; Cp, parent molecule concentration; eGFR, estimated glomerular filtration rate; MDRD, Modification of Diet in Renal Disease.

- 2 ER approaches were applied and compared: (1) ER statistical mixed models and (2) automated ML (AutoML) with feature engineering. A random intercept statistical mixed model was fitted to individual patient platelet counts. A binary target variable was defined as 1 if platelet count was <75×10⁹/L, 0 otherwise; data preparation involved outlier removal and missing value imputation
- For AutoML, 4 tree-based algorithms were applied: (1) gradient boosting machine (GBM), (2) eXtreme Gradient Boosting (XGBoost), (3) CatBoost and (4) LightGBM. Both the conventional statistical and ML models were fitted onto the training data and evaluated on the test (ie, holdout) data. The best ML models were selected based on a comparison of training data and holdout data area under the curve (AUC) values; namely, the area under the receiver operating characteristic curve
- Selected ML models were refined via grid search to optimise the hyperparameters. SHapley Additive exPlanations (SHAP) values and feature importance plots were used to describe the ML model predictions and calculate the contribution of variables to the model
- All analyses were conducted using Python (version 3.9.5) in the Databricks cloud computing environment

Results

Table 2. Statistical Model 5 Determined to Be Most Predictive of Patient Platelet Counts

Metrics	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5
AIC	387	386	388	390	392	391
BIC	450	454	461	467	470	464
AUC: training data	0.888	0.948	0.947	0.946	0.937	0.941
AUC: holdout data	0.806	0.811	0.801	0.800	0.813	0.828

AIC, Akaike information criterion; AUC, area under the curve; BIC, Bayesian information criterion.

- Model 5 was most predictive and fitted onto the full dataset; AUC value on full dataset: 0.857
- Baseline platelet count was the most statistically significant ($P<0.001$) feature, followed by age ($P=0.05$)

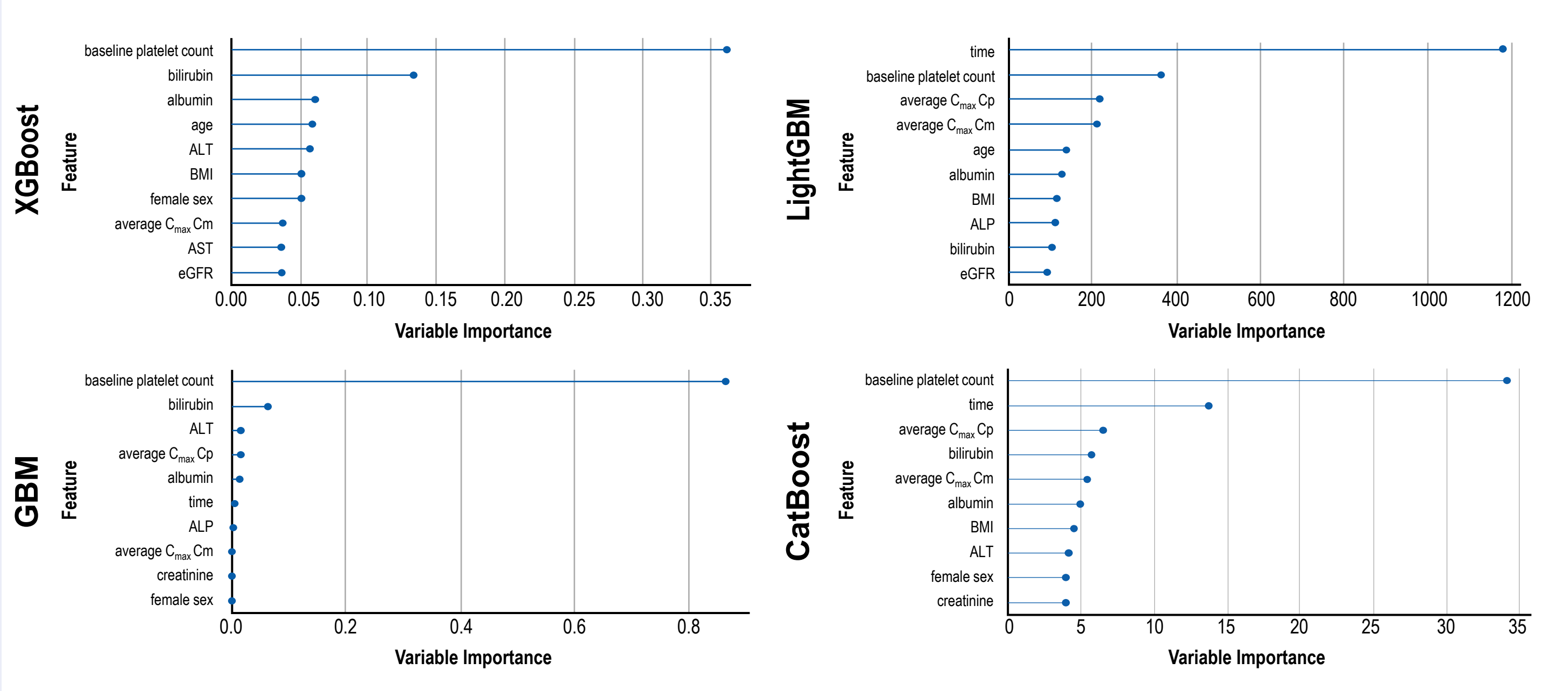
Table 3. XGBoost, LightGBM and CatBoost Were the Most Predictive ML Models

