RESEARCH ARTICLE

Predicting discontinuation of docetaxel treatment for metastatic castration-resistant prostate cancer (mCRPC) with random forest [version 1; referees: 1 approved, 1 approved with reservations]

Daniel Kristiyanto, Kevin E. Anderson, Ling-Hong Hung, Ka Yee Yeung

Institute of Technology, University of Washington, Tacoma, USA

v1 **First published:** 16 Nov 2016, **5**:2673 (doi: 10.12688/f1000research.8353.1)
Latest published: 16 Nov 2016, **5**:2673 (doi: 10.12688/f1000research.8353.1)

Abstract

Prostate cancer is the most common cancer among men in developed countries. Androgen deprivation therapy (ADT) is the standard treatment for prostate cancer. However, approximately one third of all patients with metastatic disease treated with ADT develop resistance to ADT. This condition is called metastatic castrate-resistant prostate cancer (mCRPC). Patients who do not respond to hormone therapy are often treated with a chemotherapy drug called docetaxel.
Sub-challenge 2 of the Prostate Cancer DREAM Challenge aims to improve the prediction of whether a patient with mCRPC would discontinue docetaxel treatment due to adverse effects.
Specifically, a dataset containing three distinct clinical studies of patients with mCRPC treated with docetaxel was provided. We applied the k-nearest neighbor method for missing data imputation, the hill climbing algorithm and random forest importance for feature selection, and the random forest algorithm for classification. We also empirically studied the performance of many classification algorithms, including support vector machines and neural networks. Additionally, we found using random forest importance for feature selection provided slightly better results than the more computationally expensive method of hill climbing.

Keywords
Predictive Model , Multivariate Feature Selection , Hill Climbing , Random Forest

This article is included in the DREAM Challenges gateway.

Open Peer Review

Referee Status: ✓ ?

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

1 **Vishakh Hegde**, Stanford University, USA
Karen Sachs, Stanford University, USA

2 **John E. Mittler**, University of Washington, USA

Discuss this article

Comments (0)

Corresponding author: Ka Yee Yeung (kayee@uw.edu)

Competing interests: No competing interests were disclosed.

Grant information: Ling Hong Hung and Ka Yee Yeung are supported by NIH grant U54-HL127624. Daniel Kristiyanto is sponsored by the U.S. Department of State and American-Indonesian Exchange Foundation (AMINEF) through Fulbright Scholarship, and gratefully acknowledges funding provided by the University of Washington in the form of full tuition waivers.

How to cite this article: Kristiyanto D, Anderson KE, Hung LH and Yeung KY. **Predicting discontinuation of docetaxel treatment for metastatic castration-resistant prostate cancer (mCRPC) with random forest [version 1; referees: 1 approved, 1 approved with reservations]** *F1000Research* 2016, **5**:2673 (doi: 10.12688/f1000research.8353.1)

First published: 16 Nov 2016, **5**:2673 (doi: 10.12688/f1000research.8353.1)

Motivation & background

Prostate cancer is the most common cancer affecting men. It is also one of the main causes of cancer mortality[1]. In addition to radiotherapy, androgen deprivation therapy (ADT) is a standard treatment. However, approximately one third of all patients with metastatic disease treated with ADT develop resistance to ADT. This condition is called metastatic castrate-resistant prostate cancer (mCRPC)[1,2]. Patients who do not respond to hormone therapy are often treated with a chemotherapy drug called docetaxel. The Prostate Cancer DREAM Challenge is a crowd-sourcing effort that aims to improve the prediction of survival and toxicity of docetaxel treatment in patients with mCRPC[2]. Specifically, there are two sub-challenges: 1) to predict overall survival of mCRPC patients using clinical trial data, and 2) to predict discontinuation of the docetaxel treatment due to adverse event at early time points. This paper reports our team's effort contributing to sub-challenge 2.

The data for the challenge were provided by Project Data Sphere[3], consisting of four clinical trials (ASCENT-2[4], VENICE[5], MAINSAIL[6], and ENTHUSE-33[7]) for patients with mCRPC treated by docetaxel. The training data made available to the challenge participants consisted of 1600 patients from three clinical trials (ASCENT-2, VENICE, MAINSAIL). The clinical data from ENTHUSE-33 serve as the scoring set to test the prediction accuracy of submissions and hence, were not available to participants. Each team was allowed a maximum of two online submissions. In our two submissions, we used the same methods but varied the hold-out data[8,9]. After the challenge window, the submission portal was re-opened, allowing participants to continue their effort in refining and exploring alternative methods.

As a part of the Prostate Cancer DREAM Challenge, we developed data-driven models to predict patient outcomes in mCRPC with subsequent discontinuation of docetaxel therapy. We contributed to sub-challenge 2, which aims to predict discontinuation of docetaxel treatment due to adverse events. We empirically studied and assessed the performance of various machine learning algorithms and feature selection methods using cross validation on the provided training data. We assessed our predictive models using the area under the curve (AUC)[10], which is the scoring metric adopted by the Prostate Cancer DREAM Challenge sub-challenge 2. This paper reports the predictive models we developed for the Prostate Cancer DREAM Challenge as well as further improvements we made after the challenge was closed. The methods and our challenge submission are available online from Synapse[8,9], and our post-challenge efforts are available on the GitHub repository[11].

