METHOD ARTICLE

Three general concepts to improve risk prediction: good data, wisdom of the crowd, recalibration [version 1; referees: awaiting peer review]

Ivan Kondofersky[1,2], Michael Laimighofer[1,2], Christoph Kurz[3], Norbert Krautenbacher[1,2], Julia F. Söllner[1], Philip Dargatz[4], Hagen Scherb[1], Donna P. Ankerst[2,5], Christiane Fuchs [1,2]

[1]Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
[2]Mathematical Modeling of Biological Systems, Center for Mathematics, Technical University of Munich, Garching, Germany
[3]Institute of Health Economics and Health Care Management, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
[4]Department of Hematology and Oncology, Johannes Wesling Klinikum Minden, Minden, Germany
[5]University of Texas Health Science Center at San Antonio, San Antonio, USA

v1 **First published:** 16 Nov 2016, **5**:2671 (doi: 10.12688/f1000research.8680.1)
Latest published: 16 Nov 2016, **5**:2671 (doi: 10.12688/f1000research.8680.1)

Abstract

In today's information age, the necessary means exist for clinical risk prediction to capitalize on a multitude of data sources, increasing the potential for greater accuracy and improved patient care. Towards this objective, the Prostate Cancer DREAM Challenge posted comprehensive information from three clinical trials recording survival for patients with metastatic castration-resistant prostate cancer treated with first-line docetaxel. A subset of an independent clinical trial was used for interim evaluation of model submissions, providing critical feedback to participating teams for tailoring their models to the desired target. Final submitted models were evaluated and ranked on the independent clinical trial. Our team, called "A Bavarian Dream", utilized many of the common statistical methods for data dimension reduction and summarization during the trial. Three general modeling principles emerged that were deemed helpful for building accurate risk prediction tools and ending up among the winning teams of both sub-challenges. These principles included: first, good data, encompassing the collection of important variables and imputation of missing data; second, wisdom of the crowd, extending beyond the usual model ensemble notion to the inclusion of experts on specific risk ranges; and third, recalibration, entailing transfer learning to the target source. In this study, we illustrate the application and impact of these principles applied to data from the Prostate Cancer DREAM Challenge.

Open Peer Review

Referee Status: *AWAITING PEER REVIEW*

Discuss this article

Comments (0)

This article is included in the DREAM Challenges channel.

Corresponding author: Christiane Fuchs (christiane.fuchs@helmholtz-muenchen.de)

How to cite this article: 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: awaiting peer review]** *F1000Research* 2016, **5**:2671 (doi: 10.12688/f1000research.8680.1)

Grant information: CF and IK are supported by the European Union within the ERC grant LatentCauses. CF is supported by the German Research Foundation (DFG) within the Collaborative Research Centre 1243, Subproject A17.
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: 16 Nov 2016, **5**:2671 (doi: 10.12688/f1000research.8680.1)

Introduction

Government funded clinical and research trials are currently experiencing increased pressure to publish comprehensive anonymized data in order to maximize scientific output, ushering in new challenges and opportunities for data scientists[1]. In an era of personalized medicine, scientists analyzing the results of large population-based clinical and prevention trials are further encouraged to translate results to clinical practice. With patient as the consumer, this push has led to an explosion of easy-to-use online clinical risk prediction tools for nearly all types of clinical outcomes[2,3]. In the past, single-study prediction models dominated out of convenience. In the current climate, multiple studies are available that can be combined, increasing accuracy through the wisdom-of-the-crowd philosophy, and providing more realistic estimates of variability for decision-making. Ensembles or collections of models have been shown to outperform top-nominated models[4].

Following efforts by Project Data Sphere to coordinate the release of comparative arm data from multiple pharmaceutical companies and academic medical centers, and in cooperation with the Dialogue for Reverse Engineering Assessments and Methods (DREAM) initiative, the Prostate Cancer DREAM Challenge sought to facilitate the development of survival prediction models to assist patients with metastatic castration-resistant prostate cancer (mCRPC) treated with first-line docetaxel[5,6]. Baseline and follow-up data were available from 1600 patients who had received first-line docetaxel as part of their participation on the comparator arms of three clinical trials, which formed the training set; see Guinney *et al.*[7] and https://www.synapse.org/ProstateCancerChallenge for a detailed description. This article focuses on the challenge of predicting overall patient survival (sub-challenge 1). Here, data from 157 patients from an independent trial were made available for calibration to the target, and the final model based on the training and calibration data was validated on 313 patients from the target. An open online competition format with multiple deadlines attracted researchers from around the world, encouraging efficiency and fast-paced targeted research towards a common goal of optimizing predictive accuracy of a tool on an external test set.

There is no uniform prescription for building a universally optimal risk prediction tool. In the past, researchers often focused on a small set of standard risk factors for data cleaning and inclusion in their models, either for statistical reasons or grounds content; see Kattan *et al.*[8] for the American Joint Committee on Cancer (AJCC)'s criteria for a prognostic model. The ever more commonly performed indiscriminate data-dumps from multiple clinical trials bring forth additional challenges of signal discovery, data cleaning, and missing data adjustment. Today's data scientist has to decide which datasets to use for training the models versus which to hold out for testing, as well as how to use initial information from the target population to fine-tune the model. The Prostate Cancer DREAM Challenge provided participating teams with hands-on experience in these critical areas. Through our participation in the challenge, we experimented with hundreds of models, data inclusion and missing-value adjustment options. By the end of the process, three general principles stood out that proved crucial to success: good data, wisdom of the crowds, and recalibration. Herein, we illustrate these principles and quantify their impact.

Preliminaries

The goal of sub-challenge 1 was to develop a survival prediction model using data from three different clinical trials, which was to be validated on data from a fourth independent trial. Random subsets of data from the fourth validation trial were provided at multiple interim points to guide model construction. After trying several machine learning and statistical models, the combined Cox proportional hazards and lasso model was chosen as it performed optimally on the interim validation sets[9]. The Cox proportional hazards model specifies the mortality hazard rate for an individual with covariate vector x as:

$$\lambda(t \mid x) = \lambda_0(\mathrm{t})\exp(x'\beta),$$

where β is the vector of log hazard ratios for respective covariates comprising x, and $\lambda_0(t)$ is a baseline hazard function that is left unspecified, making the model semi-parametric and more flexible than fully-specified parametric survival models. The model follows proportional hazards since the ratio of hazards for an individual with a unit increase in a single covariate relative to another individual, with all other covariates fixed, equals $\exp(\beta)$, which is constant for all times t. The non-parametric Kaplan-Meier estimator shows the empirical distribution of the observed failure times subject to censoring. Inspection of whether the curves stratified by different covariate values remain separated across the length of follow-up can be used to informally assess whether the proportional assumption holds.

