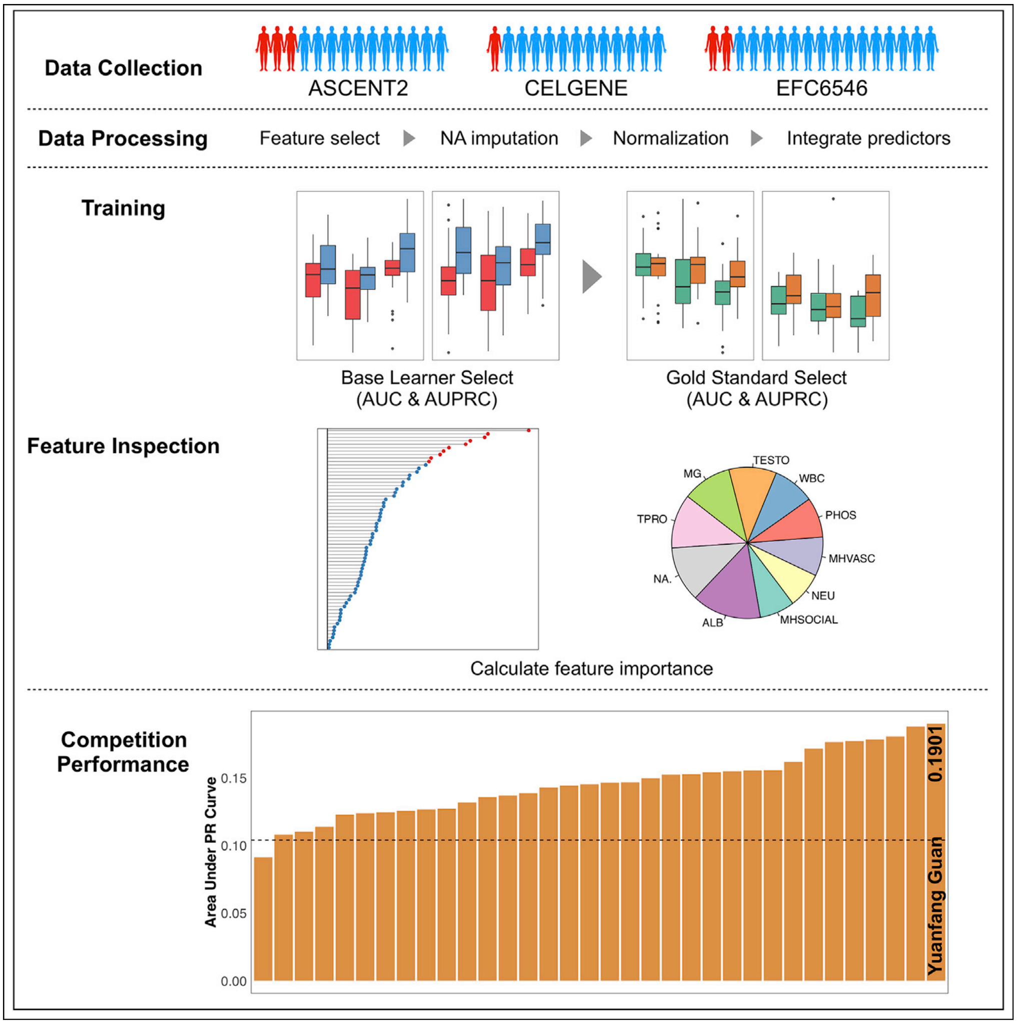
iScience

CellPress

Article

Treatment Stratification of Patients with Metastatic Castration-Resistant Prostate Cancer by Machine Learning



Kaiwen Deng,
Hongyang Li,
Yuanfang Guan

gyuanfan@umich.edu

HIGHLIGHTS

Predicting the docetaxel treatment discontinuation in prostate cancer

The winning solution in the DREAM Prostate Cancer Challenge

Integrating survival status and adverse events to stratify treatment discontinuation

Deng et al., iScience *23*, 100804
February 21, 2020 © 2019 The Author(s).
https://doi.org/10.1016/j.isci.2019.100804

iScience

CellPress

Article

Treatment Stratification of Patients with Metastatic Castration-Resistant Prostate Cancer by Machine Learning

Kaiwen Deng,[1,3] Hongyang Li,[1,3] and Yuanfang Guan[1,2,4,*]

SUMMARY

Prostate cancer is the most common cancer in men in the Western world. One-third of the patients with prostate cancer will develop resistance to hormonal therapy and progress into metastatic castration-resistant prostate cancer (mCRPC). Currently, docetaxel is a preferred treatment for mCRPC. However, about 20% of the patients will undergo early therapeutic failure owing to adverse events induced by docetaxel-based chemotherapy. There is an emergent need for a computational model that can accurately stratify patients into docetaxel-tolerable and docetaxel-intolerable groups. Here we present the best-performing algorithm in the Prostate Cancer DREAM Challenge for predicting adverse events caused by docetaxel treatment. We integrated the survival status and severity of adverse events into our model, which is an innovative way to complement and stratify the treatment discontinuation information. Critical stratification biomarkers were further identified in determining the treatment discontinuation. Our model has the potential to improve future personalized treatment in mCRPC.

INTRODUCTION

Prostate cancer is the most common cancer in men and the second leading cause of cancer-related mortality in the Western world (Gupta et al., 2014). Androgen deprivation therapy (ADT) is widely used to treat prostate cancer, but one-third of the patients will develop resistance to ADT, which is called castration-resistant prostate cancer (CRPC) (Kristiyanto et al., 2016). In general, patients with prostate cancer progress into CRPC in 18–48 months, and most deaths result from metastatic CRPC (mCRPC) with a median survival time of fewer than 2 years (Cookson et al., 2015). There are many other therapeutic options available for patients with progressive cancer. Docetaxel is one of the first-line treatments for mCRPC (Heidenreich et al., 2014). It has been shown that combined with prednisone, 75 mg/m^2 docetaxel treatment for 3 weeks significantly reduced pain and improved overall survival time and quality of life in phase III trials (Tannock et al., 2004).

Although many studies have proved the positive effect of docetaxel on population-level survival (Berthold et al., 2008; Machiels et al., 2008), some patients become resistant to the therapy and stop the treatment (Petrylak et al., 2004). Approximately 10%–20% of the patients who were initially under docetaxel treatment discontinued the therapy owing to toxicity-induced adverse events (AEs) (Templeton et al., 2013). Since docetaxel-based chemotherapy is still an essential treatment for mCRPC and hormone-sensitive metastatic prostate cancer (Sweeney et al., 2015), it is important to identify whether a patient is able to tolerate the docetaxel therapy. However, whether the early discontinuation is predictable based on the clinical characteristics of a patient remains unclear (Seyednasrollah et al., 2017).

Recognizing the critical clinical need for identifying patients who are expected to suffer from immediate adverse events upon docetaxel treatment, the international Dialogue for Reverse Engineering Assessment and Methodology (DREAM) committee organized the Prostate Cancer Challenge, which called for the research community to develop algorithms for predicting AE (Seyednasrollah et al., 2017). In this challenge, a benchmark dataset was held out to systematically evaluate different computational methods. Here we report a model based on our winning solution to this challenge, which ranked first among 34 participating teams. In this model, we use 78 features associated with the laboratory test, metastases data, patient clinical features, and medical history. We find that the survival status is highly informative, complementing the prediction of discontinuation. We further reveal important predictive features that can be directly used to guide treatment in clinical settings. The model presented in this study not only sets the state of the field for

[1] Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA

[2] Department of Internal Medicine, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA

