METHOD ARTICLE

An ensemble-based Cox proportional hazards regression framework for predicting survival in metastatic castration-resistant prostate cancer (mCRPC) patients [version 1; referees: 1 approved, 2 approved with reservations]

Richard Meier[1*], Stefan Graw[1*], Joseph Usset[1], Rama Raghavan[1], Junqiang Dai[1], Prabhakar Chalise[1], Shellie Ellis[2], Brooke Fridley[1], Devin Koestler[1]

[1]Department of Biostatistic, University of Kansas Medical Center, Kansas City, KS, USA
[2]Department of Health Policy and Management, University of Kansas Medical Center, Kansas City, KS, USA

[*] Equal contributors

v1 **First published:** 16 Nov 2016, **5**:2677 (doi: 10.12688/f1000research.8226.1)
Latest published: 16 Nov 2016, **5**:2677 (doi: 10.12688/f1000research.8226.1)

Abstract

From March through August 2015, nearly 60 teams from around the world participated in the Prostate Cancer Dream Challenge (PCDC). Participating teams were faced with the task of developing prediction models for patient survival and treatment discontinuation using baseline clinical variables collected on metastatic castrate-resistant prostate cancer (mCRPC) patients in the comparator arm of four phase III clinical trials. In total, over 2,000 mCRPC patients treated with first-line docetaxel comprised the training and testing data sets used in this challenge. In this paper we describe: (a) the sub-challenges comprising the PCDC, (b) the statistical metrics used to benchmark prediction performance, (c) our analytical approach, and finally (d) our team's overall performance in this challenge. Specifically, we discuss our curated, ad-hoc, feature selection (CAFS) strategy for identifying clinically important risk-predictors, the ensemble-based Cox proportional hazards regression framework used in our final submission, and the adaptation of our modeling framework based on the results from the intermittent leaderboard rounds. Strong predictors of patient survival were successfully identified utilizing our model building approach. Several of the identified predictors were new features created by our team via strategically merging collections of weak predictors. In each of the three intermittent leaderboard rounds, our prediction models scored among the top four models across all participating teams and our final submission ranked 9th place overall with an integrated area under the curve (iAUC) of 0.7711 computed in an independent test set. While the prediction performance of teams placing between 2nd- 10th (iAUC: 0.7710-0.7789) was better than the current gold-standard prediction model for prostate cancer survival, the top-performing team, FIMM-UTU significantly outperformed all other contestants with an iAUC of 0.7915. In summary, our ensemble-based Cox regression framework with CAFS resulted in strong overall performance for predicting prostate cancer survival and represents a promising approach for future prediction problems.

Open Peer Review

Referee Status: ? ✓ ?

	Invited Referees		
	1	**2**	**3**
version 1 published 16 Nov 2016	? report	✓ report	? report

1 **Stephen R. Piccolo**, Brigham Young University USA

2 **Ka Yee Yeung**, University of Washington USA, **Daniel Kristiyanto**, University of Washington USA

3 **Russell Greiner**, University of Alberta Canada, **Luke Kumar**, University of Alberta Canada

Discuss this article

Comments (0)

This article is included in the DREAM Challenges channel.

Corresponding author: Devin Koestler (dkoestler@kumc.edu)

How to cite this article: Meier R, Graw S, Usset J *et al.* **An ensemble-based Cox proportional hazards regression framework for predicting survival in metastatic castration-resistant prostate cancer (mCRPC) patients [version 1; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2016, **5**:2677 (doi: 10.12688/f1000research.8226.1)

Grant information: This work was supported by the National Institute of Health (NIH) grants: 1KL2TR000119 Kansas IDeA Network of Biomedical Research Excellence (K-INBRE) Bioinformatics Core supported in part by the National Institute of General Medical Science award P20GM103418, and the Biostatistics and Informatics Shared Resource supported in part by the National Cancer Institute Cancer Center Shared Grant award P30 CA168524.

Competing interests: No competing interests were disclosed.

First published: 16 Nov 2016, **5**:2677 (doi: 10.12688/f1000research.8226.1)

Introduction

Today, prostate cancer is one of the most prevalent cancers afflicting men in the Western world. In addition to the prevalence of this disease, the mortality rates for prostate cancer ranked fifth among the most common causes of cancer death worldwide in 2012 (http://www.cancerresearchuk.org/). In the US alone, approximately 137.9 out of 100,000 men were diagnosed with prostate cancer each year from 2008–2012, with an average annual mortality rate of 21.4 out of 100,000 men. (http://www.seer.cancer.gov/statfacts/html/prost.html). According to the Cancer Prevalence and Cost of Care Projections, the total annual cost of prostate cancer in 2016 has been estimated at 14.3 billion dollars (http://www.costprojections.cancer.gov/).

Over the course of the last decade in the US, approximately 15% of prostate cancer cases were initially diagnosed with metastatic disease (stage IV). Androgen deprivation therapy (ADT) is the established treatment for these cases, but one third of patients develop resistance and their disease progresses to metastatic castrate-resistant prostate cancer (mCRPC) (https://www.synapse.org/ProstateCancerChallenge). Treatment of mCRPC has been historically challenging, and while docetaxel – the current front-line therapy for mCRPC – has been effective at improving mCRPC survival at the population level, a significant fraction of patients do not respond to treatment or prematurely discontinue treatment due to adverse events (AE)[1], leading to substantial variation in the individual outcomes between mCRPC patients. For this reason, and because of the tremendous personal, societal, and economic burden associated with this disease, there is significant interest both in the identification of individual predictors for mCRPC prognosis as well as the development of prognostic models that can be used to identify high-risk mCRPC patients.

In a recent publication[2], Halabi *et al.*, utilized data from a phase III trial consisting of over one thousand mCRPC patients to develop and test a prognostic model for overall survival among patients receiving first-line chemotherapy. The time dependent area under the curve (tAUC) was > 0.73 in both testing and independent validation data sets, suggesting strong performance of the Halabi *et al.* model for identifying low- and high-risk mCRPC patients. Notwithstanding the significant advances made by Halabi *et al.*, and others toward the development of accurate prognostic models for mCRPC outcomes[2–4], there remains ample room for improved prediction performance.

