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

Justin Guinney, Tao Wang*, Teemu D Laajala*, Kimberly Kanigel Winner, J Christopher Bare, Elias Chaibub Neto, Suleiman A Khan, Gopal Peddinti, Antti Airola, Tapio Pahikkala, Tuomas Mirtti, Thomas Yu, Brian M Bot, Liji Shen, Kald Abdallah, Thea Norman, Stephen Friend, Gustavo Stolovitzky, Howard Soule, Christopher J Sweeney, Charles J Ryan, Howard I Scher, Oliver Sartor, Yang Xie†, Tero Aittokallio†, Fang Liz Zhou†, James C Costello†, and the Prostate Cancer Challenge DREAM Community‡*

Lancet Oncol 2017; 18: 132–42

Published Online November 15, 2016 http://dx.doi.org/10.1016/S1470-2045(16)30560-5

See Comment page 15

*Contributed equally as first authors

†Contributed equally as senior authors

‡Members listed in the appendix

Sage Bionetworks, Seattle, WA, USA (J Guinney PhD, J C Bare BS, E C Neto PhD, T Yu BS, B M Bot MS, T Norman PhD, S Friend MD); **Quantitative Biomedical Research Center, Department of Clinical Sciences** (T Wang PhD, Y Xie PhD), **Center for the Genetics of Host Defense** (T Wang), **The Simmons Comprehensive Cancer Center,** (Y Xie), **and Lyda Hill Department of Bioinformatics** (Y Xie), **University of Texas Southwestern Medical Center, Dallas, TX, USA; Department of Mathematics and Statistics** (T D Laajala MSc, Prof T Aittokallio PhD), **and Department of Information Technology** (A Airola PhD, T Pahikkala PhD), **University of Turku, Turku, Finland; Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland** (T D Laajala, S A Khan PhD, G Peddinti PhD, T Mirtti MD, Prof T Aittokallio); **Department of Pathology (HUSLAB), Helsinki University Hospital, Helsinki, Finland** (T Mirtti); **Department of Pharmacology and Computational Biosciences Program** (K Kanigel Winner PhD, J C Costello PhD), **and University of Colorado Comprehensive Cancer Center** (J C Costello),

Summary

Background Improvements to prognostic models in metastatic castration-resistant prostate cancer have the potential to augment clinical trial design and guide treatment strategies. In partnership with Project Data Sphere, a not-for-profit initiative allowing data from cancer clinical trials to be shared broadly with researchers, we designed an open-data, crowdsourced, DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge to not only identify a better prognostic model for prediction of survival in patients with metastatic castration-resistant prostate cancer but also engage a community of international data scientists to study this disease.

Methods Data from the comparator arms of four phase 3 clinical trials in first-line metastatic castration-resistant prostate cancer were obtained from Project Data Sphere, comprising 476 patients treated with docetaxel and prednisone from the ASCENT2 trial, 526 patients treated with docetaxel, prednisone, and placebo in the MAINSAIL trial, 598 patients treated with docetaxel, prednisone or prednisolone, and placebo in the VENICE trial, and 470 patients treated with docetaxel and placebo in the ENTHUSE 33 trial. Datasets consisting of more than 150 clinical variables were curated centrally, including demographics, laboratory values, medical history, lesion sites, and previous treatments. Data from ASCENT2, MAINSAIL, and VENICE were released publicly to be used as training data to predict the outcome of interest—namely, overall survival. Clinical data were also released for ENTHUSE 33, but data for outcome variables (overall survival and event status) were hidden from the challenge participants so that ENTHUSE 33 could be used for independent validation. Methods were evaluated using the integrated time-dependent area under the curve (iAUC). The reference model, based on eight clinical variables and a penalised Cox proportional-hazards model, was used to compare method performance. Further validation was done using data from a fifth trial—ENTHUSE M1—in which 266 patients with metastatic castration-resistant prostate cancer were treated with placebo alone.

Findings 50 independent methods were developed to predict overall survival and were evaluated through the DREAM challenge. The top performer was based on an ensemble of penalised Cox regression models (ePCR), which uniquely identified predictive interaction effects with immune biomarkers and markers of hepatic and renal function. Overall, ePCR outperformed all other methods (iAUC 0·791; Bayes factor >5) and surpassed the reference model (iAUC 0·743; Bayes factor >20). Both the ePCR model and reference models stratified patients in the ENTHUSE 33 trial into high-risk and low-risk groups with significantly different overall survival (ePCR: hazard ratio 3·32, 95% CI 2·39–4·62, p<0·0001; reference model: 2·56, 1·85–3·53, p<0·0001). The new model was validated further on the ENTHUSE M1 cohort with similarly high performance (iAUC 0·768). Meta-analysis across all methods confirmed previously identified predictive clinical variables and revealed aspartate aminotransferase as an important, albeit previously under-reported, prognostic biomarker.

Interpretation Novel prognostic factors were delineated, and the assessment of 50 methods developed by independent international teams establishes a benchmark for development of methods in the future. The results of this effort show that data-sharing, when combined with a crowdsourced challenge, is a robust and powerful framework to develop new prognostic models in advanced prostate cancer.

Funding Sanofi US Services, Project Data Sphere.

Introduction

Prostate cancer is the most common cancer among men in high-income countries and ranks third in terms of mortality after lung cancer and colorectal cancer.[1] Of more than two million men diagnosed with prostate cancer in the USA over the past 10 years, roughly 10%

Research in context

Evidence before this study

We searched PubMed between January, 2012, and July, 2015, with the terms "prognosis", "overall survival", "mCRPC", and "docetaxel". Our search yielded a 2014 study in which an updated prognostic model was described for metastatic castration-resistant prostate cancer that had been developed from the CALGB-90401 study (a randomised, double-blind, phase 3 clinical trial) and validated with data from the phase 3 ENTHUSE 33 trial. The study focused on a subset of clinical variables using datasets that were not in the public domain. Leveraging the wealth of data already generated from clinical trials is challenging on several fronts, but is complicated in particular by data access.

Added value of this study

Project Data Sphere is an independent not-for-profit initiative that aims to provide open access to historical patient-level data. The prostate cancer DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge is an open-data, crowdsourced competition to develop and assess prognostic models in metastatic castration-resistant prostate cancer. Using data from the comparator arms of four phase 3 clinical trials of chemotherapy-naive patients with metastatic castration-resistant prostate cancer, 50 independent teams—a diverse group of experts including biostatisticians, computer scientists, and clinical experts—developed prognostic models for the DREAM challenge, representing, to the best of our knowledge, the most comprehensive set of benchmarked models to date. The best-performing model was based on an ensemble of penalised Cox regression models that judged the prognostic value of interactions between predictor covariates and substantially outperformed the 2014 model. Strong support was provided for previously identified prognostic variables in the 50 models, and additional important variables were identified along with novel interactions between covariates. Data are available publicly through the Project Data Sphere initiative, and all method predictions and code are available for download through the Sage Bionetworks Synapse platform.

Implications of all the available evidence

Clinical trial data-sharing is both feasible and useful, and the DREAM challenge is an appropriate vehicle on which to build and rigorously assess prognostic or predictive models quickly, openly, and robustly. We established a new prognostic benchmark in metastatic castration-resistant prostate cancer, with applications in trial design and guidance for clinicians and patients. Robust and accurate prognostic predictors can be used to homogenise risk in clinical trials of metastatic castration-resistant prostate cancer and enable smaller trials for assessment of treatment effects.

presented with metastatic disease. For these men, the mainstay of treatment is androgen deprivation therapy, with a high proportion of response. However, responses are not durable, and nearly all tumours eventually progress to the lethal metastatic castration-resistant state. Although substantial improvements in outcome for men with metastatic castration-resistant prostate cancer have been achieved after approval of next-generation hormonal agents, an immunotherapeutic drug, a radiopharmaceutical agent, and a cytotoxic drug,[2–10] how best to deploy these treatments has not been ascertained. Elucidation of variables associated with patients' outcomes independent of treatment will facilitate the design of future trials by homogenising risk, thus enabling clinical trial questions to be answered more rapidly because smaller sample sizes will be needed.

Prognostic models in metastatic castration-resistant prostate cancer have been described[11–13] using baseline variables from independent cohort studies. A 2014 prognostic model for metastatic castration-resistant prostate cancer [14] included eight clinical factors predictive of overall survival: Eastern Cooperative Oncology Group (ECOG) performance status; disease site; use of opioid analgesics; lactate dehydrogenase; albumin; haemoglobin; prostate-specific antigen; and alkaline phosphatase. Can innovative models with improved performance be developed through a systematic search using data-driven approaches while providing insights into biological aspects of the disease that affect patients' outcomes? An example of a novel clinical factor that is underexplored in contemporary prognostic model development is interaction effects between clinical variables, even though interactions between genetic variants are used widely and known to improve genetic-based risk prediction and patients' stratification.[15,16]

Here, we present results from the prostate cancer DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge—an open-data, crowdsourced challenge in metastatic castration-resistant prostate cancer. A major contribution to this effort was removal of privacy and legal barriers associated with open access to phase 3 clinical trial data[17] by Project Data Sphere—a not-for-profit initiative of the CEO Roundtable on Cancer's Life Sciences Consortium that broadly shares oncology clinical trial data with researchers. The challenge was designed to accomplish two goals. First, we aimed to leverage open clinical trial data, enabling a community-based approach to identify the best-performing prognostic model in a rigorous and unbiased manner. Second, participating teams aimed to develop predictive models to both validate previously characterised predictive clinical variables and discover new prognostic features. Consistent with the mission of DREAM, all challenge data, results, and method descriptions from participating teams are available publicly through the open-access Synapse platform.

