







Machine Learning Application Identifies Germline Markers of Hypertension in Patients With Ovarian Cancer Treated With Carboplatin, Taxane, and Bevacizumab

Maurizio Polano¹ , Luca Bedon¹ , Michele Dal Bo¹ , Roberto Sorio², Michele Bartoletti², Elena De Mattia¹ , Erika Cecchin¹ , Carmela Pisano³, Domenica Lorusso^{4,5}, Andrea Alberto Lissoni⁶, Andrea De Censi⁷, Sabrina Chiara Cecere³, Paolo Scollo⁸, Sergio Marchini⁹, Laura Arenare¹⁰, Ugo De Giorgi¹¹, Daniela Califano¹², Elena Biagioli¹³, Paolo Chiodini¹⁴, Francesco Perrone¹⁰, Sandro Pignata^{3,†} and Giuseppe Toffoli^{1,*†} 

Pharmacogenomics studies how genes influence a person's response to treatment. When complex phenotypes are influenced by multiple genetic variations with little effect, a single piece of genetic information is often insufficient to explain this variability. The application of machine learning (ML) in pharmacogenomics holds great potential — namely, it can be used to unravel complicated genetic relationships that could explain response to therapy. In this study, ML techniques were used to investigate the relationship between genetic variations affecting more than 60 candidate genes and carboplatin-induced, taxane-induced, and bevacizumab-induced toxicities in 171 patients with ovarian cancer enrolled in the MITO-16A/MaNGO-OV2A trial. Single-nucleotide variation (SNV, formerly SNP) profiles were examined using ML to find and prioritize those associated with drug-induced toxicities, specifically hypertension, hematological toxicity, nonhematological toxicity, and proteinuria. The Boruta algorithm was used in cross-validation to determine the significance of SNVs in predicting toxicities. Important SNVs were then used to train eXtreme gradient boosting models. During cross-validation, the models achieved reliable performance with a Matthews correlation coefficient ranging from 0.375 to 0.410. A total of 43 SNVs critical for predicting toxicity were identified. For each toxicity, key SNVs were used to create a polygenic toxicity risk score that effectively divided individuals into high-risk and low-risk categories. In particular, compared with low-risk individuals, high-risk patients were 28-fold more likely to develop hypertension. The proposed method provided insightful data to improve precision medicine for patients with ovarian cancer, which may be useful for reducing toxicities and improving toxicity management.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Hypertension and other adverse drug reactions are common side effects of chemotherapy and bevacizumab therapy. Pharmacogenomics can be used to predict patient-specific variability in adverse drug effects, due to complex phenotypes simultaneously influenced by multiple genetic variations.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Can machine learning be useful to examine patients' single nucleotide polymorphism profiles to possibly comprehend complex genetic relationships that might explain drug-induced toxicities, particularly hypertension, in patients with ovarian cancer?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ This study identified germline genetic variants associated with drug-induced toxicities in patients with ovarian cancer receiving first-line therapy with bevacizumab, paclitaxel, and carboplatin. It was discovered that 43 single nucleotide polymorphisms were important for toxicity prediction. We also created a comprehensive scoring system that classified patients as having a high or low risk of toxicity.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ Decision support tools developed by machine learning to predict the likelihood of toxicity can reduce the proportion of patients with genetic drug toxicities and improve toxicity management.

¹Experimental and Clinical Pharmacology Unit, Centro di Riferimento Oncologico di Aviano, Istituto di Ricovero e Cura a Carattere Scientifico, Aviano, Italy; ²Dipartimento di Oncologia Medica, Centro di Riferimento Oncologico di Aviano, Istituto di Ricovero e Cura a Carattere Scientifico, Aviano, Italy; ³Uro-Gynecologic Oncology Unit, Istituto Nazionale Tumori Istituto di Ricovero e Cura a Carattere Scientifico Fondazione G. Pascale, Naples, Italy; ⁴Department of Women and Child Health, Division of Gynecologic Oncology, Fondazione Policlinico Universitario A. Gemelli Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy; ⁵Department of Life Science and Public Health, Catholic University of Sacred Heart Largo Agostino Gemelli, Rome, Italy; ⁶Clinica Ostetrica e Ginecologica, Istituto di Ricovero e Cura a Carattere Scientifico S. Gerardo Monza, Università di Milano Bicocca, Milano, Italy; ⁷Oncologia Medica, Ospedali Galliera, Genoa, Italy; ⁸Unità Operativa Ostetrica e Ginecologia, Dipartimento Materno-Infantile, Ospedale Cannizzaro, Catania, Italy; ⁹Molecular Pharmacology laboratory, Group of Cancer Pharmacology Istituto di Ricovero e Cura a Carattere Scientifico Humanitas Research Hospital, Rozzano, Italy; ¹⁰Clinical Trial Unit, Istituto Nazionale Tumori, Istituto di Ricovero e Cura a Carattere Scientifico, Fondazione G. Pascale, Naples, Italy; ¹¹Istituto di Ricovero e Cura a Carattere Scientifico Istituto Romagnolo per lo Studio dei Tumori Dino Amadori, Meldola, Italy; ¹²Microenvironment Molecular Targets Unit, Istituto Nazionale Tumori IRCCS, Fondazione G. Pascale, Naples, Italy; ¹³Department Of Oncology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS Milano, Milano, Italy; ¹⁴Department of Mental Health and Public Medicine, Section of Statistics, Università degli Studi della Campania Luigi Vanvitelli, Naples, Italy. *Correspondence: Giuseppe Toffoli (gtoffoli@cro.it)

†Co-last authors.

Pharmacogenomics (PGx) aims to use the genetic information of an individual's germline to predict differences in drug response in order to provide safer, more effective, and less costly treatment.¹

Machine learning (ML) is one of the “key technologies” for the development of precision medicine.² By using patients' SNP profiles as training data during model development, ML offers the potential to unravel intricate genetic relationships that explain response to therapy.³

Ovarian cancer is the most malignant gynecologic cancer, with 314,000 new diagnoses and 207,000 deaths in 2020.⁴ The 5-year overall survival rate ranges from 30% to 50%, with most patients diagnosed with advanced-stage disease (III-IV).⁵ The standard treatment approach for advanced-stage cancer was limited to surgical removal of the tumor and platinum-based chemotherapy until bevacizumab was approved as a monoclonal antibody for ovarian cancer. Bevacizumab enhanced progression-free survival in patients with ovarian cancer, as shown by a phase III trial, but additional adverse side effects were also noted, particularly hypertension.⁶ In this context, the presence of basal hypertension (prior to bevacizumab administration) was associated with a high probability of treatment-related hypertension from bevacizumab.^{7–9} A genome-wide association study (GWAS) was conducted to associate gene polymorphisms with treatment toxicity,¹⁰ and further associations were observed between the presence of several SNVs and the occurrence of toxicity, including proteinuria and nonhematological and hematological toxicity.^{9,10}

The MITO-16A/MaNGO-OV2A trial sought to evaluate clinical and molecular factors that may predict the prognosis and safety of patients receiving combined chemotherapy plus bevacizumab.¹¹ Here, through an analysis of toxicity and germline DNA sequencing data from 171 patients enrolled in the phase IV MITO-16A/MaNGO-OV2A trial,¹¹ ML techniques were used to investigate the association between germline genetic variants in over 60 candidate genes and carboplatin-induced, taxane-induced, and bevacizumab-induced toxicities in patients with ovarian cancer. The proposed multistep approach involving genotype calling, toxicity model training, variant prioritization, and variant analysis allowed us to reveal associations between germline genotypes and drug-induced toxicities, providing valuable information to improve precision medicine for patients with ovarian cancer.

METHODS

Trial description and study population.

The MITO-16A/MaNGO-OV2A trial ([ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT01706120) Identifier: NCT01706120) was designed as a single-arm, open-label, noncomparative, multicenter phase IV study in patients with advanced or recurrent previously untreated ovarian cancer who received a combination of bevacizumab, paclitaxel, and carboplatin as first-line treatment.^{11,12}

The main inclusion criteria were previously untreated patients aged at least 18 years, histologically confirmed stage IIIB-IV ovarian cancer (International Federation of Gynecology and Obstetrics, FIGO), performance status 0–2 according to the Eastern Cooperative Oncology Group (ECOG), and a life expectancy of at least 12 weeks. The inclusion criteria for diastolic and systolic pressure were ≤ 90 Hg and ≤ 140 Hg, respectively. Treatments for hypertension were allowed.