Motivated by the potential to further improve existing risk-prediction tools along with growing worldwide burden of prostate cancer, the Prostate Cancer Dream Challenge was launched in March 2015 and included the participation of nearly 60 teams from around the world. The Prostate Cancer Dream Challenge was composed of two distinct sub challenges; in sub challenge 1, teams competed in the development of prognostic models for predicting overall survival based on baseline clinical variables, whereas the objective of sub challenge 2 involved the development of models to predict short-term treatment discontinuation of docetaxel (< 3 months) due to adverse events (AE). To assist in the development and testing of prediction models, approximately 150 variables collected on over 2,000 mCRPC patients treated with first-line docetaxel in one of four different phase III clinical trials were used. Three of the four trials were combined to generate the training data set, which was used for model-building and development, while data from the remaining trial were withheld from challenge participants and used as an independent test set to evaluate each of the submitted models[5].

In the present manuscript, we focus exclusively on our methodological approach to sub challenge 1. Broadly speaking, the first step of our team's approach to sub challenge 1 involved an initial screening of the data: data cleaning and processing, creation of new variables from existing data, imputation and/or exclusion of variables with missing values, and normalization to standardize the data across trials. The final "cleaned and standardized" training data was then used to fit to an ensemble of multiple Cox proportional hazards regression models whose constituent models were developed using curated, ad-hoc, feature selection (CAFS). Models developed by our team were subjected to internal cross-validation within the training data set to identify instances of model overfitting and to assist in further refinements to our prediction models. The source code utilized for our approach can be accessed via the Team Jayhawks Prostate Cancer Dream Challenge project web page (https://www.synapse.org/#!Synapse:syn4214500/wiki/231706) or directly via the GitHub repository webpage (https://github.com/richard-meier/JayhawksProstateDream).

Materials and methods

Data

A detailed description of the datasets used in this challenge can be found on the Prostate Cancer Dream Challenge web page (https://www.synpase.org/ProstateCancerChallenge). Briefly, the training set originated from the ASCENT-2 (Novacea, provided by Memorial Sloan Kettering Cancer Center), MAINSAIL (Celgene) and VENICE (Sanofi) trials[6–8]. For the 1600 patients in the training data, baseline covariate information and clinical outcomes (i.e. time to death and time to treatment discontinuation) were provided to participating teams for the purposes of model development and training. Although baseline covariate information for a subset of patients in the ENTHUSE-33 (AstraZeneca) trial[9] scoring set was provided to participating teams (n = 157), the clinical outcomes for each of these patients were censored and withheld from teams throughout the duration of the challenge. Specifically, the ENTHUSE-33 data set (n = 470) was split into two disjoint sets that consisted of 157 and 313 patients. Whereas an undisclosed randomly selected subset of the 157 patients was used for model evaluation in each intermittent leaderboard round, the remaining 313 patients were withheld completely from participating teams and used only in the final scoring round.

Preprocessing

All aspects of our approach, from data preprocessing to model development and cross-validation, were implemented using R version 3.2.1 (2015-06-18) (https://www.r-project.org/). Baseline covariate information on subjects comprising the training data were reformatted and normalized according to the type of variable (i.e., categorical, ordinal, numeric) and feature type (i.e., medical history, laboratory values, etc). Cleaned and normalized baseline features were then used to derive additional novel risk predictors. (https://github.com/richard-meier/JayhawksProstateDream/blob/master/dataCleaningMain.R)

Several groups of binary variables representing patient specific clinical information and prior medical history reported on patients were merged into new features. Three different merging types were explored: "logical or", regular summation, and z-score weighted summation. For the latter, each individual feature in the training set was fit against survival time with a Cox proportional hazards model and their resulting z-scores were used to derive weights that were proportional to each variable's strength of association with survival (https://github.com/richard-meier/JayhawksProstateDream/blob/master/deriveHardcodedWeights.R). Summation variables were created for 3 main categories: medical history information, prior medication information and metastasis information. For each of these categories, new variables generated by merging specific subcategories (i.e. protective, harmful, total, visceral, etc.) were created.

A participant's target lesion volume (TLV) was generated by multiplying each target lesion by its size, followed by summing over all lesions within that participant (https://github.com/richard-meier/JayhawksProstateDream/blob/master/src/lesion_volume.R). To impute the TLV for the ASCENT-2 trial, we calculated the average TLV per lesion within individuals who survived or died in the MAINSAIL or VENICE trials, and multiplied these separate averages by the number of non-bone lesions found in the ASCENT-2 data. To classify whether for each category a feature was in the subcategory "protective" or "harmful", their z-scores, when individually fitting against the outcome, were used. A feature was labeled "protective" if its z-score was greater than 1.64 and "harmful" if its z-score was smaller than -1.64.

Principal component analysis (PCA) was used to split numerical laboratory values into components that best explained their variation (see above: "deriveHardcodedWeights.R"). The top PCs were then treated as new features. In order to address issues or findings involving some specific variables, additional features were created: The ECOG performance status score was both included as continuous and categorical variable. Age groups were also recoded as an ordinal age risk variable for which 0 represented patients older than 75 years, 1 represented patients younger than 65 years and 2 represented patients with ages between 65–75 years. The latter was motivated by our observation of a non-linear trend between age and survival time.

Race was recoded into a binary variable where 1 referred to patients labeled as "white" or "other" and 0 represented patients that did not fall into one of those two categories (e.g. "black", "asian", etc.). The features "harm_pro" and "harm_pro2" were created by fitting the summation variables of the medical history subgroups "harmful" and "protective" against the outcome and obtaining the z-scores of these subgroup summation variables. The difference between the two features was that harm_pro exclusively fitted the two summation variables, whereas harm_pro2 also utilized a set of important predictor variables for the initial fit. The two z-score weighted sums (corresponding to the two sets of features utilized for the previously mentioned fit) of these summation variables then correspond to the two new features. (https://github.com/richard-meier/JayhawksProstateDream/blob/master/src/add_additional_features.R)