University of Colorado, Anschutz Medical Campus, Aurora, CO, USA; AstraZeneca, Gaithersburg, MD, USA (K Abdallah MD); **IBM T J Watson Research Center, IBM, Yorktown Heights, NY, USA** (G Stolovitzky PhD); **Prostate Cancer Foundation, Santa Monica, CA, USA** (H Soule PhD); **Department of Medical Oncology, Dana-Farber Cancer Institute and Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA** (C J Sweeney MBBS); **Genitourinary Medical Oncology Program, Division of Hematology and Oncology, University of California, San Francisco, CA, USA** (Prof C J Ryan MD); **Genitourinary Oncology Services, Department of Medicine, Sidney Kimmel Center for Prostate and Urologic Cancers, Memorial Sloan-Kettering Cancer Center and Weill Cornell Medical College, New York, NY, USA** (Prof H I Scher MD); **Tulane Cancer Center, Tulane University, New Orleans, LA, USA** (Prof O Sartor MD); **and Sanofi, Bridgewater, NJ, USA** (L Shen PhD, F L Zhou MD)

Correspondence to:
Dr James C Costello, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
james.costello@ucdenver.edu

See **Online** for appendix

For more on **DREAM challenges** see http://dreamchallenges.org

For more on **Project Data Sphere** see http://www.projectdatasphere.org

For more on the **CEO Roundtable on Cancer's Life Sciences Consortium** see http://ceo-lsc.org

To access **data via the Synapse platform** see https://www.synapse.org/ProstateCancerChallenge

Methods

Trial selection

In April 2014, the DREAM challenge organising team reviewed all existing and incoming prostate cancer trial datasets (comparator arm only) in Project Data Sphere and selected four trials, which were the source of training and validation datasets for the DREAM challenge—ASCENT2,[18] MAINSAIL,[19] VENICE,[20] and ENTHUSE 33.[21] All four trials were randomised phase 3 clinical trials in which the comparator arm consisted of a docetaxel regimen and overall survival was the primary endpoint. These four trials also had similar inclusion and exclusion criteria: eligible patients were aged 18 years and older, had progressive metastatic castration-resistant prostate cancer, were chemotherapy-naive, and had an ECOG performance status of 0–2. Further details of inclusion and exclusion criteria for each trial are provided in the appendix (p 3). The patient-level trial datasets were deidentified by data providers and made available for the DREAM challenge through Project Data Sphere. No institutional review board approval was needed to access data.

Patient populations

We compiled training datasets from the comparator arms of ASCENT2, MAINSAIL, and VENICE. ASCENT2[18] is a randomised open-label study assessing DN-101 in combination with docetaxel. Patients with metastatic castration-resistant prostate cancer were randomly assigned either docetaxel and prednisone (comparator arm) or docetaxel and DN-101, stratified by geographical region and ECOG performance status. MAINSAIL[19] is a randomised double-blind study to assess efficacy and safety of docetaxel and prednisone with or without lenalidomide in patients with metastatic castration-resistant prostate cancer. Participants were randomly assigned to either docetaxel, prednisone, and placebo (comparator arm) or lenalidomide, docetaxel, and prednisone. Stratification of patients in MAINSAIL was done based on ECOG performance status (0–1 *vs* 2), geographical region (USA and Canada *vs* Europe and Australia *vs* rest of world), and type of disease progression after hormonal treatment (rising prostate-specific antigen only *vs* tumour progression). VENICE[20] is a randomised double-blind study comparing the efficacy and safety of aflibercept versus placebo, in which patients with metastatic castration-resistant prostate cancer were randomly assigned either docetaxel, prednisone or prednisolone, and placebo (comparator arm) or docetaxel, prednisone or prednisolone, and aflibercept. Participants were stratified by baseline ECOG performance status (0–1 *vs* 2). The validation dataset was from the ENTHUSE 33 trial,[21] a double-blind study in which patients with metastatic castration-resistant prostate cancer were randomly allocated (1:1) either docetaxel and placebo (comparator arm) or docetaxel with zibotentan, stratified by centre.

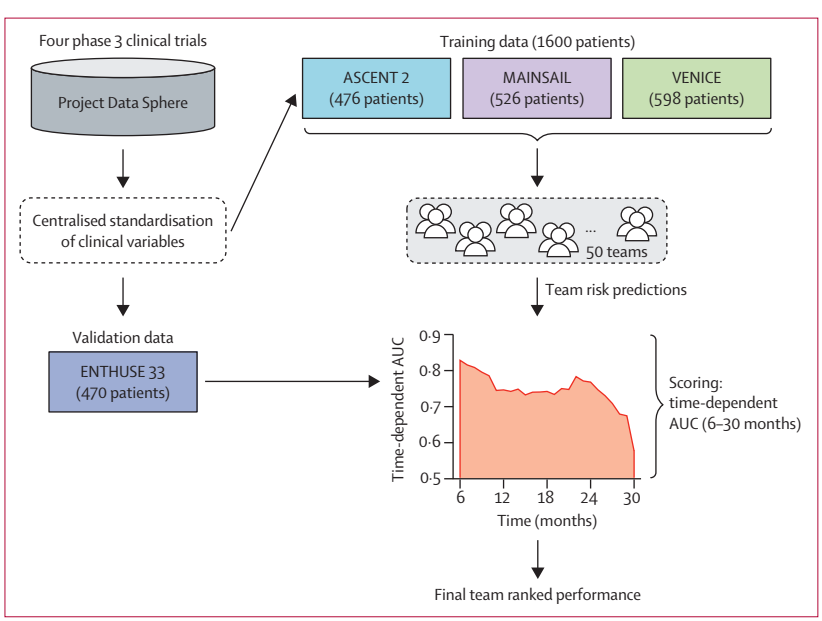


Figure 1: **Study design**
Data were acquired from Project Data Sphere and curated centrally by the organising team to provide a harmonised dataset across the four studies. Three studies were provided as training data (ASCENT2, MAINSAIL, and VENICE) and the fourth (ENTHUSE 33) was the validation dataset. Teams submitted risk scores for ENTHUSE 33, then their predictions were scored and ranked using an integrated time-dependent area under the curve (AUC) metric.

Data curation

The original datasets from Project Data Sphere contained patient-level raw tables that conformed to either Study Data Tabulation Model (SDTM) standards or company-specific clinical database standards. To optimise use of these data for the DREAM challenge, we compiled the four sets of trial data into a set of five standardised raw event-level tables, meaning all four clinical trials were combined into the same tables based on laboratory values, medical history, lesion sites, previous treatments, and vital signs. Including patients' demographic information, these tables presented most measurements made for the patient in that category. To summarise these data on a per-patient level, we created a core table, distilling the raw event-level tables and patients' demographics into 129 clinically defined baseline and outcome variables. Full details of the data curation process are provided in the appendix (pp 3, 4).

We supplied participating teams with the full set of baseline and raw variables from the core and raw event-level tables. We encouraged challenge participants to derive additional baseline clinical variables from the five standardised raw event-level tables for modeling. We also provided teams with outcome variables for the ASCENT2, MAINSAIL, and VENICE trials, but we did not release the outcome variables for the ENTHUSE 33 trial because they would serve to independently evaluate the performance of models. The primary endpoint used for model development was overall survival, defined as

	ASCENT2 (n=476)	MAINSAIL (n=526)	VENICE (n=598)	ENTHUSE 33 (n=470)	ENTHUSE M1 (n=266)
Age (years)					
18–64	111 (23%)	171 (33%)	219 (37%)	160 (34%)	58 (22%)
65–74	211 (44%)	246 (47%)	254 (42%)	217 (46%)	111 (42%)
≥75	154 (32%)	109 (21%)	125 (21%)	93 (20%)	97 (36%)
ECOG performance score					
0	220 (46%)	257 (49%)	280 (47%)	247 (53%)	196 (74%)
1	234 (49%)	247 (47%)	291 (49%)	223 (47%)	70 (26%)
2	22 (5%)	20 (4%)	27 (5%)	0 (0%)	0 (0%)
3	0 (0%)	1 (<1%)	0 (0%)	0 (0%)	0 (0%)
Missing	0 (0%)	1 (<1%)	0 (0%)	0 (0%)	0 (0%)
Metastasis					
Liver	5 (1%)	58 (11%)	60 (10%)	64 (14%)	12 (5%)
Bone	345 (72%)	439 (83%)	529 (88%)	470 (100%)	266 (100%)
Lungs	8 (2%)	74 (14%)	88 (15%)	56 (12%)	13 (5%)
Lymph nodes	163 (34%)	298 (57%)	323 (54%)	208 (44%)	80 (30%)
Analgesic use					
No	338 (71%)	347 (66%)	419 (70%)	339 (72%)	256 (96%)
Yes	138 (29%)	179 (34%)	179 (30%)	131 (28%)	10 (4%)
Lactate dehydrogenase (U/L)	202 (176–250)	210 (174–267)	NA	213 (181–287)	188 (170–219)
Missing	13 (3%)	1 (<1%)	596 (100%)	5 (1%)	7 (3%)
Prostate-specific antigen (ng/mL)	68·8 (24·2–188·4)	84·9 (32·2–271·2)	90·8 (30·8–260·6)	99·6 (33·6–236·8)	52·3 (17·3–153·0)
Missing	1 (<1%)	4 (1%)	6 (1%)	12 (3%)	4 (2%)
Haemoglobin (g/dL)	12·6 (11·6–13·6)	12·7 (11·5–13·7)	12·7 (11·7–13·5)	12·5 (11·3–13·5)	12·9 (12·2–13·7)
Missing	3 (1%)	10 (2%)	0 (0%)	4 (1%)	2 (1%)
Albumin (g/L)	NA	43 (41–45)	42 (38–45)	43 (40–46)	43 (41–45)
Missing	476 (100%)	1 (<1%)	16 (3%)	2 (<1%)	1 (<1%)
Alkaline phosphatase (U/L)	113 (80–213)	124 (81–265)	135 (85–270)	155 (98–328)	130 (83–222)
Aspartate aminotransferase (U/L)	24 (20–31)	24 (19–31)	25 (20–33)	25 (20–33)	24 (19–29)
Missing	4 (1%)	1 (<1%)	8 (1%)	3 (1%)	3 (1%)

Data are median (IQR) or number of patients (%). NA=not available. ECOG=Eastern Cooperative Oncology Group.

Table: **Patients' baseline characteristics**

the time from date of randomisation to the date of death from any cause.

We did principal component analysis to investigate systematic similarities or differences between the four clinical trials, using either all available variables or binary variables only. We visualised the principal component analysis by plotting the first principal component against the second principal component for all patients.

