Machine Learning Application in a Phase I Clinical Trial Allows for the Identification of Clinical-Biomolecular Markers Significantly Associated With Toxicity

Luca Bedon[1,2], Erika Cecchin[1], Emanuele Fabbiani[3], Michele Dal Bo[1], Angela Buonadonna[4], Maurizio Polano[1,*,†] and Giuseppe Toffoli[1,†]

Machine learning (ML) algorithms have been used to forecast clinical outcomes or drug adverse effects by analyzing different data sets such as electronic health records, diagnostic data, and molecular data. However, ML implementation in phase I clinical trial is still an unexplored strategy that implies challenges such as the selection of the best development strategy when dealing with limited sample size. In the attempt to better define prechemotherapy baseline clinical and biomolecular predictors of drug toxicity, we trained and compared five ML algorithms starting from clinical, blood biochemistry, and genotype data derived from a previous phase Ib study aimed to define the maximum tolerated dose of irinotecan (FOLFIRI (folinic acid, fluorouracil, and irinotecan) plus bevacizumab regimen) in patients with metastatic colorectal cancer. During cross-validation the Random Forest algorithm achieved the best performance with a mean Matthews correlation coefficient of 0.549 and a mean accuracy of 80.4%; the best predictors of dose-limiting toxicity at baseline were hemoglobin, serum glutamic oxaloacetic transaminase (SGOT), and albumin. The feasibility of a prediction model prototype was in principle assessed using the two distinct dose escalation cohorts, where in the validation cohort the model scored a Matthews correlation coefficient of 0.59 and an accuracy of 82.0%. Moreover, we found a strong relationship between SGOT and irinotecan pharmacokinetics, suggesting its role as surrogates' estimators of the irinotecan metabolism equilibrium. In conclusion, the potential application of ML techniques to phase I study could provide clinicians with early prediction tools useful both to ameliorate the management of clinical trials and to make more adequate treatment decisions.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Machine learning (ML) models have been employed to predict efficacy end points and drug toxicities in large phase II-III trials; however, their implementation in the early phase I clinical trial development is still an unexplored strategy.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ We investigated the application of ML in a dose escalation phase I clinical trial to test its feasibility and to highlight clinical and biomolecular predictors of irinotecan toxicity.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ ML allowed us to identify and estimate the importance of variables for predicting the dose-limiting toxicity during an irinotecan-based chemotherapy regimen despite the small number of patients enrolled. Moreover, we evidenced variables that could represent surrogates' estimators of the irinotecan metabolism equilibrium.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ The potential application of ML to phase I study enables the discovery and validation, with a reasonable degree of accuracy, of factors predicting an outcome of interest. Prediction models allow clinicians to anticipate toxicities and make decisions accordingly both to treat such toxicities and to adjust the dosing for subsequent cycles.

[1]Experimental and Clinical Pharmacology Unit, Centro di Riferimento Oncologico di Aviano, Istituto di Ricovero e Cura a Carattere Scientifico, Aviano, Italy; [2]Department of Chemical and Pharmaceutical Sciences, University of Trieste, Trieste, Italy; [3]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy; [4]Medical Oncology Unit, Centro di Riferimento Oncologico di Aviano, Istituto di Ricovero e Cura a Carattere Scientifico, Aviano, Italy. *Correspondence: Maurizio Polano (mpolano@cro.it)

[†]Equally contributed as last authors.

Dose escalation phase I trials are designed to identify the maximum tolerated dose (MTD) of a new drug or a new drug combination. The core principle of dose escalation is keeping a relatively rapid dose increase to avoid treatment of patients at subtherapeutic doses while preserving safety by limiting the frequency of toxic events (dose-limiting toxicities or DLTs). The most common dose escalation phase I strategy is based on 3 + 3 design and its variation.[1] More recent approaches are represented by the continual reassessment method and its variation.[2,3] However, regardless of dose escalation design adopted, there still are patients that experience DLTs, even in those treated at the MTD.[2,4]

Machine learning (ML) application to analyze electronic health records, diagnostic, and molecular data with the aim of reshaping key steps of clinical trial design toward increasing trial success rates is now raising interest.[5–8] Despite ML models having been employed to predict efficacy end points and drug toxicity using data sets of large phase II-III trials,[9–11] their application has not yet been considered for phase I clinical trials.

We previously conducted a phase Ib clinical trial in metastatic colorectal cancer (mCRC) guided by the UGT1A1*28 genotype to determine the MTD of irinotecan within the FOLFIRI (irinotecan plus infusional 5-fluorouracil/leucovorin) plus bevacizumab regimen.[12] However, during the study several patients experienced DLTs despite the administered dose being the safe MTD or even lower. Since irinotecan exhibits a considerable interindividual pharmacokinetics variability,[13] personalization of this treatment is necessary to guarantee a toxicity-free optimal pharmacotherapy.

In this study we apply an ML approach to highlight known and new predictors of DLT by simultaneously analyzing clinical, baseline blood biochemistry (i.e., before starting the phase I), and genetic data derived from our previous work.[12] The pipeline we used includes a step selecting the best predictors based on importance rankings; the optimal subset was then used to train models. We compared the performance of five ML classification algorithms to select the most suitable classifier.

METHODS

Patients, treatment, and toxicity assessments

This study is based on the retrospective analysis of 45 patients with mCRC enrolled in a phase I clinical trial intended to determine the MTD of irinotecan during the first cycle in *UGT1A1* *1/*1 and *1/*28 patients treated with the FOLFIRI plus bevacizumab regimen.[12]

The analyzed patients were treated at two different institutions (University of Chicago, Chicago, IL; and Centro di Riferimento Oncologico, Aviano, Italy), where the study was conducted upon the approval of each Institutional Review Board and the collection of signed consent from the participants. The signed informed consent from patients included in the study was collected at both Institutions. The ClinicalTrials.gov identifier is NCT01183494.

The study eligibility criteria were the following: confirmed diagnosis of mCRC, age ≥ 18 years, *UGT1A1* *1/*1 and *1/*28 genotypes, absolute neutrophil count ≥ 1,500/mL, platelets ≥ 100,000/mL, Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1, creatinine clearance < 1.5 times the upper limit of normal (ULN), serum glutamic oxaloacetic transaminase (SGOT) and serum glutamic-pyruvic transaminase (SGTP) < 2.5 times the ULN (< 5 times the ULN in the presence of liver metastases), and total serum bilirubin < 1.6 mg/dL.

The study followed the dose escalation of irinotecan under a 3 + 3 design: Patients were enrolled at each irinotecan dose level (260, 310, and 370 mg/m^2) in each genotype cohort and no intrapatient dose escalation was allowed. The irinotecan was given within the FOLFIRI (28 days each cycle) regimen that consisted in designated doses of irinotecan (90 min intravenous infusion on days 1 and 15), 200 mg/m^2 of leucovorin (administered concomitantly with irinotecan), and 2,800 mg/m^2 of 5-fluorouracil (400 mg/m^2 bolus + 2,400 mg/m^2 over a 46-hour intravenous infusion on days 1 and 15). Bevacizumab was given at 5 mg/kg over a 15-minute to 30-minute intravenous infusion on days 3 and 15.

Toxicity was classified and graded according to the National Cancer Institute's (NCI's) Common Terminology Criteria for Adverse Events (version 3.0). DLT was defined as recorded hematologic toxicity of grade ≥ 4 or nonhematologic toxicity of grade ≥ 3 during the first cycle.

