Title

A Pharmacogenetic Prediction Model of Progression-Free Survival in Breast Cancer using Genome-Wide Genotyping Data from CALGB 40502 (Alliance)

Permalink

https://escholarship.org/uc/item/6n1170mm

Journal

Clinical Pharmacology & Therapeutics, 105(3)

ISSN

0009-9236

Authors

Rashkin, Sara R

Chua, Katherina C

Ho, Carol

et al.

Publication Date

2019-03-01

DOI

10.1002/cpt.1241

Peer reviewed

A Pharmacogenetic Prediction Model of Progression-Free Survival in Breast Cancer using Genome-Wide Genotyping Data from CALGB 40502 (Alliance)

Sara R. Rashkin[1], Katherina C. Chua[2], Carol Ho[2], Flora Mulkey[3], Chen Jiang[3], Tasei Mushiroda[4], Michiaki Kubo[4], Paula N. Friedman[5], Hope S. Rugo[6], Howard L. McLeod[7], Mark J. Ratain[8], Francisco Castillos[9], Michael Naughton[10], Beth Overmoyer[11], Deborah Toppmeyer[12], John S. Witte[1], Kouros Owzar[3,13] and Deanna L. Kroetz[2]

Genome-wide genotyping data are increasingly available for pharmacogenetic association studies, but application of these data for development of prediction models is limited. Prediction methods, such as elastic net regularization, have recently been applied to genetic studies but only limitedly to pharmacogenetic outcomes. An elastic net was applied to a pharmacogenetic study of progression-free survival (PFS) of 468 patients with advanced breast cancer in a clinical trial of paclitaxel, nab-paclitaxel, and ixabepilone. A final model included 13 single nucleotide polymorphisms (SNPs) in addition to clinical covariates (prior taxane status, hormone receptor status, disease-free interval, and presence of visceral metastases) with an area under the curve (AUC) integrated over time of 0.81, an increase compared to an AUC of 0.64 for a model with clinical covariates alone. This model may be of value in predicting PFS with microtubule targeting agents and may inform reverse translational studies to understand differential response to these drugs.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Pharmacogenetic studies using genome-wide genotyping data typically use association testing to identify SNPs that predict treatment response or toxicity. Models developed for accurately predicting a pharmacogenetic outcome that include factors regardless of statistical association have not been widely applied to pharmacogenetic studies.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ This study used genome-wide genotyping data to develop a predictive model of PFS in patients with advanced breast cancer treated with microtubule targeting agents.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ A model that includes both clinical and genetic variables improves the prediction of PFS with microtubule targeting agents compared to the use of clinical variables alone.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ This or other similar regularized regression approaches can be applied to other genome-wide genotype datasets to develop models to predict treatment outcome based on genetic profile.

Applications of genome-wide data beyond typical variant-by-variant association testing have been gaining popularity for pharmacogenetic phenotypes. Examples include gene-based or region-based tests,[1,2] heritability analyses,[3,4] transcriptome-based analyses,[5–7] and polygenic risk scores.[8–10] Another method is meta-analysis, combining summary statistics of multiple studies,

[1]Department of Biostatistics and Epidemiology, University of California San Francisco, San Francisco, California, USA; [2]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, USA; [3]Alliance Statistics and Data Center, Duke University, Durham, North Carolina, USA; [4]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan; [5]Department of Medicine, Northwestern University, Chicago, Illinois, USA; [6]Department of Medicine, University of California San Francisco, San Francisco, California, USA; [7]DeBartolo Family Personalized Medicine Institute, Moffitt Cancer Center, Tampa, Florida, USA; [8]Department of Medicine, The University of Chicago, Chicago, Illinois, USA; [9]Nash General Hospital, Rocky, North Carolina, USA; [10]Washington University School of Medicine, St. Louis, Missouri, USA; [11]Dana-Farber/Partners Cancer Care, Boston, Massachusetts, USA; [12]Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey, USA; [13]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina, USA. Correspondence: Deanna L. Kroetz (deanna.kroetz@ucsf.edu)

Received 21 May 2018; accepted 30 August 2018; advance online publication 1 November 2018. doi:10.1002/cpt.1241

but this can be difficult for pharmacogenetic phenotypes, in which different drugs, drug combinations, or patient populations are under study.[11–13] Genome-wide data for pharmacogenetic phenotypes have not been extensively used for developing models that will predict the treatment response, often involving many single nucleotide polymorphisms (SNPs) simultaneously, regardless of their statistical association with the outcome.[14,15]

In recent years, there have been an increasing number of genetic analyses using regularized regression methods, such as elastic net or least absolute shrinkage and selection operator (LASSO) regression,[14–17] Support Vector Machine algorithms,[18] and random forest regression[19] for developing a model predicting outcome using multiple SNPs. Regularized regression methods have been applied to cell line studies for drug response,[20–23] and there have been recent applications of such analyses in the prediction of duloxetine response.[24] These models can improve prediction error and reduce overfitting as well as perform variable selection.[18,25,26] Although this type of modeling will not allow for inferences to be made regarding the strength of association between a SNP and the outcome of interest, a model can be developed that can aid in prediction. Here, we apply an elastic net model to a study focused on predicting progression-free survival (PFS) in patients with advanced breast cancer in a clinical trial of microtubule targeting agents.[27]

