METHOD ARTICLE

REVISED A predictive model of overall survival in patients with metastatic castration-resistant prostate cancer [version 2; peer review: 2 approved]

Mehrad Mahmoudian[1,2*], Fatemeh Seyednasrollah[1,2*], Liisa Koivu[3], Outi Hirvonen[3,4], Sirkku Jyrkkiö[4], Laura L. Elo[1]

[1]Turku Centre for Biotechnology, Turku, Finland
[2]Department of Information Technology, University of Turku, Turku, Finland
[3]Department of Oncology and Radiotherapy, University of Turku, Turku, Finland
[4]Department of Clinical Oncology, University of Turku, Turku, Finland

[*] Equal contributors

v2 **First published:** 16 Nov 2016, **5**:2674 (https://doi.org/10.12688/f1000research.8192.1)
Latest published: 17 May 2019, **5**:2674 (https://doi.org/10.12688/f1000research.8192.2)

Abstract
Metastatic castration resistant prostate cancer (mCRPC) is one of the most common cancers with a poor prognosis. To improve prognostic models of mCRPC, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) Consortium organized a crowdsourced competition known as the Prostate Cancer DREAM Challenge. In the competition, data from four phase III clinical trials were utilized. A total of 1600 patients' clinical information across three of the trials was used to generate prognostic models, whereas one of the datasets (313 patients) was held out for blinded validation. The previously introduced prognostic model of overall survival of chemotherapy-naive mCRPC patients treated with docetaxel or prednisone (so called Halabi model) was used as a performance baseline. This paper presents the model developed by the team TYTDreamChallenge and its improved version to predict the prognosis of mCRPC patients within the first 30 months after starting the treatment based on available clinical features of each patient. In particular, by replacing our original larger set of eleven features with a smaller more carefully selected set of only five features the prediction performance on the independent validation cohort increased up to 5.4 percent. While the original TYTDreamChallenge model (iAUC=0.748) performed similarly as the performance-baseline model developed by Halabi et al. (iAUC=0.743), the improved post-challenge model (iAUC=0.779) showed markedly improved performance by using only PSA, ALP, AST, HB, and LESIONS as features. This highlights the importance of the selection of the clinical features when developing the predictive models.

Keywords
prostate cancer, mCRPC, boosting, survival analysis

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
REVISED **version 2** published 17 May 2019		✓ report
version 1 published 16 Nov 2016	✓ report	? report

1 **Motoki Shiga**, Gifu University, Gifu, Japan

2 **Peter K Rogan**, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Canada

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the DREAM Challenges gateway.

Corresponding author: Mehrad Mahmoudian (mehmah@utu.fi)

Author roles: Mahmoudian M: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Seyednasrollah F**: Data Curation, Formal Analysis, Investigation, Methodology, Software; **Koivu L**: Data Curation; **Hirvonen O**: Data Curation; **Jyrkkiö S**: Data Curation; **Elo LL**: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Sigrid Juselius Foundation (to L.L.E) and the University of Turku Graduate School (to F.S).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Mahmoudian M, Seyednasrollah F, Koivu L *et al.* **A predictive model of overall survival in patients with metastatic castration-resistant prostate cancer [version 2; peer review: 2 approved]** F1000Research 2019, **5**:2674 (https://doi.org/10.12688/f1000research.8192.2)

First published: 16 Nov 2016, **5**:2674 (https://doi.org/10.12688/f1000research.8192.1)

REVISED **Amendments from Version 1**

In this version, we have addressed the concerns of the second Reviewer. The modifications that have been made to the article have not changed the methodology, the results, or the discussion, but focused on improving the clarity of the text. Based on the Reviewer's comments, we have:

- Used more accurate terminology for the baseline model that has been used in the Prostate Cancer DREAM Challenge
- Clarified the feature exclusion criteria
- Completely revised the caption of Figure 3C to explain in more detail how to interpret the presented partial dependence plots
- Increased the font size in Figure 3 to improve readability
- Revised the sentence regarding how the final selected features are in line with the prognosis of patients
- Provided correlation plot of a larger set of features in Supplementary Figure 1B.

See referee reports

Introduction

Prostate cancer is the second most common cancer according to the World Cancer Report 2014[1]. Hence it is one of the most studied cancer types with focus on diagnosis and prognosis. A major cause of death among prostate cancer patients is the development of metastatic castrate-resistant prostate cancer (mCRPC), which is both a persistent as well as progressing disease resistant to androgen deprivation therapy[2].

In order to boost research regarding prostate cancer, a crowdsourced competition was designed by the Dialogue for Reverse Engineering Assessments and Methods (DREAM) Consortium in collaboration with *Project Data Sphere LLC* (PDS) to improve prognostic models of mCRPC. Using data from four phase III clinical trials available through PDS, two main sub-challenges were designed. Sub-challenge 1 was aimed at improving prediction of survival risk for mCRPC patients, whereas Sub-challenge 2 was intended to predict adverse events in patients treated with docetaxel, the standard of care for mCRPC patients at the time of the trials. This paper presents the model developed by the team TYTDreamChallenge for the Sub-challenge 1 to predict survival risk scores for mCRPC at 12, 18, 24 and 30 months after diagnosis based on clinical features of each patient, as well as a post-challenge analysis to improve our initial model.

Various prognostic models for mCRPC have been previously developed[3–5]. Recently, Halabi *et al*. developed a prognostic model for mCRPC using eight clinical features (Eastern Cooperative Oncology Group performance status, disease site, lactate dehydrogenase, opioid analgesic use, albumin, hemoglobin, prostate-specific antigen, and alkaline phosphatase) and validated it on an external dataset. We refer to this model as Halabi model. It was used in the DREAM Challenge as a performance-baseline model.