Patients were treated with carboplatin (5 area under the curve (AUC)) and paclitaxel (175 mg/m²) plus bevacizumab (15 mg/kg) on Day 1 for six 3-week cycles, followed by bevacizumab monotherapy (15 mg/kg), up to a maximum of 22 cycles in total.

Toxicity was graded according to Common Terminology Criteria for Adverse Events (CTCAE) 4.03. Survival response rate was coded according to Response Evaluation Criteria in Solid Tumors (RECIST) 1.1 guidelines. Hypertension was defined as the presence of at least one of the following conditions: systolic blood pressure ≥ 140 mmHg, diastolic blood pressure ≥ 90 mmHg, or new medical intervention.

A prospective PGx study was conducted in a subset of patients who had agreed to provide blood samples for PGx analysis as part of the MITO study. The primary objective was to determine whether a patient's PGx analysis can be used to predict adverse drug reactions. In the context of the prospective PGx study, a total of 171 patients (White women, median age 59 years, interquartile range (IQR) 50–66 years) were selected as candidates for targeted next-generation sequencing (NGS) and included in the toxicity assessment.

Targeted next-generation gene sequencing

DNA (germline DNA) was extracted from blood samples of 171 patients using the BioRobot EZ1 (Qiagen SPA, Milan, Italy).

A custom panel was developed in-house using 60 genes derived from candidate ovarian cancer genes as previously shown.¹³ The pharmacogenomic germline profiles of the samples were characterized by NGS using a custom hybrid capture-based assay (Roche/NimbleGen, Madison, WI). Custom panel design was performed using NimbleDesign software based on Genome Build hg19/GRCh37 (February 2009) to capture the genetic variability of each exon and nearby splice junctions (~35 bases upstream and downstream of the exon). Sample preparation was performed according to the manufacturer's instructions as described in the SeqCap EZ Library SR User's Guide v3.0 (Roche/NimbleGen, Madison, WI).

DNA libraries were sequenced using the Miseq Illumina platform (Illumina, Inc., San Diego, CA).

From sequencing to genotype data

FASTQ files from the NGS were processed using a custom pipeline for germline calling based on GATK4 best practices. A full coverage check of the targets was also integrated into this pipeline to assess calling rate and genotype quality. To further develop our models, we selected all dbSNP (Single Nucleotide Polymorphism Database)-derived SNVs (table build 151) with a coverage of more than 50. A merged variant call format file (vcf) with high quality SNVs was converted to a PED file using PLINK2 software. The genotype was coded as 0, 1, or 2, according to the number of copies of the alternative allele.

We excluded SNVs with a cohort-calculated minor allele frequency (MAF) of ≤ 0.05 . Since strongly linked polymorphic sites lead to highly correlated variables, the genotype table was filtered to retain only one site (Cramér's $V \geq |0.9|$). The retained SNV sites isolated from germline NGS data were used for subsequent analysis.

Toxicity prediction models: Training, validation, and performance

Based on the final SNV table, we trained and validated ML algorithms to predict toxicities that occurred during the study (Figure S1). A detailed workflow summarizing the model development pipeline is shown in Figure S2.

For this task, we implemented eXtreme Gradient Boosting models (XGB) with a linear function as the booster model.¹⁴

Features were selected by implementing the Boruta algorithm with five-fold cross-validation (CV).^{15,16} The importance value was determined as mean decreased accuracy (MDA) for all features.¹⁶

The features selected from the variant importance step were used as inputs to the XGB classifier, which was then validated using leave-one-out cross-validation (LOOCV). Within LOOCV, performance metrics were calculated in terms of both Matthews correlation coefficient (MCC)¹⁷ and accuracy (ACC). Optimization of the XGB hyperparameters was performed within CV by passing a user-defined grid parameter.

The XGBoost library was used to implement the XGB classifier.¹⁸ The importance of the variables within the recursive feature elimination (RFE) process was calculated using Boruta as a z -scaled mean of reduced precision (MDA).¹⁶ The R package “caret” was used to implement and tune the XGB classifier within the CV and RFE processes.¹⁹

Variants analysis and polygenic toxicity score

A logistic regression model was used to determine the effect of SNV genotype on the occurrence of toxicity. The association between SNV genotype and toxicity was measured by the odds ratio (OR). OR greater than 1 indicates an increased incidence of toxicity in patients with an alternative genotype (1 or 2 coded genotype) compared with the reference (0 coded genotype); conversely, OR less than 1 indicates a decreased incidence of toxicity in patients with an alternative genotype (1 or 2 coded genotype) compared with the reference (0 coded genotype).

For each model, we constructed a multivariate logistic regression model using stepwise model selection (stepAIC function from the “MASS” package, USA).²⁰

The toxicity probability of each patient was calculated using the final logistic regression model and then patients were stratified into high-risk toxicity and low-risk toxicity based on the probability cutoff of 0.5.

To describe the association of SNVs with genes and pathways, each SNV was assigned to a specific gene in the panel based on its genomic position; genes were assigned to a specific reactome pathway (<http://reactome.org>) using the functional enrichment analysis of the STRING tool (<https://string-db.org>).

LDpair from the LDlinkR package, USA²¹ was used to investigate potentially correlated alleles for a pair of variants using European (“EUR”) populations (1000 Genomes Project).

Graphical representations were created using the “ggplot2” package²² and heatmap visualization was done using the “ComplexHeatmap” package.²³

RESULTS

Targeted next-generation gene sequencing and genotype calling

To characterize the genotype variants of the entire cohort of 171 cases, DNA libraries were sequenced with 75-fold mean coverage for each sample. More than 95% of each sequenced sample was covered at least 30-fold. Mutations at the 3'-UTR were the type of DNA variants most frequently detected in this panel (Table S1).

A total of 588 SNVs were mapped within the 699 sequenced genomic regions. Of these, the target bases of 348 SNVs were selected. During the filtering step, 120 SNVs were removed due to a low MAF (cohort MAF ≤ 0.05) and another 60 SNVs were removed because they were highly correlated (Cramér's $V \geq |0.9|$). Finally, 168 SNV sites (Figure 1) were isolated from the germline NGS data and used for further analysis.

The waterfall diagram in Figure S3 shows the 20 genes with the highest mutation rates in the sequenced cohort. Notably, this list includes genes well known to be associated with chemotherapy, such as *ABCC2* (100%), *HNF1* (100%), and *MMPN9* (100%).

Predictive performances of toxicity models

Adverse event data for 171 patients are presented in Table S2. Although the population studied is a subsample of the MITO-16A/MaNGO-OV2A study (398 patients), the distribution of adverse events is consistent with the original study.¹¹

The most common toxicities that occurred throughout the treatment course were hypertension (grade ≥ 3 in 38%), proteinuria (grade ≥ 1 in 27.8%), gastrointestinal toxicity (grade ≥ 1 in 67.5%), neurotoxicity (grade ≥ 1 in 62.7%), anemia (grade ≥ 1 in 52.8%), neutropenia (grade ≥ 1 in 66.3%, grade 3 in 43%), and thrombocytopenia (grade ≥ 1 in 23.9%).

In order to simplify the analysis when applying ML, we created additional new toxicity categories by combining different CTCAE toxicities of interest in the original studies (Table S2).^{11,12} These two new toxicity categories were hematological toxicity ≥ 3 , including anemia, neutropenia and thrombocytopenia, and non-hematological toxicity ≥ 3 , including hypertension, proteinuria, gastrointestinal toxicity, hepatological toxicity, and neurotoxicity. Since the high clinical relevance of hypertension and proteinuria these two toxicities were also singularly evaluated.¹⁰

Thus, we used novel categories and grade groupings to create binary predictive categories representing our research questions (Table S2); the categories were hypertension ≥ 3 (yes in 38%), hematological toxicity ≥ 3 (yes in 46%), nonhematological toxicity ≥ 3 (yes in 43%) and proteinuria ≥ 1 (yes in 27%).