RESULTS

The clinical characteristics of the 468 genetic European patients in the primary analysis closely resembled the original trial population

Table 1 Patient characteristics of clinical and pharmacogenetic cohorts

	Paclitaxel[a]			Nab-paclitaxel			Ixabepilone			Total		
	Full cohort	EUR[c] patients	non-EUR[d] patients	Full cohort	EUR patients	Non-EUR patients	Full cohort	EUR patients	Non-EUR patients	Full cohort	EUR Patients	Non-EUR patients
Total	283	160	51	271	156	47	245	152	32	799	468	130
	(0.35)[b]	(0.34)[b]	(0.39)[b]	(0.34)[b]	(0.33)[b]	(0.36)[b]	(0.31)[b]	(0.32)[b]	(0.25)[b]			
Age, years												
20–49	69	36	14	76	33	18	73	47	10	218	116	42
	(0.24)	(0.23)	(0.27)	(0.28)	(0.21)	(0.38)	(0.30)	(0.31)	(0.31)	(0.27)	(0.25)	(0.32)
50–69	183	106	31	163	104	26	154	92	21	500	300	78
	(0.65)	(0.66)	(0.61)	(0.60)	(0.67)	(0.55)	(0.63)	(0.61)	(0.66)	(0.63)	(0.64)	(0.60)
70–80+	31	18	6	32	19	3	18	13	1	81	59	10
	(0.11)	(0.11)	(0.12)	(0.12)	(0.12)	(0.06)	(0.07)	(0.09)	(0.03)	(0.10)	(0.13)	(0.08)
Taxane												
No	158	91	26	151	81	31	138	91	14	447	263	71
	(0.56)	(0.57)	(0.51)	(0.56)	(0.52)	(0.66)	(0.56)	(0.60)	(0.44)	(0.56)	(0.56)	(0.55)
Yes	125	69	25	120	75	16	107	61	18	352	205	59
	(0.44)	(0.43)	(0.49)	(0.44)	(0.48)	(0.34)	(0.44)	(0.40)	(0.56)	(0.44)	(0.44)	(0.45)
Any visceral metastases												
No	55	34	8	60	30	16	42	31	2	157	95	26
	(0.19)	(0.21)	(0.16)	(0.22)	(0.19)	(0.34)	(0.17)	(0.20)	(0.06)	(0.20)	(0.20)	(0.20)
Yes	217	126	43	205	126	31	199	121	30	621	373	104
	(0.77)	(0.79)	(0.84)	(0.76)	(0.81)	(0.66)	(0.81)	(0.80)	(0.94)	(0.78)	(0.80)	(0.80)
Disease-free interval												
≤ 2 years	121	67	22	120	64	27	92	55	14	333	186	63
	(0.43)	(0.42)	(0.43)	(0.44)	(0.41)	(0.57)	(0.38)	(0.36)	(0.44)	(0.42)	(0.40)	(0.48)
> 2 years	154	93	29	147	92	20	149	97	18	450	282	67
	(0.54)	(0.58)	(0.57)	(0.54)	(0.59)	(0.43)	(0.61)	(0.64)	(0.56)	(0.56)	(0.60)	(0.52)
Hormone receptor status												
Both negative	82	47	16	76	43	16	67	35	10	255	125	42
	(0.29)	(0.29)	(0.31)	(0.28)	(0.28)	(0.34)	(0.27)	(0.23)	(0.31)	(0.32)	(0.27)	(0.32)
Either positive	201	113	35	195	113	31	178	117	22	574	343	88
	(0.71)	(0.71)	(0.69)	(0.72)	(0.72)	(0.66)	(0.73)	(0.77)	(0.69)	(0.72)	(0.73)	(0.68)

EUR, European.

[a] Data reported as number of subjects (fraction of arm). [b] Fractions of total samples rather than of arm. [c] Genetic European samples that were genotyped and passed quality control filters. [d] Samples that were genotyped and passed quality control filters but were not genetically European.

(see **Table 1**). The median follow-up time in this European discovery sample was 32.2 months, and there were a total of 406 observed disease progression events or deaths. A standard genome-wide association study (GWAS) was performed initially, using a Cox proportional hazards model to assess the effect of each SNP on PFS, stratified by treatment arm, and controlling for previously identified clinical covariates.[27] For a significance threshold of 5E-8, no significant associations are detected. Therefore, an approach leveraging the strength of multiple SNPs collectively was applied to develop a prediction model.

Across the 10 models from the initial elastic net analyses, a median of 312 SNPs were retained (range: 263–340). The hazard ratios (HRs) ranged from 0.72–1.36, but the majority of SNPs remaining in each model had HRs between 0.95 and 1.05 (see **Figure 1**). Applying each model to the corresponding test set, the median area under the curve (AUC) collapsed across time is 0.78 (range: 0.75–0.81). A model with only clinical covariates results in a median AUC of 0.62 (range: 0.60–0.65). Comparing across models, only a few SNPs were selected in all training sets. There are 832 SNPs that were selected in just a single model, but only 13 SNPs were selected in all 10 (see **Figure 2**). Although some SNPs with larger effect sizes are selected only once or twice, these are rare exceptions. In general, the more often a SNP is selected, the less likely it is to be driven by only a few samples and the more likely it is to have a large effect size (see **Figure 2**).

