THE LANCET Oncology

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Guinney J, Wang T, Laajala TD, et al, and the Prostate Cancer Challenge DREAM Community. 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. http://dx.doi.org/10.1016/S1470-2045(16)30560-5.

Supplement to: Guinney J, Wang T, Laajala TD, et al. A prognostic model to predict overall survival for patients with metastatic castration-resistant prostate cancer: results from a crowdsourced challenge using retrospective, open clinical trial data.

Clinical trial descriptions, curation, and splitting
Challenge design, rules, and web-based resources
Evaluation of the top-performing team
Top-performing model description
Data-driven network projection for the ePCR model

Supplementary Tables

Supplementary Table 1. Full results from all 50 teams plus the Reference model across several scoring metrics from the Challenge. Performance measures were evaluated using the ENTHUSE 33 trial. Teams are listed with the links to their predictions, methods write-up, and code.

Supplementary Table 2. Comparison of risk stratification of patients in the ENTHUSE 33 trial by the ePCR and Reference models. Patients were dichotomized at median risk scores. All intervals reported are 95% confidence intervals. PPV = positive predictive value, NPV = negative predictive value. Values for Cases, Survivors, and Censored are cumulative.

Supplementary Table 3. Top 15 single and interacting variables from the final ePCR model built from the MAINSAIL and VENICE trials. Comprehensive list of evaluated variables is available at: https://www.synapse.org/#!Synapse:syn7113819

Supplementary Figures

Supplementary Figure 1. Overview of the top-performing ePCR method in comparison to the Reference model (Halabi model). (A) The benchmarking Reference model explored the LASSO model ($\alpha = 1$) in a training data cohort with respect to the regularization parameter (λ) using cross-validation (CV). (B) The top-performing ePCR approach is based on an ensemble of Penalized Cox Regression models (ePCR), which are optimized separately for each cohort or a combination of cohorts in terms of the regularization parameter (λ) as well as the full range of the L1/L2 regularization parameter ($0 <= \alpha <= 1$). The optimal model was identified with low values of α, indicating that the Ridge Regression ($\alpha = 0$)-like models performed better for modeling the complex interactions than the benchmarking Reference LASSO-model ($\alpha = 0$). (C) Ensemble predictions were generated by averaging over the predicted risk ranks from each ensemble component.

Supplementary Figure 2. (A) All data across ASCENT2, MAINSAIL, VENICE, and ENTHUSE 33– both binary and continuous data – were used in a PCA. (B) All data across the 4 studies – only binary variables – were used in PCA.

Supplementary Figure 3. (A) Density plot of follow-up times per study for the ASCENT2, MAINSAIL, VENICE, and ENTHUSE 33 trials. (B) Survival profile for each of the trials.

Supplementary Figure 4. Summary of Challenge results across all 50 teams plus the Reference model evaluated using the ENTHUSE 33 dataset. (A) Performance of submissions. Each submission underwent 1,000 paired bootstrap of final scoring patient set to calculate a Bayes factor against the top-performer a Bayes factor against the Reference model. A p value was calculated from randomization test of 1000 permutations. X-axis is iAUC and y-axis is submissions ranked by iAUC from high to low. Each team's bootstrapped iAUC scores are shown as horizontal boxplot with the black diamonds representing the point estimate of a team's performance. The colored boxes show the inter-quartile ranges and the whiskers extend to 1.5 times the corresponding interquartile ranges. Top-performer is colored in orange, other teams within Bayes factor of 20 were labeled in blue, and the rest of the

teams were labeled in green. The Reference model is labeled in purple. (B) Bayes factors of all submissions against the top-performer are shown. Bayes factors greater than 20 were truncated to 20. (C) Bayes factors of all submissions against the Reference model. Bayes factors greater than 20 were truncated to 20.

Supplementary Figure 5. Calibration plots for the ePCR model of predicted survival probability versus true survival proportion for the ENTHUSE 33 dataset at 18, 24, 30, and 36 months.

Supplementary Figure 6. Timeline for the Challenge. Five submissions were allowed per round, and only a single submission for the final validation round.

Supplementary Figure 7. Most frequently utilized variables by teams to build their final models using the ASCENT2, MAINSAIL, and VENICE trials. Results are self-reported from a post-Challenge survey over 40 teams. * variables are not used in the Reference model.

Clinical trial description, curation, data splitting

Three datasets were used to create the training dataset for the Challenge (Novacea ASCENT2[1], Sanofi VENICE[2], and Celgene MAINSAIL[3]), while one dataset (AstraZeneca ENTHUSE 33[4]) was held back for leaderboard and blinded validation. The data represented 2,070 first line mCRPC patients in four cancer trials, where all patients received docetaxel treatment in the comparator arm.

In order to perform further validation of the top-performing prognostic model algorithm, the organizing team identified a fifth trial dataset (AstraZeneca ENTHUSE M1[5]) as an independent validation dataset post-Challenge.

Due to the regulation and privacy environment of certain countries, not all patients in the comparator arm from ENTHUSE 33 and M1 were provided to PDS.

ASCENT2 (Novacea, provided by Memorial Sloan Kettering Cancer Center): ASCENT2[1] is a randomized, open-label study evaluating DN-101 in combination with docetaxel in mCRPC. Patients received docetaxel and calcitriol in comparator arm (N = 476; 138 events). Detailed inclusion/exclusion criteria is described on page 2192 from the published study.

VENICE (Sanofi): VENICE[2] is a randomized, double-blind study comparing efficacy and safety of aflibercept versus placebo in patients treated with docetaxel / prednisone for mCRPC. Patients received docetaxel, prednisone, and placebo in comparator arm (N = 598; 433 events). Detailed inclusion/exclusion criteria is described on pages 761-762 from the published study.

MAINSAIL (Celgene): MAINSAIL[3] is a randomized, double-blind study to evaluate efficacy and safety of docetaxel and prednisone with or without lenalidomide in patients with mCRPC. Patients received docetaxel, prednisone, and placebo in comparator arm (N = 526; 92 events). Detailed inclusion/exclusion criteria is described on page 418 from the published study.

ENTHUSE 33 (AstraZeneca): ENTHUSE 33[4] is a randomized, double-blind study to assess efficacy and safety of 10 mg ZD4054 combined with docetaxel in comparison with docetaxel in patients with mCRPC. Patients received docetaxel and placebo in comparator arm (N = 470; 255 events). Detailed inclusion/exclusion criteria is described on page 1741 from the published study.

ENTHUSE M1 (AstraZeneca): ENTHUSE M1[5] is a randomized, double-blind study to assess efficacy and safety of 10 mg ZD4054 versus placebo in patients with CRPC and bone metastasis who are pain free or mildly symptomatic. Patients received only placebo in comparator arm (N = 266; 133 events).

The original datasets from PDS contained patient level raw tables that conformed to either Study Data Tabulation Model (SDTM) standards or company-specific clinical database standards. In an effort to optimize the use of these data for the Challenge, four sets of raw trial data first needed to be consolidated into one set of standardized raw tables.

During initial analysis scoping, key SDTM domains were identified as targets for standardization because they covered majority of necessary information for study subjects. These domains included demographics, trial design, follow-up including survival outcomes, treatment history, lab and lesion measurement, and vital sign. The curation team converted data from each study into a common structure that then can be combined into one dataset for each domain (SDTM). Major efforts were carried out to standardize reference date, capture, and validate survival information through careful evaluation of the data, protocol, and clinical report form (CRF). Lab test names and units could vary; the way information was presented in its original structure could be dramatically different as well. Some studies came with a single table for lab, others used 6-8 tables to capture the same level of information. However, this standardization phase was critical to ensure robustness of the Challenge data.