The standard method for estimating β in the Cox model is based on the partial likelihood that specifies for each individual their relative probability of failure compared to other individuals at risk:

$$L(\beta) = \prod_{r\in D}\frac{\exp(\beta' x_r)}{\sum_{j\in R_r}\exp(\beta' x_j)}.$$

In this formulation, D is the group of distinct death times observed in the study, and R_j denotes the risk set of all individuals still alive and on-study. If multiple individuals have the same death time, modifications are needed for the likelihood, which are implemented using a choice of algorithms.

Instead of finding the β that maximizes the log likelihood $\ell(\beta) = \log L(\beta)$ itself, the lasso (least absolute shrinkage and selection operator) maximizes it subject to the constraint that $\Sigma_j |\beta_j| < s$, where s is a user-selected tuning parameter. This modification heuristically keeps model dimensionality low, with unnecessary parameters shrunk to zero as necessary.

Sub-challenge 1 was again divided into two tasks: In sub-challenge 1a, participants were asked to predict risks of death. Sub-challenge 1b asked for the prediction of exact times until death. For evaluating the proposed prediction models on the withheld test data, two criteria were used, corresponding to sub-challenges 1a and 1b, respectively. The evaluation criterion for 1a focused on discrimination, that is how well the risk prediction model differentiated a patient about to experience mortality versus not. This criterion only compared the ranks of risk scores among groups of patients, with no further regard to accuracy in terms of actual values of risk scores. The second criterion of calibration focused on accuracy in terms of how close the exact time to event (death) was to the predicted time to event.

Receiver-operator-characteristics (ROC) curves have their origin in radar technology and signal processing and remain the standard of choice for determining the discrimination capability of a diagnostic test[10]. They have been most widely used for evaluating prediction models for binary disease outcomes based on retrospective case-control studies. In this context, the idea is that risk prediction tools return a probability between 0 and 1 of an individual having a disease, and any value, say c, could be used as a threshold for making a yes/no decision concerning whether the person is diseased, warranting further diagnostic work-up. A person with predicted risk exceeding c is labeled as testing positive for disease and a person with risk less than or equal to c as negative. Given a set of diseased cases and non-diseased controls, each with a predicted risk p_r, for every threshold c there exist two measures of correct prediction, one for the cases and one for the controls, respectively:

$$\text{Sensitivity}(c) = P(p_r > c|\text{Diseased}),$$
$$\text{Specificity}(c) = P(p_r \leq c|\text{Not Diseased}),$$

The ROC curve displays the sensitivity, also termed the true positive rate, against 1-specificity, also termed the false positive rate for all possible choices of c. The area under the ROC curve (AUC) can therefore be used as a metric for model evaluation and comparison. It may be interpreted as a concordance index, where a value of 1 (100%) represents perfect accuracy (i.e. sensitivity and specificity of 1) and a value of 0.5 equals random guessing.

For extension to prediction of survival up until fixed time periods that accommodate censored observations, Heagerty *et al.*[11] proposed time-dependent ROC curves using time-specific versions of sensitivity and specificity that were based on whether individuals still on study were alive (controls) versus not (cases) at each time t, yielding as a result a plot of AUC values versus time t. Hung *et al.*[12] provided non-parametric estimators for the time-dependent AUC and Blanche *et al.*[13] provided an R package `timeROC` that was used for evaluation in sub-challenge 1a. To arrive at a single measure, integrated AUCs from 6 to 30 months were calculated and referred to as iAUCs.

Calibration measures the accuracy of numerical predictions, answering the question of how close estimates are to the truth. For sub-challenge 1b, which aimed at predicting the time to event (actual day of mortality), the root mean squared error (RMSE) was used:

$$\text{RMSE} = \sqrt{\frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i (\hat{y}_i - y_i)^2},$$

where $\hat{y}$ is a vector of n predictions for all patients in the test set, y is the vector of n observed values (which equals NA in case death is not observed), D_i is a binary variable equal to one if death is reported and zero otherwise, and subscripts denote individual predictions and observed values on the test set. Thus, the RMSE was only calculated on patients with observed death times on study, and $\sum_{i=1}^{n} D_i$ referred to the number of death event times in the test set.

Methods

First concept: good data

Figure 1 gives an overview of the Prostate Cancer DREAM Challenge data after some cleaning (see low-cost strategy in paragraph below) but before inclusion of additional variables. There were six data tables available: one core table (the basis of Figure 1), containing baseline clinical covariates at patient level, and five longitudinal data tables, containing additional information at event level. We refer to the four trials as ASCENT-2 (Novacea, provided by Memorial Sloan Kettering Cancer Center[14]), VENICE (Sanofi[15]), MAINSAIL (Celgene[16]), and ENTHUSE-33 (AstraZeneca[17]). The majority of the variables (73.95%) in the core table have been measured in all four studies. Eight variables (albumin, magnesium, sodium, total protein, phosphorus, region and presence of target and non-target lesions) were exclusive to MAINSAIL, ENTHUSE-33 and VENICE while two (red blood cells and lymphocytes) were only assessed in ENTHUSE-33 and MAINSAIL. Lactate dehydrogenase was only measured in ASCENT-2, ENTHUSE-33 and MAINSAIL but not in VENICE. The presence of neoplasms and creatinine clearance were only present in VENICE and ENTHUSE-33. Unfortunately, the interesting variable gleason score was only reported in the ASCENT-2 study. With a p-value of 0.0017 it proved to be highly significant in a univariate Cox model for those patients where the variable was available, but was removed by us due to its missingness in the test dataset ENTHUSE-33. The significance of other variables which were missing in at least one trial is presented in Table 1.

In this section we compare two strategies to secure as many data elements as possible: a relatively straightforward low-cost minimal adaptation approach versus a high-cost strategy that incorporates

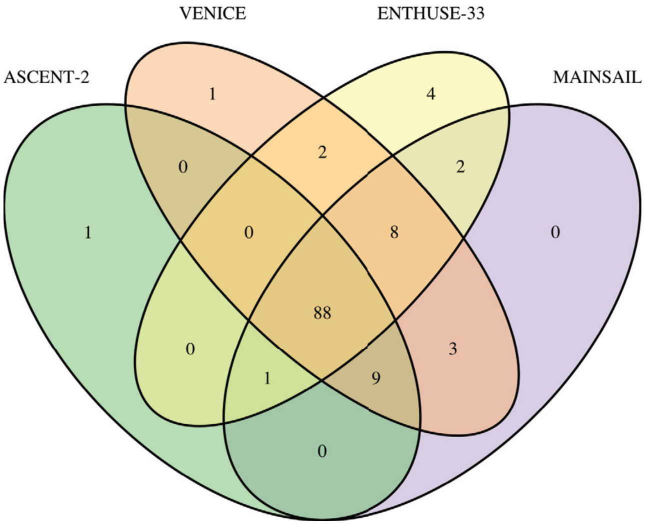

Figure 1. Variables available in the four trials ASCENT-2, MAINSAIL, VENICE, and ENTHUSE-33.

