



METHOD ARTICLE

REVISED Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients [version 3; referees: 2 approved, 1 approved with reservations]

Sebastian Pölster¹, Pankaj Gupta¹, Lichao Wang¹, Sailesh Conjeti¹, Amin Katouzian¹, Nassir Navab^{1,2}

¹Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany

²Johns Hopkins University, Baltimore, USA

v3 First published: 16 Nov 2016, 5:2676 (doi: [10.12688/f1000research.8231.1](https://doi.org/10.12688/f1000research.8231.1))
 Second version: 27 Jun 2017, 5:2676 (doi: [10.12688/f1000research.8231.2](https://doi.org/10.12688/f1000research.8231.2))
 Latest published: 06 Jul 2017, 5:2676 (doi: [10.12688/f1000research.8231.3](https://doi.org/10.12688/f1000research.8231.3))

Abstract

Ensemble methods have been successfully applied in a wide range of scenarios, including survival analysis. However, most ensemble models for survival analysis consist of models that all optimize the same loss function and do not fully utilize the diversity in available models. We propose heterogeneous survival ensembles that combine several survival models, each optimizing a different loss during training. We evaluated our proposed technique in the context of the Prostate Cancer DREAM Challenge, where the objective was to predict survival of patients with metastatic, castrate-resistant prostate cancer from patient records of four phase III clinical trials. Results demonstrate that a diverse set of survival models were preferred over a single model and that our heterogeneous ensemble of survival models outperformed all competing methods with respect to predicting the exact time of death in the Prostate Cancer DREAM Challenge.



This article is included in the **DREAM Challenges** gateway.

Open Peer Review

Referee Status: ✓ ✓ ?

	Invited Referees		
	1	2	3
REVISED version 3 published 06 Jul 2017		✓ report	
REVISED version 2 published 27 Jun 2017	✓ report	↑	
version 1 published 16 Nov 2016	? report	? report	? report

- 1 **Donna P. Ankerst**, Technical University of Munich, Germany
- 2 **Jinfeng Xiao** , University of Illinois at Urbana–Champaign, USA
- 3 **Amber L Simpson**, Memorial Sloan Kettering Cancer Center, USA

Discuss this article

Comments (0)



This article is included in the **Machine learning: life**

sciences collection.

Corresponding author: Sebastian Pölsterl (sebastian.poelsterl@tum.de)

Author roles: **Pölsterl S:** Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Gupta P:** Methodology, Software; **Wang L:** Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Conjeti S:** Methodology, Writing – Original Draft Preparation; **Katouzian A:** Project Administration, Supervision, Writing – Original Draft Preparation; **Navab N:** Funding Acquisition, Project Administration, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

How to cite this article: Pölsterl S, Gupta P, Wang L *et al.* **Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients [version 3; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2017, 5:2676 (doi: [10.12688/f1000research.8231.3](https://doi.org/10.12688/f1000research.8231.3))

Copyright: © 2017 Pölsterl S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: This work was supported by the German Research Foundation (DFG) and the Technische Universität München within the funding program Open Access Publishing.

First published: 16 Nov 2016, 5:2676 (doi: [10.12688/f1000research.8231.1](https://doi.org/10.12688/f1000research.8231.1))

REVISED Amendments from Version 2

This paper has been revised to include the [Supplementary material](#), which was missing from the previous version 2.

[See referee reports](#)

Introduction

Today, Cox's proportional hazards model¹ is the most popular survival model because of its strong theoretical foundation. However, it only accounts for linear effects of the features and is not applicable to data with multicollinearities or high-dimensional feature vectors. In addition to Cox's proportional hazards model, many alternative survival models exist: accelerated failure time model, random survival forest², gradient boosting^{3,4}, or support vector machine⁵⁻⁹. Often it is difficult to choose the best survival model, because each model has its own advantages and disadvantages, which requires extensive knowledge of each model. Ensembles techniques leverage multiple decorrelated models – called base learners – by aggregating their predictions, which often provides an improvement over a single base learner if base learners' predictions are *accurate* and *diverse*^{10,11}. The first requirement states that a base learner must be better than random guessing and the second requirement states that predictions of any two base learners must be uncorrelated. The base learners in most ensemble methods for survival analysis are of the same type, such as survival trees in a random survival forest².

Caruana *et al.*¹² proposed *heterogeneous ensembles* for classification, where base learners are selected from a library of many different types of learning algorithms: support vector machines, decision trees, k nearest neighbor classifiers, and so forth. In particular, the library itself can contain other (homogeneous) ensemble models such that the overall model is an ensemble of ensembles. The ensemble is constructed by estimating the performance of models in the library from a separate validation set and iteratively selecting the model that increases ensemble performance the most, thus satisfying the first requirement with respect to the accuracy of base learners. To ensure that models are diverse, which is the second requirement, Margineant and Dietterich¹³ proposed to use Cohen's kappa¹⁴ to estimate the degree of disagreement between any pair of classifiers. The S pairs with the lowest kappa statistic formed the final ensemble. In addition, Rooney *et al.*¹⁵ proposed a method to construct a heterogeneous ensemble of regression models by ensuring that residuals on a validation set are uncorrelated.

We present heterogeneous survival ensembles to build an ensemble from a wide range of survival models. The main advantage of this approach is that it is not necessary to rely on a single survival model and any assumptions or limitations that model may imply. Although predictions are real-valued, a per-sample error measurement, similar to residuals in regression, generally does not exist. Instead, the prediction of a survival model consists of a risk score of arbitrary scale and a direct comparison of these values, e.g., by computing the squared error, is not meaningful. Therefore, we propose an algorithm for pruning an ensemble of survival models based on the correlation between predicted risk scores on an independent test set. We demonstrate the advantage of heterogeneous survival ensembles

in the context of the Prostate Cancer DREAM Challenge¹⁶, which asked participants to build a prognostic model to predict overall survival of patients with metastatic, castrate-resistant prostate cancer (mCRPC).

In the early stages of therapy, prostate cancer patients are usually treated with androgen deprivation therapy, but for 10–20% of patients the cancer will inevitably progress from castrate-sensitive to castrate-resistant within 5 years¹⁷. The median survival time for patients with mCRPC is typically less than 2 years¹⁷. To improve our understanding of mCRPC, the Prostate Cancer DREAM Challenge exposed the community to a large and curated set of patient records and asked participants to 1) predict patients' overall survival, and 2) predict treatment discontinuation due to adverse events. In this paper, we focus on the first sub challenge, i.e., the prediction of survival. To the best of our knowledge, this is the first scientific work that uses heterogeneous ensembles for survival analysis.

The paper is organized as follows. In the methods section, we briefly describe the framework of heterogeneous ensembles proposed by Caruana *et al.*¹² and Rooney *et al.*¹⁵ and propose an extension to construct a heterogeneous ensemble of survival models. Next, we present results of three experiments on data of the Prostate Cancer DREAM Challenge, including our final submission under the name Team CAMP. Finally, we discuss our results and close with concluding remarks.

Methods

Caruana *et al.*¹² formulated four basic steps to construct a heterogeneous ensemble:

1. Initialize an empty ensemble.
2. Update the ensemble by adding a model from the library that maximizes the (extended) ensemble's performance on an independent validation (hillclimb) set.
3. Repeat step 2 until the desired size of the ensemble is reached or all models in the library have been added to the ensemble.
4. Prune the ensemble by reducing it to the subset of base learners that together maximize the performance on a validation (hillclimb) set.

By populating the library with a wide range of algorithms, the requirement of having a diverse set of base learners is trivially satisfied. In addition, each model can be trained on a separate bootstrap sample of the training data. The second step ensures that only accurate base learners are added to the ensemble, and the fourth step is necessary to avoid overfitting on the validation set and to ensure that the ensemble comprises a diverse group of base learners. These two steps are referred to as *ensemble selection* and *ensemble pruning* and are explained in more detail below.

Efficient ensemble selection

The algorithm by Caruana *et al.*¹² has the advantage that models in the library can be evaluated with respect to any performance measure. The final heterogeneous ensemble maximizes the selected performance measure by iteratively choosing the best model from the library. Therefore, the training data \mathcal{D} needs to be split into two

non-overlapping parts: one part ($\mathcal{D}_{\text{train}}$) used to train base learners from the library, and the other part (\mathcal{D}_{val}) used as the validation set to estimate model performances. Data in the biomedical domain is usually characterized by small sample sizes, which would lead to an even smaller training set if a separate validation set is used. Caruana *et al.*¹⁸ observed that if the validation set is small, the ensemble tends to overfit more easily, which is especially concerning when the library contains many models. To remedy this problem, Caruana *et al.* [18, p. 3] proposed a solution that “embed[ded] cross-validation within ensemble selection so that all of the training data can be used for the critical ensemble hillclimbing step.” Instead of setting aside a separate validation set, they proposed to use *cross-validated models* to determine the performance of models in the library (see [Algorithm 1](#)).

Algorithm 1. Ensemble selection for survival analysis

Input: Library of N base survival models, training data \mathcal{D} , number of folds K , minimum desired performance c_{min} .

Output: Ensemble of base survival models exceeding minimum performance.

```

1   $\mathcal{M} \leftarrow \emptyset$ 
2  for  $i \leftarrow 1$  to  $N$  do
3     $\mathcal{C}_i \leftarrow \emptyset$ 
4    for  $k \leftarrow 1$  to  $K$  do
5       $\mathcal{D}_{\text{train}}^k \leftarrow k$ -th training set
6       $\mathcal{D}_{\text{test}}^k \leftarrow k$ -th test set
7       $M_{ik} \leftarrow$  Train  $k$ -th sibling of  $i$ -th survival model on  $\mathcal{D}_{\text{train}}^k$ 
8       $c_k \leftarrow$  Prediction of survival model  $M_{ik}$  on  $\mathcal{D}_{\text{test}}^k$ 
9       $\mathcal{C}_i \leftarrow \mathcal{C}_i \cup \{( \mathcal{D}_{\text{test}}^k, c_k )}$  /* Store prediction and
      associated ground truth */
10   end
11    $\bar{c}_i \leftarrow$  Performance of  $i$ -th survival model based on
      predictions and ground truth in  $\mathcal{C}_i$ 
12   if  $\bar{c}_i \geq c_{\text{min}}$  then
13      $\mathcal{M} \leftarrow \mathcal{M} \cup \{(M_{i1}, \dots, M_{ik}, \bar{c}_i)\}$  /* Store  $K$ 
      siblings and performance of  $i$ -th model */
14   end
15 end
16 return Base models in  $\mathcal{M}$ 

```

A cross-validated model is itself an ensemble of identical models, termed *siblings*, each trained on a different subset of the training data. It is constructed by splitting the training data into K equally sized folds and training one identically parametrized model on data from each of the K combinations of $K - 1$ folds. Together, the resulting K siblings form a cross-validated model.

To estimate the performance of a cross-validated model, the complete training data can be used, because the prediction of a sample

i in the training data \mathcal{D} only comes from the sibling that did not see that particular sample during training, i.e., for which $i \notin \mathcal{D}_{\text{train}}$. Therefore, estimating the performance using cross-validated models has the same properties as if one would use a separate validation set, but without reducing the size of the ensemble training data. If a truly new data point is to be predicted, the prediction of a cross-validated model is the average of the predictions of its siblings. [Algorithm 1](#) summarizes the steps in building a heterogeneous ensemble from cross-validated survival models.

Note that if a cross-validated survival model is added to the ensemble, the ensemble actually grows by K identically parametrized models of the same type – the siblings (see line 13 in [Algorithm 1](#)). Therefore, the prediction of an ensemble consisting of S cross-validated models is in fact an ensemble of $K \times S$ models.

Ensemble pruning

Ensemble selection only ensures that base learners are better than random guessing, but does not guarantee that predictions of base learners are diverse, which is the second important requirement for ensemble methods^{10,11}.

In survival analysis, predictions are real-valued, because they either correspond to a risk score or to the time of an event. Therefore, we adapted a method for pruning an ensemble of regression models that accounts for a base learner’s accuracy and correlation to other base learners¹⁵, as illustrated below.

Pruning regression ensembles. Given a library of base learners, first, the performance of each base learner is estimated either from a separate validation set or via cross-validated models following [Algorithm 1](#). To estimate the diversity of a pair of regression models, Rooney *et al.*¹⁵ considered a model’s residuals as a per-sample error measurement. Given the residuals of two models on the same data, it is straightforward to obtain a measure of diversity by computing Pearson’s correlation coefficient. They defined the diversity of a single model based on the correlation of its residuals to the residuals of all other models in the ensemble and by counting how many correlation coefficients exceeded a user-supplied threshold τ_{corr} . The diversity score can be computed by subtracting the number of correlated models from the total number of models in the ensemble and normalizing it by the ensemble size. If a model is sufficiently correlated with all other models, its diversity is zero, while if it is completely uncorrelated, its diversity is one. Moreover, they defined the accuracy of the i -th model relative to the root mean squared error (RMSE) of the best performing model as $\text{accuracy}(i) = (\min_{j=1, \dots, S} \text{RMSE}(j)) / \text{RMSE}(i)$. Finally, Rooney *et al.*¹⁵ added the diversity score of each model to its accuracy score and selected the top S base learners according to the combined accuracy-diversity score. [Algorithm 2](#) summarizes the algorithm by Rooney *et al.*¹⁵, where the correlation function would compute Pearson’s correlation coefficient between residuals of the i -th and j -th model.