Objective

The training data consist of clinical variables across 1,600 mCRPC patients in three clinical trials, namely ASCENT-2 (Novacea, provided by Memorial Sloan Kettering Cancer Center, with 476 patients)[4], VENICE (Sanofi, with 598 patients)[5], and MAINSAIL (Celgene, with 526 patients)[6]. Specifically, longitudinal data from five tables were summarized into a core table consisting of 131 variables. The five tables of raw longitudinal data at patient-level include PriorMed (prior medication table records), Med-History (medical history table records patient reported diagnoses at time of patient screening), LesionMeasure (lesion table records target and non-target lesion measurement), LabValue (lab test table includes all lab data), and VitalSign (vital sign table records patient vital sign such as height and weight)[2]. We used the training data in the core table to build models predictive of treatment discontinuation (binary) for patients in a fourth clinical trial (test data), ENTHUSE-33 (AstraZeneca, with 470 patients)[7].

Data & methods

Our approach consists of four main steps: (1) data cleansing and pre-processing, (2) feature selection, (3) classification, and (4) assessment, as shown in Figure 1.

Data cleansing & pre-processing

Our analysis focused on the core data only. From the 1,600 patients, we removed 111 patients without clear discontinuation status, leaving a total of 1,489 patients across three clinical trials: ASCENT-2 with 476 patients, MAINSAIL with 420 patients, and VENICE with 593 patients. Data cleansing was performed separately within each clinical trial and later concatenated back together. Some features were only available on certain clinical trials, for instance, smoking frequency, which was only available on ASCENT-2. In contrast, features such as sodium, phosphorus, and

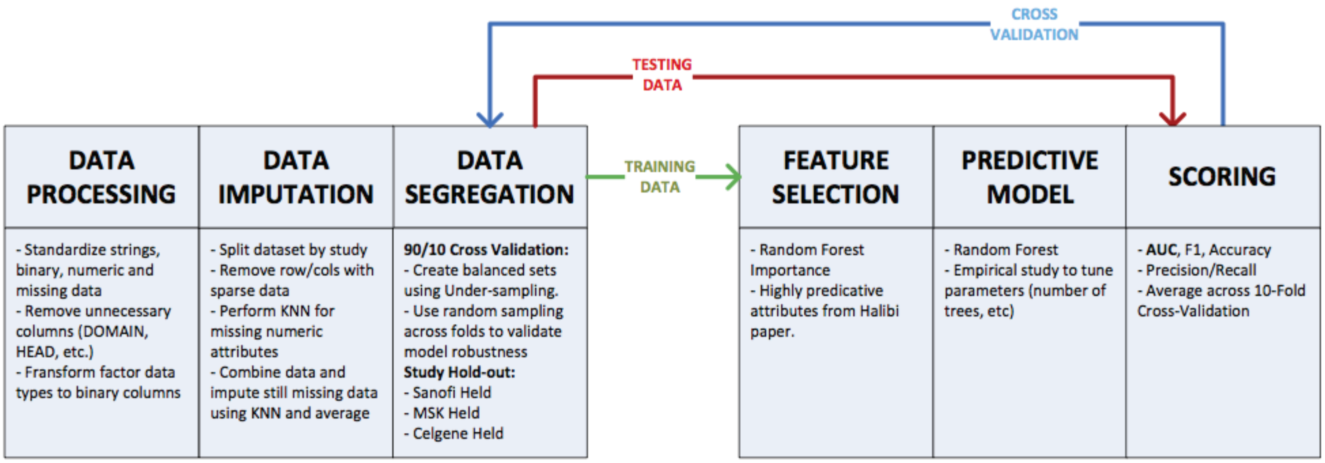
CROSS VALIDATION
TESTING DATA
TRAINING DATA
DATA PROCESSING
DATA IMPUTATION
DATA SEGREGATION
FEATURE SELECTION
PREDICTIVE MODEL
SCORING
- Standardize strings, binary, numeric and missing data
- Remove unnecessary columns (DOMAIN, HEAD, etc.)
- Fransform factor data types to binary columns
- Split dataset by study
- Remove row/cols with sparse data
- Perform KNN for missing numeric attributes
- Combine data and impute still missing data using KNN and average
90/10 Cross Validation:
- Create balanced sets using Under-sampling.
- Use random sampling across folds to validate model robustness
Study Hold-out:
- Sanofi Held
- MSK Held
- Celgene Held
- Random Forest Importance
- Highly predicative attributes from Halibi paper.
- Random Forest
- Empirical study to tune parameters (number of trees, etc)
- AUC, F1, Accuracy
- Precision/Recall
- Average across 10-Fold Cross-Validation

Figure 1. A diagram describing key properties of each major step in our pipeline.

albumin were available in two clinical trials other than ASCENT-2. The heterogeneity in the data also manifested in the different interpretation of the values. To name a few, features such as lesion locations HEAD_AND_NECK or STOMACH, only contained positive responses. In this case, we assigned all of the missing values as unknown, instead of negative values.