Further validation

After the DREAM challenge was completed using data from ENTHUSE 33 for method evaluation, we further validated the top-performing and reference models with data from a fifth trial, ENTHUSE M1,[22] to assess whether the top-performing model could be used to stratify risk for patients with metastatic castration-resistant prostate cancer who received placebo alone and no docetaxel. ENTHUSE M1 is a randomised double-blind study to assess the efficacy and safety of 10 mg zibotentan in patients with metastatic castration-resistant prostate cancer (specifically, bone metastasis). By contrast with ENTHUSE 33, the ENTHUSE M1 trial included a comparator arm of placebo alone. Patients were randomly allocated (1:1) either zibotentan or placebo and were stratified by centre. The inclusion and exclusion criteria were similar to those used for ENTHUSE 33 except that patients in ENTHUSE M1 were pain free or mildly symptomatic. To be consistent for validation, curation of ENTHUSE M1 data followed the same process as was done for ASCENT2, MAINSAIL, VENICE, and ENTHUSE 33, resulting in a core table and five raw event-level tables.

Challenge procedures

The DREAM challenge was hosted and fully managed on Synapse, a cloud-based platform for collaborative scientific data analysis, through which all model predictions were submitted. The challenge was run in two phases (appendix pp 4, 17). First, teams were allowed to train and test their models in an open testing leaderboard phase. Second, teams were permitted one last submission to the final scoring phase, after which teams were scored and

Panel: **Top-performing model construction in training datasets**

The top-performing model was based on an ensemble of penalised Cox regression models (ePCR), as shown in the equation. For each trial-specific ensemble component, the model estimation procedure identified an optimum penalisation parameter (λ), which controls for the number of non-zero coefficients in the prediction model, and simultaneously the regularisation parameter (α) with respect to the objective function:

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

Here, x are the predictors (clinical variables or their pairwise interactions), β are the model coefficients subjected to the absolute error and squared error penalisations ($|\beta|$ and β^2, respectively), p is the number of predictors, 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 for which $y_j \geq T_i$ (patients at risk at time T_i), where y_j is the observed death or right-censoring time. The set of indices R_i is redefined for each patient i using the above risk criterion incorporating y and T. With suitable regularisation, the penalised regression identifies an optimum balance between the model fit and top predictors, effectively generalising the Cox model for future predictions. To reduce the risk of overfitting and to avoid randomness bias in the binning, the final ensemble models were optimised using ten-fold cross-validation of the iAUC, averaged over multiple cross-validation runs. By modelling each trial individually as a separate ensemble component with different optima in the equation, we are able to account effectively for trial-specific variation (appendix p 12). The optimum parameters (penalisation λ and norm α) for each trial were first identified using cross-validation, after which the model coefficients (β) are estimated by optimising the above objective function.

Data processing entailed missing value imputation with a penalised Gaussian regression variant of the equation, with cross-validation when variables with non-missing values were used as predictors. Variables with missing values were inferred by training an optimum model with the non-missing variables and then imputing the missing values. Laboratory values were modelled as continuous variables. Data curation entailed unsupervised explorative analyses (appendix pp 5, 6, 12). ASCENT2 trial data were used in the imputation and unsupervised learning phases but were omitted from construction of the final supervised ensemble predictor, which was based on three components: MAINSAIL alone, VENICE alone, and their combination (appendix p 12). The final ensemble prediction was done by averaging over the ranks of the component-predicted risks for the ENTHUSE 33 dataset (appendix p 12). Averaging of risk score ranks was selected to be more robust to trial-specific variation and potential outliers. Full details of the model and its network visualisation are in the appendix (pp 5, 6, 12) with a list of chosen predictors (appendix pp 10, 11).

ranked. Accordingly, we split data from ENTHUSE 33 into two separate sets, consisting of 157 patients and 313 patients. The smaller dataset was used for the open testing phase and the larger dataset was used for the final scoring phase. Moreover, all reported performance values for the evaluated methods and all comparisons between the top-performing model and reference model used the larger set of data from the ENTHUSE 33 trial. The reference prognostic model for prediction of overall survival was a penalised Cox proportional-hazards model using the adaptive least absolute shrinkage and selection operator (LASSO) penalty.[14]

For method evaluation, we used the integrated AUC (iAUC)[23] calculated from 6–30 months as our primary scoring metric. For robust determination of the best performing team or teams, we used Bayes factor analysis and randomisation test based on iAUC (appendix pp 4, 5). For each team, we calculated the Bayes factor to directly compare the performance of a model with the reference model; coefficients for the reference model were obtained from reported hazard ratios (HRs).[14] Furthermore, we evaluated model predictions by plotting Kaplan-Meier curves, after dichotomising patients for each team separately by median risk score. We used the log-rank test to compare the two groups using the *coxph* function in the *survival R* package. We calculated CIs by inverting the Wald test statistic. The risk scores generated by each model have their own dynamic range; thus, we used the rankings of patients for scoring by iAUC or Kaplan-Meier analysis. Accordingly, we selected the median risk score as a means to compare different methods in a fair manner. A major goal of the challenge was to encourage teams to develop and test novel methods outside of standard survival analysis approaches; thus, risk score predictions across all teams varied in their range and distribution. A standard threshold could not be established fairly for all teams; therefore, we relied on rank-based scoring methods, including the iAUC, and stratifying risk scores based on the median. We also calculated other statistics, including median survival and 1-year and 2-year survival for the dichotomised high-risk and low-risk groups. We did hierarchical clustering on rank-normalised risk score predictions from all models in the challenge, using Euclidean distance and average linkage.

We used the ENTHUSE 33 dataset to assess the calibration of the top-performing model. We plotted the predicted survival probability based on the top-performing model against the observed survival proportions at 18, 24, 30, and 36 months. For each time cutoff, we divided the population into seven equally spaced categories based on the ranked predicted risk by the top-performing model. We then calculated the true survival proportion within each category and plotted it as a point estimate and 95% CI. A 45° line on the plots indicated perfect calibration.

The organisers of the DREAM challenge used SAS version 9.3 for data curation and *R* version 3.2.4 for statistical analyses. *R* packages used for challenge evaluation included *survival* version 2.38-3, *ROCR* version 1.0-7, *timeROC* version 0.3, and *Bolstad2* version 1.0-28. The top-performing model also used *glmnet* version 2.0-5 and *hamlet* version 0.9.4-2.

Clinical trial data used in the prostate cancer DREAM challenge can be accessed online.[24] Write-ups, model code, and predictions for all teams are reported in the appendix (pp 7, 8). Challenge documentation, including a detailed description of its design, overall results, scoring scripts, and the clinical trials data dictionary can be accessed via the Synapse platform.

Role of the funding source

Project Data Sphere had a collaborative role in design and logistics of the DREAM challenge but played no part in data collection, data analysis, and data interpretation or in the writing of this report. Sanofi US Services provided an in-kind contribution of human resources for curation of the raw datasets for the DREAM challenge and for clinical and scientific support of the challenge organisation, at the request of Project Data Sphere. Sanofi personnel participated in design of the DREAM challenge, in data analysis and data interpretation, and in writing of the report, but had no role in data collection. Raw clinical trial datasets for ASCENT2, MAINSAIL, and VENICE were available on the Project Data Sphere platform and were accessible by all registered users of Project Data Sphere, including all DREAM challenge participants and organisers, throughout the challenge. JG, TW, KKW, BMB, LS, KA, YX, FLZ, and JCC had access to raw data for ENTHUSE 33. JG, TW, KKW, LS, KA, FLZ, and JCC had access to raw data for ENTHUSE M1, during the post-challenge analysis. Data for ENTHUSE 33 and ENTHUSE M1 have been made freely accessible through the Project Data Sphere platform with publication of this report. The corresponding author had full access to all data in the study and had final responsibility for the decision to submit for publication.

Results

The overall DREAM challenge design is shown in figure 1, with full details in the appendix (p 4). The table presents baseline characteristics of patients in the five clinical trials included in this analysis. The training dataset included: 476 individuals from ASCENT2; 526 participants in MAINSAIL; and 598 men from VENICE. The validation dataset consisted of 470 patients from the ENTHUSE 33 trial; 528 men were initially enrolled to that trial but, because of regulatory restrictions in one country, data for 58 individuals were not made public through the challenge. The second validation dataset comprised 266 patients from ENTHUSE M1. Because of the same regulation restriction mentioned for ENTHUSE 33, some data were not provided to Project Data Sphere.

129 clinical baseline variables were measured for laboratory values, lesion site, previous medicines, medical history, and vital signs. When combined and assessed, the clinical variables for each trial were similar (appendix p 13), although when binary variables—mainly