Algorithm 2. Ensemble pruning algorithm of Rooney *et al.*¹⁵

Input: Set of base survival models \mathcal{M} and their average cross-validation performance, validation set \mathcal{D}_{val} , desired size S of ensemble, correlation threshold τ_{corr} .

Output: Aggregated predictions of S base survival models.

```

1  $c_{\text{max}} \leftarrow$  Highest performance score of any model in  $\mathcal{M}$ 
2 if  $|\mathcal{M}| > S$  then
3    $\mathcal{C} \leftarrow \emptyset$ 
4   for  $i \leftarrow 1$  to  $|\mathcal{M}|$  do
5      $p_i \leftarrow$  Prediction of data  $\mathcal{D}_{\text{val}}$  using  $i$ -th base survival
     model in  $\mathcal{M}$ 
6      $\text{count} \leftarrow 0$ 
7     for  $j \leftarrow 1$  to  $|\mathcal{M}|$  do
8        $p_j \leftarrow$  Prediction of data  $\mathcal{D}_{\text{val}}$  using  $j$ -th base
       survival model in  $\mathcal{M}$ 
9       if  $i \neq j \wedge \text{correlation}(p_i, p_j, \mathcal{D}_{\text{val}}) \geq \tau_{\text{corr}}$  then
10         $\text{count} \leftarrow \text{count} + 1$ 
11      end
12    end
13     $d_i \leftarrow (|\mathcal{M}| - \text{count}) / |\mathcal{M}|$ 
14     $\bar{c}_i \leftarrow$  Average cross-validation performance of  $i$ -th
    survival model in  $\mathcal{M}$ 
15     $\mathcal{C} \leftarrow \mathcal{C} \cup \{(i, \bar{c}_i / c_{\text{max}} + d_i)\}$ 
16  end
17   $\mathcal{M}^* \leftarrow$  Top  $S$  survival models with highest score according
  to  $\mathcal{C}$ 
18 else
19   $\mathcal{M}^* \leftarrow \mathcal{M}$ 
20 end
21 return Prediction of  $\mathcal{D}_{\text{val}}$  by aggregating predictions of base
  learners in survival ensemble  $\mathcal{M}^*$ 

```

Pruning survival ensembles. If the library consists of survival models rather than regression models, a persample error, similar to residuals in regression, is difficult to define. Instead, predictions are risk scores of arbitrary scales and the ground truth is the time of an event or the time of censoring. Hence, a direct comparison of a predicted risk score to the observed time of an event or the time of censoring, for instance via the squared error, is not meaningful. We propose to measure the diversity in an ensemble based on the correlation between predicted risk scores, i.e., independent of the ground truth. Here, we consider two correlation measures:

1. Pearson's correlation coefficient, and
2. Kendall's rank correlation coefficient (Kendall's τ).

Hence, we measure the diversity of a heterogeneous ensemble of survival models without requiring ground truth or a separate validation set. We believe this is not a disadvantage, because the combined score in line 15 of [Algorithm 2](#) already accounts for model accuracy, which could be estimated by the concordance index¹⁹ or integrated area under the time-dependent ROC curve^{20,21} on a validation set or using [Algorithm 1](#). In fact, since the diversity

score for survival models does not depend on ground truth, the pruning step can be postponed until the prediction phase – under the assumption that prediction is always performed for a set of samples and not a single sample alone. Consequently, the ensemble will not be static anymore and is allowed to change if new test data is provided, resulting in a dynamic ensemble.

In summary, for pruning an ensemble of survival models, [Algorithm 2](#) is applied during prediction with the following modifications:

1. Replace validation data \mathcal{D}_{val} by the feature vectors of the test data \mathcal{X}_{new} .
2. Compute the performance score using the concordance index¹⁹, integrated area under the time-dependent, cumulative-dynamic ROC curve^{20,21} or any other performance measure for censored outcomes.
3. Measure the correlation among predicted risk scores using Pearson's correlation coefficient or Kendall's rank correlation coefficient.

The prediction of the final ensemble is the average predicted risk score of all its members after pruning.

Experiments

Data

The Prostate Cancer DREAM Challenge¹⁶ provided access to 1,600 health records from three separate phase III clinical trials for training²²⁻²⁴, and data from an independent clinical trial of 470 men for testing (values of dependent variables were held back and not revealed to participants)²⁵. [Figure 1](#) illustrates the distribution of censoring and survival times of the respective trials. The median follow-up time for the MAINSAIL trial²³, the ASCENT-2 trial²², and VENICE trial²⁴ was 279, 357, and 642.5 days, respectively. For the test data from the ENTHUSE-33 trial²⁵, the median follow-up was 463 days.

We partitioned the training data into 7 sets by considering all possible combinations of the three trials constituting the training data (see [Table 1](#)). Each partition was characterized by a different set of features, ranging between 383 features for data from the MAINSAIL trial to 217 features when combining data of all three trials. Features were derived from recorded information with respect to medications, comorbidities, laboratory measurements, tumor measurements, and vital signs (see supplementary material for details). Finally, we used a random survival forest² to impute missing values in the data.

Validation scheme

We performed a total of three experiments, two based on cross-validation using the challenge training data, and one using the challenge test data from the ENTHUSE-33 trial as hold-out data. In the first experiment, we randomly split each of the datasets in [Table 1](#) into separate training and test data and performed 5-fold cross-validation. Thus, test and training data comprised different individuals from the same trial(s). We refer to this scenario as within trial validation. In the second experiment, referred to as between trials validation, we used data from one trial as hold-out data for

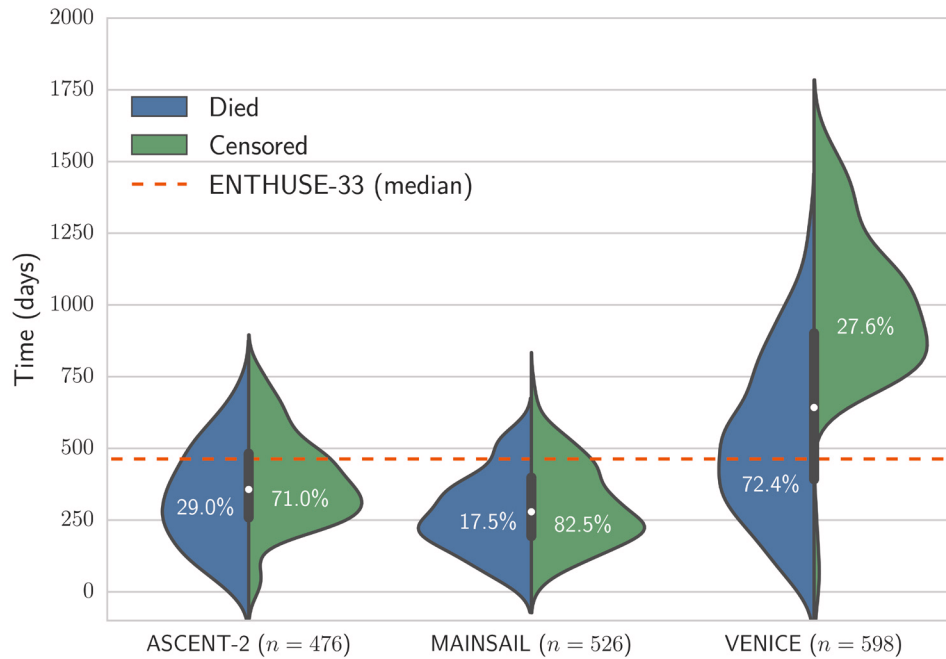


Figure 1. Overview of distribution of survival and censoring times in training data from the ASCENT-2, VENICE, MAINSAIL, and ENTHUSE-33 trial²²⁻²⁵. Numbers in brackets denote the total number of patients in the respective trial, and the dashed line is the median follow-up time in the ENTHUSE-33 trial, which was used as independent test data.

Table 1. Different sets of features that were constructed by considering the intersection between trials in the Prostate Cancer DREAM Challenge.

ASCENT-2	MAINSAIL	VENICE	Samples	Features
•	•	•	1,600	217
	•	•	1,124	345
•		•	1,074	220
•	•		1,002	221
		•	598	350
	•		526	383
•			476	223

testing and data from one or both of the remaining trials for training. This setup resembles the challenge more closely, where test data corresponded to a separate trial too. We only considered features that were part of both the training and test data. In each experiment above, the following six survival models were evaluated:

1. Cox’s proportional hazards model¹ with ridge (l_2) penalty,
2. Linear survival support vector machine (SSVM)⁹,
3. SSVM with the clinical kernel²⁶,
4. Gradient boosting of negative log partial likelihood of Cox’s proportional hazards model³ with randomized regression trees as base learners^{27,28},

5. Gradient boosting of negative log partial likelihood of Cox’s proportional hazards model³ with componentwise least squares as base learners²⁹,

6. Random survival forest².

In addition, the training of each survival model was wrapped by grid search optimization to find optimal hyper-parameters. The complete training data was randomly split into 80% for training and 20% for testing to estimate a model’s performance with respect to a particular hyper-parameter configuration. The process was repeated for ten different splits of the training data. Finally, a model was trained on the complete training data using the hyper-parameters that on average performed the best across all ten repetitions. Performance was estimated by Harrell’s concordance index (*c* index)¹⁹. All continuous features were normalized to zero mean and unit variance and nominal and ordinal features were dummy coded.

For the Prostate Cancer DREAM Challenge’s final evaluation, we built a heterogeneous ensemble from a wide range of survival models. In sub challenge 1a, the challenge organizers evaluated submissions based on the integrated area under the time-dependent, cumulative-dynamic ROC curve (iAUC)^{20,21} – integrated over time points every 6 months up to 30 months after the first day of treatment – and in sub challenge 1b, based on the root mean squared error (RMSE) with respect to deceased patients in the test data. The performance of submitted models was estimated based on 1,000 bootstrap samples of the ENTHUSE-33 trial data and the Bayes factor to the top performing model and a baseline model by Halabi *et al.*³⁰ (only for sub challenge 1a). The Bayes factor provides an

alternative to traditional hypothesis testing, which relies on p -values to determine which of two models is preferred (see e.g. 31). According to Jeffreys³², a Bayes factor in the interval [3; 10] indicates moderate evidence that the first model outperformed the second model and strong evidence if the Bayes factor is greater 10, else evidence is insufficient.

Results

With-in trial validation

Figure 2 summarizes the average cross-validation performance across all five test sets for all seven datasets in Table 1. Overall, the average concordance index ranged between 0.629 and 0.713 with a mean of 0.668. It is noteworthy that all classifiers but SSVM models performed best on data of the MAINSAIL trial, which comprised 526 subjects and the highest number of features among all trials (383 features). A SSVM was likely to have an disadvantage due to the high number of features and because feature selection is not embedded into its training as for the remaining models. In fact, SSVM models performed worst on data from the MAINSAIL and VENICE trials, which were the datasets with the most features. SVM-based models performed best if data from at least two trials were combined, which increased the number of samples and decreased the number of features. Moreover, the results show that linear survival support vector machines performed poorly. A

considerable improvement could be achieved when using kernel-based survival support vector machines with the clinical kernel, which is especially useful if data is a mix of continuous, categorical and ordinal features. For low-dimensional data, the kernel SSVM could perform equally well as or better than gradient boosting models, but was always outperformed by a random survival forest.

When considering the performance of models across all datasets (last row in Figure 2), random survival forests and Cox’s proportional hazards models stood out with an average c index of 0.681, outperforming the third best: gradient boosting with componentwise least squares base learners. Random survival forests performed better than Cox’s proportional hazards models on 4 out of 7 datasets and was tied on one dataset. The results seem to indicate that a few datasets contain non-linearities, which were captured by random survival forests, but not by gradient boosting with componentwise least squares and Cox’s proportional hazards models. Nevertheless, Cox’s proportional hazards model only performed significantly better than linear SSVM when averaging results over all datasets (see Figure 4).