We kept variables that are available from all three clinical trials in the training data. We also performed data imputation using the `impute` package version 1.42.0 from Bioconductor[12]. Data imputation was applied to the missing data in the baseline lab values (such as alkaline phosphatase, etc.) using the *k* nearest neighbor (KNN) algorithm. We varied the value of *k* from 1 to the number of patients, and evaluated the performance against a naive-bayes classifier[13] in a 10 cross-fold validation. $k = 40$ was shown to be an optimal parameter. Missing information such as the patients' weight, height and BMI were replaced by the average value of all patients. For the discrete variables, we replaced the missing values with a new variable ('unknown' value). Data augmentation was also performed by converting selected multi-label variables into binary variables, such as 'smoking frequency'. Figure 2 shows how the split and reconstruction was performed.

To avoid over-fitting, we performed data cleansing and data pre-processing for the testing data (ENTHUSE-33) separately.

Feature selection

We observed that univariate feature selection methods did not perform well in this case. We used cross validation to guide us in the search for relevant features (or clinical variables) in the data. Specifically, we assessed our models using the area under the precision-recall curve (AUC) using the `ROCR` R package version 1.0-7[10].

We adopted the multivariate hill-climbing[14] approach that optimized the AUC using 10-fold cross validation of the training data to search for relevant features among 131 features (clinical variables). The algorithm started with a random set of features against the model and returned the AUC. Depending on the AUC, the algorithm removed, kept, or added other features, and iterated until it converged. This method was used in both of our submissions to sub-challenge 2. Hill-climbing aims to maximize accuracy and is a greedy approach. However, hill-climbing also has its limitations: it is computationally intensive and prone to be stuck at a local optimum. In addition, it also tends to converge with different sets of features within each cross validation run, which makes it difficult to determine what are the corresponding factors that contribute to the discontinuation of the treatment. As one of the goals of the challenge is to identify prognostic markers in patients with mCRPC who will discontinue the treatment, hill-climbing may not the ideal approach.

Halabi *et al.* reported a list of strong predictors for overall survival of mCRPC patients[15,16]. These predictors include race, age, BMI, prior radio therapy, prior analgesics, and patient performance status. Lab results for albumin, lactate dehydrogenase, white blood cells, aspartate amminotransferase, total bilirubin, platelet count, hemoglobin, alanine transaminase, prostate specific antigen and alkaline phosphatase were also reported to be strong predictors of overall survival for patients with mCRPC. We hypothesize that the underlying molecular mechanisms that drive overall survival (the goal of sub-challenge 1) and treatment discontinuation (the goal of sub-challenge 2) are related. In addition, after the challenge was closed and the winners were announced, we checked out the winners' winning strategy. In particular, we were inspired by Team Jayhawks from University of Kansas Medical Center who stated that they "made use of the variables we derived for sub-challenge 1a and also the overall risk score for survival". Therefore, we experimented with the set of features for overall survival reported by Halabi *et al.* and used the results as the baseline. Additionally, we also performed random forest importance from `FSelector` package[17] to look for additional features to improve the prediction accuracy for treatment discontinuation (sub-challenge 2).

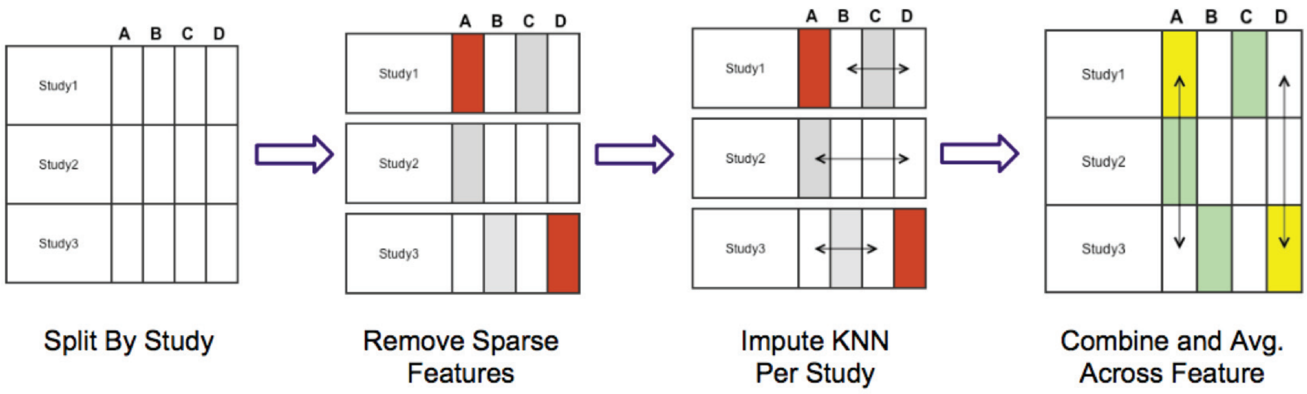

Figure 2. In step 1 (Split By Study) we split the training data into three sections, each representing data produced by a different clinical trial. In step 2 (Remove Sparse Features) considering each dataset in turn, we removed features that were missing ample data (>79%) to prevent KNN imputation (illustrated as red blocks). In step 3, also considering each clinical trial as distinct, we used KNN to impute missing data from the features that contained missing data (illustrated as grey blocks). Finally, in step 4 all datasets were concatenated. and the features removed in step 2 (illustrated now as yellow blocks) were replaced by the mean values of that feature calculated from the other two clinical trials (shown using vertical arrows).