By applying the model development pipeline (Figure S2) to the final 168 SNVs, we trained and validated the XGB classifier for predicting predefined toxicity categories (Table S2), specifically hypertension ≥ 3 , hematological toxicity ≥ 3 , nonhematological toxicity ≥ 3 , and proteinuria ≥ 1 . The results of implementation are shown in Table 1.

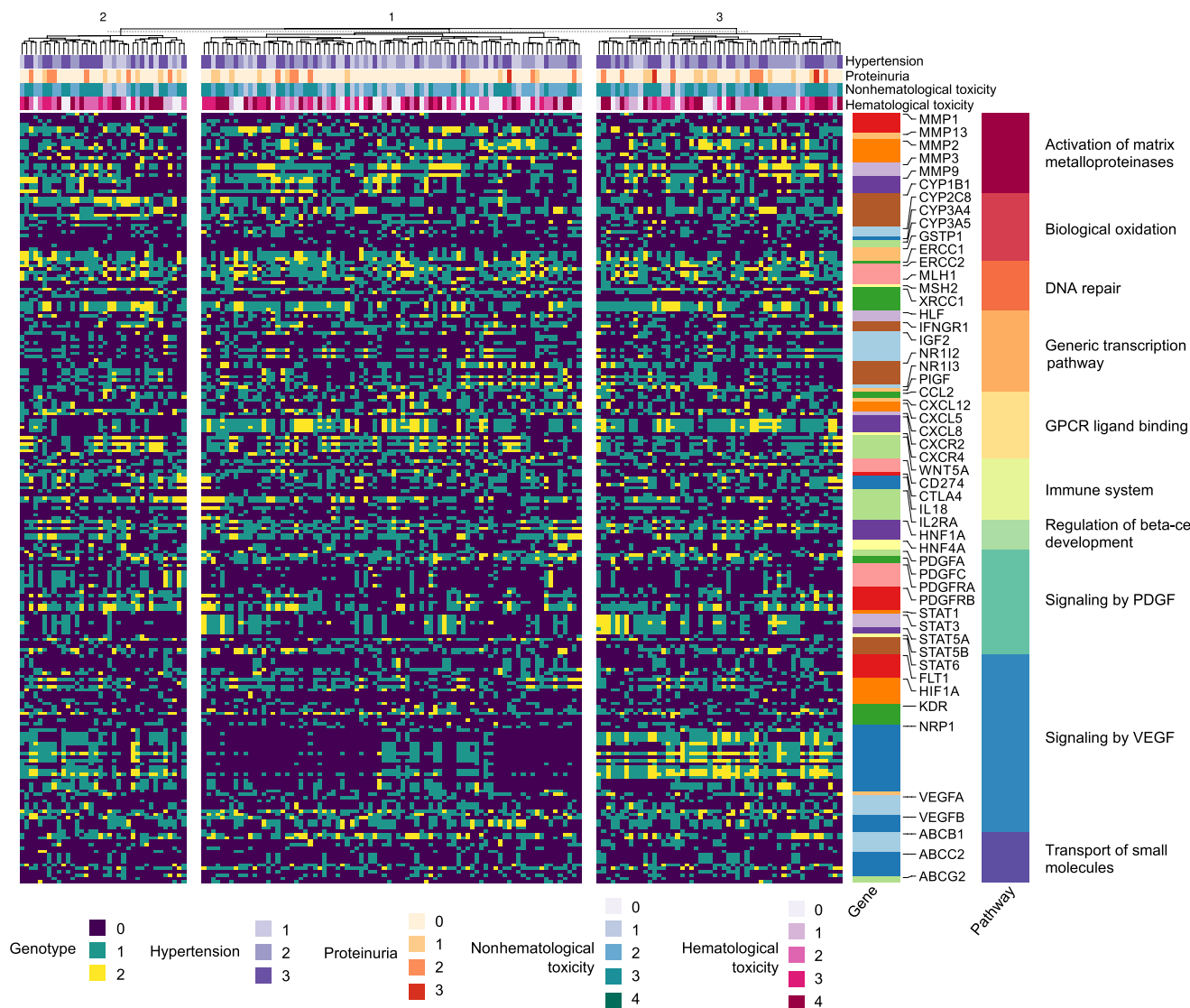


Figure 1 Heatmap of 171 patients clustered by the final 168 SNPs. The patients in the columns are divided into three main groups based on hierarchical clustering of SNV genotype (Euclidean distance, Ward method). The genotypes of the 168 SNPs are shown in rows and coded with the colors 0, 1, or 2. The gene and pathway annotation of the SNPs is given on the right-hand side. SNV, single-nucleotide variation.

Table 1 ML algorithms' performance computed within the LOOCV.

Model	nvar	nrounds	lambda	alpha	eta	MCC
Hypertension ≥ 3	20	100	1.00	0.10	0.001	0.375
Proteinuria ≥ 1	11	1	0.01	0.01	0.001	0.410
Nonhematological ≥ 3	25	100	0.01	0.10	0.001	0.400
Hematological toxicity ≥ 3	13	1	0.10	1.00	0.001	0.388

Performances of each XGB classifier computed within the LOOCV validation and expressed as MCC. The best XGB hyperparameters are also reported. alpha, Lasso regularization 1 term on weights; eta, step size shrinkage used in update; lambda, Lasso regularization 2 term on weights; LOOCV, leave-one-out cross-validation; MCC, Matthew correlation coefficient; nrounds, the number of rounds for boosting; nvar, the number of variables used to train the model; XGB, eXtreme Gradient Boosting model.

Variant importance highlights impact of individual SNVs on predicting toxicities

The contribution of each SNV to the prediction of toxicities was calculated as a z -scaled MDA using the Boruta algorithm. If an SNV performed better than the maximum MDA z score among

the shadow attributes, it was considered important for toxicity prediction (Figure 2). The list of important SNVs and their relative z -scaled MDA is shown in Table S3.

Twenty SNVs were determined to be important for the hypertension model (Figure 2a); the mean z -scaled MDA of the selected