Finally, we would like to mention that 5 out of 6 survival models performed worst on the VENICE data. Although it contained the largest number of patients, the variance of follow-up times is more

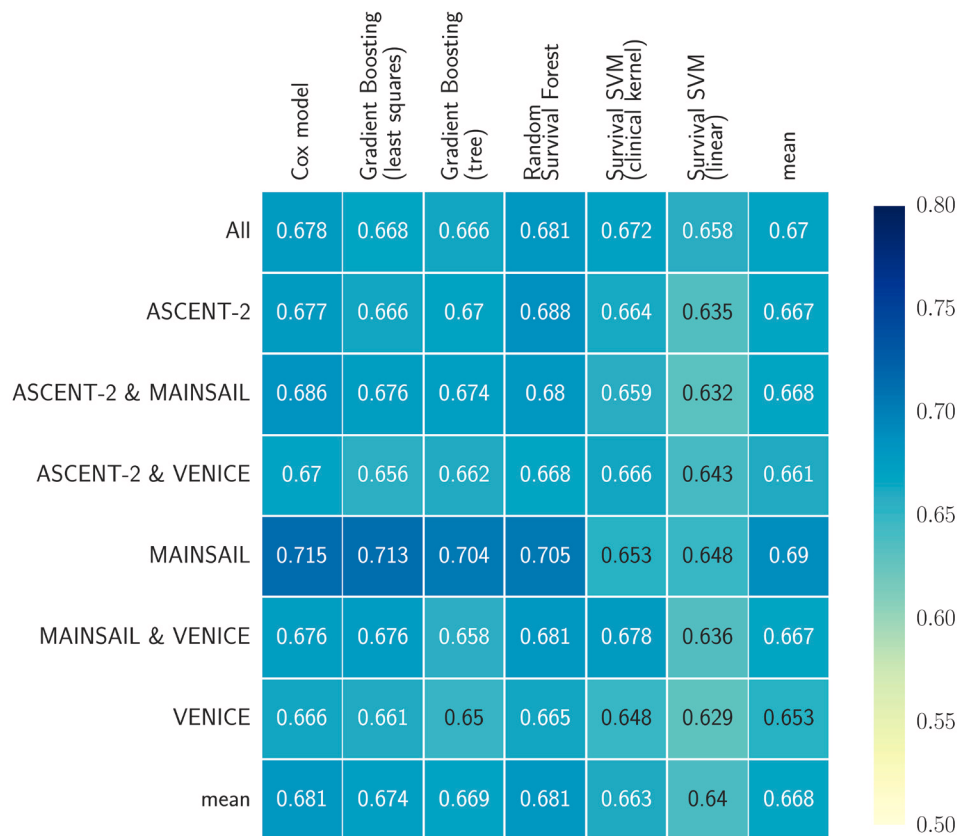


Figure 2. Cross-validation performance of survival models on data from from the ASCENT-2, VENICE, and MAINSAIL trial, as well as any combination of these datasets. The last column (mean) denotes the average performance of all models on a particular dataset and the last row (mean) denotes the average performance of a particular model across all datasets. Numbers indicate the average of Harrell’s concordance index across five cross-validation folds.

than two-fold larger compared to ASCENT-2 and MAINSAIL ($\sigma^2 \approx 342.9$ versus 165.1 for ASCENT-2 and 140.2 for MAINSAIL). Moreover, the overlap in the distribution of censoring and survival times was rather small (see [Figure 1](#)). Thus, the difference between observed time points in the training and test data based on the VENICE trial is likely more pronounced than for the data from the MAINSAIL or ASCENT-2 trials, which means a survival model has to generalize to a much larger time period. Moreover, the amount of censoring in the VENICE trial is relatively low compared to the other trials. Therefore, the observed drop in performance might stem from the fact that the bias of Harrell's concordance index usually increases as the amount of censoring increases³³. As an alternative, we considered the integrated area under the time-dependent, cumulative-dynamic ROC curve^{20,21}, which was the main evaluation measure in the Prostate Cancer DREAM Challenge. However, comparing the estimated integrated area under the ROC curve across multiple datasets is not straightforward when follow-up times differ largely among trials (see [Figure 1](#)). If the integral is estimated from time points that exceed the follow-up time of almost all patients, the inverse probability of censoring weights used in the estimator of the integrated area under the curve cannot be computed, because the estimated probability of censoring at that time point becomes zero. On the other hand, if time points are defined too conservatively, the follow-up period of most patients will end after the last time point and the estimator would ignore a large portion of the follow-up period. Hence, defining time points that lead to adequate estimates of performance in all three datasets is challenging due to large differences in the duration of follow-up periods.

Between trials validation

In the second experiment, training and test data were from separate trials, which resembled the setup of the Prostate Cancer DREAM Challenge. We included heterogeneous ensembles in the analyses, trained on a library of models that included multiple copies of each survival model, each with a different hyper-parameter configuration. The library excluded linear SSVM, because it performed poorly in previous experiments, and Cox's proportional hazards model, because its Newton-Raphson optimization algorithm used a constant step size instead of a line search, which occasionally led to oscillation around the minimum during ensemble selection. We investigated whether the observed differences in performance are statistically significant by performing a Nemenyi post-hoc test³⁴ based on the results of all train-test-set combinations in [Figure 3](#). [Figure 5](#) summarizes the results.

The results confirmed observations discussed in the previous section: 1) on average, random survival forests performed better than gradient boosting models and SSVMs, and 2) using SSVM with the clinical kernel was preferred over the linear model. Heterogeneous ensemble ranked first in our experiments, tied with Cox's proportional hazards model, and significantly outperformed linear SSVM and gradient boosting with regression trees. Among the top five models – which did not perform significantly different from each other – in [Figure 5](#), heterogeneous ensemble stands out by having the lowest variance: its performance ranged between 0.636 and 0.689 ($\Delta = 0.053$), which is a 14% reduction compared to the runner-up (Cox's model: $\Delta = 0.061$) and a 12% to 40% reduction when compared to individual base learners in the library (gradient boosting with componentwise least squares: $\Delta = 0.060$, SSVM with

clinical kernel: $\Delta = 0.088$). The results demonstrate that combining a diverse set of survival models in a heterogeneous ensemble improves performance and increases reliability.

If performance was estimated on the VENICE data, all models performed considerably worse compared to performance estimated on the other datasets. We believe the reason for these results are similar to the cross-validation results on the VENICE data described in the previous section. The bias of Harrell's concordance index due to vastly different amounts of censoring among trials could be one factor, while the other could be that the follow-up times differed drastically between training and testing. If the follow-up period is much shorter in the training data than in the testing data, it is likely that models generalize badly for time points that were never observed in the training data, which is only the case if the VENICE data is used for testing, but not if data from the MAINSAIL or the ASCENT-2 trial is used (cf. [Figure 1](#)). Interestingly, all models, except linear SSVM, performed best when trained on the maximum number of available patient records, which is different from results in the previous section, where models trained on data with more features performed better.

An unexpected result is that Cox's proportional hazards model was able to outperform many of the machine learning methods, including random survival forest, which is able to implicitly model non-linear relationships that are not considered by Cox's proportional hazards model. A possible explanation why the Cox model performed on par with more complicated machine learning methods might be the fact the effective sample size reduces if the amount of censoring increases, as kindly pointed out by one referee. If most samples are censored, the effective size of the study decreases proportionally, which in turn makes it more challenging to reliably identify non-linear effects, which would be the strength of the advanced survival models in our experiments. Following Occam's razor, the results suggest that, in this case, a simple model is preferred.

Results also indicate that models with embedded feature selection (gradient boosting and random survival forest) were not significantly better than models that take into account all features (Cox model and SSVM with the clinical kernel). The fact that models with embedded feature selection, in particular gradient boosting with componentwise least squares base learner, performed poorly might be false positive selected features, i.e., features that are actually not associated with survival. In high dimensions, methods with embedded feature selection often suffer from instability, i.e., the set of selected features can vary widely when repeatedly fitting a model, e.g., when determining optimal hyper-parameters³⁵. This problem seems to be more pronounced when evaluating models on data from a different study. The number of false positive selections could be controlled by performing stability selection³⁵.

Challenge hold-out data

To summarize, results presented in the previous two sections demonstrate that

1. SSVM should be used in combination with the clinical kernel.

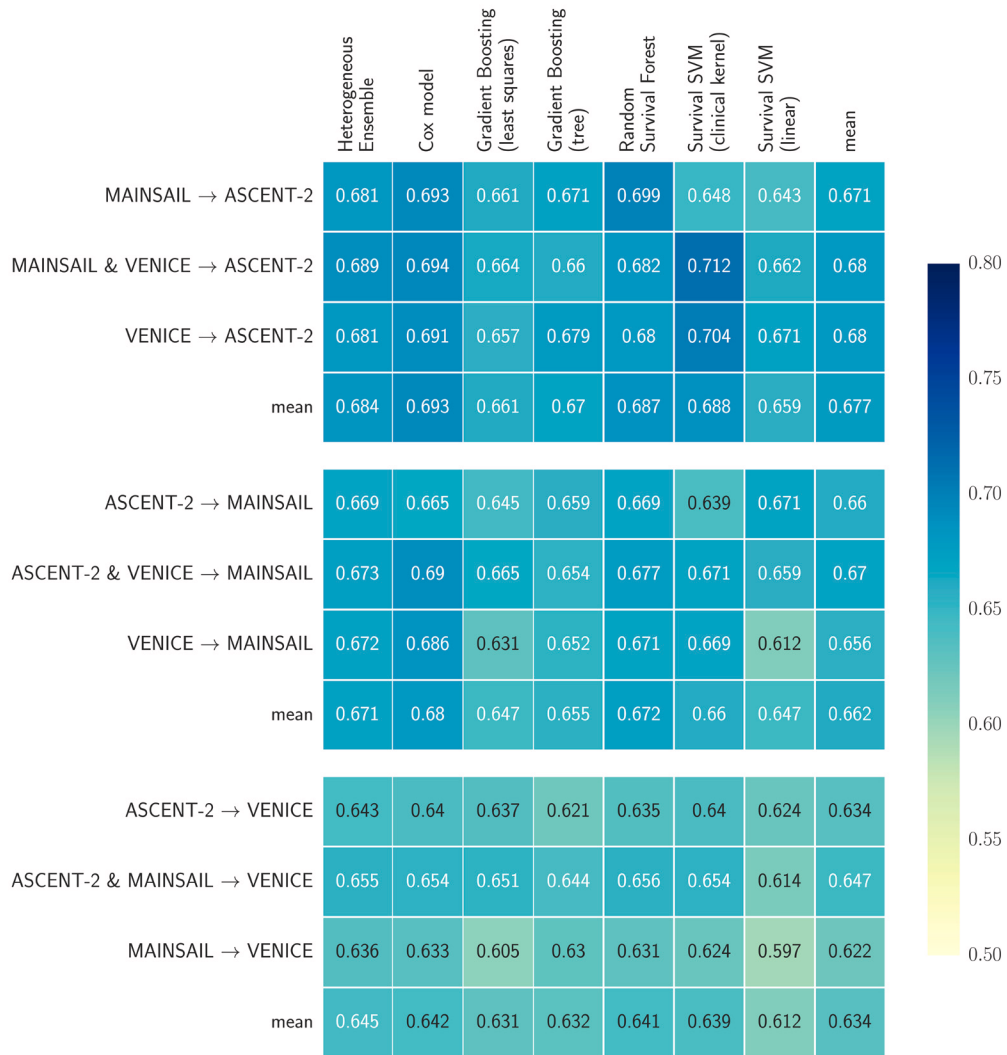


Figure 3. Performance results using hold-out data from from the ASCENT-2, VENICE, and MAINSAIL trial. One trial was used as hold-out data (indicated by the name to the right of the arrow) and one or two of the remaining trials as training data. Numbers indicate Harrell's concordance index on the hold-out data.

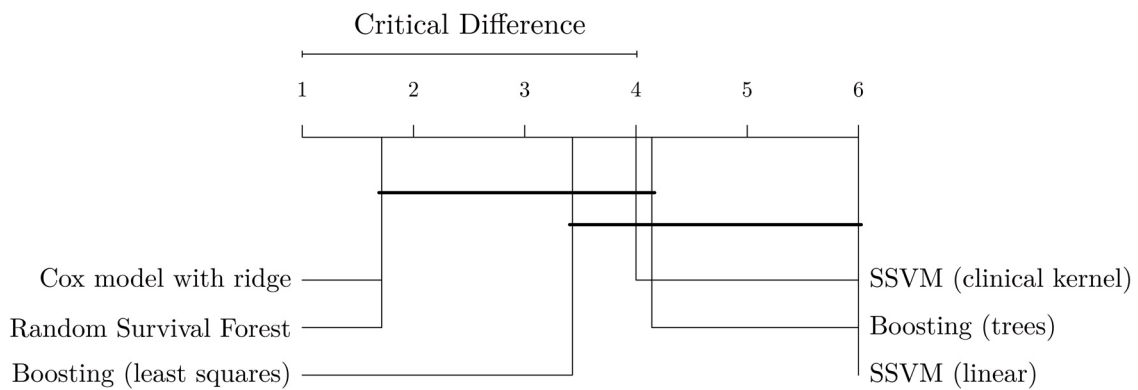


Figure 4. Comparison of methods based on experiments in Figure 2 with the Nemenyi post-hoc test³⁴. Methods are sorted by average rank (left to right) and groups of methods that are not significantly different are connected (p -value > 0.05).