Random forest measures the importance of a variable by estimating the prediction error when that variable is permuted and the rest of variables remain unchanged. We evaluated the importance of the remaining variables using random forest. By varying the number of the features and evaluating the results, medical history (neoplasms benign, malignant or unspecified), smoke, and glucose were identified as the contributing factors. In modeling this feature selection, we addressed the issue of imbalanced data by performing random sampling using the training data consisting of 0.31 positive samples (patients known to have discontinued the treatments) and 0.69 negative samples.

Classification

We applied various classification algorithms to the selected features, including support vector machine (SVM)[18], decision trees[19], neural networks[20], random forest[21,22], and ensemble methods[23]. We observed comparable performance across different classification methods, and subsequently selected random forest as the classifier in our final submissions due to its robustness in heterogeneous datasets. As an ensemble method, random forest splits the training data into a number of subset and constructed decision trees as the classification model. Although the `random forest` package from CRAN[21] usually comes with a set of sensible default parameters, we performed cross validation to optimize the classifier models. Subsequently, we adopted the following set of parameters: pick 9 random variables on each split (parameter *mtry*), and set the number of trees to 6,300 trees (parameter *ntree*). During all of the model tuning, 123 was used as the random seed.

Results

For sub-challenge 2, each team was allowed 2 submissions. In our two submissions, we used the same methods and varied the hold-out data[8,9]. See Table 1 for detailed results.

We applied hill-climbing to select features from the model space and random forest as the classification methods by varying the hold-out data (see Table 1). We performed 10-fold cross validation by randomly selecting 10% of the training data across all three clinical trials as the hold-out data, and repeated the procedures 10 times. This 10-fold cross validation procedure using hill-climbing yielded an average AUC of 0.532. However, we achieved a substantial reduction in AUC (0.129) when we applied this model to the scoring data (470 patients from ENTHUSE-33 clinical trials). This model served as our 1st submission and ranked 35th on the leaderboard.

Table 1. AUC results from models using each clinical trial as hold-out, or using 90/10 cross validation. Each column represents a different feature selection technique and/or hold-out dataset. Columns labeled ‘ENTHUSE-33’ were the test set provided by the Challenge organizers, and hence, were scored via the DREAM 9.5 submission system, while others were scored using the hold-out study as the testing dataset. BL = baseline (features selected based on the Halabi paper only[16], HC = hill-climbing, RFI = combining features identified by Halabi and random forest importance.

	Scored using hold-out			ENTHUSE-33 (AZ)	
Hold-out data	***BL***	***HC_1***	***RFI_1***	***HC_2***	***RFI_2***
All combined (90/10)	0.272	0.532	0.275	0.129	**0.146**
VENICE (Sanofi)	0.085	0.171	0.106	0.132	0.140
ASCENT-2 (MSK)	0.321	0.376	0.303	0.138	0.131
MAINSAIL (Celgene)	0.227	0.292	0.263	0.135	0.124

We went on to conduct additional empirical experiments to identify the difference between AUC from 10-fold cross validation and AUC from the scoring set. We hypothesized that this difference in AUCs from cross validation and from the scoring data is due to heterogeneity in the data collected in different clinical trials. Therefore, we studied the heterogeneity of three clinical trials in the training data by using each of three clinical trials as hold-out data. Table 1 showed that using the VENICE clinical trials as the hold-out data resulted in AUCs that are comparable to what we observed in our 1st submission. In particular, we produced an AUC of 0.171 by holding-out the VENICE clinical trial in the training data. Our 2nd submission resulted from applying hill-climbing and random forest to the ASCENT-2 and MAINSAIL clinical trials, and achieved an AUC of 0.132 on the scoring data (ENTHUSE-33). Our 2nd submission ranked 34th out of 61 submissions on the leaderboard. The AUC measured by the top performer was 0.190[24].

Figure 4 shows that albumin was consistently selected by hill-climbing as a strong predictor disregard to the hold-out data. In total, there are 27 clinical variables selected in more than one hill-climbing model, including: Na (sodium), OTHER (other lesion), ALB (albumin), ORCHIDECTOMY, CEREBACC cerebrovascular accident either hemorrhagic and or ischemic), AST (aspartate aminotransferase), HB (hemoglobin), Mg (magnesium), LYMPH_NODES (lesion), PROSTATE (lesion), BILATERAL_ORCHIDECTOMY, CORTICOSTEROID (medication), CREAT (creatinine), PSA (prostate specific antigen), CREACL (creatinine clearance).

After the challenge ended, we continued to fine tune our models and submitted predictions to be scored. We adopted the Halabi model[16] and combined it with random forest importance (RFI) to improve prediction. Random forest was kept as the classification method, and we also varied the hold-out data as the assessment. Random forest importance was computed by excluding features from the Halabi model which resulted in different sets of features depending on the hold-out data (see Figure 5).