Further details of the study are reported in the previous study.[12] Patient characteristics are shown in **Table 1**.

Feature definition and data organization

The predictive value of 37 features was assessed in 45 patients with mCRC treated with FOLFIRI plus bevacizumab regimen. The features collected for this study are represented in **Figure 1**.

These features were divided in three categories:

- *Clinical* patient data: age, sex, body surface area, and ethnicity.
- *Blood baseline* laboratory analysis performed prior to the beginning of chemotherapy: hemoglobin, hematocrit, erythrocytes, total leucocytes, neutrophils, lymphocytes, platelets, SGOT (also known as AST for aspartate transaminase), SGTP (also known as ALT for alanine transaminase), total protein, albumin, alkaline phosphatase, bilirubin, glucose, sodium, potassium, calcium, and creatinine.
- *Genotype* of genes that affect irinotecan and fluorouracil disposition: 14 polymorphisms in *UGT1* genes family, *CYP3A* family, and *DPYD* gene were included (**Table S1**).

The output was defined as patients' DLT status recorded at the end of first cycle; patients were classified in DLT = *Yes* and DLT = *No*. *Pharmacokinetic* parameters of irinotecan and its metabolites during first cycle (irinotecan, SN-38, SN-38-G, APC) were also collected for further analysis. To perform the following steps, all categorical variables were converted to dummies.

Machine learning approach

We used clinical, blood baseline, and genotype features of 45 patients with mCRC to predict the DLT event after the first cycle of FOLFIRI plus bevacizumab treatment.

Given the limited size of the cohort, and to avoid model overfitting, a fivefold cross-validation (CV) repeated five times coupled with an encapsulated Recursive Feature Elimination process (RFE), were employed to develop and tune the models.[14] The detailed workflow is depicted in **Figure 2**.

The entire data set was randomly split in five folds and then, at each iteration of the outer CV loop, one fold was taken as test set and all the remaining folds were used as training set. The algorithm firstly fits the models to all features using the training set and calculates the model performances using the holdout test set, then the features are ranked by importance using the training set. We defined subsets of i features $S = \{S_2, ..., S_i\}$, which are a set of values that define the number of most important variables to keep; at each iteration of the inner feature selection loop, the S_i top ranked are used to refit and assess the model performances. The performances over the S_i subsets for each CV iteration are finally gathered to determine the appropriate number of S_i features to keep in the final models. The variable importance within the RFE process was computed

Table 1 Characteristics of patients involved in the study

DLT	No	Yes
Observations	31	14
Center		
Aviano, Italy	22 (71%)	6 (43%)
Chicago, IL	9 (29%)	8 (57%)
Age		
Mean (SD)	55 (9.3)	56 (14)
Sex		
F	12 (39%)	7 (50%)
M	19 (61%)	7 (50%)
Ethnicity		
Asian	0 (0%)	1 (7%)
Black	4 (13%)	2 (14%)
White	27 (87%)	11 (79%)
BSA (m^2)		
Mean (SD)	1.8 (0.27)	1.9 (0.30)
ECOG		
0	28 (90%)	12 (86%)
1	3 (10%)	2 (14%)
Primary site		
Colon	22 (71%)	12 (86%)
Rectum	9 (29%)	2 (14%)
UGT1A1		
*1/*1	16 (52%)	6 (43%)
*1/*28	15 (48%)	8 (57%)
Dose irinotecan (mg/m^2)		
260	15 (48%)	3 (21%)
310	13 (42%)	7 (50%)
370	3 (10%)	4 (29%)

Clinical demographic characteristics of patients that did not experience (No) and that did experience (Yes) DLT during the first cycle of treatment. Description of clinical characteristics of the cohort with their respective categorization and percentages.
BSA, body surface area; ECOG, Eastern Cooperative Oncology Group; SD, standard deviation.

using Boruta[15] as *z*-scaled mean decreased accuracy (MDA). The "caret" R package[16] was used to implement and tune the classifiers within the CV and the RFE processes.

We implemented and compared five machine learning classification algorithms, namely Random Forest (RF),[17] Generalized Linear Models (GLM),[18] eXtreme Gradient Boosting (XGB),[19] Support Vector Machine (SVM),[20] and K-Nearest Neighbors (KNN).[21]

Due to the imbalance of the output DLT classes, to avoid overoptimistic results the models' performances were assessed in term of Matthews Correlation Coefficient[22] (MCC) in addition to accuracy (ACC) and area under the curve (AUC) of the receiver operating characteristic (ROC). The metrics were computed as averages across fivefold CV repeated five times with 95% studentized CV confidence intervals (CIs). The importance of the predictive features was evaluated by the mean importance of features obtained from Boruta algorithm over all of the CV cycles.

The same workflow described in **Figure 2** was used to train RF models intended to predict a DLT event using the *1/*28 study cohort as training set (model development and evaluation); the *1/*1 study cohort was kept apart as validation set to evaluate the ability of the models to predict a DLT event on unseen observations (**Figure S1**). The evaluation metrics were computed across a threefold CV repeated five times using the training set, whereas the validation metrics were computed by comparing the predicted and the real DLT class of the validation set.

Statistical analysis

All statistical analyses were performed in R environment (v3.6.3).[23] Statistical independence or association between categorical variables was tested using the Pearson's χ^2 test. Normality was inspected using the Shapiro-Wilk's test. Comparisons of the means of independent groups were performed using the unpaired two samples Student's *t*-test (normally distributed data) and unpaired two samples Wilcoxon–Mann–Whitney test (not normally distributed data). Box plots coupled with Student's *t*-test results were used to graphically support the differences found in feature values between DLT and non-DLT patients. Correlation was assessed using Pearson's correlation test, and correlation strength description was based on the correlation coefficient value.[24] The threshold for statistical significance was set at $P < 0.05$.

RESULTS

Dose escalation results and toxicities

We started from the data reported in a previous study of ours,[12] in which the 310 mg/m^2 irinotecan dose was declared the MTD in *1/*1 patients and the 260 mg/m^2 irinotecan dose was declared the MTD in *1/*28. Fourteen DLTs were recorded during the study (**Table S2**), of which four DLTs (4/7; 57%) were observed at 370 mg/m^2, seven DLTs (7/20; 35%) were observed at 310 mg/m^2, and three DLTs (3/18; 17%) were observed at 260 mg/m^2. In the *1/*28 study cohort, eight patients experienced DLTs (8/23; 35%), whereas in the *1/*1 study cohort the DLTs recorded were six (6/22; 27%). No dependency was found between *UGT1A1* genotype and irinotecan dosage (mg/m^2) (Pearson's χ^2 test, $P = 0.8424$), as well as between *UGT1A1* genotype and DLT (Pearson's χ^2 test, $P = 0.8244$). A slightly more significant dependency, but still not reaching the level of significance, was found between irinotecan dosage and DLT (Pearson's χ^2 test, $P = 0.1283$).

ML implementation predicts DLT events and highlights DLT-related variables

The results of each ML algorithm to predict if a patient is going to manifest a DLT are reported in **Table 2**. In this approach the whole patients' cohort of 45 cases was used during the model development and hence the evaluation metrics were computed through a fivefold CV repeated 5 times (**Figure 2**). The RF algorithm showed superior performance compared with the other models in terms of both MCC score (MCC = 0.549, CI = 0.429–0.670) and accuracy (ACC = 0.804, CI = 0.755–0.853). Linear SVM (AUC = 0.859) and GLM (AUC = 0.48) showed higher AUC values than the RF model; however, MCC was chosen as a reference metric due to its robustness in producing an informative and truthful score in evaluating binary classifications.[22] The best subsets of predictors S_i ranged from two to four variables (**Table 2**), and the best RF model reached the highest CV performance with three variables, specifically hemoglobin, SGOT, and albumin.