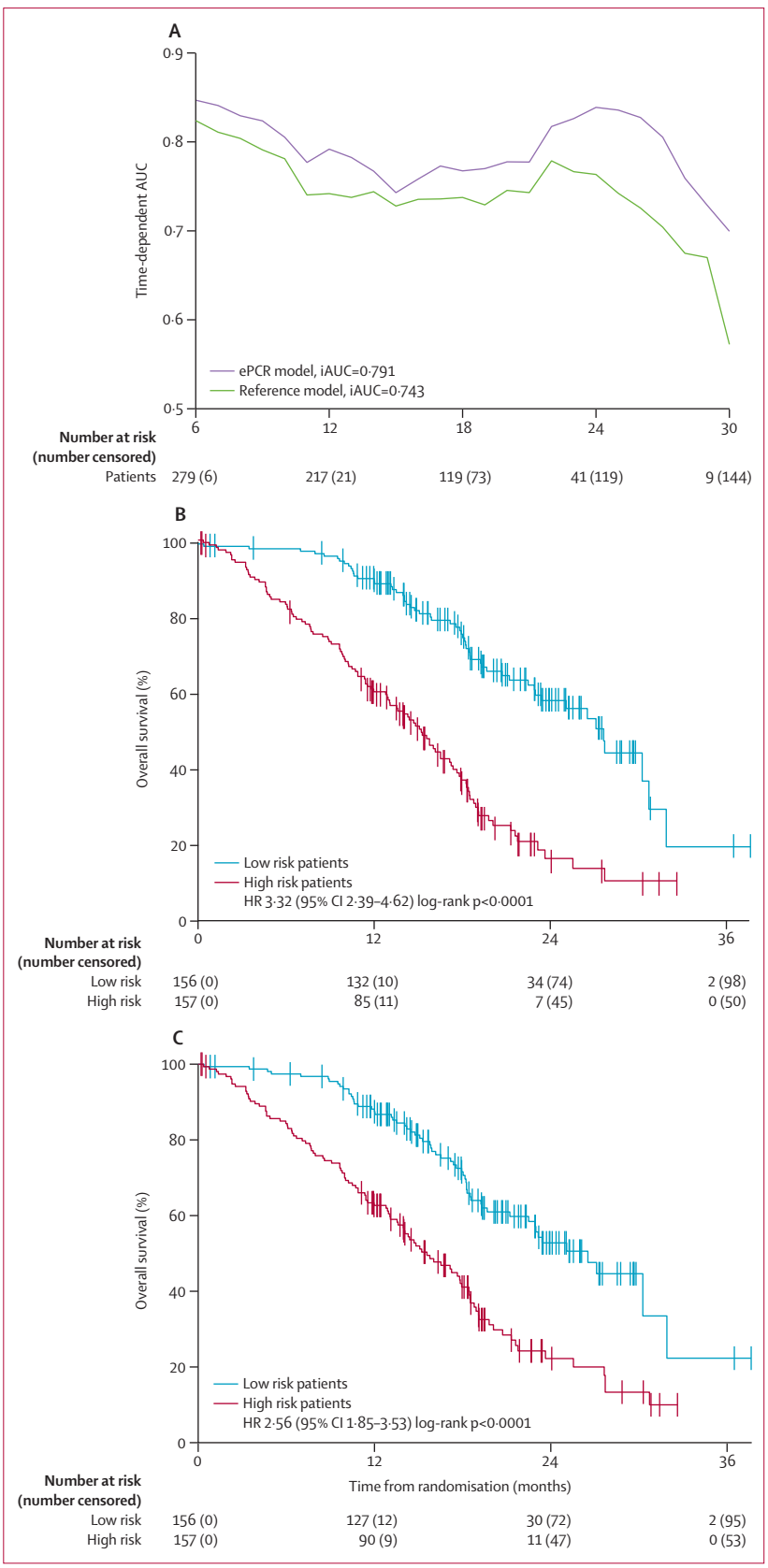


Figure 2: **Performance of ePCR model, using data from ENTHUSE 33**
(A) Time-dependent AUC was measured from 6 months to 30 months at 1-month intervals, reflecting the performance of predicting overall survival at different timepoints. (B, C) Overall survival was assessed by the Kaplan-Meier method, stratified by the median in the top-performing ePCR model (B) and the reference model (C). The log-rank test was used to compare risk groups. ePCR=ensemble of penalised Cox regression models. iAUC=integrated time-dependent area under the curve. HR=hazard ratio.

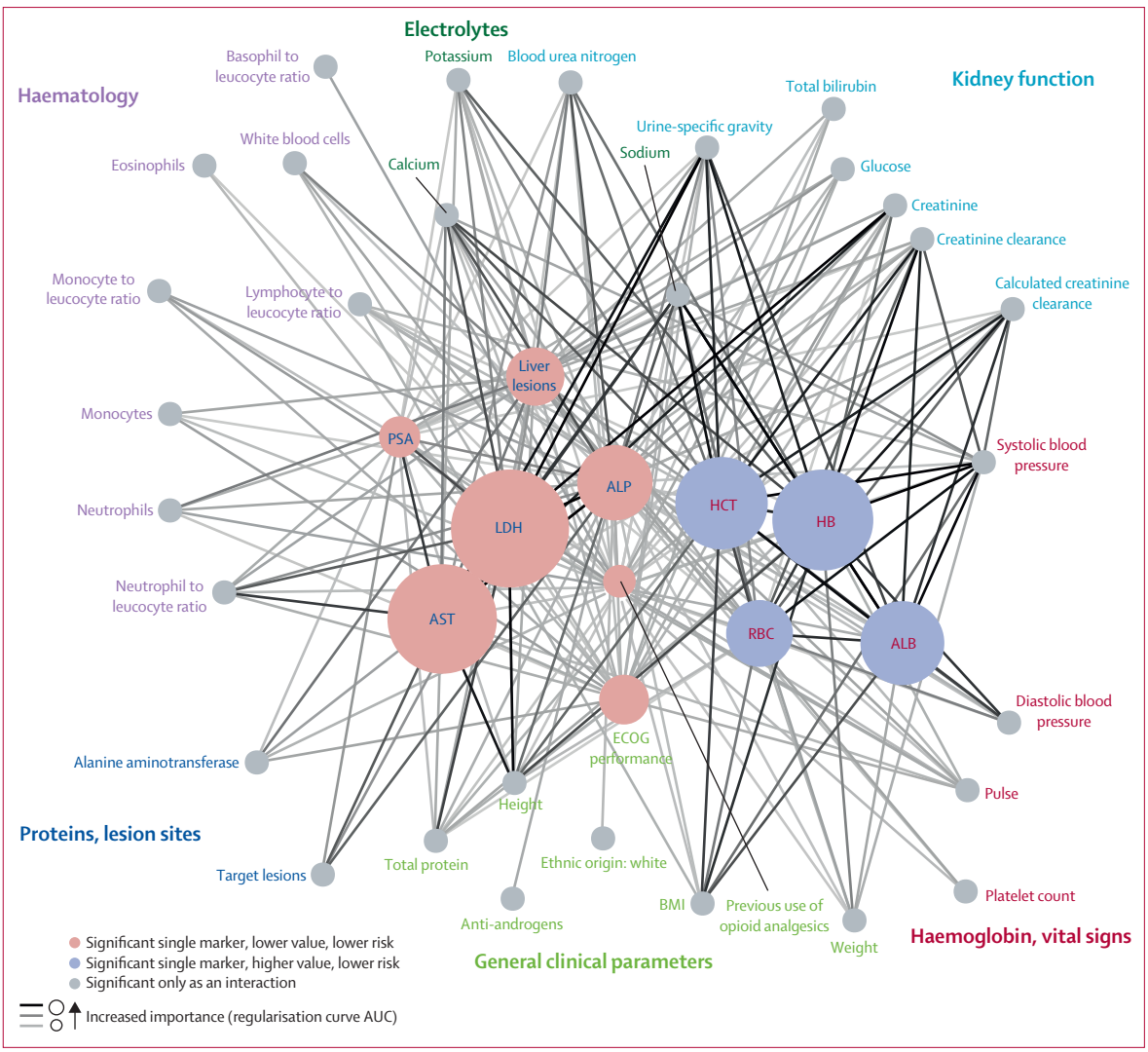


Figure 3: **Projection of the most important variables and interactions in the ePCR model**
Automated data-driven network layout of the most significant model variables, according to their interconnections with other model variables. Node size and colour indicate the importance of the variable alone for prediction of overall survival and its coefficient sign, respectively. This importance was calculated as the area under the curve (AUC) of the penalised model predictors, as a function of penalisation parameter λ. Edge colour indicates the importance of an interaction between two model variables, with a darker colour corresponding to a stronger interaction effect. Coloured subnetwork modules annotate the variables based on expert curated categories. Variable and interaction statistics can be found in the appendix (pp 10, 11). ALB=albumin. ALP=alkaline phosphatase. AST=aspartate aminotransferase. BMI=body-mass index. ECOG=Eastern Cooperative Oncology Group. ePCR=ensemble of penalised Cox regression models. HB=haemoglobin. HCT=haematocrit. LDH=lactate dehydrogenase. PSA=prostate-specific antigen. RBC=red blood cell count.

representing lesion sites—were judged separately, differences in clinical trials were recorded (appendix p 13). ASCENT2 had a lower frequency of patients with visceral metastases (1·1% liver and 1·7% lung) compared with individuals in the other three trials (10–14% liver, 11–15% lung). By contrast, the proportion of patients with bone metastases was high across the four trials (72–100%). Median follow-up differed among the four studies: 11·7 months (IQR 8·6–15·8) in ASCENT2; 9·2 months (6·4–13·1) in MAINSAIL; 21·1 months (12·9–29·6) in VENICE; and 15·3 months (10·9–20·8) in ENTHUSE 33. Risk profiles for each of the trials—specifically, mortality—were similar among the four trials (proportionality of hazards, p>0·5; appendix p 14). The proportion of patients who died in each trial was 138 (29%) of 476 in ASCENT2, 92 (17%) of 526 in MAINSAIL, 433 (72%) of 598 in VENICE, and 255 (54%) of 470 in ENTHUSE 33.

50 international teams—comprising 163 individuals—submitted predictions from their models to the challenge; with the reference model, the total number of models is 51. The distribution of all team scores by iAUC is shown in the appendix (p 15). The top-performing model was developed by a collaborative team from the

Institute for Molecular Medicine Finland and the University of Turku. The method was based on an ensemble of penalised Cox regression (ePCR) models. The ePCR model extended beyond the LASSO-based reference model by using an elastic net to select additional correlated groups of clinical variables and their interactions, modelled as interaction terms (panel). The risk predictions from the trial-specific ensemble components were rank-averaged to produce the final ensemble risk score predictions and to avoid trial-specific variation.

The top-scoring ePCR model reported an iAUC of 0·791 and outscored all other teams, with a Bayes factor greater than 5, surpassing the threshold that defines significantly different performances (Bayes factor >3). The reference model achieved an iAUC of 0·743, with a significant difference in scores between the ePCR model and the reference model (Bayes factor >20). With a time-dependent AUC metric, the ePCR model outperformed the reference model at every timepoint, with the biggest difference in performance at later timepoints between 18 and 30 months (figure 2A). A median split of patients into low-risk and high-risk groups for the ePCR model resulted in a low-risk group comprising 156 patients and 56 deaths (median follow-up 27·6 months [IQR 18·2–31·9]) and a high-risk group containing 157 patients and 107 deaths (15·1 [8·5–20·1]). Similarly for the reference group, a low-risk group including 156 patients and 59 deaths (median follow-up 26·5 months [IQR 17·2–31·9]) and a high-risk group with 157 patients and 104 deaths (15·6 [8·6–21·8]) were generated. Kaplan-Meier analysis showed that low-risk and high-risk groups had significantly different overall survival in each model (ePCR, HR 3·32, 95% CI 2·39–4·62, p<0·0001; reference, 2·56, 1·85–3·53, p<0·0001; figure 2B, 2C). A full comparison is provided in the appendix (p 9). We assessed the calibration of the ePCR model by comparing predicted probabilities versus actual probabilities at multiple timepoints (appendix p 16).