Model building and feature selection

Our methodological framework utilized an ensemble of Cox proportional hazards regression models that were found to be individually competitive in predicting survival. For each patient, the ensemble-based risk scores were generated as a weighted sum of the individually estimated risk scores from separate Cox-regression models, fit using the "coxph" function in the "survival" R-package[10] (Figure 1C). Feature selection among the competitive risk-prediction models that constituted our ensemble was undertaken by a method we call curated, ad-hoc, feature selection (CAFS). This method attempts to maximize the prediction performance of a given model by iteratively including and excluding features from a baseline initial model. The method is greedy in the sense that in each step of the algorithm, only the model candidates that achieve the current "local best" performance are selected. Each iteration started with a group of experts making two executive decisions based on a set of possible model candidates for which performance was evaluated in prior iterations. First, one model was nominated as the best current model and a decision was made whether to expand or shrink the model, or terminate the procedure and keep the model, in its current form (Figure 1A). Choosing the current best model was guided by a candidate's estimated performance, performance of the previous best model, as well as knowledge of the researchers as to whether the form and components of a given model were reasonable in the context of the problem at hand. An example for the latter case would be that a newly introduced interaction term between completely unrelated features might be rejected after evaluation, even though it technically achieved the current best performance.

Model reduction was done via ad-hoc backward selection (Figure 1B). In this procedure a set of new models was generated by individually excluding each parameter or feature present in the current model. For each of these models, performance was evaluated based on a previously chosen optimization criterion, i.e., integrated time-dependent area under the curve (iAUC). The criterion was estimated via a cross-validation procedure in which the training set was repeatedly split into two random subsets of a fixed size. The first subset was used to estimate parameters of a given model, whereas the second subset was used to predict the outcome using the previously estimated parameters and to calculate the optimization criterion based on comparing the prediction with the true outcome. In our study, we utilized two-thirds for the parameter estimation subset, i.e., first subset, while the remaining one third comprised the second subset. The average of the calculated optimization criterion values, obtained from all random splits, served then as a performance estimate. We used 10,000 cross-validation steps for each model in our study to ensure stability of the average

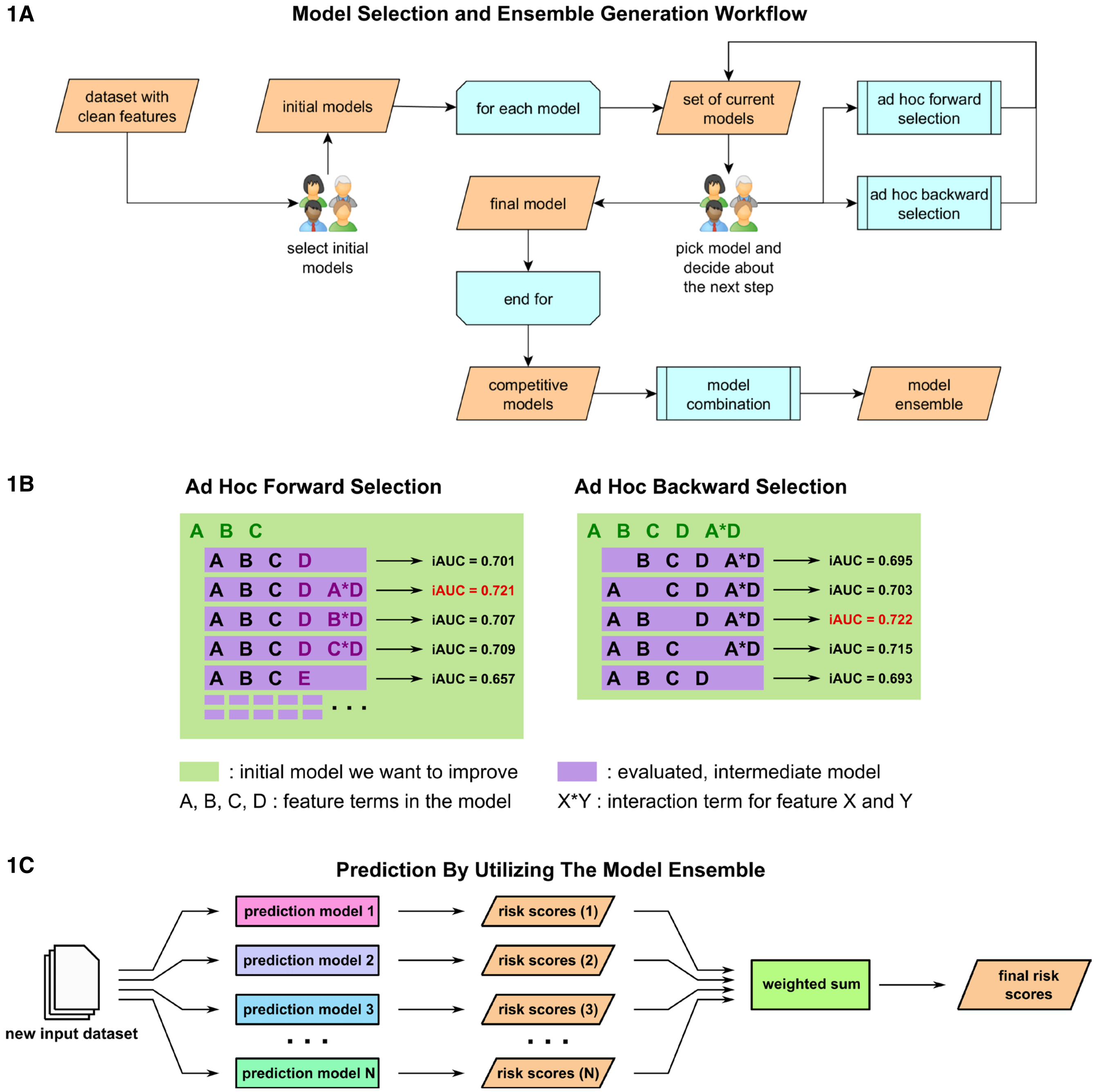


Figure 1. Model building and model ensemble utilization. (**1A**) Competitive prediction models were built individually by a curated, ad-hoc feature selection procedure. In each step researchers picked a new best model from the set of current models based on an optimization criterion and decided how it would be processed. (**1B**) Models were optimized by either forward selection, in which a new feature was added, or backward selection, in which a feature that had become obsolete was removed. Both selection methods generated a set of new models for which performance was predicted via in-depth cross-validation. (**1C**) Once a variety of competitive prediction models had been created, models were combined into an ensemble, which averaged their individual predictions in order to increase performance.

performance. The new models and performance estimates were then used as the basis for subsequent iterations.