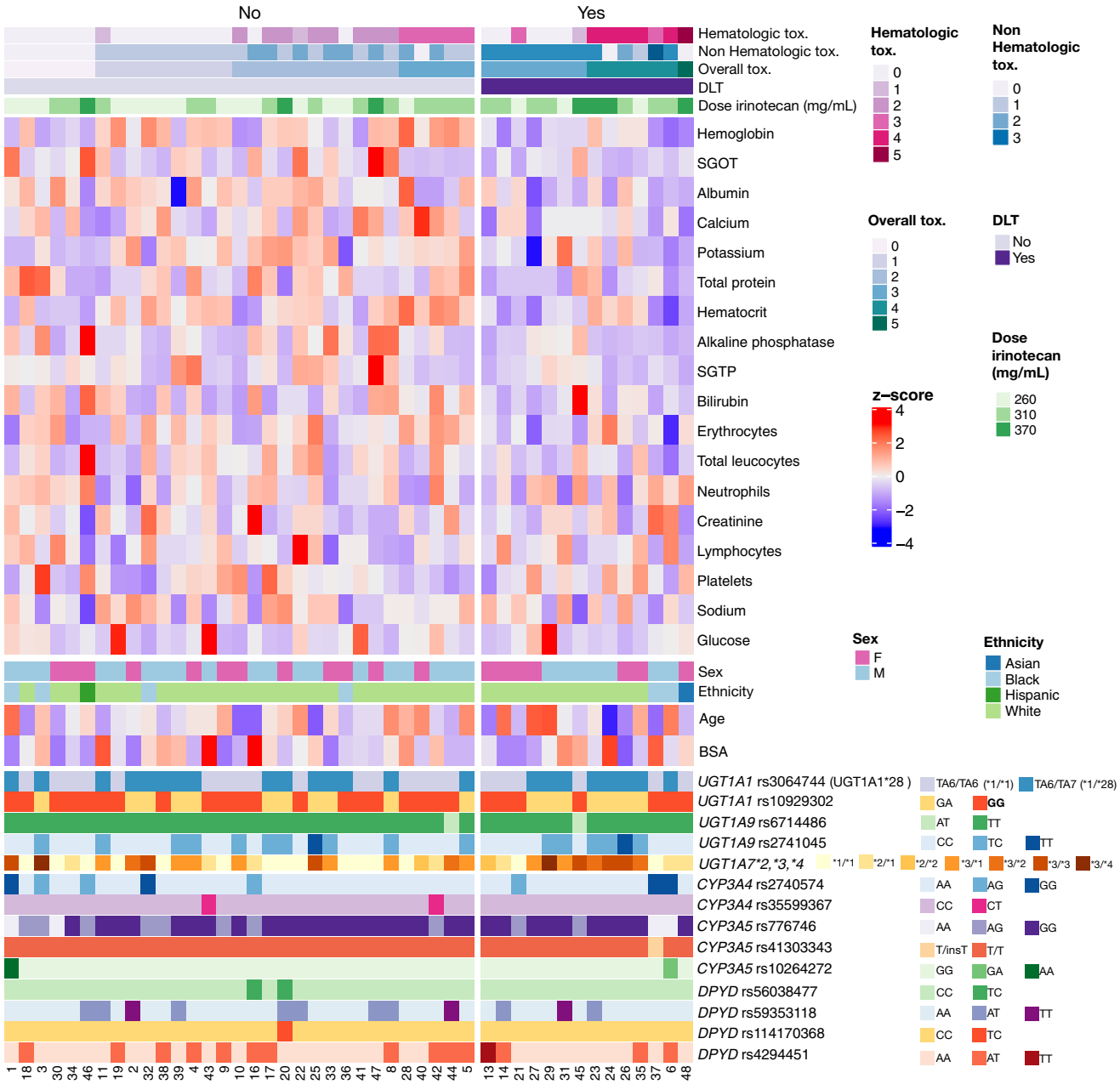

Figure 1 Heat map representation of the data used to develop the models. Patients in columns are ordered in accordance with the overall toxicity grade (increasing grade from left to right). The variables used are represented in rows. Continuous variables have been standardized and represented as *z* scores (color scale from blue to red). Discrete and categorical values are represented as colored annotations with their own legend. BSA, body surface area; DLT, dose-limiting toxicity; SGOT, serum glutamic oxaloacetic transaminase; SGTP, serum glutamic-pyruvic transaminase; tox., toxicity.

The features' importance to predict the DLT was calculated using Boruta within the RFE process across all CV training folds (**Figure 2**). The most important features that scored a CV averaged *z*-scaled MDA higher than 1 are represented in **Figure 3a**. Hemoglobin was the most important feature (MDA = 6.83, N Folds = 322), followed by SGOT (MDA = 4.14, N Folds = 240), and albumin (MDA = 3.57, N Folds = 209).

Our approach selected mainly blood baseline laboratory variables as best predictors of DLT (**Figure 3a**), such as blood count variables (hemoglobin, hematocrit, erythrocytes), liver enzymes and blood proteins (SGOT, SGTP, bilirubin albumin, total protein), and basic metabolic electrolytes (calcium, potassium). To assess if these variables indeed were important to predict DLT events, we evaluated RF models with different variables categories (**Table S3**). In this cohort Clinical and Genotype data seem to not contain relevant predictive information as opposed to blood baseline variables. In fact, when the latter were excluded the MCC dropped to 0.068. In this case adding more biology data (such as