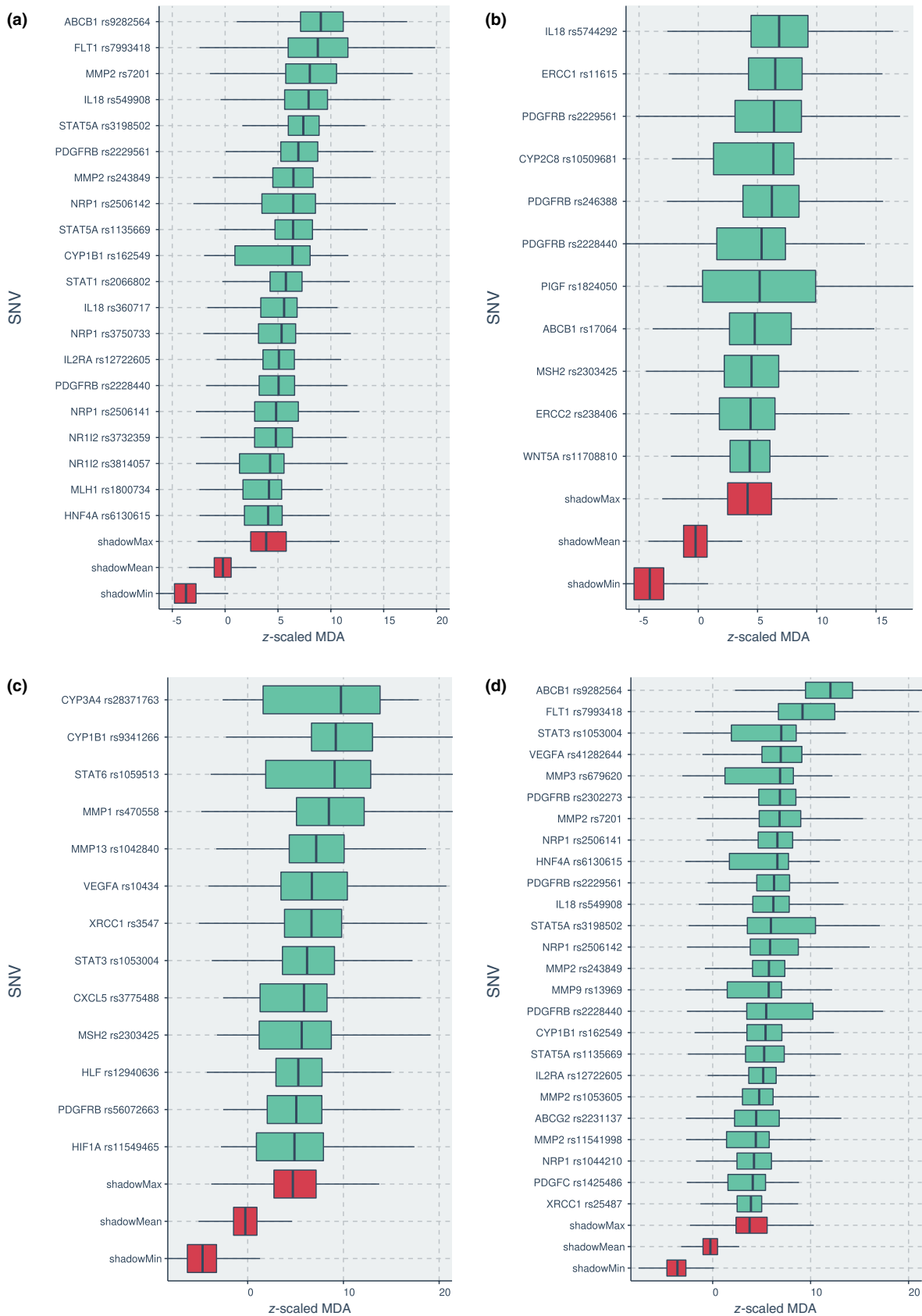


Figure 2 Importance plots of toxicity models. (a) Hypertension; (b) proteinuria; (c) hematological toxicity; (d) nonhematological toxicity. Boruta plot with box plots of SNV importance over the course of cross-validation. Importance is given as z-scaled MDA, and SNVs are sorted by the median of the z-scaled MDA. The red box plots correspond to the minimum (shadowMin), average (shadowMean), and maximum (shadowMax or MZSA) z score of the shadow attributes. The green box plots represent confirmed important SNVs that perform better than the shadowMax attribute. MDA, mean decreased accuracy; SNV, single-nucleotide variation.

SNVs was 6.13 (IQR = 4.99–7.04), and *ABCB1* rs9282564 was the best (z -scaled MDA = 9.05). Eleven SNVs were determined to be important for the proteinuria model (Figure 2b); the mean z -scaled MDA of the selected SNVs was 5.52 (IQR = 4.62–6.36), and *IL18* rs5744292 was the best (z -scaled MDA = 6.82). Thirteen SNVs were identified as important in the hematological toxicity model (Figure 2c); the mean z -scaled MDA of the selected SNVs was 6.92 (IQR = 5.64–8.48), and *CYP3A4* rs28371763 was the best (z -scaled MDA = 9.74). Twenty-five SNVs were identified as important for the nonhematological toxicity model (Figure 2d); the mean z -scaled MDA of the selected SNVs was 6.07 (IQR = 5.16–6.83), and *ABCB1* rs9282564 was the best (z -scaled MDA = 12.02). It is noteworthy that the mean value of the z -scaled MDA, which expresses the mean significance of the selected SNVs, is comparable for all models, resulting in comparable performance of the models.

To assess whether significant SNVs were primarily detected in a particular gene or pathway, each SNV was assigned to a single gene

in the panel based on its genomic position, and the genes were then assigned to a particular reactome pathway (Figure 3). For each toxicity model, the list of genes containing important SNVs is summarized in Table S4.

Relationships between important SNVs of examined toxicities

A total of 43 SNVs were used for the models out of the 168 SNV sites originally extracted from the germline NGS data. The non-hematological ≥ 3 models and hypertension ≥ 3 have 14 SNVs in common (nonhematological toxicity \cap hypertension) (Figure 4), which are probably sufficient to explain the hypertension cases with nonhematological toxicity. Most of the common SNVs were identified in the platelet derived growth factor (PDGF)(4 SNVs) and vascular endothelial growth factor (VEGF) (3 SNVs) pathways, suggesting a key role for these pathways in the development of hypertension. Of note, *ABCB1* rs9282564 was present in both

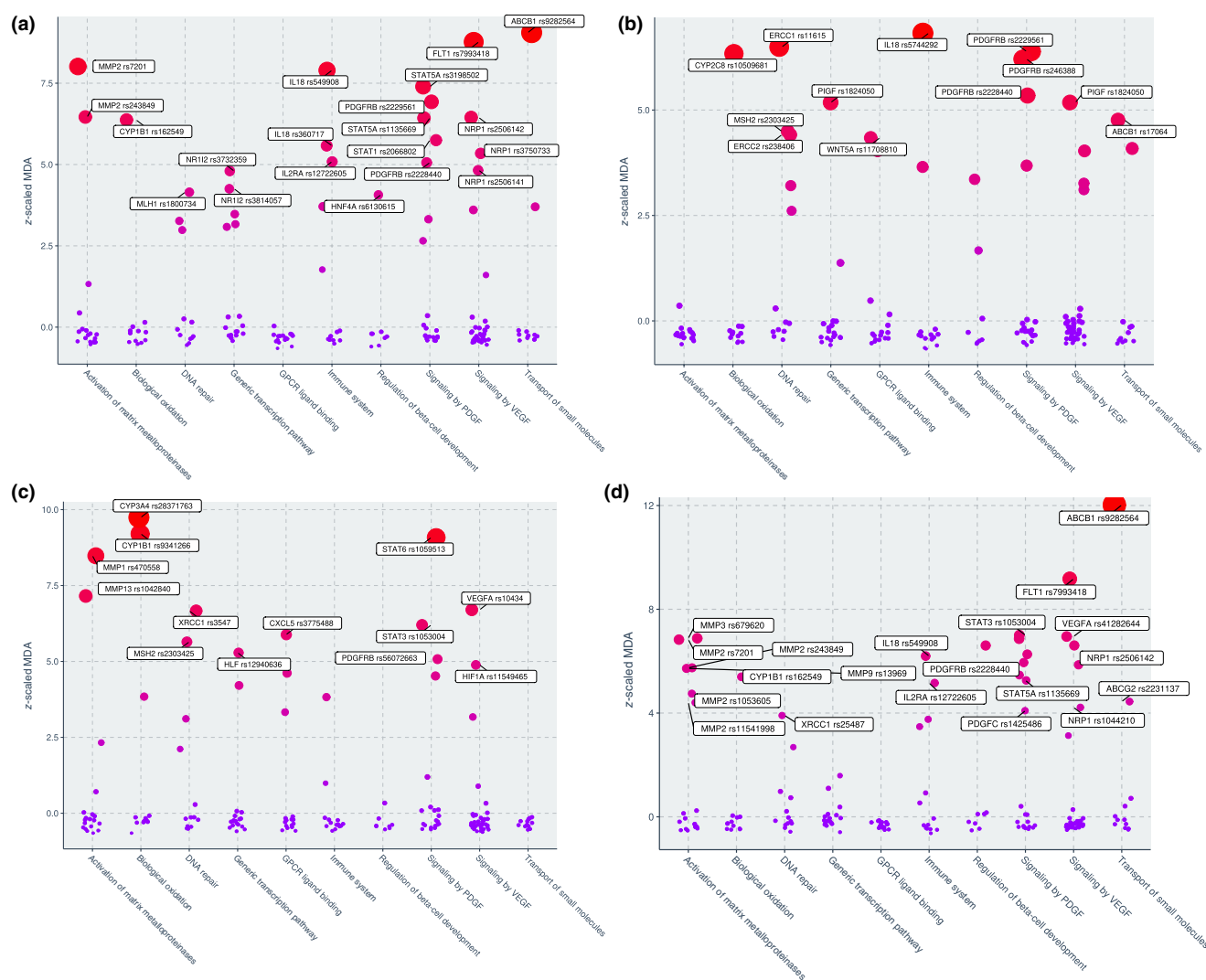


Figure 3 Enrichment plots of important SNVs. (a) Hypertension; (b) proteinuria; (c) hematological toxicity; (d) nonhematological toxicity. An enrichment plot indicates the pathways in which the most important SNVs are found. Importance as z -scaled MDA is reported in the y-axis while pathways are given in the x-axis. After SNVs were mapped to genes based on their genomic position, genes were assigned to a reactome pathway. MDA, mean decreased accuracy; MZSA, maximum MDA Z score among shadow attributes; SNV, single-nucleotide variation.