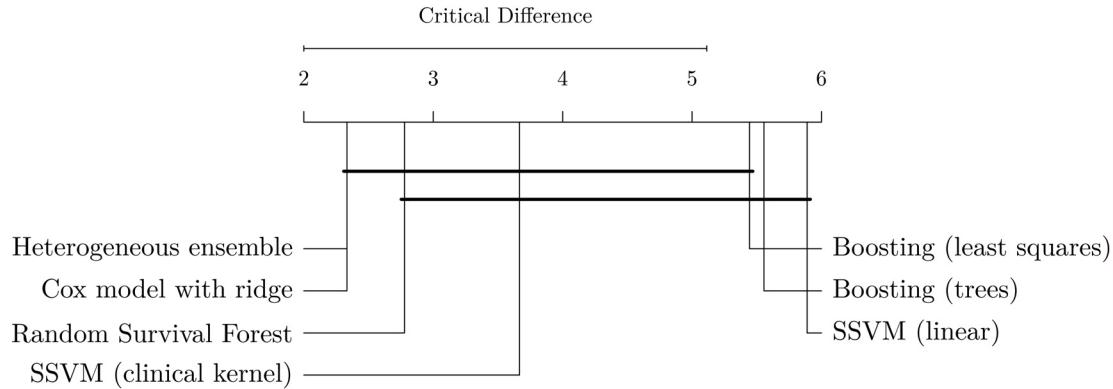


Figure 5. Comparison of methods based on experiments in Figure 3 with the Nemenyi post-hoc test³⁴. Methods are sorted by average rank (left to right) and groups of methods that are not significantly different are connected (p -value > 0.05).

2. Increasing the number of samples is preferred over increasing the number of features, especially if follow-up periods are large.
3. There is no single survival model that is clearly superior to all other survival models.

From these observations, we concluded that employing heterogeneous survival models, trained on all 1,600 patient records in the training data, would be most reliable. We built two ensembles using [Algorithm 1](#) and [Algorithms 2](#): one maximizing Harrell's concordance index¹⁹, and one minimizing the RMSE. The former was constructed from a library of 1,801 survival models for sub challenge 1a ($K = 5$, $c_{\min} = 0.66$, $\tau_{\text{corr}} = 0.6$, $S = 90$) and the latter from a library of 1,842 regression models for sub challenge 1b ($K = 5$, $c_{\min} = 0.85$, $\tau_{\text{corr}} = 0.6$, $S = 92$). We submitted predictions based on these two models to the Prostate Cancer DREAM Challenge. The results in the remainder of this section refer to the final evaluation carried out by the challenge organizers.

Sub challenge 1a. Four of the six survival models evaluated in the cross-validation experiments formed the basis of the ensemble (see [Table 2](#)). [Figure 6](#) depicts scatter plots comparing models' performance and diversity. Most of the gradient boosting models with regression trees as base learners were pruned because their predictions were redundant to other models in the ensemble ([Figure 6A](#)). In contrast, all random survival models remained in the ensemble throughout ([Figure 6C](#)). We observed the highest diversity for gradient boosting models (mean = 0.279) and the highest accuracy for random survival forests (mean = 0.679). The final ensemble comprised all types of survival models in the library, strengthening our conclusion that a diverse set of survival models is preferred over a single model.

In the challenge's final evaluation based on 313 patients of the ENTHUSE-33 trial, 30 out of 51 submitted models outperformed the baseline model by Halabi *et al.*³⁰ by achieving a Bayes factor greater than 3¹⁶. There was a clear winner in team FIMM-UTU and the performance of the remaining models were very close to each

other; there was merely a difference of 0.0171 points in integrated area under the ROC curve (iAUC) between ranks 2 and 25¹⁶.

The proposed heterogeneous ensemble of survival models by Team CAMP achieved an iAUC score of 0.7646 on the test data and was ranked 23rd according to iAUC and 20th according to Bayes factor with respect to the best model (FIMM-UTU). When considering the Bayes factor of the proposed ensemble method to all other models, there is only sufficient evidence (Bayes factor greater 3) that five models performed better (FIMM-UTU, Team Cornfield, TeamX, jls, and KUstat). The Bayes factor to the top two models was 20.3 and 6.6 and ranged between 3 and 4 for the remaining three models. With respect to the model by Halabi *et al.*³⁰, there was strong evidence (Bayes factor 12.2; iAUC 0.7432) that heterogeneous ensembles of survival models could predict survival of mCRPC patients more accurately.

Sub challenge 1b. In subchallenge 1b, participants were tasked with predicting the exact time of death rather than ranking patients according to their survival time. Similar to subchallenge 1a, our final model was a heterogeneous ensemble, but based on a different library of models (see [Table 3](#)).

[Figure 7](#) illustrates the RMSE and diversity of all 1,281 models after the first pruning step (cf. [Table 3](#)). In contrast to the ensemble of survival models used in subchallenge 1a, the ensemble in this subchallenge was characterized by very little diversity: the highest diversity was 0.064. In fact, all 92 models included in the final ensemble had a diversity score below 0.001, which means that pruning was almost exclusively based on the RMSE. Gradient boosting models with componentwise least squares base learners were completely absent from the final ensemble and only two hybrid survival support vector machine models had a sufficiently low RMSE to be among the top 5%.

The evaluation of all submitted models on the challenge's final test data from the ENTHUSE-33 trial revealed that our proposed heterogeneous ensemble of regression models achieved the lowest root mean squared error (194.4) among all submissions¹⁶. The difference

Table 2. Heterogeneous ensemble of survival models used in sub challenge 1a of the Prostate Cancer DREAM Challenge. *All* denotes the initial size of the ensemble, *Pruned* the size after pruning models with Harrell's concordance index below 0.66, and *Top 5%* to the final size of the ensemble corresponding to the top 5% according the combined accuracy and diversity score in [Algorithm 2](#).

Survival model	Configurations		
	All	Pruned	Top 5%
Gradient boosted Cox model (tree) ^{3,27}	1,728	936	56
Gradient boosted Cox model (least squares) ^{3,29}	36	36	7
Random survival forest ²	24	24	24
Ranking-based survival SVM (clinical kernel) ^{9,26}	13	3	3
Σ	1,801	999	90

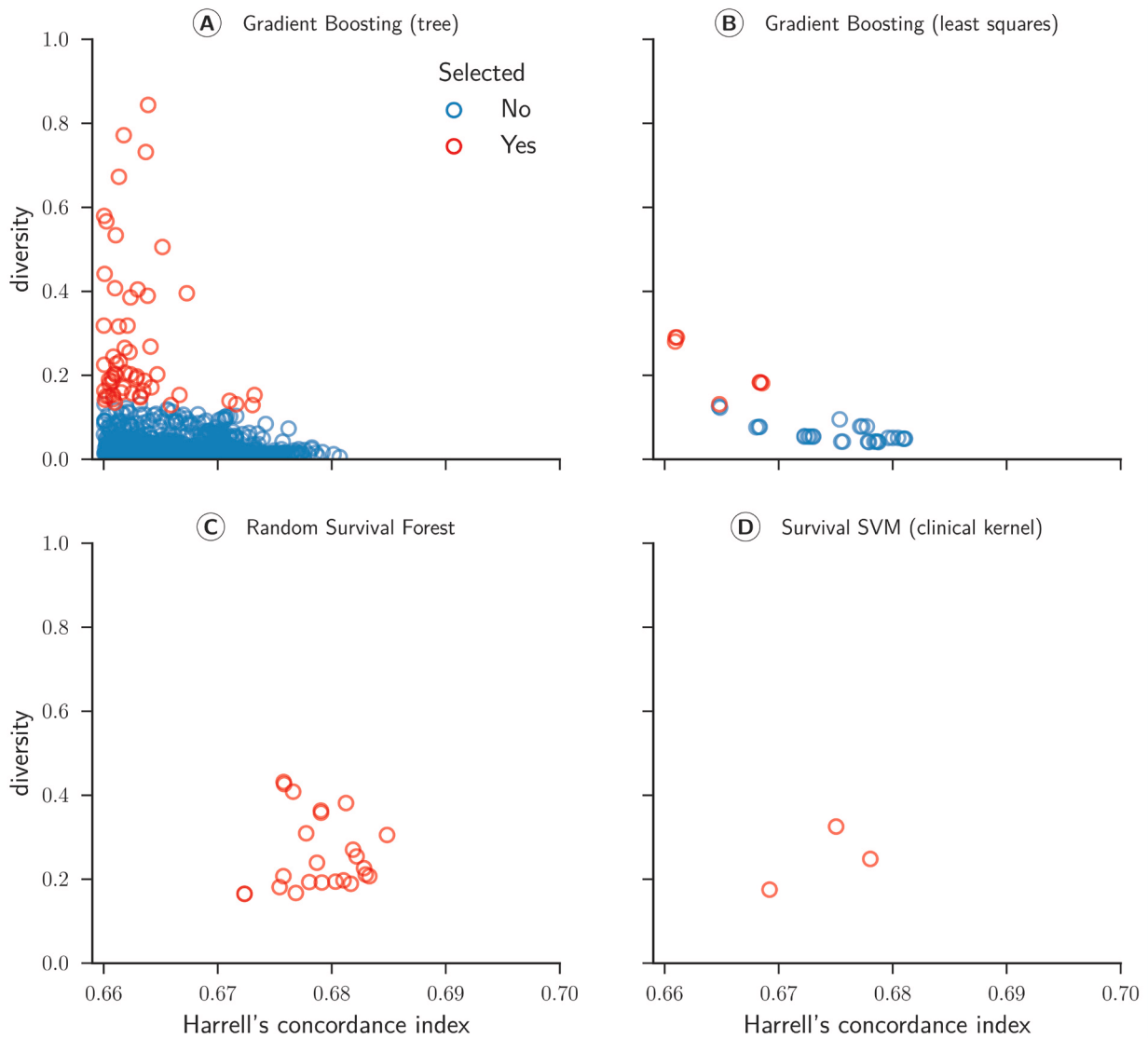


Figure 6. Concordance index and diversity score of 999 survival models for sub challenge 1a. The concordance index was evaluated by cross-validated models on the training data from the from the ASCENT-2, VENICE, and MAINSAIL trial. Diversity was computed based on Pearson's correlation coefficient between predicted risk scores for 313 patients of the ENTHUSE-33 trial (final scoring set).

Table 3. Heterogeneous ensemble used in sub challenge 1b of the Prostate Cancer DREAM Challenge. *All* denotes the initial size of the ensemble, *Pruned* the size after pruning models with a root mean squared error more than 15% above the error of the best performing model, and *Top 5%* to the final size of the ensemble corresponding to the top 5% according the combined accuracy and diversity score in *Algorithm 2*. AFT: Accelerated Failure Time.

Regression model	Configurations		
	All	Pruned	Top 5%
Gradient boosted AFT model (tree) ^{4,27}	1,728	1,236	90
Gradient boosted AFT model (least squares) ^{4,29}	36	36	0
Hybrid survival SVM (clinical kernel) ^{9,28}	78	9	2
Σ	1,842	1,281	92

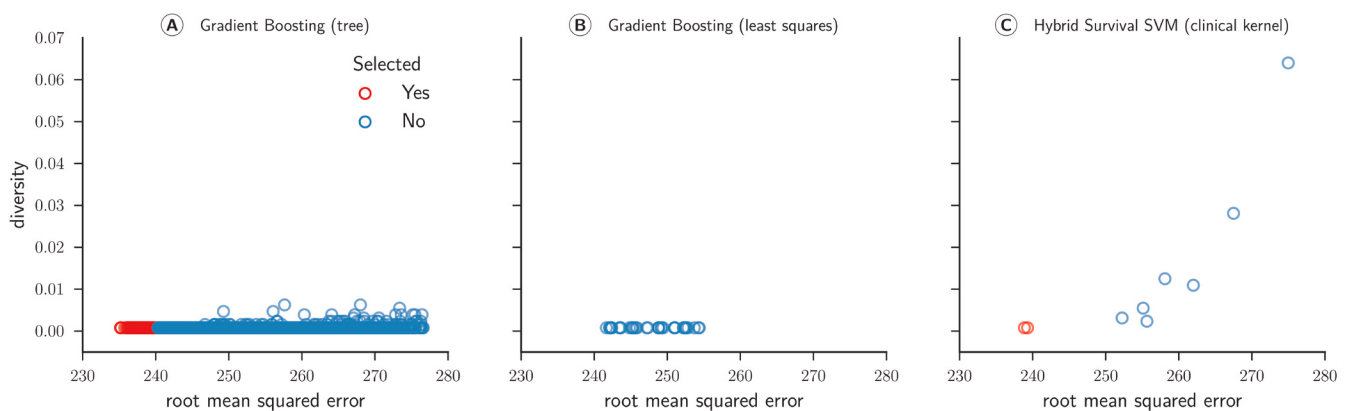


Figure 7. Root mean squared error (RMSE) and diversity score of 1,281 regression models for sub challenge 1b. The RMSE was evaluated by cross-validated models on the training data from the ASCENT-2, VENICE, and MAINSAIL trial. Diversity was computed based on Pearson’s correlation coefficient between residuals on the training data.

in RMSE between the 1st placed model and the 25th placed model was less than 25. With respect to our proposed winning model, there was insufficient evidence to state it outperformed all other models, because the comparison to five other models yielded a Bayes factor less than three (Team Cornfield, M S, JayHawks, Bmore Dream Team, and A Bavarian dream).