Next, we compared the features chosen by hill-climbing, RFI and the Halabi model. We observed that SMOKE and REGION overlapped among the hill-climbing and RFI results as shown in Figure 6. By varying the number of top features ranked by andom forest importance combined with the Halabi model, 4 top ranking clinical variables resulted from random forest importance yielded the best average accuracy (see Figure 3). Applying this predictive model in 10-fold cross validation resulted in an average AUC of 0.275. We also repeated the process by selecting each individual clinical trial as the hold-out data, which yielded AUCs of 0.106 (VENICE as the

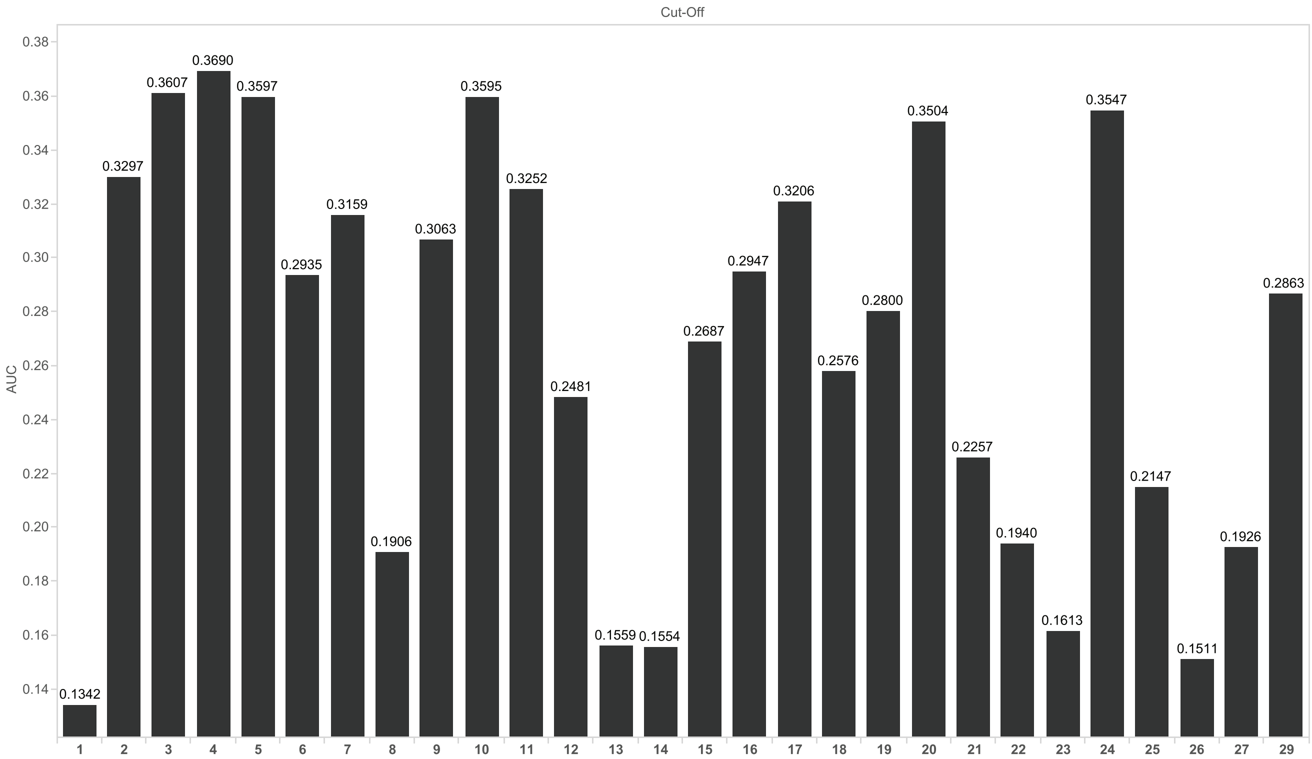

Figure 3. Resulting AUC by varying the number of selected features from random forest importance combined with features described by the Halabi model. Random forest was used as the classification method, and 10% of randomly selected data across all sample as the holdout data.

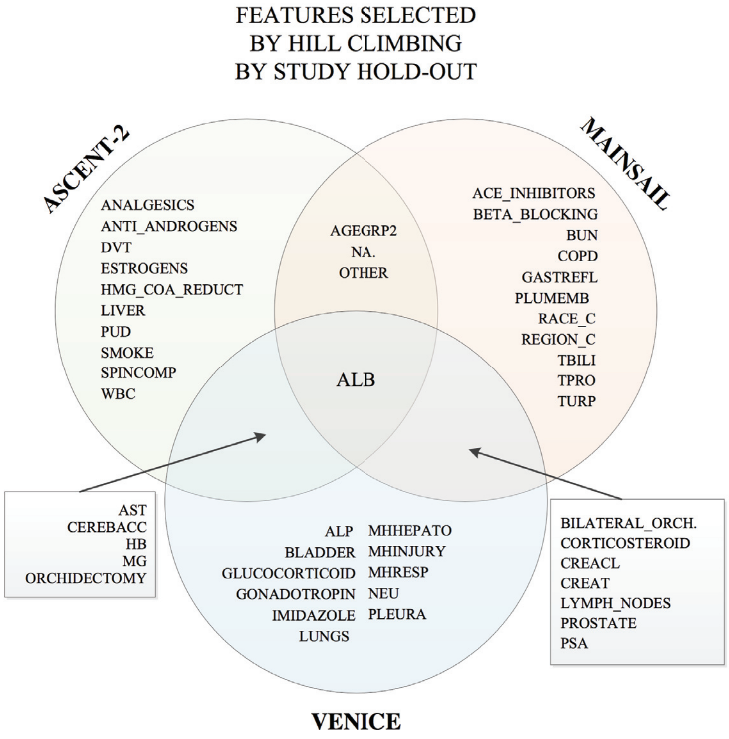

Figure 4. Venn Diagram comparing features identified using the hill-climbing method, using each study as the hold-out data during training. Random forest was used as the classification model.