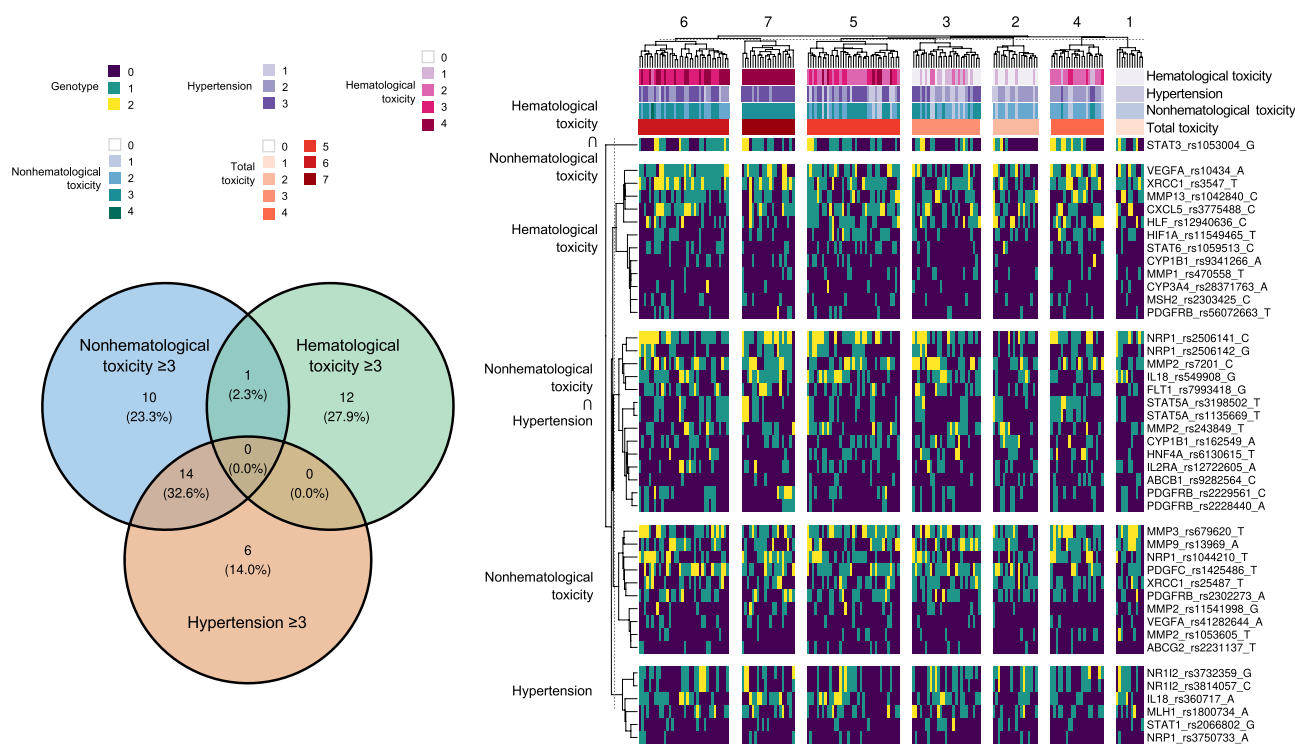


Figure 4 Relationships between important SNVs. The association between significant 43 SNVs from the toxicity analyzed is illustrated on the left by the Venn diagram. The heatmap reflecting the 43 SNVs is shown on the right, with patients sorted into seven primary groups by severity of overall toxicity (grade sum of hematological and nonhematological toxicities) and the SNVs in the rows are grouped by toxicity and relative intersections. Row and column groups are clustered by hierarchical clustering of SNV genotypes (Euclidean distance, Ward method). Genotypes of 43 SNVs are displayed in rows and color-coded as 0, 1, or 2. SNV, single-nucleotide variation.

toxicities and proved to be the most significant SNV in both toxicities.

We used the 14 SNVs shared by the nonhematological toxicity ≥ 3 and hypertension ≥ 3 models (nonhematological toxicity \cap hypertension) to determine whether the smaller set of SNVs was still sufficient to predict such toxicities. The MCC score for LOOCV prediction of hypertension was 0.406 ($ACC = 0.725$), while the MCC score for prediction of nonhematological toxicity was 0.387 ($ACC = 0.702$); the set of 14 SNVs not only provides sufficient information to predict hypertension, but also slightly increases the MCC score. Focusing on SNVs that can explain hypertension cases within nonhematological toxicity leads, as expected, to a slightly lower MCC value for the prediction of global nonhematological toxicity.

Association variants: Toxicity and polygenic toxicity risk score.

To investigate the influence of SNV genotype on the occurrence of toxicity, we used a logistic regression model; odds ratios were used to evaluate the relationship between SNV genotype and toxicity. The results of the univariate regression modeling are shown in [Table S5](#).

We then created a polygenic toxicity risk score using multivariate logistic regression with stepwise model selection to reduce Akaike information criteria to further investigate the cumulative effects of significant SNVs on the development of toxicity ([Table S6](#), [Figure 5](#)). When calculating the polygenic risk score

for hypertension ≥ 3 , basal hypertension values were taken into account because basal hypertension was associated with an increased risk of developing hypertension toxicity ([Table S7](#)). The corresponding log odds of each patient was calculated based on the patient's genotype, and then a hypertension ≥ 3 probability was calculated. Finally, patients were stratified into high-risk hypertension and low-risk hypertension ([Figure 6a](#)). Log odds for the other toxicities were derived using the same method, using the odds ratios of their SNVs ([table S6](#)) to stratify individuals into high-risk and low-risk groups for each toxicity ([Figure 6b–d](#)). The ability of the scoring system to correctly classify patients into high-risk or low-risk is summarized as the odds ratio for high-risk patients in [Table S8](#). Compared with low-risk patients, high-risk patients were 8.44 to 45.3 times more likely to develop toxicity.

DISCUSSION

In this study, a rule-based ML approach was used to identify relationships between genome-related characteristics and treatment-related toxicity, particularly hypertension, the most common nonhematologic adverse effect reported with bevacizumab/chemotherapy treatment.^{8,24}

A basal hypertension state (prior to bevacizumab administration) was associated with a higher incidence of hypertension toxicity.^{11,25,26} Several studies, including a GWAS with four clinical trials, have shown an association between different germline variants and bevacizumab-related toxicities such as hypertension.¹⁰ In the present study, we specifically focused on a limited gene panel of