Discussion

From experiments on the challenge training data, we concluded that it would be best to combine data from all three clinical trials to train a heterogeneous ensemble, because maximizing the number of distinct time points was preferred. Interestingly, the winning team of sub challenge 1a completely excluded data from the ASCENT-2 trial in their solution. They argued that it was too dissimilar to data of the remaining three trials, including the test data³⁶. Therefore, it would be interesting to investigate unsupervised approaches that could deduce a similarity or distance measure between patients, which can be used to decrease the influence of outlying patients during training.

The second important conclusion from our experiments is that no survival model clearly outperformed all other models in all the evaluated scenarios. Our statistical analysis based on results of the

between trials validation revealed that Cox’s proportional hazards model performed significantly better than the linear survival support vector machine and gradient boosting with regression trees as base learners, and that the random survival forest performed significantly better than linear survival support vector machines; the remaining differences were deemed statistically insignificant. Therefore, we constructed a heterogeneous ensemble of several survival models with different hyper-parameter configurations and thereby avoided relying only on a single survival model with a single hyper-parameter configuration. In total, we considered two libraries, each consisting of over 1,800 different models, which were pruned to ensure accuracy and diversity of models – we observed only minor differences when substituting Pearson’s correlation for Kendall’s rank correlation during ensemble pruning.

The proposed ensemble approach was able to outperform all competing models in sub challenge 1b, where the task was to predict the exact time of death. In sub challenge 1a, participants had to provide a relative risk score and our ensemble approach was significantly outperformed by five competing models¹⁶. Due to large differences in teams’ overall solutions it is difficult to pinpoint the reason for the observed performance difference: it could be attributed to the choice of base learners, or to choices made during pre-processing

or filtering the data. From our experience of the three intermediate scoring rounds before the final submission, we would argue that identifying the correct subset of patients in the training data that is most similar to the test data is more important than choosing a predictive model. By training a survival model on data combined from three trials and applying it to patients from a fourth trial, inconsistencies between trials inevitably lead to outliers with respect to the test data, which in turn diminishes the performance of a model – if not addressed explicitly during training.

A possible explanation why the heterogeneous ensemble worked better for survival time prediction (sub challenge 1b) than for risk score prediction (sub challenge 1a) might be that we maximized the concordance index during ensemble construction and not the area under the time-dependent ROC curve, which was used in the challenge's final evaluation. In addition, we aggregated predictions of survival models by averaging, although predictions of survival models are not necessarily on the same scale. In regression, the prediction is a continuous value that directly corresponds to the time of death, which allows simple averaging of individual predictions. In survival analysis, semantics are slightly different. Although predictions are real-valued as well, the prediction of a survival model does generally not correspond to the time of death, but is a risk score on an arbitrary scale. A homogeneous ensemble only consists of models of the same type, therefore predictions can be aggregated by simply computing the average. A problem arises for heterogeneous ensembles if the scale of predicted risk scores differs among models. To illustrate the problem, consider an ensemble consisting of survival trees as used in a random survival forest² and ranking-based linear survival support vector machines⁹. The prediction of the former is based on the cumulative hazard function estimated from samples residing in the leaf node a new sample was assigned to. Thus, predictions are always positive due to the definition of the cumulative hazard function (see e.g. 37). In contrast, the prediction of a linear SSVM is the inner product between a model's vector of coefficients and a sample's feature vector, which can take on negative as well as positive values. It is easy to see that, depending on the scale difference, simply averaging predicted risk scores favors models with generally larger risk scores (in terms of absolute value) or positive and negative predicted risk scores cancel each other out. Instead of simply averaging risk scores, the problem could be alleviated if model risk scores were first transformed into ranks, thereby putting them on a common scale, before averaging the resulting ranks. We evaluated this approach after the Prostate Cancer DREAM Challenge ended: averaging ranks instead of raw predicted risk scores increased the iAUC value from 0.7644 to 0.7705 on a random sub sample of the ENTHUSE-33 trial.

Finally, we want to pay particular attention to the challenge of combining multiple patients populations for risk prediction. As mentioned above, the follow-up periods and the information collected for the four studies considered here differed vastly.

Figure 5 illustrates that there is no single model equally suitable for all cohorts. This problem arises if prediction models are badly calibrated with respect to the target cohort. If outcome information for the target cohort is available, recalibration methods can be used to improve calibration and discrimination of the risk score^{38–41}. In the context of the Prostate Cancer DREAM Challenge, Kondofersky *et al.*⁴² showed that employing simple recalibration models significantly improved prediction performance for subchallenge 1b. Moreover, researchers developed models specifically designed to amalgamate diverse patient cohorts by utilizing ideas from machine learning^{43–45}.

Conclusions

We proposed heterogeneous survival ensembles that are able to aggregate predictions from a wide variety of survival models. We evaluated our method using data from an independent fourth trial from the Prostate Cancer DREAM Challenge. Our proposed ensemble approach could predict the exact time of death more accurately than any competing model in sub challenge 1b and was significantly outperformed by 5 out of 50 competing solutions in sub challenge 1a. We believe this result is encouraging and warrants further research in using heterogeneous ensembles for survival analysis. The source code is available online <https://www.synapse.org/#!/Synapse:syn3647478>.

Author contributions

SP prepared the raw datasets, implemented the survival models, and wrote the manuscript. PG and SP performed analyses to establish the final models. LW and SC contributed to establishing the final models. AK and NN supervised the analysis.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the German Research Foundation (DFG) and the Technische Universität München within the funding program Open Access Publishing.

Acknowledgements

We thank Sage Bionetworks, the DREAM organization, and Project Data Sphere for developing and supplying data for the Prostate Cancer DREAM Challenge. We thank the Leibniz Supercomputing Centre (LRZ, www.lrz.de) for providing the computational resources for our experiments. 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.

Supplementary material

Data pre-processing, missing imputations, and hyper-parameter configurations.

[Click here to access the data.](#)

References

1. Cox DR: **Regression models and life tables.** *J R Stat Soc Series B.* 1972; **34**(2): 187–220.
[Reference Source](#)
2. Ishwaran H, Kogalur UB, Blackstone EH, *et al.*: **Random survival forests.** *Ann Appl Stat.* 2008; **2**(3): 841–860.
[Publisher Full Text](#)
3. Ridgeway G: **The state of boosting.** *Comput Sci Stat.* 1999; **31**: 172–181.
[Reference Source](#)
4. Hothorn T, Bühlmann P, Dudoit S, *et al.*: **Survival ensembles.** *Biostatistics.* 2006; **7**(3): 355–373.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Van Belle V, Pelckmans K, Suykens JAK, *et al.*: **Support vector machines for survival analysis.** In *Proc of the 3rd International Conference on Computational Intelligence in Medicine and Healthcare.* 2007; 1–8.
[Reference Source](#)
6. Shivaswamy PK, Chu W, Jansche M: **A support vector approach to censored targets.** In *7th IEEE International Conference on Data Mining.* 2007; 655–660.
[Publisher Full Text](#)
7. Khan FM, Zubek VB: **Support vector regression for censored data (SVRC): A novel tool for survival analysis.** In *8th IEEE International Conference on Data Mining.* 2008; 863–868.
[Publisher Full Text](#)
8. Eleuteri A: **Support vector survival regression.** In *4th IET International Conference on Advances in Medical, Signal and Information Processing.* 2008; 1–4.
[Publisher Full Text](#)
9. Pölsterl S, Navab N, Katouzian A: **Fast training of support vector machines for survival analysis.** In *Machine Learning and Knowledge Discovery in Databases.* volume 9285 of Lecture Notes in Computer Science. 2015; 243–259.
[Publisher Full Text](#)
10. Hansen LK, Salamon P: **Neural network ensembles.** *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1990; **12**(10): 993–1001.
[Publisher Full Text](#)
11. Dietterich TG: **Ensemble methods in machine learning.** In *1st International Workshop on Multiple Classifier Systems.* 2000; **1857**: 1–15.
[Publisher Full Text](#)
12. Caruana R, Niculescu-Mizil A, Crew G, *et al.*: **Ensemble selection from libraries of models.** In *22nd International Conference on Machine Learning.* 2004.
13. Margineantu DD, Dietterich TG: **Pruning adaptive boosting.** In *14th International Conference on Machine Learning.* 1997; 211–218.
[Reference Source](#)
14. Cohen J: **A coefficient of agreement of nominal scales.** *Educ Psychol Meas.* 1960; **20**(1): 37–46.
[Publisher Full Text](#)
15. Rooney N, Patterson D, Anand S, *et al.*: **Dynamic integration of regression models.** In *Proc of the 5th International Workshop on Multiple Classifier Systems.* 2004; **3077**: 164–173.
[Publisher Full Text](#)
16. 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.* 2017; **18**(1): 132–142.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Kirby M, Hirst C, Crawford ED: **Characterising the castration-resistant prostate cancer population: a systematic review.** *Int J Clin Pract.* 2011; **65**(11): 1180–1192.
[PubMed Abstract](#) | [Publisher Full Text](#)
18. Caruana R, Munson A, Niculescu-Mizil A: **Getting the most out of ensemble selection.** In *6th IEEE International Conference on Data Mining.* 2006; 828–833.
[Publisher Full Text](#)
19. Harrell FE Jr, Califf RM, Pryor DB, *et al.*: **Evaluating the yield of medical tests.** *JAMA.* 1982; **247**(18): 2543–2546.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Uno H, Cai T, Tian L, *et al.*: **Evaluating prediction rules for t-year survivors with censored regression models.** *J Am Stat Assoc.* 2007; **102**(478): 527–537.
[Publisher Full Text](#)
21. 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](#)
22. 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](#)
23. Petrylak DP, Vogelzang NJ, Budnik N, *et al.*: **Docetaxel and prednisone with or without lenalidomide in chemotherapy-naïve 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](#)
24. 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](#)
25. 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](#)
26. Daemen A, Timmerman D, Van den Bosch T, *et al.*: **Improved modeling of clinical data with kernel methods.** *Artif Intell Med.* 2012; **54**(2): 103–114.
[PubMed Abstract](#) | [Publisher Full Text](#)
27. Breiman L, Friedman JH, Stone CJ, *et al.*: **Classification and Regression Trees.** Wadsworth International Group, 1984.
[Reference Source](#)
28. Breiman L: **Random forests.** *Mach Learn.* 2001; **45**(1): 5–32.
[Publisher Full Text](#)
29. Bühlmann P, Yu B: **Boosting with the L_2 loss.** *J Am Stat Assoc.* 2003; **98**(462): 324–339.
[Publisher Full Text](#)
30. 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](#)
31. Wasserman L: **Bayesian Model Selection and Model Averaging.** *J Math Psychol.* 2000; **44**(1): 92–107.
[PubMed Abstract](#) | [Publisher Full Text](#)
32. Jeffreys H: **The Theory of Probability.** Oxford University Press, 1961.
[Reference Source](#)
33. Antolini L, Boracchi P, Biganzoli E: **A time-dependent discrimination index for survival data.** *Stat Med.* 2005; **24**(24): 3927–3944.
[PubMed Abstract](#) | [Publisher Full Text](#)
34. Demšar J: **Statistical comparisons of classifiers over multiple data sets.** *J Mach Learn Res.* 2006; **7**: 1–30.
[Reference Source](#)
35. Meinshausen N, Bühlmann P: **Stability selection.** *J Roy Stat Soc B.* 2010; **72**(4): 417–473.
[Publisher Full Text](#)
36. Laajala TD, Khan S, Airola A, *et al.*: **Predicting patient survival and treatment discontinuation in DREAM 9.5 mCRPC challenge.**
[Reference Source](#)
37. Klein JP, Moeschberger ML: **Survival Analysis: Techniques for Censored and Truncated Data.** Springer, 2nd edition, 2003.
[Publisher Full Text](#)
38. Steyerberg EW, Borsboom GJ, van Houwelingen HC, *et al.*: **Validation and updating of predictive logistic regression models: a study on sample size and shrinkage.** *Stat Med.* 2004; **23**(16): 2567–2586.
[PubMed Abstract](#) | [Publisher Full Text](#)
39. Janssen KJ, Moons KG, Kalkman CJ, *et al.*: **Updating methods improved the performance of a clinical prediction model in new patients.** *J Clin Epidemiol.* 2008; **61**(1): 76–86.
[PubMed Abstract](#) | [Publisher Full Text](#)
40. Toll DB, Janssen KJ, Vergouwe Y, *et al.*: **Validation, updating and impact of clinical prediction rules: A review.** *J Clin Epidemiol.* 2008; **61**(11): 1085–1094.
[PubMed Abstract](#) | [Publisher Full Text](#)
41. Su TL, Jaki T, Hickey GL, *et al.*: **A review of statistical updating methods for clinical prediction models.** *Stat Methods Med Res.* 2016; pii: 0962280215626466.
[PubMed Abstract](#) | [Publisher Full Text](#)
42. Kondofersky I, Laimighofer M, Kurz C, *et al.*: **Three general concepts to improve risk prediction: good data, wisdom of the crowd, recalibration [version 1; referees: 2 approved with reservations].** *F1000Res.* 2016; **5**: 2671.
[Publisher Full Text](#)
43. Wiens J, Gutttag J, Horvitz E: **A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions.** *J Am Med Inform Assoc.* 2014; **21**(4): 699–706.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
44. Gong JJ, Sundt TM, Rawn JD, *et al.*: **Instance weighting for patient-specific risk stratification models.** In *Proc. of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, USA, ACM Press. 2015; 369–378.
[Publisher Full Text](#)
45. Nori N, Kashima H, Yamashita K, *et al.*: **Learning implicit tasks for patient-specific risk modeling in ICU.** In *Proc. of the 31st AAAI Conference on Artificial Intelligence.* 2017; 1481–1487.
[Reference Source](#)