[3] These authors contributed equally

[4] Lead Contact

*Correspondence: gyuanfan@umich.edu

https://doi.org/10.1016/j.isci.2019.100804

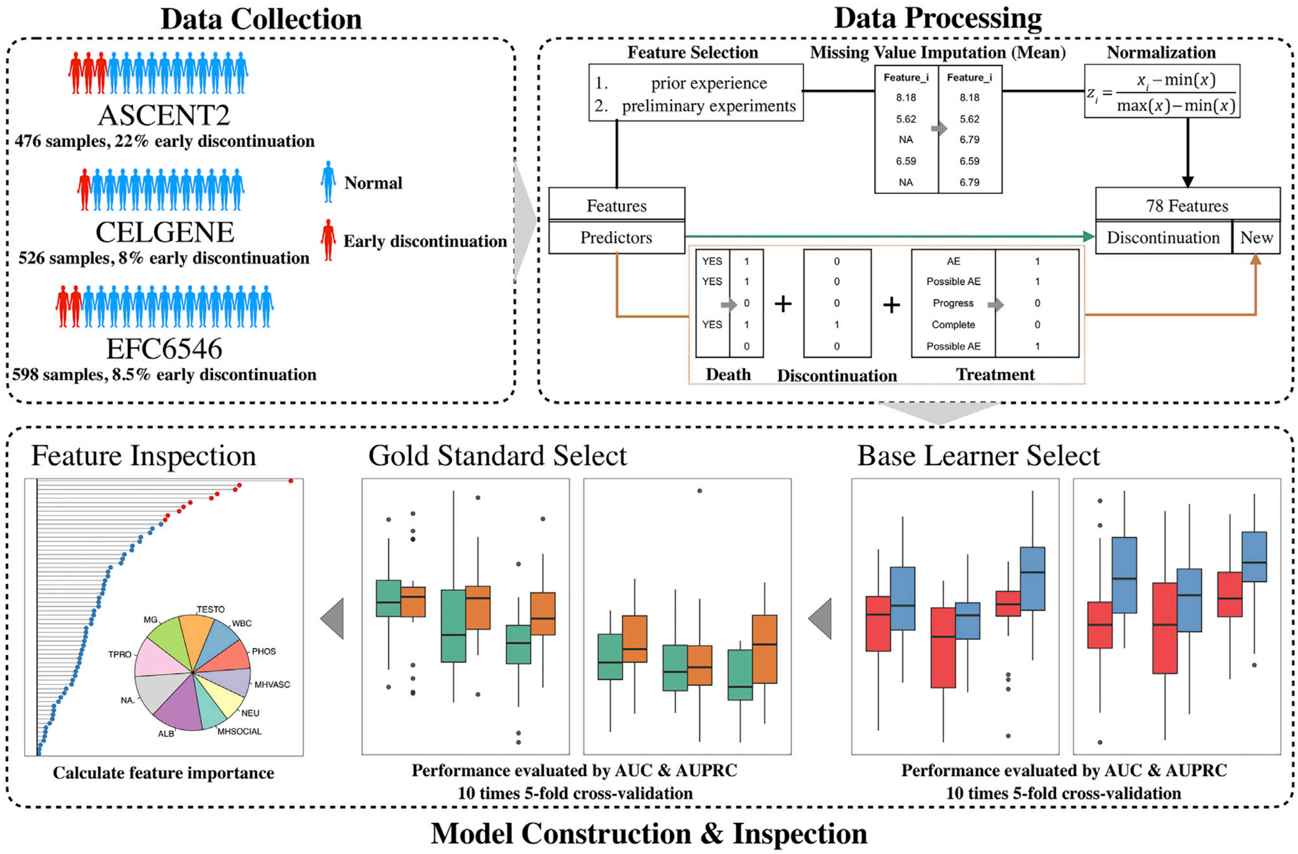


Figure 1. The Workflow of the Algorithm Design for Predicting Early Discontinuation of Treatment in Prostate Cancer

The overall workflow consists of three main parts: (1) Data Collection; (2) Data Processing; (3) Model Construction & Inspection. Data were collected from three cohorts. In data processing, 78 features were used and normalized. Missing values were imputed. We generated two sets of gold standards for the subsequent model fitting, in which "New" (right in the "Data Processing" part) represents the gold standard assembled by the death, discontinuation, and treatment statuses. The "Model Construction & Inspection" part has three steps: (step 1) choose the best base learner; (step 2) construct a combinatory gold standard; (step 3) calculate the feature importance. All the performances were evaluated by randomly partitioning the data 10 times into 5-fold cross-validation, either on the full dataset or on an individual cohort. The final model submitted during the challenge was trained on the full dataset, which combined three cohorts.

personalized modeling of mCRPC treatment but also provides potential biomarkers and risk factors in the future prostate cancer treatment.

RESULTS

Overview of the Experimental Design for Predicting Treatment Discontinuation in Prostate Cancer

The schematic illustration of our experimental design is shown in Figure 1. In this study, we used the data collected at three different cohorts from a total of 1,600 patients in phase III prostate cancer clinical trials Table 1. To address the cohort and batch effects, the original data were processed through missing value imputation and normalization (see details in Methods). We tested and compared different types of machine learning models via the standard 5-fold cross-validation experiments. The prediction performance was evaluated by area under the receiver operating characteristic curve (AUC) (Bradley, 1997) and area under the precision-recall curve (AUPRC). To understand the crucial features in predicting treatment discontinuation, we further performed feature importance analysis and identified top contributing features.

We first tested the prediction performance of five base learners: (1) linear regression, (2) logistic regression, (3) Cox regression (using AE as an endpoint), (4) bootstrap aggregation classification and regression trees (BAG-CART) (Sutton, 2005), and (5) random forest (RF) (Chen et al., 2004). The first three methods are linear

Cohorts	ASCENT2 (ASC)	CELGENE (CEL)	EFC6546 (VEN)
Basic Information			
# Sample	476	526	598
Median age (years)[a]	71	68	68
Discontinuation Status			
% Discontinuation	22.05	7.79	8.52
Median time to discontinuation (days)	153.0	211.0	202.5
% Discontinuation records missing	0.00	18.06	0.00
Death Status			
% Death	28.99	17.49	72.41
Median time to death (days)	357.0	279.0	642.5
Treatment Status			
% AE	9.03	13.31	21.07
% Possible AE	46.85	45.24	22.24
% Progression	19.96	21.29	55.68
% Complete	24.16	0.00	0.17
% Treatment status records missing	0.00	20.15	0.84

Table 1. Patient Characteristics of the Three Cohorts

[a]"≥85" is converted to 85.

or generalized linear statistical models, and the last two methods are nonlinear tree-based models. For this task of predicting treatment discontinuation primarily based on laboratory test and patient clinical features, we anticipated that the nonlinear interactions between input features would be crucial. We, therefore, investigated two tree-based methods, which have slightly different implementations—when building trees, BAG-CART considers all features for each node for a split and RF considers a random subset of features for each node for a split. We performed 5-fold cross-validation for ten times in two scenarios: (1) using the combined dataset of three cohorts (hereafter referred to as "the full dataset"; Figures 2A and 2B) and (2) using each cohort individually (Figures 2C and 2D).

In general, the discontinuation status can be predicted by the models trained from clinical features with those widely used base learners. When trained and evaluated on the full dataset, all the five base learners achieved acceptable performances, and the tree-based methods, including BAG-CART and RF, performed relatively better than the other three (Figures 2A and 2B). Similarly, when trained and evaluated on each cohort individually, the tree-based methods consistently achieved good performance, whereas the linear, logistic regression and Cox regression model performed even worse than the random prediction baseline in EFC6546(VEN) (Figure 2C; Table S1). The highest AUC scores were achieved by RF, RF, BAG-CART, RF, respectively in the full dataset and each of the three cohorts, where the median AUCs were 0.6269, 0.5701, 0.6562, and 0.5378 (Figure 2C). They also outperformed other base learners in terms of AUPRC, where the medians were 0.1999, 0.3080, 0.1792, and 0.0959 (Figure 2D). Of note, the average baseline AUPRCs of random prediction were 0.1308, 0.2146, 0.0952, and 0.0885, reflecting the overall ratio of positive cases. Since BAG-CART and RF are similar tree-based models and RF achieved the highest performance in two of the three cohorts and the full dataset, we chose RF as our base learner in the following analysis.

Cross-Reference Multiple Labels Improves Prediction Performance

To improve the performance of our model, we carefully inspected the dataset and found that multiple labels provided non-redundant information. The original definition of discontinuation was whether an AE occurred within 92 days. According to this definition, patients with AE at 93 days (high risk to discontinue) or 393 days (low risk to discontinue) were indistinguishable and their labels were identical. A major