In the TYTDreamChallenge model and in our post-challenge model, we used generalized boosted models implemented in the R[6] package gbm (generalized boosted regression models) to predict overall survival of mCRPC patients with Cox proportional hazard model as the underlying regression model[7]. The gbm package is an extension to the Freund and Schapire's AdaBoost algorithm[8] and Friedman's gradient boosting machine[9]. In general, boosting is a concept in supervised machine learning with the goal of generating multiple relatively weak learner models, which each individually work slightly better than random guess, and use them all in corporation to have a highly accurate overall model[10].

Methods

Our methodology consisted of two major steps[11]. The first step was data preparation. The second step was model building utilizing generalized boosted models. All the validations of the model predictions were performed through the submission system of the DREAM Challenge, where the true response values in the validation data remained hidden.

Data

The data used in this study was collected from mCRPC patients by four institutes. The datasets were based on a cancer treatment trial in which patients received docetaxel treatment. Details of the four trials are shown in Table 1. In the Prostate

Table 1. The four clinical trial datasets used for the mCRPC predictions.

Data Provider	ID	Number of patients	Reference
Novacea, provided by Memorial Sloan Kettering Cancer Center	ASCENT-2	476	Scher *et al.*[13]
Celgene	MAINSAIL	526	Petrylak *et al.*[14]
Sanofi	VENICE	598	Tannock *et al.*[15]
AstraZeneca	ENTHUSE-33	470	Fizazi *et al.*[16]

Cancer DREAM Challenge, three (ASCENT-2[12], MAINSAIL[13] and VENICE[14]) out of the four datasets were available as training sets. The remaining dataset (ENTHUSE-33[15]) was used for validation by the DREAM Challenge organizers without releasing the survival data to the participants of the competition. All the data were gathered into five major tables (Supplementary Table 1). Additionally, a sixth table, called CoreTable, was provided by the challenge organizers. The CoreTable was a collection of features from the other five tables that summarized the baseline (day 0) values. The clinical features in CoreTable contained treatment variables, cancer staging based on AJCC[16], Gleason Score[17], ECOG Performance Status[18], and lesion details. This table was curated by challenge organizers and was considered as the main table in the Challenge.

Out of all the data provided, we focused on the CoreTable and LabValue tables to form the training and validation datasets. The LabValue table was an event level longitudinal data table which contained all the lab tests performed along with the sampling date and reference range of each lab test. The CoreTable consisted of 131 features, of which two were for identification, five were dependent variables and 124 were independent variables. The two dependent variables we used in this study were DEATH and LKADT_P. The variable DEATH indicates the death status of a patient and has value "YES" for patients who died from mCRPC and value "NO" otherwise. The variable LKADT_P is the last day that the patient was known to be alive.

The full set of the Challenge data is available under the standard Synapse Terms and Conditions of Use and the Prostate Cancer DREAM Challenge Rules and can be downloaded from Synapse web interface. The links and authentication information are available in the following URL: https://www.synapse.org/ProstateCancerChallenge

Data preparation

Processing of the laboratory values (table LabValue) consisted of a sequence of actions. First, it was observed that there were some duplicate rows in the data; hence 2545 rows were removed. Secondly, based on consultation with oncologists, rows with measurements of 13 lab tests were extracted, including ALT, AST, ALP, LDH, MG, PHOS, ALB, TPRO, PSA, HB, WBC, NEU and LYM (Table 2). After this step, the number of rows left in the data was 80744. Thirdly, we removed 603 rows marked with "NOT DONE" status in the LBSTAT column, which specifies completion status of the lab test, and with missing value in their LBSTRESC column, which contains standardized format of the test results. Finally, only the 17015 baseline measurements from the 1599 patients were kept in the study, while removing the other follow-up measurements over time as they were unavailable in the validation data. During the steps explained above, one patient (ASC-518-0003) was completely removed from the analysis because of having "NOT DONE" status in most of the lab tests, including ALT, AST, ALP and LDH.

The measurement values for all lab tests, except PSA, were standardized based on their minimum and maximum ranges as

$$x' = 2 \cdot \frac{x - \alpha}{\beta - \alpha} - 1 \quad (1)$$

where x is the observed value of the lab test and β and α are the corresponding upper and lower limit of the reference range. The

Table 2. Definitions of the 13 lab tests selected on the basis of consultation with oncologists.

Lab Test Abbreviation	Definition
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
ALP	Alkaline phosphatase
LDH	Lactate dehydrogenase
MG	Magnesium
PHOS	Phosphorus
ALB	Albumin
TPRO	Total protein
PSA	Prostate specific antigen
HB	Hemoglobin
WBC	White blood cells
NEU	Neutrophils
LYM	Lymphocytes

standardized values are between -1 and 1 if the lab test value is within the normal range. The PSA values were only $\log_2$ transformed and the issue of $\log_2 0$ was bypassed by adding e^{-4} to the values before $\log_2$ transformation. The ALP and NEU values were truncated to 10 and 5, respectively.

In the validation dataset, there were two patients (AZ-00131 and AZ-00383) that had no records in the LabValue table nor in the CoreTable. To predict their survival using the laboratory values, we extracted medians of the 13 lab test features across all patients and used them for these two patients.

In addition to lab tests, we considered some additional features from the CoreTable. These included ECOG_C and ANALGESICS as well as four derived features that were summarized to reduce the variation and existing noise in the data. These included LESIONS, DRUGS, DISEASES and PROCEDURES, which were defined as arithmetic sums of the numbers of lesions, medicines, diseases or medical operations, respectively. LKADT_P and DEATH were also directly adopted from the CoreTable.

As the final step in pre-processing, the resulting training and validation datasets were checked for features having large proportions of missing values or having missing values for a particular data provider. The missingness in data is shown in Supplementary Figure 1A. Based on this, six features including MG, ALB, TPRO, LYM, PHOS, and LDH were excluded from the training and validation sets with maximum missingness of ~100% in at least one of the datasets. Additionally, to minimize the number of highly correlated features in the training data, we further removed WBC and ALT, which showed high correlation with NEU and AST, respectively (Supplementary Figure 1B).