Once standardized raw tables were in place, clinically important baseline covariates and dependent variables relevant to the draft research questions were then created to form the "Core Table". A list of prostate cancer related prognostic factors was pre-identified through literature review. The analysis expanded beyond the list to cover more than 150 variables including patient demographics, risk factors, functional status, prostate cancer treatment history, concomitant medicine, prevalent comorbidity, and condition by body system, major hematology/urology test, lesion measure/location, and vital sign. Variable creation was intended to be extensive yet not exhaustive to encourage independent thinking from the DREAM community.

Six data tables were released for this Challenge. The Core Table was the main table that was summarized at the patient level with dependent variables and clinical covariates. The remaining five tables were standardized raw event-level tables (lab, lesion, prior medicine, medical history, vital sign) used to create the Core Table that was at the event level and could be used for additional variable creation and/or exploration.

Challenge design, rules, and web-based resources

The Challenge was hosted on Synapse (www.synapse.org), a cloud-based platform for collaborative scientific data analysis. Synapse was used to allow access to Challenge data and to track participant agreements to the appropriate data use agreements (https://www.synapse.org/#!Synapse:syn3348040) and the Challenge rules (https://www.synapse.org/#!Synapse:syn3348041).

The Challenge was designed to have several rounds, including real-time leaderboard rounds and a final scoring round. A timeline for the Challenge can be found in Supplementary Figure 6. The leaderboard rounds provided teams the ability to build their models, make predictions, submit their predictions, and get real-time feedback on their performance. A total of three leaderboard rounds were run and teams were limited to five submissions per leaderboard round. For every submission made, an email was returned to the team with several performance metrics, including the iAUC, concordance index, and the AUC for 12, 18, and 24 months. At the end of a leaderboard round, a public leaderboard was updated with the best team score for that round.

For final submissions to the final scoring round, Challenge participants created Synapse projects containing predictions from their best model together with the code used to derive them and wikis in which participants describe their methods in text and figures. Teams were only allowed one submission to the final scoring round. To ensure reproducibility of the Challenge results, the Challenge organizers ran the code of the best performing methods and reviewed team write-ups. Team scores were not released until the top performing models were verified to reproduce the predictions that the team submitted. After the final method vetting, final scores were posted publicly on the final scoring leaderboard (Supplementary Table 1).

The ASCENT2, MAINSAIL, and VENICE datasets were used as training datasets, while the ENTHUSE 33 dataset was used as the validation dataset. The ENTHUSE 33 dataset was split in a non-overlapping manner into one 157-patient leaderboard set and one 313-patient final scoring round set. To choose this separation, we generated 100 random splits and manually chose one that yielded moderately different performance accuracy between the two sets. The 157-patient leaderboard set was further split into three overlapping smaller sets for the three leaderboard rounds. Each smaller set had 126 patients. We chose the three groups by generating 100 random splits and manually chose three that were dissimilar in patient membership and each yielded a moderate difference in performance accuracy between the chosen 126 patients and the other 31 patients. Together the three groups covered the whole set of 157 patients in the leaderboard set.

Evaluation of the top-performing team

Teams were evaluated using several criteria to rank and determine the top-performing team(s). Principally, we were interested in the three following evaluations: a team's prediction is meaningfully (a) better than random, (b) better than the existing Reference model, and (c) better than the next best performing team.

Both (b) and (c) were evaluated using the Bayes factor measurement[6,7]. To calculate the Bayes factor, we used paired bootstrap sampling of the final set of patients 1,000 times and scored each new sample using the designated scoring metrics to obtain a distribution for each submission. Using these distributions, we tested the hypothesis H1 (defined as submission A is better than submission B) versus H0 (defined as submission A is no better than submission B). To be more specific, the Bayes factor of submission B versus submission A is calculating as the posterior probability of H1 as the fraction of bootstrap replications in which submission A is better than submission B divided by the posterior probability of H0 as the fraction of bootstrap replications in which submission A is no better than submission B. The Bayes factor will decide against H0 if the calculated posterior odds is larger than a pre-specified cutoff (three in this Challenge).

Better than random. To assess whether team predictions were better than random (a), a team's score was compared against an empirical null distribution from 1,000 resamplings of the dependent variable. One-sided p values were computed and corrected for multiple testing using the Benjamini-Hochberg procedure.

Better than the Reference model. The prognostic model from Halabi, et al[8] was used as the Reference model for predicting OS in mCRPC. The Reference model consists of 8 clinical variables: ECOG performance status, disease site, opioid analgesic use, LDH > 1 x ULN, albumin, hemoglobin, PSA, and alkaline phosphatase. Beta coefficients used in implementing this model were obtained from hazard ratios as reported in Table 2 from Halabi, et al.[8] The Bayes factor was calculated from 1,000 resamplings to compare the Reference model against each submission.

Better than next-best performer. We compared each submission against the top-performing submission using the Bayes factor, calculated using 1,000 resamplings. Submissions within Bayes factor < 3 from the top-performing team were declared indistinguishable from each other. In this Challenge, the top-performing team had a Bayes factor > 3 for when compared to all other teams.

In addition to the above listed evaluation methods, we evaluated the top-performing (ePCR) method using Kaplan Meier curves with patients stratified on median risk score. For the ePCR model, the high risk group was defined as score > 0.487 and low risk group as score ≤ 0.487 for the ENTHUSE 33 dataset. For the Reference model, the high risk group was defined as score > 1.05 and low risk group as score ≤ 1.05 for the ENTHUSE 33 dataset. The log rank test was used to statistically compare the high and low risk groups. Further analysis between the ePCR and Reference model was done using the ENTHUSE M1 dataset in the same manner as was done with ENTHUSE 33. For the ePCR model, the high risk group was defined as score > 0.501 and low risk group as score ≤ 0.501 for the ENTHUSE M1 dataset. For the Reference model, the high risk group was defined as score > 0.80 and low risk group as score ≤ 0.80 for the ENTHUSE M1 dataset. The log rank test was used to statistically compare the high and low risk groups.

Top-performing method description

The key phases of the team FIMM-UTU method included*:* (*i*) processing of raw data input, imputation of missing values, filtering, and truncation; (*ii*) utilizing unsupervised learning to identify most relevant patterns in the training datasets; (*iii*) fitting study-wise optimized penalized Cox regression models; and (*iv*) constructing the ensemble collection of study-wise optimized components for performing the final predictions.

(*i*) In addition to the refined Core Table provided by the Challenge organizers, a number of additional variables were manually extracted from the available additional data tables, namely the vital signs and lab values for markers such as blood pressure and hematocrit. After an initial data matrix was composed, imputation of missing data values was carried out using penalized regression model in two steps. In the first phase, missing at random (MAR) variables were imputed, and in the second phase, structural study-wise imputation was conducted for the study-specific variables. All the variables were then truncated where appropriate and log-transformed (Supplementary Fig. 1A). (*ii*)

Study-wise differences or redundancies were observed for some features, which were dealt with by omitting or further transforming the selected variables. Interactions were introduced between the extracted single markers to derive new covariates. Principal Component Analysis (PCA) revealed systematic differences between the four studies (Supplementary Fig. 2), which was later accounted for by modeling study-specific components through ensemble learning. Further, clinical expertise within the team was utilized by omitting non-relevant or confounding factors. Initial data matrix included 124 variables and after removing clinically irrelevant ones, redundant, or highly skewed variables, 101 variables were left for use in the predictive modeling. Modeling of non-linearity through pairwise interactions resulted in a final data matrix with 3,422 features. (*iii*) Based on the unsupervised explorative analyses, two of the most representative studies (MAINSAIL and VENICE) were utilized in the supervised model learning. Three separate ensemble components were composed: MAINSAIL-specific ensemble component, VENICE-specific ensemble component and a combined ensemble component, which simultaneously modeled the two selected studies (Supplementary Fig. 1B). To reduce the risk of overfitting and avoid randomness bias in the binning, the final ensemble models were optimized using 10-fold cross-validation as well as averaged over multiple cross-validation runs. The model estimation procedure identified an optimal penalization parameter (λ), which controlled for the number of non-zero coefficients in the model. Simultaneously, the L_1/L_2 norm regularization parameter (α) was explored throughout the full model spectrum, ranging from Ridge Regression ($\alpha = 0$) to Elastic Net ($0 < \alpha < 1$) and LASSO ($\alpha = 1$) in penalized regression with respect to the objective function:

$$argmax_{\beta}[\frac{2}{n}\sum_{i=1}^{n}(x_{j(i)}^{T}\beta - log\,(\sum_{j\in R_i} e^{x_j^T\beta})) - \lambda(\alpha\sum_{i=1}^{p}|\beta_i| + \frac{1}{2}(1-\alpha)\sum_{i=1}^{p}\beta_i^2)]\ (Eqn.\ 1)$$