A final model for prediction of PFS was created among all genetic European samples with the clinical covariates and only the 13 SNPs selected by all 10 initial models (see **Table 2**). For a validation set of 130 non-European samples, AUC was calculated across varying time points. The median follow-up time in this non-European validation sample was 24.1 months, and there were a total of 113 observed disease progression events or deaths. A model with only clinical covariates was compared to a model with clinical covariates and the 13 SNPs (see **Figure 3**). Across the time period from 0–50 months, the model with only clinical covariates had an integrated AUC of 0.64 with a 95% confidence interval (CI) of 0.60–0.67. Addition of the 13 SNPs to the model yielded an integrated AUC of 0.81 (95% CI: 0.77–0.83). Across 10,000 simulations of equivalent models with 13 randomly selected SNPs, only 2,934 resulted in models retaining SNP covariates, none of which performed as well as our final model (median integrated AUC: 0.64; range: 0.63–0.70).

Because the elastic net was used to construct our final model, many model diagnostics are not feasible. Using a standard Cox proportional hazard regression in lieu of the actual final model leads to similar effect sizes for all covariates. The proportionality assumption for all variables was tested and verified. Additionally, a deviance-like test was performed, and there was a significant ($P < 2.2E\text{-}16$) reduction in model log-likelihood with the addition of the 13 SNPs compared with the model with clinical covariates alone.

DISCUSSION

Here, an elastic net approach was applied to develop a prediction model of PFS in a clinical trial of paclitaxel, nab-paclitaxel, and ixabepilone in 468 genotyped patients with advanced breast cancer with European ancestry. The model was then validated in 130 genotyped patients with advanced breast cancer with non-European ancestry. Elastic net models reduce overfitting and perform variable selection, resulting in improved prediction error. Including genetic covariates improves the predictive ability (AUC = 0.81) compared with a model using only clinical covariates (AUC = 0.64). Additionally, the SNP effect sizes are comparable to or greater than those of the clinical covariates. The final model also validates the importance of controlling for multiple SNPs simultaneously, as it allows for inclusion of SNPs that are associated with both improved and worsened clinical outcome (see **Table 2**).

One limitation of this approach includes computational resources. Although it is theoretically possible to include all SNPs genome-wide in the elastic net model, this is computationally burdensome. In order to reduce the computational load, only the top 1,000 SNPs from a univariate analysis were used in the elastic net

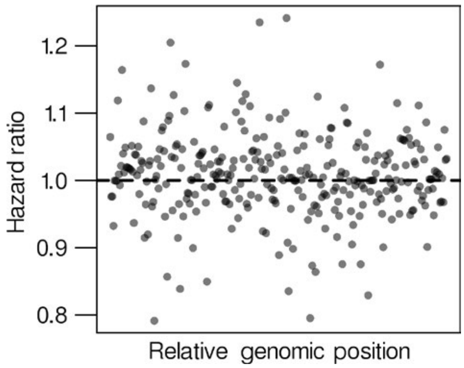

Figure 1 Distribution of hazard ratios for single nucleotide polymorphisms remaining in the model following elastic net regularization. The representative data are for one example fold from the cross-validation process. For ease of plotting, data were ordered by genomic position, but actual positions are not represented.

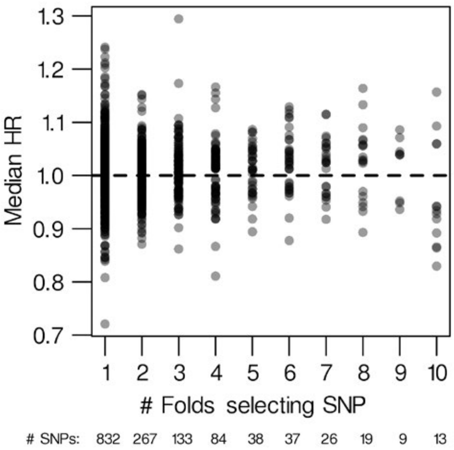

Figure 2 Comparison of median single nucleotide polymorphism (SNP) effect size by the frequency of SNP selection. The median hazard ratios (HRs) for each SNP selected in the models from the 10-fold cross-validations are plotted as a function of the number of folds selecting each SNP. The SNPs selected in all 10 folds were included in the final model.