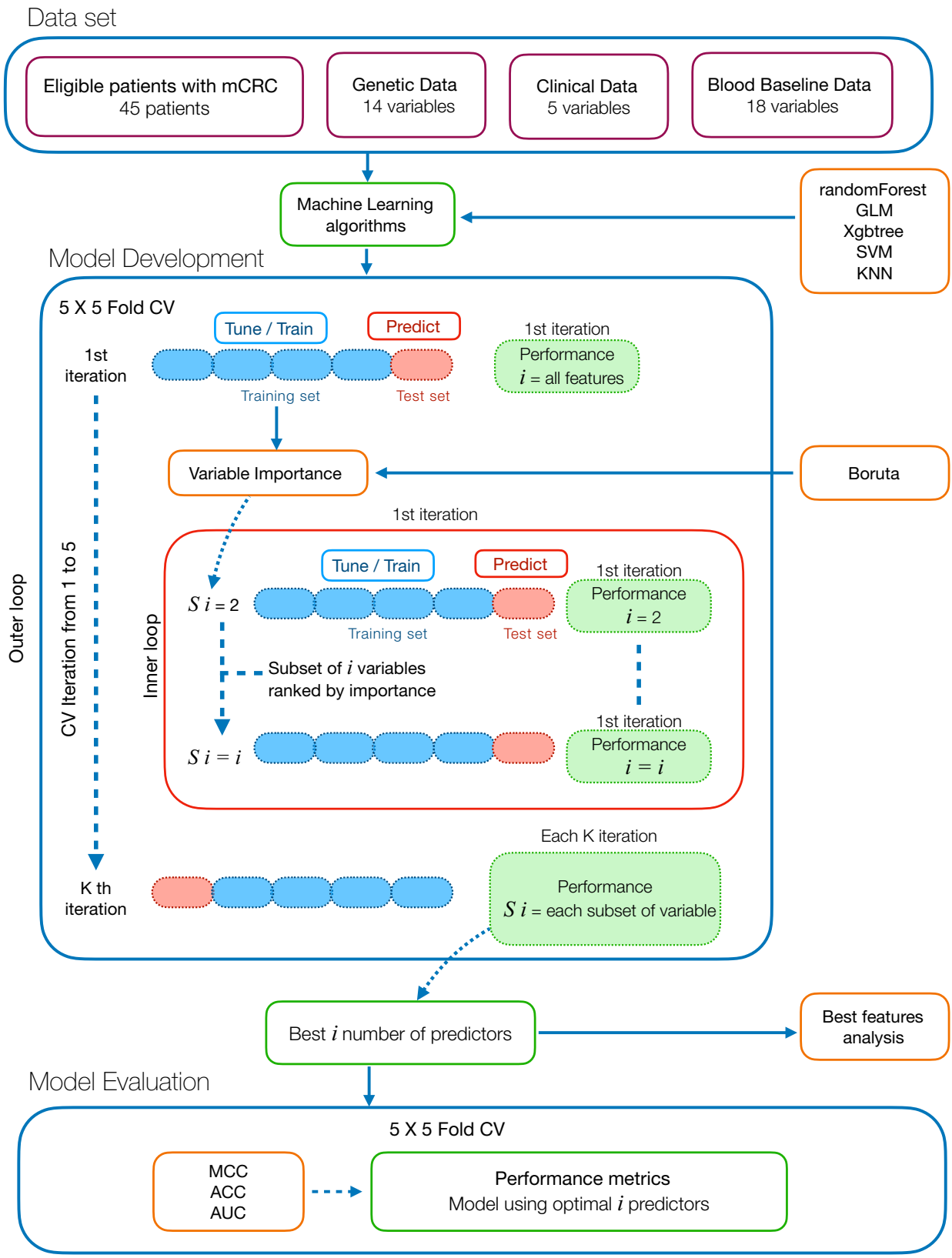

Figure 2 Development of the DLT prediction model using the whole patient cohort. ACC, accuracy; AUC, area under the curve of the receiver operating characteristic; CV, cross-validation; GLM, Generalized Linear Model; *i*, integer representing the number of variables considered; KNN, K-Nearest Neighbors; MCC, Matthews Correlation Coefficient; mCRC, metastatic colorectal cancer; S, subset of variables; SVM, Support Vector Machine; Xgbtree, eXtreme Gradient Boosting tree.

Table 2 ML algorithms' performance computed within the fivefold CV repeated five times

Algorithm	Best S_i	MCC (CI)	ACC (CI)	AUC (CI)
RF	Hemoglobin SGOT Albumin	0.549 (0.429–0.670)	0.804 (0.755–0.853)	0.822 (0.745–0.898)
GLM	Hemoglobin SGOT Albumin Calcium	0.461 (0.345–0.577)	0.772 (0.728–0.815)	0.848 (0.797–0.899)
XGB	Hemoglobin SGOT	0.401 (0.264–0.539)	0.769 (0.727–0.812)	0.819 (0.758–0.880)
SVM	Hemoglobin SGOT Albumin Calcium	0.446 (0.341–0.551)	0.773 (0.736–0.810)	0.859 (0.809–0.908)
KNN	Hemoglobin SGOT	0.315 (0.182–0.448)	0.723 (0.677–0.769)	0.635 (0.542–0.728)

ML algorithms' performance computed within the fivefold CV repeated five times (mean with confidence intervals) and relative best subset of variables. ACC, accuracy; AUC, area under the curve of the ROC; CI, 95% studentized bootstrap confidence interval; CV, cross-validation; GLM, Generalized Linear Model; KNN, K-Nearest Neighbors; MCC, Matthews Correlation Coefficient; ML, machine learning; RF, Random Forest; ROC, receiver operating characteristic; SGOT, serum glutamic oxaloacetic transaminase; S_i, subset of variables; SVM, Support Vector Machine; XGB, eXtreme Gradient Boosting.

