METHOD ARTICLE

DWCox: A density-weighted Cox model for outlier-robust prediction of prostate cancer survival [version 1; referees: 1 approved, 2 approved with reservations]

Jinfeng Xiao, Sheng Wang, Jingbo Shang, Henry Lin, Doris Xin, Xiang Ren, Jiawei Han, Jian Peng

Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana-Champaign, USA

v1 **First published:** 01 Dec 2016, **5**:2806 (doi: 10.12688/f1000research.9434.1)
Latest published: 01 Dec 2016, **5**:2806 (doi: 10.12688/f1000research.9434.1)

Abstract

Reliable predictions on the risk and survival time of prostate cancer patients based on their clinical records can help guide their treatment and provide hints about the disease mechanism. The Cox regression is currently a commonly accepted approach for such tasks in clinical applications. More complex methods, like ensemble approaches, have the potential of reaching better prediction accuracy at the cost of increased training difficulty and worse result interpretability. Better performance on a specific data set may also be obtained by extensive manual exploration in the data space, but such developed models are subject to overfitting and usually not directly applicable to a different data set. We propose DWCox, a density-weighted Cox model that has improved robustness against outliers and thus can provide more accurate predictions of prostate cancer survival. DWCox assigns weights to the training data according to their local kernel density in the feature space, and incorporates those weights into the partial likelihood function. A linear regression is then used to predict the actual survival times from the predicted risks. In the 2015 Prostate Cancer DREAM Challenge, DWCox obtained the best average ranking in prediction accuracy on the risk and survival time. The success of DWCox is remarkable given that it is one of the smallest and most interpretable models submitted to the challenge. In simulations, DWCox performed consistently better than a standard Cox model when the training data contained many sparsely distributed outliers. Although developed for prostate cancer patients, DWCox can be easily re-trained and applied to other survival analysis problems. DWCox is implemented in R and can be downloaded from https://github.com/JinfengXiao/DWCox.

This article is included in the DREAM Challenges channel.

Open Peer Review

Referee Status: ✓ ? ?

	Invited Referees		
	1	**2**	**3**
version 1 published 01 Dec 2016	✓ report	? report	? report

1 **Motoki Shiga**, Gifu University Japan

2 **Sebastian Pölsterl**, Insitute for Cancer Research UK

3 **Riku Klen**, Department of Mathematics and Statistics Finland, **Mehrad Mahmoudian**, University of Turku Finland

Discuss this article

Comments (0)

Corresponding authors: Jinfeng Xiao (jxiao13@illinois.edu), Jian Peng (jianpeng@illinois.edu)

How to cite this article: Xiao J, Wang S, Shang J *et al.* **DWCox: A density-weighted Cox model for outlier-robust prediction of prostate cancer survival [version 1; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2016, **5**:2806 (doi: 10.12688/f1000research.9434.1)

Grant information: Research was sponsored in part by grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), National Science Foundation IIS-1320617, IIS-1354329 and IIS 16-18481, and the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA). The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the opinions of the funding agencies.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 01 Dec 2016, **5**:2806 (doi: 10.12688/f1000research.9434.1)

Introduction

Prostate cancer is the 2nd leading cause of cancer death in men in the United States[1] and the 6th worldwide[2]. In the past 10 years more than 2 million men in the US suffered from prostate cancer, and about 5% of those patients had metastatic castrate-resistant prostate cancer (mCRPC), an advanced form of the disease whose outcomes are poor and treatment remains unclear. Survival analysis based on clinical records has attracted researchers' attention, since it can hopefully direct cancer treatment and help elucidate the disease mechanism.

The Cox regression[3], also known as the proportional hazards model, is a classic model in survival analysis. The simplicity and interpretability of the Cox model come from the proportional hazards assumption, which basically states that the risk can be estimated based on a linear combination of the predictive variables. A trained Cox model can calculate a relative risk score for a new patient based on his/her clinical information, and is thus able to rank patients with their expected order of death. It cannot, though, directly predict the expected time to death.

The Cox-based model proposed by Halabi *et al.* in 2014[4] (referred to as **Halabi's model** in the rest of this manuscript) is a state-of-the-art method for clinical prediction of prostate cancer survival. Halabi's model is outlined in Figure 1(a). It starts with 22 features ("Halabi's 22 features"), including some previously defined predictors of overall survival and some clinical parameters, picks out the eight most important features ("Halabi's 8 features") using L_1 regularization, and predicts patients' risks using those eight features only.

We propose **DWCox, a density-weighted Cox model** for predicting prostate cancer survival. DWCox was a best-performing method in the **2015 Prostate Cancer DREAM Challenge** (**PCDC**), with performance better than or comparable to the best ensemble approaches. Simulations have shown that DWCox can achieve better performance than a standard Cox model when many sparsely distributed outliers exist in training data. DWCox is implemented in R in a way such that it can be easily re-trained and applied to other survival analysis problems, not restricted to prostate cancer. Please refer to the section "Data and software availability" for a download link and a citable link to the software.

Methods

DWCox assigns weights to the training data according to their local kernel density in the feature space, and then trains an adopted Cox model with those weights incorporated into the loss function, as demonstrated in Figure 1(b). DWCox can also predict the actual survival time from the predicted risk score using a linear regression.

The development of DWCox underwent two phases. It was first developed and tested during the PCDC, and then further refined after its success. In this paper, unless something is stated to happen during the PCDC, DWCox should be understood as what it is now after the post-challenge refinements.

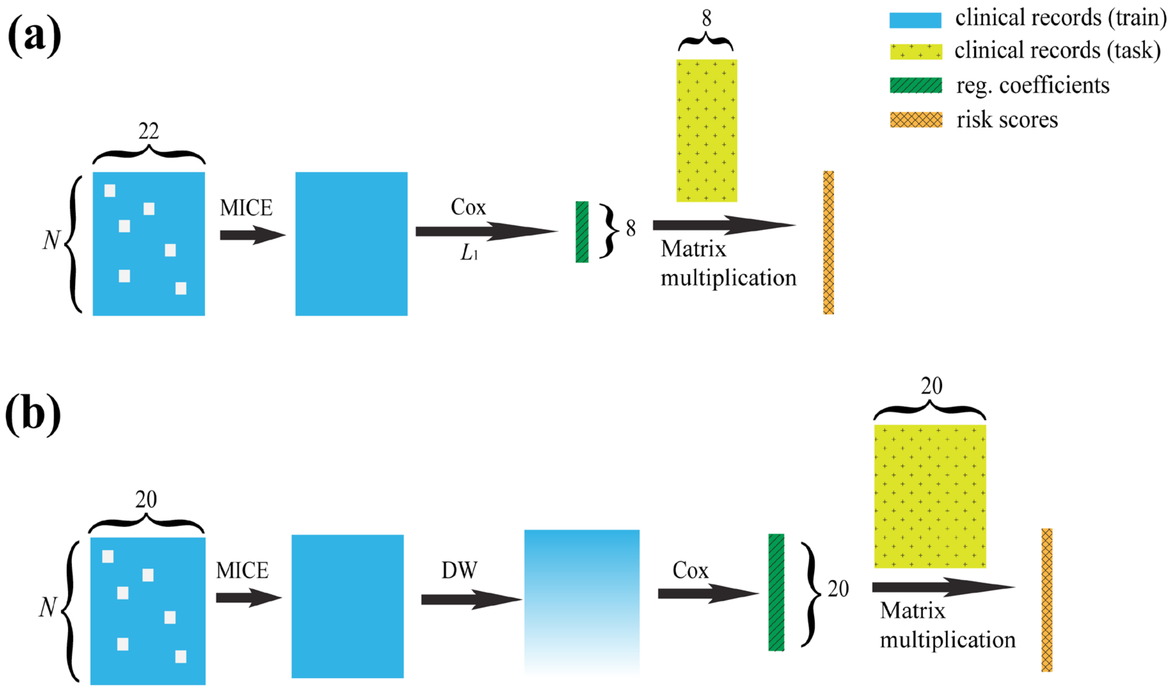

Figure 1. Illustration of how Halabi's model (a) and DWCox (b) predict the risk scores. DWCox is also able to predict the days to death using linear regression with the risk scores (not demonstrated in this figure). *N*: number of patients. MICE: Multivariate Imputation by Chained Equations. L_1: Lasso regularization. DW: Density-based weighting. Note that the objective functions in the Cox step of (**a**) and (**b**) are different, as discussed in the main text.