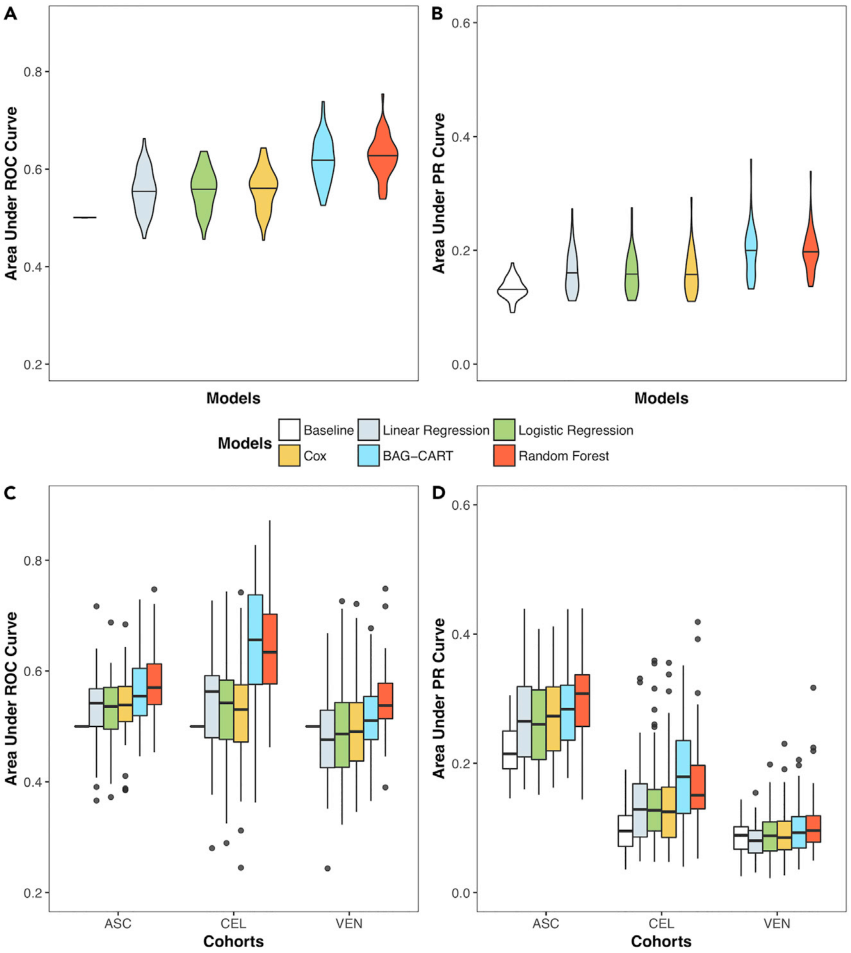


Figure 2. Evaluation of Different Base Learning Using AUC and AUPRC

The (A) AUCs and (B) AUPRCs of five base learners trained on the full dataset are shown in different colors. The (C) AUCs and (D) AUPRCs of these base learners trained on each cohort individually were also calculated. The colored boxes or violins represent the values of 10 times 5-fold cross-validation of five models, and the white ones represent the baselines. Of note, the AUC baseline of random prediction is always 0.5 in different cohorts. In contrast, the AUPRC baseline will change, which equals the percentage of early discontinuation events in a particular cohort.

challenge was how to integrate this hidden information into a model. In fact, the death and treatment statuses were highly associated with the discontinuation. And a patient who died early (within 3 months) could be, although not definitely, associated with an adverse event. Therefore, we treated early death as another type of "adverse event" and assumed that patients labeled by "AE" and "possible AE" in treatment status should have a higher risk to discontinue than those labeled by "progress" and "complete." These labels are intermediate variables reflecting diverse categories of the discontinuation reason, where "progress" means the treatment is in progress without AE and "complete" means the treatment has finished without AE.

Therefore, we integrated death status, discontinuation status. and treatment status (AE, possible AE, progression and complete) and created a new gold standard, retaining the original labels for a fair comparison. Specifically, we re-labeled the original discontinuation status: if a patient had AE or possible AE, or died early during the study, we treated this patient as a "discontinuation" case (see Methods). The distributions of the original and new labels are shown in Figures 3A and 3B, respectively. Notably, the high-risk patients (red) to discontinue treatment are separated from the low-risk ones (yellow and orange). Then we re-trained our model using this new gold standard and compared it with the model trained on the original discontinuation label. In all three cohorts, the performance was improved using the new gold standard (Table S2,

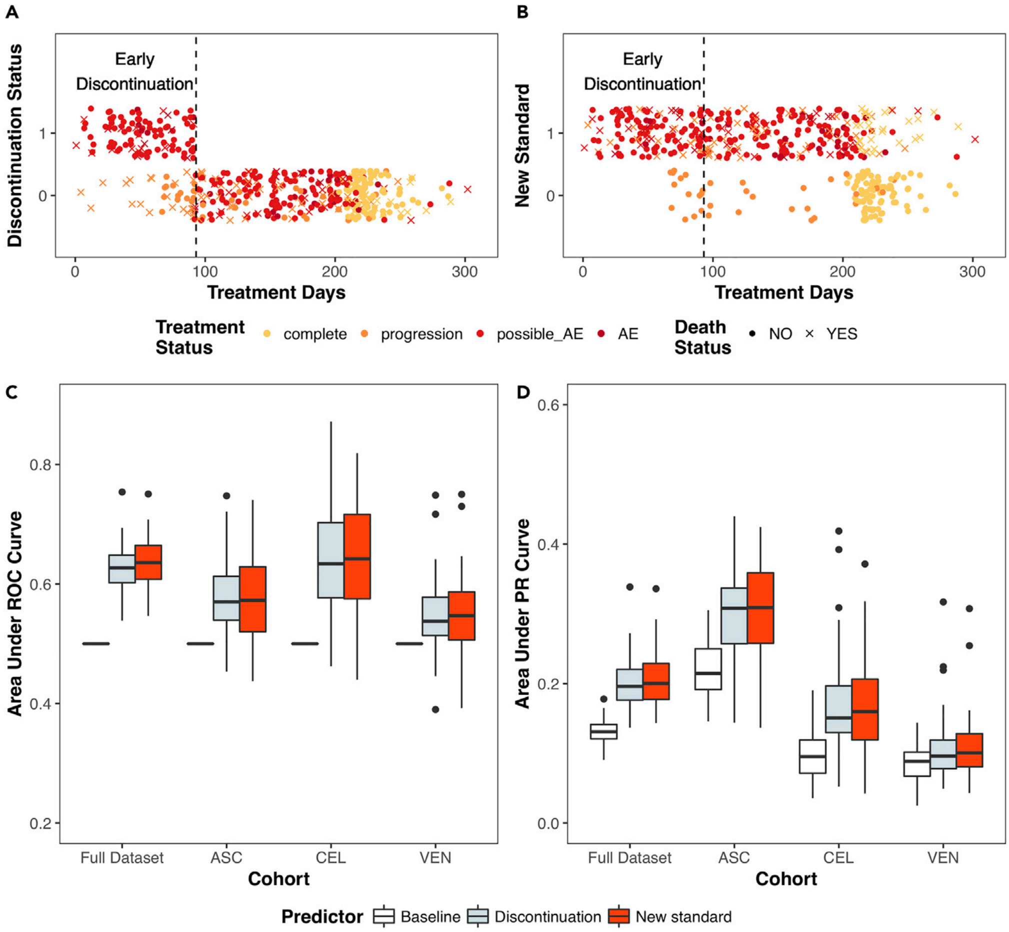


Figure 3. Comparison of Models Using the Original Discontinuation Labels and New Labels

We evaluated the performance of our model using two types of gold standards: (A) the original discontinuation status and (B) our new gold standard integrating both treatment and death status. The colors represent the treatment status of each sample, and the shapes represent the death status. The vertical dash line is placed at 93 treatment days, which is the original time threshold of early discontinuation. The early death cases are re-labeled as positive, which are more likely caused by the treatment adverse events. Models using labels defined by different gold standards are evaluated by (C) AUC and (D) AUPRC. "Discontinuation" represents the model using labels of the discontinuation status, and the "New standard" represents the model using labels of the new gold standard. Notably, the AUC baseline is a constant 0.5 (so that it is a horizontal line instead of a box), whereas the AUPRC baseline changes along with the partition of the dataset.

Figures 3C and 3D). The AUC scores were increased to 0.6356, 0.5726, 0.6420, and 0.547, and the AUPRC scores were increased to 0.2001, 0.3089, 0.1598, and 0.1006, in the full dataset, ASCENT(ASC), CELGENE (CEL), and EFC6546(VEN) cohorts, respectively. This result indicates that the discontinuation label alone cannot fully reflect the status of a patient; the death and treatment status provides important and complementary information for more accurate predictions.

Inspection of Important Features in Predicting Treatment Discontinuation

To find the key determinants of treatment discontinuation, we investigated the feature importance by calculating the delta-error of each feature in our random forest model (see Methods), on both the full dataset and subsets (ASC + CEL; CEL + VEN; VEN + ASC). We also colored the features according to their categories defined in the previous study (Seyednasrollah et al., 2017).