Table 1. **Significance in univariate Cox models for variables which were not available in at least one training study.**

Variable	p-value	Significance	in ASCENT-2	in MAINSAIL	in VENICE
Albumin	< 0.0001	***	no	yes	yes
Lactate dehydrogenase	< 0.0001	***	(no)	yes	yes
Red blood cells	< 0.0001	***	no	yes	no
Sodium	< 0.001	**	no	yes	yes
Phosphorus	< 0.05	.	no	yes	yes
Region North America	< 0.05	.	no	yes	yes
Region Western Europe	< 0.05	.	no	yes	yes
Target lesions	< 0.05	.	no	yes	yes
Region Other	0.0651		no	yes	yes
Region South America	0.0882		no	yes	yes
Non-target lesions	0.2155		no	yes	yes
Blood urea nitrogen	0.2163		no	yes	yes
Lymphocytes	0.2223		no	yes	no
Neoplasms	0.3242		no	no	yes
Total protein	0.3832		no	yes	yes
Calculated creatinine clearance	0.5784		no	yes	yes
Magnesium	0.7174		no	yes	yes
Glucose	0.7910		no	yes	yes

subject-matter knowledge into the procedure. The minimal adaptation approach followed recommendations typically provided in statistical packages. We excluded variables with more than 10% missing values in either the training or test set, while for variables with less than 10% missing values, we used imputation, replacing the missing values with the mean value among observations that were not missing. For the second more intensive strategy, we performed subject-matter informed data cleaning, such as including additional information from the event tables, preprocessing the data, including new variables such as principal components, a toxicity score and interaction effects. The extra effort for the second approach paid off in terms of substantially increasing validation accuracy on the external test set as shown in Table 3. Details of the second approach are provided below.

High-cost data cleaning and preprocessing. An essential component for developing the final predictions for both sub-challenges 1a and 1b was a comprehensive interdisciplinary exploration of the data. We built a cleaned and preprocessed dataset comprising information from the provided covariate and event tables as described in this section.

Cleaning of core table. In a first data cleaning, we identified incomplete (e. g. more than 70% missing values in either the training data or the test data), inconsistent (e. g. different levels between trials for categorical data) or irrelevant (e. g. the same value for all or almost all patients) covariables in the core table and modified the datasets accordingly: We unified categories for height, weight, race and region and removed variables with either very large fractions of missing values, redundant information or hardly any variability.

Event tables. We derived baseline patient information from the event tables as follows: The *PriorMed table* contained information about the medication that patients received prior to their participation in the clinical trials. Categorical assignments for medications were often missing, sometimes erroneous, and categories differed between trials. Based on our clinical expertise, we assigned appropriate categories to each medication. We then introduced new variables counting for each patient the number of medications from each category. Studies substantially differed in distributions of numbers of prior medications. We suspected that this was due to reporting biases. We hence scaled the new variables such that they had identical mean and variance across studies. The *MedHistory table* contained information about medical diagnoses that patients got prior to their participation in the clinical trials. For each patient, we counted the number of diagnoses in the various categories. We excluded categories which we assumed not to be clinically relevant for death or treatment discontinuation. We also deleted categories where diagnoses were reported for less than 2% of the training or test patients. From the *LesionMeasure table*, we extracted information such as the number of target and non-target lesions, counts per tissue and maximum target size. We noticed systematic differences in numbers of reported lesions between studies. We suspected that these differences were due to different reporting behaviour rather than different patient properties. In compliance with the guidelines by Eisenhauer *et al.*[18], we only used the five largest target lesions for covariable generation and limited the number of target lesions per tissue to two. From the *VitalSign table*, we used patient-specific information about pulse and blood pressure. From the *LabValue table*, we derived covariables with additional lab test results. Difficulties were different units and truncated lab values.

Preprocessing. There were a number of values that appeared to be outliers in the statistical sense. However, though being extreme, many of these values were clinically not impossible. In order to not throw away important information, we only removed values where hemoglobin was less than five or the prostate specific antigen or platelet count were equal to zero. For ASCENT-2, there was no event data on lesions. Hence, we set the variables for the presence of target or non-target-lesions to NA ("no information") rather than NO ("no lesions found"). We log-transformed the most skewed continuous variables (prostate specific antigen, alkaline phosphatase, aspartate aminotransferase, lactate dehydrogenase and testosterone). We included selected interactions of covariables in the model, based on the results of all pairwise Cox models with two main effects and an interaction. If the coefficient of the interaction term was larger than 0.1 in its absolute value, and the p-value of the coefficient was less than 0.05 after multiple testing correction, the combination was included in the list. From the final dataset, we removed variables such that afterwards all pairwise Pearson correlations were below 0.95 in absolute value.

Several covariables were generally observed in one or several of the studies but missing for single patients. We imputed these missing values with 5-fold multivariate imputations by chained equations (MICE) using the R package `mice`[19] with default settings, R version 3.2.1. This imputation approach has proven to be successful for a variety of cancer specific data[20–23].

New variables. We introduced a number of additional newly-derived variables to the set already described above: First, we aimed to represent the information from the large number of newly derived covariables from the event data tables by a smaller number. To that end, we performed a principal component analysis (PCA) once on the new variables from MedHistory and once on the new variables from LesionMeasure. We included the most important principal components as additional covariables until 95% of the variance was explained. The original variables derived from the event tables remained in the dataset as well. As a second measure, we introduced a toxicity score for each patient based on lab value information. In this variable, we combined all toxicity grades which were either provided in the LabValue table or which we derived from literature research, using databases from the U.S. Department of Health and Human Services, Food and Drug Administration (http://www.fda.gov/downloads/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/Vaccines/ucm091977.pdf), The International Clinical Studies Support Center (ICSSC, http://www.icssc.org/Documents/Resources/AEManual2003AppendicesFebruary_06_2003 final.pdf), and HSeT - Health Teaching Portal (http://hset.bio-med.ch/cms/Default.aspx?Page=12173). Third, as the reference method by Halabi *et al.*[24] was successful, we included their risk score as an additional covariable.