At the end of the pre-processing, the training set consisted of 1599 patients and validation set of 313 patients. Both datasets had 15 features out of which two were for identification, two were response variables and the other 11 were independent predictor variables (ECOG_C, ANALGESICS, LESIONS, DRUGS, DISEASES, PROCEDURES, AST, ALP, PSA, HB, and NEU).

Machine learning and survival prediction

To develop a model of overall survival in mCRPC, we utilized a gradient boosting algorithm based on regression trees, with a Cox proportional hazard model as the underlying regression model. The R package gbm[19] was used with 5000 trees, 10-fold cross-validation, minimum 3 observations in the trees' terminal nodes, and step-size reduction value of 0.007.

In the DREAM Challenge competition, we considered a separate risk score for 12, 18, 24 and 30 months. For 18, 24 and 30 months, we built a separate model for each data provider, and the mean of the three individual risk score predictions was calculated as the final risk score at each time point. For 12 months, all the training data were used to create a single model and risk score prediction. After the challenge, we also tested these two strategies in constructing a single overall risk score for each patient: 1) average of risk scores obtained separately for each data provider (referred to as PostSeparate), or 2) a single risk score obtained by combining data from all the providers in the modelling (referred to as PostCombined).

Performance evaluation

The performance of the predictions was measured using the integrated area under the ROC curve (iAUC) from 6 to 30 months, as well as separate AUC values at 12, 18, and 24 months. The iAUC was calculated using the R package timeROC (version 0.3)[20]. The performance measures were obtained from blinded validation by the DREAM Challenge organizers.

Results and discussion

The performance of the TYTDreamChallenge model (iAUC= 0.748) was significantly better than random. However, it did not perform statistically significantly better than the performance baseline Halabi model (iAUC=0.743, Bayes factor < 3), as determined by the DREAM Challenge organizers[21].

To further investigate the possibility to improve our model after the challenge, we considered in our post-challenge analysis the impact of calculating an overall risk score instead of our original strategy of having separate scores for the different time points. Interestingly, this had a marked effect on the performance of our model (Figure 1). When the average model across the different data providers was considered, the iAUC improved to 0.757 (model PostSeparate). When all the data were used together for model building, the iAUC increased further to 0.777 (model PostCombined).

Next, we examined the relative importance of the different features on the predictions in the PostCombined model, as determined by the boosting algorithm (Figure 2A). As expected, many of the features used in the Halabi model had high importance also in our model (PSA, ALP, and HB). However, additional features were found (AST, NEU). On the other hand, ECOG_C was not as important in our model as it was in the Halabi model. We also tested the effect of removing one variable at a time when building the model (Figure 2B). This supported further the importance of ALP, HB, AST, PSA and LESIONS, whereas the removal of NEU actually improved the performance further (iAUC=0.780). Removal of PROCEDURES, ANALGESICS, ECOG_C, DISEASES or DRUGS did not have a marked impact on the performance.

Finally, we applied the same boosting strategy to build a model using only five features ALP, HB, LESIONS, AST and PSA (Figure 3A; referred to as PostFive). Notably, the performance in the validation data did not decrease markedly from that with a larger set of features (iAUC=0.779). Among the features, PSA and ALP had the largest relative importance in predicting the survival, whereas LESIONS had the lowest relative importance (Figure 3B). To assist in understanding the contribution of the