Feature construction

Training DWCox requires a training group of N patients whose clinical features $\mathbf{X}$ and survival outcomes $\mathbf{Y}$ are known. $\mathbf{X}$ is an N-by-M matrix, where M is the number of clinical features and each element X_{ij} is the value of the jth clinical feature of the ith patient. $\mathbf{Y}$ is an N-by-2 matrix, where each row gives the survival outcome of a patient. The 1st column of $\mathbf{Y}$ is a vector of the last observed survival time $\mathbf{t}$, and the 2nd column is a vector of binary event indicators $\mathbf{d}$. A patient i with $d_i = TRUE$ is known to die at time t_i. Oppositely, one with $d_i = FALSE$ is known to be alive at time t_i, but no information is available after t_i. In the latter case, the record of that patient is said to be censored. In the data sets used in the PCDC, $\mathbf{Y}$ is known, while $\mathbf{X}$ needs to be constructed from clinical data.

To ensure fair comparison with Halabi's model, DWCox constructed $\mathbf{X}$ in line with the way Halabi defined his 22 features, as summarized in Table 1 and described in details in the Supplementary material. Note that two features Halabi's model started with, namely the Charlson comorbidity index and the Biopsy Gleason score, were not considered by DWCox since during the PCDC the former was not available in the training data and the latter was 100% missing in the leaderboard data. (Data were split into training, leaderboard and final validation sets. Details will be described in the Experiments section). That means $M = 20$.

At this stage $\mathbf{X}$ was not complete (i.e. there were many missing elements in that matrix) due to missing information in the raw clinical records. Those missing values in $\mathbf{X}$ were imputed with the algorithm Multivariate Imputation by Chained Equations (MICE)[5,6]. The idea of MICE is to use Bayesian statistics to iteratively infer the missing values from other known and previously inferred values. Missing values in the training data were imputed with knowledge about the survival outcome, since it was argued that the outcomes could help generate less biased imputations[7]. The survival outcome was incorporated into the imputation in the form of the Nelson–Aalen estimator as suggested by White and Royston[8]. Imputation on the leaderboard and final validation data were done without using the survival outcome.

During the PCDC, three more binary features were used to indicate the trial ID (described in the Experiments section) of each patient. Those features were removed in post-challenge analysis so that the performance of DWCox does not depend on prior knowledge about the data source.

Density-based weighting

After the imputation, the N-by-M matrix $\mathbf{X}$ can be represented by N points scattered in a M-dimensional space $\mathbb{F}$ ("**feature space**"). Each point represents a patient whose each coordinate is the value of one of his/her M clinical features. We assign each patient i a weight $w_i \in [0, 1]$ proportional to the estimated local Gaussian kernel density in the feature space. To calculate w_i, we used the default settings of the function *kepdf* in the R package *pdfCluster*[9]. These weights were then divided by the maximum value. Thus a patient with a higher weight indicates there are more other patients with similar clinical features.

Table 1. Clinical features used by DWCox, ordered with decreasing $|\beta_i|$.

Variable name	Description	β_i	In Halabi's 8 features?
ast	aspartate aminotransferase level	0.567	
liver	liver metastases	0.497	
bmi	body mass index	-0.439	
alp	alkaline phosphatase level	0.260	Yes
ecogps	Eastern Cooperative Oncology Group performance status	0.204	Yes
alt	alanine transaminase level	-0.197	
race	race	-0.168	
hb	hemoglobin level	-0.117	Yes
lung	lung metastases	0.101	
analgesics	prior analgesics use	0.099	Yes
ds	disease site	-0.087	Yes
plt	platelet count	0.055	
psa	prostate-specific antigen level	0.050	Yes
wbc	white blood cell count	0.045	
bili	bilirubin level	0.041	
radio	prior radiotherapy	-0.040	
testo	testosterone level	0.038	
alb	albumin level	-0.027	Yes
ldh	lactate dehydrogenase level	-0.011	Yes
age	age	-0.008	

Model training

After density-based weighting, we used the R package *glmnet*[10] to maximize the weighted partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{e^{\sum_{j \in D_i} w_j \left(\mathbf{X}_j^T \boldsymbol{\beta}\right)}}{\left(\sum_{k:t_k \geq t_i} w_k e^{\mathbf{X}_k^T \boldsymbol{\beta}}\right)^{\sum_{l \in D_i} w_l}} \qquad (1)$$

where $\boldsymbol{\beta}$ is a vector of the regression coefficients, $\mathbf{X}_j$ denotes the jth row of $\mathbf{X}$, $t_1 < t_2 < \ldots < t_i < \ldots < t_m$ is an increasing list of death times in $\mathbf{Y}$, and D_i is the set of patients died at time t_i.

During the PCDC, L_2 regularization was imposed to the objective function. The penalty weight was chosen to optimize the model performance (more specifically, iAUC, as defined in the next subsection) averaged over 100 repeated random sub-sampling validation on the training data. In each random sub-sampling validation experiment, 2/3 of all the training data were randomly selected to train the model with a wide range of possible penalty weights, and the iAUC was evaluated for each possible penalty weight on the remaining 1/3 of the training data. After the PCDC, the regularization was removed from DWCox since its contribution to the model performance was not obvious during the challenge and its removal sped up training.

After model training, the risk vector **r** of the training patients were calculated as

$$\mathbf{r} = \mathbf{X}\hat{\boldsymbol{\beta}}. \qquad (2)$$

A higher risk indicates a shorter expected remaining lifetime for a patient. A linear regression $\mathbf{t} = \hat{k}\mathbf{r} + \hat{\mathbf{b}} + \mathbf{e}$ was then performed to correlate **t** to **r**, where $\hat{k}$ and $\hat{\mathbf{b}}$ were the regression coefficients, and **e** was the error between the actual survival time and the estimated value (i.e. $\mathbf{t} = \hat{\mathbf{t}} + \mathbf{e},$ where $\hat{\mathbf{t}} = \hat{k}\mathbf{r} + \hat{\mathbf{b}}$).

Prediction & evaluation

The trained model was used to predict the risk $\mathbf{r}_{\text{test}}$ and the remaining lifetime $\mathbf{t}_{\text{test}}$ for a new group of patients whose clinical features $\mathbf{X}_{\text{test}}$ could be constructed from clinical data while the outcome $\mathbf{Y}_{\text{test}}$ was not seen by the model. The model performance was then evaluated by comparing $\mathbf{r}_{\text{test}}$ and $\mathbf{t}_{\text{test}}$ to $\mathbf{Y}_{\text{test}}$.

