pISSN 1598-2998, eISSN 2005-9256

Cancer Res Treat. 2019;51(2):672-684

https://doi.org/10.4143/crt.2018.137

Original Article

Open Access

Prediction of Acquired Taxane Resistance Using a Personalized Pathway-Based Machine Learning Method

Young Rae Kim, MD
Dongha Kim, MS
Sung Young Kim, MD, PhD

Department of Biochemistry, Konkuk University School of Medicine, Seoul, Korea

Correspondence: Sung Young Kim MD, PhD
Department of Biochemistry,
Konkuk University School of Medicine,
120 Neungdong-ro, Gwangjin-gu,
Seoul 05029, Korea
Tel: 82-2-2049-6060
Fax: 82-2-2049-6192
E-mail: palelamp@kku.ac.kr

Received March 2, 2018
Accepted August 4, 2018
Published Online August 10, 2018

Purpose

This study was conducted to develop and validate an individualized prediction model for automated detection of acquired taxane resistance (ATR).

Materials and Methods

Penalized regression, combined with an individualized pathway score algorithm, was applied to construct a predictive model using publically available genomic cohorts of ATR and intrinsic taxane resistance (ITR). To develop a model with enhanced generalizability, we merged multiple ATR studies then updated the learning parameter via robust cross-study validation.

Results

For internal cross-study validation, the ATR model produced a perfect performance with an overall area under the receiver operating curve (AUROC) of 1.000 with an area under the precision-recall curve (AUPRC) of 1.000, a Brier score of 0.007, a sensitivity and a specificity of 100%. The model showed an excellent performance on two independent blind ATR cohorts (overall AUROC of 0.940, AUPRC of 0.940, a Brier score of 0.127). When we applied our algorithm to two large-scale pharmacogenomic resources for ITR, the Cancer Genome Project (CGP) and the Cancer Cell Line Encyclopedia (CCLE), an overall ITR cross-study AUROC was 0.70, which is a far better accuracy than an almost random level reported by previous studies. Furthermore, this model had a high transferability on blind ATR cohorts with an AUROC of 0.69, suggesting that general predictive features may be at work across both ITR and ATR.

Conclusion

We successfully constructed a multi-study–derived personalized prediction model for ATR with excellent accuracy, generalizability, and transferability.

Key words
Taxoids, Paclitaxel, Docetaxel, Drug resistance, Molecular diagnosis, Machine learning

Introduction

Taxanes, notably paclitaxel (PTX) and docetaxel (DTX), are cytotoxic microtubule-stabilizing agents used in various types of cancers, including gynaecological cancers (ovarian, cervical, and endometrial cancer) and breast cancers, with proven survival benefits [1]. Intrinsic or acquired resistance (AR) to chemotherapy are major clinical obstacles, resulting in poor response and lower overall survival rates. Yet, there is no efficient predictive model for resistance due to its complexities. Drug resistance is a result of complex biochemical and molecular processes. Moreover, crosstalks between different signaling pathways adds an additional layer of intricacy. Identifying the genetic and pathway alterations for resistant tumor cells and predicting resistance using genomic data will be valuable in cancer research and clinical management.

To predict anti-cancer drug responses, recent large-scale pharmacogenomic projects, notably Cancer Cell Line Ency-

clopedia (CCLE) and Cancer Genome Project (CGP), published genomic data and dose responses for drugs across cancer cell lines. CCLE and CGP analyzed responses of over 1,000 cell lines to 24 anti-cancer drugs and over 700 cell lines to 138 drugs, respectively. Two studies tested over 400 cell lines and 15 drugs in common. One of the common drugs is PTX and CGP has additional DTX response data. Despite the pharmacogenomic significance in medical research, inconsistency between two studies has been controversial recently. A past study reported a discordance in measured pharmacologic drug response between CCLE and CGP and inconsistent correlation between genomic profile and drug response, potentially undermining researches based on the database [2]. Another study obtained fair statistical consistency by revised metrics but attributed PTX to the majority of inconsistent drug/cell line pairs [3]. Besides, drug resistance of cell lines in CCLE and CGP is known as intrinsic, posing a difficulty in modeling AR. Past studies on acquired taxane resistance (ATR) often individually investigated single cell lines and drug treatments, and the generalizability and transferability of their findings remain undetermined.

High-throughput technologies such as array and sequencing have drastically altered biological research. Since the high dimensionality of genomic features renders the conventional regression limited, analyzing large-scale bioinformatic data became particularly challenging. Standard statistical models require independent assumption, which is violated by the highly correlated nature of genomic features. Regularized machine learning such as penalized regression has been developed for high dimension data structures. Penalized regression with its versatility in data mining and machine learning quickly became one of the most widely used ensemble learning methods. The regression is highly data adaptive, suitable for high dimension low sample size data, and sensitive for interactions and correlations among features. In addition, penalized regression is more interpretable and hence advantageous than "black-box machine learning models," especially in the field of medicine.

In this study, we developed and validated a highly accurate multi-study–derived, multivariable predictive model for ATR using personalized pathways and sophisticated machine learning algorithms.

Materials and Methods

1. Study selection

We searched for relevant articles on the PubMed and EMBASE using the following search term combinations:

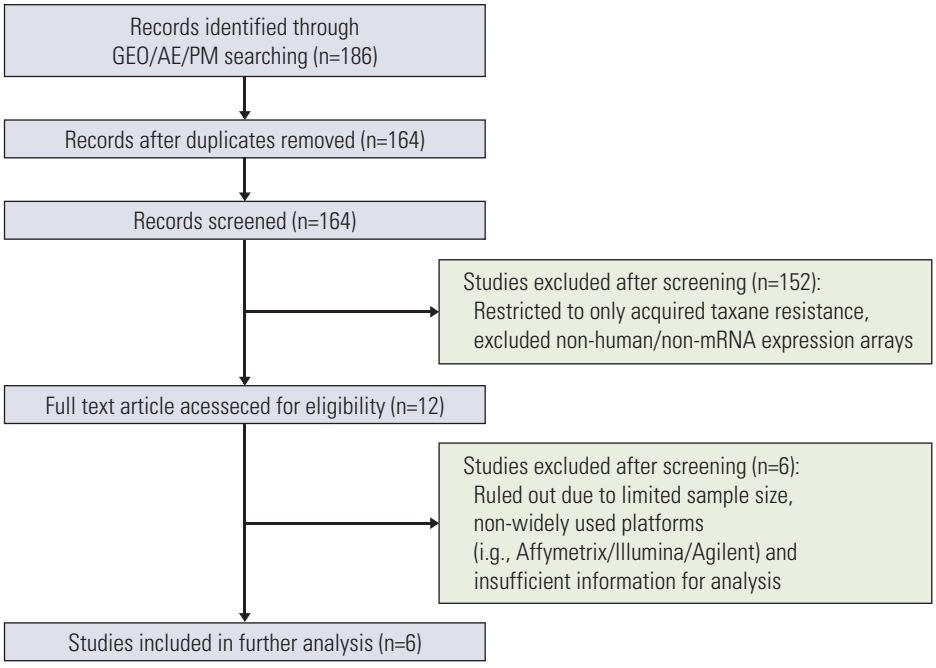

Fig. 1. Flow diagram describing the selection process of genomic studies for acquired taxane resistance. GEO, Gene Expression Omnibus; AE, ArrayExpress; PM, PubMed.