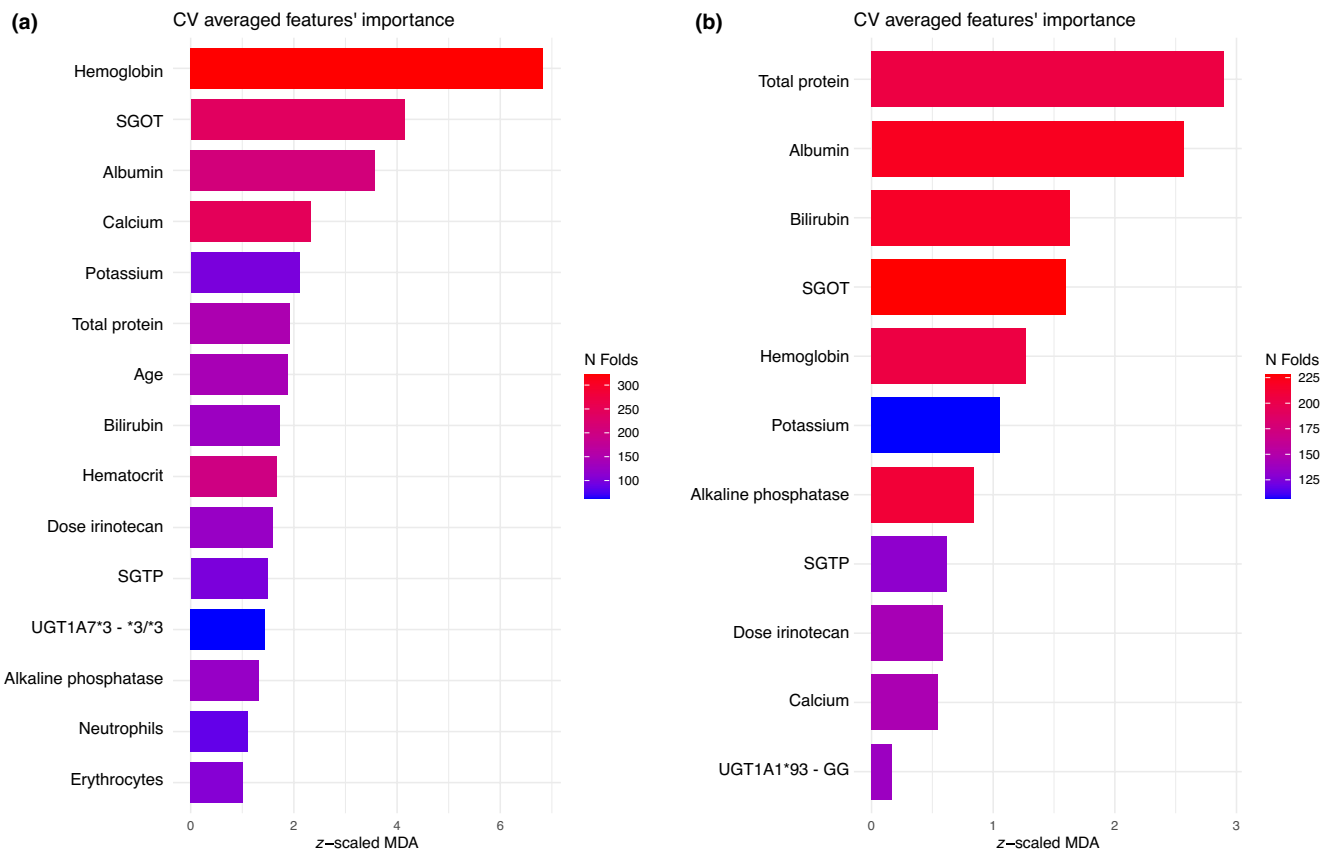

Figure 3 Variable importance computed using Boruta within the RFE (Recursive Feature Elimination) process and across all CV training folds. Bars' length represents the importance reported as scaled mean decreased accuracy (MDA). The features are ranked by importance from the most important (top) to the least important (bottom) and only features that scored a *z*-scaled MDA higher than 1 are represented. Color scale represents the number of iterations within the model development in which a variable was kept as most important variables. (**a**) Variables' importance computed during CV using the whole patient cohort. (**b**) Variables' importance computed during CV using the *1/*28 cohort. CV, cross-validation; SGOT, serum glutamic oxaloacetic transaminase; SGTP, serum glutamic-pyruvic transaminase.

genotype) did not improve the performance probably due the concurrence of small sample size and low-frequency genetic variants.

We also compared different CV techniques during the model evaluation step depicted in **Figure 2** (**Table S4**). The MCC score of RF algorithm was comparable between 5-fold CV repeated five times (MCC = 0.549, CI = 0.429–0.670), 10-fold CV (MCC = 0.597, CI = 0.300–0.894), and leave one out cross validation (LOOCV) (MCC = 0.564). Considering bias-variance trade-off we chose fivefold CV repeated five times, since with $k = 5$ test error rate, estimates do not suffer either from high bias or from high variance.[25]

Principle validation using two distinct dose escalation cohorts

As the two *UGT1A1* cohorts followed two distinct dose escalation studies, we attempted to validate the principle of implementing ML models to predict DLT events by using the *1/*28 cohort (23 out of 45 cases) to develop and train the models and the *1/*1 cohort (22 out of 45 cases) to assess the prediction ability on a new upcoming dose escalation study (**Figure S1**).

During CV using the *1/*28 cohort, the best RF model achieved a mean MCC of 0.403 (CI = 0.221–0.586), a mean ACC of 0.748 (CI = 0.679–0.816), and a mean AUC of 0.744 (CI = 0.633–0.856). The best subset of variables S_i was composed of eight features: hemoglobin, SGOT, SGTP, total protein, albumin, alkaline phosphatase, bilirubin, and potassium (**Figure 3b**). Total protein was the most important feature (MDA = 2.90, N Folds = 204), followed by albumin (MDA = 2.56, N Folds = 220), and bilirubin (MDA = 1.63, N Folds = 217).

On the *1/*1 validation cohort, the model scored an MCC of 0.59, an ACC of 0.82, and an AUC of 0.81. As reported in **Table S5**, the model correctly classified 18 out of 22 patients, of which 5 out of 6 were DLT patients (DLT = Yes) and 13 out of 16 were no DLT patients (DLT = No). Notably, the Patients 10, 40, and 41 that were incorrectly classified as DLT patients with a DLT probability of 0.61, 0.74, and 0.61, respectively, consistently experienced a moderate overall toxicity after the first cycle of treatment (grade 2, 3, and 2, respectively). Conversely, Patient 45 was incorrectly classified as no DLT patient despite the onset of a grade 3 overall toxicity.

Statistical analysis of predictive DLT variables

The Boruta built-in variable importance measure allowed us to rank the features with respect to their relevance for DLT prediction. Nevertheless, they only represent the strength of this dependency. To capture the distribution pattern of the models' relevant variables with respect to the DLT event, we performed additional statistical analysis.

In **Figure 4a** variables' standardized values are depicted compared with the DLT status; DLT patients are mainly characterized by negative z scores, which means that raw variables' values are below the mean cohort average, especially for patients that suffered the most severe toxicities (rightmost columns).

DLT patients had significantly lower baseline hemoglobin (Student's t-test, $P = 0.004$), lower baseline albumin (Student's t-test, $P = 0.025$), and lower baseline total proteins (Student's t-test, $P = 0.025$) (**Figure 4b**). The hepatic parameters SGOT, SGTP, and alkaline phosphatase were also significantly lower in DLT patients (Student's t-test, $P = 0.028$; 0.030; 0.016, respectively) (**Figure 4b**).

Relationships between DLT predictors and pharmacokinetic parameters

We investigated the relationship between the models' relevant variables and the pharmacokinetic parameters of irinotecan and its metabolites by conducting Pearson correlation analysis (**Figure 5a**). We found mostly negligible correlations between potassium, total proteins, and any of the pharmacokinetic parameters ($0.0 < |r| < 0.2$). Albumin and hemoglobin showed a similar correlation pattern; both were weakly negatively correlated ($-0.2 \geq r > -0.4$) with the area under the concentration time curve extrapolated to infinity (AUC_{inf}) of irinotecan and weakly positively correlated ($0.2 < r < 0.4$) with irinotecan clearance and distribution, suggesting that lower values of albumin and hemoglobin are associated with higher irinotecan exposure.

A series of hepatic parameters (SGOT, SGTP, alkaline phosphatase, and bilirubin) were moderately positively correlated ($0.4 \leq r < 0.7$) with AUC_{inf} values of irinotecan and weakly negatively correlated ($-0.2 \geq r > -0.4$) with irinotecan clearance and distribution, this time suggesting that higher values of hepatic parameters are associated with a higher irinotecan exposure. These hepatic parameters were also moderately and strongly positively correlated ($0.4 \leq r < 0.9$) with AUC_{inf} values of SN-38G and APC, respectively. This behavior was also reflected by the AUC_{inf} ratios, representing the relative amount of irinotecan converted into its metabolites (**Figure 5a**). Indeed, hepatic parameters' values were positively correlated with APC/irinotecan and SN-38G/SN-38 AUC_{inf} ratios and negatively correlated with SN-38/irinotecan AUC_{inf} ratio, suggesting that patients characterized by higher hepatic parameters are more prone to convert irinotecan toward SN-38G and APC rather than production of the cytotoxic SN-38 metabolite (**Figure 5b**). This pattern might explain the reason underlying the unexpected distribution of hepatic parameters that were found higher in patients that did not experience DLTs (**Figure 4b**). In fact, DLT patients were characterized by a significantly lower APC/irinotecan AUC_{inf} ratio (Wilcoxon–Mann–Whitney test, $P = 0.019$) and also a significantly lower APC/SN38, AUC_{inf} ratio (Wilcoxon–Mann–Whitney test, $P = 0.024$) (**Figure 5c**).

DISCUSSION

We collected clinical, blood baseline, and genotype variables of 45 patients with mCRC entered in a previous phase Ib dose escalation study[12] that were used to train and compare the performance of five ML classification algorithms and to identify the best DLT predictors. During CV the RF model achieved the best performance with a mean MCC of 0.549 (CI = 0.429–0.670) and a mean ACC of 0.804 (CI = 0.755–0.853); the best subset of variables S_i was composed of three features, specifically hemoglobin, SGOT, and albumin.