Figure 3 shows a network visualisation of the significant groups of variables identified in the ePCR model and their predictive relations, based on the importance of the model covariates and their interactions. Although many of the variables used in the reference model were also included in the ePCR model, aspartate aminotransferase was identified as a new important predictor. We also recorded a number of factors that were included as interaction terms, and of particular note were those reflecting the immunological or renal function of the patient. Prostate-specific antigen was an independent but weak prognostic factor that interacted strongly with lactate dehydrogenase and aspartate aminotransferase.

In addition to identifying the top-performing model, the challenge also tested the other independent models, with 30 of 50 outperforming the reference model (Bayes factor >3; appendix p 15). We performed hierarchical clustering of risk scores from the 51 models to identify three distinct risk groups (figure 4A), with 98 patients (77 deaths) in group A (high risk), 131 patients (61 deaths) in group B (moderate risk), and 84 patients (25 deaths) in group C (low risk). Differences in overall survival among these three groups were significant (log-rank p<0·0001), with median overall survival of 12·9 months (95% CI 10·7–15·3) for group A, 20·8 months (18·3–25·6) for group B, and 27·7 months (26·6–not available) for group C (figure 4B).

40 of 50 teams provided a list of common clinical factors that were incorporated into their final models; the frequencies with which a feature was reported as being important or significant in a team's model are summarised in the appendix (p 18). The results not only confirmed the variables identified previously in the reference model but also highlighted several factors that were not. Of note, aspartate aminotransferase was included in more than half the team models. Other novel variables that were included in at least 15% of the

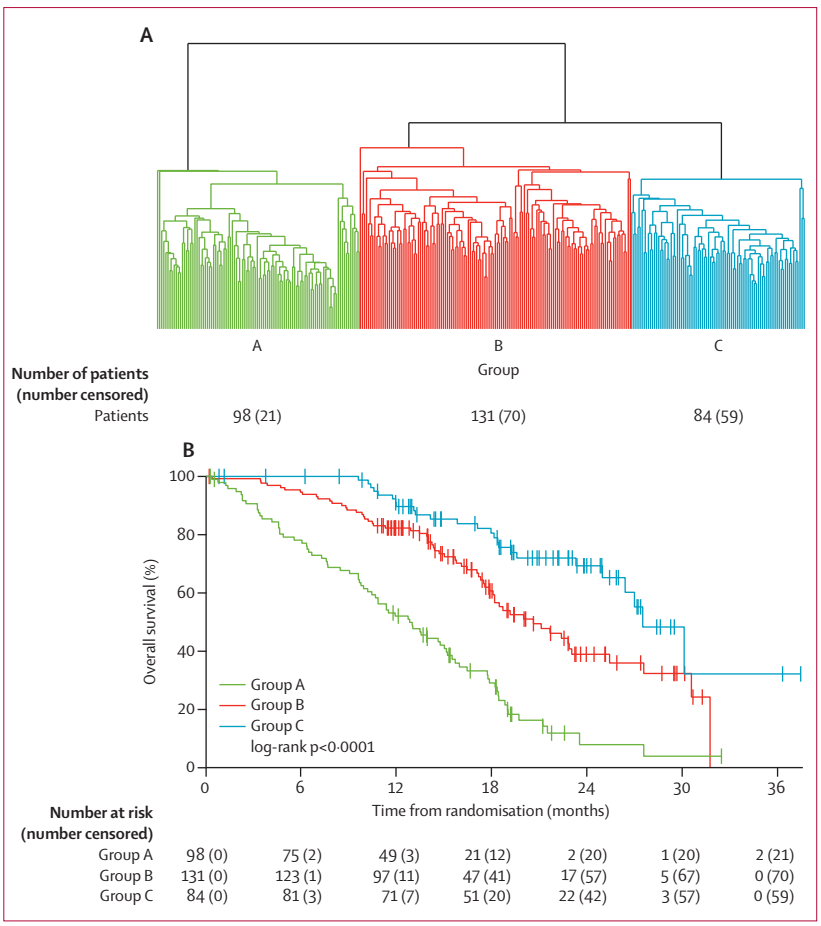


Figure 4: **Challenge meta-analysis**
(A) Hierarchical clustering of patients (Euclidean distance, average linkage) by rank-normalised prediction scores from all 51 models using the ENTHUSE 33 data. (B) Kaplan-Meier plot of survival probability for the three patient clusters from (A). Group A=high risk. Group B=moderate risk. Group C=low risk.

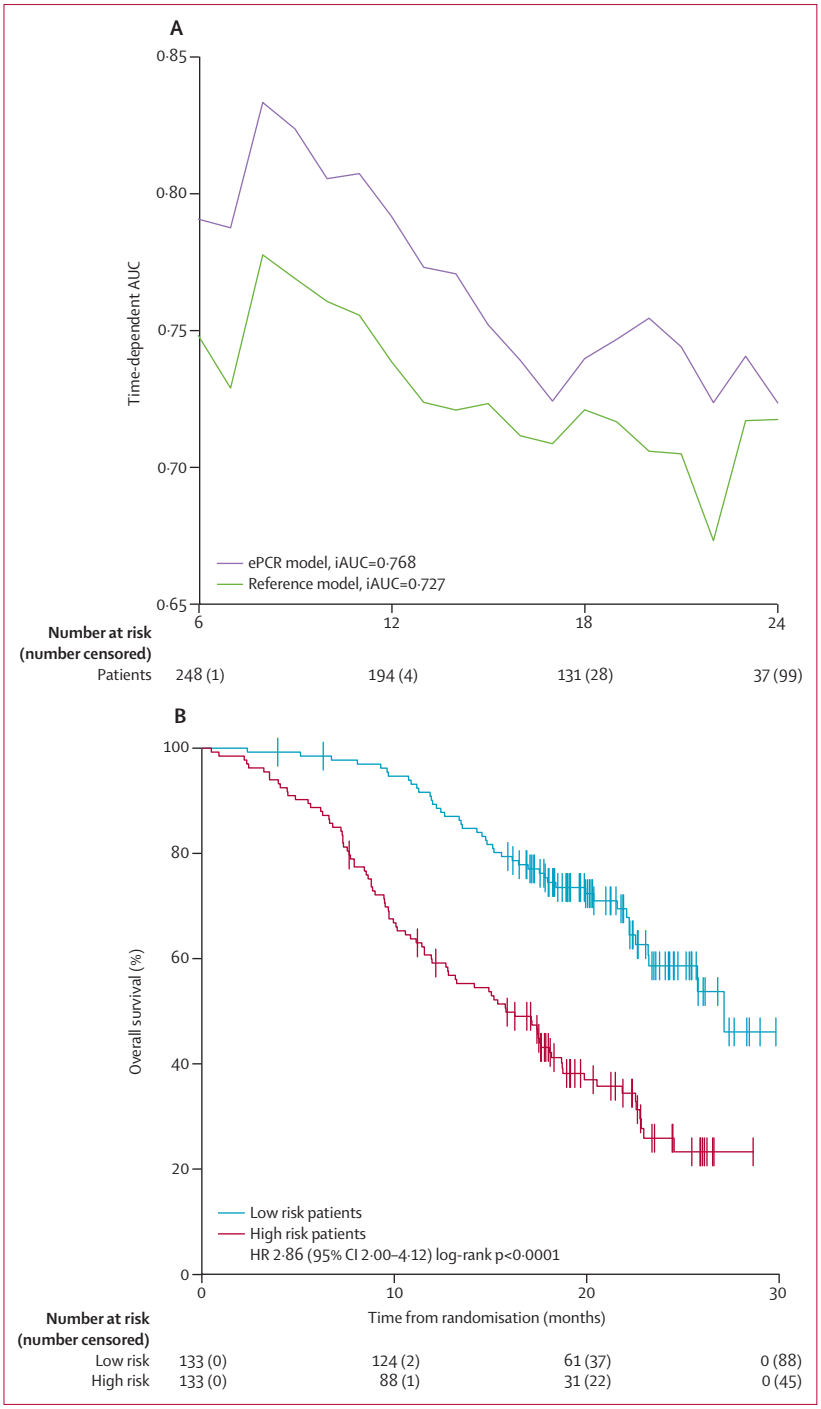


Figure 5: **Performance of ePCR model, using data from ENTHUSE M1**
(A) Time-dependent AUC was measured from 6 months to 24 months at 1-month intervals, reflecting the performance of predicting overall survival at different timepoints. The top-performing model (ePCR) is shown compared with the reference model. (B) Overall survival was assessed by the Kaplan-Meier method, stratified by median risk score. The log rank test was used to compare risk groups. ePCR=ensemble of penalised Cox regression models. iAUC=integrated time-dependent area under the curve. HR=hazard ratio.

models are total white blood cell count, absolute neutrophil count, red blood cell count, region of the world, body-mass index, and creatinine.

Application of the ePCR and reference models to the ENTHUSE M1 dataset showed model performances comparable with the primary challenge, with an iAUC of 0·768 for the ePCR model and 0·727 for the reference model (figure 5A). A median split of risk scores in the ePCR model led to a high-risk group of 133 patients, of which 45 were right-censored, and a low-risk group of 133 patients, of which 88 were right-censored. Kaplan-Meier analysis of the ENTHUSE M1 data showed significant separation of the high-risk and low-risk predicted patients (p<0·0001), with median survival of 15·8 months (95% CI 12·8–18·7) for high-risk patients and 27·1 months (23·2–not available) for low-risk patients (figure 5B).

Discussion

The prostate cancer DREAM challenge resulted in one prognostic model to predict overall survival significantly outperforming all other methods, including a reference model reported by Halabi and colleagues,[14] and led to a network perspective of predictive biological variables and their interactions. The results from the top-performing team's model pointed to important interaction effects with immune biomarkers and markers of hepatic function (potentially reflected in the increased amounts of aspartate aminotransferase) and renal function. The network visualisation of the prediction model suggests a complex relation and dependency structure among many of the predictive clinical variables. Many of these noted interactions, although not significant as independent variables, might be important modulators of key clinical traits—eg, haematology-related measurements such as haemoglobin and haematocrit. Although further investigation is necessary to determine the clinical implication of these associations and provide new insights into tumour–host interaction, these findings shed light on the complex and interwoven nature of prognostic factors on patients' survival.