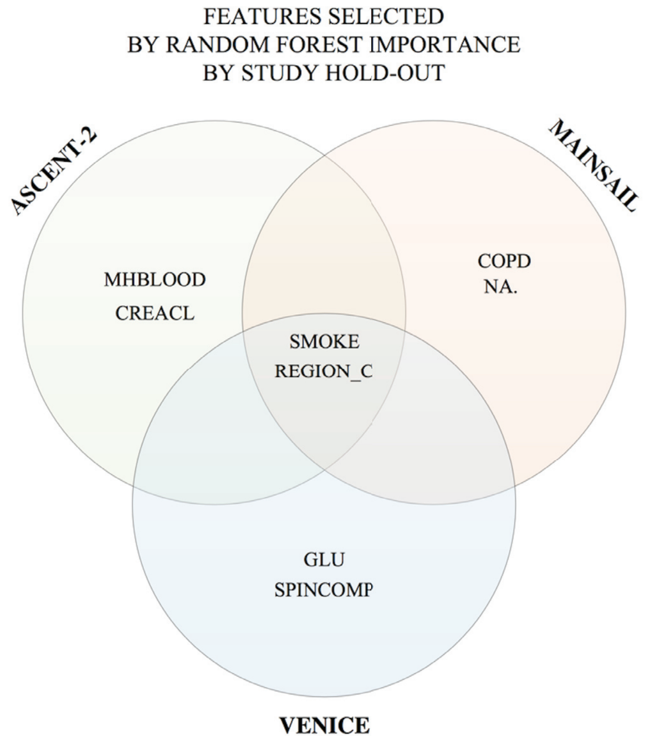

Figure 5. Venn diagram comparing features identified by random forest importance (RFI), using each study as the hold-out data. Features identified by the Halabi model was excluded during the computation.

hold-out), 0.303 (ACCENT-2), and 0.263 (MAINSAIL). Compared to the hill-climbing method, this model produced better and more consistent AUCs across clinical trials. Using each of the trained models to predict the discontinuation of docetaxel treatments for 470 patients from ENTHUSE-33 (AstraZeneca) clinical trials resulted in AUCs of 0.140 (VENICE), 0.132 (ACCENT-2), 0.124 (MAINSAIL), as also shown in Table 1.

Discussion

A major challenge of the Prostate Cancer DREAM Challenge was the unbalanced class sizes and the heterogeneity of the clinical trials. Subsequent to data cleansing, there remained only 197 positive samples (patients who discontinued the treatments) and 1,292 of negative samples. We observed that data cleaning and augmentation improved the AUC and F1 (before we augmented the data, our AUCs were in the range of 0.2 in 10-fold cross validation). Lastly, we were delighted to find that our computationally intensive hill-climbing algorithm, designed to find an optional feature set, provided strong results when testing using the datasets provided to us. When scored using the ENTHUSE-33 dataset, hill-climbing performed weakly against the combination of Halabi model and random forest importance.

We experimented with various feature selection and classification methods. We observed that some clinical variables were selected more consistently across feature selection methods, see Figure 6. These variables include: AGEGRP2 (Age Group), ALB (Albumin), ALP (Alkaline Phosphatase), ANALGESICS, AST (Aspartate Aminotransferase), BLADDER (Lesion), COPD (Blood and Lymphatic System), CREACL (Creatinine Clearance), HB (Hemoglobin), LIVER (Leison), LUNGS (Lesion), LYMPH_NODES (Lesion), NA. (Sodium), OTHER (Other Lesion), PLEURA (Pleura), PROSTATE (Lesion), PSA (Prostate Specific Antigen), RACE_C, REGION_C, SMOKE, SPINCOMP (Spinal Cord Compression), TBILI (Total Bilirubin), and WBC (White Blood Cells).

In this study, we only looked at the core table that was precompiled from detailed longitudinal tables. Given more time and resources, a close look at the raw longitudinal data may yield additional insight in discovering the clinical variables that predict the discontinuation of treatment for mCRPC patients.

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/10.7303/syn4601848[8] and http://dx.doi.org/10.7303/syn4729761[9]. The method and results are also presented as a poster[25].