The top contributing features of models trained the full dataset and three individual cohorts are shown in Figure 4A and Figure S1, respectively. In general, laboratory test features play more important roles in this

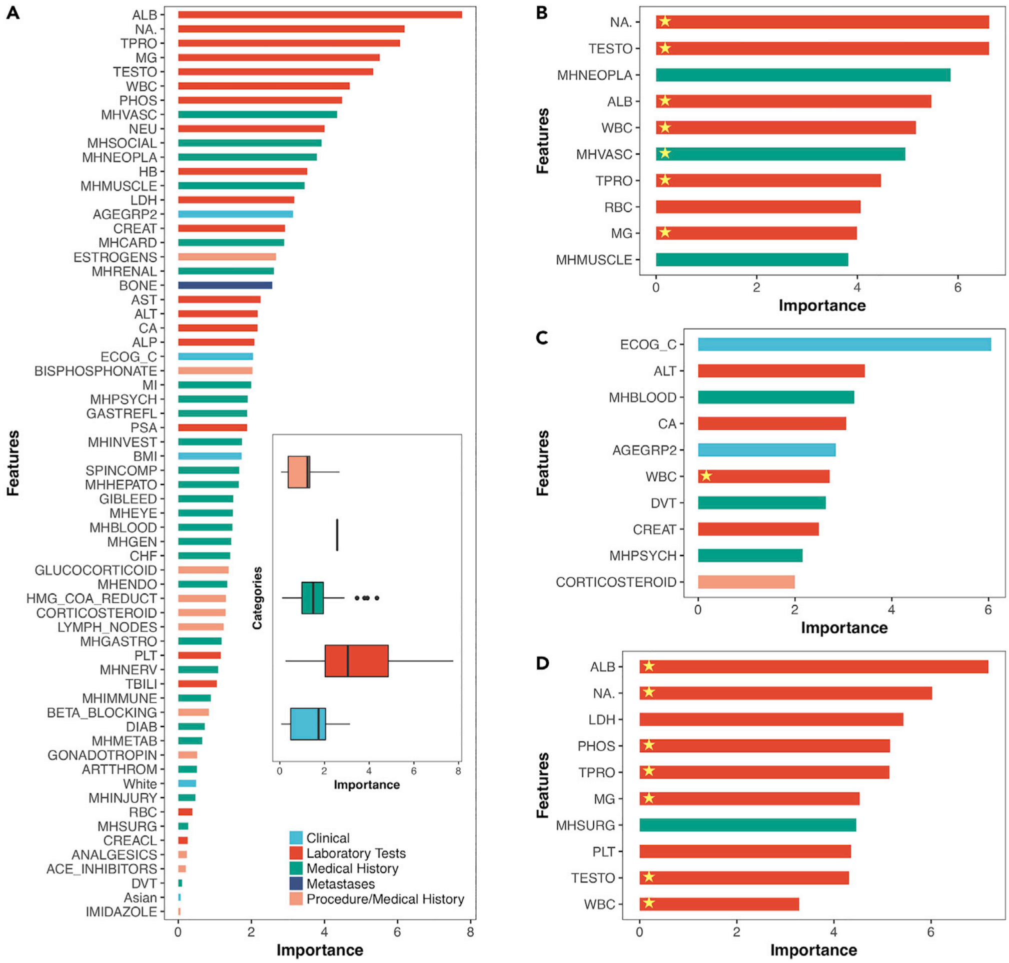


Figure 4. The Feature Importance Map of All 78 Features used in Our Model

(A) The importance of all features (features with negative importance have been removed). The top 10 important features are ALB, NA., TPRO, MG, TESTO, WBC, PHOS, MHVASC, NEU, and MHSOCIAL, which represent albumin, sodium, total protein, magnesium, testosterone, white blood cell, phosphorus, medical history of vascular disorders, neutrophils, and medical history of social circumstances, respectively. The colors of the bars represent the categories that the features belong to, and the sub-boxplot summaries the feature importance in each category. Note that there is only one feature that belongs to metastases.

(B–D) Top 10 important features in different subsets of data (B: ASC + CEL, C: CEL + VEN, D: VEN + ASC) are also presented here, where the stars represent the overlaps between them and the overall top 10 in panel A.

model, and the top 10 important features are consistent among different datasets (Figures 4B–4D, starred features). However, we also notice that, when trained on the combined dataset of CEL and VEN, the situation is different (Figure 4C). This may result from the cohort effects, which we will discuss in detail later. To investigate the impact of these features on patients with mCRPC, we first tested the linear relationship (Pearson's correlations) between the top 10 features and the risk of discontinuation on patients with mCRPC (Table S3). We observed that the impact of an individual feature on the discontinuation risk was weak because these clinical features had complex nonlinear interactions, which cannot be reflected by linear Pearson's correlation. This is also why the random forest model outperformed other base learners for this prediction task since the random forest was able to learn the non-linear interactions between features and robust to outliers or noises when the dataset is relatively small. We further calculated the correlations between the prediction of the random forest model using only the top 10 features and the treatment discontinuation (the last row in Table S3). This prediction can be regarded as a new non-linearly "combined" feature based on the top 10 features, and it has a much higher and significant correlation with the discontinuation risk on mCRPC patients.

The top 10 important features are ALB (Albumin), NA. (Sodium), TPRO (Total Protein), MG (Magnesium), TESTO (Testosterone), WBC (White Blood Cells), PHOS (Phosphorus), MHVASC (Medical History: Vascular Disorders), NEU (Neutrophils), and MHSOCIAL (Medical History: Social Circumstance). Most of them play important roles in cancer-related diagnosis, prognosis, and therapy in previous studies, especially the biological features including serum albumin, sodium, magnesium, testosterone, phosphorus, and neutrophils (Halabi et al., 2014). Our findings are not incidental; in fact, many previous studies have reported these features for predicting the survival of patients with mCRPC. Serum albumin can be used to predict prognosis and recurrence for patients with prostate cancer (Sejima et al., 2013; Wang et al., 2017); neutrophils is also regarded as a prognostic factor (Langsenlehner et al., 2015; Sonpavde et al., 2014; Templeton et al., 2014), especially for the patients with mCRPC with docetaxel chemotherapy (Yao et al., 2015). Sodium (Na) and magnesium (Mg) regulate and influence cancer development as ions and compounds: sodium compounds can affect prostate cancer cell activities and the function of related proteins (Berggren et al., 2009; Kim et al., 2014), whereas the rate of Ca^{2+}/Mg^{2+} is significantly associated with the risk and proliferation (Dai et al., 2011; Sun et al., 2013). Phosphorus, as another important chemical element for humans, has also been proven to be associated with the risk of high-grade prostate cancer (Wilson et al., 2014). And testosterone is also used as a therapeutic strategy and a predictive feature (Baillargeon et al., 2015; Mearini et al., 2013; Pastuszak et al., 2015), although its relationship with prostate cancer is controversial (Klap et al., 2015). Of note, these top features are highly associated with predicting treatment discontinuation, yet they do not imply causation. In this study, we focused on developing accurate machine learning models based on large datasets from three independent cohorts and identified the cross-cohort putative risk factors associated with (instead of causing) treatment discontinuation. Identifying etiological features requires further data collection and experimental validation, which is beyond the scope of this paper. Nevertheless, we here summarize literature evidence about these potential causal factors, which could provide guidance for future experimental design.

Validation on an External Cohort by Blind Test via the DREAM Platform

The validation of prediction was carried out through an objective evaluation venue. The international DREAM consortium had created a framework to unbiasedly evaluate the accuracy of different models (Guan, 2019). The participants were provided with the training data and the gold standard, and the test data, whereas the validation gold standard was hidden from participants. Then, through an online platform, the participants submitted a one-shot prediction file, which was evaluated automatically on the server.

Different gold standard creations and machine learning methods were explored by 34 teams during the challenge (Figure 5A). The basic gold standard labels are the discontinuation status (DISCONT) and censoring date (ENTRT_PC) of a patient, which were prevalently used by most teams (blue tiles in Figure 5A left). However, the death status (DEATH), date (LKADT_P), and the adverse effect (ENDTRS_C) of a patient were only used in our method, except for the death status was used once by another team. We integrated all these features and reconstructed the gold standard, which achieved the highest AUPRC for predicting discontinuation (the red model in Figure 5A). This comparison together with our cross-validation result suggests that gold standard reconstruction by capturing multiple sources of information was an important component of our high performance. Moreover, the base learners used by the participants were summarized and RF was used most by 14 of 34 teams. This suggests that the curation of the gold standard, rather than the exact machine learning base learner, distinguished our method from others.

The validation cohort of this challenge comes from the comparative arm (docetaxel-placebo) of ENTHUSE33 (Fizazi et al., 2013), a phase III trial previously designed for studying the effect of Docetaxel in Combination With Zibotentan in Patients With mCRPC, including 470 patients. Although the failure of the trial results was previously published, the data were not publicly available, allowing it to be used as an independent validation cohort in this study. In this validation cohort, our algorithm achieved the top performance among the competing teams with an AUPRC of 0.190, compared with the random baseline of 0.104 ($p = 0.003$, Figure 5). It means that if today's therapy would have mistakenly put an expected 104 out of 1,000 patients to docetaxel treatment, with our algorithm deployed in place, we will be able to move (0.190–0.104)/(1–0.104)*104, approximately 10 patients out of a wrong therapy. As with any personalized medicine algorithm using machine learning, a perfect prediction is our DREAM to pursue, but it is still a dream. However, translating this 10 of 104 patients into a type of cancer that affects the largest number of men in this world, it is expected to have a significant impact on the care to our patients.