Open-data, crowdsourced, scientific challenges have been highly effective at drawing together large cross-disciplinary teams of experts to solve complex problems.[25–30] To our knowledge, this DREAM challenge represented the first public collaborative competition[31] to use open-access registration trial datasets in cancer with the intention of improving outcome predictions. In total, 163 individuals comprising 50 teams participated in the challenge, applying state-of-the-art machine learning and statistical modelling methods. The contribution of five clinical trial datasets from industry and academic institutions to Project Data Sphere, and their subsequent use in an open challenge, enabled the advancement of prognostic models in metastatic castration-resistant prostate cancer that up to now was not possible. Modellers had access to several independent clinical trial cohorts with subtle differences in eligibility that increased the diversity (heterogeneity) of the total patient population considered for model development. Access was also

provided to data for 150 independent and standardised variables over the trials; by contrast, only 22 variables were considered by the reference model.[14] The challenge resulted in creative data-mining approaches that used standardised raw event-level tables, which are rarely leveraged for prognostic model development, and enabled innovative clinical features to be derived for modelling. Several teams—including the top-performing team—made use of these event-level tables. Finally, evaluation of the 50 methods (validated by an independent and neutral party) provided the most comprehensive assessment of prognostic models in metastatic castration-resistant prostate cancer. These results are both a benchmark for future prognostic model development and a rich source of information that can be mined for additional insights into both patients' stratification and the robustness of clinical predictive factors.

This study has shown the benefits of open data access at a time when clinicians, researchers, and the public are advocating for improved platforms and policies that encourage sharing of clinical trial data.[32,33] Project Data Sphere has overcome major barriers to data sharing with support of data providers, to allow broad access to cancer clinical trial data. To researchers who are interested in leveraging open-access cancer trial data, this study represents a novel research approach that encompassed scientific rigor and a deep understanding of clinical data through effective collaboration of multidisciplinary teams of experts. The top-performing ePCR model was free of any a-priori clinical assumptions, with the exception of exclusion of non-relevant variables in early data curation. The data-driven modelling process identified automatically the best combination of predictors through cross-validation. Furthermore, the ePCR modelling process was fully agnostic to the variables used in the previous reference model; however, many of the same predictors were identified, in addition to novel ones. Such data-driven, unbiased modelling approaches can mine effectively the predictive variables and their combinations from large-scale and open clinical trial data.

The trials used here represent the standard of care at the time when the trials were done, which is a limitation of this study. Since 2010, several treatments have become available, for use both before and after first-line chemotherapy, and new trials have changed the way clinicians approach this disease.[34] Abiraterone and enzalutamide—both approved for first-line treatment of metastatic castration-resistant prostate cancer—are not included within the scope of this challenge because of a limitation of control arm data; both COU-AA-302[5] and PREVAIL[10] have placebo or prednisone controls, and comparative trials using these agents as control have not been done. Accordingly, trial sponsors should be encouraged to contribute data from the experimental arm (particularly for approved drugs) to an active and engaged research community. Although sponsors are concerned that virtual comparisons might be made between treatments in experimental arms of different trials, there is far more benefit in leveraging these data to validate prognostic factors and models and to investigate intermediate clinical endpoints predictive of survival.

The DREAM challenge described here has shown that there is opportunity to further optimise prognostic models in metastatic castration-resistant prostate cancer using baseline clinical variables. For substantial advances beyond the work presented here, clinical trial data must be made available that reflects current advancements in treatment paradigms, including new data-capture techniques such as genomics, immunogenomics, and metabolomics that might more accurately describe the malignant state of the tumour and its microenvironment. Vital to either of these will be the need to share patient-level oncology data with the research community for the development of the next generation of prognostic and predictive models in cancer.

Contributors

JG, TW, JCB, ECN, TY, BMB, KA, TN, SF, GS, HS, CJS, CJR, HIS, OS, YX, FLZ, and JCC designed the DREAM Challenge. FLZ and LS led the efforts by Project Data Sphere to obtain and process the clinical trial data. TDL, SAK, GP, AA, TP, TM, and TA designed the top-performing method. JG, TW, TDL, KKW, JCB, ECN, GS, TA, FLZ, and JCC did the post-challenge data analysis and interpretation. HS, CJS, CJR, HIS, and OS assisted in clinical variable interpretation. All members of the Prostate Cancer Challenge DREAM Consortium (appendix pp 20–22) submitted prognostic models to the Challenge and provided method write-ups and the code to reproduce their predictions. JG, TW, TDL, HS, CJS, CJR, HIS, OS, TA, FLZ, and JCC wrote the report.

Declaration of interests

CJS reports personal fees from Sanofi, Janssen, Astellas, and Bayer, outside the submitted work. FLZ and LS are employed by and have stock in Sanofi. KA is employed by and has stock in AstraZeneca, outside the submitted work. HIS reports non-financial support from AstraZeneca, Bristol-Myers Squibb, Ferring Pharmaceuticals, Medivation, and Takeda Millennium, outside the submitted work; consultancy fees from Astellas, BIND Pharmaceuticals, Blue Earth Diagnostics, Clovis Oncology, Elsevier (PracticeUpdate website), Genentech, Med IQ, Merck, Roche, Sanofi Aventis, WCG Oncology, and Asterias Biotherapeutics, outside the submitted work; and grants from Illumia, Innocrin Pharma, Janssen, and Medivation, outside the submitted work. GP reports grants from the Academy of Finland (decision number 265966), during the conduct of the study and outside the submitted work. SAK reports grants from the Academy of Finland (decision number 296516), during the conduct of the study and outside the submitted work. TDL reports research grants from the Finnish Cultural Foundation and Drug Research Doctoral Programme, during the conduct of the study; and a research contract from the National Cancer Institution, during the conduct of the study. TA reports grants from the Academy of Finland and the Cancer Society of Finland, during the conduct of the study; and a research contract from the National Cancer Institute, during the conduct of the study. OS reports grants and personal fees from Sanofi, outside the submitted work. AA, JCB, BMB, JCC, SF, JG, TM, ECN, TN, TP, CJR, HS, GS, TW, KKW, TY, and YX declare no competing interests.

Acknowledgments

This report is based on research using information obtained from Project Data Sphere. Neither Project Data Sphere nor the owners of any information from the website have contributed to, approved, or are in any way responsible for the contents of this report. We thank the Sage Bionetworks Synapse team for the development and design of the DREAM challenge website. This work is supported in part by: the National Institutes of Health, National Library of Medicine (2T15-LM009451), National Cancer Institute (16X064; 5R01CA152301), Boettcher Foundation, Academy of Finland (grants 265966, 296516, 292611, 269862, 272437,

279163, 295504), Cancer Society of Finland, Drug Research Doctoral Programme (DRDP) at the University of Turku, and Finnish Cultural Foundation. Sanofi US Services provided an in-kind contribution of human resources for curation of raw datasets for the challenge and for clinical and scientific support of challenge organisation, at the request of Project Data Sphere.

References

1 Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin* 2011; **61:** 69–90.

2 Tanimoto T, Hori A, Kami M. Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *N Engl J Med* 2010; **363:** 1966.

3 Berruti A, Pia A, Terzolo M. Abiraterone and increased survival in metastatic prostate cancer. *N Engl J Med* 2011; **365:** 766.

4 Fizazi K, Carducci M, Smith M, et al. Denosumab versus zoledronic acid for treatment of bone metastases in men with castration-resistant prostate cancer: a randomised, double-blind study. *Lancet* 2011; **377:** 813–22.

5 Ryan CJ, Smith MR, de Bono JS, et al. Abiraterone in metastatic prostate cancer without previous chemotherapy. *N Engl J Med* 2013; **368:** 138–48.

6 Scher HI, Fizazi K, Saad F, et al. Increased survival with enzalutamide in prostate cancer after chemotherapy. *N Engl J Med* 2012; **367:** 1187–97.

7 de Bono JS, Oudard S, Ozguroglu M, et al, for the TROPIC investigators. Prednisone plus cabazitaxel or mitoxantrone for metastatic castration-resistant prostate cancer progressing after docetaxel treatment: a randomised open-label trial. *Lancet* 2010; **376:** 1147–54.

8 Parker C, Nilsson S, Heinrich D, et al. Alpha emitter radium-223 and survival in metastatic prostate cancer. *N Engl J Med* 2013; **369:** 213–23.

9 Kantoff PW, Higano CS, Shore ND, et al. Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *N Engl J Med* 2010; **363:** 411–22.

10 Beer TM, Armstrong AJ, Rathkopf DE, et al. Enzalutamide in metastatic prostate cancer before chemotherapy. *N Engl J Med* 2014; **371:** 424–33.

11 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:** 1232–37.

12 Smaletz O, Scher HI, Small EJ, et al. Nomogram for overall survival of patients with progressive metastatic prostate cancer after castration. *J Clin Oncol* 2002; **20:** 3972–82.

13 Armstrong AJ, Garrett-Mayer ES, Yang Y-CO, de Wit R, Tannock IF, Eisenberger M. A contemporary prognostic nomogram for men with hormone-refractory metastatic prostate cancer: a TAX327 study analysis. *Clin Cancer Res* 2007; **13:** 6396–403.

14 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:** 671–77.

15 Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet* 2014; **10:** e1004754.

16 Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic classification and prognosis in acute myeloid leukemia. *N Engl J Med* 2016; **374:** 2209–21.

17 Longo DL, Drazen JM. Data sharing. *N Engl J Med* 2016; **374:** 276–77.

18 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:** 2191–98.

19 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.

20 Tannock IF, Fizazi K, Ivanov S, et al, on behalf of the VENICE investigators. 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:** 760–68.

21 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:** 1740–47.

22 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). *Proc Am Soc Clin Oncol* 2011; **29:** abstr 117.