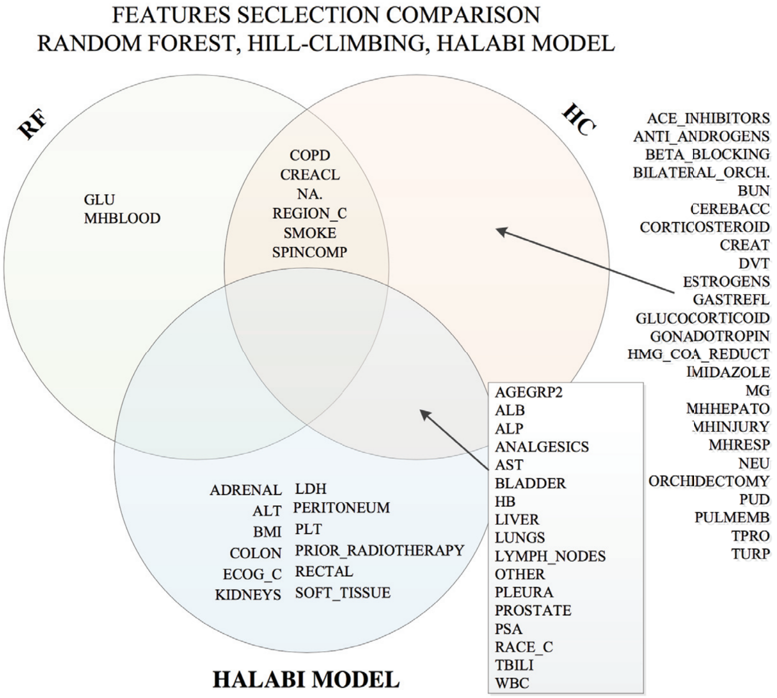

Figure 6. Venn diagram comparing the union of features identified via random forest importance (RFI), hill-climbing (HC), and the features identified by Halabi in 15. Random forest was used as the classification method. RFI: the union of all features selected by random forest importance by varying the hold-out data after excluding features described by the Halabi model. HC: is the union of all features selected by hill-climbing by varying the hold-out data.

Author contributions

DK served as team captain of team ‘yoda’ in the DREAM9.5 Prostate Challenge, wrote the first draft of the manuscript and was responsible for code consolidation and submission. KA contributed data cleansing scripts, created figures and assisted in the writing of this paper. KYY supervised the project through its duration. All authors contributed to the writing of the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

Ling Hong Hung and Ka Yee Yeung are supported by NIH grant U54-HL127624. Daniel Kristiyanto is sponsored by the U.S. Department of State and American-Indonesian Exchange Foundation (AMINEF) through Fulbright Scholarship, and gratefully acknowledges funding provided by the University of Washington in the form of full tuition waivers.

Acknowledgements

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. This project used computing resources provided by Microsoft Azure.

We also would like to thank all students in TCSS 588 Bioinformatics in the Spring Quarter of 2015 at University of Washington Tacoma who contributed to this project, and extend our acknowledgments to Sage Bionetworks, the DREAM organization, and Project Data Sphere for developing and supplying data for the challenge.

References

1. Gupta E, Guthrie T, Tan W: **Changing paradigms in management of metastatic Castration Resistant Prostate Cancer (mCRPC).** *BMC Urol.* 2014; **14**(1): 55.
PubMed Abstract | Publisher Full Text | Free Full Text
2. **Dream 9.5 Prostate Cancer dream challenge**. 2015.
Publisher Full Text
3. **Home** | **share, integrate & analyze cancer research data** | **project data sphere**. 2016.
Reference Source
4. 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
5. 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
6. 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
7. 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
8. Anderson K, Sina Khankhajeh S, Kristiyanto D, *et al.*: **Prostate cancer 9.5 submission 1 - syn4601848.** 2015. Accessed on 02/22/2016.
Publisher Full Text
9. Anderson K, Sina Khankhajeh S, Kristiyanto D, *et al.*: **Prostate cancer 9.5 submission 2 - syn4729761.** 2015. Accessed on 02/22/2016.
Publisher Full Text
10. Sing T, Sander O, Beerenwinkel N, *et al.*: **ROCR: visualizing classifier performance in R.** *Bioinformatics.* 2005; **21**(20): 3940–3941.
PubMed Abstract | Publisher Full Text
11. Kristiyanto D, Andersonn K: **Predicting discontinuation of docetaxel treatment for metastatic castration-resistant prostate cancer (mCRPC).** *GitHub repository.* 2016.
Reference Source
12. Hastie T, Tibshirani R, Narasimhan B, *et al.*: **impute: Imputation for microarray data.** R package version 1.44.0. 2016.
Reference Source
13. Dimitriadou E, Hornik K, Leisch F, *et al.*: **Misc functions of the department of statistics (e1071), tu wien.** R package. 2008; **1**: 5–24.
14. Romanski P: **Fselector: Selecting attributes.** Vienna: R Foundation for Statistical Computing. 2009.
15. 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): 671–677.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Halabi S, Small EJ, Kantoff PW, *et al.*: **Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer.** *J Clin Oncol.* 2003; **21**(7): 1232–1237.
PubMed Abstract | Publisher Full Text
17. Romanski P, Kotthoff L: **FSelector: Selecting attributes.** R package version 0.20. 2014.
Reference Source
18. Hearst MA, Dumais ST, Osman E, *et al.*: **Support vector machines.** *Intelligent Systems and their Applications, IEEE.* 1998; **13**(4): 18–28.
Publisher Full Text
19. Ripley B: **tree: Classification and Regression Trees.** R package version 1.0-37. 2016.
Reference Source
20. Fritsch S, Guenther F: **Training of neural networks.** R package version 1.32. 2012.
Reference Source
21. Liaw A, Wiener M: **Classification and regression by randomforest.** *R News.* 2002; **2**(3): 18–22.
Reference Source
22. Breiman L: **Random forests.** *Mach Learn.* 2001; **45**(1): 5–32.
Publisher Full Text
23. Dietterich TG: **Ensemble methods in machine learning.** In *Multiple classifier systems.* Springer. 2000; **1857**: 1–15.
Publisher Full Text
24. **Dream9.5 - prostate cancer dream challenge final scoring round - syn2813558**. 2016.
Reference Source
25. Kristiyanto D, Anderson K, Sina Khankhajeh S, *et al.*: **Predicting discontinuation of docetaxel treatment for metastatic castration-resistant prostate cancer (mcrpc) with hill-climbing and random forest.** *F1000Research.* 2015; 4:1383 (poster).
Publisher Full Text

Open Peer Review

Current Referee Status: ✓ ?