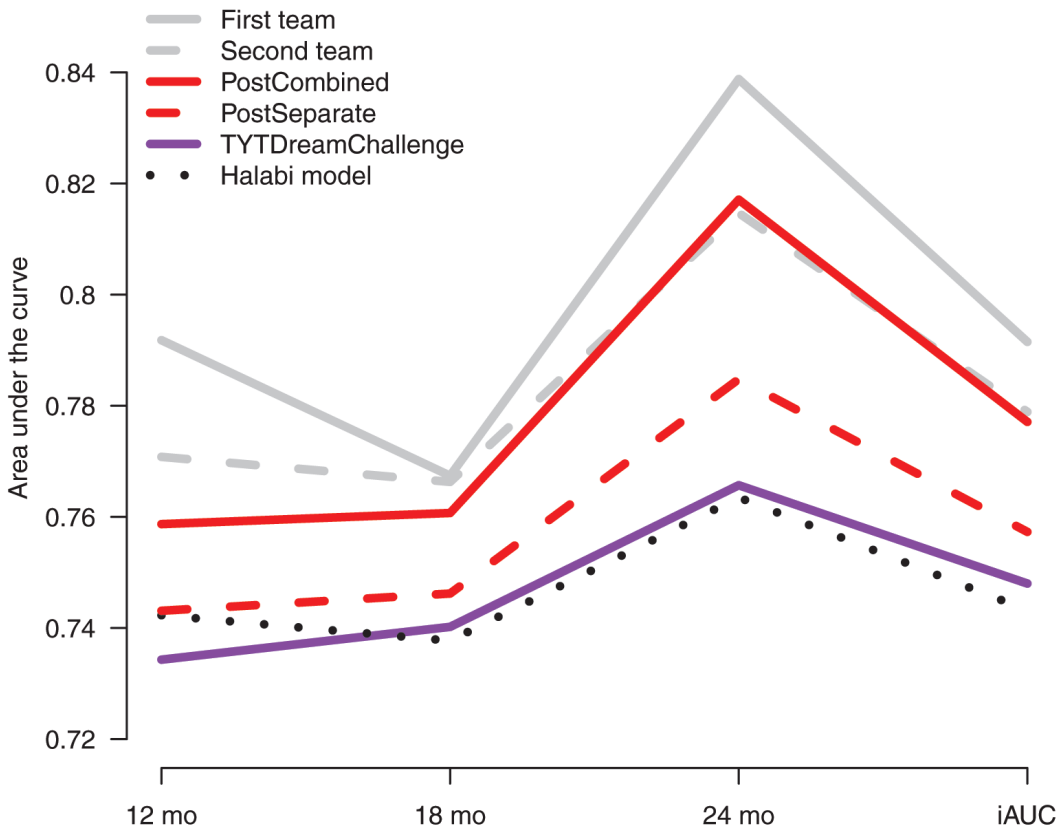


Figure 1. Performance of the different models compared to the performance baseline Halabi model. The integrated area under the ROC curve (iAUC) from 6 to 30 months, as well as separate AUC values at 12, 18, and 24 months are shown. The performance measures were obtained from blinded validation by the DREAM organizers. The first and second team refer to the top-ranked models in the DREAM Challenge; PostCombined and PostSeparate refer to two different post-challenge analyses using the same modelling strategy and same features as in our original DREAM Challenge submission (TYTDreamChallenge) but, instead of having time-specific models, a single overall risk score was calculated for each patient either as an average risk score across the data providers (PostSeparate) or as a single risk score obtained by combining data from all the providers in the modelling (PostCombined).

identified features, partial dependence plots were examined, which illustrate the partial dependence of the risk scores on each feature after accounting for the effects of the other features (Figure 3C). Similarly as in the Halabi model, the risk increases with high values of PSA and ALP, high numbers of LESIONS, and low values of HB[22]. Additionally, our model suggests that high values of AST increase the risk. These findings are well in line with the general hypothesis that these factors are basic values that associate with patients' prognosis.

Taken together, based on the blinded validations it can be concluded that the proposed post-challenge model in this paper (PostCombined) was markedly better than the Halabi model, which is considered as the state-of-the-art standard method in the field. The post-challenge analysis suggested that a single overall risk score performed better than our original strategy of time-specific risk scores by better targeting the overall survival pattern of patients. A model based on only five features ALP, HB, AST, PSA and LESIONS produced a relatively high accuracy compared to the Halabi model with eight features or the model of the winning team involving a large number of features and their interactions. Thus the five-feature model (PostFive) presented here provides an efficient option in terms of practical clinical use.

The present study focused on clinical features only. Additional possibilities to improve the performance of the models would be to add molecular level information, such as gene expression data to training and test sets.

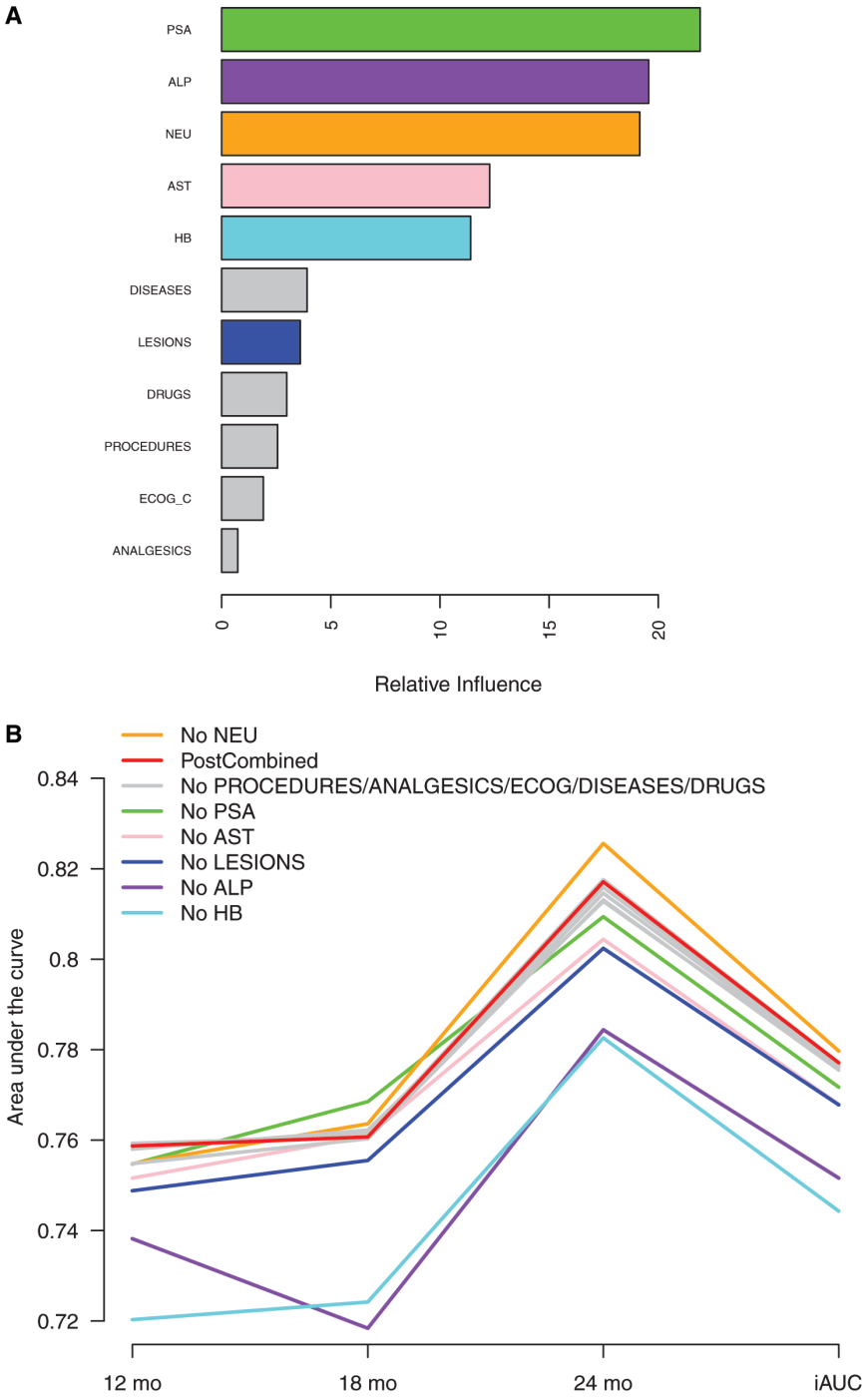


Figure 2. Relative importance of the different features on the predictions. (**A**) Relative influence in the post-challenge model where a single overall risk score was calculated for each patient by combining data from all the providers in the modelling (PostCombined). (**B**) Effect of removing one feature at a time when building the model. The integrated area under the ROC curve (iAUC) from 6 to 30 months, as well as separate AUC values at 12, 18, and 24 months are shown. The performance measures were obtained from blinded validation by the DREAM organizers.