Here, x are the predictors (selected clinical variables or their pairwise interactions), β are the model coefficients subjected to the L_1 and L_2 norm penalization, p is the number of dimensions in the data, n is the number of observations, $j(i)$ is the index of the observation event at time t_i, and R_i is the set of indices j with $y_j \geq t_i$ (those patients at risk at time t_i). Each ensemble component resulted in a different optimum in Eqn. 1, as investigated by 10-fold cross-validated iAUC, although the resulting Elastic Net models closely resembled Ridge Regression. The penalized regression model was based on Cox proportional hazards (Eqn. 1). (*iv*) An ensemble prediction was performed by averaging the ranks over the component-wise predicted risk for the ENTHUSE 33 study (Supplementary Fig. 1C). Overall, the highest and lowest risk patients were concordantly predicted in each component. A few patient cases resulted in a moderate ensemble risk score, even if a particular ensemble component predicted a high or a low risk. Such challenging cases were controlled by not allowing any single study-specific effects to dominate the final predictions, through averaging over all the ensemble components.

Data-driven network projection for the ePCR model

The top-performing model's ensemble dual-study component was summarized by network visualization to create a clinically relevant representation of the most important markers and interaction effects (Figure 3). Each model coefficient β_i was given an importance score by computing the Elastic Net area under or above the regularization curve in the penalization and coefficient $\{\lambda, \beta_i\}$-space. Absolute values of the areas were used to rank each coefficient, which yielded a simultaneous scoring of both the effect size of the covariate as well as the importance of the feature in relation to the penalization. Statistical significance of each coefficient was then assessed by re-fitting to 10,000 bootstrapped datasets, and empirical p values were computed as the proportion of bootstrapped coefficients where $|\beta_{i,bootstrap}| \leq 10^{-10}$ or where $\beta_{i,bootstrap}$ flipped sign. A stringent threshold of $P < 1e^{-3}$ was used to select the coefficients as network nodes (single marker) or edges (interaction effects). Ensemble p values were averaged over all the components. Variable and interaction weighting was computed according to the average rank of the integrated regularization area over all ensemble components. The automated network layout was performed using attracting and repelling forces among the vertices, and the physical system (*graphopt*) was simulated until it reached the equilibrium (Figure 3). Top variables and interactions presented in this graph are available in the Supplementary Table 3, with the full variable and interaction list available at (https://www.synapse.org/#!Synapse:syn7113819).

Supplementary Tables

Supplementary Table 1. Full results from all 50 teams plus the Reference model across several scoring metrics from the Challenge. Performance measures were evaluated using the ENTHUSE 33 trial. Teams are listed with the links to their predictions, methods write-up, and code.

Team	Risk score predictions	Method write-up & code	iAUC	c-index	AUC12	AUC18	AUC24
FIMM-UTU (ePCR)	syn4732198	syn4227610	0.7915	0.7307	0.7918	0.7674	0.8388
Team Cornfield	syn4732339	syn4732274	0.7789	0.7263	0.7708	0.7663	0.8147
TeamX	syn4732955	syn4732218	0.7778	0.7157	0.7492	0.7645	0.8369
jls	syn4732934	syn4732827	0.7758	0.7212	0.7713	0.7553	0.8085
PC LEARN	syn4733119	syn3822697	0.7743	0.7205	0.7577	0.762	0.8258
KUstat	syn4741808	syn4260742	0.7732	0.7126	0.7436	0.7533	0.8376
A Bavarian dream	syn4732177	syn5592405	0.7725	0.7237	0.7721	0.7664	0.8019
qiuyulian1994	syn4732213	syn4732205	0.7716	0.711	0.7423	0.7506	0.8297
JayHawks	syn4731663	syn4214500	0.7711	0.7193	0.7717	0.7607	0.8124
Wind	syn4731647	syn4731645	0.771	0.7181	0.7625	0.7688	0.8124
Alvin	syn4732814	syn4229406	0.7707	0.7136	0.7586	0.7568	0.7927
brainstorm	syn4730818	syn3821841	0.7706	0.718	0.7617	0.7614	0.8175
uci-cbcl	syn4731657	syn4227279	0.7704	0.717	0.76	0.7716	0.8206
DreamOn	syn4731710	syn4731708	0.7704	0.712	0.7559	0.7582	0.8245
Clinical Persona	syn4681602	syn4681529	0.7704	0.7149	0.7533	0.7545	0.8328
Murat Dundar	syn4595033	syn4595029	0.7701	0.7305	0.7763	0.7773	0.773
Mistral	syn4622079	syn4622016	0.7689	0.7073	0.7382	0.7624	0.8268
UNC-BIAS	syn4731768	syn4731674	0.7685	0.717	0.7559	0.7568	0.8293
Team Marie	syn4731882	syn4485029	0.7682	0.7142	0.7519	0.7705	0.8151
A Elangovan	syn4643159	syn4212102	0.7677	0.7135	0.7655	0.7461	0.7977
M S	syn4730601	syn4229266	0.7671	0.707	0.7372	0.7652	0.8256
Jeevomics	syn4733845	syn4074987	0.7651	0.719	0.7733	0.7526	0.7917
CAMP	syn4731373	syn3647478	0.7646	0.7077	0.7331	0.758	0.8143
DAL_LAB	syn4731755	syn4731746	0.7642	0.7103	0.7521	0.7486	0.8305
Yuanfang Guan	syn7152471	syn7152438	0.7618	0.7143	0.7545	0.7631	0.8005
Bmore Dream Team	syn4733165	syn3616830	0.761	0.7121	0.7464	0.766	0.7948
Brigham Young University	syn4733391	syn4382527	0.7578	0.7048	0.7381	0.7685	0.7599
Team Simon	syn4733651	syn4732901	0.7573	0.7033	0.7278	0.7611	0.827
alan.saul	syn4731492	syn4587469	0.7568	0.7078	0.7464	0.7606	0.7961
BiSBII-UM	syn4733056	syn4229636	0.7561	0.6992	0.7394	0.7397	0.8007
RUBME6	syn4733262	syn4590933	0.7547	0.6994	0.7419	0.7198	0.7866
Jing Zhou	syn4646618	syn3685423	0.7507	0.6994	0.7361	0.7491	0.803
TYTDreamChallenge	syn4733257	syn4228911	0.748	0.7002	0.7343	0.7402	0.7657