Table 2 Clinical variables and SNPs included in the final model for progression-free survival

	Elastic net		MAF[a]				
Variable	Beta	HR	EUR[b]	Non-EUR[c]	Chromosome	Position	RefSeq annotation
Nab-paclitaxel[d]	0.19	1.21	—	—	—	—	—
Ixabepilone[d]	0.51	1.67	—	—	—	—	—
Disease-free interval[e]	−0.24	0.78	—	—	—	—	—
Visceral metastases[f]	0.28	1.33	—	—	—	—	—
Prior taxane[g]	0.50	1.65	—	—	—	—	—
Hormone receptor status[h]	−0.34	0.71	—	—	—	—	—
rs10490308	−0.38	0.68	0.12	0.096	2	75482375	*TACR1*
rs10516451	−0.32	0.72	0.40	0.25	4	100626549	*MTTP*
rs12440889	−0.41	0.66	0.12	0.096	15	38506692	*SPRED1*
rs12953016	0.31	1.37	0.14	0.49	17	52000308	*KIF2B*
rs17590916	−0.28	0.76	0.38	0.30	6	91344010	*MAP3K7*
rs419463	−0.26	0.77	0.32	0.16	3	128741642	*CCDC48*
rs4944458	−0.53	0.59	0.20	0.33	11	72088912	*CLPB*
rs5008836	0.32	1.37	0.38	0.51	1	79658861	*ELTD1*
rs540407	0.28	1.32	0.48	0.45	9	135343839	*C9orf171*
rs7812482	−0.27	0.77	0.37	0.18	8	76013093	*CRISPLD1*
rs897102	−0.31	0.74	0.17	0.24	11	8331494	*LMO1*
rs9859426	0.40	1.49	0.067	0.050	3	161523859	*OTOL1*
exm1431132	0.39	1.48	0.30	0.32	19	12774208	*MAN2B1*

EUR, European; HR, hazard ratio; MAF, minor allele frequency; SNP, single nucleotide polymorphism.
[a]Minor allele frequency = observed frequency of least common allele in sample. [b]Genetic European samples that were genotyped and passed quality control filters (n = 468). [c]Samples that were genotyped and passed quality control filters but were not genetically European (n = 130). [d]Compared to paclitaxel. [e]Greater than 2 years, compared to ≤ 2 years. [f]Presence of visceral metastases compared to no visceral metastases. [g]Prior taxane compared to no prior taxane. [h]Either positive compared to both negative.

model. Whereas this ensures that the SNPs with the smallest P values are included in the elastic net model, SNPs with larger P values but important for prediction may be omitted from the final model. Screening based on univariate effect size rather than P value was considered, but the limited range of effect sizes led to more SNPs with lower effect sizes and fewer SNPs selected by all initial models. We also investigated the ideal number of SNPs to use in the elastic net and concluded that, for our data, 1,000 SNPs is optimal. Inclusion of < 1,000 SNPs led to a model that did not improve integrated AUC as much as the final model presented above, and > 1,000 SNPs had a negative effect on results, likely due to the addition of noise, hindering model optimization. Additionally, some of the SNPs remaining in the model have different minor allele frequencies in the European training and non-European validation sets (see **Table 2**), which could potentially affect the prediction accuracy. However, the increase in AUC observed with the non-European samples suggests this prediction model is robust to population differences and genotype frequencies. Caveats of this approach are that elastic net methods require complete data, and, for the Cox proportional hazards implementation, it is not possible to stratify on a predictor. Finally, application of this method to other data may require optimization of the elastic net constraint parameter α.

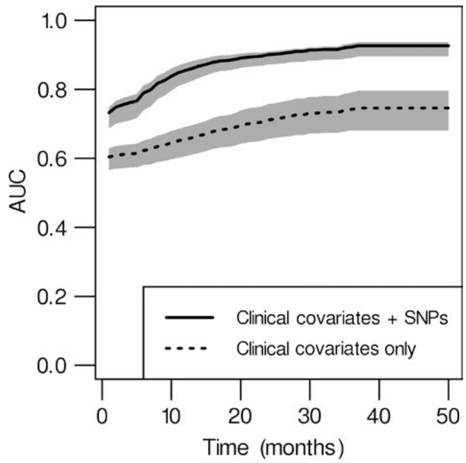

Figure 3 Model prediction of progression-free survival in non-European samples. The data represent the time-dependent area under the curve (AUC) and 95% confidence interval by month for the model with genetic information and clinical covariates compared to the model with only clinical covariates. SNP, single nucleotide polymorphism.

The Cancer and Leukemia Group B (CALGB) 40502 tested whether newer microtubule-targeting agents showed improved PFS and/or toxicity in patients with locally recurrent or metastatic breast cancer. The conclusion was that ixabepilone was inferior and nab-paclitaxel was not superior, with a trend toward inferiority, compared with paclitaxel. Toxicity was also increased in experimental arms. Because the two alternative therapies will not replace the standard-of-care paclitaxel, the prediction model developed here will have limited clinical utility in making treatment decisions between these three options. However, there may be some utility in making a prediction estimate for PFS in this setting when treated with paclitaxel.

Pharmacogenetic findings can also inform regarding the underlying mechanisms of drug response. Of the 13 SNPs remaining in the final model, several are annotated to genes that are plausibly related to the pharmacology of microtubule targeting agents, including a kinesin family member and genes involved in mitogen-activated protein kinase signaling (see **Table 2**). The focus in prediction modeling is not on determining direct associations between a single SNP and the outcome of interest but, rather, on the collective predictive ability of multiple SNPs simultaneously. However, these genomic regions may be good candidates for further exploration in clinical and molecular studies. Whether these variants or genomic regions might be interrogated for their contribution to the poor response in the treatment arms of CALGB 40502 would require further study. These same variables could also be the focus of reverse-translational studies to understand microtubule-targeting agents.

