**AI-based predictive biomarker discovery via contrastive learning retrospectively improves clinical trial outcome**

Arango-Argoty G, Bikiel DE, Sun GJ, Kipkogei E, Smith KM, Jacob E

**Supplementary Information**

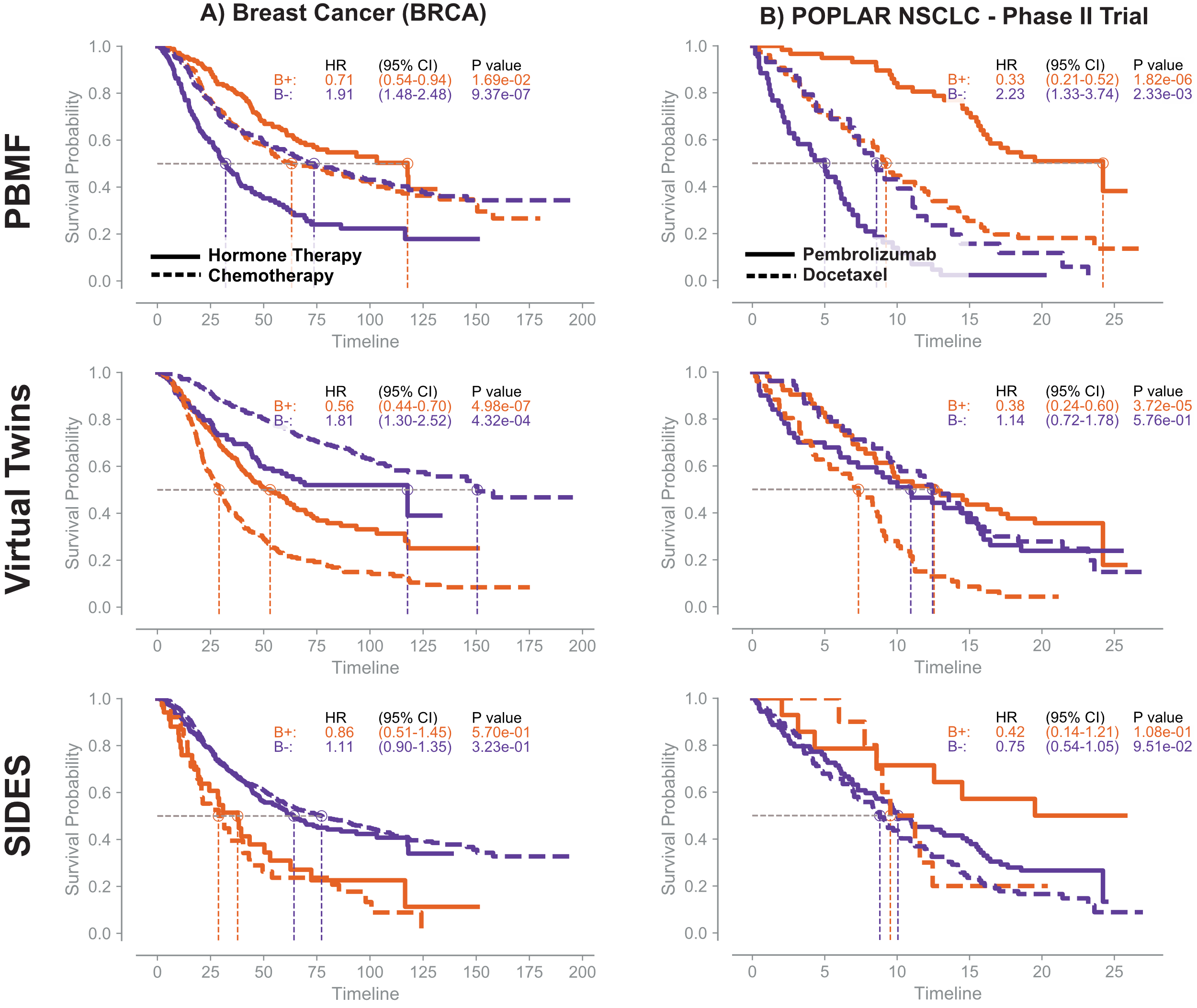
# Supplementary methods

## Creating uncorrelated covariate matrix for simulations

To properly control the relationship between covariates, we simulated random multivariate normal with specific covariance matrix. Although induced covariance may be interesting in some cases, we forced all the features to be fully uncorrelated by using an identity matrix as a covariance matrix. To ease the definitions of the hazard ratio group in the next stage, a binarization process was performed in each feature. To create a particular prevalence, the binarization was done at the specific percentile desired in each feature. This step was followed by multiplication of the features and treatment to the level needed (i.e., first-order interaction, two feature interactions or higher). The next step is the definition of the underlying hazard ratios structure, meaning the definition of how many different groups will be in the data and the relative hazard between them. Once the values of the parametric model are defined, the covariate extended matrix is used to create survival times by sampling from the inverted hazard function. Once the times are obtained, it is possible to increase the complexity of the data set by inducing extra