UoB_Prostate	syn4733441	syn4591879	0.7478	0.7057	0.7468	0.7367	0.7699
Junmei Wang	syn4732891	syn4225820	0.7475	0.694	0.7319	0.7332	0.7955
Halabi Model	syn4770841	syn3324780	0.7429	0.6985	0.7418	0.7375	0.7634
Trishna	syn4730580	syn4730570	0.742	0.6922	0.7285	0.7383	0.774
CQB	syn4732202	syn3566822	0.7412	0.6914	0.7185	0.7293	0.7686
Ye Li	syn4731357	syn4731355	0.74	0.6907	0.7258	0.7249	0.806
Zhang Chihao	syn4748861	syn4259433	0.7376	0.7063	0.7561	0.7426	0.745
Guoping Feng	syn4730823	syn4730561	0.7261	0.6781	0.7073	0.707	0.7504
Y P	syn4732913	syn4732909	0.7241	0.6799	0.732	0.7057	0.7594
RainLab	syn4730829	syn4238316	0.7232	0.6708	0.7141	0.7394	0.7821
forPro	syn4707761	syn4707464	0.7219	0.6839	0.7267	0.7249	0.739
Marat Kazanov	syn4731369	syn4730567	0.7215	0.6675	0.7089	0.7112	0.7524
Jing Lu	syn4732498	syn4556277	0.7035	0.6689	0.6931	0.7073	0.7154
orion	syn4733693	syn4732963	0.6837	0.6457	0.717	0.7359	0.7952
limax	syn4732094	syn4721051	0.6756	0.6484	0.7033	0.6685	0.689
ECOP	syn4647266	syn4647259	0.6746	0.6554	0.6774	0.6881	0.6949
Massimiliano Zanin	syn4732241	syn4732239	0.6171	0.6081	0.6206	0.432	0.3852
The Data Wizard	syn4229053	syn4228992	0.5945	0.5815	0.6039	0.5824	0.6085
Compiled set of all predictions		syn7071669					

Supplementary Table 2. Comparison of risk stratification of patients in the ENTHUSE 33 trial by the ePCR and Reference models. Patients were dichotomized at median risk scores. All intervals reported are 95% confidence intervals. PPV = positive predictive value, NPV = negative predictive value. Values for Cases, Survivors, and Censored are cumulative.

ePCR model	Patient count	Event count	Median survival time, month (CI)	1 year survival rate (CI)	2 year survival rate (CI)	
Low risk group	156	56	27.6 (23.4-NA)	90.20% (85.5%-95.00%)	58.60% (49.7%- 69.00%)	
High risk group	157	107	15.1 (13.0-17.2)	59.90% (52.55%-68.20%)	15.70% (9.28%- 26.70%)	
Reference model	**Patient count**	**Event count**	**Median survival time, month (CI)**	**1 year survival rate (CI)**	**2 year survival rate (CI)**	
Low risk group	156	59	26.5 (22.5-NA)	87.40% (82.30%-92.90%)	52.80% (43.90%-63.50%)	
High risk group	157	104	15.6 (14.0-18.4)	62.70% (55.50%-70.80%)	22.20% (15.00%-32.90%)	
	Time (months)	t=6	t=12	t=18	t=24	t=30
	Cases	28	75	121	153	160
	Survivors	279	214	118	41	9
	Censored	6	24	74	119	144
Sensitivity (%)	ePCR	92.89	81.32	72.63	65.86	60.67
	Reference	85.73	75.94	67.43	61.19	61.21
Specificity (%)	ePCR	54.48	60.28	68.64	82.93	66.67
	Reference	53.76	57.94	64.41	73.17	44.44
PPV (%)	ePCR	16.96	40.15	64.2	86.31	82.41
	Reference	15.65	37.17	59.46	78.85	73.93
NPV (%)	ePCR	98.71	90.78	76.41	59.78	39.7
	Reference	97.41	88.02	71.86	53.57	30.8

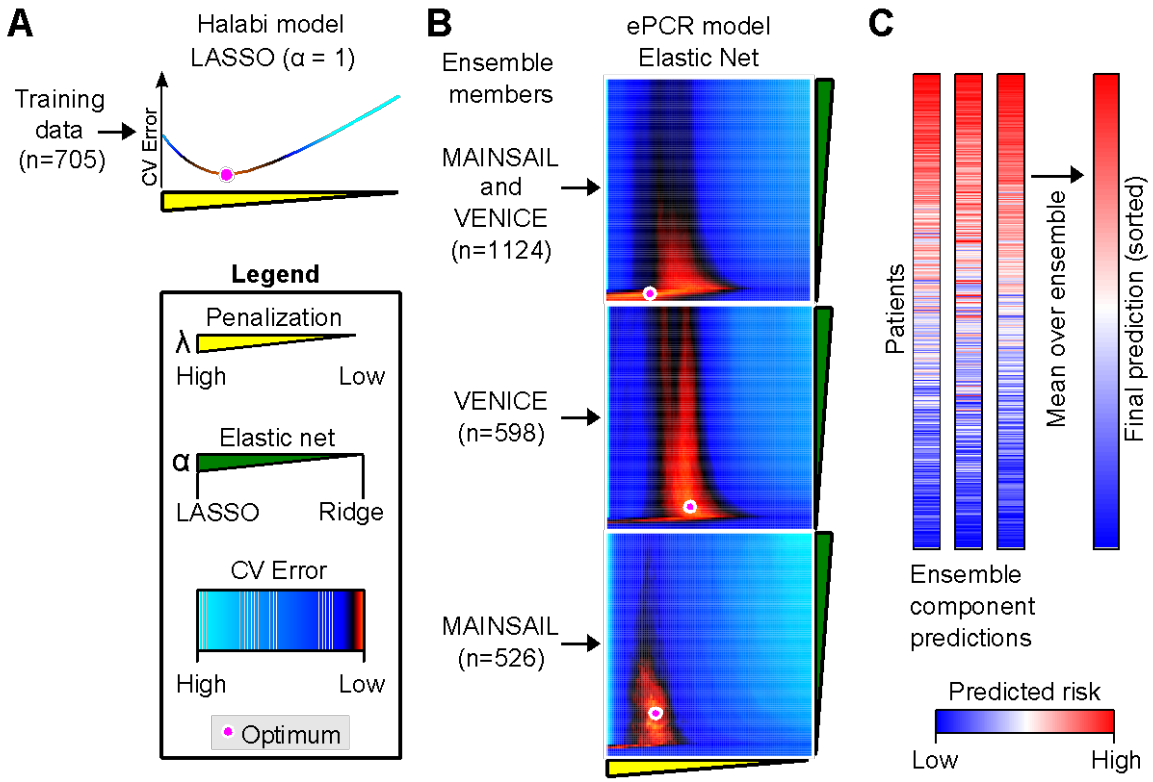
Supplementary Table 3. Top 15 single and interacting variables from the final ePCR model built from the MAINSAIL and VENICE trials. Comprehensive list of evaluated variables is available at: https://www.synapse.org/#!Synapse:syn7113819

Top 15 single variables in the ePCR model	Ensemble p value	Ensemble effect size
Lactate dehydrogenase (LDH)	< 0.0001	3405.667
Aspartate aminotransferase (AST)	< 0.0001	3376.667
Hemoglobin (HB)	< 0.0001	3369.667
Hematocrit (HCT)	< 0.0001	3354.333
Albumin (ALB)	0.0004	3316.667
Alkaline phosphatase (ALP)	< 0.0001	3291.333
Red blood cell count (RBC)	< 0.0001	3237.333
Systolic blood pressure (SYSTOLICBP)	0.0012	3192.000
Lesions at liver (LIVER)	< 0.0001	3184.000
Sodium (NA)	0.0205	3032.000
Lesions at target site (TARGET)	0.0118	3001.000
ECOG performance status (ECOG_C)	0.0003	2923.000
Medical history: cardiac disorders (MHCARD)	0.1100	2827.667
Lymphocyte/Leukocyte ratio (LYMperLEU)	0.0143	2684.333
Body mass index (BMI)	0.0214	2679.333