The predicted risks $\mathbf{r}_{\text{test}}$ were evaluated with the integrated area under the ROC curve (iAUC) as described below. After obtaining $\hat{\boldsymbol{\beta}}$ by maximizing Equation (1), we can estimate the risks of the patients $\mathbf{r}_{\text{test}}=\mathbf{X}_{\text{test}}\hat{\beta}$. Then an estimated order of death $\hat{\mathbf{o}}$ can be constructed by sorting $\mathbf{r}_{\text{test}}$ (i.e. $\hat{\mathbf{o}}_i = j$ where $i = 1, 2, \ldots, N$ and $r_{\text{test},i}$ is the jth smallest element of $\mathbf{r}_{\text{test}}$). By comparing $\hat{\mathbf{o}}$ with the actual outcome $\mathbf{Y}_{\text{test}}$, at any given time threshold t_i we can calculate the area under the receiver operating characteristic curve AUC_{t_i}. If we integrate AUC_{t_i} with respect to t_i from the 6th to the 30th month, we get the integrated area under curve iAUC $\in$ [0, 1]. The greater the iAUC, the better the predicted risks reflect the actual order of death.

DWCox also gives the estimated time to death of the test set: $\hat{\mathbf{t}}_{\text{test}} = \hat{k}\mathbf{r}_{\text{test}} + \hat{\mathbf{b}}$. In the PCDC $\hat{\mathbf{t}}_{\text{test}}$ was evaluated by its RMSE from $\mathbf{t}_{\text{test}}$.

Extended applications

The open-source release of DWCox is coded in a way such that it can be easily re-trained and applied to other survival analysis problems, not restricted to prostate cancer. To re-train and apply DWCox to a new dataset, users simply need to:

- Format their data into the three matrices **X**, **Y** and $\mathbf{X}_{\text{test}}$.
- Hit enter and get some coffee.
- Now they get the predicted risk $\mathbf{r}_{\text{test}}$ and time to event $\hat{\mathbf{t}}_{\text{test}}$.

Here **X** and $\mathbf{X}_{\text{test}}$ can have as many rows (i.e. subjects) and columns (i.e. features) as needed. They can have missing values as well. More details can be found in the documentation inside the package.

Experiments

Challenge data & context

DWCox has been developed and evaluated with data from the comparator arms of four phase III clinical trials with over 2,000 mCRPC patients in total treated with first-line docetaxel. Those four trials and the corresponding data providers are:

- ASCENT-2 (Novacea, provided by Memorial Sloan Kettering Cancer Center)[11],
- MAINSAIL (Celgene)[12],
- VENICE (Sanofi)[13], and
- ENTHUSE-33 (AstraZeneca)[14].

During the PCDC those trials were referred to with their study IDs (Table 2).

The development and evaluation of DWCox began with the 2015 Prostate Cancer DREAM Challenge and continued after the

Table 2. Clinical trial used to develop DWCox.

Trial	Study ID	Number of patients	Survival outcome
ASCENT-2	ASCENT2	476	Known
MAINSAIL	CELGENE	526	Known
VENICE	EFC6546	598	Known
ENTHUSE-33	AZ	470	Hidden

challenge. The full anonymized information about the patients in trials ASCENT-2, MAINSAIL and VENICE was released to the challenge participants. As for trial ENTHUSE-33, the participants only knew the clinical records available at the beginning of the trial ("baseline clinical records"), while data obtained after the start of the trial including the survival outcome were visible only to the challenge organizers. The challenge goal was to develop models that used the baseline clinical records to predict the patients' relative risk (**sub-challenge 1a**), days till death (**sub-challenge 1b**), and treatment discontinuation (sub-challenge 2) (Table 3).

DWCox was trained on Trials ASCENT-2, MAINSAIL and VENICE ("**PCDC training data**") by the authors, and evaluated on Trial ENTHUSE-33 ("**PCDC validation data**") by the challenge organizers. Trial ENTHUSE-33 was further divided into a leaderboard set (157 patients) and a validation set (313 patients). The leaderboard set was used to run three leaderboard rounds. In each round, the challenge organizers randomly subsampled 80% patients from the leaderboard set, evaluated the participants' models on that random sample, and returned the feedback to the participants. After the 3rd leaderboard round, each participating team submitted a final model, whose performance on the validation set was used to rank the teams. Bootstrapping was performed by the challenge organizers to make sure the winning teams gave statistically significantly better predictions than other teams and Halabi's model. DWCox was involved in the leaderboard rounds of sub-challenge 1a and the final scoring round of sub-challenges 1a & 1b.

Table 3. Three sub-challenges of the PCDC.

Sub-challenge	What to predict	Evaluation metrics	Our participation
1a	Relative risk	iAUC	Yes
1b	Days to death	RMSE	Yes
2	Treatment discontinuation	(Irrelevant to this paper)	No

Simulations

Simulation experiments were performed to evaluate the contribution of density-based weighting to the model performance. DWCox was trained and evaluated on 100 simulated data sets (one example is given in Figure 2) separately, each of which was designed to mimic the real challenge data to some extent, while the randomness in the data generation process assured the variation across simulations. In each simulation, three groups of patients were simulated. Each patient had 20 features and an outcome.

One group ("**signal group**") represented a group of 1,000 patients that reflected the true correlation between the outcome and the features. The features were sampled from Gaussian distributions:

$$\mathbf{X}_{\text{signal},ij} \sim \mathcal{N}(\mu_j, \sigma_j^2) \quad (3)$$

where μ_j and σ_j were the mean and standard deviation of the jth feature in the PCDC training data. Following the idea of R. Bender *et al.*[15], we simulated the survival time of each patient i with a Weibull distribution:

$$t_{\text{signal},i} = \left(-\frac{\log\left(u_{\text{signal},i}\right)}{\lambda e^{\mathbf{X}^T_{\text{signal},i*}\boldsymbol{\beta}}} \right)^{\frac{1}{\nu}} \quad (4)$$

where $u_{\text{signal},i} \sim U(0, 1)$, $\lambda > 0$, $\nu > 0$, and the subscript $i*$ takes the ith row of the matrix. $U(\ ,\ )$ denotes uniform distributions.

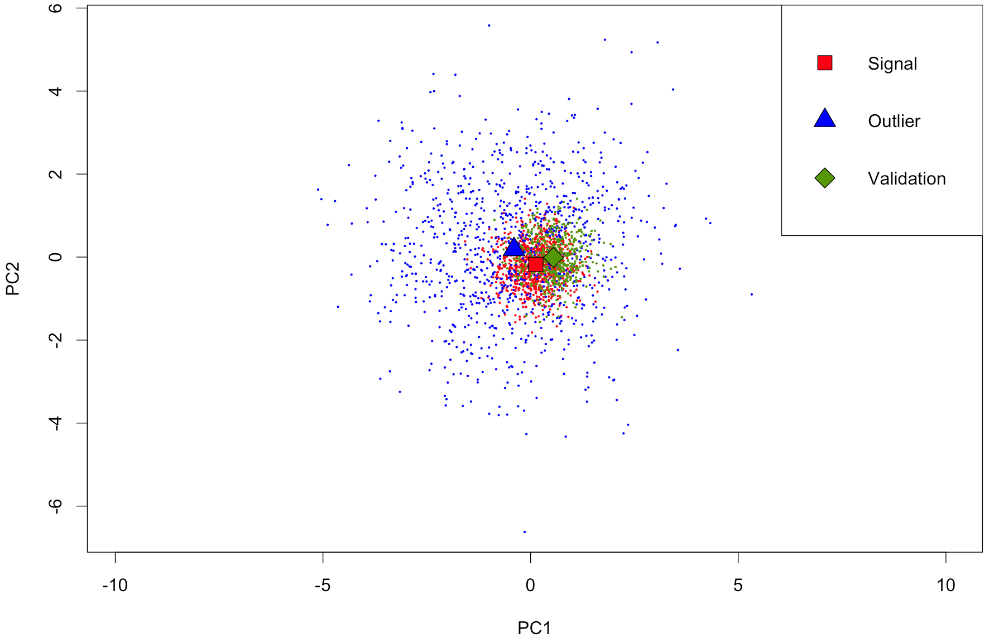

Figure 2. Scatter plot of the first two principle components of the signal, noise and validation groups in a simulated data set. Each point represents a patient. The shapes mark the mean of each group. (Best viewed in color).