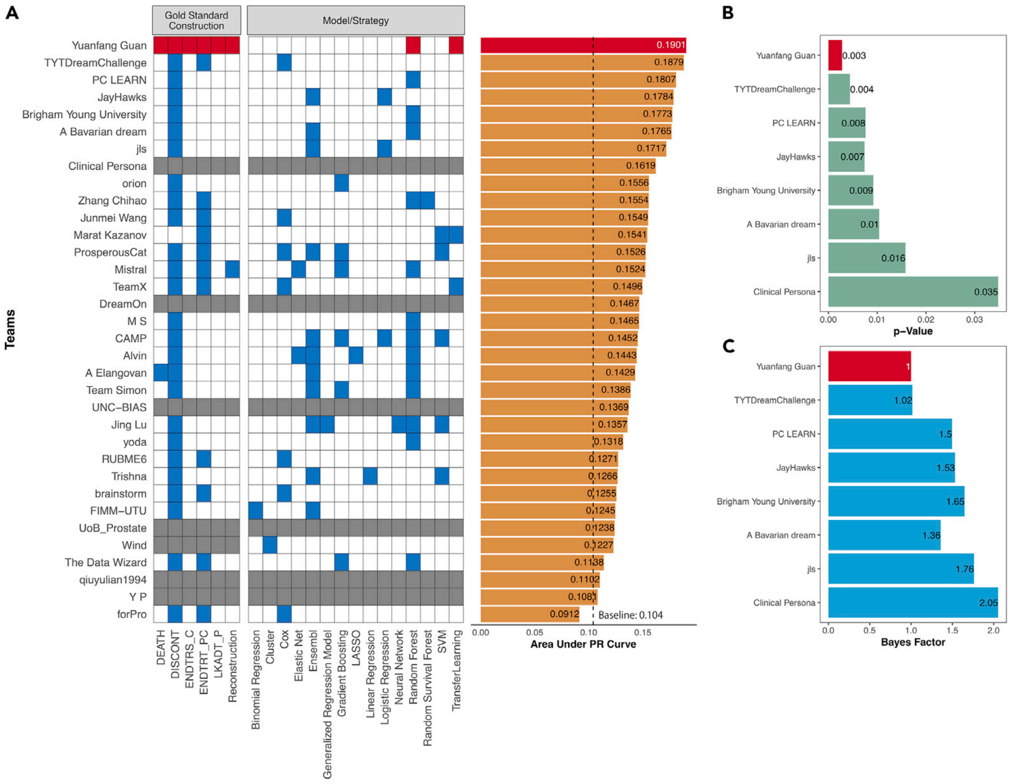


Figure 5. A Summary of the Challenge Results

All data were the final round results and retrieved from the challenge official website. Considering the potential privacy problems, all the team names except ours are hidden in this figure.

(A) Information of gold standard constructions, model categories, and learning strategies of the 34 teams (the left heatmap), as well as the final AUPRC estimation results (the right barplot). In summary, there were five variables used in the gold standard construction and 13 types of base learners. The participants mainly used ensemble study (Ensembl in heatmap x axis) to improve their models, and some of them reconstructed the gold standard (Reconstruction in heatmap x axis) for better performance. The blue tiles in the heatmap represent the variables or base learners used by the teams, whereas the white ones represent variables/learners that were not used. The gray tiles represent the missing information, and the red ones highlight our method. The barplot on the right shows the final AUPRC scores, where the dashed line represents the baseline of the validation data: 0.104.

(B) The p value of how significant a model is better than the score of 5,000 random predictions.

(C) The Bayes factor for the top performers. The Bayes factor is calculated by comparing each team with the best performer.

DISCUSSION

In this work, we report a machine learning model for predicting the docetaxel treatment discontinuation in mCRPCs. We find that tree-based methods are suited to this problem and outperform the Cox regression, one of the most popular methods in survival analysis. This phenomenon originated from the nature of the clinical dataset, where the features had complicated high-level interactions with each other in determining the treatment discontinuation. Tree-based methods are good at constructing models on a complex or non-linear dataset, and random forest applies the strategy of training on different parts of the dataset and averaging multiple decision trees to reduce the variance and avoid overfitting (Breiman, 2001; Li and Guan, 2019); thus, tree-based methods, especially random forest, are more suitable for such problems.

Apart from the best base learner selection, we also observed that the assembled gold standard would generate better predictions than using discontinuation status alone. This result reveals that the mCRPC clinical features have potential relationships with the risk of discontinuation of patients, which was reflected by the treatment status and the discontinuation time in this dataset. A better method to evaluate the risk should give even better performance than ours because the binary combination cannot reflect the risk order entirely.

Our work provides an attempt to integrate multiple information to address this problem. A dataset with continuous confidence labels instead of binary labels would facilitate method development in the future.

One limitation of this work results from the data availability. Most of the features in this dataset have more than 25% of their records missed (Figure S2). This situation can be even more severe when considering the cohort individually, such as albumin (ALB), which is totally missing in ASC. To impute the missing values, we just simply replaced them with the average of the corresponding feature, which may introduce biases, and intensify the batch effects among cohorts and complexity among features. This should be another reason why linear, logistic regression and Cox regression failed to generate acceptable models. A more intact dataset should generate more precision models.

The clinical data itself is largely biased by cohort effects. The inconsistent results show not only in the feature importance calculation but also in our previous attempts to evaluate models cross cohorts even after normalizing. So we used t-distributed stochastic neighbor embedding (t-SNE) (Van Der Maaten and Hinton, 2008) to visualize the differences among cohorts (Figure S3). This figure shows the separation of these three cohorts, in which CEL and VEN have a closer relationship than to ASC. This explains why the results are inconsistent between the datasets with and without ASC.

Overall, our research provides a positive answer to the question of whether the early discontinuation can be predicted from basic clinical information and points out that random forest is the most suitable base learner. Our assembled gold standard performance also suggests that considering the risk of discontinuation for each patient rather than the binary label of early discontinuation can further improve the accuracy of prediction.

Limitation of the Study

1. Feature availability. Most of the features have more than 25% missing values. And some of them are even completely missing in one cohort, like albumin (ALB) in the ASC cohort. Our imputation using the average of non-missing values may introduce biases into their original distributions.
2. Cohort effects. Data in the competition were provided by three providers, which could include cohort effects. Although we have normalized the whole data, our results indicate the existence of these effects. The performances vary when we built and tested models with and without the ASC cohort.
3. Base model. In recent years, releases of Xgboost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) provide us more efficient and powerful tools for constructing and training tree-based models. They may achieve better performances than the random forest model.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.100804.

ACKNOWLEDGMENTS

This work is supported by NSF-US14-PAF07599 (#1452656; CAREER: On-line Service for Predicting Protein Phosphorylation Dynamics Under Unseen Perturbations NSF), R35-GM133346-01 (Machine Learning for Drug Response Prediction), AWD007950 (Digital Biomarkers in Voices for Parkinson's Disease American Parkinson's Disease Association), University of Michigan O'Brien Kidney Translational Core Center, and Michael J. Fox Foundation #17373 to Y.G. This work is also supported by 19AMTG34850176 (American Heart Association and Amazon Web Services 3.0 Data Grant Portfolio: Artificial Intelligence and Machine Learning Training Grants) to H.L. We thank the GPU donation from Nvidia and the AWS donation from Amazon.

We thank the Prostate Cancer DREAM Challenge organizers:

Justin Guinney (Sage Bionetworks); Tao Wang (UT Southwestern); Brian Bot (Sage Bionetworks); Christopher Bare (Sage Bionetworks); Mike Kellen (Sage Bionetworks); Stephen Jones (Covance); Elise Blaese

(IBM); Joseph Morrell (Project Data Sphere, LLC); Thea Norman (Sage Bionetworks); Kald Abdallah (Project Data Sphere, LLC); Stephen Friend (Sage Bionetworks); Yang Xie (UT Southwestern); James Costello (University of Colorado Anschutz Medical Campus); Gustavo Stolovitzky (IBM Research & Icahn School of Medicine at Mt. Sinai); Liz Zhou (Project Data Sphere, LLC)

We thank the Prostate Cancer DREAM Challenge data providers:

AstraZeneca; Celgene; Sanofi; Memorial Sloan Kettering Cancer Center.

AUTHOR CONTRIBUTIONS

Y.G. designed and implemented the best-performing algorithm in the Prostate Cancer DREAM Challenge. Post-challenge, K.D. and H.L. performed computational analysis of the observed and predicted data and analyzed model performance and feature importance. K.D. prepared figures. K.D., H.L., and Y.G. wrote this manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

Y.G. receives personal payment from Eli Lilly and Company, Genentech Inc, F. Hoffmann-La Roche AG, and Cleerly Inc; holds equity shares at Cleerly Inc and Ann Arbor Algorithms Inc; receives research support from Merck KGaA as research contracts and Ryss Tech as unrestricted donations.

Received: September 4, 2019
Revised: November 22, 2019
Accepted: December 19, 2019
Published: February 21, 2020

REFERENCES

Baillargeon, J., Kuo, Y.-F., Fang, X., and Shahinian, V.B. (2015). Long-term exposure to testosterone therapy and the risk of high grade prostate cancer. J. Urol. *194*, 1612–1616.

Berggren, M., Sittadjody, S., Song, Z., Samira, J.-L., Burd, R., and Meuillet, E.J. (2009). Sodium selenite increases the activity of the tumor suppressor protein, PTEN, in DU-145 prostate cancer cells. Nutr. Cancer *61*, 322–331.