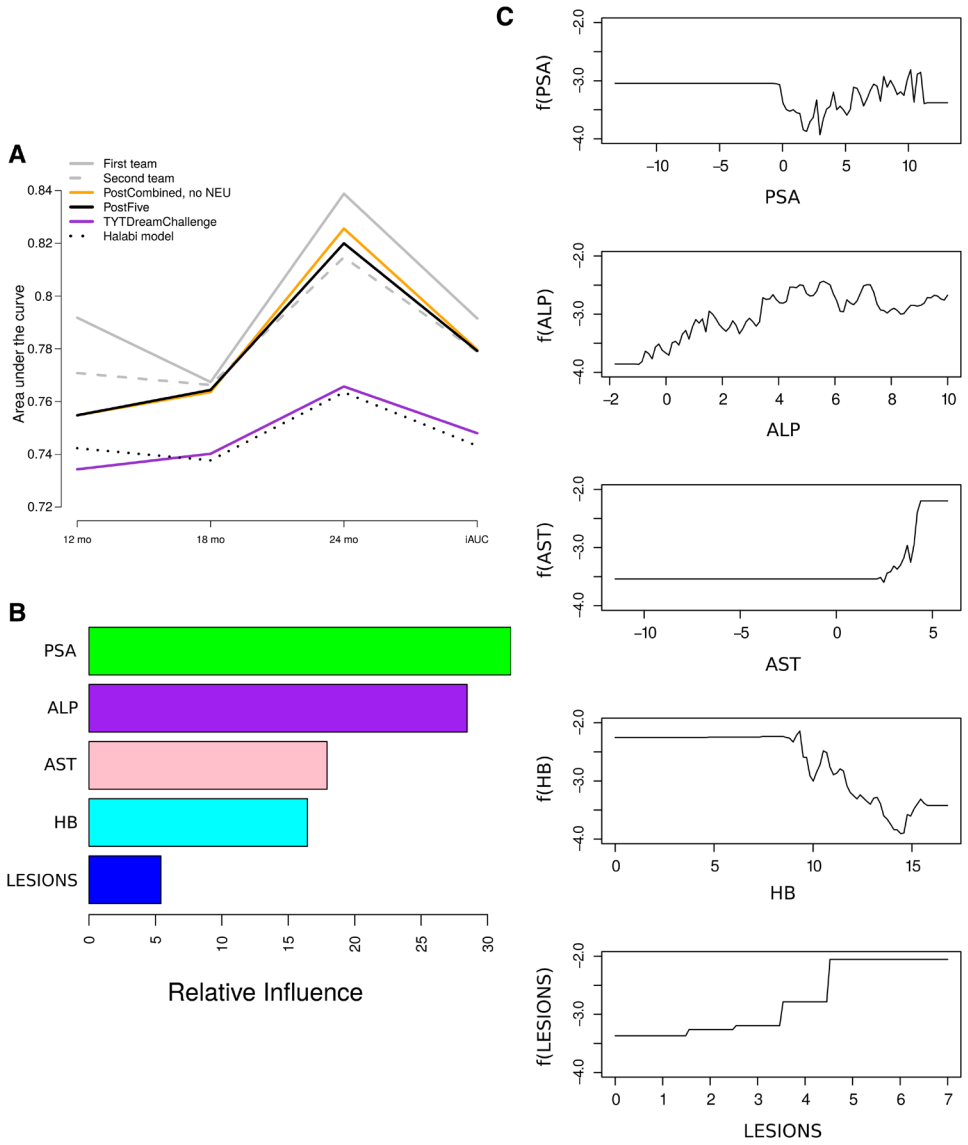


Figure 3. PostFive model. (**A**) Performance of the boosting strategy using only five features ALP, HB, LESIONS, AST and PSA, as compared to the DREAM Challenge models and our post-challenge models. The integrated area under the ROC curve (iAUC) from 6 to 30 months, as well as separate AUC values at 12, 18, and 24 months are shown. The performance measures were obtained from blinded validation by the DREAM organizers. (**B**) Relative importance of the different features on the predictions. (**C**) Partial dependence plots illustrating the marginal effect of changing the value of each feature (x-axis) on the value of the hazard function (y-axis), while averaging out the other variables. In general, higher hazard value can be interpreted as lower survival probability.

Data and software 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.47706[23]

The latest source code is available at: https://bitbucket.org/mehrad_mahmoudian/dream-prostate-cancer-challenge-q.1a

Author contributions

MM participated in the pre-processing of the data, performed all the post-challenge analyses and drafted the manuscript. FS pre-processed the data and developed the TYTDreamChallenge model. LK, OH and SJ participated in the pre-processing and provided the clinical insights. LLE designed and supervised the study, participated in the analyses and drafted the manuscript.