Expansion of a model was accomplished using an ad-hoc forward selection procedure (Figure 1B). In this procedure several new models were created for each feature within the feature space. Each subset of new models contained one base model that included only main effect terms for new features, i.e., no interaction terms included. All other models in the subset further expanded this base model by individually introducing an interaction term with each element already in the previous best model.

Performance of each new model was again assessed via the cross-validation procedure. Since this step iterated over the feature space, it created a large amount of different models. To make this step computationally feasible, the number of cross-validation iterations had to be reduced. In our study, 500 cross-validation steps per new model were utilized. (https://github.com/richard-meier/JayhawksProstateDream/blob/master/src/modelTuning.R)

Finally, since the variances of these performance estimates were much higher than in the shrinkage step, the top 30 performing models were chosen and performance was re-estimated via 10,000 fold cross-validation. This set of new models and performance estimates was then used in the next iteration. Once iterations provided only marginal performance increases, the procedure was terminated and a final model was declared. Different models for the ensemble were found either by choosing different intermediate models as the current best and branching off a certain path, or by choosing different initial models.

Model evaluation

Each of the sub challenges in the Prostate Cancer Dream Challenge had its own prediction scoring metrics. In sub challenge 1A, participants were asked to submit a global risk score and time dependent risk scores, optimized for 12, 18 and 24 months. These risk scores were evaluated utilizing two scoring metrics: a concordance index (cIndex), and an integrated time dependent area under the curve (iAUC; 6–30 months). The time specific risk scores were assessed using AUC's computed using Hung and Chiang's estimator of cumulative AUC[11]. In sub challenge 1B, participants were asked to predict the time to event (death). The predictions of time to event were scored utilizing the root mean squared error (RMSE), using patients with known days to death.

When applying CAFS, we utilized the iAUC calculated from the predicted risk scores as an optimization criterion. This measure was also used by the challenge organizers for performance assessment in the scoring rounds for sub challenge 1A. While participants were asked to predict the risk score for overall survival based on patients' clinical variables, they were also tasked to predict the time to event (TTE) in sub challenge 1B. We used the risk score for each patient to model the TTE:

$$TTE_i = f(riskScore_i) + \epsilon_i$$

Where $\texttt{riskScore}_i$ corresponds to the risk score calculated in sub challenge 1A for the ith patient and f is an unknown smoothing function. We estimated f using a Generalized Additive Model (GAM) via the "gam" function within the "mgcv" package in R[12]. When regressing TTEs on risk we used only the subset of individuals who died.

Results

The principal component analysis with all laboratory values revealed that the first principal component was highly correlated with patient survival. Furthermore, across all laboratory values, only a subset of six features (baseline levels of: albumin, alkaline phosphatase, aspartate aminotransferase, hemoglobin, lactate dehydrogenase and prostate specific antigen) contributed significantly to explaining the variation in said first component. Thus, in the first PC only these six laboratory values were used during model building and development. In addition to the first principal component, several other newly created metavariables were identified as clinically relevant predictors by our model building procedure. Three z-score weighted sums merging metastases locations, medical history and prior medication were included in our prediction models. The "logical or" merged variable, whether or not a patient had any known medical history issues, was also utilized. The protective versus harmful subcategorization was only included in the models in the form of the sum of protective medical history features. However, this category only included a single feature, vascular disorders (yes/no).

We developed 5 competitive prediction models (M1 – M5) that were used in our Cox proportional hazards regression ensemble (Figure 2). All models were developed by either refining a previous model via CAFS or by building a model from the bottom up via CAFS. M1 used the best model found by manually selecting promising features as its initial model. M2 used an intermediate model from the CAFS procedure of M1 to deliberately branch off and provide a similar, yet different model. M3 and M5 were both built by using an initial model solely utilizing the strong predictors target lesion volume and principal component 1, but branching off in early iterations. M4 was built by using an initial model utilizing target lesion volume and the alkaline phosphatase level under the restriction that principal component 1 was excluded from the feature space.

While no single feature was utilized in every model M1–M5, five different features were shared between four models, six features between three models, four features between two models and eight features were unique to a model (Figure 2A). Each model had at least one unique feature. Between two and four interaction terms (two-way interaction terms) were present in all of the observed models (Figure 2B). One interaction was shared between the models M3, M4 and M5, while two interactions were shared between two models M1 and M2. Including components of newly derived features, eight features that were included in the original model by Halabi *et al.* in some form, were also utilized in the model ensemble. In total, the ensemble contained 38 coefficients, out of which 11 were pairwise interaction terms across all models.

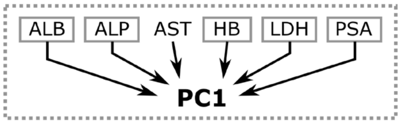
The estimated iAUC during performance assessment was found to be stable up to approximately three decimals when using 10,000 fold cross-validation. Similar estimated performance within the range of 0.005 iAUC difference was achieved between the competitive prediction models, the highest total iAUC being