Data	Statistical Mixed Model (Model 5)	GBM	CatBoost	XGBoost	LightGBM
Average cross-validation AUC (training data)		0.783	0.776	0.785	0.769
Predicted AUC (training data)	0.941	0.958	0.994	0.998	0.999
Predicted AUC (holdout data)	0.828	0.850	0.903	0.882	0.904
Predicted AUC (complete dataset)	0.857	0.931	0.976	0.972	0.971

AUC, area under the curve; ML, machine learning.

- AUC values >0.79 indicate a better discriminating performance
- Based on AUC values, XGBoost, LightGBM and CatBoost were the most predictive models and selected for further analysis

Figure 2. Baseline Platelet Count and Age Identified as Important Features Predictive of Patient Platelet Counts



ALP, alkaline phosphatase; ALT, alanine transaminase; AST, aspartate transferase; BMI, body mass index; Cm, metabolite concentration; C_{max}, maximum observed serum concentration; Cp, parent molecule concentration; eGFR, estimated glomerular filtration rate.

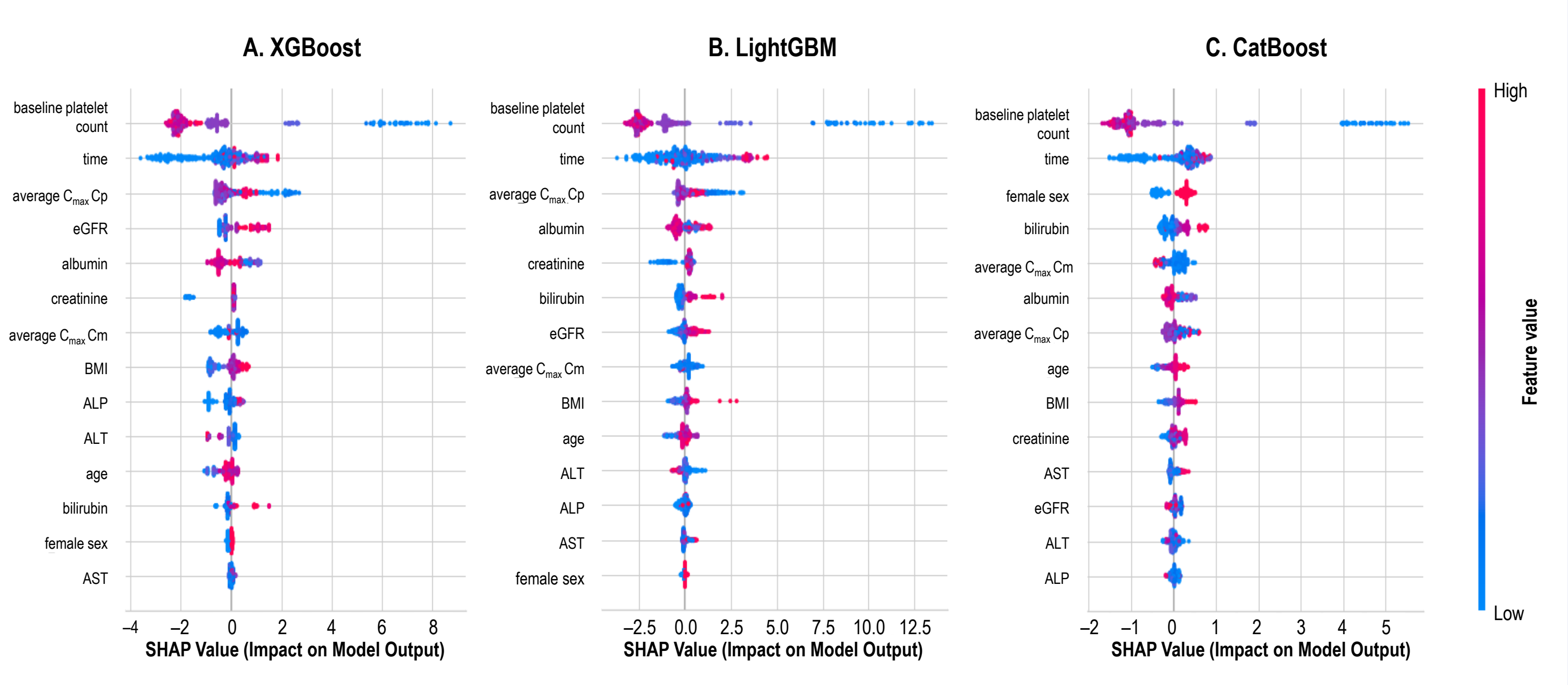
- In line with conventional statistical analysis methods, ML ER models consistently identified baseline platelet count and age as significant covariates or top features, suggesting both are important in relation to patient platelet counts
- XGBoost and LightGBM models, but not the conventional mixed models, repeatedly identified baseline albumin levels, body mass index (BMI), maximum observed serum concentration (C_{max}), bilirubin levels and renal function (derived using the Modification of Diet in Renal Disease [MDRD] equation) as important features that impact patient platelet counts
- Baseline alanine transaminase (ALT), aspartate transferase (AST) and sex were identified by XGBoost modelling method alone

Table 4. ML Model-Derived Hierarchy of Important Features Predictive of Patient Platelet Counts

Statistical Mixed Model ($P<0.1$)	GBM	CatBoost	XGBoost	LightGBM
Baseline platelet count	Baseline platelet count	Baseline platelet count	Baseline platelet count	Time (since first dose)
Age	Bilirubin	Time (since first dose)	Bilirubin	Baseline platelet count
	ALT	Average C _{max} Cp	Albumin	Average C _{max} Cp
	Average C _{max} Cp	Bilirubin	Age	Average C _{max} Cm
	Albumin	Average C _{max} Cm	ALT	Age
	Time (since first dose)	Albumin	BMI	Albumin
	ALP	BMI	Female sex	BMI
	Average C _{max} Cm	ALT	Average C _{max} Cm	ALP
	Creatinine	Female sex	AST	Bilirubin
	Female sex	Creatinine	eGFR	eGFR

ALP, alkaline phosphatase; ALT, alanine transaminase; AST, aspartate transferase; BMI, body mass index; Cm, metabolite concentration, C_{max}, maximum observed serum concentration; C_{min}, minimum observed serum concentration; Cp, parent molecule concentration; eGFR, estimated glomerular filtration rate; ML, machine learning.