METHODS

Patients and drug response phenotype

Data collection was conducted by the Alliance Statistics and Data Center. Genotype and phenotype data were collected from CALGB 40502 (NCT00785291), a phase III randomized three-arm study comparing nanoparticle albumin-bound (nab) paclitaxel or ixabepilone to paclitaxel (all given with bevacizumab) as first-line therapy for patients with advanced breast cancer. Trial design and outcome have been previously described.[27] CALGB is now part of the Alliance for Clinical Trials in Oncology. PFS was the primary end point. The analyses described here tested whether incorporating both clinical risk variables and genetic risk alleles provides an improved prediction model for PFS.

Genotyping and quality control

Genotyping and analysis of existing samples was approved by the National Cancer Institute Adult Central Institutional Review Board and by the Institutional Review Board at the University of California San Francisco. All participants provided written informed consent for pharmacogenetic sample procurement and analysis, and all studies were conducted in accordance with recognized ethical guidelines. From the 799 patients randomized to the parent study, 633 consented to the pharmacogenetic substudy and had DNA available for genotyping (see **Figure 4**). DNA samples were genotyped using the Illumina HumanOmniExpressExome-8 BeadChip at the Riken Center for Genomic Medicine, interrogating 964,055 SNPs with coverage of common variants and additional exonic content. All samples passed a low call rate filter (> 0.99). Two unintended duplicates were excluded. An X chromosome heterozygosity estimation identified three genetic males that were removed.

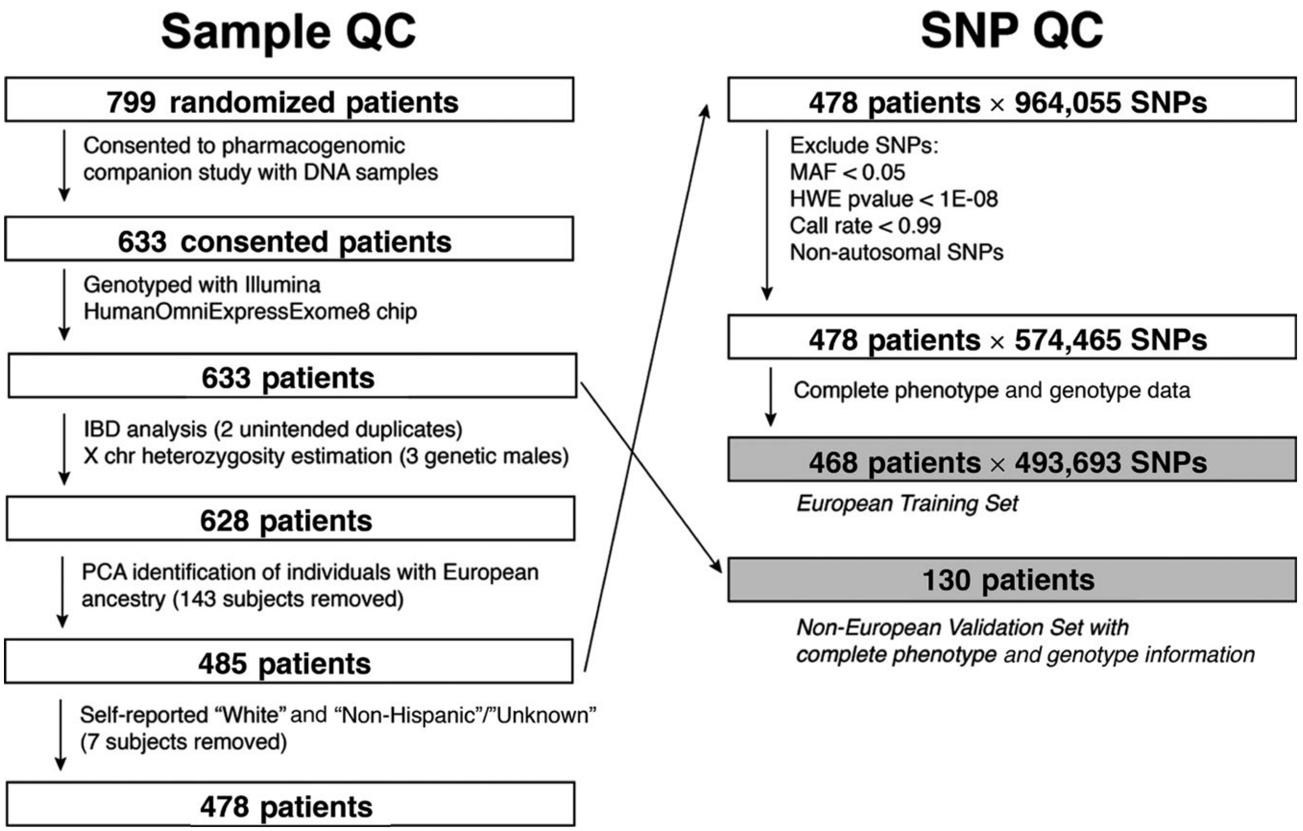
Sample QC
799 randomized patients
Consented to pharmacogenomic companion study with DNA samples
633 consented patients
Genotyped with Illumina HumanOmniExpressExome8 chip
633 patients
IBD analysis (2 unintended duplicates)
X chr heterozygosity estimation (3 genetic males)
628 patients
PCA identification of individuals with European ancestry (143 subjects removed)
485 patients
Self-reported "White" and "Non-Hispanic"/"Unknown" (7 subjects removed)
478 patients
SNP QC
478 patients × 964,055 SNPs
Exclude SNPs:
MAF < 0.05
HWE pvalue < 1E-08
Call rate < 0.99
Non-autosomal SNPs
478 patients × 574,465 SNPs
Complete phenotype and genotype data
468 patients × 493,693 SNPs
European Training Set
130 patients
Non-European Validation Set with complete phenotype and genotype information