23 Hung H, Chiang C-T. Estimation methods for time-dependent AUC models with survival data. *Can J Stat* 2010; **38:** 8–26.

24 Project Data Sphere. Prostate cancer DREAM challenge. https://www.projectdatasphere.org/projectdatasphere/html/pcdc (accessed Oct 21, 2016).

25 Costello JC, Stolovitzky G. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clin Pharmacol Ther* 2013; **93:** 396–98.

26 Bender E. Challenges: crowdsourced solutions. *Nature* 2016; **533:** S62–64.

27 Bansal M, Yang J, Karan C, et al. A community computational challenge to predict the activity of pairs of compounds. *Nat Biotechnol* 2014; **32:** 1213–22.

28 Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 2014; **32:** 1202–12.

29 Margolin AA, Bilal E, Huang E, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci Transl Med* 2013; **5:** 181re1.

30 Ewing AD, Houlahan KE, Hu Y, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* 2015; **12:** 623–30.

31 Saez-Rodriguez J, Costello JC, Friend SH, et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet* 2016; **17:** 470–86.

32 Merson L, Gaye O, Guerin PJ. Avoiding data dumpsters: toward equitable and useful data sharing. *N Engl J Med* 2016; **374:** 2414–15.

33 Bierer BE, Li R, Barnes M, Sim I. A global, neutral platform for sharing trial data. *N Engl J Med* 2016; **374:** 2411–13.

34 Lewis B, Sartor O. The changing landscape of metastatic prostate cancer. *Am J Hematol* 2015; **11:** 11–20.

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

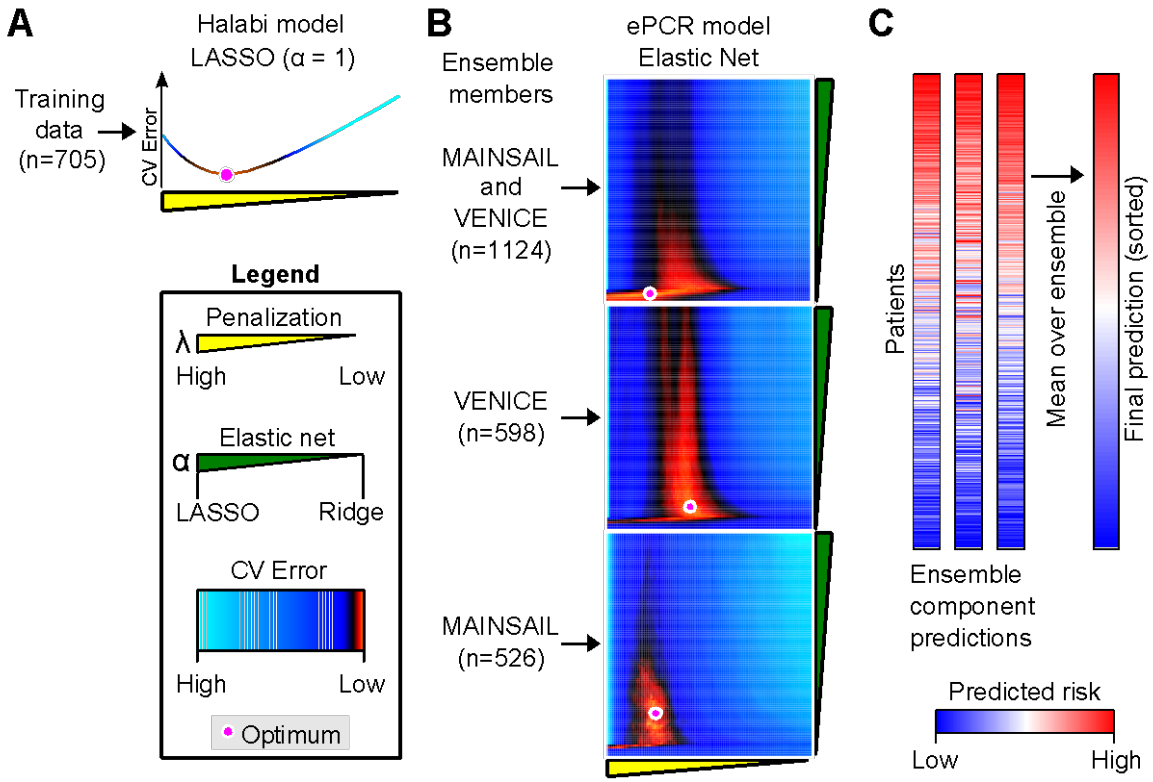
Supplementary Table 3. Top 15 single and interacting variables from the final ePCR model built from the MAINSAIL and VENICE trials. Comprehensive list of evaluated variables is available at: https://www.synapse.org/#!Synapse:syn7113819

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

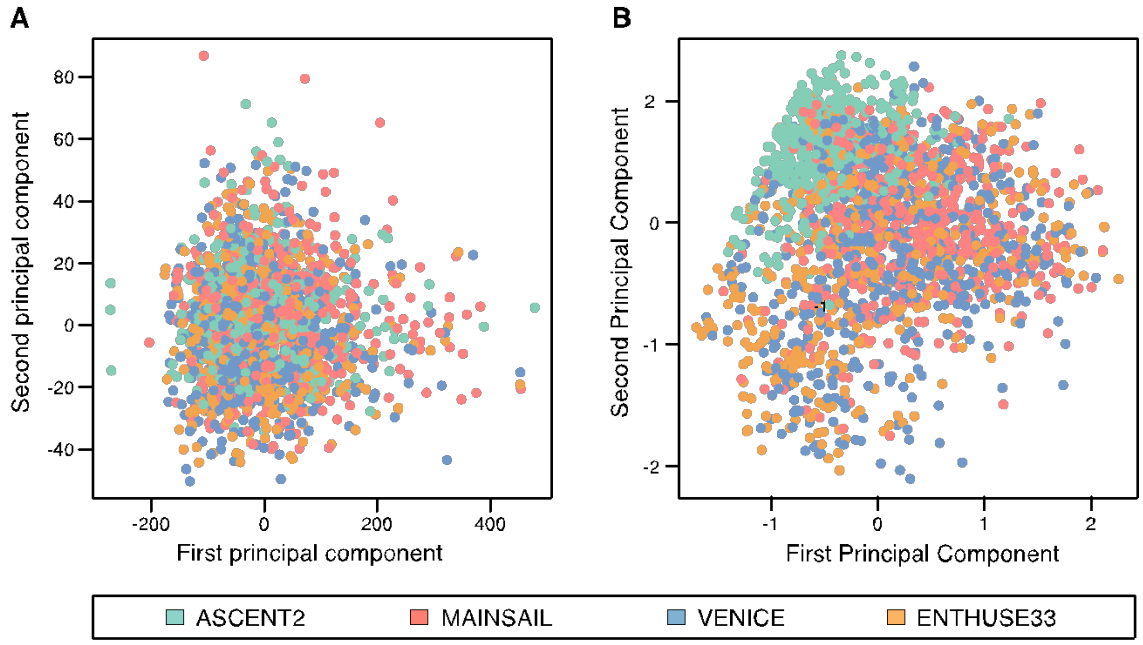
Top 15 interactions in the ePCR model		Ensemble p value	Ensemble effect size
AST	LDH	< 0.0001	3408.333
ALP	LDH	< 0.0001	3406.667
ALP	AST	< 0.0001	3404.333
HB	SYSTOLICBP	< 0.0001	3402.333
LDH	Urine Specific Gravity	< 0.0001	3400.667
SYSTOLICBP	HCT	< 0.0001	3400.333
Creatinine	LDH	< 0.0001	3397.333
LDH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