2A

M1	M2	M3	M4	M5	Feature type	Description of the information type
🟩 ×	🟩 ×	🟩	🟩	🟩	□ AGE	Patient age group (3 groups): years coded as 18-64, 65-74, 75+
🟨	🟨	🟨 ×	🟨 ×	🟨 ×	□ ALP	Alkaline Phosphatase level
🟦	🟦	🟦	🟦 ×	🟦	AST	Aspartate Aminotransferase level
🟦	🟦 ×	🟦	🟦	🟦	BISPH	Prior Bisphosphonate medication (Yes / No)
🟥 ×	🟥 ×	🟥 ×	🟥 ×	🟥	□ ECOG	ECOG patient performance status (0,1,2,3,4)
🟨 ×	🟨 ×	🟨 ×	🟨	🟨	ESTRO	Prior Estrogen medication (Yes / No)
🟦	🟦	🟦 ×	🟦	🟦	GLCC	Prior Glucocortocoid medication (Yes / No)
🟩	🟩	🟩 ×	🟩 ×	🟩	HAPRO	Z-score metavariable merging "harmful" and "protective" medical history sums
🟨 ×	🟨	🟨	🟨 ×	🟨 ×	□ HB	Hemoglobin level
🟦	🟦	🟦	🟦 ×	🟦	□ LDH	Lactate Dehydrogenase level
🟥	🟥 ×	🟥 ×	🟥 ×	🟥 ×	MHIST_OR	Logical or merged medical history data (asks if any medical history issues are known)
🟥	🟥 ×	🟥 ×	🟥 ×	🟥 ×	MHIST_ZW	Z-score metavariable merging medical history data
🟨 ×	🟨 ×	🟨	🟨 ×	🟨	METAS_ZW	Z-score metavariable merging metastases location data
🟩 ×	🟩 ×	🟩	🟩	🟩	MI	Myocardial infarction diagnosed in medical history (Yes / No)
🟦 ×	🟦	🟦	🟦	🟦	NA.	Sodium level
🟨	🟨 ×	🟨	🟨 ×	🟨 ×	NEU	Neutrophil level
🟥 ×	🟥 ×	🟥 ×	🟥	🟥 ×	⬚ PC1	Principal component 1 of the most relevant laboratory values
🟨	🟨	🟨 ×	🟨 ×	🟨 ×	PHOS	Phosphorus level
🟦	🟦	🟦	🟦	🟦 ×	PRIMED_ZW	Z-score metavariable merging prior medication information
🟦	🟦	🟦 ×	🟦	🟦	PROST	Prostate lesions present (Yes / No)
🟦 ×	🟦	🟦	🟦	🟦	PROT_MH	"Protective" medical history diagnosis (Vascular Disorders: Yes / No)
🟩 ×	🟩 ×	🟩	🟩	🟩	□ RACE	Race information: white (Yes / No)
🟥	🟥 ×	🟥 ×	🟥 ×	🟥 ×	TLV	Target lesion volume: metavariable merging target lesions by volume
Features only in models as part of PC1					□ ALB	Albumin level
					□ PSA	Prostate Specific Antigen level

Feature in 🟥 4/5 🟨 3/5 🟩 2/5 🟦 1/5 of models

□ Feature was also used in the model proposed by Halabi et al. [2]



2B

Model	Parameters	Interaction terms		iAUC (cv)
M1	14	4	AGE*MI, METAS_ZW*RACE, METAS_ZW*NA., METAS_ZW*HB	0.743
M2	15	3	AGE*MI, METAS_ZW*RACE, NEU*BISPH	0.745
M3	12	2	MHIST_OR*PHOS, ALP*GLUCOCORTICOID	0.743
M4	14	2	MHIST_OR*PHOS, ALP*ECOG	0.740
M5	12	4	MHIST_OR*PHOS, TLV*NEU, PC1*ALP, PRIMED_ZW*HB	0.741
Ensemble	38	11		0.757

Figure 2. **Generated models utilized in the final challenge submission.** (**2A**) The ensemble consisted of five different models, M1 to M5, which ended up sharing many feature types even though they were individually generated under different conditions. (**2B**) All models made use of a similar number of parameters and achieved comparable performance in cross-validation. Performance further increased when using the model ensemble.

0.745. Optimal weights were chosen based on randomly initializing weights 100 times and estimating performance. Performance tended to be optimized the smaller the maximum pairwise difference between weights in an ensemble was. The best possible performance was estimated when choosing equal weights for all models. This ensemble was chosen as the best model. Utilizing the ensemble led to an estimated performance increase of 0.012 iAUC.

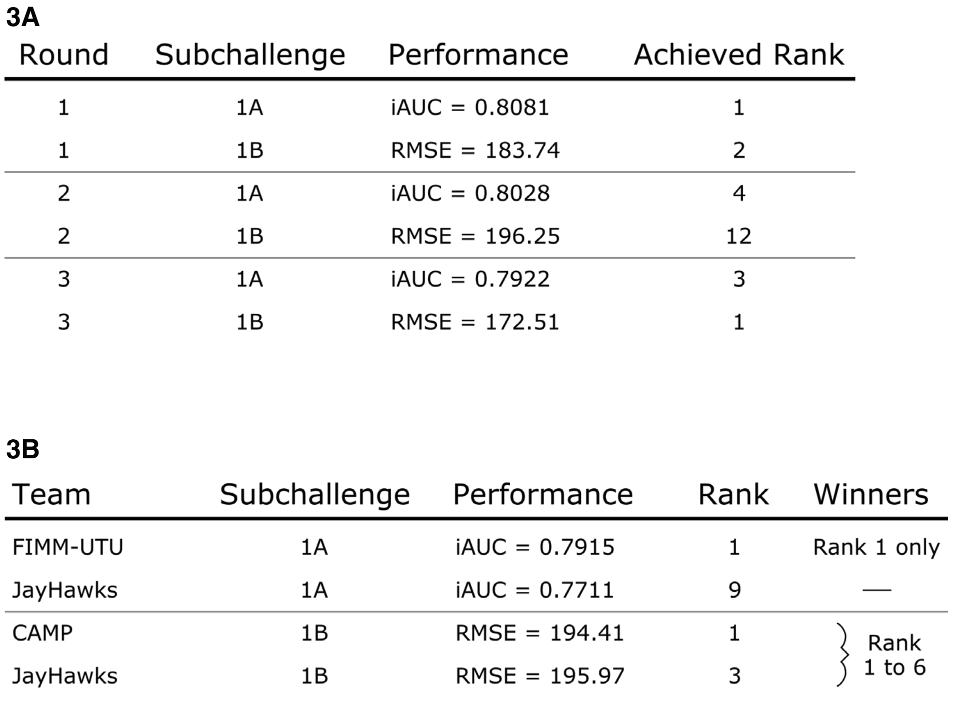
During the three leaderboard rounds the team explored and submitted various methodologies. Top performing submissions were always Cox proportional hazards models that outperformed more sophisticated approaches such as generalized boosted regression models and random survival forests. From scoring round 2 onward, single models utilizing CAFS were also submitted. In all intermittent leaderboard rounds, at least one of our submitted entries ranked among the top 4 performing models of sub challenge 1A (Figure 3A). In sub challenge 1B, at least one submission was within the top 3 performing models, with the exception of the second leaderboard round were our best model ranked number 12. Our models achieved performances ranging from 0.792 to 0.808 iAUC in 1A and from 172.51 to 196.25 RMSE in 1B. In the final scoring round, team FIMM-UTU[5] significantly outperformed all other contestants with an iAUC of 0.7915 (Figure 3B). Our submission for 1A that utilized the CAFS ensemble achieved rank 9 with an iAUC of 0.7711. The performances of teams ranking from 2nd to 10th were very similar. While the difference in performance between rank 1 and 2 was 0.0126 iAUC, the difference in performance between our method and rank 2 was only 0.0078 iAUC. Our submitted model ensemble also successfully outperformed the previous best model by Halabi *et al.*[2], which was placed at rank 36 with an iAUC of 0.7429. Sub challenge 1B was won simultaneously by 6 teams out of which our method achieved rank 3.