Second concept: wisdom of the crowd

Wisdom of the crowd philosophically asserts that a prediction gauged among a group of experts will be more accurate than any single prediction; the readable book by Surowiecki provides tantalizing historical and contemporary examples[25]. Wisdom of the crowds underpins the Sage Bionetworks DREAM challenge efforts behind crowdsourcing and citizen science, the opening of challenges to mass numbers of competitive teams on the internet or active members of the public, which has brought about improvements in breast cancer prognostic modeling among other efforts[26,27]. Wisdom of the crowds also underpins the accepted notion that ensembles of models confer better predictive accuracy than single models, are more robust than single methods, and have the added advantage of appropriately accounting for uncertainty[28]. The ability to test models on parts of the withheld test set influenced the choice of which models should be contained in the ensemble; one could term this supervised ensemble construction. We herein describe the approach.

Model averaging. In the first concept we described our multiple imputation approach for missing values. However, we noticed that distributions of variables differed between trials. Hence, we decided to only impute within the trials, not across. In other words, values were imputed based on covariable information only from patients within the same study.

Our question was then how to deal with variables that were (almost) completely missing in one entire training study but measured in other training studies. Our solution was to estimate seven different models, each taking into account a different subset of the three studies ASCENT-2, MAINSAIL and VENICE (see Figure 2). Depending on the set of studies, different covariables could be included in the model. For example, lactate dehydrogenase was completely missing in VENICE, but not in ASCENT-2 and MAINSAIL. It was hence excluded in every model based on VENICE but not otherwise. Table 2 contains two more examples. Once we had fixed the model and corresponding data, we jointly scaled the explanatory variables of all training and test studies to mean zero and variance one.

Third concept: recalibration

Recalibration of a model encompasses any manner of change to the model using data or information from the target model. In the Prostate Cancer DREAM Challenge, patient-level data from the test set did not include the target variables. Recalibration was still possible as described in the following.

High-risk and low-risk recalibration (sub-challenge 1a). With the averaged Cox model described in the previous section, we expected to predict the risks for "average patients" satisfyingly well. For high-risk or low-risk patients, however, we aimed to further improve the predictions. Hence, we adapted the scores by estimating two more models: (i) a high-risk model, where we modified the target variable DEATH (indicating whether death had been observed) such that it only counted events that happened prior to 14 months, and (ii) a low-risk model, where we only considered events that occurred after 18 months. We then recalibrated the risk scores for the

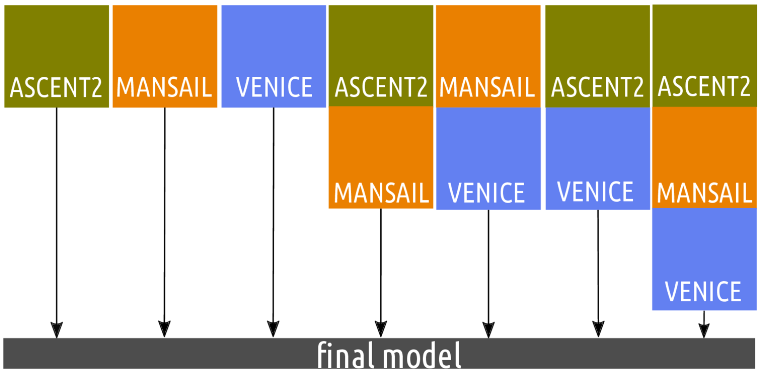

Figure 2. Datasets contained in different models to be averaged.

Table 2. Datasets contained in different models to be averaged, exemplified on selected variables.

Model name Studies included	1 ASCENT-2	2 MAINSAIL	3 VENICE	12 ASCENT-2, MAINSAIL	13 ASCENT-2, VENICE	23 MAINSAIL, VENICE	123 ASCENT-2, MAINSAIL, VENICE
Lactate dehydrogenase	yes	yes	no	yes	no	no	no
Albumin	no	yes	yes	no	no	yes	no
Prostate specific antigen	yes	yes	yes	yes	yes	yes	yes

Table 3. Comparison of prediction performance for minimal-adaptation vs. high-cost data preprocessing.

	Minimal-adaptation (standard) data	High-cost (improved) data
iAUC (1a)	0.7535	0.7642
RMSE (1b)	304.79	292.15

following patients: (i) For patients with risk score above the median, we calculated the average between the initial prediction and the high-risk score and considered this as the new risk score, and (ii) for those patients whose risk score was below the 25-percentile, we calculated the average between the initial model and the low-risk model. In both cases, we made sure that the modifications only altered the ranks of patients within the defined ranges, i.e. above the median and below the 25-percentile with respect to the initial risk score. Figure 3 shows the former (x-axis) vs. the new rank (y-axis) for each patient, where a low rank means a low risk of dying.

Quantile recalibration (sub-challenge 1b). As described above, we estimated a Cox model with lasso regularization. Based on the estimated coefficients from the training datasets, we predicted a survival curve for each of the patients in the test data. From each survival curve we derived a point estimate for the time of death as follows: A typical estimate would have been the median. However, in the training data this estimate was not optimal with respect to RMSE. We hence determined from the training data the value of α such that

$$\sum_{i=1}^{n} D_i (Q_{\alpha i} - y_i)^2$$

was minimized. In this formula, n is the number of patients in the training data, $Q_{\alpha i}$ denotes the $\alpha \cdot 100\%$-quantile in the survive curve for patient i, y_i is the observed time of death for patient i (can be any value if death is not observed), and $D_i = 1$ is the indicator that death of patient i has been observed (otherwise $D_i = 0$). The resulting value of α was 0.69. We hence derived the 69%-quantiles from the survival curves as final prediction as illustrated in Figure 4.