We would like to clarify a few things about Equation (4). Readers may get confused if they see an online manuscript with the same title and authors as those of Reference 15, where the minus sign of Equation (4) is outside the parenthesis. Obviously it is a typo, and it has been corrected in the version cited here. Although Equation (4) may not look like a Weibull distribution at first glance, the proof is a very straightforward and standard procedure. The shape and scale parameters of the Weibull distribution is ν and $\left(\lambda e^{\mathbf{X}^T_{\text{signal},i*}\boldsymbol{\beta}}\right)^{-1/\nu}$ respectively.

Such generated survival times follow a Cox model with the baseline hazard function $h_0(t) = \lambda \nu t^{\nu-1}$[15]. The parameters λ, ν and $\boldsymbol{\beta}$ were estimated from the uncensored part of the PCDC training data as follows. First, we assumed $\boldsymbol{\beta} = \mathbf{0}$ and fit a Weibull distribution to the distribution of $\mathbf{t}_{\text{uncensored}}$ to estimate ν and λ. Then DWCox was applied to the PCDC training data to obtain $\hat{\boldsymbol{\beta}}$. At this stage $\hat{\boldsymbol{\beta}}$ did not include $\hat{\beta}_0$, a constant term that affected $\hat{\mathbf{t}}$ but not iAUC, since $\hat{\beta}_0$ played no role during the maximization process of Equation (1). We chose a $\hat{\beta}_0$ value such that the mean of the survival times simulated with Equation (4) was close to the mean of the uncensored survival times in the PCDC training data. After getting the estimates of λ, ν and $\boldsymbol{\beta}$, $\mathbf{t}_{\text{signal}}$ was simulated with Equation (4).

We then generated 1,000 more patients ("**noise group**") to represent outliers, or noises, in the training data. We made the outliers more sparsely and widely distributed in the feature space than the signal group by simulating

$$\mathbf{X}_{\text{noise},ij} \sim \mathcal{N}\,(c_{1j}\,\mu_j\,,\,(c_{2j}\,\sigma_j)^2) \qquad (5)$$

where $c_{1j} \sim U(0.5,\ 1.5)$ and $c_{2j} \sim U(2,\ 4)$. In this section, identical mathematical symbols present in multiple equations (e.g. μ_j in Equation (3) and Equation (5)) share the same definitions and values.

The survival times of the noise group were simulated with a Weibull distribution independent of $\mathbf{X}_{\text{noise}}$:

$$t_{\text{noise},i} = \left(-\frac{\log(u_{\text{noise},i})}{\lambda}\right)^{\frac{1}{\nu}} \qquad (6)$$

where $u_{\text{noise},i} \sim U(0,\ 1)$.

A 3rd group of 500 patients ("**validation group**") was generated in a fashion similar to that of the signal group.

We let

$$\mathbf{X}_{\text{vali},ij} \sim \mathcal{N}\,(c_{3j}\,\mu_j\,,\,(c_{4j}\,\sigma_j)^2) \qquad (7)$$

where $\mathrm{c}_{3j} \sim U(0.5,\ 1.5)$ and $\mathrm{c}_{4j} \sim \mathrm{U}(0.8,\ 1.2)$. The survival times are generated with

$$t_{\text{vali},i} = \left(-\frac{\log(u_{\text{vali},i})}{\lambda e^{\mathbf{X}^T_{\text{vali},i*}\boldsymbol{\beta}}}\right)^{\frac{1}{\nu}} \qquad (8)$$

where $u_{\text{vali},i} \sim U(0,\ 1)$.

After simulating the three groups of patients, we mixed the signal and noise groups together to form a training set. DWCox and a 20-feature standard Cox model were trained on this training set, and evaluated with iAUC on the validation group.

Results

DWCox was submitted to the sub-challenges 1a & 1b (Table 3) of the 2015 Prostate Cancer DREAM Challenge. Sub-challenge 1a aimed at better predictions on the relative risks and order of death, evaluated with iAUC. Sub-challenge 1b evaluated the models using the RMSE between the predicted days to death and the actual time. While this manuscript is focused on our method, more details about other teams' methods and performance can be found in papers from the challenge organizers and individual teams.

Heterogeneity in the PCDC data

Analysis of the PCDC data suggests that there exists rather high heterogeneity across the three training trials and the validation trial. The missing-rate profile of the 20 clinical features varies across trials (Figure 3). The average values of the first two principle components of the 20 features of Trial ASCENT-2 is farther away from those of the validation trial, compared to those of the other two training trials (Figure 4). Leave-one-trial-out cross-validation (i.e. to train with two training trials and evaluate with the left-out training trial) gives very different results when different trials are left out (Table 4).

Those facts give such a clue: If we consider the "true model" underlying the validation trial as the signal, it is very likely that the PCDC training data contain many outliers. Those outliers do not follow the "true model", and thus tend to bring down the validation-set performance of models that failed to deal with the outliers properly during training. Therefore robustness against outliers is probably important to models aimed at winning the PCDC.

Indeed, several other winning teams of the PCDC tried hard to deal with the outliers in the training data. For example, the top performer (FIMM-UTU) of sub-challenge 1a decided to discard the entire ASCENT-2 trial, because after some manual exploration in the data space they found significant differences in clinical variables that set this trial apart from the other trials. Our team (Team Cornfield) instead used all available data and let DWCox automatically handle the outliers.

Results on the PCDC data

DWCox obtained the best average ranking in sub-challenges 1a & 1b among about 50 models (Figure 5). On the PCDC validation data, DWCox gave an iAUC of 0.7789 and a RMSE of 194.8650 days, out-performing Halabi's model which gave an iAUC of 0.7581 and a RMSE of 196.6704 days. Bootstrapping has shown that DWCox outperforms Halabi's model with a Bayes Factor (BF) > 3. Note that while the other numbers in this paragraph are official results provided by the challenge organizers, the Halabi RMSE is not. In order to get the Halabi RMSE, we implemented a Halabi's model and appended to it a linear regression step similar to the one in DWCox. After applying bootstrapping and the BF > 3 threshold against other teams' submissions, the challenge organizers reported DWCox as a winner in sub-challenge 1b and a runner-up in sub-challenge 1a. The winner of sub-challenge 1a, FIMM-UTU, obtained