Discussion

Many feature types present in the original model by Halabi *et al.*[2] were also independently picked up and retained by CAFS. This solidifies the idea that these might be key components influencing survival. Considering that five out of these eight were also involved in the first principal component, which was one of the strongest predictors, does also support this. Another set of potentially interesting predictors are those shared between three or more models.

It is debatable whether the fact that a lot of overlap exists between the various sub-models points towards the validity of selected features and the developed approach, or a potential bias in the feature selection procedure. However, the former appears more likely in the light of the approach's good performance on new data in the competition.

The included interaction terms are difficult to interpret. There is no guarantee that an interaction is modeling a direct relationship and some terms might be artifacts of higher order interactions or confounding issues. Also, when solely including terms into the model based on the optimization criterion in each step of CAFS, there is a bias to include interaction terms. Since they introduce



3A

Round	Subchallenge	Performance	Achieved Rank
1	1A	iAUC = 0.8081	1
1	1B	RMSE = 183.74	2
2	1A	iAUC = 0.8028	4
2	1B	RMSE = 196.25	12
3	1A	iAUC = 0.7922	3
3	1B	RMSE = 172.51	1

3B

Team	Subchallenge	Performance	Rank	Winners
FIMM-UTU	1A	iAUC = 0.7915	1	Rank 1 only
JayHawks	1A	iAUC = 0.7711	9	—
CAMP	1B	RMSE = 194.41	1	Rank 1 to 6
JayHawks	1B	RMSE = 195.97	3	

Figure 3. Team performance during the challenge. (**3A**) Submitted models were consistently ranked at the top of the leaderboards during the scoring rounds before the final submission. Models build via the CAFS procedure were submitted starting with the second leaderboard round. (**3B**) The final challenge submission made use of the described model ensemble approach and was placed at rank 9 in sub challenge 1A and at rank 3 in sub challenge 1B.

more parameters into the model than a main effect, they have more opportunity to improve the model within each step, even though including two different main effects in a row might be more beneficial. While our team was aware of this issue and cautious with the selection of sub-models, this still leaves potential for making suboptimal choices. This weakness could potentially be addressed in the future by switching to a parameter count based iteration, rather than a feature type based iteration.

The performed recoding of the age groups is still problematic. Intuitively, it does not make sense that the order "oldest, youngest, in-between" would be related to the outcome when disease progression usually worsens with age. A possible explanation might be that the oldest patient group is confounded with a subset of people that are resistant to the disease and have already survived for a long time. Further research is required to validate this effect.

Overall the presented method successfully built a robust predictor for the target outcome. Evidence for this is provided by the fact that the estimated performance via in-depth cross validation (iAUC = 0.757) was close to the reported performance on the larger, final leaderboard set (iAUC = 0.771) and the fact that our models were among the top performing candidates throughout the entire challenge. It should also be highlighted that the required human intervention in each selection step gives the team of researchers a lot of control, which can be very useful to introduce knowledge about the feature space into the selection process. An example of this benefit is that despite the pointed out weakness in the implementation, the team was able to account for it by rejecting inclusions of interactions that did not have a great enough impact. If desirable, early branches of the selection process can be tailored towards features with a known connection to the outcome, when multiple feature inclusions provide similar performance benefits.

Conclusion

The presented method generated a model ensemble that was able to outperform the previous best efforts to predict survival in prostate cancer patients. The developed model ensemble also successfully competed with the top performing research teams in the Prostate Cancer Dream Challenge and was among the winning teams in sub challenge 1B. We attribute this success to careful data cleaning, our efforts to derive novel features and the fact that skeptic, human decision making is integral to each iteration of the curated ad-hoc feature selection. Due to its general applicability to model building, especially in exploratory settings, the approach is promising in being useful for researchers around the world. Future studies will need to validate the presented, potentially disease associated features and potential weaknesses in the CAFS procedure should be investigated and addressed.

Data availability

The Challenge datasets can be accessed at: https://www.projectdatasphere.org/projectdatasphere/html/pcdc

Challenge documentation, including the detailed description of the Challenge design, overall results, scoring scripts, and the clinical trials data dictionary can be found at: https://www.synapse.org/ProstateCancerChallenge

The code and documentation underlying the method presented in this paper can be found at: http://dx.doi.org/10.5281/zenodo.49063[13]

Author contributions

PC thoroughly investigated the use of LASSO, elastic-net, and ridge regression as an alternative for both model selection and improving prediction. JD helped with overall code cleaning and documentation and also built predictive models for treatment discontinuation (sub challenge 2). SE acted as a liaison between the Urology Department at KUMC and the Department of Biostatistics at KUMC. She provided indispensable advice about variables to consider and those that could be combined in the creation of novel risk predictors. BF looked into non-missing at random imputation schemes for the lab values, proposed new variables/models to consider for sub challenge 1, and assisted in the development of prediction models for treatment discontinuation. SG helped with model building, code organization and presentation, wrote data cleaning function for medication history, and created new treatment history variables. DK created code to identify novel interactions between risk predictors and identified several variables to improve cross-validation performance on the training data. RM led organization of the overall code, generated code to perform intensive stepwise selection based on iAUC, and assisted in the cleaning and transforming lab data. RR helped with model building, code organization and presentation, wrote data cleaning function for medical history, and created new medical history variables. JU focused on initial model building, cleaning and deriving metastases variables, PCA for lab values, sub challenge 1b, and write-up of the synapse wiki.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the National Institute of Health (NIH) grants: 1KL2TR000119 Kansas IDeA Network of Biomedical Research Excellence (K-INBRE) Bioinformatics Core supported in part by the National Institute of General Medical Science award P20GM103418, and the Biostatistics and Informatics Shared Resource supported in part by the National Cancer Institute Cancer Center Shared Grant award P30 CA168524.