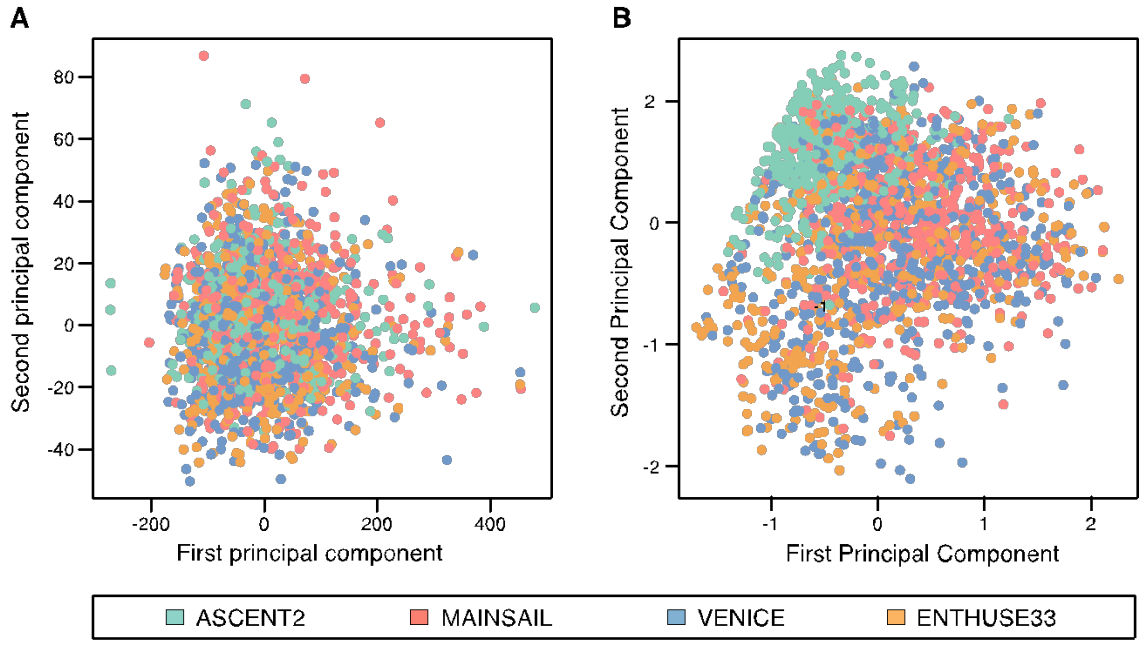
Top 15 interactions in the ePCR model		Ensemble p value	Ensemble effect size
AST	LDH	< 0.0001	3408.333
ALP	LDH	< 0.0001	3406.667
ALP	AST	< 0.0001	3404.333
HB	SYSTOLICBP	< 0.0001	3402.333
LDH	Urine Specific Gravity	< 0.0001	3400.667
SYSTOLICBP	HCT	< 0.0001	3400.333
Creatinine	LDH	< 0.0001	3397.333
LDH	LDH	< 0.0001	3392.000
HB	ALB	< 0.0001	3387.333

AST	AST	< 0.0001	3384.333
HB	NA	< 0.0001	3382.667
Height	LDH	< 0.0001	3381.667
ALB	SYSTOLICBP	< 0.0001	3379.333
HB	Creatinine clearance	< 0.0001	3378.000
ALB	HCT	< 0.0001	3377.333

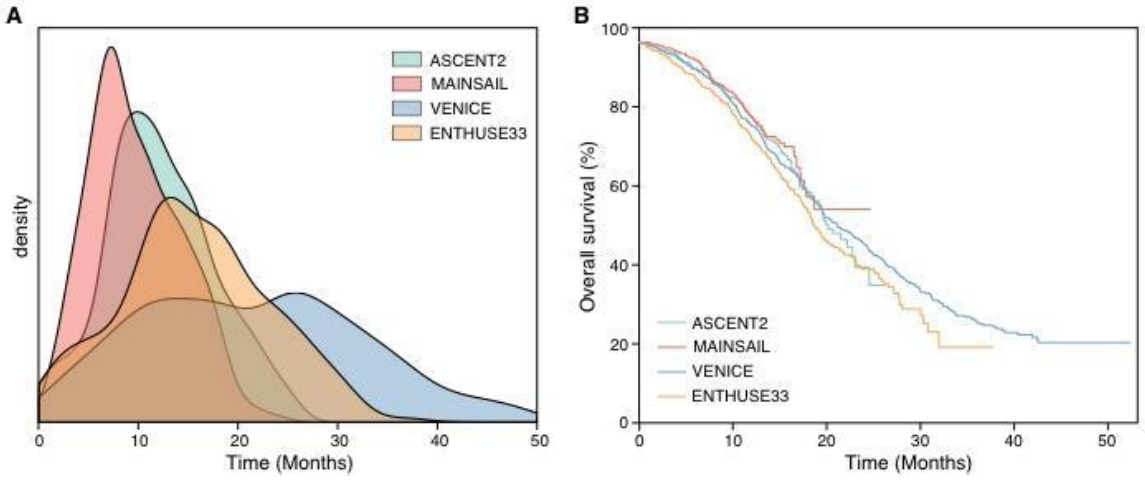
Supplementary Figures



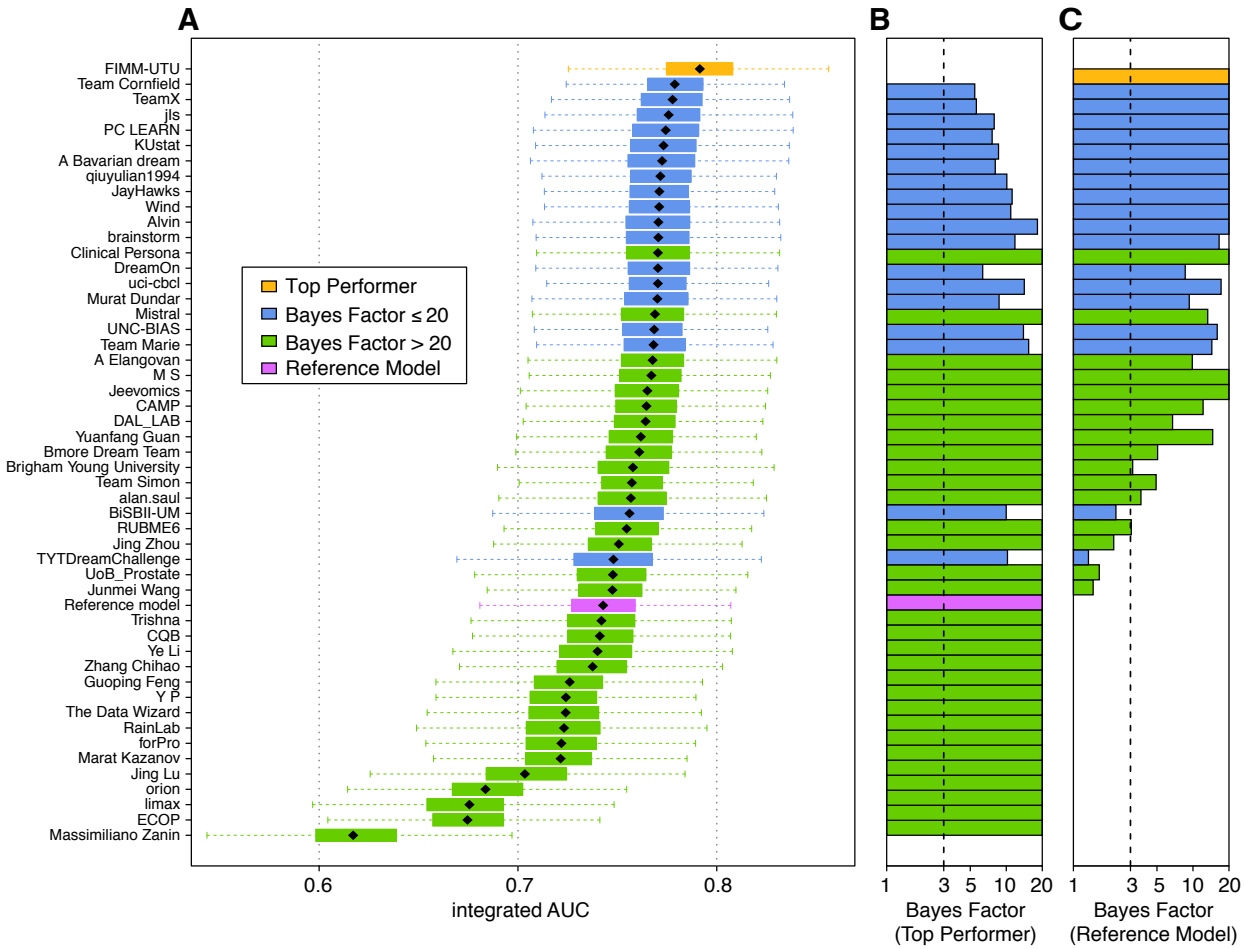
Supplementary Figure 1. Overview of the top-performing ePCR method in comparison to the Reference model (Halabi model). (A) The benchmarking Reference model explored the LASSO model (α = 1) in a training data cohort with respect to the regularization parameter (λ) using cross-validation (CV). (B) The top-performing ePCR approach is based on an ensemble of Penalized Cox Regression models (ePCR), which are optimized separately for each cohort or a combination of cohorts in terms of the regularization parameter (λ) as well as the full range of the L1/L2 regularization parameter (0 <= α <= 1). The optimal model was identified with low values of α, indicating that the Ridge Regression (α = 0)-like models performed better for modeling the complex interactions than the benchmarking Reference LASSO-model (α = 0). (C) Ensemble predictions were generated by averaging over the predicted risk ranks from each ensemble component.