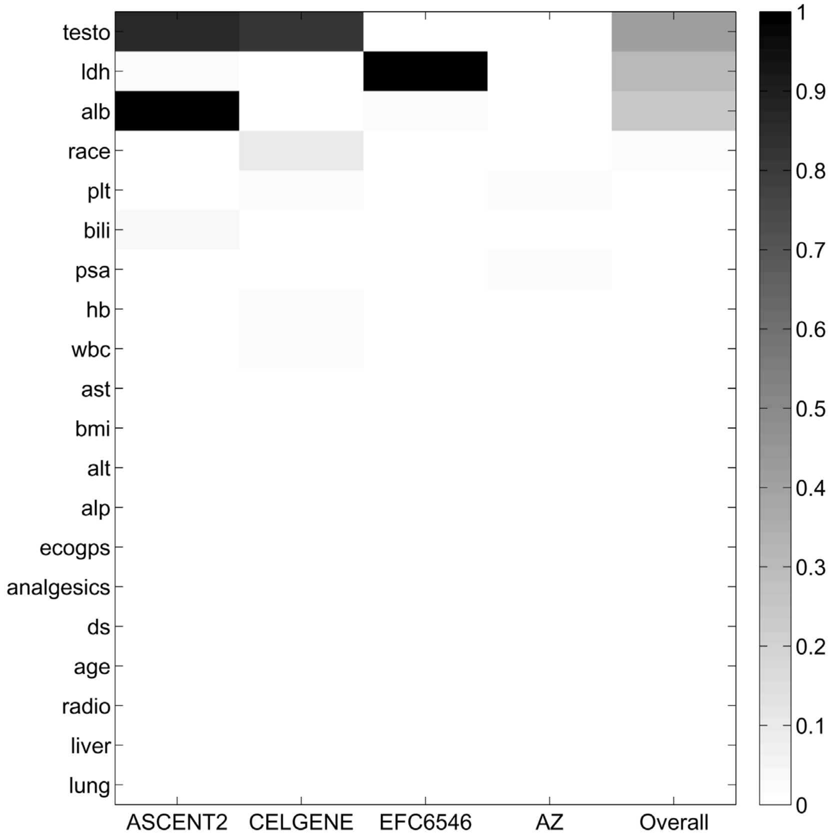

Figure 3. Heatmap of the percentage missing of the 20 clinical features used in DWCox.

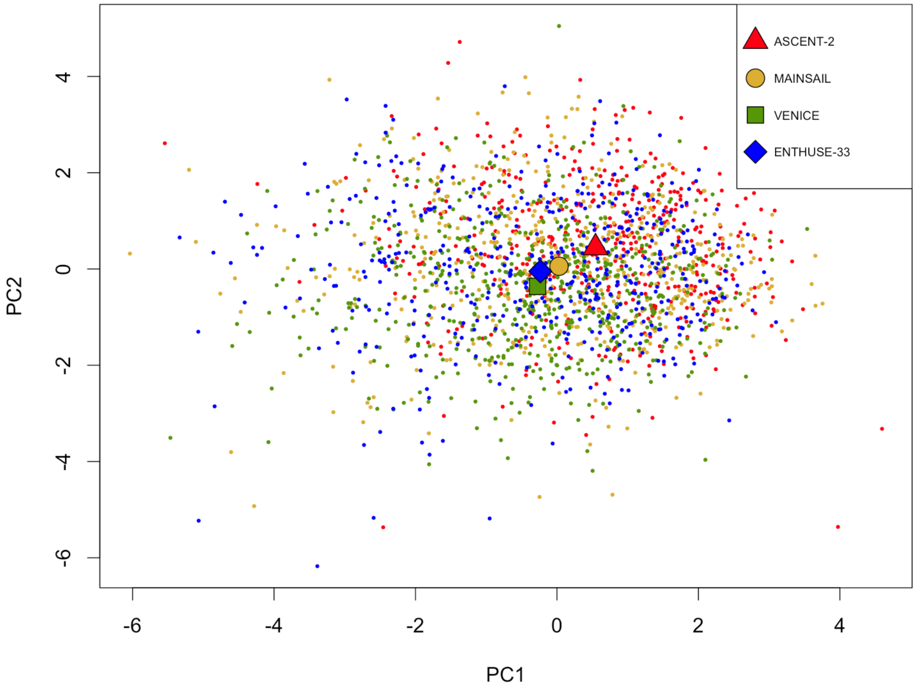

Figure 4. Scatter plot of the first two principle components of the four prostate cancer trials. Each point represents a patient. The shapes mark the average values of each trial.

an iAUC of 0.7915 and a RMSE of 201.3779 days. Their model is an ensemble of penalized cox regressions developed with extensive manual exploration in the data space. More details about the challenge results can be found at https://www.synapse.org/#!Synapse:syn2813558/wiki/232674. Table 1 gives the regression coefficients determined by DWCox.

Table 4. Results of DWCox leave-one-trial-out cross-validation.

Left-out trial	iAUC
ASCENT-2	0.572
MAINSAIL	0.567
VENICE	0.685

An inverse correlation between the actual survival time **t** and risk scores **r** was observed (Figure 6). Note that the adjusted R^2 of the linear regression $\hat{\mathbf{t}} = \hat{k}\mathbf{r} + \hat{\mathbf{b}}$ is small (0.1513), and the shape of the **t** vs **r** plot implies that there may exist models better than a linear regression for capturing their correlation.

Results on simulated data

In the 100 repeated simulations (described in the Experiments section), DWCox performed better than a standard Cox model when as many as half of the training data were outliers. DWCox

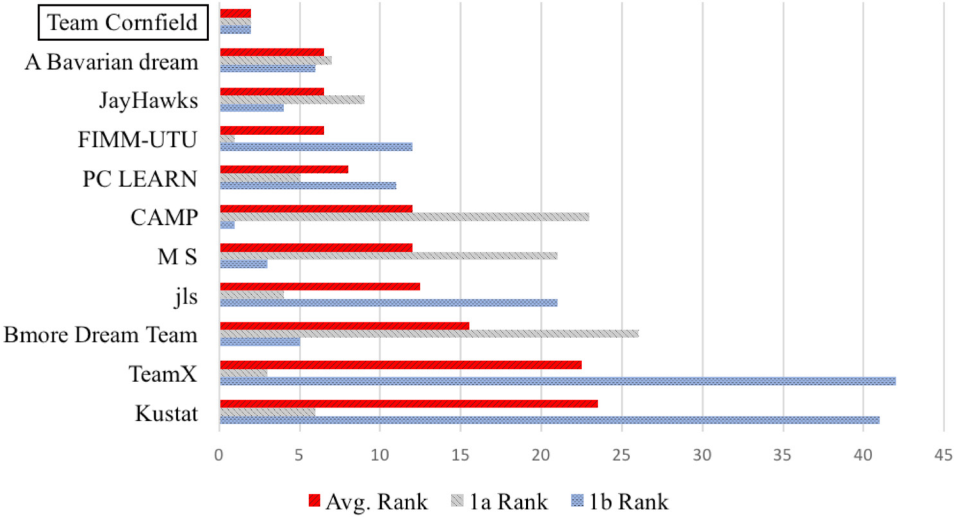

Figure 5. Ranking of the top teams in sub-challenges 1a & 1b. The six best teams of each sub-challenge are included. DWCox was submitted by the authors' Team Cornfield.

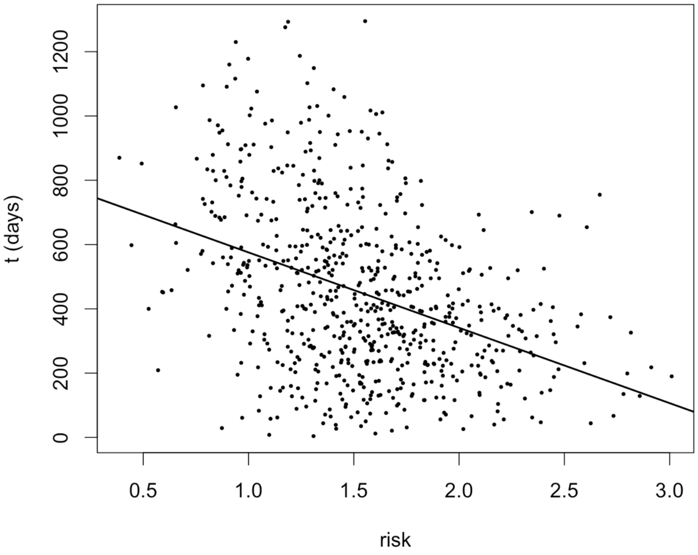

Figure 6. Scatter plot of the uncensored survival time versus the predicted risk on the PCDC training data. The straight line is the linear regression line with slope = -234.6, intercept = 810.3 and adjusted R^2 = 0.1513.

not only gave better average performance over the 100 experiments (Table 5, Figure 7), but also performed consistently better in each experiment (Figure 8, paired t-test p-value = 2.1×10^{-20}). The improvement in performance clearly resulted from the density-based weighting, since everything else was the same across the two models.