Berthold, D.R., Pond, G.R., Soban, F., de Wit, R., Eisenberger, M., and Tannock, I.F. (2008). Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer: updated survival in the TAX 327 study. J. Clin. Oncol. *26*, 242–245.

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. *30*, 1145–1159.

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32, https://doi.org/10.1023/A:1010933404324.

Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data, *vol. 110*. (University of California, Berkeley), pp. 1–12.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: ACM), pp. 785–794.

Cookson, M.S., Lowrance, W.T., Murad, M.H., and Kibel, A.S.; American Urological Association (2015). Castration-resistant prostate cancer: AUA guideline amendment. J. Urol. *193*, 491–499.

Dai, Q., Motley, S.S., Smith, J.A., Jr., Concepcion, R., Barocas, D., Byerly, S., and Fowke, J.H. (2011). Blood magnesium, and the interaction with calcium, on the risk of high-grade prostate cancer. PLoS One *6*, e18237.

Fizazi, K., Fizazi, K.S., Higano, C.S., Nelson, J.B., Gleave, M., Miller, K., Morris, T., Nathan, F.E., McIntosh, S., Pemberton, K., and Moul, J.W. (2013). Phase III, randomized, placebo-controlled study of docetaxel in combination with zibotentan in patients with metastatic castration-resistant prostate cancer. J. Clin. Oncol. *31*, 1740–1747.

Guan, Y. (2019). Waking up to data challenges. Nat. Mach. Intell. *1*, 67.

Gupta, E., Guthrie, T., and Tan, W. (2014). Changing paradigms in management of metastatic castration resistant prostate cancer (mCRPC). BMC Urol. *14*, 55.

Halabi, S., Lin, C.-Y., Kelly, W.K., Fizazi, K.S., Moul, J.W., Kaplan, E.B., Morris, M.J., and Small, E.J. (2014). Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. J. Clin. Oncol. *32*, 671–677.

Heidenreich, A., Bastian, P.J., Bellmunt, J., Bolla, M., Joniau, S., van der Kwast, T., Mason, M., Matveev, V., Wiegel, T., Zattoni, F., and Mottet, N.; European Association of Urology (2014). EAU guidelines on prostate cancer. Part II: treatment of advanced, relapsing, and castration-resistant prostate cancer. Eur. Urol. *65*, 467–479.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems 30, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.), pp. 3146–3154.

Kim, Y., Jeong, I.G., You, D., Song, S.H., Suh, N., Jang, S.-W., Kim, S., Hwang, J.J., and Kim, C.-S. (2014). Sodium meta-arsenite induces reactive oxygen species-dependent apoptosis, necrosis, and autophagy in both androgen-sensitive and androgen-insensitive prostate cancer cells. Anticancer Drugs *25*, 53–62.

Klap, J., Schmid, M., and Loughlin, K.R. (2015). The relationship between total testosterone levels and prostate cancer: a review of the continuing controversy. J. Urol. *193*, 403–413.

Kristiyanto, D., Anderson, K.E., Hung, L.-H., and Yeung, K.Y. (2016). Predicting discontinuation of docetaxel treatment for metastatic castration-resistant prostate cancer (mCRPC) with random forest. F1000Res. *5*, https://doi.org/10.12688/f1000research.8353.1.

Langsenlehner, T., Thurner, E.-M., Krenn-Pilko, S., Langsenlehner, U., Stojakovic, T., Gerger, A., and Pichler, M. (2015). Validation of the neutrophil-to-lymphocyte ratio as a prognostic factor in a cohort of European prostate cancer patients. World J. Urol. *33*, 1661–1667.

Li, H., and Guan, Y. (2019). Machine learning empowers phosphoproteome prediction in cancers. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz639.

Machiels, J.-P., Mazzeo, F., Clausse, M., Filleul, B., Marcelis, L., Honhon, B., D'Hondt, L., Dopchie, C., Verschaeve, V., Duck, L., et al. (2008). Prospective randomized study comparing docetaxel, estramustine, and prednisone with docetaxel and prednisone in metastatic hormone-refractory prostate cancer. J. Clin. Oncol. *26*, 5261–5268.

Mearini, L., Zucchi, A., Nunzi, E., Villirillo, T., Bini, V., and Porena, M. (2013). Low serum testosterone levels are predictive of prostate cancer. World J. Urol. *31*, 247–252.

Pastuszak, A.W., Khanna, A., Badhiwala, N., Morgentaler, A., Hult, M., Conners, W.P., Sarosdy, M.F., Yang, C., Carrion, R., Lipshultz, L.I., and Khera, M. (2015). Testosterone therapy after radiation therapy for low, intermediate and high risk prostate cancer. J. Urol. *194*, 1271–1276.

Petrylak, D.P., Tangen, C.M., Hussain, M.H.A., Lara, P.N., Jr., Jones, J.A., Taplin, M.E., Burch, P.A., Berry, D., Moinpour, C., Kohli, M., et al. (2004). Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. N. Engl. J. Med. *351*, 1513–1520.

Sejima, T., Iwamoto, H., Masago, T., Morizane, S., Yao, A., Isoyama, T., Kadowaki, H., and Takenaka, A. (2013). Low pre-operative levels of serum albumin predict lymph node metastases and ultimately correlate with a biochemical recurrence of prostate cancer in radical prostatectomy patients. Cent. European J. Urol. *66*, 126–132.

Seyednasrollah, F., Koestler, D.C., Wang, T., Piccolo, S.R., Vega, R., Greiner, R., Fuchs, C., Gofer, E., Kumar, L., Wolfinger, R.D., et al. (2017). A DREAM challenge to build prediction models for short-term discontinuation of docetaxel in metastatic castration-resistant prostate cancer. JCO Clin. Cancer Inform. *1*, 1–15.

Sonpavde, G., Pond, G.R., Armstrong, A.J., Clarke, S.J., Vardy, J.L., Templeton, A.J., Wang, S.L., Paolini, J., Chen, I., Chow-Maneval, E., et al. (2014). Prognostic impact of the neutrophil-to-lymphocyte ratio in men with metastatic castration-resistant prostate cancer. Clin. Genitourin. Cancer *12*, 317–324.

Sun, Y., Selvaraj, S., Varma, A., Derry, S., Sahmoun, A.E., and Singh, B.B. (2013). Increase in serum Ca2+/Mg2+ ratio promotes proliferation of prostate cancer cells by activating TRPM7 channels. J. Biol. Chem. *288*, 255–263.

Sutton, C.D. (2005). 11-Classification and regression trees, bagging, and boosting. In Handbook of Statistics, *vol. 24*, C.R. Rao, E.J. Wegman, and J.L. Solka, eds. (Elsevier), pp. 303–329.

Sweeney, C.J., Chen, Y.-H., Carducci, M., Liu, G., Jarrard, D.F., Eisenberger, M., Wong, Y.N., Hahn, N., Kohli, M., Cooney, M.M., et al. (2015). Chemohormonal therapy in metastatic hormone-sensitive prostate cancer. N. Engl. J. Med. *373*, 737–746.

Tannock, I.F., de Wit, R., Berry, W.R., Horti, J., Pluzanska, A., Chi, K.N., Oudard, S., Théodore, C., James, N.D., Turesson, I., et al.; TAX 327 Investigators (2004). Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. N. Engl. J. Med. *351*, 1502–1512.

Templeton, A.J., Pezaro, C., Omlin, A., McNamara, M.G., Leibowitz-Amit, R., Vera-Badillo, F.E., Attard, G., de Bono, J.S., Tannock, I.F., and Amir, E. (2014). Simple prognostic score for metastatic castration-resistant prostate cancer with incorporation of neutrophil-to-lymphocyte ratio. Cancer *120*, 3346–3352.

Templeton, A.J., Vera-Badillo, F.E., Wang, L., Attalla, M., De Gouveia, P., Leibowitz-Amit, R., Knox, J.J., Moore, M., Sridhar, S.S., Joshua, A.M., et al. (2013). Translating clinical trials to clinical practice: outcomes of men with metastatic castration resistant prostate cancer treated with docetaxel and prednisone in and out of clinical trials. Ann. Oncol. *24*, 2972–2977.

Van Der Maaten, L., and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. J. Machine Learn. Res. *9*, 26.

Wang, Y., Chen, W., Hu, C., Wen, X., Pan, J., Xu, F., Zhu, Y., Shao, X., Shangguan, X., Fan, L., et al. (2017). Albumin and fibrinogen combined prognostic grade predicts prognosis of patients with prostate cancer. J. Cancer *8*, 3992–4001.

Wilson, K.M., Shui, I.M., Mucci, L.A., and Giovannucci, E. (2014). Calcium and phosphorus intake and prostate cancer risk: a 24-y follow-up study–. Am. J. Clin. Nutr. *101*, 173–183.