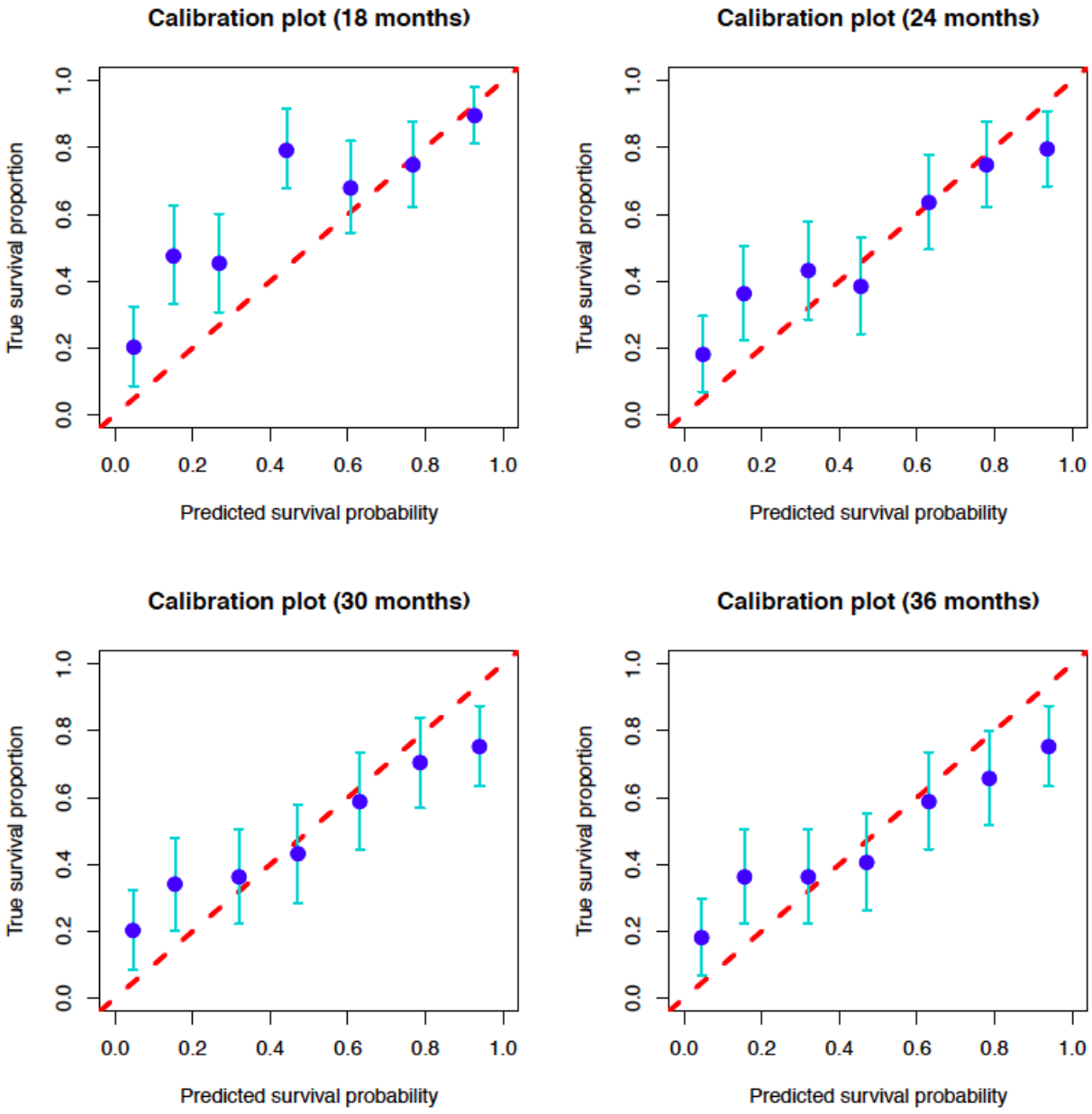
Supplementary Figure 2. (A) All data across ASCENT2, MAINSAIL, VENICE, and ENTHUSE 33– both binary and continuous data – were used in a PCA. (B) All data across the 4 studies – only binary variables – were used in PCA.



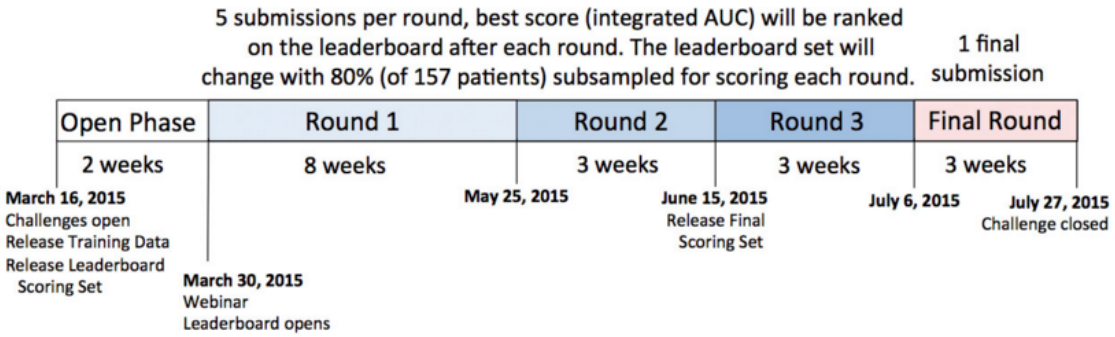
Supplementary Figure 3. (A) Density plot of follow-up times per study for the ASCENT2, MAINSAIL, VENICE, and ENTHUSE 33 trials. (B) Survival profile for each of the trials.