Open Peer Review

Current Referee Status:




Version 3

Referee Report 25 July 2017

doi:[10.5256/f1000research.13099.r24079](https://doi.org/10.5256/f1000research.13099.r24079)



Jinfeng Xiao 

Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, IL, USA

The authors have made a great revision and addressed all my concerns in my previous comments. The revised version is scientifically solid, and provides readers with a lot of insights and inspirations about the topic. Good job!

Competing Interests: No competing interests were disclosed.

Referee Expertise: Machine learning, data mining, bioinformatics

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 2

Referee Report 28 June 2017

doi:[10.5256/f1000research.13006.r23825](https://doi.org/10.5256/f1000research.13006.r23825)



Donna P. Ankerst

Department of Mathematics, Technical University of Munich, Garching, Germany

The authors are to be congratulated for revising the article and thereby improving its important contribution to the science of prediction modeling. No further comments.

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.

Version 1

Referee Report 14 March 2017

doi:10.5256/f1000research.8853.r19237



Amber L Simpson

Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA

The authors describe an ensemble approach for predicting survival in prostate cancer patients as part of the 2015 Prostate Cancer DREAM (Dialogue for Reverse Engineering Assessments and Methods) Challenge. Patients included in the study had metastatic, castrate-resistant prostate cancer, an advanced cancer with poor prognosis. The authors are commended for undertaking a difficult problem and providing an elegant solution incorporating Cox, gradient boosting, random survival forest, and SVM for time-dependent analysis.

Addressing these points would improve the manuscript:

- 1) The luxury of a challenge is that authors are positioned to use knowledge gained from the challenge to improve their prediction model. The intent of sharing these datasets is to develop the best biomarker that can be used to change patient selection for therapy. Can the authors comment on what they would do differently now that they have considered methods proposed by other groups in the challenge? How can others use the lessons learned in this challenge to make the best biomarker possible?
- 2) The authors should also comment on the generalizability of their methods to other problems.
- 3) The paper is a good technical companion paper to the overview paper that was recently released, which should be cited ¹.
- 4) For those unfamiliar with the challenge, it is important to note that the challenge organizers confirmed performance on the validation data as noted by a reviewer above. This information should be incorporated into the manuscript, as it is not readily apparent.
- 5) How does the model perform relative to published clinical nomograms? For example, the Armstrong nomogram achieved a concordance index of 0.69. Can the authors comment on the improvement over existing methods? One could argue that the slight improvement is not worth the overhead of employing ensemble methods ²⁻⁴.
- 6) How does predicting survival change the management of these patients? For example, would bad actors be selected for a different treatment or spared from treatment? If so, it may be appropriate to calculate positive and negative predictive value for specific time points. Maximizing positive and negative predictive value may also make sense. The proposed method could aid in chemoprevention, as an example.
- 7) Is it possible to make the model publicly available as a nomogram (see nomograms.org)? Clinicians will not have the ability to download and install the code, but they may be interested in the results for individual patients.
- 8) How does the ensemble method compensate for highly correlated variables?
- 9) How was feature selection performed?

10) Listing the features would be helpful for clinicians looking to refine/improve existing nomograms.

References

1. Guinney J, Wang T, Laajala T, Winner K, Bare J, Neto E, Khan S, Peddinti G, Airola A, Pahikkala T, Mirtti T, Yu T, Bot B, Shen L, Abdallah K, Norman T, Friend S, Stolovitzky G, Soule H, Sweeney C, Ryan C, Scher H, Sartor O, Xie Y, Aittokallio T, Zhou F, Costello J: 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. *The Lancet Oncology*. 2017; **18** (1): 132-142 [Publisher Full Text](#)
2. Armstrong AJ, Garrett-Mayer ES, Yang YC, de Wit R, Tannock IF, Eisenberger M: A contemporary prognostic nomogram for men with hormone-refractory metastatic prostate cancer: a TAX327 study analysis. *Clin Cancer Res*. 2007; **13** (21): 6396-403 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Armstrong AJ, Garrett-Mayer E, de Wit R, Tannock I, Eisenberger M: Prediction of survival following first-line chemotherapy in men with castration-resistant metastatic prostate cancer. *Clin Cancer Res*. 2010; **16** (1): 203-11 [PubMed Abstract](#) | [Publisher Full Text](#)
4. Halabi S, Lin CY, Kelly WK, Fizazi KS, Moul JW, Kaplan EB, Morris MJ, Small EJ: 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](#)

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.

Author Response 22 Jun 2017

Sebastian Pölsterl, Technische Universität München, Germany

The authors describe an ensemble approach for predicting survival in prostate cancer patients as part of the 2015 Prostate Cancer DREAM (Dialogue for Reverse Engineering Assessments and Methods) Challenge. Patients included in the study had metastatic, castrate-resistant prostate cancer, an advanced cancer with poor prognosis. The authors are commended for undertaking a difficult problem and providing an elegant solution incorporating Cox, gradient boosting, random survival forest, and SVM for time-dependent analysis.

Addressing these points would improve the manuscript:

1. The luxury of a challenge is that authors are positioned to use knowledge gained from the challenge to improve their prediction model. The intent of sharing these datasets is to develop the best biomarker that can be used to change patient selection for therapy. Can the authors comment on what they would do differently now that they have considered methods proposed by other groups in the challenge? How can others use the lessons learned in this challenge to make the best biomarker possible?

1. **Response:** *Several teams, including the winning team (FIMM-UTU), implemented methods to carefully select patients from the three studies constituting the training data such that they are not too different from the target study, which was used for the final evaluation. We believe that a considerable improvement can be gained by discarding outliers from the training data.*

2. The authors should also comment on the generalizability of their methods to other problems.
 1. **Response:** *Our proposed solution relies on heterogeneous ensembles, which are comprised of survival models to predict the risk of death. Hence, our approach is directly applicable to any data with right censored survival times. For other problems, such as classification or regression, the ensemble selection and ensemble pruning need to be adapted by choosing an appropriate performance measure (see line 11 of algorithm 1 and line 14 of algorithm 2). In fact, the original authors of heterogeneous ensembles investigated ten performance metrics for classification and Rooney et al. proposed to using the mean squared error for regression problems. Therefore, heterogeneous ensembles are applicable to a wide range of learning problems.*
3. The paper is a good technical companion paper to the overview paper that was recently released, which should be cited ¹.
 1. **Response:** *We updated reference 16 to refer to the paper in The Lancet Oncology.*
4. For those unfamiliar with the challenge, it is important to note that the challenge organizers confirmed performance on the validation data as noted by a reviewer above. This information should be incorporated into the manuscript, as it is not readily apparent.
 1. **Response:** *We updated the last paragraph of the “Validation scheme” section and the first paragraph of the “Challenge hold-out data” to emphasize that validation was carried out by the challenge organizers.*
5. How does the model perform relative to published clinical nomograms? For example, the Armstrong nomogram achieved a concordance index of 0.69. Can the authors comment on the improvement over existing methods? One could argue that the slight improvement is not worth the overhead of employing ensemble methods²⁻⁴.
 1. **Response:** *In subchallenge 1a, submissions of all participating teams were compared to the model by Halabi et al., which was considered the state-of-the-art risk prediction model prior to the challenge. Only submissions with a statistically better performance than the model by Halabi et al. were considered for the final evaluation (see section Validation scheme in our manuscript for further details). Our proposed model achieved an iAUC score of 0.7646 on the challenge’s hold data, whereas the model by Halabi et al. achieved a score of 0.7432, which is significantly worse: the Bayes factor of the proposed model vs. Halabi et al. model, is 12.2, which indicates strong evidence.*
6. How does predicting survival change the management of these patients? For example, would bad actors be selected for a different treatment or spared from treatment? If so, it may be appropriate to calculate positive and negative predictive value for specific time points. Maximizing positive and negative predictive value may also make sense. The proposed method could aid in chemoprevention, as an example.
 1. **Response:** *We agree that ultimately the focus should be on improving patient treatment, but at the same time computational methods can only hint at potentially interesting biomarkers or patient subgroups, whether this information is useful in the clinic requires additional research, e.g., to rule-out harmful side-effects. Data in the Prostate Cancer DREAM Challenge are collated based on comparator arm data sets of Phase III prostate cancer clinical trials, where all patients received docetaxel and prednisone in the comparator arm. Therefore, we could not determine whether differences in survival can be attributed to different treatment types. If outcome information from multiple treatments were available, it would indeed be very*

interesting to infer the optimal treatment by maximizing positive and negative predictive value over time instead of specificity and sensitivity as the iAUC metric used in the challenge does.

7. Is it possible to make the model publicly available as a nomogram (see nomograms.org)? Clinicians will not have the ability to download and install the code, but they may be interested in the results for individual patients.
 1. **Response:** *Unfortunately, it is often difficult to understand how an ensemble method relates the input to variables to each other in order to form a prediction, which is especially true for heterogeneous ensembles, because of their non-linear nature. A nomogram describes a non-linear model only inadequately, because it gives each variable only a single weight and usually lacks high-order interactions. However, there are several alternative ways to obtain insight. For instance, Breiman (Machine Learning, 45:1, 2001. <http://dx.doi.org/10.1023/A:1010933404324>) suggested a variable importance measure for random forests that could be adapted. The j-th feature is randomly permuted for all out-of-bag samples and run down the corresponding tree. The output is the relative increase in prediction error as compared to if the j-th feature is intact. Feature with a larger increase in prediction error, are considered more important to the ensemble. If one wants to infer which interactions among features the ensemble considers, more sophisticated methods are available (e.g. Henelius et al., SLDS 2015. http://dx.doi.org/10.1007/978-3-319-17091-6_5).*
8. How does the ensemble method compensate for highly correlated variables?
 1. **Response:** *Whether the ensemble compensates for highly correlated variables depends on choice of base learners. Here, all base learners account for multicollinearities. The penalized Cox model and survival support vector machine use a ridge (L2) penalty, gradient boosting with regression trees and random survival forests recursively split the data based on a single feature, and gradient boosting with componentwise least squares selects only one feature in each iteration such that the error is maximally reduced. Hence, all models can be trained despite highly correlated variables in the data.*
9. How was feature selection performed?
 1. **Response:** *We did not perform feature selection prior to constructing the heterogeneous ensemble, however, the ensemble comprised base learners that implicitly perform feature selection when trained on high-dimensional data, namely random survival forest and gradient boosting models. The remaining models (penalized Cox model and survival support vector machine) do not perform feature selection and only account for multicollinearities.*
10. Listing the features would be helpful for clinicians looking to refine/improve existing nomograms.
 1. **Response:** *We trained models on different subsets of the data, ranging from 383 features for data from the MAINSAIL trial to 217 features when combining data of all three trials (see table 1). More details on the extracted features are available from the supplementary material and at <https://www.synapse.org/#!Synapse:syn4650470>.*

Competing Interests: No competing interests were disclosed.

doi:10.5256/f1000research.8853.r20214



Jinfeng Xiao

Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, IL, USA

This paper is written by CAMP, a winning team of the 2015 Prostate Cancer DREAM Challenge (“the PCDC”, or “the challenge”), to introduce their winning method. The authors built heterogeneous ensembles with the training data (Trials ASCENT-2, MAINSAIL, VENICE) and the unlabeled part of the validation data (Trial ENTHUSE-33). The high performance of their method, especially in predicting patients’ days to death, was confirmed by the challenge organizers on the validation data. This manuscript contains sufficient details about the actual method they used in the PCDC. Achilles heels for this paper include: 1) To show the necessity of using ensembles; 2) To establish the generalizability of the proposed ensemble models to new data sets. See the major issues below for more detailed comments.