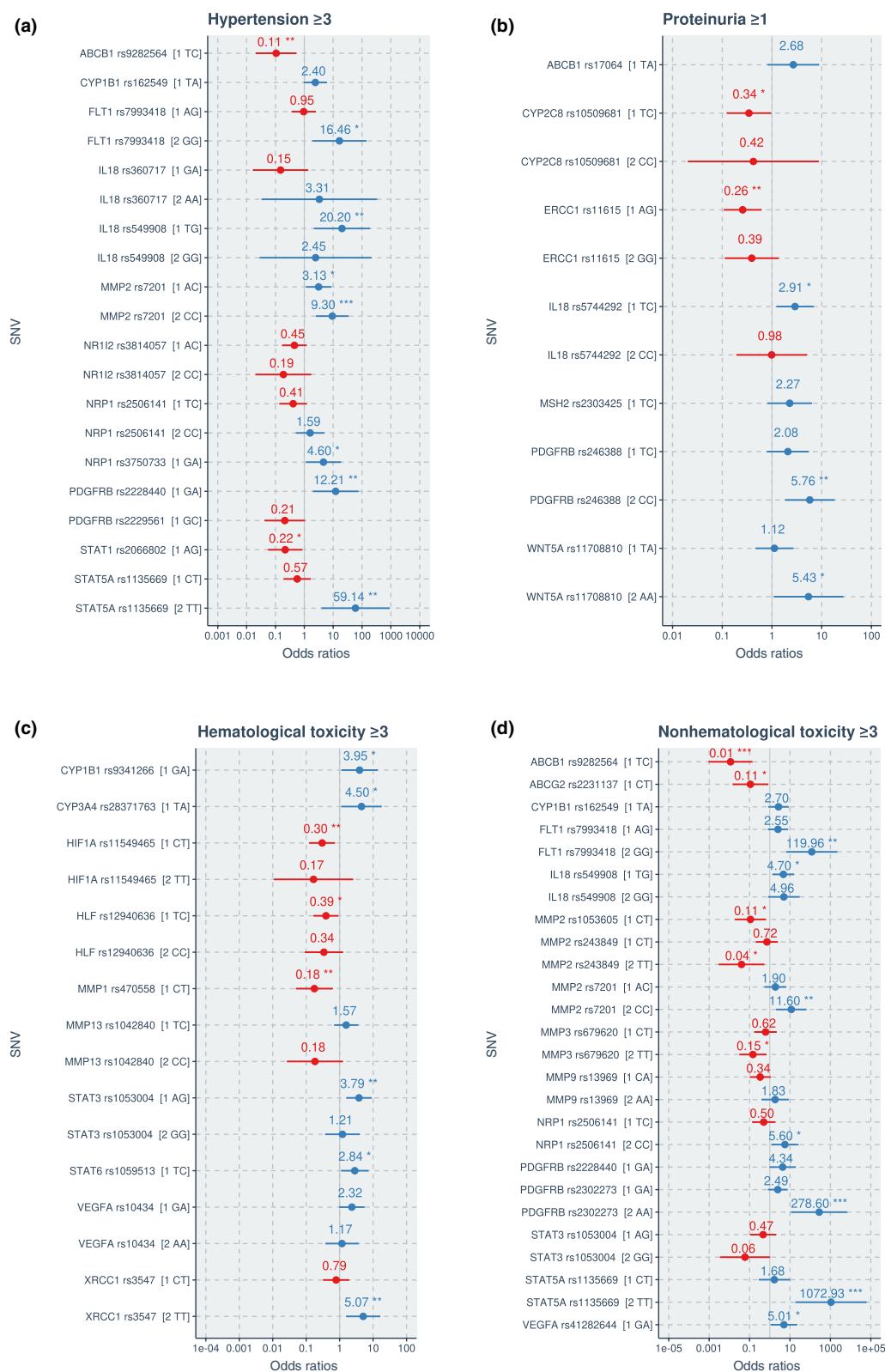


Figure 5 Odds ratios of each polygenic risk score. **(a)** Hypertension; **(b)** proteinuria; **(c)** hematological toxicity; **(d)** nonhematological toxicity. We used multivariate logistic regression to further investigate the cumulative effects of significant SNVs on the development of toxicity. Blue dots and relative confidence interval bars represent odds ratios greater than 1 indicating increased incidence of toxicity in patients carrying that particular genotype; conversely, red dots and relative confidence interval bars represent odds ratios less than 1 indicating decreased incidence of toxicity in patients carrying that particular genotype. SNV, single-nucleotide variation. * $P < 0.05$; ** $P < 0.01$, *** $P < 0.001$.

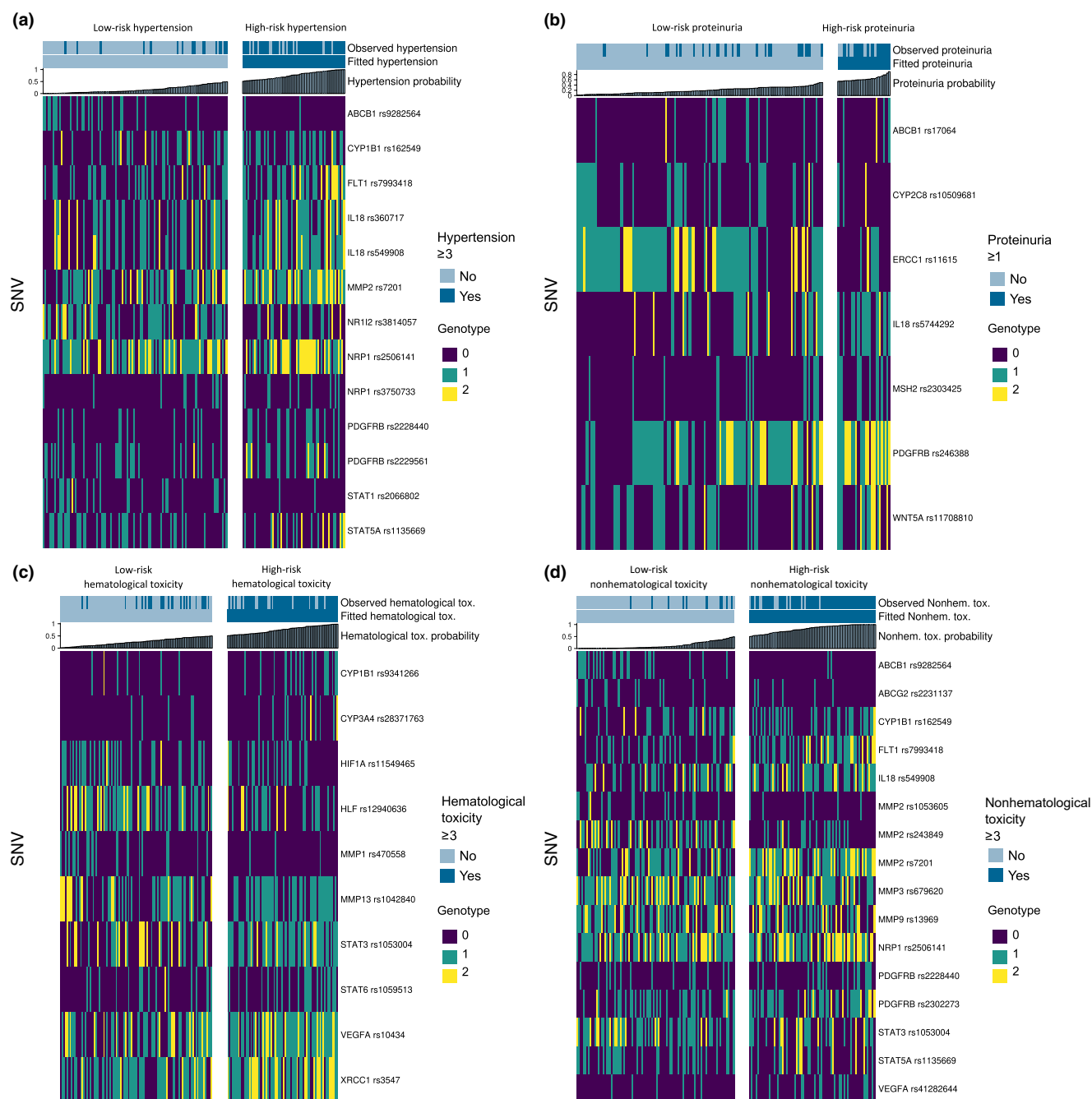


Figure 6 Heatmap of patients stratified as low and high risk for each toxicity. (a) Hypertension; (b) proteinuria; (c) hematological toxicity; (d) nonhematological toxicity. The patients are arranged in the columns according to increasing probability of toxicity calculated from each polygenic risk score, from left to right. The genotypes of the SNVs used to classify patients into high and low risk are shown in rows and color-coded as 0, 1, or 2. SNV, single-nucleotide variation. Nonhem., nonhematological; tox., toxicity.

60 genes reported to be involved in the pharmacodynamic effect of bevacizumab, carboplatin, and taxane.²⁷ It should be noted that genes belonging to the panel used for this study are involved in several signaling pathways, some of which are related to VEGF pathways, others to drug metabolism, and still others to the immune system and inflammation.²⁸

Then, starting from 168 polymorphic DNA sites, we investigated and ranked genetic germline variants associated with drug-induced toxicities, specifically hypertension ≥ 3 , hematological

toxicity ≥ 3 , nonhematological toxicity ≥ 3 , and proteinuria ≥ 1 , using ML techniques. In this context, we have chosen grade 3 as the cutoff point for hypertension, as generally only severe hypertension toxicities are considered the most clinically relevant.¹⁰

Our ML approach has two advantages. First, the SNV-based toxicity model provides reliable risk predictions and can therefore help in clinical decision making. In fact, CV was used to build the model and prioritize the variants, which provides accurate power estimates regardless of sample size.²⁰ The MCC was used to assess

classification performance in our binary classification. This metric outperforms the accuracy measure by preventing overly optimistic results.¹⁷ Second, this technique can be used to infer putative biological pathways underlying the development of toxicity in ovarian cancer patients treated with a combination of carboplatin, taxane and bevacizumab.