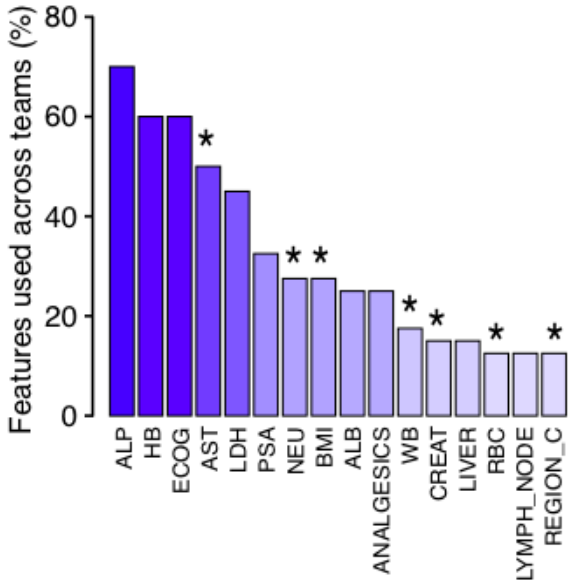
Supplementary Figure 4. Summary of Challenge results across all 50 teams plus the Reference model evaluated using the ENTHUSE 33 dataset. (A) Performance of submissions. Each submission underwent 1,000 paired bootstrap of final scoring patient set to calculate a Bayes factor against the top-performer a Bayes factor against the Reference model. A p value was calculated from randomization test of 1000 permutations. X-axis is iAUC and y-axis is submissions ranked by iAUC from high to low. Each team's bootstrapped iAUC scores are shown as horizontal boxplot with the black diamonds representing the point estimate of a team's performance. The colored boxes show the inter-quartile ranges and the whiskers extend to 1.5 times the corresponding interquartile ranges. Top-performer is colored in orange, other teams within Bayes factor of 20 were labeled in blue, and the rest of the teams were labeled in green. The Reference model is labeled in purple. (B) Bayes factors of all submissions against the top-performer are shown. Bayes factors greater than 20 were truncated to 20. (C) Bayes factors of all submissions against the Reference model. Bayes factors greater than 20 were truncated to 20.



Supplementary Figure 5. Calibration plots for the ePCR model of predicted survival probability versus true survival proportion for the ENTHUSE 33 dataset at 18, 24, 30, and 36 months.



Supplementary Figure 6. Timeline for the Challenge. Five submissions were allowed per round, and only a single submission for the final validation round.



Supplementary Figure 7. Most frequently utilized variables by teams to build their final models using the ASCENT2, MAINSAIL, and VENICE trials. Results are self-reported from a post-Challenge survey over 40 teams. * variables are not used in the Reference model.

References

1 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: 2191–8.

2 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: 760–8.

3 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: 417–25.

4 Fizazi K, Fizazi KS, Higano CS, 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: 1740–7.

5 Nelson JB, Fizazi K, Miller K, et al. Phase III study of the efficacy and safety of zibotentan (ZD4054) in patients with bone metastatic castration-resistant prostate cancer (CRPC). J Clin Oncol 2011; 29:abstract 117.

6 Kass RD, Raftery AE. Bayes factors. JASA. 1995; 90: 773-795.

7 Goodman SN. Towards evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med 1999; 130: 1005-13.

8 Halabi S, Lin C-Y, 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: 671–7.

Prostate Cancer Challenge DREAM Community
Kald Abdallah[82], Tero Aittokallio[21,22], Antti Airola[23], Catalina Anghel[6], Helia Azima[45], Robert Baertsch[35], Pedro J Ballester[39,78], Chris Bare[78], Vinayak Bhandari[72], Brian M Bot[78], Cuong C Dang[39,78], Maria Bekker-Nielsen Dunbar[34], Ann-Sophie Buchardt[34], Ljubomir Buturovic[77], Da Cao[10], Prabhakar Chalise[28], Junwoo Cho[20], Tzu-Ming Chu[32], R Yates Coley[8], Sailesh Conjeti[13], Sara Correia[15,16], James C Costello[81,87], Ziwei Dai[26], Junqiang Dai[28], Philip Dargatz[3], Sam Delavarkhan[45], Detian Deng[8], Ankur Dhanik[27], Yu Du[8], Aparna Elangovan[14], Shellie Ellis[29], Laura L Elo[53,55], Shadrielle M Espiritu[72], Fan Fan[72], Ashkan B Farshi[45], Ana Freitas[16], Brooke Fridley[28], Stephen Friend[78], Christiane Fuchs[1,4], Eyal Gofer[43], Gopalacharyulu Peddinti[22], Stefan Graw[28], Russ Greiner[41,42], Yuanfang Guan[56], Justin Guinney[78], Jing Guo2[5,65], Pankaj Gupta[13], Anna I Guyer[12], Jiawei Han[47], Niels R Hansen[34], Billy HW Chang[40], Outi Hirvonen[52], Barbara Huang[72], Chao Huang[58], Jinseub Hwang[19], Joseph G Ibrahim[58], Vivek Jayaswal[50], Jouhyun Jeon[6], Zhicheng Ji[8], Deekshith Juvvadi[30], Sirkku Jyrkkiö[52], Kimberly Kanigel-Winner[81], Amin Katouzian[13], Marat D Kazanov[37], Suleiman A Khan[22], Shahin Khayyer[45], Dalho Kim[20], Agnieszka K Golińska[59], Devin Koestler[28], Fernanda Kokowicz[17], Ivan Kondofersky[1,4], Norbert Krautenbacher[1,4], Damjan Krstajic[76,77], Luke Kumar[41], Christoph Kurz[2], Matthew Kyan[74], Teemu D Laajala[21,22], Michael Laimighofer[1,4], Eunjee Lee[58], Wojciech Lesiński[59], Miaozhu Li[11], Ye Li[61,68], Qiuyu Lian[44], Xiaotao Liang[61,62], Minseong Lim[20], Henry Lin[47], Xihui Lin[6], Jing Lu[31], Mehrad Mahmoudian[53], Roozbeh Manshaei[45], Richard Meier[28], Dejan Miljkovic[13], Tuomas Mirtti[22,24], Krzysztof Mnich[60], Nassir Navab[13], Elias C Neto[78], Yulia Newton[35], Thea Norman[78], Tapio Pahikkala[23], Subhabrata Pal[51], Byeongju Park[20], Jaykumar Patel[41], Swetabh Pathak[30], Alejandrina Pattin[13], Donna P Ankerst[4,5], Jian Peng[47], Anne H Petersen[34], Robin Philip[30], Stephen R Piccolo[12], Sebastian Pölsterl[13], Aneta Polewko-Klim[59], Karthik Rao[9], Xiang Ren[47], Miguel Rocha[15,16], Witold R. Rudnicki[59,60,66], Charles J Ryan[71], Hyunnam Ryu[20], Oliver Sartor[67], Hagen Scherb[1], Raghav Sehgal[30], Fatemeh Seyednasrollah[53,55], Jingbo Shang[47], Bin Shao[26], Liji Shen[86], Howard Sher[88], Motoki Shiga[36], Artem Sokolov[35], Julia F Söllner[1], Lei Song[48], Howard Soule[69], Gustavo Stolovitzky[83], Josh Stuart[35], Ren Sun[6,7], Christopher J Sweeney[70], Nazanin Tahmasebi[41], Kar-Tong Tan[25], Lisbeth Tomaziu[34], Joseph Usset[28], Yeeleng S Vang[57], Roberto Vega[41], Vitor Vieira[16], David Wang[72], Difei Wang[49], Junmei Wang[33], Lichao Wang[13], Sheng Wang[47], Tao Wang[79,80], Yue Wang[58], Russ Wolfinger[32], Chris Wong[35], Zhenke Wu[8], Jinfeng Xiao[46], Xiaohui Xie[57], Yang Xie[79,84,85], Doris Xin[47], Hojin Yang[58], Nancy Yu[6], Thomas Yu[78], Xiang Yu[10], Sulmaz Zahedi[73,75], Massimiliano Zanin[38], Chihao Zhang[64], Jingwen Zhang[58], Shihua Zhang[64], Yanchun Zhang[61,68], Fang Liz Zhou[86], Hongtu Zhu[58], Shanfeng Zhu[61,62,63] and Yuxin Zhu[8]