Version 1

Referee Report 26 January 2017

doi:10.5256/f1000research.8982.r19698

?

John E. Mittler
Department of Microbiology, University of Washington, Seattle, WA, USA

The authors summarize the work they have done to improve computational methods for predicting which prostate cancer patients will discontinue the cancer drug docetaxel. The methods were tested using clinical data given in one of the DREAM competitions, a series of competitions in which computational biologists are challenged to submit predictions for quantitative biological and biomedical questions for which answers have been withheld from the competitors during the competition phase.

I am marking this as “Approved with Reservations” because of oversimplification in the abstract and page 7 concerning the performance of Hill Climbing (HC) and Random Forest Importance (RFI). (The journal specifically asks us to comment on whether the “abstract represents a suitable summary of the work”). Although RFI arguably did better than HC on the ENTHUSE-33 dataset, HC did substantially better in the hold-out experiments in Table 1. On page 7 they write: “Compared to the hill-climbing method, [RFI] produced better and more consistent AUCs across clinical trials.” Having higher AUC in 2/4 ENTHUSE-33 tests doesn’t strike me as consistently better.

To get me to “Approve” this, I would like full attention to my main point above (this shouldn’t be too hard) and some subset of the comments below (some of which are optional).

1. With respect to my main comment above, it would be nice if the authors could give some measure (e.g., 95% CIs) of the variation in their AUC estimates. Is 0.146 (AUC for RFI on the ENTHUSE-33 dataset) significantly higher than 0.129 (AUC for HC)?

2. A bit more discussion about the lessons that you learned from this exercise would be helpful. Based on your research, what advice, if any, would you give to someone entering a similar competition next year? If you feel uncomfortable providing more advice, please explain what it is that makes you uncomfortable.

3. With regard to comment #2, one of the key points I derived from this paper was the importance of properly cleaning and augmenting the data. However, the quantitative data that support this conclusion was tucked into a parenthetical remark in the discussion. These data belong in the results.

4. Also, in the discussion, please provide whatever thoughts you may have as to why HC didn’t do as well on the ENTHUSE-33 dataset as it did in the other experiments.

5. The data in Figure 3 are highly erratic. As far as I can tell, 3, 5, 10, 20, or 24 features would be indistinguishable from 4. Please comment.

6. I concur with one of previous reviewers about the value of adding the AUC score for BL on the ENTHUSE-33 dataset. Also, consider adding a row (first three entries of which would be blank) giving AUCs for ENTHUSE-33 for models trained on the full dataset.

7. I wonder if the statement in the abstract “We also empirically studied the performance of many classification algorithms, including support vector machines and neural networks” could be converted into some kind of result (which would need to supported in the results section with AUC values). Maybe cut back on the background to make room from this.

Minor: “random” came out as “andom” on page 5. “cutoff” should be under the x-axis in Figure 3. Bold entry in Table 1 isn’t explained.

Competing Interests: No competing interests were disclosed.

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.

Referee Report 20 January 2017

doi:10.5256/f1000research.8982.r19549

✓ **Vishakh Hegde** [1], **Karen Sachs** [2]

[1] Stanford University, Palo Alto, CA, USA

[2] Baxter Laboratory for Stem Cell Biology, Department of Microbiology & Immunology, Stanford University, Palo Alto, CA, USA

Vishakh Hegde:
The title, abstract and introduction reflects the core content of the article. The authors clearly specify the problem and provide a sound technical overview of their approach and solution, with diagrams clearly illustrating it. However, we would like to see the following:

1. While they authors provide the AUC for baseline (BL) scored on hold-out data, we would like to see the same for the test data as well (ENTHUSE-33). This will provide a metric to compare their algorithms (HC and RFI) with respect to BL

2. In the ‘Classification’ subsection, we would like to see how the AUC compare across various classification algorithms they claim to have tried.

Karen Sachs:
The authors present an exploration of a feature selection and classification problem in prostate cancer from multiple clinical trials. Overall an interesting exploration. I did find a few points confusing:

1. The hill climbing features selection was described as nonoptimal because it does not concur from iteration to iteration – a fair point, also described in Figure 4. I did not understand why it was

nonetheless employed in the results? It was a bit confusing which of the fs procedures described were used for which result.

2. Also, it was not clear to me if the hill climbing feature selection was done with the entire dataset? If so it will overfit and reduce test performance. In fact I wonder if this is the reason that performance degraded for Enthuse-33. Can the authors comment on this/clarify this point? Was the entire pipeline—feature selection through classifier – performed on a subset of the data, such that the hold out (test) data had not been used in any part of the process before it was used to assess AUC?

3. Minor point – Text in the paper “Figure 4 shows that…disregard to..” should instead be “irrespective of”.

Competing Interests: No competing interests were disclosed.

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.