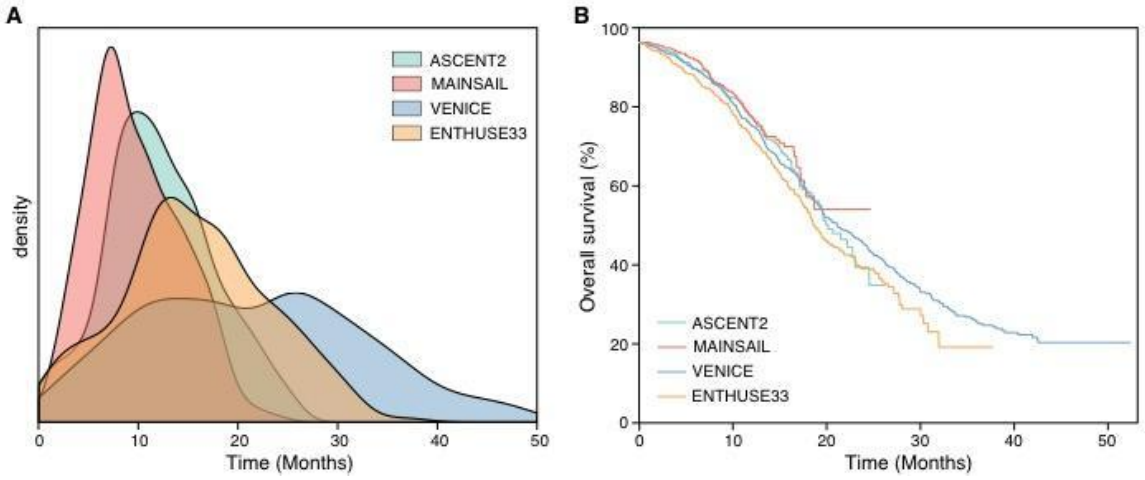
Supplementary Figures



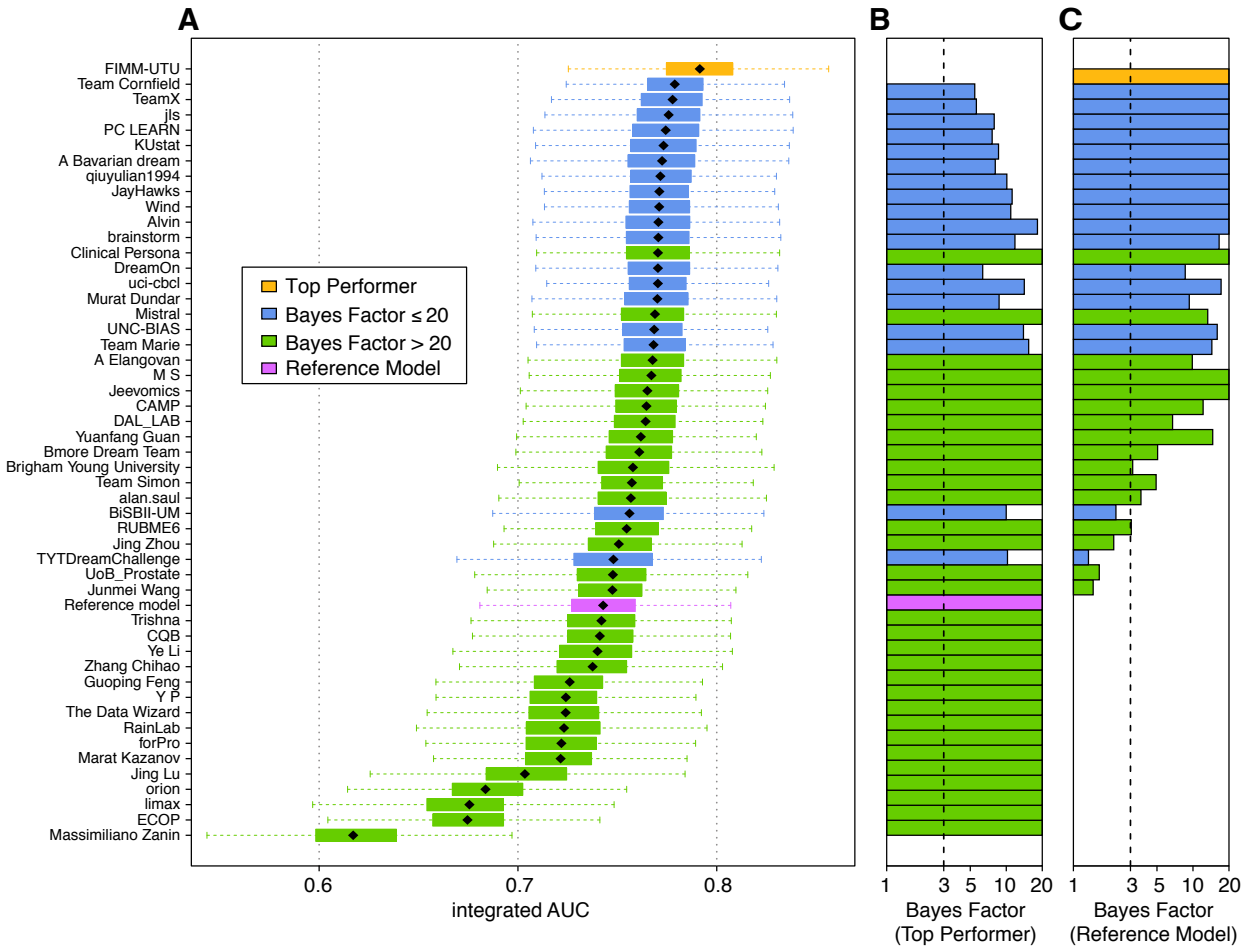
Supplementary Figure 1. Overview of the top-performing ePCR method in comparison to the Reference model (Halabi model). (A) The benchmarking Reference model explored the LASSO model (α = 1) in a training data cohort with respect to the regularization parameter (λ) using cross-validation (CV). (B) The 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.



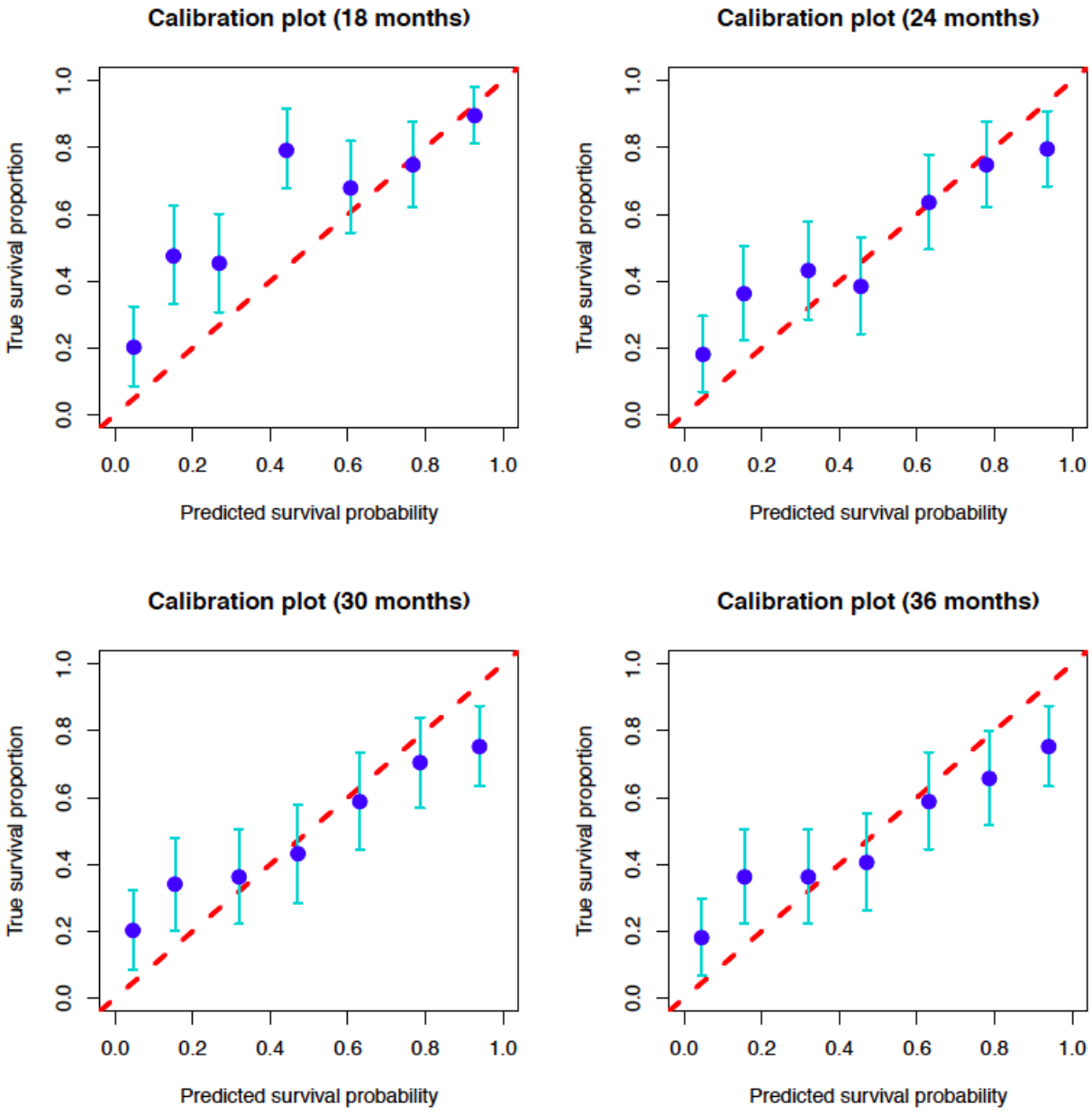
Supplementary Figure 2. (A) All data across ASCENT2, MAINSAIL, VENICE, and ENTHUSE 33– both binary and continuous data – were used in a PCA. (B) All data across the 4 studies – only binary variables – were used in PCA.



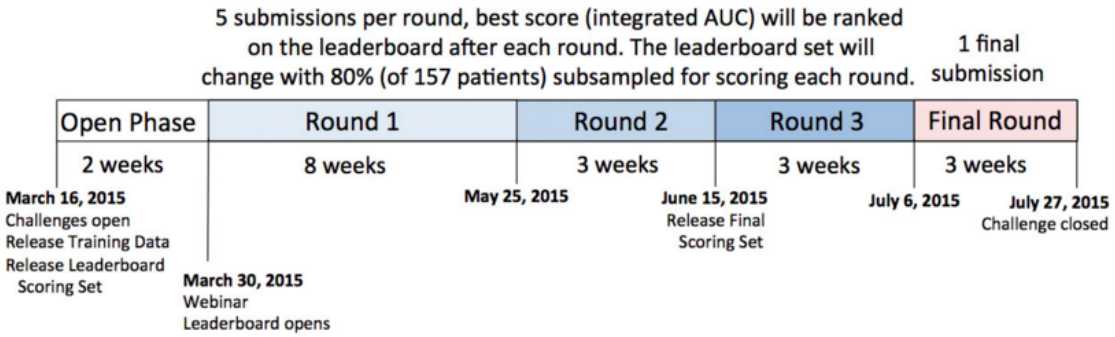
Supplementary Figure 3. (A) Density plot of follow-up times per study for the ASCENT2, MAINSAIL, VENICE, and ENTHUSE 33 trials. (B) Survival profile for each of the trials.



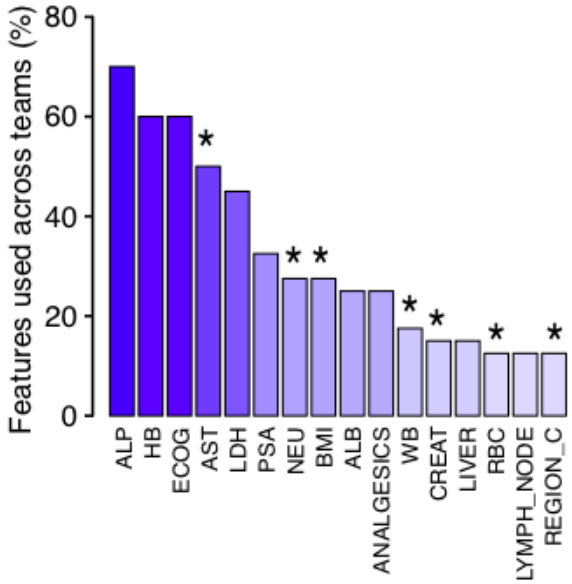
Supplementary Figure 4. Summary of Challenge results across all 50 teams plus the Reference model evaluated using the ENTHUSE 33 dataset. (A) Performance of submissions. Each submission underwent 1,000 paired bootstrap of final scoring patient set to calculate a Bayes factor against the top-performer 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.