Table 1. Characteristics of individual studies

	Cohort	Drug sensitivity		Cancer cell lines	Centre	Platform
		S	R			
Intrinsic drug resistance	CCLE–PTX[a]	239[a]	240	240 Human cancer cell lines	Broad Institute	Affymetrix HG U133 Plus 2.0 Array
	CGP–PTX, DTX[a] (E-MTAB-783)	143 244	143 244	388 Human cancer cell lines	Wellcome Sanger Institute	Affymetrix HG U133A
Aquired drug resistance	GSE36135 (Domingo-Domenech et al. [4])	6	6	Parental docetaxel-sensitive prostate cancer cell lines (DU145 and 22Rv1) and selected docetaxel-resistant cells (DU145-DR and 22Rv1-DR)	Mount Sinai School of Medicine	Affymetrix HG U133 Plus 2.0 Array
	GSE28784 (Zwart and Rechache [5])	3	6	Docetaxel and paclitaxel resistant MDA-MB-231 (breast cancer) cells	Georgetown University	Affymetrix HG U133A Array
	GSE12791 (Luo et al. [7])	8	8	Parental paclitaxel-sensitive breast cancer cell lines (MDA-MB-231) and paclitaxel resistant cells (MDA-PR)	Denovo Biopharma	Affymetrix HG U133A Array
	GSE33455 (Marin-Aguilera et al. [8])	6	6	Parental docetaxel-sensitive prostate cancer cell lines (DU145 and PC3) and selected docetaxel-resistant cells (DU145-DR and PC3-DR)	Fundació Clínic per a la Recerca Biomèdica	Affymetrix HG U133 Plus 2.0 Array
	GSE23779 (Landen et al. [6])	3	3	Parental ovarian cancer cell lines (SKOV3ip1) and paclitaxel-resistant SKOV3TRip2	University of Alabama at Birmingham	Illumina Human Ref-8 v2.0

S, sensitive; R, resistant; CCLE, Cancer Cell Line Encyclopedia; PTX, paclitaxel; CGP, Cancer Genome Project; DTX, docetaxel.
[a]The resistant/sensitive phenotypes to taxane were classified as follows: cell lines in the below median IC_{50} or area under curve (referred to as ActArea in CCLE) values were classified as sensitive and those above median IC_{50} or area under curve values were classified as resistance.

"(taxane OR taxoids OR paclitaxel OR docetaxel OR cabazitaxel) AND (drug resistance OR chemoresistance)." The gene expression datasets were screened and retrieved from NCBI's Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) or ArrayExpress (http://www.ebi.ac.uk/arrayexpress) at the European Bioinformatics Institute by the following query: "('taxane' [All Fields] OR 'taxoids' [MeSH Terms] OR AND 'drug resistance' [MeSH Terms]) AND expression profiling by array." Studies with insufficient sample sizes, animal data and inadequate control groups were excluded. We only included samples that acquired taxane resistance via stepwise selection and the datasets from microarray platforms from Affymetrix GeneChip, Agilent one-color microarrays and Illumina BeadArray. These platforms are widely used with publicly available annotation information and more consistent in quality. The flow chart of study selection is in Fig. 1 and the selected study cohorts are summarized in Table 1.

2. Data processing

All data sets (GSE36135 [4], GSE28784 [5], GSE23779 [6], GSE12791 [7], GSE33455 [8], CCLE, and CGP) used in this study are publicly accessible from the GEO via National Center for Biotechnology Information (http://www.ncbi.nih.gov/geo) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress) at the European Bioinformatics Institute, CCLE (http://www.broadinstitute.org/ccle/), and CGP (http://www.cancerrxgene.org/). Raw gene expression profiles for CCLE and CGP cell lines were publicly accessible from both

the CCLE website and ArrayExpress under the accession number E-MTAB-783, respectively. For CCLE and CGP datasets, the resistant/sensitive phenotype to taxane were classified as follows: cell lines in the below median IC_{50} or area under curve (referred to as ActArea in CCLE) values were classified as sensitive and those above median IC_{50} or area under curve values were classified as resistance. Detailed step-by-step normalization methods and procedures have been documented previously [9]. Expression values from each data set were normalized and log-transformed. Raw data from Affymetrix platforms, if available, were pre-processed using Robust Multi-array Average (RMA) [10]. Otherwise, we used pre-processed data as provided by the original authors. To generate gene level summarization, we utilized an interquartile range (IQR) method. This allowed us to designate the probe set ID with the largest IQR of expression values out of all multiple probe set IDs as the representative of the gene. Missing expression values are designated using nearest neighbor imputation (R package impute) [11]. To achieve correct batch effect and cross-study normalization, ComBat, an empirical Bayes method, was applied [12].

3. Development of the algorithm

The pathway dysregulation scores (PDS) for each individual sample point were calculated using Pathifier algorithm which is designed to quantify the degree of pathway abnormality [13]. This method uses the algorithm by Hastie and Stuetzle to find a principal curve which is nonparametric, nonlinear generalization of the first principal component for dimension reduction [14].

Consider one-dimensional curve f which is a vector $f(s)$ of n functions of a single parameter variable s in n−dimensional space. Given a finite n−dimensional random vector $X=(X_1, X_2, ..., X_n)$, the projection index is defined as:

$$V_f(x)=\sup_s\{v: ||x-f(s)|| =\inf_\mu ||x-f(\mu)||\}$$

and the condition for self-consistency is simply $f(v)=E(X | v_f(X)=s)$. The PDS of sample i is defined as the distance along the curve between principal curve f_i and a reference end point, defined as the centroid of control set of samples (i.e., sensitive cells). Every sample is analyzed in relation to this principal curve and PDSs are assigned using the normalized projection distance for each sample's pathway. Pathway information used to design the PDS matrix was obtained from three curated pathway databases (the Kyoto Encyclopedia of Genes and Genomes, the BioCarta and the National Cancer Institute–Nature Pathway Interaction Database). Then we used regularized regression on these PDS matrices to fit the model. Regularization techniques have been described in detail in previous reviews [15,16]. The elastic net is a regularized regression method that linearly combines the penalties of the lasso and ridge regression methods and is defined as $p_{\alpha,\lambda}(\beta_j)=\lambda(\alpha||\beta_j||_1+(1-\alpha)_2||\beta_j||_2)$ [15,16]. It combines L2 norm (ridge) and L1 norm (lasso) penalty with a tuning parameter α, where $\alpha \in [0, 1]$ that can control the proportion of ridge/lasso penalty.