Grant information

This work was supported by the Sigrid Juselius Foundation (to L.L.E) and the University of Turku Graduate School (to F.S).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

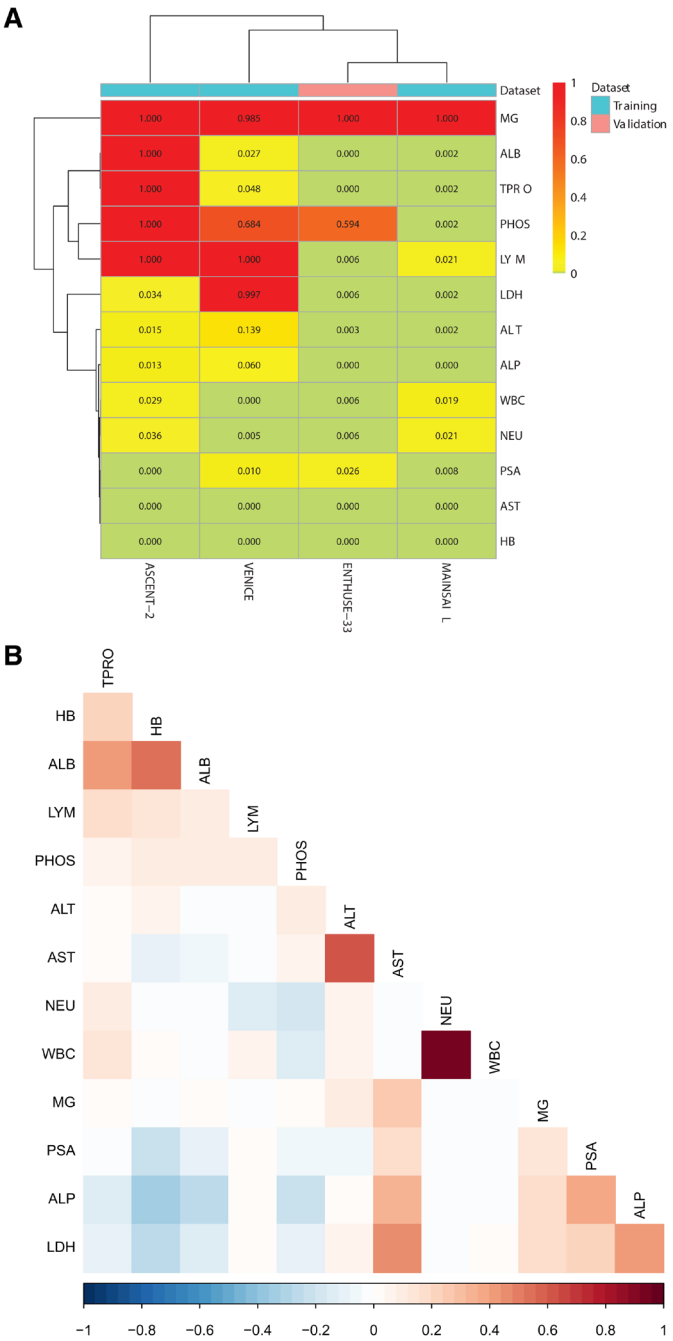
Acknowledgments

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.

We would also thank the Sage Bionetworks, the DREAM organization, and Project Data Sphere for developing and supplying data for the Challenge.

We would also like to thank Riku Klén and Mikko Venäläinen for their thoughtful input and technical insight during the manuscript writing phase.

Supplementary material



Supplementary Figure 1. Missing values and correlations in the training data. (**A**) Proportions of missing values for the 13 lab tests selected on the basis of consultation with oncologists. The proportion of missing values is shown separately for each clinical trial dataset (columns), with red indicating large proportions of missing values. (**B**) Pearson correlation between the different features across all studies. The darker the color, the higher the absolute value of correlation.

Supplementary Table 1. Raw data available in the Challenge.

Table name	Table description
PriorMed	Prior Medication table records the medication that the patients have taken before the first treatment date of the trial.
MedHistory	Medical History table records the diagnoses reported by the patients (co-existing diseases) at the time of patient screening to participate in the trial.
LesionMeasure	Lesion table records target and non-target lesion measurements.
LabValue	Lab test table includes all data from lab tests (hematology and urinary lab).
VitalSign	Vital Sign table records vital signs of the patients (e.g. height, weight, etc.).

References

1. Stewart B, Wild CP (eds.): **World Cancer Report 2014**. I. A. for R. on C. W. 2014.
Reference Source
2. Hotte SJ, Saad F: **Current management of castrate-resistant prostate cancer.** *Curr Oncol.* 2010; **17**(Suppl 2): S72–9.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Lucarelli G, Ditonno P, Bettocchi C, *et al.*: **Serum sarcosine is a risk factor for progression and survival in patients with metastatic castration-resistant prostate cancer.** *Future Oncol.* 2013; **9**(6): 899–907.
PubMed Abstract | Publisher Full Text
4. Pond GR, Armstrong AJ, Galsky MD, *et al.*: **Efficacy of docetaxel-based chemotherapy following ketoconazole in metastatic castration-resistant prostate cancer: implications for prior therapy in clinical trials.** *Urol Oncol.* 2013; **31**(8): 1457–63.
PubMed Abstract | Publisher Full Text
5. Qu YY, Dai B, Kong YY, *et al.*: **Prognostic factors in Chinese patients with metastatic castration-resistant prostate cancer treated with docetaxel-based chemotherapy.** *Asian J Androl.* 2013; **15**(1): 110–5.
PubMed Abstract | Publisher Full Text | Free Full Text
6. 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
7. R Core Team: **R: A Language and Environment for Statistical Computing**. 2015.
Reference Source
8. Cox DR: **Regression Models and Life-Tables.** *J R Stat Soc Ser B.* 1972; **34**(2): 187–220.
Reference Source
9. Freund Y, Schapire R, Abe N: **A short introduction to boosting.** *Journal-Japanese Soc Artif Intell.* 1999; **14**(5): 1612.
Reference Source
10. Friedman JH: **Greedy Function Approximation: A Gradient Boosting Machine.** *Ann Stat.* 2001; **29**(5): 1189–1232.
Publisher Full Text
11. Schapire RE: **The strength of weak learnability.** *Mach Learn.* 1990; **5**(2): 197–227.
Publisher Full Text
12. **TYT Prostate Cancer DREAM Challenge writeup - syn4228911**.
Reference Source
13. 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
14. 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
15. 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): 760–8.
PubMed Abstract | Publisher Full Text
16. 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
17. Schmoll HJ, Greene FL, Page DL, *et al.* (eds): **AJCC Cancer Staging Manual, 6th edition.** *Ann Oncol.* 2003; **14**(2): 345–346.
Publisher Full Text
18. Gleason DF, Mellinger GT: **Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging.** *J Urol.* 1974; **111**(1): 58–64.
PubMed Abstract | Publisher Full Text
19. Oken MM, Creech RH, Tormey DC, *et al.*: **Toxicity and response criteria of the Eastern Cooperative Oncology Group.** *Am J Clin Oncol.* 1982; **5**(6): 649–55.
PubMed Abstract | Publisher Full Text
20. Ridgeway G: **gbm: Generalized Boosted Regression Models**. 2015.
Reference Source
21. Blanche P, Dartigues JF, Jacqmin-Gadda H: **Estimating and Comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks.** *Stat Med.* 2013; **32**(30): 5381–5397.
PubMed Abstract | Publisher Full Text
22. **DREAM9.5 - Prostate Cancer DREAM Challenge - syn2813558**. 2015.
Reference Source
23. Mahmoudian M, Seyednasrollah F, Koivu L, *et al.*: **Source code for "A predictive model of overall survival in patients with metastatic castration-resistant prostate cancer"**. 2016.
http://www.doi.org/10.5281/zenodo.47706