Validation by calibration (sub-challenge 1b). In addition to the above calibration of times to event, we also applied the *validation-by-calibration* method by Van Houwelingen[29] in sub-challenge 1b. This method adjusts the original predictions by rescaling them to the range of the observed outcomes using linear regression. The original method splits the training data into two subsets for model

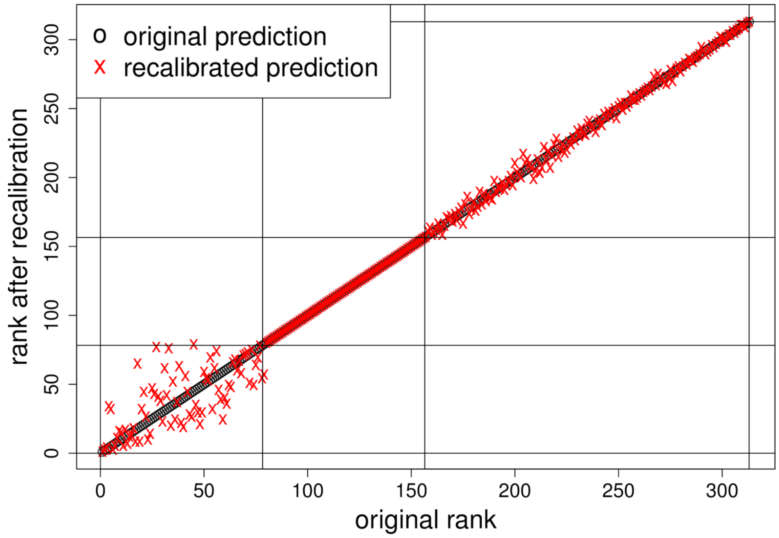

Figure 3. Recalibration for high and low risks: former (x-axis) vs. the new rank (y-axis) for each patient, where a low rank means a low risk of dying.

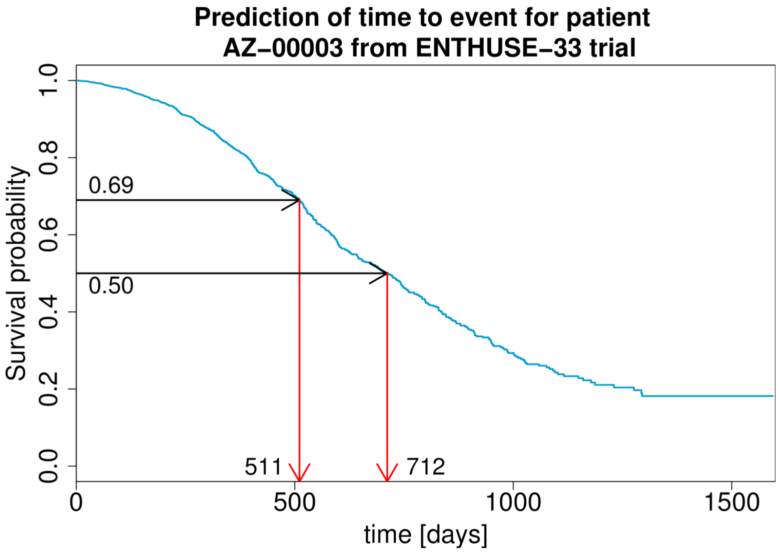

Figure 4. Time to event prediction via 69%-quantile vs. 50%-quantile for one selected patient.

building and validation. For computational reasons, we omit this step. Adapted to the context here, validation-and-calibration works as follows:

1. Estimate a Cox-lasso model based on the three training trials (ASCENT-2, MAINSAIL, VENICE) as described above. From this model, compute survival curves and estimate the times to event for each patient in the training data. Let $\hat{y}$ be the predictions for those patients where death was observed, and y be the corresponding observed times of death. From the same model, estimate the times to event for all patients in the test set (ENTHUSE-33) and denote them by $\hat{z}$.

2. Plot y versus $\hat{y}$ and decide whether a linear relationship of the two variables can be assumed. If so, proceed.

3. Estimate a linear model $y = \beta_0 + \beta_1\hat{y} + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the estimated intercept and slope coefficient.

4. Recalibrate the predictions for the test patients to $\hat{z}_c = \hat{\beta}_0 + \hat{\beta}_1 \hat{z}$.

Figure 5 illustrates the procedure on our training data.

Results

We applied the three general concepts to the prediction problems of sub-challenges 1a and 1b. The benefit of applying each of the principles on the iAUC and RMSE in the Prostate Cancer DREAM Challenge is quantified in Figure 7 and Figure 8, respectively. Details are given in the following.

Impact of good data

In order to assess the gain of the elaborate data preprocessing as compared to the low-cost minimal adaptation approach, we predicted the risk of death (sub-challenge 1a) and the time to death (sub-challenge 1b) for both data preparations. Table 3 shows the respective validation measures iAUC and RMSE when a Cox model with lasso regularization is applied as described above. For sub-challenge 1b, we used median survival times from the estimated survival curves. Prediction improved substantially for the high-cost data preparation with respect to both measures: The iAUC (sub-challenge 1a) increased by more than 0.01 units from 0.7535 to 0.7642. The RMSE (sub-challenge 1b) decreased by more than 10 units from 304.79 to 292.15.

Impact of wisdom of the crowd

We estimated the seven models described in the model averaging section above (see also Figure 2) on the training data and got a risk prediction for the test data for each of these. We then took the average of the seven predictions (each of which was again an average over five imputed datasets) to arrive at a final risk score. Compared to the standard approach (no splitting into submodels), this model averaging approach yielded improvements in terms of iAUC and RMSE measures. This is shown in Table 4 where both the standard approach model and the averaging approach employ the improved data as described in the first concept. While the increase in iAUC is again around 0.1 units from 0.7642 to 0.7733, the improvement of the prediction of time to event is considerably

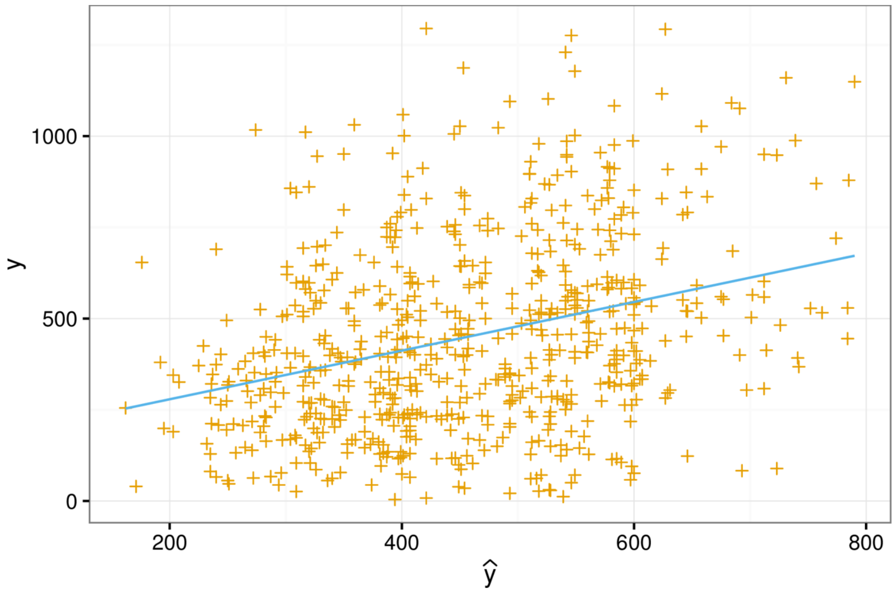

Figure 5. Linear relationship between predicted and observed times to event for the training data, used as a basis for the validation-by-calibration method.