The main criticality of ML application in phase I clinical trials derives from an intrinsic characteristic of these studies, that is they

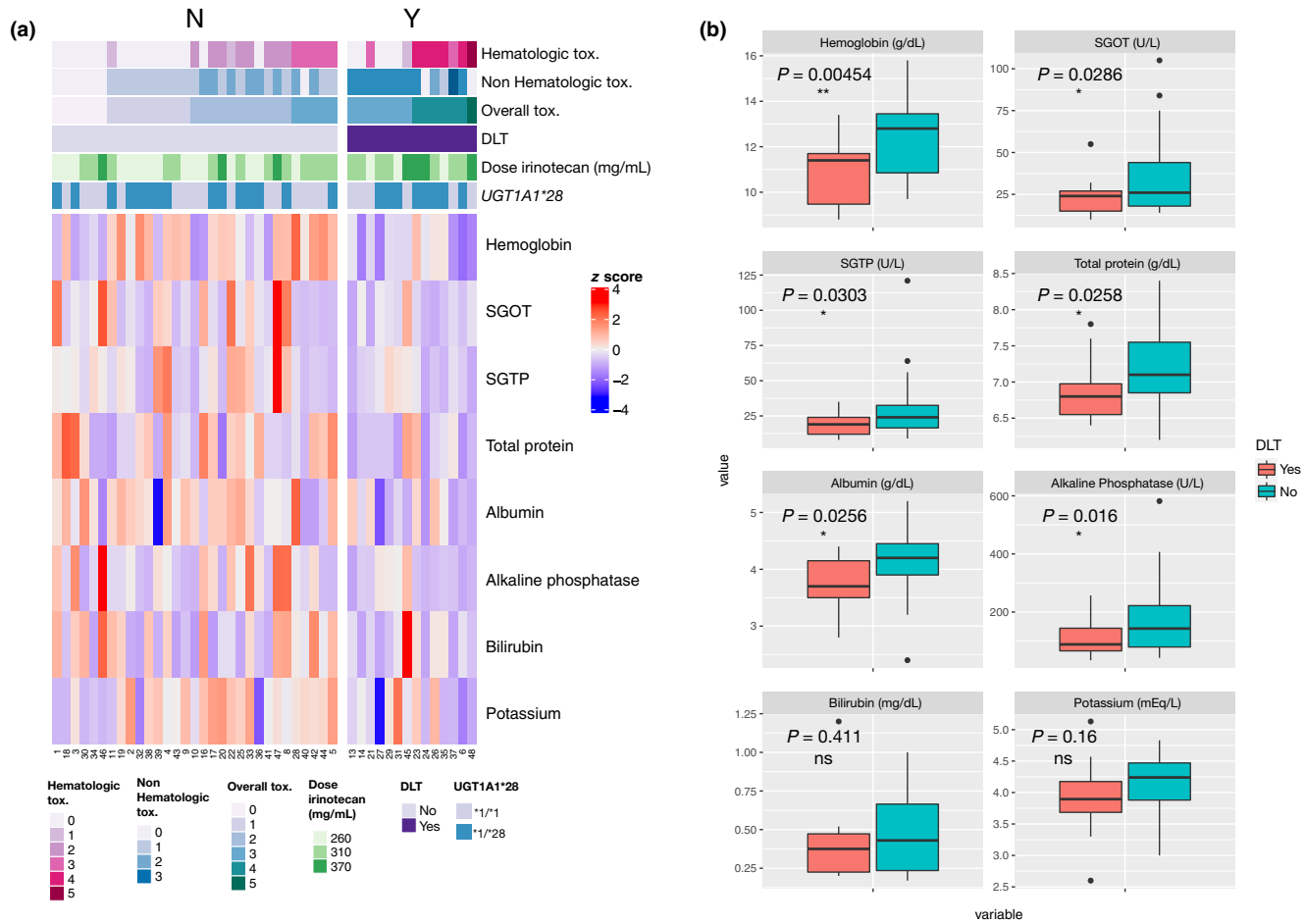

Figure 4 Graphical representation of DLT relevant variables used by the models and their difference between patients DLT = Yes and patients DLT = No. (**a**) Heat map representation of the DLT relevant variables used by the models. Patients in columns are ordered in accordance with the overall toxicity grade (increasing grade from left to right). Variables in rows have been standardized and represented as *z* scores (color scale from blue to red). (**b**) Box plots of the DLT-relevant variables between DLT groups, *P* value of each comparison has been computed from a two-sample Student's *t*-test. DLT, dose-limiting toxicity; N, no DLT; SGOT, serum glutamic oxaloacetic transaminase; SGTP, serum glutamic-pyruvic transaminase; tox., toxicity; Y, yes DLT.

typically involve about 20 to 60 subjects. The limited number of patients does not allow for the most common reliable way to validate ML model's performance, i.e., Train/Test Split, consisting of randomly splitting a portion of data before developing an ML model and to use only test data for validation. Recent studies focused on the importance of predicting error on small sample size and the definition of straightforward methods to prevent biased results.[14,26–28] Our approach employs a nested CV, meaning that a portion of data was repetitively split at each CV iteration and the model was then developed inside the CV process on the reduced training set from scratch, including feature selection and parameter tuning (**Figure 2**), providing robust and unbiased performance estimates regardless of sample size.[14,26] The use of robust performance estimators and confidence intervals computed within the CV process is of particular importance in studies with limited sample size. In our two-class classification problem, the classifier performance was measured using the MCC; this metric overcomes the accuracy measure by preventing the overoptimistic inflated results on imbalanced data sets[22] and by avoiding the probabilistic chance of resulting in 50% accuracy by random classification.[28] In fact, obtaining high MCC values occurs only when prediction is good on all four confusion matrix categories (true positives, false negatives, true negatives, and false positives); conversely, obtaining MCC values close to zero signals that classifier performance is no better than random guessing. Moreover, our model accuracy (ACC = 80.4%) is greater than the minimal statistically significant accuracy threshold (ACC = 62.5%) as function of sample size ($n = 40$), class number (c = 2), and significance levels ($P < 0.05$).[28]

The ML approach allowed us to estimate the importance of baseline variables for predicting the DLT, and taken together, hemoglobin, SGOT, and albumin were the most important (**Table 2**). Moreover, we found that their levels were significantly lower in patients that experienced DLT (**Figure 4**). These results are in keeping with previous studies reporting that the variables used by our model affect the irinotecan disposition and induced toxicity. Irinotecan is moderately bounded to albumin and erythrocytes, whereas SN-38, the active metabolite, is highly associated with albumin and blood cells.[29]

Previous data identified low baseline hemoglobin as an independent predictor for grade 3-4 hematologic toxicity and