Figure 4 Quality control (QC) flowchart. The stepwise protocol for quality control analysis for samples and genotype data are presented. HWE, Hardy-Weinberg Equilibrium; IBD, identity-by-descent; MAF, minor allele frequency; PCA, principal component analysis; SNP, single nucleotide polymorphism.

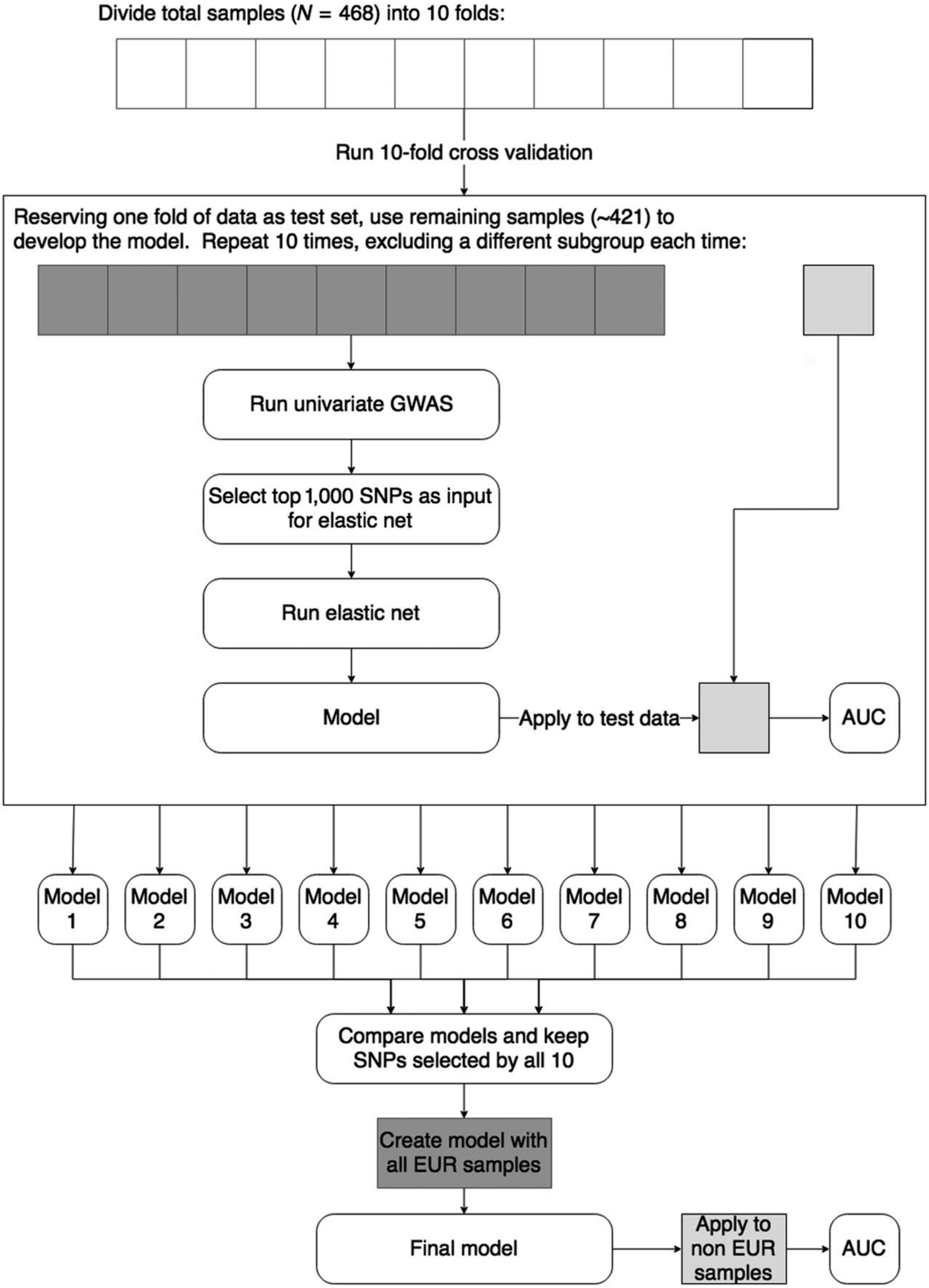
Divide total samples (*N* = 468) into 10 folds:
Run 10-fold cross validation
Reserving one fold of data as test set, use remaining samples (~421) to develop the model. Repeat 10 times, excluding a different subgroup each time:
Run univariate GWAS
Select top 1,000 SNPs as input for elastic net
Run elastic net
Model
Apply to test data
AUC
Model 1
Model 2
Model 3
Model 4
Model 5
Model 6
Model 7
Model 8
Model 9
Model 10
Compare models and keep SNPs selected by all 10
Create model with all EUR samples
Final model
Apply to non EUR samples
AUC

Figure 5 Overview of the statistical analysis. A 10-fold cross-validation in European (EUR) samples was used to create a final prediction model that was then tested in non-European samples. AUC, area under the curve; GWAS, genome-wide association study; SNP, single nucleotide polymorphism.