Here, a total of 43 SNVs were identified as critical for predicting toxicity, 14 of which are present in both hypertension and nonhematological toxicity and are likely sufficient to explain the cases of hypertension with nonhematological toxicity. These 14 SNVs could improve the prediction of hypertension. In addition, considering all nonhematological toxicities, including hypertension, as a single group increased the performance of the model. This argues for limiting the analysis to this approach without examining the predictability of the individual toxicities considered singularly. The different results obtained in our study compared with previous studies might depend on the different sequencing techniques used (gene panel vs. GWAS) and on the different statistical vs. ML approaches. Our approach allowed us to define SNVs important for hypertension and associated with PDGF and VEGF signaling. Moreover, we found that *ABCB1* rs9282564 was most significant in both hypertension and nonhematological toxicities.^{29,30} Conversely, despite that certain significant SNVs associated with hematological toxicity are associated with PDGF and VEGF signaling, the most significant and prominent SNVs in hematological toxicity are those associated with the biological oxidation pathway, most notably *CYP3A4* rs28371763 and *CYP1B1* rs9341266.^{31,32}

In our study, the most frequent and significant association between SNVs and toxicity was found between coding synonym changes and noncoding variations in the 5' and 3' sides of the untranslated regions (or UTRs) of genes. The regulatory sites required for the modulation of transcription by proteins, microRNAs, and long noncoding RNAs can be altered by genetic variations in the UTRs as well as by variations in the coding regions.³³ In addition, association signals that appear meaningless at first glance may actually be SNVs that are in linkage disequilibrium (LD) with other unobserved SNVs that are likely to be relevant to the event under study.

To further assess the cumulative impact of genetic risk variations, we also developed a polygenic toxicity risk score that allowed us to divide patients into those at high and low risk of toxicity. The polygenic risk scores we created reflect the likelihood of toxicity for patients quite well, as the odds of developing toxicity was higher in high-risk patients than in low-risk patients.

Most of the confirmed SNVs are found in the PDGF and VEGF signaling pathways. VEGF signaling regulates a number of factors that may be associated with systemic hypertension.²⁵ The *FLT1* (*VEGFR1*) rs7993418 G allele enhances the efficiency of *VEGFR1* messenger RNA translation.³² Platelet-derived growth factor (PDGF) stimulates smooth muscle cell migration and proliferation in the pulmonary artery and may contribute to the development of pulmonary arterial hypertension.³⁰ An *in silico* study identified rs2229561 at the 3'-UTR site of the *PDGFRB* gene as a target of microRNA; variations at this site may alter PDGF signaling by increasing expression of the corresponding gene.³⁴

Remodeling of the extracellular matrix by matrix metalloproteinases (MMPs) affects the mechanical properties of conducting vessels. Higher MMP2 expression has been associated with intima and medial thickness of guide vessels.³⁵ *MMP2* rs7201 has been shown to be located in a microRNA-binding region, and the risk C allele acts as a loss-of-function mutation causing higher expression of MMP2.³⁶ Interleukin 18 (IL-18) is capable of altering endothelial function or triggering vascular changes associated with hypertension, either directly or indirectly via oxidative stress pathways and induction of MMPs; clinical studies have consistently found a significant association between blood pressure and IL-18 levels.³⁷ Alterations in serum levels of IL-18 have been linked to the *IL18* promoter polymorphism rs360717.³⁸ In our analysis, the *IL18* coding synonym variant rs549908 was identified as important. Although no data on this site can be found in the literature, rs549908 is linked to rs5744292 ($R^2 = 0.12$, $P \leq 0.0001$), which has been associated with diastolic blood pressure.³⁹

The transmembrane efflux permeability glycoprotein *ABCB1*, also known as multidrug resistance 1 (MDR1), is widely known for its function in the transport of various drugs and other xenobiotics.⁴⁰ In our study, missense C variation at rs9282564 was associated with a low risk of hypertension (OR = 0.11, $P = 0.002$). Cells with genotype CT were shown to have a slightly reduced ability to efflux paclitaxel compared with cells with genotype TT.⁴¹ Accordingly, human *ABCB1* genes may be involved in blood pressure regulation, and several antihypertensive drugs are *ABCB1* substrates; rs9282564 is linked to rs2032582 ($R^2 = 0.10$, $P \leq 0.0001$), which influences the pharmacokinetics of antihypertensive drugs in healthy subjects.⁴² Cytochrome P450 Family 1 Subfamily B Member 1 mediates pathways involved in blood pressure control, migration, proliferation, and hypertrophy of vascular smooth muscle cells; its overexpression causes hypertension and associated pathologies.⁴³ The alternative 3' UTR variant at the site of the *CYP1B1* gene rs162549 has been associated with a 1.5-fold increase in *CYP1B1* expression.⁴⁴

NR1I2 (nuclear receptor subfamily 1 group I member 2) encodes a protein that is a known regulator of *CYP3A4* expression, an enzyme involved in the metabolism of 40–50% of all drugs. The polymorphic locus rs3814057 was selected as a significant candidate for predicting hypertension, and patients with AC and CC have a lower risk of developing hypertension; the C allele rs3814057 is correlated with the G allele rs6785049 ($R^2 = 0.34$, $P \leq 0.0001$). Sunitinib is a multicenter receptor tyrosine kinase inhibitor that acts similarly to bevacizumab by blocking VEGF and PDGF signaling pathways. Patients carrying the rs6785049 GG genotype may have a lower risk of hypertension when treated with sunitinib.⁴⁵

We are fully aware of the limitations of our ML approach. First, the selection of the study population was limited to White women who participated in the clinical trial. In addition, the number of patients involved is relatively small compared with GWAS. Nevertheless, the validation techniques we used are the best solution when dealing with a limited sample size.⁴⁶ Second, our discovery approach was unbiased, as it was a panel-based approach, with the possibility that some genes outside the sequencing panel were missed. However, the fact that we selected only a specific subset of genes related to the

metabolism of the drugs used in the study at least overcame this limitation by deeply photographing the germline mutation landscape. Together with the limited panel of genetic variants, the limited number of patients and the unfeasibility to include different ethnic populations, the application of the proposed polygenic risk score may not be readily transferable to other ethnicities.⁴⁷

The results of this study suggest that ML can be used to develop a decision support tool to predict the risk of toxicity, especially in relation to hypertension. This may help to reduce the number of women affected by genetic toxicities and improve the management of this toxicity. Furthermore, the analytical framework here used is generalizable and can be adapted to outcomes and pathologies that go far beyond the field of oncology and pharmacology.

SUPPORTING INFORMATION

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

ACKNOWLEDGMENTS

We acknowledge Roldano Fossati and Simona Stupia with the Pandora Biobank organization. We acknowledge all the staff help to develop a genotype of the samples.

FUNDING

This research was partly supported by the Italian Association for Cancer Research (AIRC) IG 2016 (ID 18921) and IG 2021 (ID 25932) to S.P., IRCCS-G. Pascale Ricerca Corrente 2022 grant L3/13 from Ministero della Salute to S.P. This work was in part supported by the Italian Ministry of Health-Ricerca Corrente.

CONFLICT OF INTEREST

All authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

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

© 2023 The Authors. *Clinical Pharmacology & Therapeutics* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

[Correction added on 18 Oct 2023, after first online publication: The copyright line was changed.]