Elastic net is optimal for high dimension, low sample size (HDLSS) genomic data with highly correlated predictors because L1 reduces model complexity and L2 prevents over-simplifying. To make the multi-study–derived classifier, the leave-one-out cross validation (LOOCV) procedure, which is repeated N times (the total number of samples), is used to estimate the average standard error and identify an optimal value of the regularization parameter with minimum deviance. The efficient parameter selection via global optimization (EPSGO) algorithm was then used to further optimize the parameters [17]. EPSGO, based on learning an online Gaussian process, is a meta-heuristic algorithm which selects its parameters according to maximum likelihood. This algorithm, robust against local minima, is far more computationally efficient than the commonly used grid search method. For variable selection, the optimal parameter values were then utilized. We used R package pathifier to calculate PDS and glmnet package to construct the model and modified the methods of Hughey and Butte [9] and the functions from R package C060 [17]. We utilized the caret R package which implements e1071 and randomForest packages for support vector machines (SVM) and random forest (RF), respectively, using its default optimization by grid search on set parameter ranges [18].

4. Evaluation strategies

The performance evaluation metrics used in this study were the area under the receiver operating curve (AUROC), the precision-recall curve (AUPRC), Brier score (BS), precision, recall, accuracy (ACC), Matthews correlation coefficient (MCC), and F1 score. Receiver operating curve is a plot of test sensitivity (true positive [TP]/(TP+false negative [FN]) along the y axis versus 1-specificity (1−true negative [TN]/(false positive [FP]+TN)) along the x axis. Area under the curve (AUC) value ranges from 0.5 (random prediction) to 1 (perfect prediction). Precision-recall curve is a plot characterized by different set of precision (TP/(TP+FP)) and recall (sensitivity) of the model evaluated with selected thresholds. BS is calculated as:

$$BS=n^{-1}\sum_{i=1}^{n}(o_i-p_i)^2$$

where p_i is the predicted probability and o_i is the actual outcome of the event and n is the sample size. BS is essentially

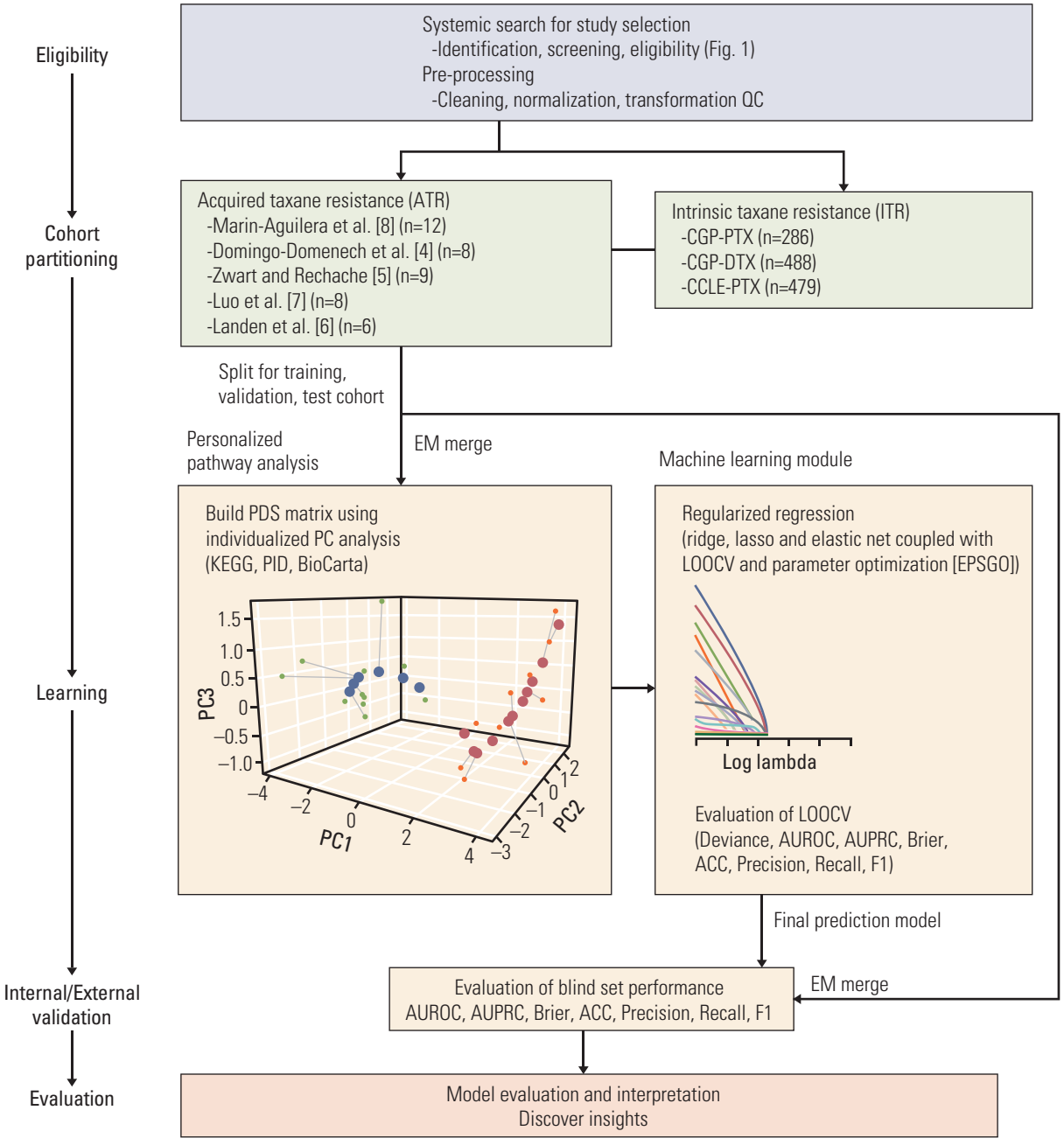

Fig. 2. Workflow for the development and validation of machine learning model for predicting acquired taxane resistance (ATR). The pipeline consists of three main parts: cross-study normalization, transformation into pathway information and model construction. The study cohort was preprocessed and splited into an internal development and validation cohort and an external blind validation cohort. An empirical Bayes approach (Combat) method was used for cross-study normalization. Transforming gene expression level information into pathway-level score for each individual sample was conducted using three curated pathway databases (Kyoto Encyclopedia of Genes and Genomes [KEGG], Pathway Interaction Database [PID], and BioCarta). Using these pathway-level score matrix, penalized regression model was constructed. Parameter optimization of the prediction model was conducted using leave-one-out cross validation (LOOCV) with Efficient Parameter Selection via Global Optimization (EPSGO) algorithm. QC, quality control; CGP, Cancer Genome Project; PTX, paclitaxel; DTX, docetaxel; CCLE, Cancer Cell Line Encyclopedia; EM, Empirical Bayes Method; PDS, pathway dysregulation scores; PC, principal component; AUROC, area under the receiver operating curve; AUPRC, area under the precision-recall curve; ACC, accuracy.