noise in several ways. For example, while fully random features can be added at the beginning in the covariate matrix and then have their parameters in the hazard function set to 1, extra noise can be added to the predictive and prognostic features to transform the features from binary back to continuous.

# Optimizing Virtual Twins

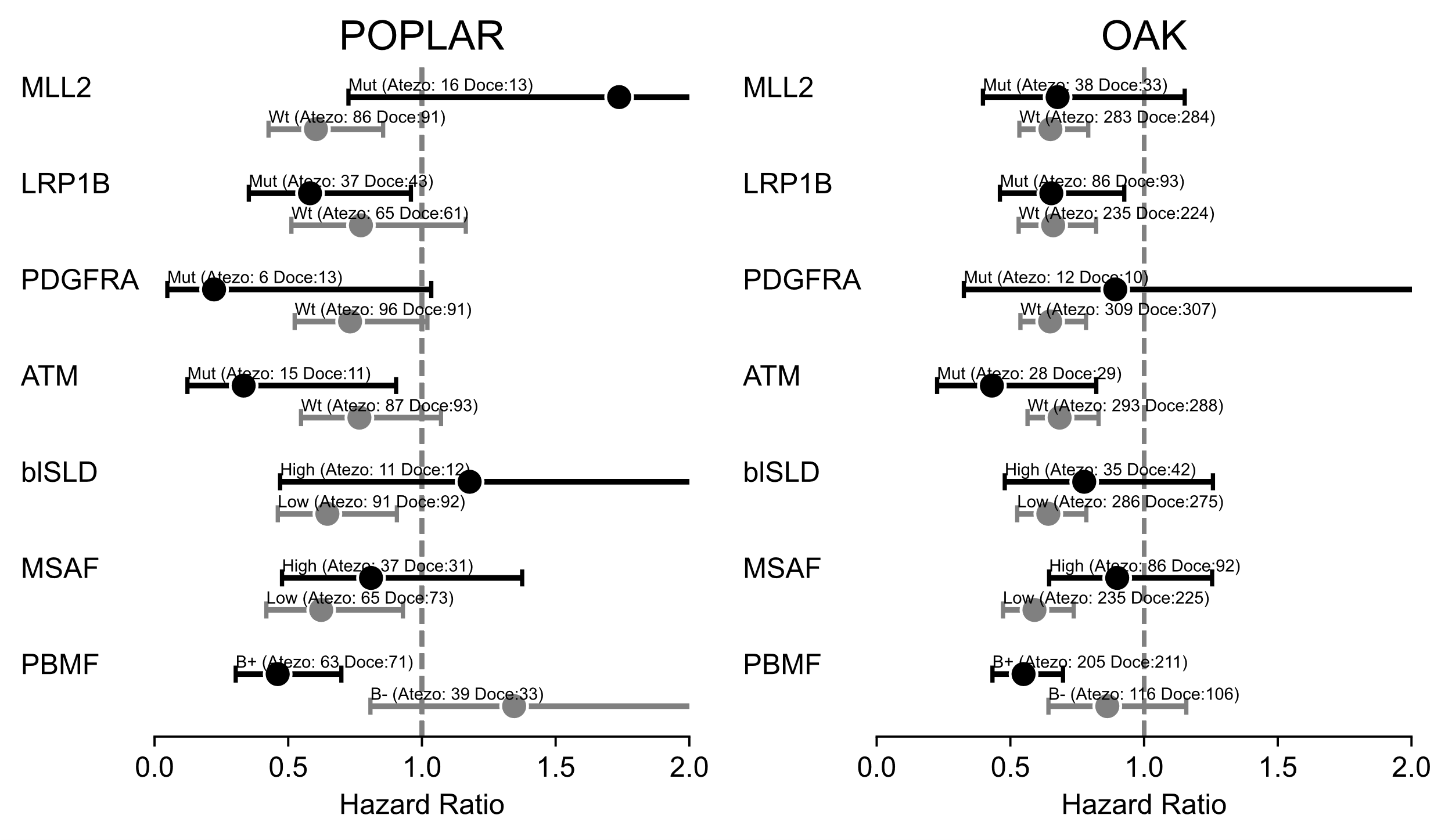
To perform a fair comparison between the performance of virtual twins and the PBMF, a grid search was conducted over various hyperparameters for the random forest used in the virtual twins model. Different values for the number of features chosen in each split, the maximum depth of each tree, and the number of trees in the forest were evaluated (Table S6). Similar data to the one used during the experiments was created. In this case a dataset containing the predictive biomarker (2 features) in addition to 3 random features was used to train and evaluate these models. For each combination of hyperparameters, 10 different splits of the training and test set were generated from 80% of the data. After evaluating the results for each model trained on a different split of the data, we recorded the mean and standard deviation of the AUPRC for both the training and test sets. The best-performing model from the search was trained with 4 features in each split, 200 trees in the ensemble, and a maximum depth of 5, which resulted in a training AUPRC of 0.946 ± 0.014 and a testing AUPRC of 0.933 ± 0.015.