To avoid potential population stratification, principal component analysis (PCA) was performed using genotypes of all 628 patients to determine genetic ancestry with the GenABEL R package.[28] A total of 485 patients of European genetic ancestry were identified and confirmed with a second PCA using the EIGENSTRAT method.[29] Samples were excluded if any of the first three principal component vectors were outside two SDs of the mean values for patients self-reporting "White" race and "Non-Hispanic"/"Unknown" ethnicity, resulting in 478 samples for the primary analysis. SNPs were excluded using the following quality-control filters: minor allele frequency (MAF) < 0.05, deviation from Hardy-Weinberg Equilibrium with P value < 1E-8, call rate < 0.99, and nonautosomal, leaving 574,465 SNPs for analysis.

The analysis was further limited to the 468 genetic European patients who had complete data for PFS and all previously identified clinical covariates in the primary analysis from the clinical trial (prior taxane status, hormone receptor status, disease-free interval, and presence of visceral metastases).[27] PFS was defined as the time from registration until the date of the first disease progression or death from any cause, censoring at 50 months post-treatment in absence of an event or at the time of treatment discontinuation for any cause other than progression. To maintain complete data across all loci as required by the elastic net, SNPs with incomplete data were excluded, leaving 493,693 SNPs for analysis. An additional 130 non-European patients with complete data for the final model were reserved to evaluate model prediction.

Statistical analysis

To reduce the number of SNPs remaining in the model by chance alone, 10-fold cross-validation was performed to create a final prediction model (see **Figure 5**). Ten initial models were created by subdividing the samples into 10 groups of approximately equal size. For each model, one group of samples was reserved for validation, so that each sample was used in exactly one test set. Each model was developed with the remaining samples as the training set.

To reduce the dimensionality in the elastic net regularization, the 1,000 SNPs showing the strongest association with the trait were preselected for each model as follows. A variant-by-variant GWAS, using Cox proportional hazards regression with PFS as the outcome of interest and stratified by treatment arm, was performed on the training data. Regardless of whether a SNP attained genome-wide significance, the 1,000 SNPs with the smallest P values were retained.

For each training set, an elastic net Cox proportional hazards model was applied to the top 1,000 SNPs along with previously identified clinical covariates.[27] Both variable selection and regularization to enhance prediction accuracy was performed via elastic net regularization using the glmnet R package.[30,31] The elastic net model has two tuning parameters, α and λ.[31] An $\alpha = 1$ corresponds to the LASSO penalty and an $\alpha = 0$ is the ridge penalty. To balance equally between ridge and LASSO regressions—and both improve prediction by shrinking estimates to reduce overfitting and perform variable selection—a constraint parameter of $\alpha = 0.5$ was used. Cross-validation determined the optimal value of λ, using a pathwise coordinate descent. Solutions for $\hat{\beta}$ were computed for a decreasing sequence of λ, starting with the smallest value for which all $\hat{\beta} = 0$. To account for the randomness in the cross-validation, the model was run 100 times, and the error curves were averaged for each value of λ. The value of λ with the smallest mean cross-validation error was selected as the optimal λ. The AUC of each initial model was calculated using the reserved test data using the Song and Zhou estimator in the survAUC R package.[32]

The models created across all 10 training sets were compared, and SNPs remaining in all 10 models were retained. These SNPs were used to create a final model using all genetic European samples. The AUC of this model was calculated using 130 non-European samples from patients with complete data across all covariates using the same estimator described above. A bootstrap method was used to estimate a 95% CI for AUC. All analyses were conducted at the University of California, San Francisco.

FUNDING

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Numbers R01CA1922156 (D.L.K.), U10CA180821, U10CA180882, and U24CA19617 (to the Alliance for Clinical Trials in Oncology), U10CA180833, U10CA180836, and U10CA180857, U10CA180867. This research was also supported in part by funds from The Breast Cancer Research Foundation. S.R.R. was supported by R25 CA112355 and K.C. was supported by T32 GM007175. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CONFLICT OF INTEREST/DISCLOSURE

The parent study in which phenotype data and biospecimens were collected was supported in part by Celgene. Deborah Toppmeyer is a consultant for Merck and has a family member employed by Novartis. John Witte is a consultant for Pfizer and Navigate. Deanna Kroetz is a consultant for Genentech. Michael Naughton is a speaker for Genentech and Celgene.