Table 4. Comparison of prediction performance for standard approach vs. model averaging.

	Standard approach	Model averaging
iAUC (1a)	0.7642	0.7733
RMSE (1b)	292.15	263.37

Table 5. Effect of low-risk and high-risk calibration on iAUC in sub-challenge 1a.

iAUC	w/o low-risk calibration	w/ low-risk calibration
w/o high-risk calibration	0.7642	0.7668
w/ high-risk calibration	0.7642	0.7668

more dramatic as it decreases the RMSE by almost 30 units from 292.15 to 263.37.

Impact of recalibration

We applied the three proposed recalibration techniques to our predictions for risk of death (sub-challenge 1a) and time of death (sub-challenge 1b) and validated the effects of these measures on the test data (ENTHUSE-33).

Sub-challenge 1a. For the risks of dying, we once applied the low-risk calibration only, the high-risk calibration only, and both measures simultaneously. Table 5 summarizes the results. It shows that neither the low-risk nor the high-risk calibration had a substantial effect on the prediction performance in terms of iAUC: The low-risk calibration led to a small increase of iAUC by approximately 0.003 units from 0.7642 to 0.7668. The high-risk calibration did not improve the prediction accuracy at all, although the ranks of patients changed.

Sub-challenge 1b. Recalibration of times to event caused a highly convincing improvement of prediction accuracy. Table 6 shows RMSE values for the 69%-quantile recalibration only, for Van Houwelingen's validation-by-calibration approach only, and for the two measures combined (i. e. first applying quantile recalibration and then the validation-by-calibration method). All recalibration approaches decreased the RMSE substantially by as much as around 100 days as compared to the non-calibrated predictions.

The choice of $\alpha = 0.69$ for the α-quantile recalibration had resulted from the training data only. As a further post-challenge analysis, we investigated whether this was also a good choice for the test dataset. Figure 6 shows RMSEs for the α-quantile recalibration as well as the combination of quantile recalibration and validation-by-calibration for a grid of α values between 0.6 and 0.8. On this grid, $\alpha = 0.72$ was the optimal choice when applying the quantile recalibration only, but $\alpha = 0.69$ was also reasonable. When followed by validation-by-calibration, the effect of α was hardly visible anymore. This makes validation-by-calibration an appealing approach for the prediction of times to event.

Conclusion

As data, computation, and statistical methods reach new horizons for the clinical risk prediction dreamers, this study reminds us of some timeless basics we should not forget: good data, wisdom of the crowds and recalibration. In this study we translated

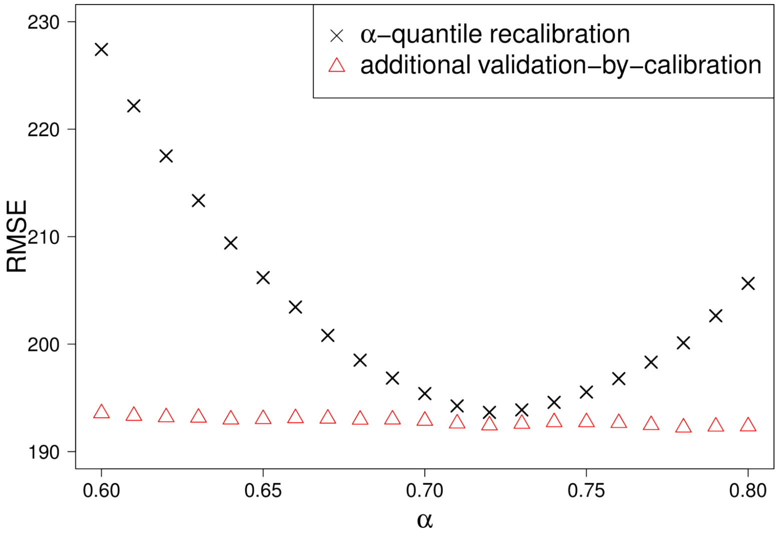

Figure 6. Effect of α-quantile recalibration (followed by validation-by-calibration or not) on RMSE for varying α.

Table 6. Effect of 69%-quantile recalibration and validation-by-calibration on RMSE in sub-challenge 1b.

	w/o quantile calibration	w/ quantile calibration
w/o validation-by-calibration	292.15	196.84
w/ validation-by-calibration	194.18	192.99

and enhanced these principles for use in developing survival risk prediction tools based on multiple heterogeneous clinical trials with large and non-overlapping sets of covariates. The impact of individual components of our proposed strategy can be quantified by their incremental influence on the assessment criteria.

The AUC is the most widely used endpoint for measuring the discrimination capability of a biomarker or risk prediction tool. However, it is limited by a lack of clinical relevance for the individual patient, defined as a comparative probability of ranks for pairs of patients, as well as a lack of statistical power being based on ranks, making it insensitive (it is invariant to monotonic changes) and notoriously difficult to budge[30]. Accordingly, Figure 7 shows small gains of 0.0105 points for improved data, 0.0091 additional points for model averaging, and 0.0004 additional points for recalibration, taking the best-performing option for each principle. The bottom line for implementing the three principles was to increase the iAUC from 0.7535 to 0.7768, a minor improvement, but comparable to laudable improvements in published risk prediction tools given the robust nature of the iAUC.

The RMSE measures accuracy of a risk prediction, in other words how close a projected risk is to what actually happened to the patient, and based on the continuous measures of risk, has greater statistical power to detect differences due to technical improvements. Accordingly, more significant gains are more readily apparent in the RMSE in Figure 8, with 12.64 points for improved data, 28.78 additional points for model averaging or 99.01 additional points for recalibration, resulting in a net reduction from 111.57 points on the square root prediction scale after implementation of all three principles. Missing a large gain such as this by hastily fitting a single model without regard to the data, without averaging and without recalibration, would have cost us the challenge. But, most importantly, skipping these time-consuming basics would result in a less accurate prognosis for the individual patient.