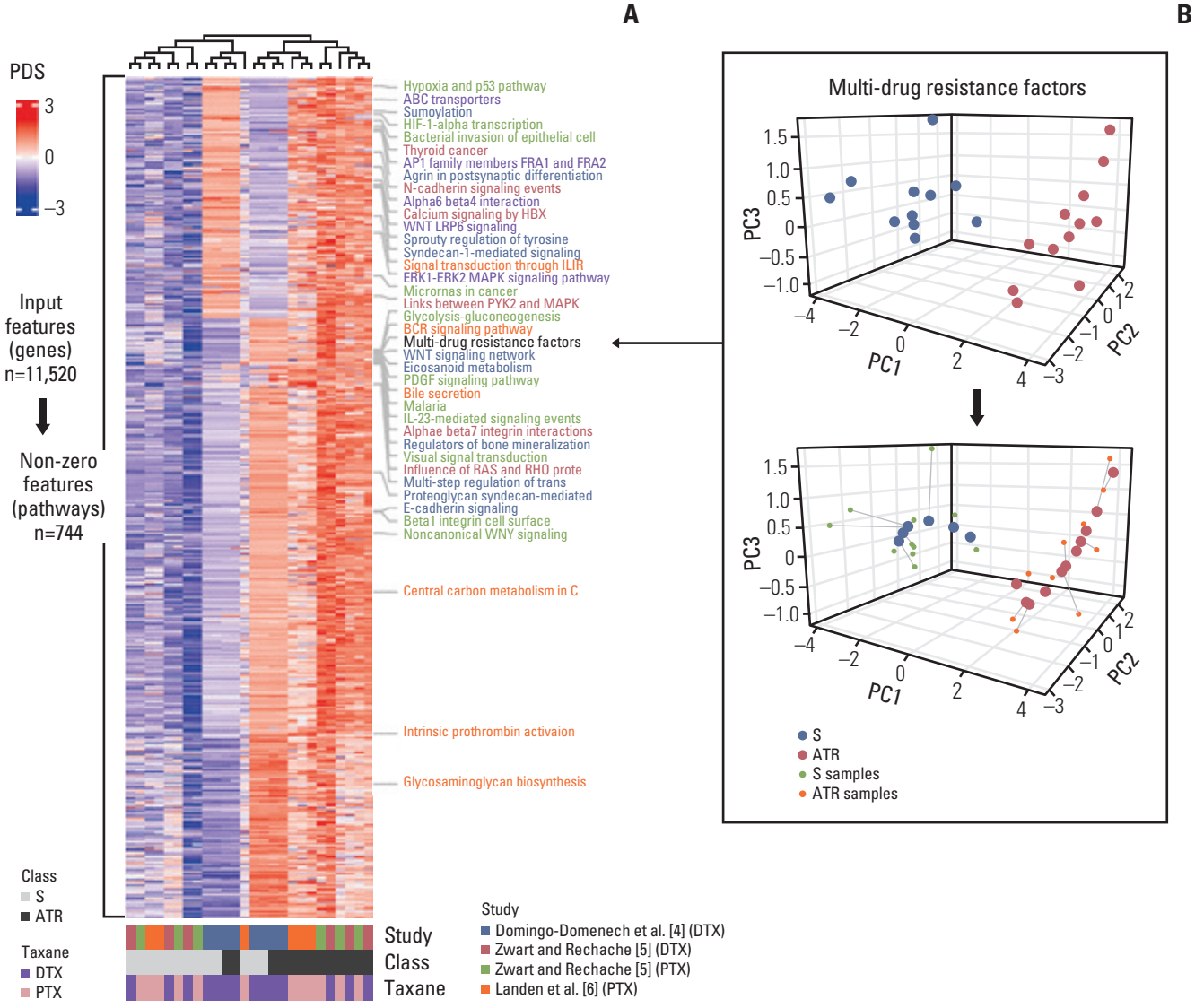

Fig. 3. Multi-study–derived, individualized pathway learning model for predicting acquired taxane resistance (ATR). (A) Pathway deregulation score (PDS) matrix for the three development cohorts (GSE36135, GSE28784, GSE23779). Each row (744 pathway features from 11,520 input gene features) represents the z-score-normalized PDS for each individual sample in each cohort. The color bars in the bottom indicate drug sunsitivity status, type of taxane and study cohort. (B) An example of principal curve of the pathway. The principal curve is individually learned with each pathways of the development cohorts. The data points and the principal curve are projected onto the three principal components (PCs). The principal curve goes through the cloud of samples and is directed so that control samples (sensitive to taxane) are near the beginning of the curve. *(Continued to the next page)*

the mean squared error of the probability forecast of a dichotomous event. Hence, a small BS corresponds to a good calibration of predictions. F1 score is a weighted mean of precision and recall, ranging from 0 (worst value) to 1 (best value). ACC is defined as (TP+TN)/(TP+TN+FP+FN). MCC, considered as a balanced measure, is a geometric mean corrected for chance agreement ((TP×TN)–(FP×FN)/square root ((TP+FP)×(TP+FN)×(TN+FP)×(TN+FN))). Prediction performances of all parameters except BS are directly proportional, ranging from 0 to 1. Higher BS denotes worse performance. MCC ranges from –1 (completely incorrect) to 1 (completely correct). All statistical analyses were performed