Note that in the simulations we used iAUC but not the RMSE to evaluate model performance. There are three reasons for that. 1. iAUC evaluates model performance on the validation data in a more comprehensive manner, while RMSE is based on individual predictions which are independent of each other. 2. DWCox's time-to-event prediction is dependent on its predicted risks. 3. A standard Cox model does not directly give the predicted time-to-event.

Table 5. iAUC statistics of 100 simulations.

	DWCox	Cox
Mean(iAUC)	0.674	0.643
SD(iAUC)	0.033	0.033

Discussion

We propose DWCox, a density-weighted Cox model for survival analysis that is more robust against overfitting outliers from the training data. In our simulations DWCox outperformed the standard Cox when as many as half of the training data were noise. In the 2015 Prostate Cancer DREAM Challenge (the PCDC), DWCox obtained the best average ranking in sub-challenge 1, which was to predict the risk and survival time of prostate cancer patients from clinical data available at the beginning of trials.

DWCox was one of the only two models among the seven winners of the PCDC sub-challenge 1 that did not use super-learners (or ensemble methods). (The other model[16] of the two was a standard Cox trained with different features. In Figure 5 the corresponding team name is M S.) This is a remarkable achievement, since super-learners usually give better results than single methods. Given that now DWCox gives results comparable to or better than ensemble methods, there are even more reasons to prefer DWCox over ensemble ones in real-world applications. During the training of ensemble methods, there often exist some empirical parameters (e.g. the number of base learners to use) that require more hyper-parameter tuning, because people do not know exactly which value works best and why. In addition, some ensemble methods

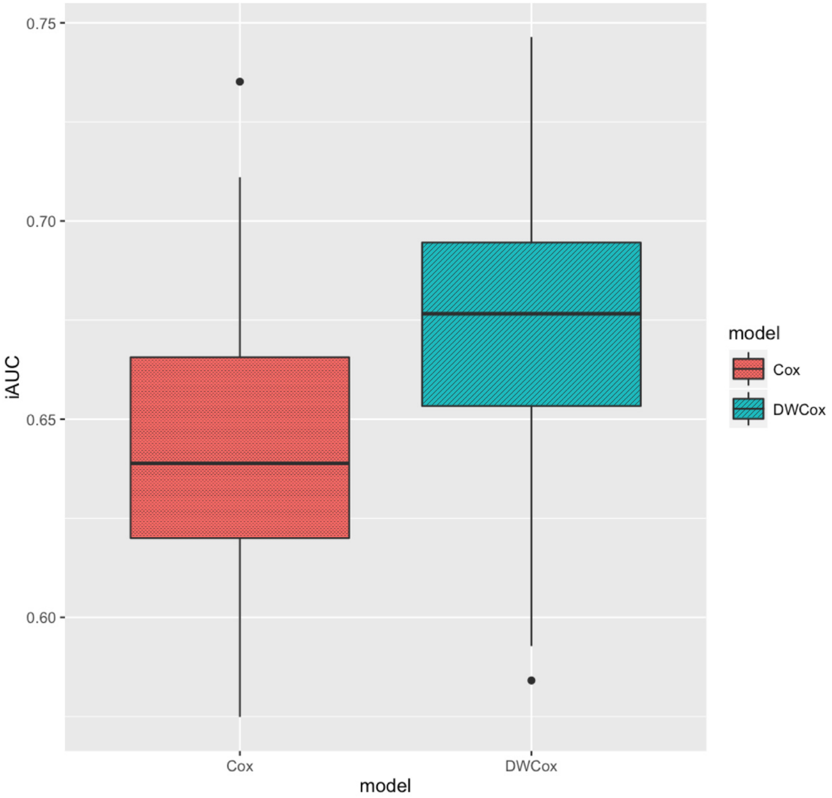

Figure 7. Boxplot of the iAUC of DWCox and a standard Cox model in 100 simulations. The boxes show the medians and inter-quartile ranges (IQR). The vertical black lines extends from the boxes by at most 1.5 IQR. Black points represent experiments whose iAUC is more than 1.5 IQR away from the boxes.

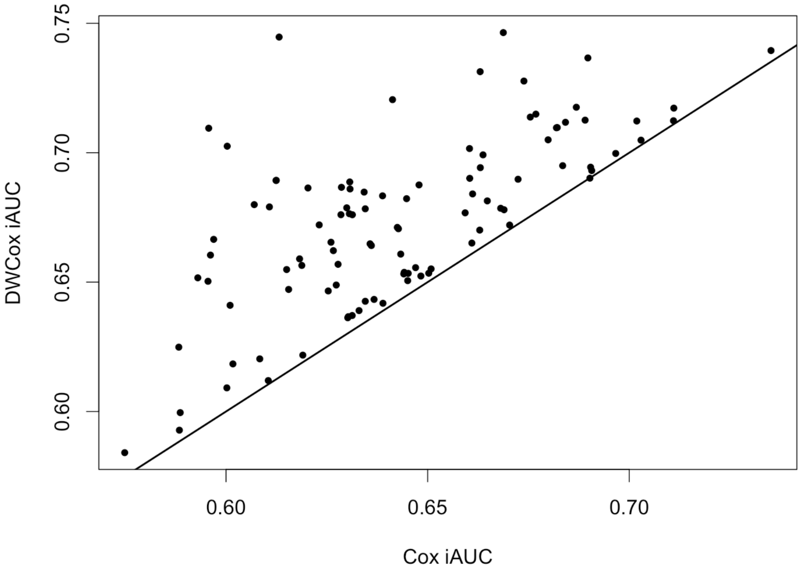

Figure 8. DWCox iAUC vs the standard Cox iAUC in 100 simulations. Each point is given by a simulation. The straight line has slope = 1 and intercept = 0.

(e.g. random forests) have great built-in randomness and produce very complex models, and thus it is sometimes hard to interpret and understand the results they give. Oppositely, the training phase of DWCox involves no empirical parameters or built-in randomness (except when the user wants DWCox to impute the missing data with MICE), and the results can be easily interpreted.

DWCox's success in the PCDC should be credited to its density-based weighting mechanism. There exists inter-trial heterogeneity in the PCDC data, which implies some training trials may contribute more signals than others, while some may contain more outliers. It turned out that several top-performing methods of the PCDC recognized such problem and tried to handle it properly. DWCox achieved this by taking in all training data and automatically weighting away outliers with the local Gaussian kernel density. DWCox can be easily re-trained and applied to other data sets, not restricted to prostate cancer survival data.

Perhaps the greatest limitation of DWCox also lies in its density-based weighting mechanism. Such mechanism cannot weight away outliers falling inside the signaling region of the feature space, or outliers that happen to cluster together in the feature space and thus give a local kernel density similar to those of the signals. In another extreme case where the data contain few outliers and follow a standard Cox model rather well, introducing weights into the partial likelihood function can make the performance worse. Therefore it is better to apply DWCox to cases where the data are expected to contain some sparsely distributed outliers.

Data and software availability

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.

The clinical trial data used in the PCDC, in its raw and processed format, can be accessed at: https://www.projectdatasphere.org/projectdatasphere/html/content/149?pcdc=true. Challenge documentation, including the detailed description of the Challenge design, overall results, scoring scripts, and the clinical trials data dictionary can be found at: https://www.synapse.org/ProstateCancerChallenge.