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 04 June 2019

https://doi.org/10.5256/f1000research.20848.r48714

✔ **Peter K Rogan**
Dept of Biochemistry, Schulich School of Medicine and Dentistry, University of Western Ontario, London, ON, N6A 5C1, Canada

I approve the revised article based on the revisions made by the authors.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine learning, mathematical modeling, gene signature analysis, cancer research

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.

Version 1

Reviewer Report 26 June 2017

https://doi.org/10.5256/f1000research.8812.r23094

? **Peter K Rogan**
Dept of Biochemistry, Schulich School of Medicine and Dentistry, University of Western Ontario, London, ON, N6A 5C1, Canada

In general, the paper is understandable and well organized. The use of boosting to improve the Cox proportional hazard and regression models is interesting. Despite the issues the authors address regarding missing data, there appears to be adequate information in this dataset to test this approach.

Abstract

“as well as an improvement of the model developed after the challenge”
This is vague statement that adds little to the abstract. It is irrelevant when the improvement was made. The abstract should describe the clinical features and any improvements in an organized manner so that the reader can determine whether the method proposed differs from previous approaches.

“TYTDreamChallenge model performed similarly as the gold-standard Halabi model”
What was the performance accuracy or misclassification rate?

The “Halabi” model terminology is jargonistic and should be described briefly in conventional oncology criteria (ie. an abbreviated form of the statement in the introduction would be sufficient). Was the increase in the ROC by 3% significant? If so, please explain why.

Introduction

The authors term the paper by Halabi *et al.* to be the “gold standard” prognostic model. While it performs well and is a reasonable comparator, it does not meet the criteria for a gold standard. It is a conventionally trained and tested and validated with a single external dataset. Gold standards, on the other hand, have been reproduced by other investigators using other patient cohorts multiple times with similar findings, may have fulfilled ISO or other standards and have been recommended by internationally recognized authorities (for example, the dicentric chromosome assay for radiation dosimetry).

Methods

“some features were removed from the study due to missing values, high correlations with other features or being unimportant for such survival analysis as determined by clinical experts. “
Please indicate which features were removed. What proportions were attributable to missing values, etc.

“based on consultation with oncologists, rows with measurements of 13 important lab tests”
Define important vs unimportant lab tests, and reasons for selecting them.

Did the authors evaluate whether their methods were very sensitive to assumptions made about the missingness mechanism or about the distributions of the variables with missing data? If so, please state.

Results

The authors don’t clearly distinguish which methods were used in their submission of results to the DREAM challenge vs. how or why the “improvements” were made after the submission to the challenge. Furthermore, the overall risk score vs separate scores at different time points can simply be reported, without making this artificial distinction. Thus, the distinction made in the abstract between these lacks context, and I would recommend removing it.

There are many other measures for evaluating the models that the authors could report besides AUC, including Matthews Correlation Coefficient, F-measure, Precision, and Accuracy. They may consider

computing and reporting these.

Figure 3 is unacceptable quality. Even at 200% magnification, the labels on each of the graphs are barely readable. Panel C requires some further explanation in order to interpret it. The text indicates “This suggested intuitive interpretations for the different features,” which does not explain the results or whether the partial dependence can be used for feature selection or interpretation.

“These findings are well in line with the general hypothesis that these factors are basic values representing the volume of the disease.”
By volume, are the authors referring to the extent of the disease? The extent of the disease is not the same as the survival risk, which is what the authors state they are modeling in the introduction. In cancer, volume refers to the size of the tumor and quantitative distribution (Castro-Mesta *et al.* 2016[1]). In fact, the authors use the numbers of lesions to infer survival risk, so it would appear that this is a circular argument.

References

1. Castro-Mesta J, González-Guerrero J, Barrios-Sánchez P, Villarreal-Cavazos G: Bases and foundations of the treatment of peritoneal carcinomatosis: Review article. *Medicina Universitaria*. 2016; **18** (71): 98-104 Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine learning, mathematical modeling, gene signature analysis, cancer research

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.

Author Response 23 Apr 2019

Mehrad Mahmoudian, Turku Centre for Biotechnology, Turku, Finland

We would like to thank the reviewer for the time and effort he has put into our manuscript and the insightful comments. We have addressed all the comments in this response as well as in the main text and figures accordingly.

Abstract

> "as well as an improvement of the model developed after the challenge"
> This is vague statement that adds little to the abstract. It is irrelevant when the improvement was made. The abstract should describe the clinical features and any improvements in an organized manner so that the reader can determine whether the method proposed differs from previous approaches.