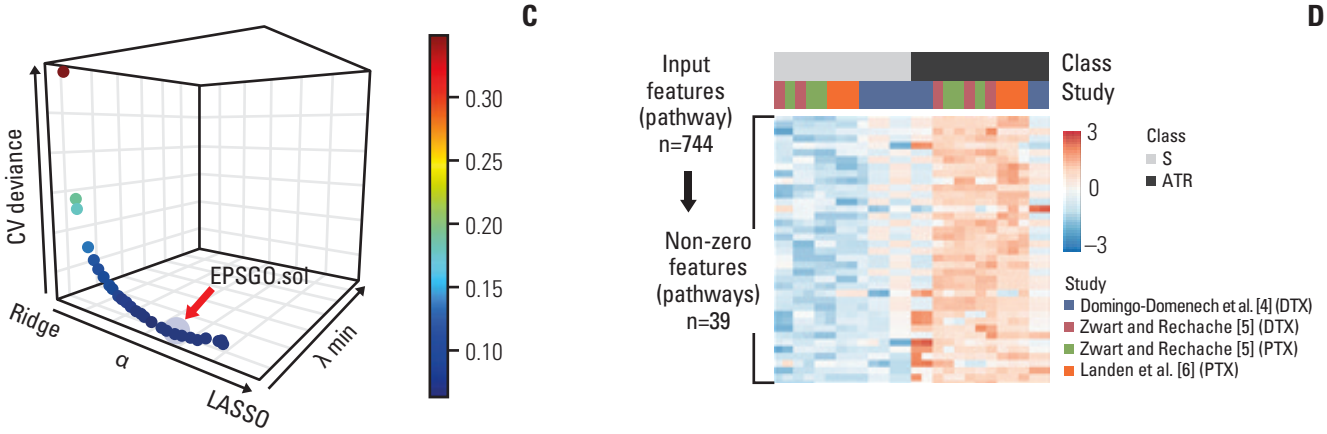

Fig. 3. *(Continued from the previous page)* (C) Hyperparameter optimization for elastic-net with Efficient Parameter Selection via Global Optimization (EPSGO). Cross-study validation deviance as a function of both tuning hyperparameters α and λ is shown. α controls the tradeoff between the ridge and lasso penalties, whereas λ controls the overall amount of penalization. The red arrow highlights the final EPSGO solution where the deviance is within 1SE of the minimum (α=0.682 and λ=0.004). (D) Heatmap of the pathways with non-zero coefficient. From 744 input pathways, 39 pathways with non-zero coefficients were selected. The names of the final pathways are labelled on the right side of PDS matrix shown in panel A. S, sensitive; DTX, docetaxel; PTX, paclitaxel.

using R ver. 3.2.3 (R Foundation for Statistical Computing Platform, Vienna, Austria).

Results

To develop a robust and generalized ATR prediction model based on personalized pathway information, we devised a workflow that integrates multi-study, and multi-platform penalized machine learning method with individual pathway deregulations in taxane treatments (PTX, DTX) in various cancer cell lines (Fig. 2). Three microarray studies (GSE36135-DTX-prostate, GSE28784-DTX and PTX-breast, GSE23779-PTX-ovarian) were used as a development study set for model construction. For external blind validation, two independent cohorts (GSE12791-PTX-breast and GSE33455-DTX-prostate) were used to test the algorithm's generalizability and transferability. To explore possible transferability between intrinsic taxane resistance (ITR) and ATR, we used CCLE and CGP cohorts as a development set and tested ATR cohorts as an external blind. Detailed descriptions of cohorts and the technical variables used in the studies are in Table 1.

We merged the three discovery study cohort using the ComBat method [12]. These merged gene expression level data were then transformed into pathway-level information using the Pathifier algorithm which generated a one-dimensional principal curve from a cloud of data points in a high-dimensional space (Fig. 3B) and yields a PDS for each individual sample in a context-specific manner (Fig. 3A, see Materials and Methods section) [13]. Using pathway information extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [19], Pathway Interaction Database (PID) [20] and the BioCarta [21], we calculated a principal curve for each pathway and obtained 744PDS from 11520 merged genes (Fig. 3A and B). We then applied regularized regression to PDS matrix to build a prediction model for ATR. Elastic-net regularization linearly combines the ridge and lasso regression [16]. Hyperparameter α adjusts ridge (L2-norm) and lasso penalties (L1-norm), whereas and λ controls the total level of penalization. The hyperparameters are fine-tuned for an optimal elastic-net penalty function. We used an EPSGO algorithm to optimize α and λ with minimum binomial deviance (Fig. 3C) [17]. At the value that regularization parameter gave the lowest binomial deviance, EPSGO-tuned elastic-net selected a parsimonious set of 39 predictors with non-zero pathway dysregulation coefficients (Fig. 3D). Detailed descriptions of 39 non-zero pathways and their gene components are in S1 Table and S2A Fig. Out of 39 non-zero pathways, five most informative pathways with coefficients greater than 1 were Biocarta's "HYPOXIA AND P53 IN THE CARDIOVASCULAR SYSTEM," BioCarta's "MULTI-DRUG RESISTANCE FACTORS," KEGG's "BILE SECRETION," PID's "ALPHAE BETA 7 INTEGRIN CELL SURFACE INTERACTIONS," and KEGG's "ABC TRANSPORTERS"

Table 2. Performance measures in ATR cross-study validation

	GSE36135 (Domingo-Domenech et al. [4])	GSE28784 (Zwart and Rechache [5])	GSE23779 (Landen et al. [6])
AUROC	1.000	1.000	1.000
AUPRC	1.000	1.000	1.000
Brier	0.022	0	0.001

AUROC, area under the receiver operating curve; AUPRC, area under the precision-recall curve.

Table 3. Performance measures in overall cross-study validation and external validation for ATR

	Cross study validation	Blind study validation
AUROC	1.000 (1.000-1.000)	0.940 (0.841-1.000)
AUPRC	1.000	0.940
Brier score	0.007	0.127
Confusion matrix metrics		
Sensitivity (Recall/TPR)	1.000	0.900
Specificity	1.000	0.800
Precision (PPV)	1.000	0.818
Likelihood ratio positive (LR+)	Inf	4.500
Likelihood ratio negative (LR–)	0.000	0.125
F1	1.000	0.857
Cohort	GSE3613 (Domingo-Domenech et al. [4]), GSE28784 (Zwart and Rechache [5]), GSE23779 (Landen et al. [6])	GSE33455 (Marin-Aguilera et al. [8]), GSE12791 (Luo et al. [7])

Values in parentheses are 95% confidence intervals. ATR, acquired taxane resistance; AUROC, area under the receiver operating curve; AUPRC, area under the precision-recall curve; TPR, true positive rate; PPV, positive predictive value.

(S2B Fig.). The final elastic-net model produced a perfect performance on leave-one-out cross-study validation. The overall AUROC for the three development cohorts were 1.000 with a AUPRC of 1.000, a BS of 0.007 and a sensitivity, specificity of 100% (Tables 2, 3, Fig. 4A). We further validated the generalizability of our model using two external validation cohorts. Our algorithm showed excellent performances on both independent test sets. The overall AUROC for the two external blind cohorts were 0.940 (95% confidence interval [CI], 0.841 to 1.000) with an AUPRC of 0.940 and a BS of 0.127 (Fig. 4B). The sensitivity and specificity of the algorithm were 90.0% and 80.0%, respectively (Table 3). The algorithm showed excellent performance on leave-one-out cross-validation compared to RF or SVM (S3 Fig.).