Major issues:

1. The power of averaging over the base learners was taken for granted in the paper without experimental evidence. Training an ensemble costs much more effort than training a single model, and therefore it has to be shown that such effort is worth it. Direct comparison in performance between the ensemble and base learners is needed to make this point clear.
2. The training of the ensembles, in particular the ensemble pruning step, used information from the validation set. Although only the features, but not the outcomes, of the validation data were seen by the model, this practice is still not encouraged. A generalizable model should not use the validation data in any way during training. Therefore, whether the proposed method is generalizable to new data sets is in doubt. I would suggest the authors to prune the ensemble on the training set and check the performance on the validation set.
3. There is no instruction in the code documentation about how to apply the code to new data sets. Adding such information can greatly increase the chance that the code will be used by other researchers.
4. Can the authors mine some knowledge from the trained model? For example, what are the most important features? Where are the baseline (i.e. Halabi’s model¹) features in the ranked list? Such analysis of the model can be helpful to biomedical researchers and doctors.

Minor issues:

1. In Algorithms 1 & 2, how did the authors choose the minimum desired performance c_{\min} and the desired set of ensemble S ?
2. Page 6, paragraph 2, line 3: “Median” should be changed to “standard deviation” or some other measures of variance, because in a within-trial validation the “median” is not directly related to “the difference between observed time points in the training and test data” (lines 5-6).
3. Page 8, paragraph 2, the last 8 lines: This example is not very convincing. A model considering all features trained on the first dataset will assign a very small (if not zero) weight to feature 3, which will compensate little for the fact that feature 3 is important in the second dataset.
4. Page 8, paragraph 5: What numerical difficulties did the authors encounter so that they could not include the Cox regression in the ensembles? Is there anything special about Cox model that makes it harder to train than other base learners?
5. It is not explicitly stated in the paper that the authors are from Team CAMP.

Grammar:

Page 4, last paragraph: “within-in trial validation” should be “within-trial validation”; “between trials

validation” should be “between-trial validation”.

References

1. Halabi S, Lin CY, Kelly WK, Fizazi KS, Moul JW, Kaplan EB, Morris MJ, Small EJ: 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](#)

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.

Author Response 22 Jun 2017

Sebastian Pölsterl, Technische Universität München, Germany

This paper is written by CAMP, a winning team of the 2015 Prostate Cancer DREAM Challenge (“the PCDC”, or “the challenge”), to introduce their winning method. The authors built heterogeneous ensembles with the training data (Trials ASCENT-2, MAINSAIL, VENICE) and the unlabeled part of the validation data (Trial ENTHUSE-33). The high performance of their method, especially in predicting patients’ days to death, was confirmed by the challenge organizers on the validation data. This manuscript contains sufficient details about the actual method they used in the PCDC. Achilles heels for this paper include: 1) To show the necessity of using ensembles; 2) To establish the generalizability of the proposed ensemble models to new data sets. See the major issues below for more detailed comments.

Major issues:

1. The power of averaging over the base learners was taken for granted in the paper without experimental evidence. Training an ensemble costs much more effort than training a single model, and therefore it has to be shown that such effort is worth it. Direct comparison in performance between the ensemble and base learners is needed to make this point clear.
 1. **Response:** *We included heterogeneous ensembles in the between trials validation (see figures 3 and 5) and in our discussion of the results.*
2. The training of the ensembles, in particular the ensemble pruning step, used information from the validation set. Although only the features, but not the outcomes, of the validation data were seen by the model, this practice is still not encouraged. A generalizable model should not use the validation data in any way during training. Therefore, whether the proposed method is generalizable to new data sets is in doubt. I would suggest the authors to prune the ensemble on the training set and check the performance on the validation set.
 1. **Response:** *For survival data, the pruning step is delayed until prediction is performed, because predictions are risk scores on an arbitrary scale for which a per-sample error measure is not readily available. This is in contrast to ensemble pruning for regression problems, where a per-sample error can be easily computed and models having highly correlated errors are pruned. As referee 2 suggests, the pruning step could be performed via cross-validation on the training data. However, our pruning step does not take survival times or censoring status into account, therefore we prefer to delay the pruning step as long as possible as to avoid overfitting on the training data. If the additional costs associated with storing the*

ensemble before pruning are prohibitive, we recommend that pruning should be performed via cross-validation on the training data.

3. There is no instruction in the code documentation about how to apply the code to new data sets. Adding such information can greatly increase the chance that the code will be used by other researchers.
 1. **Response:** *Our source code is accompanied by a README file explaining all steps necessary to reproduce our results. The source code can be downloaded from Synapse (<http://dx.doi.org/10.7303/syn3647478>) as well as GitHub (<https://github.com/tum-camp/dream-prostate-cancer-challenge>).*
4. Can the authors mine some knowledge from the trained model? For example, what are the most important features? Where are the baseline (i.e. Halabi's model¹) features in the ranked list? Such analysis of the model can be helpful to biomedical researchers and doctors.
 1. **Response:** *Please see our response to question 8 of referee 1.*

Minor issues:

1. In Algorithms 1 & 2, how did the authors choose the minimum desired performance c_{min} and the desired set of ensemble S ?
 1. **Response:** *We chose $c_{min} = 0.66$ based on results of the with-in trial validation (figure 2): approximately 30% of the experiments performed worse. The final ensemble consists of all base learners in the top 5% according the combined accuracy and diversity score (see table 2 and algorithm 2). Both c_{min} and S remained fixed throughout our experiments and were not optimised.*
2. Page 6, paragraph 2, line 3: "Median" should be changed to "standard deviation" or some other measures of variance, because in a within-trial validation the "median" is not directly related to "the difference between observed time points in the training and test data" (lines 5-6).
 1. **Response:** *Thank you for the suggestion, we replaced median by standard deviation in the manuscript.*
3. Page 8, paragraph 2, the last 8 lines: This example is not very convincing. A model considering all features trained on the first dataset will assign a very small (if not zero) weight to feature 3, which will compensate little for the fact that feature 3 is important in the second dataset.
 1. **Response:** *We agree that the example was inadequate to explain this observation. We replaced it by referencing the work by Meinshausen and Bühlmann, who showed that models with embedded feature selection suffer from false positive selections in high dimensions.*
4. Page 8, paragraph 5: What numerical difficulties did the authors encounter so that they could not include the Cox regression in the ensembles? Is there anything special about Cox model that makes it harder to train than other base learners?
 1. **Response:** *Please see our response to question 7 of referee 1.*
5. It is not explicitly stated in the paper that the authors are from Team CAMP.
 1. **Response:** *We mentioned that we participated under the name "Team CAMP" at the end of the introduction and in the section "Challenge hold-out data".*

Grammar:

Page 4, last paragraph: "within-in trial validation" should be "within-trial validation"; "between trials validation" should be "between-trial validation".

Response: *Thank you, we corrected these errors in the manuscript.*

Competing Interests: No competing interests were disclosed.

Referee Report 19 December 2016

doi:10.5256/f1000research.8853.r18609



Donna P. Ankerst

Department of Mathematics, Technical University of Munich, Garching, Germany

The authors are to be congratulated for landing among the circle of winners of the Prostate Cancer DREAM Challenge and for clearly describing their innovative methods in this paper. An informative discussion critically appraises their approach, providing suggestions for advancing the field of clinical risk prediction. Instead of relying on one survival model, their approach hinges on heterogeneous ensembles that invoke a variety of model types, including gradient boosting (least squares versus trees), random survival forests, and survival support vector machines (linear versus clinical kernels), thereby hedging against sub-optimality of any single model for any single test set. I have only minor comments.

1. It is argued throughout that heterogeneous ensembles have been shown to be optimal compared to single models for this challenge, but I did not see a head-to-head comparison illustrating this. For example, could one not add ensemble methods as an extra column to the within- and between-trial validations in Figures 2 and 3, respectively?
2. I greatly appreciated Figure 4 that showed which of the multiple comparisons in Figure 3 (the between-trial validation) were actually critically different, as many of the iAUCs only differed out to the second decimal (which is by the way a clinically meaningless difference). It would be nice to also have such a comparison for Figure 2 (the within-trial validation) that could definitely show whether or not the Cox model was statistically indistinguishable from random forests, and to temper the Results section concerning the comparison of the methods. One method only beats another if the confidence intervals of the respective AUCs do not overlap. Given their similar performance, the comparison among the different individual survival models might not be as relevant as whether or not the ensemble outperformed any one of them.
3. As nicely pointed out in the Discussion, it is a surprise and a great pity that the concordance statistic c was used for the training of the models instead of the iAUC, the criterion used for evaluation for the challenge. While easy to compute, the concordance statistic suffers greatly from censored observations, they essentially are discarded in the evaluation. This means that only a minority of the data in the ASCENT and MAINSAIL trials were used (71% and 82.5% of the data censored). The iAUC, however, also suffers from censored data, but from what I understand, to a lesser extent. Is it possible to redo Figures 2 and 3 using the iAUC instead of the concordance statistic, to see if similar conclusions held?
4. In the discussion of the within-trial internal cross-validation of Figure 2 it is mentioned that some of the methods may have performed poorly because of a difference in follow-up between the random partitions of the trial into training and test sets. In medical studies, this is often controlled using stratified randomization, which ensures the proportion of observed events (deaths in this case) or follow-up remains equal across the sets. Would it be possible to implement to see if it improved the outcome for VENICE, in order to help explain the poor behavior there? It of course, does not help the between-trial validation, the subject of the next point.

5. The problem of recalibration to different trials is becoming more and more recognized in medicine; searching for “recalibration risk score” or “recalibration risk model” in PubMed reveals hundreds of suggestions and applications. The authors do a nice job of illustrating the particular difficulties with survival data – a look at Figure 1 shows that median follow-up in the held out ENTHUSE-33 trial was longer than two of the trials used for training. In our analysis for the challenge we showed that recalibration made a big difference for the root-mean-squared-error in Subchallenge 1b but not the iAUC in Subchallenge 1a, matching previous results we have obtained in proposals to dynamically update risk models (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4532612/>). Recalibration means any method to tweak an existing risk model to conform to a particular target population, but has the problem that it requires data from the intended target population, something not generally possible for clinical practice. I agree with the authors that this could have improved their models and would like to see more discussion of the literature from recalibration of survival models.
6. In the Discussion of the between-trials validation, the authors try to explain the surprising result that the simpler Cox model with its stringent proportional hazards and linear assumption performs as well as some of the other models that incorporate non-linearity. I think lack of statistical power, i.e., small sample size, may be another culprit here. The effective information size for survival data (defined as the size of the information matrix) is only proportional to the number of observed events and not the total sample size, this is an issue that clinical trial statisticians who design trials understand well, but unfortunately not the rest of the community. It was a point I tried to raise at the first Challenge webinar, foreseeing that there would be many ties among winners due to the high censoring. While for training it seemed like there were trials of size 476, 526 and 598 patients for the respective trials in Figure 1, with a total of 1600 patients, the effective information content was only 138, 92, and 432, respectively, for a total of 662 patients. Simulation studies would reveal what sample size would be needed to detect nonlinearities of different magnitudes. My point is not to suggest doing these, but rather to modify the discussion that the high-performance of Cox’s simpler model may be due to the Occam’s Razor principle, that if there exists two explanations for data, the simpler is preferred.
7. In light of Point 6.), it is a pity that the well-performing Cox’s proportional hazards model was eventually dropped because of numerical problems. Our team used this model without much difficulty. Can the authors elaborate here or propose suggestions for overcoming the numerical difficulties? For example, could it be that the input data contained a lot of features with anomalies that should have been cleaned out?
8. I realize it was not the point of this paper, but it is a pity that there is no discussion of the specifics of the 90 features that ultimately made it into the prediction models. Were they the same as the ones found by Halabi *et al.*? 90 features are a lot and not generally implementable in online risk tools designed to help patients – would there be a way to summarize the features that are most important in order to help clinicians understand the important indicators?
9. Looking back at the Halabi paper, which has a simple Cox model with a handful of predictors that is immediately interpretable, the AUC obtained there on the test set (0.76) seems close to those obtained in this challenge. The AUC is a rank-based discrimination measure, that reflects the probability that for a randomly selected pair of patients, the patient that died later had a lower risk score and differences have to be interpreted relative to this meaning. I would like to hear the authors’ reflection as to whether the DREAM Challenge has proven the case for the large-scale

methods used in the Challenge or against them. What future directions are needed to improve prediction? Some, like myself, would argue that new markers need to be discovered rather than bigger models.