****

**Fig. S1.** Training performance of the predictive biomarker modeling framework (PBMF) ensemble.

****

**Figure S2. a**,Distribution of ensemble versus pruned model scores. **b**, Using the pruned model, identify the best cutoff for getting high confidence biomarker pseudo-labels. **c**,Decision tree depicting a set of rules that simplifies the predictive biomarker. **d**, Estimation of the minimum number of patients required to validate a predictive biomarker discovered in a phase 2 (POPLAR) clinical trial. Data were subsampled from the OAK trial. **e**, Evaluation of consistency across endpoints. PBMF was trained on overall survival (OS) endpoint and evaluated by using the tree-derived predictive biomarker on progression-free survival (PFS).

****

**Figure S3** Forest plot depicting the predictive value of individual tree selected features. Cutoffs for continuous features are taken from the tree classifier. Mutational status was defined as wild type (*wt*) or mutated (*mut*).

**Table S1.** Model hyperparameters

|  |  |  |  |
| --- | --- | --- | --- |
| **Data set** | **Rotterdam and German breast cancer studies** | **POPLAR and OAK** | **Simulations** |
| Seed | 0 | 0 | 0 |
| Number of models | 100 | 100 | 128/1024 |
| Subset training data | 80% | 80% | 80% |
| Ignore patients during loss computation | 10% | 10% | 10% |
| Use only *n* features (for each model if using an ensemble) | 4 of 7 | 28 of 29 | \* |
| Learning rate | 0.01 | 0.01 | 0.01 |
| Layers [neurons] | [64] | [64] | [64] |
| Training epochs | 500 | 500 | 2000 |
| Minimum population | 0.75 | 0.0 | 0.5 |
| w1, w2 | 1, 1 | 1, 0\*\* | 1, 1 |
| L1, L2 normalization | 0, 0 | 0, 0 | 0, 0 |

\* Value varied depending on the simulation experiment run.

\*\* Prevalence loss deactivated.