An R implementation of DWCox can be downloaded from https://github.com/JinfengXiao/DWCox. A citable snapshot of that GitHub repository has also been archived with the DOI: 10.5281/zenodo.167143[17].

Author contributions

JX and JP designed DWCox and the simulations. JX, SW, JS, HL, DX and XR tested the performance of DWCox against other methods. JX, JS and HL drafted the manuscript, and the other authors helped revise it. JP and JH advised the team.

Competing interests

No competing interests were disclosed.

Grant information

Research was sponsored in part by grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), National Science Foundation IIS-1320617, IIS-1354329 and IIS 16-18481, and the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA). The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the opinions of the funding agencies.

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

Acknowledgments

The authors would like to thank Sage Bionetworks, the DREAM organization, and Project Data Sphere for developing and supplying data for the Challenge.

Supplementary material

This section gives details about the 20 predictors used by DWCox in the PCDC.

- age: Categorical variable with 3 levels. 18–64 years old = 1; 65–74 years old = 2; at least 75 years old = 3.
- alb: Albumin level in g/L. Continuous variable.
- alp: Natural logarithm of the alkaline phosphatase level in U/L. Continuous variable.
- alt: Natural logarithm of the alanine transaminase level in U/L. Continuous variable.
- analgesics: Prior analgesics use. Binary variable. 1 means yes; 0 means no. Note that this is not exactly the "opioid analgesic use" as appeared in the baseline paper, since the latter is not contained in the challenge data set.
- ast: Natural logarithm of the aspartate aminotransferase level in U/L. Continuous variable.
- bili: Natural logarithm of the total bilirubin level in μmol/L. Continuous variable.
- bmi: Natural logarithm of the body mass index in kg/m^2. Continuous variable.
- ds: Disease site. Categorical variable with 3 levels. 0 means the disease sites are not at bones or viscera. 1 means the disease sites are at bones but not at viscera. 2 means at least some disease sites are at viscera.
- ecogps: Eastern Cooperative Oncology Group performance status. Categorical variable with 3 levels (0, 1 and 2). The greater the value is, the more severe the situation is for the patient. Technically this variable should have 6 levels (0, 1, …, 5), but Halabi's model only considers the first 3 levels. Besides, in the challenge training data there is only 1 patient whose ecogps is greater than 2 (and it is 3). Therefore DWCox sets all ecogps > 2 to 2.
- hb: Hemoglobin level in g/dL. Continuous variable.
- ldh: Lactate dehydrogenase level. Binary variable. 1 means the lactate dehydrogenase level is greater than 200 units/liter, which is considered as the value of the upper limit of normal (ULN)[18]. 0 means the opposite.
- liver: Liver metastases. Binary variable. Yes = 1; No = 0.
- lung: Lung metastases. Binary variable. Yes = 1; No = 0.
- plt: Natural logarithm of the platelet count in 10^9/L. Continuous variable.
- psa: Natural logarithm of the prostate-specific antigen level in ng/mL. Continuous variable. The reason for taking logarithm is to make the distribution less skewed.
- race: categorical variable with 4 levels. White = 1; Asian = 2; Black = 3; Other or Hispanic = 4.
- radio: Prior radiotherapy. Binary variable. Yes = 1; No = 0.
- testo: Testosterone level in nmol/L. Continuous variable.
- wbc: Natural logarithm of the white blood cell count in 10^9/L. Continuous variable.

References

1. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2015.** *CA Cancer J Clin.* 2015; **65**(1): 5–29.
PubMed Abstract | Publisher Full Text

2. Garcia M, Jemal A, Ward EM, *et al.*: **Global cancer facts & figures 2007.** *Atlanta, GA: American cancer society.* 2007; **1**(3): 52.
Reference Source

3. Cox DR: **Regression models and life-tables.** In *Breakthroughs in statistics.* Springer, 1992; 527–541.
Publisher Full Text

4. 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

5. van Buuren S, Boshuizen HC, Knook DL, *et al.*: **Multiple imputation of missing blood pressure covariates in survival analysis.** *Stat Med.* 1999; **18**(6): 681–694.
PubMed Abstract | Publisher Full Text

6. van Buuren S, Groothuis-Oudshoorn K: **mice: Multivariate imputation by chained equations in R.** *J Stat Softw.* 2011; **45**(3).
Publisher Full Text

7. Moons KG, Donders RA, Stijnen T, *et al.*: **Using the outcome for imputation of missing predictor values was preferred.** *J Clin Epidemiol.* 2006; **59**(10): 1092–1101.
PubMed Abstract | Publisher Full Text

8. White IR, Royston P: **Imputing missing covariate values for the Cox model.** *Stat Med.* 2009; **28**(15): 1982–1998.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Azzalini A, Menardi G: **Clustering via nonparametric density estimation: the R package pdfCluster.** *arXiv preprint arXiv: 1301.6559.* 2013.
Reference Source

10. Simon N, Friedman J, Hastie T, *et al.*: **Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent.** *J Stat Softw.* 2011; **39**(5): 1–13.
PubMed Abstract | Publisher Full Text | Free Full Text

11. 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

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

13. 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

14. 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

15. Bender R, Augustin T, Blettner M: **Generating survival times to simulate Cox proportional hazards models.** *Stat Med.* 2005; **24**(11): 1713–1723.
PubMed Abstract | Publisher Full Text

16. Shiga M: **Two-step feature selection for predicting survival time of patients with metastatic castrate resistant prostate cancer [version 1; referees: awaiting peer review].** *F1000Res.* 2016; **5**: 2678.
Publisher Full Text

17. Xiao J: **DWCox: A Density-Weighted Cox Model for Outlier-Robust Prediction of Prostate Cancer Survival [Data set].** *Zenodo.* 2016.
Data Source

18. Joseph J, Badrinath P, Basran GS, *et al.*: **Is the pleural fluid transudate or exudate? A revisit of the diagnostic criteria.** *Thorax.* 2001; **56**(11): 867–870.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✓ ? ?

Version 1

Referee Report 17 February 2017

doi:10.5256/f1000research.10162.r20019

?

Riku Klen[1], **Mehrad Mahmoudian**[2]

[1] University of Turku, Department of Mathematics and Statistics, Turku, Finland

[2] Turku Centre for Biotechnology, University of Turku, Turku, Finland

General comments

The article introduces a density-weighted Cox model (DWCox). The model was created by Team Cornfield in the 2015 Prostate Cancer DREAM Challenge for outlier-robust prediction of survival. The article is well written and the introduced method is novel. The only major comment about the article is that the comparison of the new method and the existing methods could be more complete.

The authors test the DWCox method with the 2015 Prostate Cancer DREAM Challenge data and simulated data. They compare DWCox method with the Cox model and Halabi's model introduced in reference 4. I suggest that the authors also consider as a fourth alternative method the Smaletz method [1] due to the fact that it was also used in Halabi et al. article. To make the study complete these 4 methods should be compared and the results reported for the simulated data and the 2015 Prostate Cancer DREAM Challenge. It would be interesting to see complete results in the spirit of Table 5 and Figure 7.

Table 1 shows the 22 featured ranked by DWCox. It would be interesting to know how DWCox behaves when the features with missing values (namely race, testo, ldh or alb) were omitted. Based on the Table 1 these features have small weight and they might have little effect on the prediction results.

Additionally, it would be interesting to see how the method behaves on other data sets. However, this general study might be out of the scope of this article.

Furthermore, it would be more descriptive to explain how the missing values are handled in DWCox approach in page 5.