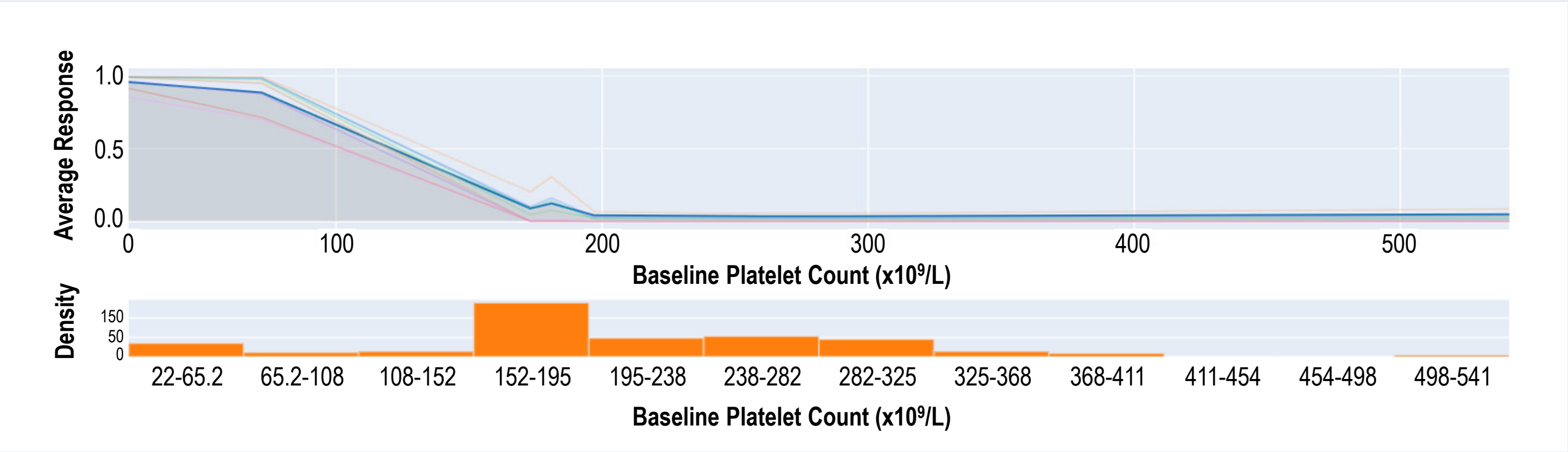
Figure 3. SHAP Analysis Identified Features Predictive of Increased and Reduced Likelihood of Grade 2 Thrombocytopenia



ALP, alkaline phosphatase; ALT, alanine transaminase; AST, aspartate transferase; BMI, body mass index; Cm, metabolite concentration; C_{max}, maximum observed serum concentration; Cp, parent molecule concentration; eGFR, estimated glomerular filtration rate; SHAP, SHapley Additive exPlanations.