Thank you for your comment. We understand your concern that the text was not clear enough to convey the concept we wanted to the full extent. Therefore, we have modified the text to improve clarity and transparency and included information about the clinical features in the abstract as suggested by the reviewer. Regarding the time of analysis, we feel that it should be mentioned to comply with the DREAM Challenge procedure and evolution of developing the ultimate model.

> "TYTDreamChallenge model performed similarly as the gold-standard Halabi model"
> What was the performance accuracy or misclassification rate?

All the validations of the predictions were done through submitting the predicted values to the DREAM Challenge evaluation system as have been explained in detail in https://www.synapse.org/#!Synapse:syn2813558/wiki/209591 . As the true response values of the validation dataset were kept hidden from the participants, unfortunately, the requested information cannot be provided. We have modified the text to clarify this fact accordingly.

> The "Halabi" model terminology is jargonistic and should be described briefly in conventional oncology criteria (ie. an abbreviated form of the statement in the introduction would be sufficient). Was the increase in the ROC by 3% significant? If so, please explain why.

We have clarified the term "Halabi model" in the abstract via:
"The previously introduced prognostic model of overall survival of chemotherapy-naive mCRPC patients treated with docetaxel or prednisone (so called Halabi model) was used as a performance baseline."
Regarding the comment about significance, unfortunately as explained above, we do not have access to the true response values and therefore the significance between the ROC curves cannot be computed.

Introduction

> The authors term the paper by Halabi et al. to be the "gold standard" prognostic model. While it performs well and is a reasonable comparator, it does not meet the criteria for a gold standard. It is a conventionally trained and tested and validated with a single external dataset. Gold standards,

on the other hand, have been reproduced by other investigators using other patient cohorts multiple times with similar findings, may have fulfilled ISO or other standards and have been recommended by internationally recognized authorities (for example, the dicentric chromosome assay for radiation dosimetry).

We agree that the term “gold-standard” was not a good choice. Therefore, we have changed all instances to “performance baseline”.

Methods

> “some features were removed from the study due to missing values, high correlations with other features or being unimportant for such survival analysis as determined by clinical experts. “
> Please indicate which features were removed. What proportions were attributable to missing values, etc.

We agree that the description of the feature selection process was not very clear. We have now modified the text. The revised Supplementary Figure 1 now shows the proportions of missing values and the correlations between the 13 lab tests considered. We also provide the source codes of this study publicly available, containing the in-depth details of the feature exclusion along with the exact procedure to reproduce the results. The DOI of the published code is doi:10.5281/zenodo.47706 which is also presented as the last reference of the manuscript.

> “based on consultation with oncologists, rows with measurements of 13 important lab tests”
> Define important vs unimportant lab tests, and reasons for selecting them.

“Important” referred to the educated estimation of the oncologists about the importance of investigating these features. We have now rephrased the text.

> Did the authors evaluate whether their methods were very sensitive to assumptions made about the missingness mechanism

The features used in the modelling contained only relatively few missing values. Therefore, in the current study, we did not investigate the missingness mechanisms.

Results

> The authors don’t clearly distinguish which methods were used in their submission of results to the DREAM challenge vs. how or why the “improvements” were made after the submission to the challenge. Furthermore, the overall risk score vs separate scores at different time points can simply be reported, without making this artificial distinction. Thus, the distinction made in the abstract between these lacks context, and I would recommend removing it.

The main difference between the original submission and the post-challenge submission was a

more systematic feature selection approach in the latter. This is now clarified in the revised manuscript. According to the suggestion by the Reviewer, we have now removed from the abstract the distinction between the overall risk score vs separate scores at different time points.

> There are many other measures for evaluating the models that the authors could report besides AUC, including Matthews Correlation Coefficient, F-measure, Precision, and Accuracy. They may consider computing and reporting these.

The true response values of the independent validation set are not publicly available. Therefore, it is not possible to calculate other performance measures than those available from the DREAM evaluation system.

> Figure 3 is unacceptable quality. Even at 200% magnification, the labels on each of the graphs are barely readable.

We agree that the font size could be larger and therefore we have increased the font size in Figure 3.

> Panel C requires some further explanation in order to interpret it. The text indicates "This suggested intuitive interpretations for the different features," which does not explain the results or whether the partial dependence can be used for feature selection or interpretation.

We agree that the legend did not address the interpretation of the partial dependence plots. Therefore, we have completely revised the text for panel C and have elaborated on the interpretation.

> "These findings are well in line with the general hypothesis that these factors are basic values representing the volume of the disease."
> By volume, are the authors referring to the extent of the disease? The extent of the disease is not the same as the survival risk, which is what the authors state they are modeling in the introduction. In cancer, volume refers to the size of the tumor and quantitative distribution (Castro-Mesta et al. 2016[1]). In fact, the authors use the numbers of lesions to infer survival risk, so it would appear that this is a circular argument.

We agree with the reviewer that the term "volume" was not correct in this sentence. We have modified the sentence accordingly.

Competing Interests: No competing interests were disclosed.

Reviewer Report 26 June 2017

https://doi.org/10.5256/f1000research.8812.r23226

Motoki Shiga
Department of Electrical, Electronic and Computer Engineering, Gifu University, Gifu, Japan

This paper offers a survival time prediction method based on a gradient boosting algorithm for a Cox proportional hazard model. The analysis of the Prostate Cancer DREAM Challenge dataset found that the prediction performance of a model built from all three clinical trials is much better than the performance of models built from each clinical trial.

Major comments:

Additional comparison with the prediction performance of the Halabi model built from all clinical trials and the performance of the models built using each trial should help to evaluate method PostCombined in more detail. And the description of procedures of the first and second winning teams is cessary.

What is the test data to compute iAUC? It was clinical trial AstraZeneca in the challenge. The manuscript should describe the procedure of the performance evaluation in detail.

Minor comments:

Please explain the machine learning method used in this work in more detail.

The threshold value of the feature selection should be described in the third paragraph of the right column in page 4.

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

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.