- van der Lee, M., Kriek, M., Guchelaar, H.J. & Swen, J.J. Technologies for pharmacogenomics: a review. *Genes (Basel)* **11**, 1–16 (2020).
- Mesko, B. The role of artificial intelligence in precision medicine. *Expert Rev. Precis. Med. Drug Dev.* **2**, 239–241 (2017).
- Bedon, L. *et al.* Machine learning application in a phase I clinical trial allows for the identification of clinical-biomolecular markers

- significantly associated with toxicity. *Clin. Pharmacol. Ther.* **111**, 686–696 (2022).
- Cabasag, C.J. *et al.* Ovarian cancer today and tomorrow: a global assessment by world region and human development index using GLOBOCAN 2020. *Int. J. Cancer* **151**, 1535–1541 (2022).
- Nahshon, C., Barnett-Griness, O., Segev, Y., Schmidt, M., Ostrovsky, L. & Lavie, O. Five-year survival decreases over time in patients with BRCA-mutated ovarian cancer: a systemic review and meta-analysis. *Int. J. Gynecol. Cancer* **32**, 48–54 (2022).
- Perren, T.J. *et al.* A phase 3 trial of bevacizumab in ovarian cancer. *N. Engl. J. Med.* **365**, 2484–2496 (2011).
- Camarda, N., Travers, R., Yang, V.K., London, C. & Jaffe, I.Z. VEGF receptor inhibitor-induced hypertension: emerging mechanisms and clinical implications. *Curr. Oncol. Rep.* **24**, 463–474 (2022).
- Ranpura, V., Pulipati, B., Chu, D., Zhu, X. & Wu, S. Increased risk of high-grade hypertension with bevacizumab in cancer patients: a meta-analysis. *Am. J. Hypertens.* **23**, 460–468 (2010).
- Schneider, B.P. *et al.* Association of vascular endothelial growth factor and vascular endothelial growth factor receptor-2 genetic polymorphisms with outcome in a trial of paclitaxel compared with paclitaxel plus bevacizumab in advanced breast cancer: ECOG 2100. *J. Clin. Oncol.* **26**, 4672–4678 (2008).
- Quintanilha, J.C.F. *et al.* Bevacizumab-induced hypertension and proteinuria: a genome-wide study of more than 1000 patients. *Br. J. Cancer* **126**, 265–274 (2022).
- Daniele, G. *et al.* Bevacizumab, carboplatin, and paclitaxel in the first line treatment of advanced ovarian cancer patients: the phase IV MITO-16A/MaNGO-OV2A study. *Int. J. Gynecol. Cancer* **31**, 875–882 (2021).
- Pignata, S. *et al.* Carboplatin-based doublet plus bevacizumab beyond progression versus carboplatin-based doublet alone in patients with platinum-sensitive ovarian cancer: a randomised, phase 3 trial. *Lancet Oncol.* **22**, 267–276 (2021).
- De Mattia, E. *et al.* Rare genetic variant burden in DPYD predicts severe fluoropyrimidine-related toxicity risk. *Biomed. Pharmacother.* **154**, 113644 (2022).
- Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
- Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
- Kursa, M.B. & Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
- Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
- Chen, T. *et al.* xgboost: Extreme Gradient Boosting <<https://CRAN.R-project.org/package=xgboost>> (2022).
- Kuhn, M. Caret: Classification and Regression Training <<https://CRAN.R-project.org/package=caret>> (2020).
- Vabalas, A., Gowen, E., Poliakov, E. & Casson, A.J. Machine learning algorithm validation with a limited sample size. *PLoS One* **14**, e0224365 (2019).
- Machiela, M.J. & Chanock, S.J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
- Wickham, H. *ggplot2: elegant graphics for data analysis*. Springer-Verlag, New York. ISBN 978-3-319-24277-4. (2016).
- Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
- Behravan, H. *et al.* Machine learning identifies interacting genetic variants contributing to breast cancer risk: a case study in Finnish cases and controls. *Sci. Rep.* **8**, 13149 (2018).
- Li, M. & Kroetz, D.L. Bevacizumab-induced hypertension: clinical presentation and molecular understanding. *Pharmacol. Ther.* **182**, 152–160 (2018).
- Plummer, C. *et al.* Expert recommendations on the management of hypertension in patients with ovarian and cervical cancer receiving bevacizumab in the UK. *Br. J. Cancer* **121**, 109–116 (2019).

27. Whirl-Carrillo, M. *et al.* An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **110**, 563–572 (2021).
28. Khrunin, A. *et al.* Pharmacogenomics of cisplatin-based chemotherapy in ovarian cancer patients of different ethnic origins. *Pharmacogenomics* **13**, 171–178 (2012).
29. Pandey, A.K. *et al.* Mechanisms of VEGF (vascular endothelial growth factor) inhibitor-associated hypertension and vascular disease. *Hypertension* **71**, e1–e8 (2018).
30. Solinc, J., Ribot, J., Soubrier, F., Pavoine, C., Dierick, F. & Nadaud, S. The platelet-derived growth factor pathway in pulmonary arterial hypertension: still an interesting target? *Life (Basel)* **12**, 1–19 (2022).
31. Frederiks, C.N., Lam, S.W., Guchelaar, H.J. & Boven, E. Genetic polymorphisms and paclitaxel- or docetaxel-induced toxicities: a systematic review. *Cancer Treat. Rev.* **41**, 935–950 (2015).
32. Lambrechts, D. *et al.* VEGF pathway genetic variants as biomarkers of treatment outcome with bevacizumab: an analysis of data from the AVITA and AVOREN randomised trials. *Lancet Oncol.* **13**, 724–733 (2012).
33. Steri, M., Idda, M.L., Whalen, M.B. & Orrù, V. Genetic variants in mRNA untranslated regions. *Wiley Interdiscip. Rev. RNA* **9**, e1474 (2018).
34. Manikandan, M. & Munirajan, A.K. Single nucleotide polymorphisms in microRNA binding sites of oncogenes: implications in cancer and pharmacogenomics. *OMICS* **18**, 142–154 (2014).
35. Lacerda, L., de Faria, A.P., Fontana, V., Moreno, H. & Sandrim, V. Role of MMP-2 and MMP-9 in resistance to drug therapy in patients with resistant hypertension. *Arq. Bras. Cardiol.* **105**, 168–175 (2015).
36. Tsai, E.M. *et al.* A microRNA-520 mirSNP at the MMP2 gene influences susceptibility to endometriosis in Chinese women. *J. Hum. Genet.* **58**, 202–209 (2013).
37. Rabkin, S.W. The role of interleukin 18 in the pathogenesis of hypertension-induced vascular disease. *Nat. Clin. Pract. Cardiovasc. Med.* **6**, 192–199 (2009).
38. Al-Khateeb, G.M. *et al.* Analysis of interleukin-18 promoter polymorphisms and changes in interleukin-18 serum levels underscores the involvement of interleukin-18 in recurrent spontaneous miscarriage. *Fertil. Steril.* **96**, 921–926 (2011).
39. Kostis, W.J. *et al.* Relationships between selected gene polymorphisms and blood pressure sensitivity to weight loss in elderly persons with hypertension. *Hypertension* **61**, 857–863 (2013).
40. Haque, A. *et al.* MDR1 gene polymorphisms and its association with expression as a clinical relevance in terms of response to chemotherapy and prognosis in ovarian cancer. *Front. Genet.* **11**, 516 (2020).
41. Kimchi-Sarfaty, C., Gribar, J.J. & Gottesman, M.M. Functional characterization of coding polymorphisms in the human MDR1 gene using a vaccinia virus expression system. *Mol. Pharmacol.* **62**, 1–6 (2002).
42. Kim, K.A., Park, P.W. & Park, J.Y. Effect of ABCB1 (MDR1) haplotypes derived from G2677T/C3435T on the pharmacokinetics of amlodipine in healthy subjects. *Br. J. Clin. Pharmacol.* **63**, 53–58 (2007).
43. Malik, K.U. *et al.* Contribution of cytochrome P450 1B1 to hypertension and associated pathophysiology: a novel target for antihypertensive agents. *Prostaglandins Other Lipid Mediat.* **98**, 69–74 (2012).
44. Serre, D. *et al.* Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet.* **4**, e1000006 (2008).
45. Narjoz, C. *et al.* Role of the lean body mass and of pharmacogenetic variants on the pharmacokinetics and pharmacodynamics of sunitinib in cancer patients. *Invest. New Drugs* **33**, 257–268 (2015).
46. Tsamardinos, I. Don't lose samples to estimation. *Patterns (NY)* **3**, 100612 (2022).
47. Araújo, D.S. & Wheeler, H.E. Genetic and environmental variation impact transferability of polygenic risk scores. *Cell Rep. Med.* **3**, 100687 (2022).