Acknowledgements

We would like to offer a special thanks to the organizing committee of the Prostate Cancer Dream Challenge and participants ASCENT-2 (Novacea, provided by Memorial Sloan Kettering Cancer Center), MAINSAIL (Celgene), VENICE (Sanofi), and ENTHUSE-33 (AstraZeneca) trials. This publication is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this **publication**.

References

1. Schallier D, Decoster L, Braeckman J, *et al.*: **Docetaxel in the treatment of metastatic castration-resistant prostate cancer (mCRPC): an observational study in a single institution.** *Anticancer Res.* 2012; **32**(2): 633–41.
PubMed Abstract

2. Halabi S, Lin CY, Kelly WK, *et al.*: **Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer.** *J Clin Oncol.* 2014; **32**(7): 671–7.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Chang K, Kong YY, Dai B, *et al.*: **Combination of circulating tumor cell enumeration and tumor marker detection in predicting prognosis and treatment effect in metastatic castration-resistant prostate cancer.** *Oncotarget.* 2015; **6**(39): 41825–41836.
PubMed Abstract | Publisher Full Text | Free Full Text

4. van Soest RJ, Templeton AJ, Vera-Badillo FE, *et al.*: **Neutrophil-to-lymphocyte ratio as a prognostic biomarker for men with metastatic castration-resistant prostate cancer receiving first-line chemotherapy: data from two randomized phase III trials.** *Ann Oncol.* 2015; **26**(4): 743–9.
PubMed Abstract | Publisher Full Text

5. Guinney J, Wang T, Laajala TD, *et al.*: **Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data.** *Lancet Oncol.* 2016.
Publisher Full Text

6. Scher HI, Jia X, Chi K, *et al.*: **Randomized, open-label phase III trial of docetaxel plus high-dose calcitriol versus docetaxel plus prednisone for patients with castration-resistant prostate cancer.** *J Clin Oncol.* 2011; **29**(16): 2191–2198.
PubMed Abstract | Publisher Full Text

7. Petrylak DP, Vogelzang NJ, Budnik N, *et al.*: **Docetaxel and prednisone with or without lenalidomide in chemotherapy-naive patients with metastatic castration-resistant prostate cancer (MAINSAIL): a randomised, double-blind, placebo-controlled phase 3 trial.** *Lancet Oncol.* 2015; **16**(4): 417–425.
PubMed Abstract | Publisher Full Text

8. Tannock IF, Fizazi K, Ivanov S, *et al.*: **Aflibercept versus placebo in combination with docetaxel and prednisone for treatment of men with metastatic castration-resistant prostate cancer (VENICE): a phase 3, double-blind randomised trial.** *Lancet Oncol.* 2013; **14**(8): 760–768.
PubMed Abstract | Publisher Full Text

9. Fizazi K, Higano CS, Nelson JB, *et al.*: **Phase III, randomized, placebo-controlled study of docetaxel in combination with zibotentan in patients with metastatic castration-resistant prostate cancer.** *J Clin Oncol.* 2013; **31**(14): 1740–1747.
PubMed Abstract | Publisher Full Text

10. Therneau TM: **A Package for Survival Analysis in S.** version 2.38, 2015.
Reference Source

11. Hung H, Chiang CT: **Estimation methods for time-dependent AUC models with survival data.** *Can J Stat.* 2010; **38**(1): 8–26.
Publisher Full Text

12. Wood SN: **Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models.** *J R Stat Soc Series B Stat Methodol.* 2011; **73**(1): 3–36.
Publisher Full Text

13. Chalise P, Dai J, Ellis S, *et al.*: **JayhawksProstateDream: First release (PCDC submission).** *Zenodo.* 2016.
Data Source

Open Peer Review

Current Referee Status: ? ✓ ?

Version 1

Referee Report 20 March 2017

doi:10.5256/f1000research.8848.r17707

?

Russell Greiner, **Luke Kumar**
Department of Computing Science, University of Alberta, Edmonton, AB, Canada

The authors describe the model they submitted to the recent “Prostate Cancer DREAM Challenge - Sub-Challenge - 1”, which was ranked among the top models in the challenge. In particular, they describe their feature selection process, which they mostly credit for their success in the challenge. The manuscript summarized their approach and the results from the challenge in an adequate manner. However, we have some concerns:

Main Critique:
Their CAFS feature selection (FS) process seems to related to the wrapper feature selection methods, but includes a human expert in the loop. This greatly reduces the reproducibility of this work. Also, the authors have not clearly listed the guideline followed by the experts when deciding on features, which appears to further reduce the usefulness of this approach in general.

Minor Points:

1. The final selected model from the FS process is not explicit from Figure 1B. It would also be better if the authors explicate CAFS’s boundary between expert intervention vs data-driven selection.

2. The basic algorithm embodied several assumptions -- eg, p6 mentions 6 features of PC1. Why 6? Also why just use sum of "protective medical history features"? Why not include harmful features? We think we understand these decisions but the paper would be improved if it better motivated these various decisions.

3. Their approach of combining the ensemble with randomly selected weights seem to introduce instability to the final prediction in different runs. It would be worthwhile to give more details on this step, describing how the proposed method compares with a simple mean (or sum) and listing the motivations for this choice -- and relate this to the claim that equal weights for all models gives the best performance

4. It was great that the authors listed the results from the winning models to give the reader a good idea about the challenge itself.

5. Figure 2A was a well thought-out table, which gives the reader insights in understanding the selected features.

6. The flow diagram in Figure 1A was difficult to process as it does not show a single flow (eg: left-right or top-down). Perhaps it would be improved with a more streamlined flow diagram of the FS process.