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.

Author Response 22 Jun 2017

Sebastian Pölsterl, Technische Universität München, Germany

The authors are to be congratulated for landing among the circle of winners of the Prostate Cancer DREAM Challenge and for clearly describing their innovative methods in this paper. An informative discussion critically appraises their approach, providing suggestions for advancing the field of clinical risk prediction. Instead of relying on one survival model, their approach hinges on heterogeneous ensembles that invoke a variety of model types, including gradient boosting (least squares versus trees), random survival forests, and survival support vector machines (linear versus clinical kernels), thereby hedging against sub-optimality of any single model for any single test set. I have only minor comments.

1. It is argued throughout that heterogeneous ensembles have been shown to be optimal compared to single models for this challenge, but I did not see a head-to-head comparison illustrating this. For example, could one not add ensemble methods as an extra column to the within- and between-trial validations in Figures 2 and 3, respectively?
 1. **Response:** *We included heterogeneous ensembles in the between trials validation (see figures 3 and 5) and in our discussion of the results.*
2. I greatly appreciated Figure 4 that showed which of the multiple comparisons in Figure 3 (the between-trial validation) were actually critically different, as many of the iAUCs only differed out to the second decimal (which is by the way a clinically meaningless difference). It would be nice to also have such a comparison for Figure 2 (the within-trial validation) that could definitely show whether or not the Cox model was statistically indistinguishable from random forests, and to temper the Results section concerning the comparison of the methods. One method only beats another if the confidence intervals of the respective AUCs do not overlap. Given their similar performance, the comparison among the different individual survival models might not be as relevant as whether or not the ensemble outperformed any one of them.
 1. **Response:** *As suggested, we added a plot for the results presented in figure 2. It shows that the Cox model and random survival forest only significantly outperform linear SVM, whereas the performance of the other methods lies within the critical difference interval.*
3. As nicely pointed out in the Discussion, it is a surprise and a great pity that the concordance statistic c was used for the training of the models instead of the iAUC, the criterion used for evaluation for the challenge. While easy to compute, the concordance statistic suffers greatly from censored observations, they essentially are discarded in the evaluation. This means that only a minority of the data in the ASCENT and MAINSAIL trials were used (71% and 82.5% of the data censored). The iAUC, however, also suffers from censored data, but from what I understand, to a lesser extent. Is it possible to redo Figures 2 and 3 using the iAUC instead of the concordance statistic, to see if similar conclusions held?

1. **Response:** *We did perform the same analyses as depicted in figures 2 and 3 using iAUC as evaluation criteria. When ranking methods according to average iAUC, one arrives at the same conclusion as when ranking according to average c-index. However, the average performance with respect to the test datasets are quite different. As we pointed out in the main text, this is due to the definition of the iAUC used in the Prostate Cancer Dream Challenge, which is the integral over time points every 6 months up to 30 months after the first day of treatment. This would cover most time points in ASCENT-2 and MAINSAIL, but would miss out many events occurring after 30 months for VENICE (cf. figure 1). Consequently, it appears that all methods perform considerably better when tested on the VENICE data. Usually, it is recommended to choose the limits of the interval to integrate over from the data, e.g. the 5% to 90% percentile of observed time points. However, the iAUC would be based on a different interval for each study, making inter-study comparisons difficult to interpret. Therefore, we believe that the c-index is easier to interpret when considering the inter-study comparison. In addition, we re-trained our heterogeneous ensemble using the iAUC metric in algorithm 2 and submitted its prediction after the challenge concluded. We obtained an iAUC of 0.7636 compared to 0.7644 when using c-index, and 0.7537 for the Halabi model.*
4. In the discussion of the within-trial internal cross-validation of Figure 2 it is mentioned that some of the methods may have performed poorly because of a difference in follow-up between the random partitions of the trial into training and test sets. In medical studies, this is often controlled using stratified randomization, which ensures the proportion of observed events (deaths in this case) or follow-up remains equal across the sets. Would it be possible to implement to see if it improved the outcome for VENICE, in order to help explain the poor behavior there? It of course, does not help the between-trial validation, the subject of the next point.
 1. **Response:** *We implemented the suggested modification to perform stratified cross-validation and repeated the experiment. The results are very similar to figure 2: the average performance of all methods is still worst when trained and tested on data from the VENICE study.*
5. The problem of recalibration to different trials is becoming more and more recognized in medicine; searching for “recalibration risk score” or “recalibration risk model” in PubMed reveals hundreds of suggestions and applications. The authors do a nice job of illustrating the particular difficulties with survival data – a look at Figure 1 shows that median follow-up in the held out ENTHUSE-33 trial was longer than two of the trials used for training. In our analysis for the challenge we showed that recalibration made a big difference for the root-mean-squared-error in Subchallenge 1b but not the iAUC in Subchallenge 1a, matching previous results we have obtained in proposals to dynamically update risk models (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4532612/>). Recalibration means any method to tweak an existing risk model to conform to a particular target population, but has the problem that it requires data from the intended target population, something not generally possible for clinical practice. I agree with the authors that this could have improved their models and would like to see more discussion of the literature from recalibration of survival models.
 1. **Response:** *We agree with the referee that calibration is an important aspect besides discrimination that should be considered for prognostic models. The focus of the referenced article is on calibration of binary classification models, which is very different from calibration with respect to survival models. In contrast to binary classification, predicted risk scores of a survival model are only the relative risk, but*

not absolute risks (Royston and Altman. BMC Medical Research Methodology, 13(1), 2013. <http://doi.org/10.1186/1471-2288-13-33>). Although, a measure of absolute risk can be derived for the Cox model by estimating the baseline hazard function, to the best of our knowledge, there is no standard approach to statistically assess calibration for arbitrary survival models. We could only visually assess calibration by constructing low, medium, and high risk groups corresponding to cut-offs at the 33% and 66% percentile of predicted risk scores. Using the same cut-offs on predicted risk scores from hold-out dataset, we constructed two Kaplan-Meier plots, each stratified by risk group. For a well calibrated model, we would expect the Kaplan-Meier curves derived from the training and hold-out data to agree with each other. The disadvantage of this approach is that we cannot precisely quantify the lack of calibration.

6. In the Discussion of the between-trials validation, the authors try to explain the surprising result that the simpler Cox model with its stringent proportional hazards and linear assumption performs as well as some of the other models that incorporate non-linearity. I think lack of statistical power, i.e., small sample size, may be another culprit here. The effective information size for survival data (defined as the size of the information matrix) is only proportional to the number of observed events and not the total sample size, this is an issue that clinical trial statisticians who design trials understand well, but unfortunately not the rest of the community. It was a point I tried to raise at the first Challenge webinar, foreseeing that there would be many ties among winners due to the high censoring. While for training it seemed like there were trials of size 476, 526 and 598 patients for the respective trials in Figure 1, with a total of 1600 patients, the effective information content was only 138, 92, and 432, respectively, for a total of 662 patients. Simulation studies would reveal what sample size would be needed to detect nonlinearities of different magnitudes. My point is not to suggest doing these, but rather to modify the discussion that the high-performance of Cox's simpler model may be due to the Occam's Razor principle, that if there exists two explanations for data, the simpler is preferred.
 1. **Response:** *Thank you for pointing out that effective sample size could be limiting the performance of more complicated models. We added a paragraph discussing this issue to the "Between trials validation" section.*
7. In light of Point 6.), it is a pity that the well-performing Cox's proportional hazards model was eventually dropped because of numerical problems. Our team used this model without much difficulty. Can the authors elaborate here or propose suggestions for overcoming the numerical difficulties? For example, could it be that the input data contained a lot of features with anomalies that should have been cleaned out?
 1. **Response:** *We fit Cox's proportional hazards model using a Newton-Rhapson algorithm with constant step size, i.e., without a line search. We observed that optimization sometimes diverged, which can be attributed to choosing a constant step size. If the chosen step size is too large, it can lead to oscillation around the minimum due to overshooting and the minimum is never reached. A crude solution would be to increase the tolerance that determines convergence or choose a different starting point. A better solution would be to determine the optimal step size in each iteration of Newton's method via line search or employing a trust region method (see e.g. Boyd and Vandenberghe, Convex Optimization, Cambridge University Press, 2009). Unfortunately, we were unable to implement this modification before the challenge was closed.*
8. I realize it was not the point of this paper, but it is a pity that there is no discussion of the specifics of the 90 features that ultimately made it into the prediction models. Were they the

same as the ones found by Halabi *et al.*? 90 features are a lot and not generally implementable in online risk tools designed to help patients – would there be a way to summarize the features that are most important in order to help clinicians understand the important indicators?

1. **Response:** *Our final ensemble considered 217 features (see table 1), which included those found by Halabi et al., and was comprised of 90 different base learners (see table 3). We agree that the raw predictive performance of a model often provides insufficient information in medical research. Clearly, there is a trade-off between model complexity and how well a model can be interpreted. As we mentioned in the introduction of our manuscript, an ensemble approach is only beneficial if base learners perform better than random prediction and are diverse. The latter can be achieved by using base learners that are based on different loss functions (heterogeneous ensemble) or by using the same loss for all base learners and forcing each base learner to consider different subsets of features (homogeneous ensemble). We chose to utilize both types by selecting base learners from a large library of different models and identical, but differently parametrized, models. Consequently, we encourage base learners to weight features differently, which makes creating a universal ranking of features challenging. Although feature importances are not directly available from our final ensemble, there are several alternative ways to obtain insight. For instance, Breiman (Machine Learning, 45:1, 2001. <http://dx.doi.org/10.1023/A:1010933404324>) suggested a variable importance measure for random forests that could be adapted. The j -th feature is randomly permuted for all out-of-bag samples and run down the corresponding tree. The output is the relative increase in prediction error as compared to if the j -th feature is intact. Feature with a larger increase in prediction error, are considered more important to the ensemble. If one wants to infer which interactions among features the ensemble considers, more sophisticated methods are available (e.g. Henelius et al., SLDS 2015. http://dx.doi.org/10.1007/978-3-319-17091-6_5).*
9. Looking back at the Halabi paper, which has a simple Cox model with a handful of predictors that is immediately interpretable, the AUC obtained there on the test set (0.76) seems close to those obtained in this challenge. The AUC is a rank-based discrimination measure, that reflects the probability that for a randomly selected pair of patients, the patient that died later had a lower risk score and differences have to be interpreted relative to this meaning. I would like to hear the authors' reflection as to whether the DREAM Challenge has proven the case for the large-scale methods used in the Challenge or against them. What future directions are needed to improve prediction? Some, like myself, would argue that new markers need to be discovered rather than bigger models.
1. **Response:** *The Prostate Cancer DREAM challenge provided high-quality data to a large group of researchers, which led to improved prediction performance compared to the model by Halabi et al. and highlighted interesting problems for future research. In particular, we believe that future research should focus on how to best utilize data from multiple clinical trials and how to adapt a model to new patient cohorts. As described in our manuscript, the Prostate Cancer DREAM challenge compiled data from four clinical trials, with each trial having its own characteristics, ranging from different follow-up periods to different clinical information collected. In light of these differences, just combining all the data and learning a model on top of it is likely to lead to a poor model, despite an increase in sample size. Multiple teams identified this problem and tried to address it. Most importantly, the winning team (FIMM-UTU) selected a subset of the provided patient data as to obtain a coherent patient sample*

for training their model. By identifying and omitting patients that appear considerably different from the remaining patients, they successfully lessened the effect of study-specific batch-effects. Another interesting approach has been proposed by team A Bavarian dream (as pointed out by the reviewer above). They used recalibration methods to adapt their model to the target study, which was used for final evaluation. In conclusion, we believe that the biggest improvements in risk prediction were not due to identifying new risk markers, but by choosing methods that account for sub-structures in the data. More research is needed to reliably detect such sub-structures and to overcome the problems they attend.

Competing Interests: No competing interests were disclosed.