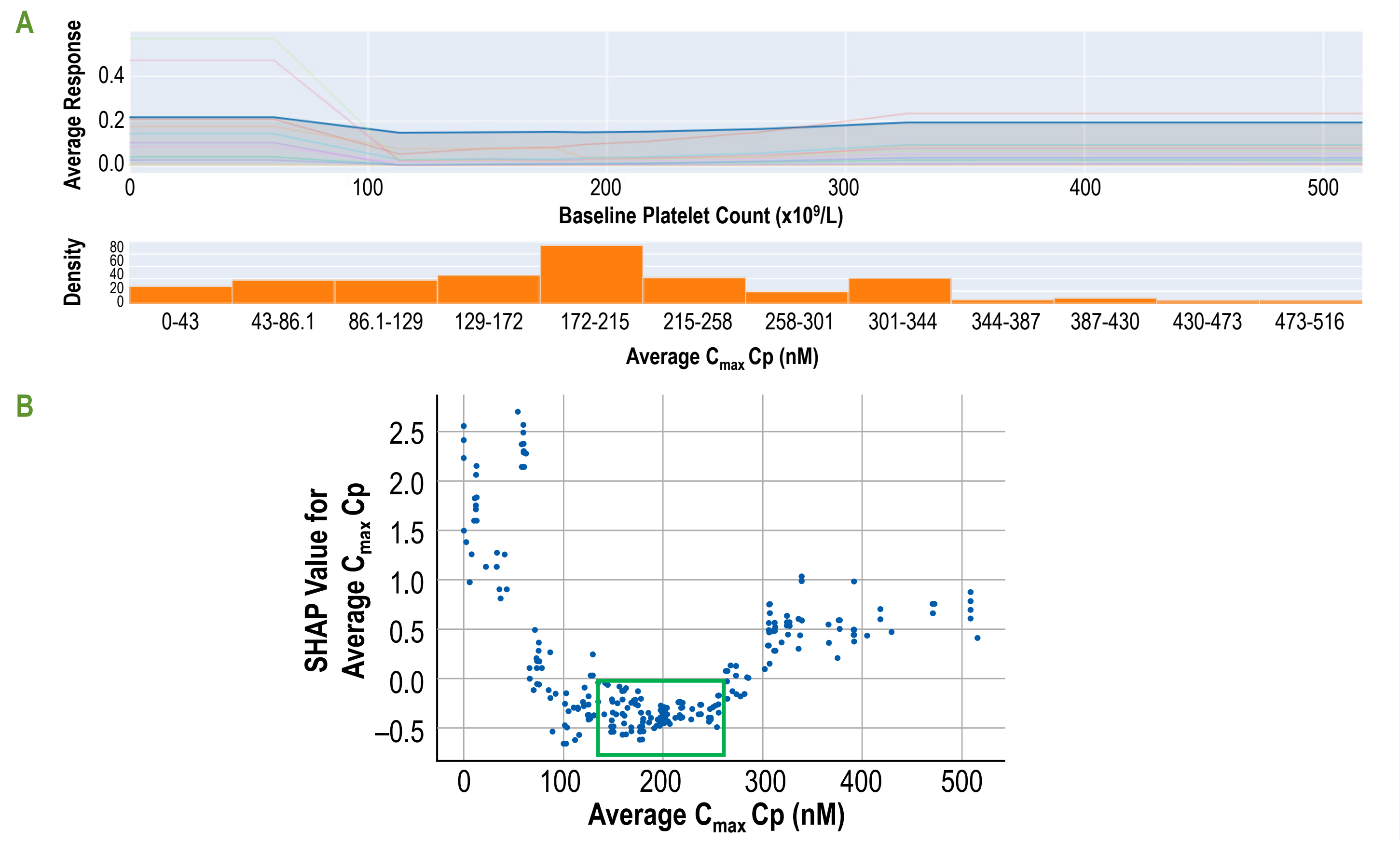
- Lower levels of baseline platelets, average C_{max} of parent molecule (C_{max} Cp) and metabolite (C_{max} Cm), albumin and ALT are predictive of grade 2 thrombocytopenia
- Increased time since first drug dose, and higher levels of renal function (derived using the MDRD equation), higher BMI and higher levels of bilirubin are predictive of grade 2 thrombocytopenia
- Lower levels of creatinine, alkaline phosphatase (ALP) and younger age are predicted to prevent grade 2 thrombocytopenia
- Females are predicted to be more likely to develop grade 2 thrombocytopenia

Figure 4. Partial Dependence Plot of Baseline Platelet Counts Using XGBoost Modelling



- Baseline platelet count >175×10⁹/L is associated with reduced likelihood of grade 2 thrombocytopenia
- Using partial dependence plots to determine baseline platelet counts predictive of grade 2 thrombocytopenia, a clear cutoff was identified around 175×10⁹/L

Figure 5. Partial Dependence Plot (A) and SHAP Value (B) of Average C_{max} Cp Using XGBoost Modelling



C_{max}, maximum observed serum concentration; Cp, parent molecule concentration; SHAP, SHapley Additive exPlanations.