7. The authors mention they experimented with other survival prediction models, such as random survival forests and generalized boosted regression models. It would be useful to show results from those models, for a better comparison.

8. The authors have used links to certain web pages in the text. It would be more in line with academic publications if proper citations were used.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Competing Interests: We also competed in the same Prostate Cancer DREAM Challenge, but our participation did not influence our review. (We did not submit an F1000 submission on our work.)

Referee Report 27 February 2017

doi:10.5256/f1000research.8848.r20536

✓ **Ka Yee Yeung**, **Daniel Kristiyanto**
University of Washington, Tacoma, WA, USA

The authors documented the strategies they developed and used to predict the risk scores (Sub-challenge 1a) and survival times (Sub-challenge 1b) of prostate cancer patients as part of their participation in the DREAM 9.5 Prostate Cancer Challenge. The paper was well-written with ample of details for each step of the pipeline, with very nice figures and tables. The models performed well and ranked well in the challenge. However, there are still some areas that need further elaborations:

- The authors mentioned missing data in the "Introduction" section. However, the data imputation techniques used to replace the missing data, such as the lesion volume in the ASCENT-2 trial, are not described in "Materials and Methods". The reviewers would like to request the authors to explain what have been done to replace the missing data in all the clinical trial studies with missing data and the rationale of their strategies.

- The authors developed a "curated, ad-hoc, feature selection" (CAFS) strategy to identify predictors. The reviewers would like to request additional details on how this method selects the features, and especially on how the weights of different features are computed.

- Figure 2 showed the representation of each model and the variables involved. If the 6 considered highly correlated variables with patients survival (ALB, ALP, etc.) are already included in the first principal component (PC1) (which is used in 4/5 of the cross validation models), wouldn't it be redundant to have the same features included in the model? What were the considerations taken for this decision?

- Also related to the previous questions, model M4 was built with restrictions to not include PC1 which is claimed as the significant predictor. However, as seen in Fig 2B, this model still performed reasonably well --any observations or comments on this?

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 06 December 2016

doi:10.5256/f1000research.8848.r17706

? **Stephen R. Piccolo**

Department of Biology, Brigham Young University, Provo, UT, USA

The authors describe their participation in Subchallenge 1 of the Prostate Cancer DREAM Challenge. Their model performed well, even though it was not considered a top performer. They were creative in the way they designed their approach and tried many different options, which helped to provide insights into this particular problem as well as general strategies for model selection and optimization. Overall I was pleased with the quality of the writing and the level of detail used in the descriptions of methods and results. In particular, I like that they mentioned specific software versions and provided direct links to the code that they used for specific tasks. I did have a few questions and noticed a few gaps, which I have outlined below.

Major points:

- The manuscript provides context about the challenge as a whole. This was helpful. For example, the authors described how their approach performed in comparison to the other approaches. However, it would have been much more insightful if the authors had provided at least a brief description of the approach used by FIMM-UTU and how that approach compared to their own and what this team might have done to perform better. In hindsight, what can they learn from this?

- The manuscript describes the CAFS approach in fairly vague terms. It makes sense that the authors used intuition to optimize the feature selection. Figure 2 also provides some insight into feature and model selection. However, it is difficult to understand much about the thought process that went into these decisions. If someone else wanted to repeat this approach, how would they go about it? Are there any general guidelines that they used in making these decisions? Maybe they could provide an example that illustrates this process. Because of this, I am hesitant to accept the claim that "the approach is promising in being useful for researchers around the world."

- The manuscript mentions imputation and dealing with missing values in a couple places. But very little, if anything, is stated in the methods (or results) about how missing values were actually handled. The authors should be more explicit in describing this.

Minor points:

- In some cases, features may have been correlated strongly with each other. For example, the z-score weighted sum values and “logical or” merged variables were derived from the same underlying data. Did the authors account for these dependencies in their models in any way? If so, how?

- The authors used the class labels extensively in the training set to optimize their models. For example, their z-scores were generated based on the class labels, and they trained a large number of different models on the same data set. Thus it is impressive that their iAUC values generalized as well as they did on the validation set. However, it was unclear (or perhaps I missed it) whether the authors set aside any part of the training set as a pseudo-validation set. Figure 1C suggests that they did, but I didn't see any explicit explanation of this.

- For the " weighted sum" approach, it was a bit unclear exactly how the weights were calculated. In addition, the manuscript states that, "Optimal weights were chosen based on randomly initializing weights 100 times and estimating performance." What range of weights were used and how were they varied?

- The authors state that, "Different models for the ensemble were found either by choosing different intermediate models as the current best and branching off a certain path, or by choosing different initial models." At an abstract level, this makes sense, but it is hard to know exactly what this means. It would help to be more explicit on this part.

- In the Discussion, it says, "Another set of potentially interesting predictors are those shared between three or more models." But it is unclear what these predictors are (or perhaps I missed it). Mentioning these predictors explicitly would be helpful.

- It's a little confusing to have source code in two different locations (Zenodo and GitHub). I'd suggest just pointing people to Zenodo since the data files are there, in addition to the code. Or maybe the two are integrated? But again, if that is the case, I would suggest just using one or the other.

- I am not sure you really need to mention the top-performing team in the abstract. My recommendation would be to focus the abstract more on your solution rather than on the challenge results.

- The authors use URLs as citations in the Introduction (e.g., http://www.cancerresearchuk.org/ and http://www.seer.cancer.gov/statfacts/html/prost.html and https://www.synapse.org/ProstateCancerChallenge). It seems that some of these should instead be references to peer-reviewed publications.

- The authors state that, "suggesting strong performance of the Halabi et al. model for identifying low- and high-risk mCRPC patients." What does this mean, more specifically, from a clinical standpoint?

- The author contributions section states, "assisted in the development of prediction models for treatment discontinuation." This doesn't seem relevant to this paper.

- I tried to install the R package dependencies that are described in the README file. However, it gave me an error message saying that some of the packages could not be found. To solve this, I had to specify a repository in the code ("repos" parameter of install.packages). It would be helpful if the authors changed this part of the code so that it will run out of the box.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: I was also a competitor in the Prostate Cancer DREAM Challenge. However, I did not submit an F1000 report describing my own algorithmic approach and feel that my participation did not cause a bias in my review.