**Table S2.** AUPRC on the test set for different training size comparing VT and PBMFa

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Size** | **VT (AUPRC)** | **PBMF (AUPRC)** | **PBMFp75 (AUPRC)** | **PBMFp90 (AUPRC)** |
| 250 | 0.752 ± 0.091 | 0.786 ± 0.066 | 0.785 ± 0.070 | 0.787 ± 0.068 |
| 500 | 0.803 ± 0.055 | 0.849 ± 0.065 | 0.844 ± 0.074 | 0.845 ± 0.073 |
| 1000 | 0.858 ± 0.029 | 0.918 ± 0.047 | 0.918 ± 0.060 | 0.917 ± 0.059 |
| 2000 | 0.858 ± 0.030 | 0.966 ± 0.028 | 0.964 ± 0.044 | 0.963 ± 0.045 |
| 4000 | 0.845 ± 0.026 | 0.985 ± 0.019 | 0.981 ± 0.038 | 0.980 ± 0.041 |

AUPRC, area under the precision-recall curve; VT, virtual twins.

a A total of 128 ensembles, 3 features from a data set containing 3 features. p75 and p90 are the corresponding models pruned to the top 25 and top 10 models in the ensemble.

**Table S3.** Patient stratification using most important tree-derived featuresa

| **HR** | **HR\_lo** | **HR\_hi** | ***P*** | **Atezo (*N*)** | | **Doce (*N*)** | | **Data set** | **Biomarker** | **Group** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.460543 | 0.303604 | 0.698606 | 2.65E-04 | 63 |  | 71 |  | POPLAR | PBMF | B+ |
| 1.344858 | 0.807394 | 2.240101 | 2.55E-01 | 39 |  | 33 |  | POPLAR | PBMF | B– |
| 0.549327 | 0.432999 | 0.696908 | 8.05E-07 | 205 |  | 211 |  | OAK | PBMF | B+ |
| 0.862457 | 0.642394 | 1.157906 | 3.25E-01 | 116 |  | 106 |  | OAK | PBMF | B– |
| 0.809564 | 0.47673 | 1.374771 | 4.34E-01 | 37 |  | 31 |  | POPLAR | MSAF | High |
| 0.62323 | 0.4182 | 0.928778 | 2.02E-02 | 65 |  | 73 |  | POPLAR | MSAF | Low |
| 0.899465 | 0.64495 | 1.254421 | 5.32E-01 | 86 |  | 92 |  | OAK | MSAF | High |
| 0.590117 | 0.47252 | 0.73698 | 3.30E-06 | 235 |  | 225 |  | OAK | MSAF | Low |
| 1.178311 | 0.469664 | 2.956195 | 7.27E-01 | 11 |  | 12 |  | POPLAR | blSLD | High |
| 0.64637 | 0.461001 | 0.906275 | 1.14E-02 | 91 |  | 92 |  | POPLAR | blSLD | Low |
| 0.776102 | 0.478992 | 1.257506 | 3.03E-01 | 35 |  | 42 |  | OAK | blSLD | High |
| 0.641929 | 0.52561 | 0.783989 | 1.39E-05 | 286 |  | 275 |  | OAK | blSLD | Low |
| 1.737477 | 0.725244 | 4.162495 | 2.15E-01 | 16 |  | 13 |  | POPLAR | MLL2 | Mut |
| 0.603687 | 0.42629 | 0.854906 | 4.47E-03 | 86 |  | 91 |  | POPLAR | MLL2 | Wt |
| 0.676654 | 0.397474 | 1.151929 | 1.50E-01 | 38 |  | 33 |  | OAK | MLL2 | Mut |
| 0.649714 | 0.533546 | 0.791174 | 1.78E-05 | 283 |  | 284 |  | OAK | MLL2 | Wt |
| 0.333364 | 0.123031 | 0.90328 | 3.08E-02 | 15 |  | 11 |  | POPLAR | ATM | Mut |
| 0.766136 | 0.547877 | 1.071342 | 1.19E-01 | 87 |  | 93 |  | POPLAR | ATM | Wt |
| 0.430885 | 0.226066 | 0.821272 | 1.05E-02 | 28 |  | 29 |  | OAK | ATM | Mut |
| 0.684175 | 0.564204 | 0.829655 | 1.14E-04 | 293 |  | 288 |  | OAK | ATM | Wt |
| 0.581416 | 0.352588 | 0.958752 | 3.36E-02 | 37 |  | 43 |  | POPLAR | LRP1B | Mut |
| 0.771924 | 0.511864 | 1.164113 | 2.17E-01 | 65 |  | 61 |  | POPLAR | LRP1B | Wt |
| 0.653204 | 0.460844 | 0.925856 | 1.67E-02 | 86 |  | 93 |  | OAK | LRP1B | Mut |
| 0.660094 | 0.530953 | 0.820647 | 1.84E-04 | 235 |  | 224 |  | OAK | LRP1B | Wt |
| 0.222601 | 0.04788 | 1.0349 | 5.53E-02 | 6 |  | 13 |  | POPLAR | PDGFRA | Mut |
| 0.731323 | 0.524079 | 1.020521 | 6.57E-02 | 96 |  | 91 |  | POPLAR | PDGFRA | Wt |
| 0.892299 | 0.326007 | 2.44227 | 8.24E-01 | 12 |  | 10 |  | OAK | PDGFRA | Mut |
| 0.648723 | 0.537572 | 0.782855 | 6.39E-06 | 309 |  | 307 |  | OAK | PDGFRA | Wt |