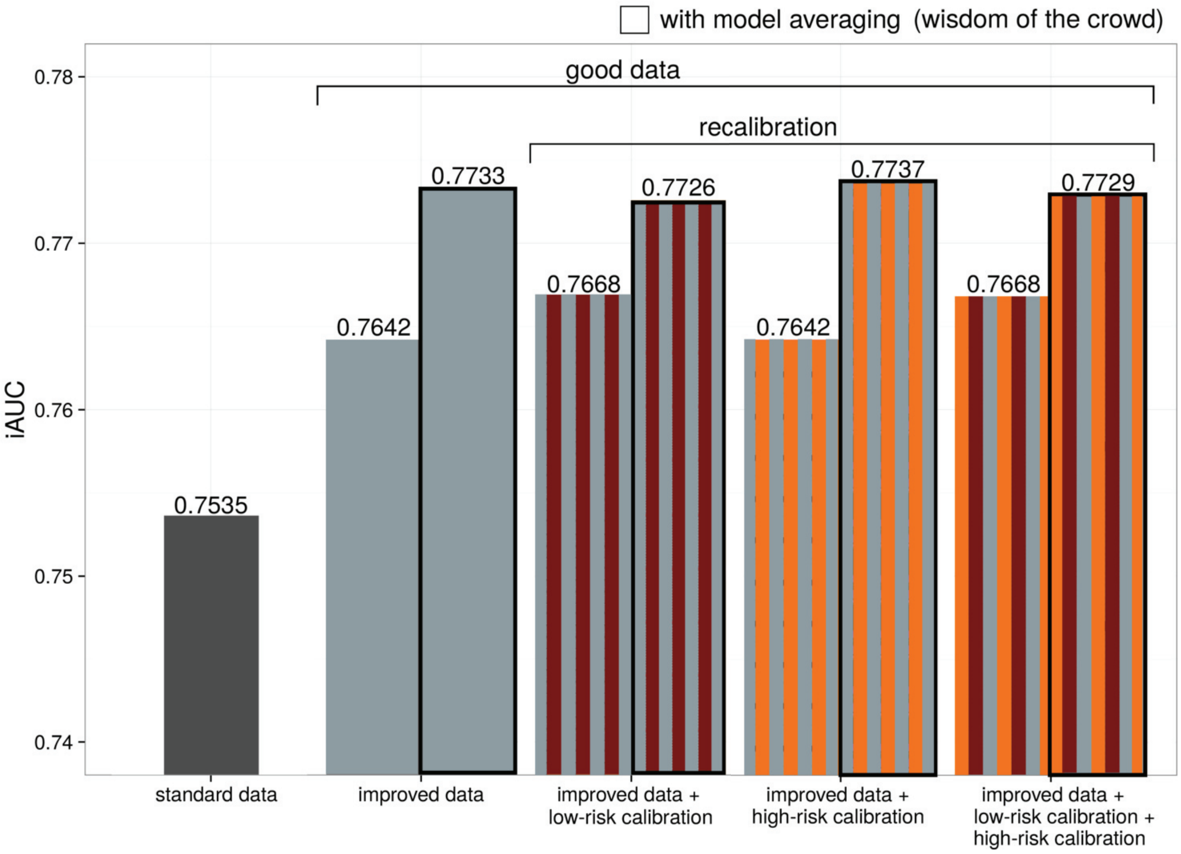

Figure 7. iAUC values resulting from different combinations of core principles to sub-challenge 1a.

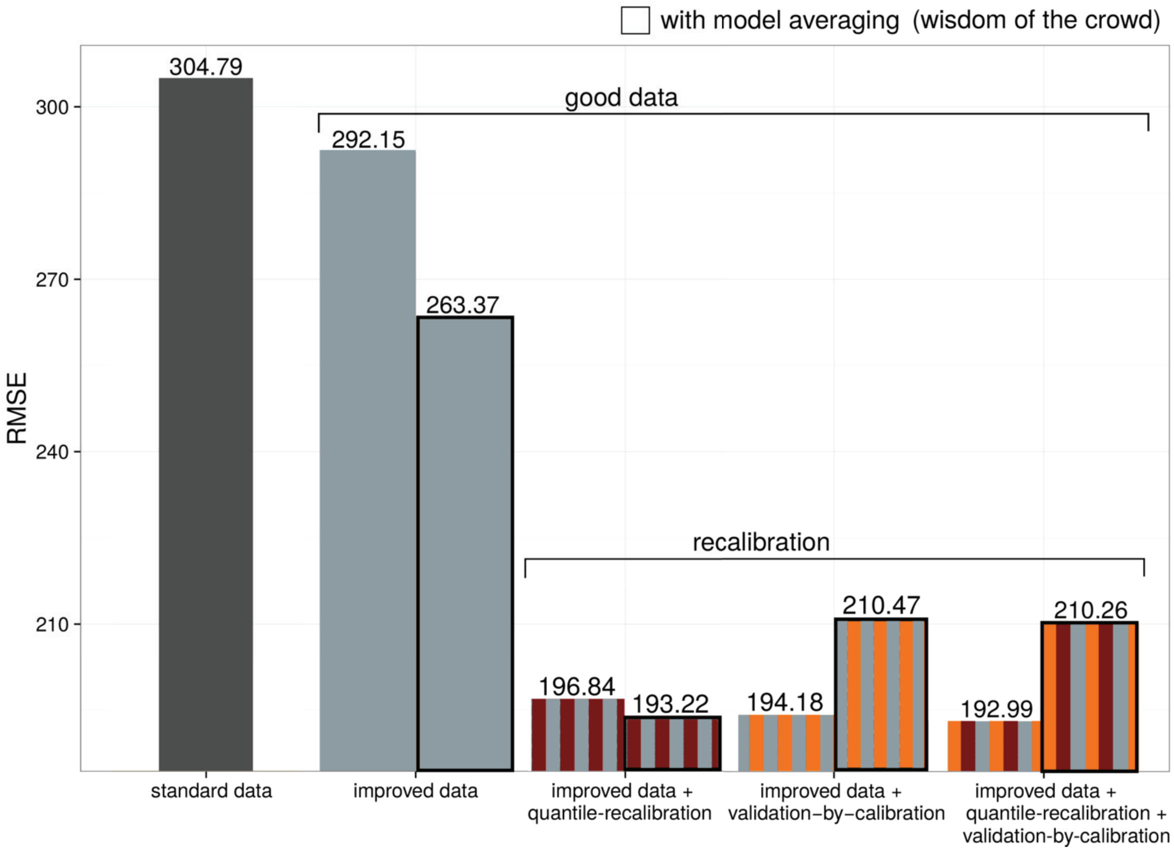

Figure 8. RMSE values resulting from different combinations of core principles to sub-challenge 1b.