Next, we explored whether a transferability exists between ITR to and ATR. We classified the CCLE and CGP data as sensitive (below the median IC_{50} or area under curve values) or resistant (above the median IC_{50} or area under curve values). After classification, the total of 239, 143, 244 samples were obtained in the drug-sensitive (S) group and 240, 143, 244 in the resistant (R) groups of CCLE-PTX, CGP-PTX, CGP-DTX, respectively. In leave-one-out cross-study validation for CCLE and CGP, our algorithm showed a high discrimination ability with an overall AUROC of 0.703 (95% CI, 0.674 to 0.731), an AUPRC of 0.712, a BS of 0.218, a sensitivity of 61.7% and a specificity of 67.1% (Tables 4, 5, Fig. 5A). Considering the accuracies previously reported on consistency between CGP and CCLE were close to random level at 0.5, our model's performance was remarkable. Next, we tested whether this ITR-based model could predict ATR. Surprisingly, our ITR-based model had a good prediction performance on ATR (overall AUROC, 0.688 [95% CI, 0.539 to 0.837]; AUPRC, 0.735; BS, 0.226; sensitivity, 68.0%; and specificity, 64.0%), suggesting high transferability between ITR and ATR (Table 5, Fig. 5B).

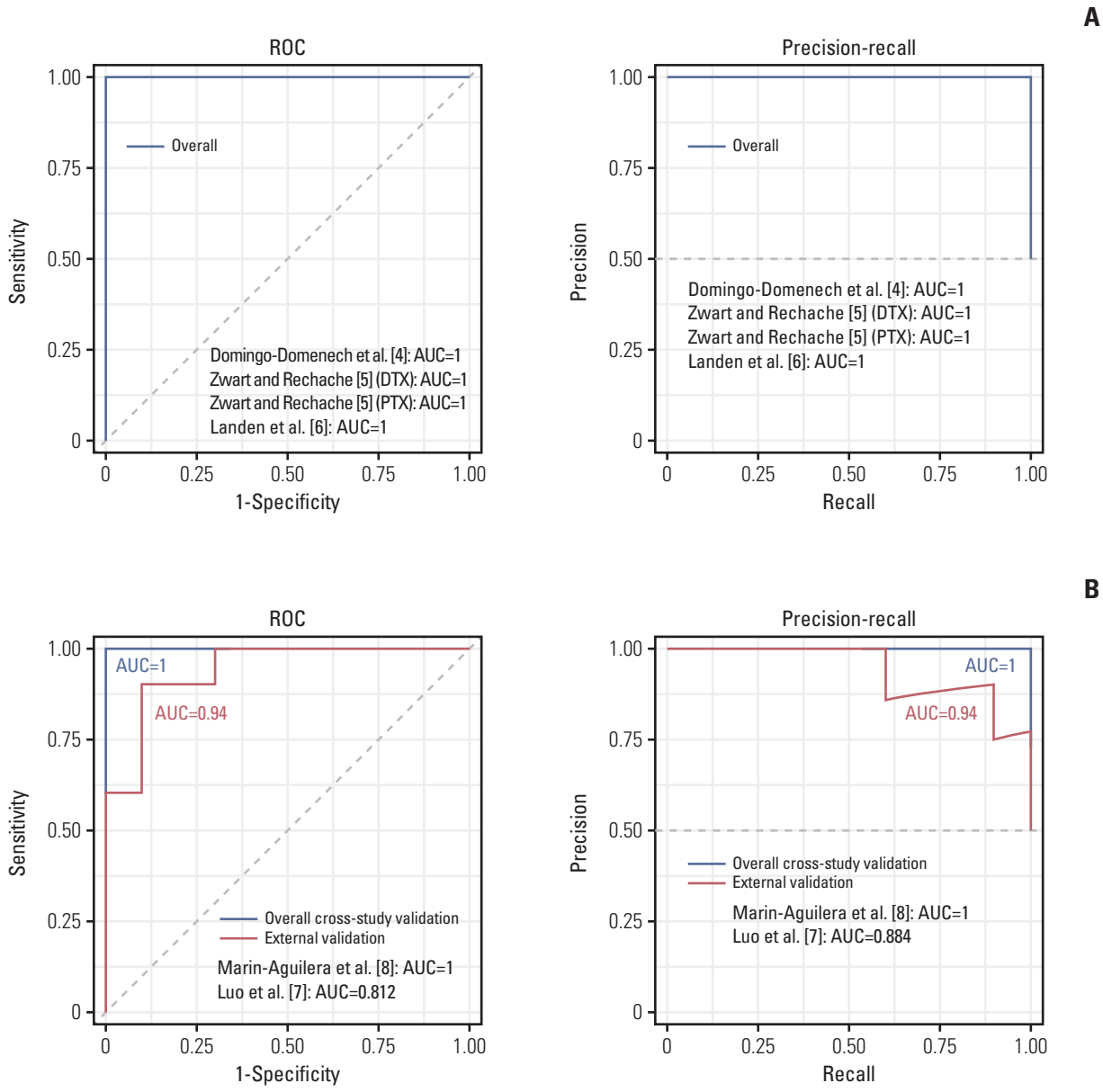

Fig. 4. Acquired taxane resistance (ATR)–trained model performances on internal and external validation ATR cohorts. Receiver operating characteristic (ROC) and precision-recall curve are used to show ability to predict. (A) Model performances on internal cross-validation ATR cohorts. (B) Model performances on external blind ATR cohorts. DTX, docetaxel; PTX, paclitaxel; AUC, area under the curve.

Discussion

The transferability and generalizability of our model may be attributed to using multi-study–derived and pathway-based regularized regression. Our model showed high generalizability for ATR by producing near-perfect performance in both internal cross-study validation and external validation. We associate the model's robustness with our two-step approach of multi-study–derived pathway mapping and penalized regression to maximize generalizability. The first step was to convert genomic information into pathway information because pathways represent multifactorial nature of cancer and drug resistance better than individual genes. The second step was to use penalized regression to avoid overfitting and secure interpretability.

Table 4. Performance measures in ITR cross-study validation

	CCLE-PTX	CGP-PTX	CGP-DTX
AUROC	0.739 (0.695-0.783)	0.660 (0.597-0.722)	0.692 (0.645-0.738)
AUPRC	0.735	0.679	0.708
BRIER	0.208	0.23	0.221

Values in parentheses are 95% confidence intervals. ITR, intrinsic taxane resistance; CCLE, Cancer Cell Line Encyclopedia; PTX, paclitaxel; CGP, Cancer Genome Project; DTX, docetaxel; AUROC, area under the receiver operating curve; AUPRC, area under the precision-recall curve.