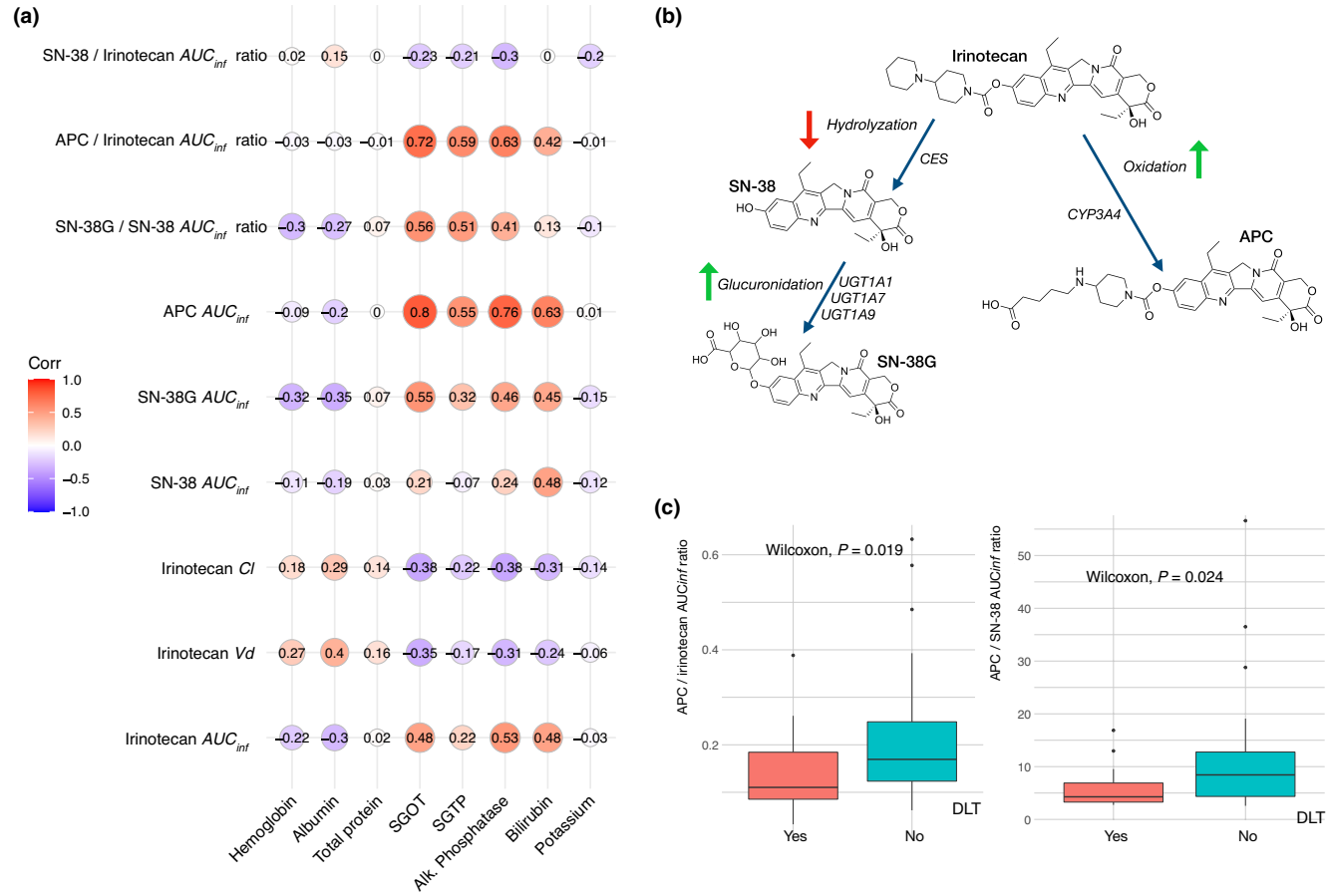

Figure 5 Relationship between the models' relevant variables and the irinotecan's pharmacokinetic. (**a**) Correlation plot of the models' relevant variables and the pharmacokinetic parameters of irinotecan and its metabolites (SN-38, SN-38G, and APC). Values in circles represent the Pearson correlation coefficient (*r*) computed between each pair of variables, and circle's color indicates the correlation direction (blue scale for negative correlation and red scale for positive correlation). (**b**) Metabolism of irinotecan. Trending arrows indicate the irinotecan metabolism direction associated with hepatic parameters' values. (**c**) Box plots of APC/irinotecan AUC_{inf} ratio and APC/SN-38 AUC_{inf} ratio between DLT groups. Alk. phosphatase, alkaline phosphatase; AUC_{inf}, area under the concentration-time curve extrapolated to infinity; *Cl*, clearance; Corr, correlation; DLT, dose-limiting toxicity; SGOT, serum glutamic oxaloacetic transaminase; SGTP, serum glutamic-pyruvic transaminase; *Vd*, distribution volume.

nonhematologic toxicity in patients treated with irinotecan-based therapies;[30–32] such suggestions agree with the results of our ML model that highlight the importance of baseline albumin and baseline hemoglobin as DLT predictors. Our model integrates also the evaluation of the SGOT liver enzyme. High concentrations of liver enzymes in plasma are observed in liver injury and is a common reason for cautiously treating patients with cancer due to reduced liver metabolic functions.[33] In our study we observed an opposite association in which DLT patients were characterized by lower values of SGOT. It must be considered that this trend was observed in patients with hepatic parameters within the normal range and not in extreme situations. In fact, in our patients the products of irinotecan metabolism SN-38G and APC, the irinotecan inactive metabolites, showed an increase in dose-normalized AUC_{inf} with increasing hepatic parameters (**Figure 5a**), suggesting that in this range the hepatic functionality, represented by UGT1A1, UGT1A7, UGT1A9, and CYP3A4 isoenzymes activity, appears to remain conserved. Moreover, the positive correlation between plasma liver enzymes and the AUC_{inf} of APC was also noticed in a previous study.[34] Presumably, our model selected the SGOT because within the normal range it does not represent a biomarker of liver injury, but rather a surrogate estimation of the irinotecan metabolism equilibrium in which higher SGOT values are associated with a higher irinotecan conversion toward the less cytotoxic SN-38G and APC metabolites (**Figure 5b**).

An ML approach like the one adopted in the present study has potential implications both for a better understanding of the study in which it is applied and in other similar contexts derived from phase I clinical trials. In fact, it could be translated to other phase I studies and surely represent a starting point for the next phases of the drug evaluation.

This possibility could be very interesting especially in the case of large sample size availability. Nevertheless, our approach is consistent with the "small data paradigm"[35] for precision medicine that suggests how ML application to small data sets could produce transportable knowledge useful for integration in the next phases of clinical trials.

The results obtained from an ML strategy, as those reported in this study, may have potential relevance for regulatory agencies by

improving the understanding of the phase I clinical trial results. It may provide, in the very early phase (phase I) of drug development, predictive markers of toxicity that could become of mandatory consideration (i.e., companion diagnostics) in the later drug developmental phases. This could also have important regulatory implications based on the possibility to highlight during the phase I study that a treatment is not safe in specific groups of patients (according to the genotype, ethnicity, or other characteristics) and that the use should therefore be limited to specific groups of patients.

We are fully aware of the limitations of our ML study. First, the number of patients involved is relatively small, although the employed validation techniques are the best solution when dealing with limited sample size.[14] Moreover, the small sample size could, at least in part, explain the lack of a significant correlation between biological parameters such as total proteins and any of the pharmacokinetic parameters. A second limitation is the lack of an external validation cohort, which is implicit due to the unrepeatability of the phase I studies. The limited panel of genetic variants that may be included in a phase I study with a limited number of patients and the unfeasibility to cover different ethnic populations could represent a potential source of bias. Moreover, additional aspects related to concomitant treatments or comorbidities could be missed.

However, implementing ML analysis to phase I clinical trial with a small number of patients is of critical importance for an early identification of biomarkers and pilot work, but bearing in mind that it could lead to biased ML performance estimates.[26]

Finally, we chose not to integrate pharmacokinetic parameters in our ML models because pharmacokinetic data can be obtained only once the dose is administered, preventing its usage as baseline prechemotherapy predictors.