Detailed comments:

- page 1, line 6 of abstract: The result will not be worse in interpretation, but the model's interpretability will decrease. Hence the last word "result" should be substituted by "model"
- page 3, line 18: It should be specified that the reason Cox is not appropriate for testing time dependency is due to the nature of semi-parametric models that are have no assumption on the shape of the hazard function.[2]
- page 3, line 30: It would be nice to have a citation to clarify the statement thatDWCox was performing

"better than or comparable to the best ensemble approaches"
- page 10, line 3: the comparison represented in Figure 8 was done using t-test. Maybe Wilcoxon test would have been more appropriate.

References

1. Smaletz O, Scher HI, Small EJ, Verbel DA, McMillan A, Regan K, Kelly WK, Kattan MW: Nomogram for overall survival of patients with progressive metastatic prostate cancer after castration.*J Clin Oncol*. 2002; **20** (19): 3972-82 PubMed Abstract | Publisher Full Text
2. Garson, GD: Paremetric survival analysis. *Statistical associates blue book series 17*. 2012.

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

Competing Interests: Mahmoudian is co-author with the authors of this article in the following community article DOI: 10.1016/S1470-2045(16)30560-5 There are no other competing interests.

Referee Report 13 February 2017

doi:10.5256/f1000research.10162.r20018

?

Sebastian Pölsterl
Insitute for Cancer Research , London, UK

The authors present an interesting extension of the well-known Cox proportional hazards model if data contain outliers. They demonstrate the advantage over the traditional Cox model on synthetic data and applied their proposed model to a real world problem in the context of the 2015 Prostate Cancer DREAM challenge.

Major issues

Unfortunately, the author only provides little insight in the motivation for choosing a kernel density estimator to determine the sample weights. In particular, traditional kernel density estimation is only applicable to continuous random variables, whereas feature vectors comprised of clinical variables can contain continuous as well as categorical variables. It is unclear how density estimation was performed when feature vectors are a mix of continuous and categorical variables. Moreover, I strongly suggest to explicitly mention the assumption of the proposed density-weighted Cox model. The authors state that their proposed model is suitable when data "contain some sparsely distributed outliers." A more systematic approach to thoroughly formulate this assumption would be highly appreciated.

Minor issues

1. Page 2, paragraph 2:
    1. Reference 3, please cite the original paper Cox 1972.
2. Page 4, paragraph 3:
    1. I would suggest to change reference 5 to the original work on multiple imputation by Rubin: D. B. Rubin, Multiple imputation for nonresponse in surveys, John Wiley & Sons Inc., 1987.[1]
3. Page 5, paragraph 2:
    1. How were the candidate values for the L2 penalty chosen?
4. Page 5, paragraph 3:

    1. It is not clear what the coefficients \hat{b} represent. It seems they are not associated with any features, only \hat{k} is.
    2. Is the error e assumed to be normally distributed? If yes, such a choice might be problematic, because survival times usually follow a skewed distribution. Representing the log survival time as a linear model, as in the case of the accelerated failure time model, is usually preferred.
5. Page 5, paragraph 5:
    1. Please cite the original work on iAUC:
        - H. Hung and C. T. Chiang, “Estimation methods for time-dependent AUC models with survival data,” Canadian Journal of Statistics, vol. 38, no. 1, pp. 8–26, 2010.[2]
        - H. Uno, T. Cai, L. Tian, and L. J. Wei, “Evaluating prediction rules for t-year survivors with censored regression models,” Journal of the American Statistical Association, vol. 102, pp. 527–537, 2007 [3]
    2. It should be mentioned that the RMSE used in the challenge was only with respect to uncensored survival times in the test set.
6. Page 5, extended applications:
    1. I would suggest to remove this short section, because it is already clear from the description in the text, that the author’s propose a general model that can be applied to any survival data.
7. Page 7, paragraph 2:
    1. Please cite the recently published paper summarising the Prostate Cancer DREAM challenge
8. Page 7, Results on the PCDC data:
    1. The author’s stated earlier that the Halabi model is based on a Cox model and that a Cox model is not able to directly predict time to death. However, the authors mention that the Halabi model achieved an RMSE of 196.6704. How was this value obtained, if the model is not applicable for this task?
    2. It would be helpful of the authors could add the exact Bayes factor of the proposed model.
9. Page 9, Results on simulate data
    1. To better understand the benefit of the proposed density-weighted Cox model, it might be interesting to plot the the i-th weight against the i-th residual in the unweighted Cox model. I would assume that samples with high residuals are assigned a low weight, leading to an overall better prediction.

Grammar

The text contains several grammatical errors and convoluted formulations, which damps the overall presentation. I strongly suggest to improve the grammar and wording. I’m only highlighting some obvious errors below.

Page 3, paragraph 2:

- The simplicity and interpretability of the Cox model come from the proportional hazards assumption

Page 4, paragraph 5:

- Each point represents a patient whose each coordinate is the value of one of his/her *M* clinical features.

Page 6, paragraph 1:

- The challenge goal was to develop models

Page 7, paragraph 1:

- The shape and scale parameters of the Weibull distribution is

References

1. Rubin DB: Multiple Imputation for Nonresponse in Surveys. *John Wiley and Sons, Inc*. 1987. Publisher Full Text

2. Hung H, Chiang C: Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*. 2009. Publisher Full Text

3. Uno H, Cai T, Tian L, Wei L: Evaluating Prediction Rules fort -Year Survivors With Censored Regression Models. *Journal of the American Statistical Association*. 2007; **102** (478): 527-537 Publisher Full Text

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

Competing Interests: No competing interests were disclosed.

Referee Report 13 December 2016

doi:10.5256/f1000research.10162.r18159

✓

Motoki Shiga

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

This paper proposed a weighted Cox proportional hazards model (DWCox) to reduce the effects of outliers. Experimental results demonstrated that DWCox outperforms the standard Cox model. The proposed method is interesting. This manuscript is well-written.

Major comments:

1. Table 1 shows that the selected features by Halabi's model (Halabi's 8 features) and DWCox are quite different. A performance comparison the proposed model with a Cox model using Halabi's 8 features would be a good demonstration of the proposed method.

2. Performance of DWCox and a standard Cox model were compared using only simulated data. A performance comparison using real datasets by leave-one-trial-out CV such as Table 4 is an important experiment to evaluate the proposed method.

Minor comments:

1. p. 5 (in the above section of Eq. (2)): “sped up training” -> “speed up training”.

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.

Competing Interests: No competing interests were disclosed.

Author Response 09 Jan 2017

Jinfeng Xiao, University of Illinois at Urbana-Champaign, USA

Dear Dr. Shiga,

Thank you for reviewing our manuscript! We appreciate your feedback. Here is our response to your major comments.

1. DWCox versus Halabi's model
   Halabi's model is the baseline method of the Prostate Cancer DREAM Challenge (PCDC). As described in the "Results on the PCDC data" subsection under the "Results" section, the better performance of DWCox compared to Halabi's model was validated by the challenge organizers using bootstrapping.

2. DWCox versus Cox in leave-one-trail-out cross-validation
   We tried both DWCox and a standard Cox in leave-one-trail-out cross-validations. The difference in iAUC is less than 1%, which is much smaller than the difference across the three leave-one-trial-out cross-validation experiments (Table 4). In this case the difference in iAUC is dominated by the inter-trial heterogeneity, and thus the contribution of density-based weighting is masked. It is also interesting that DWCox's iAUC (0.779) on the validation data set is much higher than its highest iAUC (0.685) in leave-one-trail-out cross-validation experiments. It indicates that the validation trial is better represented by the three training trials, compared to how well each training trial is represented by the other two.

Competing Interests: No competing interests were disclosed.