Table 5. Performance measures in overall cross-study validation for ITR and external validation with ATR cohorts

	Cross study validation	Blind study validation
AUROC	0.703 (0.674, 0.731)	0.688 (0.539, 0.837)
AUPRC	0.712	0.735
Brier score	0.218	0.226
Confusion matrix metrics		
Sensitivity (Recall/TPR)	0.617	0.680
Specificity	0.671	0.640
Precision (PPV)	0.653	0.654
Likelihood ratio positive (LR+)	1.876	1.889
Likelihood ratio negative (LR–)	0.571	0.500
F1	0.634	0.667
Cohort	CCLE-PTX CGP-PTX CGP-DTX	GSE3613 (Domingo-Domenech et al. [4]), GSE28784 (Zwart and Rechache [5]), GSE23779 (Landen et al. [6]), GSE33455 (Marin-Aguilera et al. [8]), GSE12791 (Luo et al. [7])

Values in parentheses are 95% confidence intervals. ITR, intrinsic taxane resistance; ATR, acquired taxane resistance; AUROC, area under the receiver operating curve; AUPRC, area under the precision-recall curve; TPR, true positive rate; PPV, positive predictive value; CCLE, Cancer Cell Line Encyclopedia; PTX, paclitaxel; CGP, Cancer Genome Project; DTX, docetaxel.

CCLE and CGP have provided, since their publications, pharmacogenomic information for prediction of drug sensitivity. Yet, recent studies documented that the inconsistency of two studies, largely on pharmacologic response to antitumor drugs, may be problematic for studies based on these datasets. A past study observed very poor correlation of IC_{50} between CCLE and CGP (Pearson rho of 0.18 in IC_{50} for PTX with SVM classifier) [2]. Another study reported a similar finding, citing PTX among the major cause for drug/cell line inconsistency (Spearman's rank correlation coefficient 0.1-0.2) [3]. Recently, Dong et al. [22] applied linear (SVM) and non-linear (random forest) modeling and addressed that, although SVM achieves better performance (0.55) for PTX than RF predicting model (0.482), the values were not much higher than random prediction. Our internal cross-study validation using CCLE and CGP cohorts shows the model with pathway mapping approach is highly consistent between CCLE and CGP (overall cross-study AUC, 0.703), compared to the almost random levels reported in previous studies. Applying this model on ATR cohorts, we further tested whether a model built from intrinsic resistance dataset can predict AR. Surprisingly, prediction parameters were as high, suggesting the model's generalizability over intrinsic and acquired resistant cell lines (overall AUC, 0.688).

By using pathifier and penalized regression, we achieved parsimony, narrowing 11,520 input gene features down into 39 non-zero pathways. Interestingly, compared to ATR, the number of coefficients (features) for ITR are much greater than for ATR (S4A and S4B Fig.). This is suggesting ITR had a broader feature selection. Although intrinsic and acquired resistances are explained by different mechanisms, the algorithm may have captured the common pathways shared by

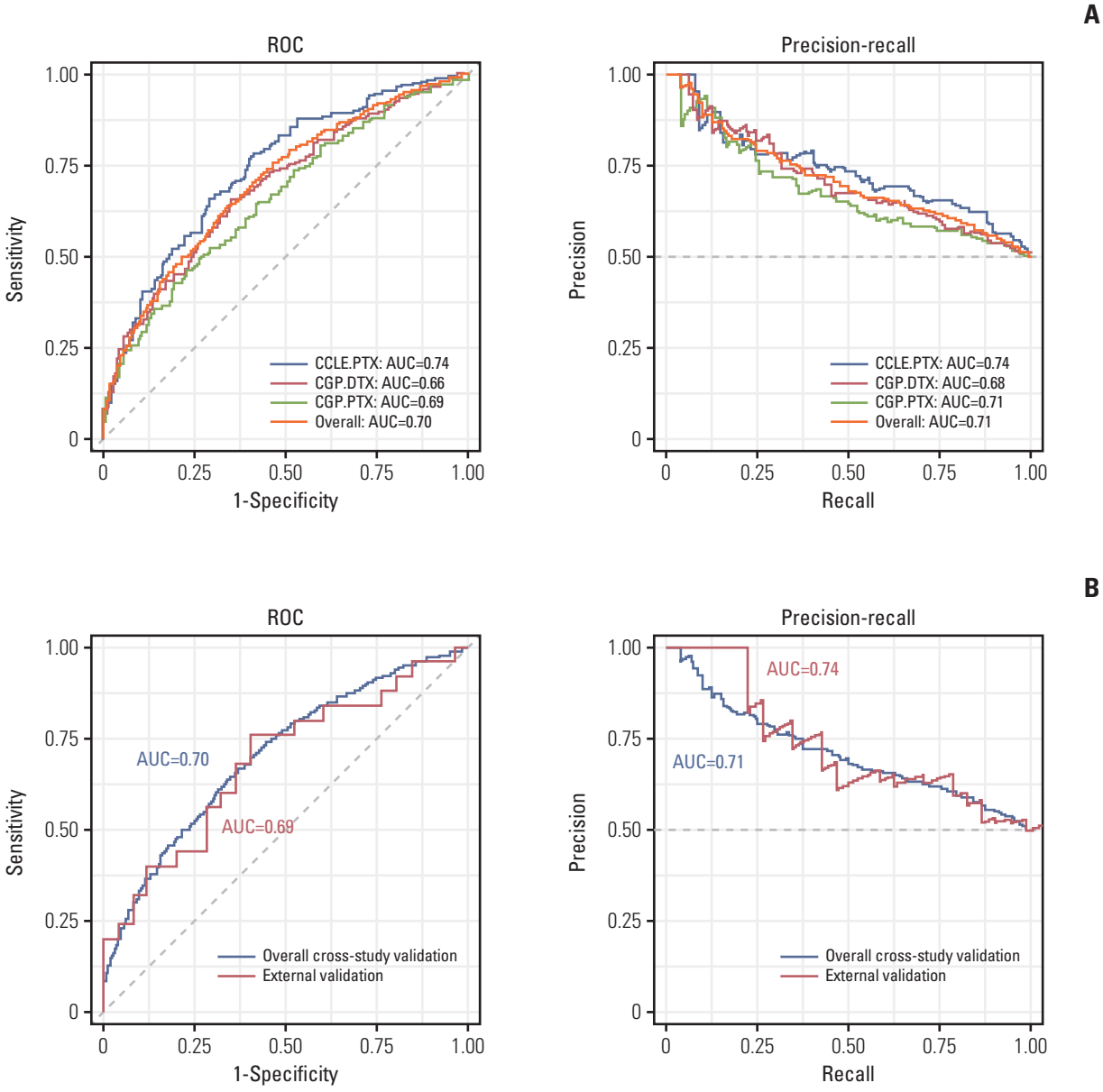

Fig. 5. Intrinsic taxane resistance (ITR)–trained model performances on internal (ITR) and external validation (acquired taxane resistance [ATR]) cohorts. Receiver operating characteristic (ROC) and precision-recall curve are used to show ability to predict. (A) Model performances on internal validation of ITR cohorts (CCLE-PTX, CGP-DTX, and CGP-PTX). (B) Model performances on external ATR cohorts. AUC, area under the curve; CCLE, Cancer Cell Line Encyclopedia; PTX, paclitaxel; CGP, Cancer Genome Project; DTX, docetaxel.