In conclusion, this study provides a proof of concept of the potential application of ML techniques to phase I studies. This implementation enables the discovery and validation, with a reasonable degree of accuracy, of factors predicting induced severe toxicities by simultaneously analyzing multiple heterogeneous patient-related variables at baseline. This could be useful for subsequent phase II and III studies.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutic*s website (www.cpt-journal.com).

ACKNOWLEDGMENTS

This work was supported by the Italian Ministry of Health (Ricerca Corrente) (no grant number provided).

FUNDING

No funding was received for this work.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

L.B., E.C., E.F., M.D.B., A.B., M.P., and G.T. wrote the manuscript. L.B., M.P., and G.T. designed the research. L.B. and M.P. performed the research. L.B. and M.P. analyzed the data. E.C. and G.T. contributed new reagents/analytical tools.

1. Le Tourneau, C., Lee, J.J. & Siu, L.L. Dose escalation methods in phase I cancer clinical trials. *J. Natl. Cancer Inst.* **101**, 708–720 (2009).
2. Zhou, H., Yuan, Y. & Nie, L. Accuracy, safety, and reliability of novel phase I trial designs. *Clin. Cancer Res.* **24**, 4357–4364 (2018).
3. Iasonos, A., Wilton, A.S., Riedel, E.R., Seshan, V.E. & Spriggs, D.R. A comprehensive comparison of the continual reassessment method to the standard 3 + 3 dose escalation scheme in phase I dose-finding studies. *Clin. Trials* **5**, 465–477 (2008).
4. van Brummelen, E.M.J., Huitema, A.D.R., van Werkhoven, E., Beijnen, J.H. & Schellens, J.H.M. The performance of model-based versus rule-based phase I clinical trials in oncology: a quantitative comparison of the performance of model-based versus rule-based phase I trials with molecularly targeted anticancer drugs over the last 2 years. *J. Pharmacokinet. Pharmacodyn.* **43**, 235–242 (2016).
5. Harrer, S., Shah, P., Antony, B. & Hu, J. Artificial intelligence for clinical trial design. *Trends Pharmacol. Sci.* **40**, 577–591 (2019).
6. Bedon, L., Dal Bo, M., Mossenta, M., Busato, D., Toffoli, G. & Polano, M. A novel epigenetic machine learning model to define risk of progression for hepatocellular carcinoma patients. *Int. J. Mol. Sci.* **22**, 1075 (2021).
7. Polano, M. *et al*. A new epigenetic model to stratify glioma patients according to their immunosuppressive state. *Cells* **10**, 576 (2021).
8. Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B. & Goldenberg, A. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis. Oncol.* **4**, 19 (2020).
9. Rajpurkar, P. *et al*. Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: a prespecified secondary analysis of a randomized clinical trial. *JAMA Netw. Open* **3**, e206653 (2020).
10. Xu, J., Zhang, H., Zhang, H., Bian, J. & Wang, F. Machine learning enabled subgroup analysis with real-world data to inform better clinical trial design. *medRxiv* 2021.05.11.21257024 (2021). https://doi.org/10.1101/2021.05.11.21257024.
11. Gibson, W.J. *et al*. Machine learning versus traditional risk stratification methods in acute coronary syndrome: a pooled randomized clinical trial analysis. *J. Thromb. Thrombolysis* **49**, 1–9 (2020).
12. Toffoli, G. *et al*. Genotype-guided dosing study of FOLFIRI plus bevacizumab in patients with metastatic colorectal cancer. *Clin. Cancer Res.* **23**, 918–924 (2017).
13. de Man, F.M., Goey, A.K.L., van Schaik, R.H.N., Mathijssen, R.H.J. & Bins, S. Individualization of irinotecan treatment: a review of pharmacokinetics, pharmacodynamics, and pharmacogenetics. *Clin. Pharmacokinet.* **57**, 1229–1254 (2018).
14. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A.J. Machine learning algorithm validation with a limited sample size. *PLoS One* **14**, e0224365 (2019).
15. Kursa, M.B. & Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **36**, 1–13 (2010).
16. Kuhn, M. *et al*. Caret: classification and regression training. *R project* <https://CRAN.R-project.org/package=caret> (2020).
17. Breiman, L. Random forests. *Mach. Learn* **45**, 5–32 (2001).
18. Nelder, J.A. & Baker, R.J. Generalized linear models. In *Encyclopedia of Statistical Sciences* (eds. Kotz, S., Read, C.B., Balakrishnan, N., Vidakovic, B., & Johnson N.L.) (John Wiley & Sons, Hoboken, NJ, 2004). https://doi.org/10.1002/0471667196.ess0866.
19. Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
20. Cristianini, N. & Schölkopf, B. Support vector machines and Kernel methods: the new generation of learning machines. *AI Mag.* **23**, 31–41 (2002).

21. Zhang, Z. Introduction to machine learning: k-nearest neighbors. *Ann. Transl. Med.* **4**, 218 (2016).
22. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **21**, 6 (2020).
23. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2020) <https://www.R-project.org/>.
24. Schober, P., Boer, C. & Schwarte, L.A. Correlation coefficients: appropriate use and interpretation. *Anest. Analg.* **126**, 1763–1768 (2018).
25. James, G., Witten, D., Hastie, T. & Tibshirani, R. Resampling methods. In *An Introduction to Statistical Learning: With Applications in R* (James, G., Witten, D., Hastie, T. & Tibshirani, R.) 175–201 (Springer, New York, NY, 2013). https://doi.org/10.1007/978-1-4614-7138-7_5.
26. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91 (2006).
27. Shaikhina, T. & Khovanova, N.A. Handling limited datasets with neural networks in medical applications: a small-data approach. *Artif. Intell. Med.* **75**, 51–63 (2017).
28. Combrisson, E. & Jerbi, K. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J. Neurosci. Methods* **250**, 126–136 (2015).
29. Combes, O. *et al*. *In vitro* binding and partitioning of irinotecan (CPT-11) and its metabolite, SN-38, in human blood. *Invest. New Drugs* **18**, 1–5 (2000).
30. Glimelius, B. *et al*. Prediction of irinotecan and 5-fluorouracil toxicity and response in patients with advanced colorectal cancer. *Pharmacogenomics J.* **11**, 61–71 (2011).
31. Freyer, G. *et al*. Prognostic factors for tumour response, progression-free survival and toxicity in metastatic colorectal cancer patients given irinotecan (CPT-11) as second-line chemotherapy after 5FU failure. *Br. J. Cancer* **83**, 431–437 (2000).
32. Díaz, R. *et al*. Clinical predictors of severe toxicity in patients treated with combination chemotherapy with irinotecan and/or oxaliplatin for metastatic colorectal cancer. *Med. Oncol.* **23**, 347–357 (2006).
33. Raymond, E. *et al*. Dosage adjustment and pharmacokinetic profile of irinotecan in cancer patients with hepatic dysfunction. *J. Clin. Oncol.* **20**, 4303–4312 (2002).
34. Rouits, E. *et al*. Pharmacokinetic and pharmacogenetic determinants of the activity and toxicity of irinotecan in metastatic colorectal cancer patients. *Br. J. Cancer* **99**, 1239–1245 (2008).
35. Hekler, E.B., Klasnja, P., Chevance, G., Golaszewski, N.M., Lewis, D. & Sim, I. Why we need a small data paradigm. *BMC Med.* **17**, 133 (2019).