AUTHOR CONTRIBUTIONS

S.R.R., K.C.C., F.M., P.N.F., H.L.M., M.J.R., J.S.W., K.O., and D.L.K. wrote the manuscript. S.R.R., J.S.W., K.O., and D.L.K. designed the research. S.R.R., K.C.C., C.H., F.M., C.J., T.M., M.K., H.S.R., F.C., M.N., B.O., and D.T. performed the research. S.R.R., K.C.C., J.S.W., K.O., and D.L.K. analyzed data.

1. Daneshjou, R. *et al.* Genetic variant in folate homeostasis is associated with lower warfarin dose in African Americans. *Blood* **124**, 2298–2305 (2014).
2. Mak, A.C. *et al.* Whole genome sequencing of pharmacogenetic drug response in racially diverse children with asthma. *Am. J. Respir. Crit. Care Med.* **197**, 1552–1564 (2018).
3. Chhibber, A. *et al.* Genomic architecture of pharmacological efficacy and adverse events. *Pharmacogenomics* **15**, 2025–2048 (2014).
4. McGeachie, M.J. *et al.* Polygenic heritability estimates in pharmacogenetics: focus on asthma and related phenotypes. *Pharmacogenet. Genomics* **23**, 324–328 (2013).
5. Dolan, M.E. *et al.* Clinical and genome-wide analysis of cisplatin-induced peripheral neuropathy in survivors of adult-onset cancer. *Clin. Cancer Res.* **23**, 5757–5768 (2017).
6. Clarelli, F. *et al.* Pharmacogenetic study of long-term response to interferon-β treatment in multiple sclerosis. *Pharmacogenomics J.* **17**, 84–91 (2017).
7. Maranville, J.C., Baxter, S.S., Witonsky, D.B., Chase, M.A. & Di Rienzo, A. Genetic mapping with multiple levels of phenotypic information reveals determinants of lymphocyte glucocorticoid sensitivity. *Am. J. Hum. Genet.* **93**, 735–743 (2013).
8. Natarajan, P. *et al.* Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* **135**, 2091–2101 (2017).
9. Chung, S. *et al.* A genome-wide association study of chemotherapy-induced alopecia in breast cancer patients. *Breast Cancer Res.* **15**, R81 (2013).
10. Low, S.-K. *et al.* Genome-wide association study of chemotherapeutic agent-induced severe neutropenia/leucopenia for patients in Biobank Japan. *Cancer Sci.* **104**, 1074–1082 (2013).
11. Postmus, I. *et al.* Meta-analysis of genome-wide association studies of HDL cholesterol response to statins. *J. Med. Genet.* **53**, 835–845 (2016).
12. Mahmoudpour, S.H. *et al.* Meta-analysis of genome-wide association studies on the intolerance of angiotensin-converting enzyme inhibitors. *Pharmacogenet. Genomics* **27**, 112–119 (2017).

13. Gong, Y. *et al.* Pharmacogenomic genome-wide meta-analysis of blood pressure response to β-blockers in hypertensive African Americans. *Hypertension* **67**, 556–563 (2016).
14. Kooperberg, C., LeBlanc, M. & Obenchain, V. Risk prediction using genome-wide association studies. *Genet. Epidemiol.* **34**, 643–652 (2010).
15. Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C. & Sölkner, J. Evaluation of the Lasso and the elastic net in genome-wide association studies. *Front. Genet.* **4**, 270 (2013).
16. Cho, S. *et al.* Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Ann. Hum. Genet.* **74**, 416–428 (2010).
17. Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E. & Lange, K. Genome-wide association analysis by Lasso penalized logistic regression. *Bioinformatics* **25**, 714–721 (2009).
18. Wei, Z. *et al.* From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet.* **5**, e1000678 (2009).
19. Lee, S. *et al.* Machine learning on a genome-wide association study to predict late genitourinary toxicity after prostate radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **101**, 128–135 (2018).
20. Azuaje, F. Computational models for predicting drug responses in cancer research. *Brief. Bioinform.* **18**, 820–829 (2017).
21. Neto, E.C., Jang, I.S., Friend, S.H. & Margolin, A.A. The stream algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications to pharmacogenomic prediction of cancer cell line sensitivity. *Pac. Symp. Biocomput.* 27–38 (2014).
22. Zhang, N. *et al.* Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput. Biol.* **11**, e1004498 (2015).
23. Sokolov, A., Carlin, D.E., Paull, E.O., Baertsch, R. & Stuart, J.M. Pathway-based genomics prediction using generalized elastic net. *PLoS Comput. Biol.* **12**, e1004790 (2016).
24. Maciukiewicz, M. *et al.* GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. *J. Psychiatr. Res.* **99**, 62–68 (2018).
25. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
26. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
27. Rugo, H.S. *et al.* Randomized phase III trial of paclitaxel once per week compared with nanoparticle albumin-bound nab-paclitaxel once per week or ixabepilone with bevacizumab as first-line chemotherapy for locally recurrent or metastatic breast cancer: CALGB 40502/NCCTG N063H (alliance). *J. Clin. Oncol.* **33**, 2361–2369 (2015).
28. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
29. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904 (2006).
30. Friedman, J.H., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
31. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).
32. Song, X. & Zhou, X.-H. A semiparametric approach for the covariate specific ROC curve with survival outcome. *Stat. Sin.* **18**, 947–965 (2008).