both ATR and ITR. This may be one of the reasons that our model developed from ITR data (CCLE, CGP) was transferable to predict ATR with good accuracy. We additionally examined whether ATR-based model predicted ITR and observed near-random performance (data not shown). This implies the smaller number of feature selection for ATR might be due to the presence of AR-specific pathways. Further knowledge of resistant mechanisms is needed to fully understand this outcome.

One advantage of using regression-based models in medical sciences is results are highly interpretable, compared to "black-box" models such as deep learning, SVM and random forest. It is interesting to note that KEGG's "BILE SECRETION" and PID's "ALPHAE BETA 7 INTEGRIN CELL SUR-

FACE INTERACTIONS" were among the top 5 most informative pathways with coefficient greater than 1 (S2B Fig.). While the other top pathways (Biocarta's "HYPOXIA AND P53 IN THE CARDIOVASCULAR SYSTEM," BioCarta's "MULTI-DRUG RESISTANCE FACTORS," and KEGG's "ABC TRANSPORTERS") have been consistently associated with chemoresistance with previous studies to suggest robustness of the model [23], no direct physiological relationship between bile secretion, integrin β7 and taxane resistance has been reported in previous studies [24,25]. Integrins are reported to be implicated in cell adhesion-mediated drug resistance and the examples include integrin β1, which was among 39 non-zero features, for erlotinib resistance in lung cancer and lapatinib/trastuzumab in breast cancer [26]. Yet, the role of integrin β7, one of the most informative feature according to our algorithm, in drug resistance has not been thoroughly studied. Further investigations into bile secretion and integrin β7 pathway may provide novel molecular target candidates for ATR.

An accurate model will be valuable for clinical decision making, providing most effective and least toxic drug choices. In this study, we developed a multi-study–derived personalized prediction model for ATR with excellent accuracy, generalizability and transferability.

Electronic Supplementary Material

Supplementary materials are available at Cancer Research and Treatment website (https://www.e-crt.org).

Conflicts of Interest

Conflict of interest relevant to this article was not reported.

Acknowledgments

This paper was supported by the National Research Foundation (NRF)-2016R1A1A1A05921984.

References

1. Nussbaumer S, Bonnabry P, Veuthey JL, Fleury-Souverain S. Analysis of anticancer drugs: a review. Talanta. 2011;85:2265-89.
2. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJ, et al. Inconsistency in large pharmacogenomic studies. Nature. 2013;504:389-93.
3. Bouhaddou M, DiStefano MS, Riesel EA, Carrasco E, Holzapfel HY, Jones DC, et al. Drug response consistency in CCLE and CGP. Nature. 2016;540:E9-10.
4. Domingo-Domenech J, Vidal SJ, Rodriguez-Bravo V, Castillo-Martin M, Quinn SA, Rodriguez-Barrueco R, et al. Suppression of acquired docetaxel resistance in prostate cancer through depletion of notch- and hedgehog-dependent tumor-initiating cells. Cancer Cell. 2012;22:373-88.
5. Zwart A, Rechache N. Gene expression data of sensitive, Docetaxel resistant and paclitaxel resistant MDA-MB-231 cells. GEO DataSets. Series GSE28784 [Internet]. National Center for Biotechnology Information; 2011 [cited 2018 May 2]. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28784.
6. Landen CN Jr, Goodman B, Katre AA, Steg AD, Nick AM, Stone RL, et al. Targeting aldehyde dehydrogenase cancer stem cells in ovarian cancer. Mol Cancer Ther. 2010;9:3186-99.
7. Luo W, Schork NJ, Marschke KB, Ng SC, Hermann TW, Zhang J, et al. Identification of polymorphisms associated with hypertriglyceridemia and prolonged survival induced by bexarotene in treating non-small cell lung cancer. Anticancer Res. 2011;31:2303-11.
8. Marin-Aguilera M, Codony-Servat J, Kalko SG, Fernandez PL, Bermudo R, Buxo E, et al. Identification of docetaxel resistance genes in castration-resistant prostate cancer. Mol Cancer Ther. 2012;11:329-39.
9. Hughey JJ, Butte AJ. Robust meta-analysis of gene expression using the elastic net. Nucleic Acids Res. 2015;43:e79.
10. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4:249-64.
11. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17:520-5.
12. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118-27.
13. Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. Proc Natl Acad Sci U S A. 2013;110:6388-93.
14. Hastie T, Stuetzle W. Principal curves. J Am Stat Assoc. 1989;84:502-16.
15. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1-22.
16. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005;67:301-20.
17. Sill M, Hielscher T, Becker N, Zucknick M. c060: extended inference with Lasso and elastic-net regularized Cox and generalized linear models. J Stat Softw. 2014;62:1-22.
18. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28:1-26.

19. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27-30.
20. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. Nucleic Acids Res. 2009;37:D674-9.
21. Nishimura D. BioCarta. Biotech Softw Internet Rep. 2001;2: 117-20.
22. Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer. 2015;15:489.
23. Greenberger LM, Sampath D. Resistance to taxanes. In: Teicher BA, ed. Cancer drug resistance. Totowa, NJ: Humana Press; 2006. p. 329-58.
24. Zeng L, Kizaka-Kondoh S, Itasaka S, Xie X, Inoue M, Tanimoto K, et al. Hypoxia inducible factor-1 influences sensitivity to paclitaxel of human lung cancer cell lines under normoxic conditions. Cancer Sci. 2007;98:1394-401.
25. Szakacs G, Paterson JK, Ludwig JA, Booth-Genthe C, Gottesman MM. Targeting multidrug resistance in cancer. Nat Rev Drug Discov. 2006;5:219-34.
26. Seguin L, Desgrosellier JS, Weis SM, Cheresh DA. Integrins and cancer: regulators of cancer stemness, metastasis, and drug resistance. Trends Cell Biol. 2015;25:234-40.