Our data preparation included the generation of additional clinical variables. Post-challenge analyses showed that the newly introduced toxicity score was especially beneficial for good predictions in all sub-challenges, and so were the variables derived from the event data tables on lesion measures. We thus propose to generally capture such information in any clinical trials on prostate cancer. As more data become publicly available as a resource for expanding clinical risk tools, it becomes tempting to think that the art of risk prediction can be automated, eliminating the need for interdisciplinary scientists to work together. This study concludes that interdisciplinary subject-matter knowledge remains essential and that building optimal risk prediction tools remains as much an art as a process.

Data availability

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

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

The code and documentation underlying the method presented in this paper can be found at: http://dx.doi.org/10.7303/syn5592405[31]

Author contributions

CF, IK, CK and JS preprocessed the data. CF, IK, NK, CK, ML, HS and JS established first analyses. IK, NK, CK and ML performed in-depth analysis. CF, IK and ML proposed novel modeling refinements. PD assisted the team with respect to clinical questions. DPA, CF and HS advised the team with respect to statistical questions. IK was responsible for code integration. DPA and CF wrote the manuscript, with contributions from IK, NK, CK, ML and JS. All members proofread the manuscript. CF supervised and guided the work. All authors contributed to discussions and decision-making processes.

Competing interests

No competing interests were disclosed.

Grant information

CF and IK are supported by the European Union within the ERC grant LatentCauses. CF is supported by the German Research Foundation (DFG) within the Collaborative Research Centre 1243, Subproject A17.

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

Acknowledgements

This publication is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this publication. We are grateful to Sage Bionetworks, the DREAM organization, and Project Data Sphere for developing and supplying data for the Challenge. We thank Fabian Theis and Rolf Holle for supporting us to participate in the Challenge.

References

1. Koenig F, Slattery J, Groves T, *et al.*: **Sharing clinical trial data on patient level: opportunities and challenges.** *Biom J.* 2015; **57**(1): 8–26.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Halabi S, Small EJ, Kantoff PW, *et al.*: **Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer.** *J Clin Oncol.* 2003; **21**(7): 1232–1237.
PubMed Abstract | Publisher Full Text

3. Thompson IM, Ankerst DP, Chi C, *et al.*: **Assessing prostate cancer risk: Results from the prostate cancer prevention trial.** *J Natl Cancer Inst.* 2006; **98**(8): 529–534.
PubMed Abstract | Publisher Full Text

4. Chen M, Shi L, Kelly R, *et al.*: **Selecting a single model or combining multiple models for microarray-based classifier development?--a comparative analysis based on large and diverse datasets generated from the MAQC-II project.** *BMC Bioinformatics.* 2011; **12**(Suppl 10): S3.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Hede K: **Project data sphere to make cancer clinical trial data publicly available.** *J Natl Cancer Inst.* 2013; **105**(16): 1159–60.
PubMed Abstract | Publisher Full Text

6. Rozengauz DE: **[Tumor of the left temporal lobe of the brain simulating an otogenic abscess].** *Zh Ushn Nos Gorl Bolezn.* 1965; **25**(3): 83–4.
PubMed Abstract

7. Guinney J, Wang T, Laajala TD, *et al.*: **Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data.** *Lancet Oncol.* 2016; published online Nov 15.
Publisher Full Text

8. Kattan MW, Hess KR, Amin MB, *et al.*: **American Joint Committee on cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine.** *CA Cancer J Clin.* 2016.
PubMed Abstract | Publisher Full Text

9. Tibshirani R: **The lasso method for variable selection in the Cox model.** *Stat Med.* 1997; **16**(4): 385–95.
PubMed Abstract | Publisher Full Text

10. Metz CE: **Basic principles of ROC analysis.** *Semin Nucl Med.* 1978; **8**(4): 283–298.
PubMed Abstract | Publisher Full Text

11. Heagerty PJ, Zheng Y: **Survival model predictive accuracy and ROC curves.** *Biometrics.* 2005; **61**(1): 92–105.
PubMed Abstract | Publisher Full Text

12. 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

13. Blanche P, Dartigues JF, Jacqmin-Gadda H: **Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks.** *Stat Med.* 2013; **32**(30): 5381–5397.
PubMed Abstract | Publisher Full Text

14. 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

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

16. 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

17. 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

18. Eisenhauer EA, Therasse P, Bogaerts J, *et al.*: **New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1).** *Eur J Cancer.* 2009; **45**(2): 228–247.
PubMed Abstract | Publisher Full Text

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

20. Clark TG, Altman DG: **Developing a prognostic model in the presence of missing data: an ovarian cancer case study.** *J Clin Epidemiol.* 2003; **56**(1): 28–37.
PubMed Abstract | Publisher Full Text

21. Royston P, Parmar MK, Sylvester R: **Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer.** *Stat Med.* 2004; **23**(6): 907–926.
PubMed Abstract | Publisher Full Text

22. Barosi G, Bergamaschi G, Marchetti M, *et al.*: **JAK2 V617F mutational status predicts progression to large splenomegaly and leukemic transformation in primary myelofibrosis.** *Blood.* 2007; **110**(12): 4030–4036.
PubMed Abstract | Publisher Full Text

23. Fernandes AS, Fonseca JM, Jarman IH, *et al.*: **Evaluation of missing data imputation in longitudinal cohort studies in breast cancer survival.** *Int J Knowl Eng Soft Data Paradig.* 2009; **1**(3): 257.
Publisher Full Text

24. 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

25. Surowiecki J: **The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.** Doubleday. 2004.
Reference Source

26. Bain R: **Citizen science and statistics: Playing a part.** *Significance.* 2016; **13**(1): 16–21.
Publisher Full Text

27. McCarthy N: **Prognostic models: rising to the challenge.** *Nat Rev Cancer.* 2013; **13**(6): 378.
PubMed Abstract | Publisher Full Text

28. Hoeting JA, Madigan D, Raftery AE, *et al.*: **Bayesian model averaging: A tutorial.** *Stat Sci.* 1999; **14**(4): 382–417.
Publisher Full Text

29. van Houwelingen HC: **Validation, calibration, revision and combination of prognostic survival models.** *Stat Med.* 2000; **19**(24): 3401–3415.
PubMed Abstract | Publisher Full Text

30. Ware JH: **The limitations of risk factors as prognostic tools.** *N Engl J Med.* 2006; **355**(25): 2615–2617.
PubMed Abstract | Publisher Full Text

31. Kondofersky I, Laimighofer M, Kurz C, *et al.*: **A Bavarian Dream: Methods for Challenges 1a, 1b and 2.** *Synapse Storage*, 2016.
Publisher Full Text