- Partial dependence plot of average C_{max} Cp and correlation plot of SHAP versus average C_{max} Cp, derived using XGBoost modelling, also suggest the nonlinear relationship between average C_{max} Cp and the likelihood of grade 2 thrombocytopenia

Conclusions

- This study has established an efficient ML development framework for predicting platelet counts, including feature engineering, selection of ML models and interpretation of outputs
- Our findings highlight that ML methods can complement conventional ER analysis, providing additional insights to assist clinical data interpretation

Disclosures

Raheem, Cheng, Yang, Sheng: Employment and stock ownership – Incyte Corporation.

Acknowledgments

The study was sponsored by Incyte Corporation (Wilmington, DE, USA). Editorial and graphics support was provided by Envision Pharma Group (Fairfield, CT, USA) and funded by Incyte Corporation.

References

- Blanchet B, et al. *Br J Cancer*. Published online Jan 25, 2024.
- Liu C, et al. *Clin Pharmacol Ther*. 2017;101:657-666.
- van der Bij S, et al. *Cancer Causes Control*. 2013;24:1-12.
- Dou Y, et al. *Comput Biol Med*. 2024;170:108066.
- Feng D, et al. *J Biomol Struct Dyn*. Published online Jan 10, 2024.
- Pozdnyakova O, et al. *J Clin Pathol*. 2023;76:624-631.



To download a copy of this poster, scan code.