[1]Institute of Computational Biology, Helmholtz Zentrum München, Munich, Germany
[2]Institute of Health Economics and Health Care Management, Helmholtz Zentrum München, Munich, Germany
[3]Department of Hematology and Oncology, Johannes Wesling Klinikum Minden, Germany
[4]Department of Mathematics, Technische Universität München, Munich, Germany
[5]University of Texas Health Science Center at San Antonio, TX, USA
[6]Informatics and Biocomputing Program, Ontario Institute for Cancer Research (OICR), Toronto, Canada
[7]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Canada
[8]Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA
[9]School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA
[10]University of Pennsylvania, Philadelphia, PA, USA
[11]Biodemography of Aging Research Unit, Center for Population Health and Aging, Social Science Research Institute, Duke University, Durham, NC, USA
[12]Department of Biology, Brigham Young University, Provo, UT, USA
[13]Computer Aided Medical Procedures, Technische Universität München, Germany
[14]Computer Science Department, University of Melbourne, Melbourne, Australia
[15]Department of Informatics, University of Minho, Portugal
[16]Centre of Biological Engineering, University of Minho, Portugal
[17]Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, Florianpolis, Brazil
[18]Johns Hopkins University, Baltimore, MA, USA
[19]Department of Computer science and Statistics, Daegu University 712-714, Daegu, South Korea
[20]Department of Statistics, Kyungpook National University, 702-701 Daegu, South Korea
[21]Department of Mathematics and Statistics, University of Turku, Finland
[22]Institute for Molecular Medicine Finland, University of Helsinki, Finland

[23]Department of Information Technology, University of Turku, Finland
[24]Department of Pathology, Helsinki University Hospital, Finland
[25]Cancer Science Institute of Singapore, National University of Singapore, Singapore
[26]Center for Quantitative Biology, Peking University, Beijing 100871, China
[27]Regeneron Pharmaceuticals Inc, Tarrytown, New York, NY, USA
[28]Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS, USA
[29]Department of Health Policy and Management, University of Kansas Medical Center, Kansas City, KS, USA
[30]Jeevomics Pvt. Ltd.
[31]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
[32]JMP Life Sciences Division, SAS Institute Inc., Cary, NC, USA
[33]UT Southwestern, Dallas, TX, USA
[34]University of Copenhagen, Copenhagen, Denmark
[35]Department of Biomolecular Engineering and Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA, USA
[36]Department of Electrical, Electronic and Computer Engineering, Gifu University, Gifu, Japan
[37]Research and Training Center on Bioinformatics, Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia
[38]INNAXIS Foundation & Research Institute, Madrid, Spain
[39]Cancer Research Centre of Marseille, Marseille, France
[40]Division of Biostatistics, Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong
[41]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada
[42]Alberta Innovates Centre for Machine Learning, Edmonton, Alberta, Canada
[43]The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel
[44]Tsinghua University, Beijing 100084, China
[45]Electrical and Computer Engineering Dept., Ryerson University, Toronto, Canada
[46]Center for Biophysics and Quantitative Biology, The University of Illinois at Urbana-Champaign, IL, USA
[47]Department of Computer Science, The University of Illinois at Urbana-Champaign, IL, USA
[48]National Cancer Institute, National Institutes of Health, 9609 Medical Center Dr., Rockville, MD, USA
[49]Department of Biochemistry and Molecular and Cellular Biology, Georgetown University Medical Center, 4000 Reservoir Rd NW, Washington DC, USA
[50]Biocon Bristol-Myers Squibb Research Centre, Bangalore, India
[51]Centre for Cellular and Molecular Platforms, Bangalore, India
[52]The Department of Oncology and Radiotherapy, Turku University Central Hospital, Turku, Finland
[53]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland
[54]The Department of Clinical Oncology, University of Turku, Turku, Finland
[55]Department of Mathematics and Statistics, University of Turku, Turku, Finland
[56]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
[57]Department of Computer Science, University of California Irvine, Irvine, CA, USA
[58]Biostatistics and Imaging Analysis Lab, University of North Carolina at Chapel Hill, NC, USA
[59]Faculty of Mathematics and Informatics, University of Bialystok, Poland
[60]Computational Centre, University of Bialystok, Poland
[61]School of Computer Science, Fudan University, Shanghai 200433, China
[62]Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China
[63]Centre for Computational Systems Biology,Fudan University, Shanghai 200433, China
[64]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 100190 Beijing, China
[65]Research and development department, Annoroad Gene Technology Co. Ltd, Beijing, China
[66]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Poland
[67]Tulane Cancer Center, Tulane University, New Orleans, LA, USA
[68]Shanghai Key Lab of Data Science, Fudan University, Shanghai 200433, China
[69]Prostate Cancer Foundation, Santa Monica, CA, USA

[70]Department of Medical Oncology, Dana-Farber Cancer Institute and Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[71]Genitourinary Medical Oncology Program, Division of Hematology & Oncology, University of California, San Francisco, CA, USA
[72]Ontario Institute for Cancer Research, Toronto, Ontario, Canada
[73]The Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Canada
[74]Electrical Engineering and Computer Science Dept., York University, Toronto, Canada
[75]iBEST - Li Ka Shing Institute of Knowledge, St. Michael's Hospital, Toronto, Canada
[76]Research Centre for Cheminformatics, Jasenova 7, 11030 Beograd, Serbia
[77]Clinical Persona Inc, 932 Mouton Circle, East Palo Alto, CA, USA
[78]Sage Bionetworks, Seattle, WA, USA
[79]Quantitative Biomedical Research Center, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, USA
[80]Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, Texas, USA
[81]Department of Pharmacology & Computational Biosciences Program, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA
[82]AstraZeneca, Gaithersburg, MD, USA
[83]IBM T.J. Watson Research Center, IBM, Yorktown Heights, NY, USA
[84]The Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA
[85]Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, USA
[86]Sanofi, Bridgewater, NJ, USA
[87]University of Colorado Comprehensive Cancer Center, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA
[88]Sidney Kimmel Center for Prostate and Urologic Cancers, Memorial Sloan-Kettering Cancer Center and Weill Cornell Medical College, New York, NY, USA

Funding Support:
European Union within the ERC grant LatentCauses supported the work of C.F and I.K. German Research Foundation (DFG) within the Collaborative Research Centre 1243, subproject A17 awarded to C.F. German Federal Ministry of Education and Research (BMBF) through the Research Consortium e:AtheroMED (Systems medicine of myocardial infarction and stroke) under the auspices of the e:Med Programme (grant # 01ZX1313C) supported the work of D.P.A., P.D., C.F., C.K., I.K., N.K., M.L., H.S. and J.F.S. at the Institute of Computational Biology. NIH Grants RR025747-01, MH086633 and 1UL1TR001111, and NSF Grants SES-1357666, DMS-14-07655 and BCS-0826844 supported the work of C.H., J.I., E.L., Y.W., H.Y., H.Z. and J.Z. NSFC Grant Nos. 61332013, 61572139 supported the work of X.L, Y.L, Y.Z., and S.Z. National Natural Science Foundation of China grants [Nos. 61422309, 61379092] was awarded to S.Z. The Patrick C. Walsh Prostate Research Fund and the Johns Hopkins Individualized Health Initiative supported the work of R.Y.C., D.D., Y.D., Z.J., K.R., Z.W. and Y.Z. FCT Ph.D. Grant SFRH/BD/80925/2011 was awarded to S.C. Clinical Persona Inc., East Palo Alto, CA supported the work of L.B. and D.K. The Finnish Cultural Foundation and the Drug Research Doctoral Programme (DRDP) at the University of Turku supported T.D.L. The National Research Foundation Singapore and the Singapore Ministry of Education, under its Research Centres of Excellence initiative, supported the work of J.G. and K.T. A grant from the Russian Science Foundation 14-24-00155 was awarded to M.D.K. A*MIDEX grant (no. ANR-11-IDEX-0001-02) was awarded to P.J.B. NSERC supported the work of R.G. The Israeli Centers of Research Excellence (I-CORE) program (Center No. 4/11) supported the work of E.G. Academy of Finland (grants 292611, 269862, 272437, 279163, 295504), National Cancer Institute (16X064), and Cancer Society of Finland supported the work of T.A. Academy of Finland (grant 268531) supported the work of T.M.