Yao, A., Sejima, T., Iwamoto, H., Masago, T., Morizane, S., Honda, M., and Takenaka, A. (2015). High neutrophil-to-lymphocyte ratio predicts poor clinical outcome in patients with castration-resistant prostate cancer treated with docetaxel chemotherapy. Int. J. Urol. *22*, 827–833.

iScience, Volume *23*

Supplemental Information

Treatment Stratification of Patients with Metastatic Castration-Resistant Prostate Cancer by Machine Learning

Kaiwen Deng, Hongyang Li, and Yuanfang Guan

Figure S1. The feature importances evaluated by delta-error in the random forest model, Related to Figure 4. (A) ASC; (B) CEL; (C) VEN

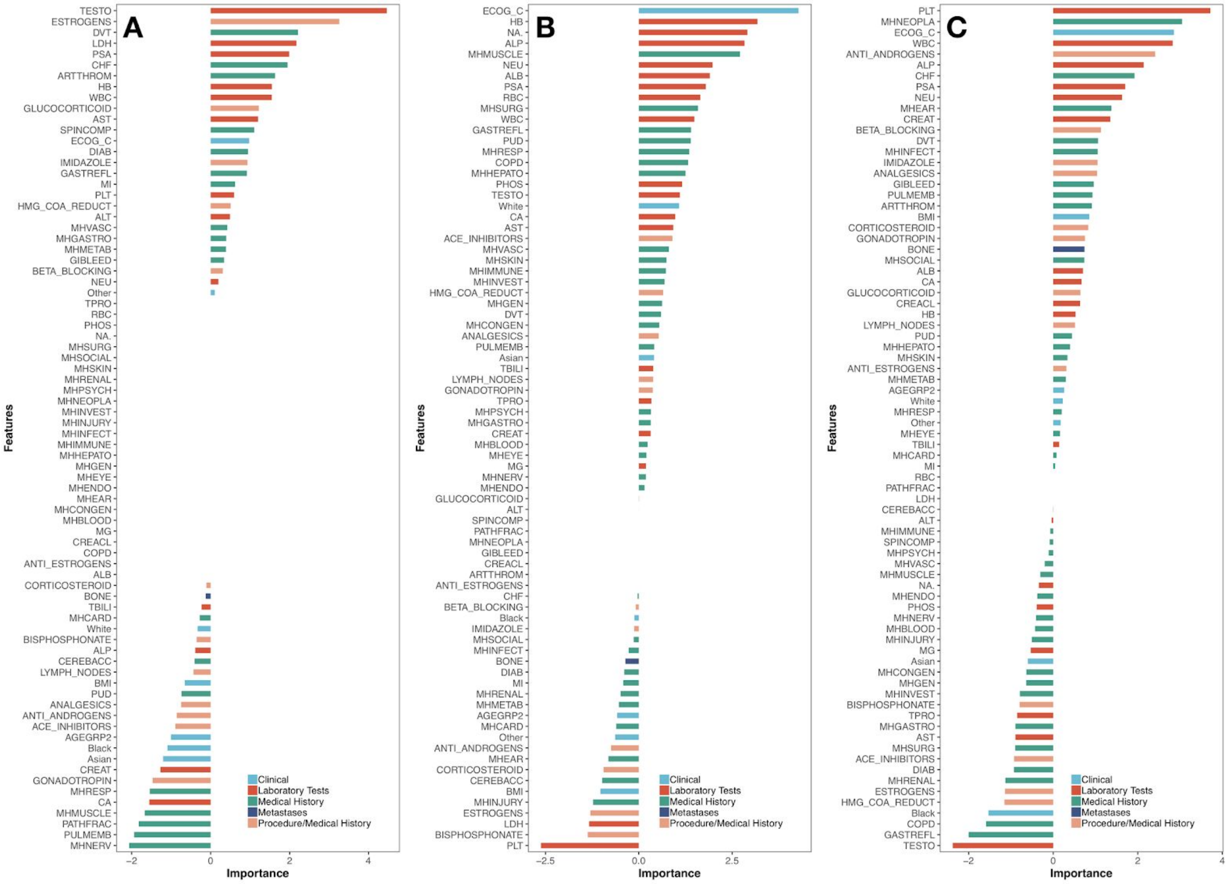


Figure S2. The proportions of missing of the features. Related to Figure 1

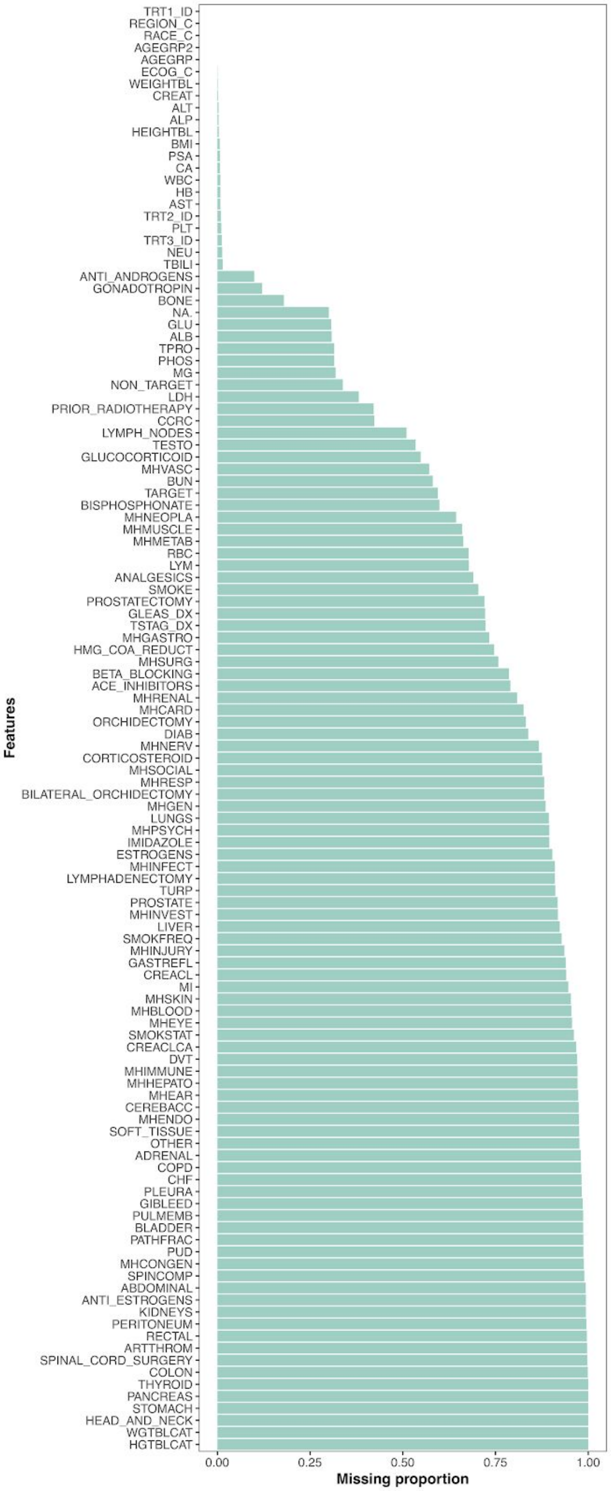


Figure S3. t-SNE map for three cohorts. Related to Figure 1

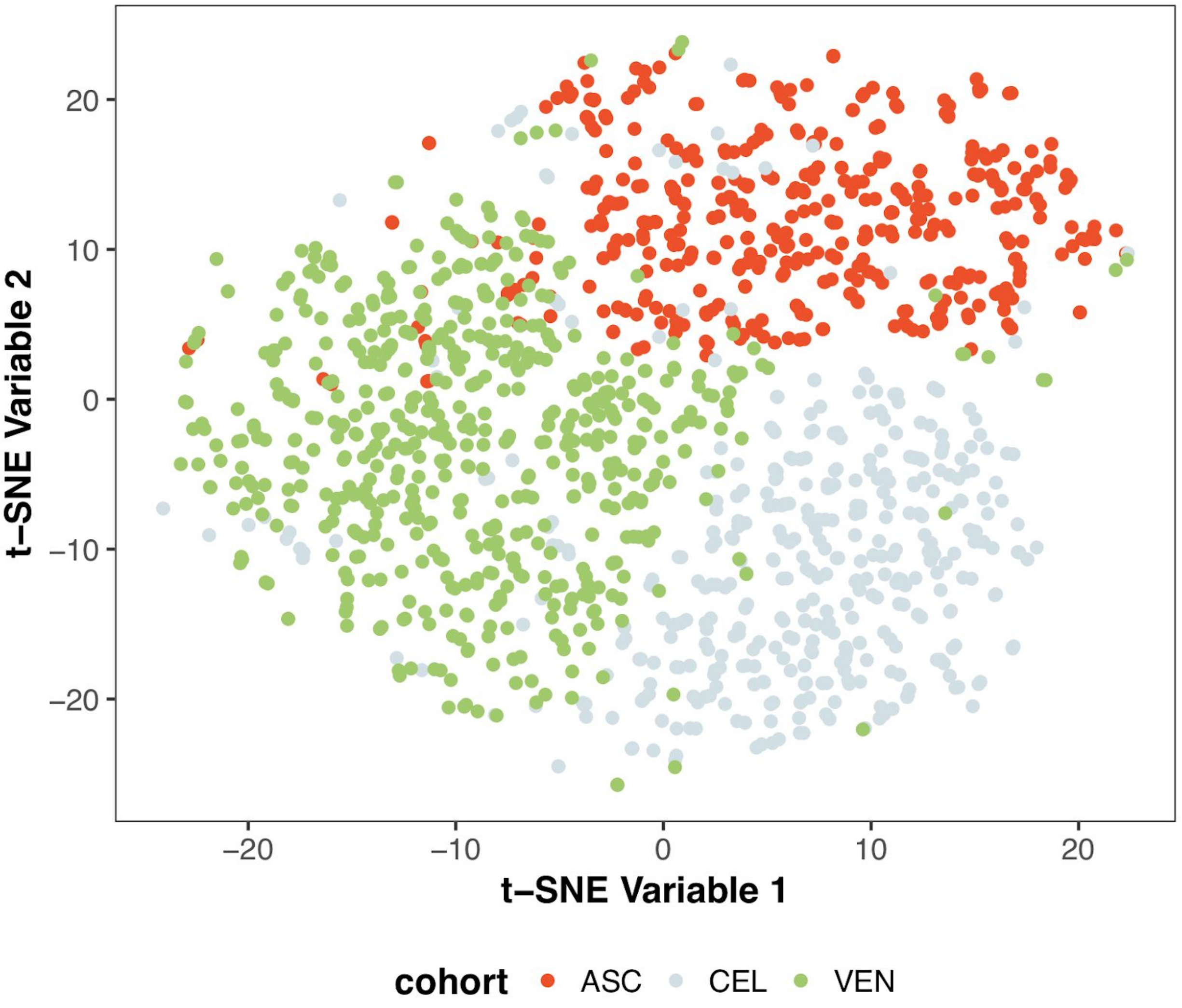


Table S1. Median performance scores of different base learners, Related to Figure 2

Curves	**Cohorts**	BAG-CART	Random Forest	Cox	Linear	Logistic	*Baseline*
AUC	Full Dataset	0.6146	0.6269	0.5633	0.5541	0.5608	0.5
	ASC	0.5545	0.5701	0.5386	0.5419	0.5360	0.5
	CEL	0.6562	0.6338	0.5304	0.5630	0.5424	0.5
	VEN	0.5104	0.5378	0.4906	0.4761	0.4863	0.5
AUPRC	Full Dataset	0.1999	0.1961	0.1527	0.1527	0.1527	0.1308
	ASC	0.2836	0.3080	0.2703	0.2649	0.2603	0.2146
	CEL	0.1792	0.1507	0.1250	0.1286	0.1273	0.0952
	VEN	0.0926	0.0959	0.0851	0.0803	0.0880	0.0885

Table S2. Median performance scores of models with different gold standards, Related to Figure 3

Curves	**Cohorts**	DISCONT	NEW	*Baseline*
AUC	Full Dataset	0.6269	0.6356	0.5
	ASC	0.5701	0.5726	0.5
	CEL	0.6338	0.6420	0.5
	VEN	0.5378	0.5470	0.5
AUPRC	Full Dataset	0.1961	0.2001	0.1309
	ASC	0.3080	0.3089	0.2146
	CEL	0.1507	0.1598	0.0952
	VEN	0.0959	0.1006	0.0885

Table S3. Correlations Between Features and Discontinuation. Related to Figure 4

Feature	Correlation Estimation	Correlation p-value
ALB	-0.0552	0.0331
NA.	-0.0438	0.0904
TPRO	-0.0039	0.8784
MG	-0.0083	0.7468
TESTO	0.0092	0.7232
WBC	0.0175	0.4993
PHOS	-0.0417	0.1073
MHVASC	-0.0433	0.0948
NEU	0.0126	0.6277
MHSOCIAL	-0.0582	0.0246
rf.model	**0.1329**	**2.6334e-07**

Table S4. Information and characteristics of clinical features, Related to Figure 1
(See the Supplementary Dataset Table_S4.csv)

Transparent Methods

Data Collection

Data were collected from the provider-deidentified comparator arm datasets of phase III prostate cancer clinical trials, including ASCENT2 (ASC) from Memorial Sloan Kettering Cancer Center, with 105 patients discontinuing docetaxel due to adverse event or possible adverse event (Scher et al., 2011), CELGENE (CEL) from Celgene, with 41 discontinued patients (Petrylak et al., 2015), and EFC6546 (VEN) from Sanofi, with 51 discontinuations (Tannock et al., 2013) (Table 1).

All of the competition data can be accessed at:
https://www.synapse.org/#!Synapse:syn2813558/wiki/209590
The codes for our model are available at:
https://github.com/GuanLab/prostate_discontinuation

Feature pre-processing

The list of all features used in this study is shown in Table S4. The nominal features were converted into 0 and 1 based on their missing status. Other missing values in each feature were filled by average across all samples.

Our operations in special cases are listed below:

- LDH: we converted this feature into 0 and 1 with the threshold of 250 after filling in the missing values.
- BONE: this feature was labeled referring to other 11 features including "RECTAL", "KIDNEYS", "LUNGS", "LIVER", "PLEURA", "ADRENAL", "BLADDER", "COLON", "STOMACH", "PANCREAS" and "ABDOMINAL". For each record:
    - if all of them were missing, gave “BONE” 0;
    - if all but “BONE” were missing, gave 1;
    - if all but “BONE” were not missing, gave 2.
- RACE_C: we separated it into four dummy variables: “White”, “Asian”, “Black” and “Other”. “Hispanic” was categorized into “Other”. Missing values were also filled by the average of each category.

Features were normalized by Min-Max Normalization (Gopal Krishna Patro & Sahu, 2015).

$$normalized\,feature(x_i) = \frac{x_i - min(x)}{max(x) - min(x)}$$

AUC and AUPRC

Since the discontinuation status is highly unbalanced with only a small portion of patients have early discontinuation (22%, 8% and 10% in ASC, CEL and VEN), the differences are not apparent in the ROC space. This is because the number of negative cases largely exceeds the number of positive ones, thus the false positive (FP) rate will change slightly even when FP increase largely. But precision is able to capture the differences by calculating the true positive (TP) overall positive prediction. Therefore, the precision-recall (PR) curve and AUPRC are commonly used in the evaluation of predictive performance in unbalanced data accomplished with AUC (Li, Li, Quang, & Guan, 2018; Li, Quang, & Guan, 2018). Both AUC and AUPRC are calculated using the R package "PRROC" (Grau, Grosse, & Keilwagen, 2015). The R version is 3.4.0 (2017-04-21).

New gold-standard assembling

The new gold standard was constructed based on early DEATH (death status within three months), DISCONT (discontinuation status), and ENDTRS_C (treatment status). All non-missing records in DEATH and ENDTRS_C were converted into 0 and 1 following their labels, where "YES", "AE" and "possible AE" were set as 1, and others 0. Only when all of the three variables equaled 0, was the new gold standard labeled as 0, otherwise, 1.

$$New\ standard = max(death\ status,\ discontinuation,\ treatment\ status)$$

Model construction and performance evaluation

Models were constructed by R with packages or functions ipred (Peters, Hothorn, & Lausen, 2002), randomForest (Liaw & Wiener, 2002), survival (Therneau, 2016), lm() and glm(family = binomial(link = "logit")) respectively, with default settings.

Models were tested by cross-validation across cohorts and 10 times 5-fold cross-validation for each cohort and they were stacked to test on the final test cohort in the DREAM challenge for performance validation. Their performances were evaluated by the ROC curve and PR curve with package PRROC (Grau et al., 2015). The areas under the curves were visualized by ggplot2 (Wickham, 2016).

The batch effect among the cohorts was analyzed and visualized by t-distributed stochastic neighbor embedding using package Rtsne (Krijthe, 2015).

Feature importance

Since RF achieved the best performance for predicting the discontinuation of docetaxel treatment, we estimated the feature importance based on the RF model instead of other models. In particular, we used the out-of-bag methods to estimate the feature importance (Breiman, 2001; Li, Panwar, Omenn, & Guan, 2017). For each feature, we calculated the increase of the prediction error ("delta-error") for the RF model without this feature. As a result, a feature is "important" if the out-of-bag "delta-error" is large. In this study, we performed this feature importance estimation using the R package “caret” (Kuhn, 2015) with 10 times 5-fold cross-validation, and the scoring metric for “delta-error” was the mean squared error (MSE). This method was employed on the full dataset and three subsets (ASC + CEL; CEL + VEN; VEN + ASC). The features are re-ordered by the errors from high to low.