Atezo, atezolizumab; Doce, docetaxel.

aEndpoint for all data sets was overall survival.

**Table S4.** Efficacy comparison of PBMF against blood TMB (16 bp/Mb cutoff)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **HR** | **HR\_lo** | **HR\_hi** | ***P*** | **Atezo (*N*)** | **Doce (*N*)** | **Data set** | **Endpoint** | **Biomarker** | **Group** |
| 0.460543 | 0.303604 | 0.698606 | 2.65E-04 | 63 | 71 | POPLAR | OS | PBMF | B+ |
| 0.737623 | 0.509153 | 1.068614 | 1.08E-01 | 63 | 71 | POPLAR | PFS | PBMF | B+ |
| 1.344858 | 0.807394 | 2.240101 | 2.55E-01 | 39 | 33 | POPLAR | OS | PBMF | B– |
| 1.335188 | 0.813648 | 2.19103 | 2.53E-01 | 39 | 33 | POPLAR | PFS | PBMF | B– |
| 0.549327 | 0.432999 | 0.696908 | 8.05E-07 | 205 | 211 | OAK | OS | PBMF | B+ |
| 0.787426 | 0.640296 | 0.968365 | 2.35E-02 | 205 | 211 | OAK | PFS | PBMF | B+ |
| 0.862457 | 0.642394 | 1.157906 | 3.25E-01 | 116 | 106 | OAK | OS | PBMF | B– |
| 1.11819 | 0.849369 | 1.472092 | 4.26E-01 | 116 | 106 | OAK | PFS | PBMF | B– |
| 0.55635 | 0.311286 | 0.994344 | 4.78E-02 | 25 | 38 | POPLAR | OS | bTMB | High |
| 0.572667 | 0.330899 | 0.991082 | 4.64E-02 | 25 | 38 | POPLAR | PFS | bTMB | High |
| 0.766408 | 0.521741 | 1.12581 | 1.75E-01 | 77 | 66 | POPLAR | OS | bTMB | Low |
| 1.117139 | 0.78627 | 1.587241 | 5.36E-01 | 77 | 66 | POPLAR | PFS | bTMB | Low |
| 0.622256 | 0.429643 | 0.901219 | 1.21E-02 | 78 | 82 | OAK | OS | bTMB | High |
| 0.657654 | 0.469874 | 0.920479 | 1.46E-02 | 78 | 82 | OAK | PFS | bTMB | High |
| 0.671064 | 0.542226 | 0.830514 | 2.45E-04 | 243 | 235 | OAK | OS | bTMB | Low |
| 0.988343 | 0.817333 | 1.195133 | 9.04E-01 | 243 | 235 | OAK | PFS | bTMB | Low |

Atezo, atezolizumab; Doce, docetaxel; OS, overall survival; PFS, progression-free survival; bTMB, blood tumor mutation burden.

**Table S5.** Results of grid search for the optimum hyperparameters for the virtual twins model. Hyperparameters that gave the best performance in boldface.

| **No. of features** | **No. of trees** | **Maximum depth** | **AUPRC** | |
| --- | --- | --- | --- | --- |
| **Training set** | **Testing set** |
| 2 | 10 | 3 | 0.842 ± 0.034 | 0.819 ± 0.039 |
| 2 | 10 | 5 | 0.880 ± 0.024 | 0.857 ± 0.026 |
| 2 | 10 | 7 | 0.870 ± 0.031 | 0.840 ± 0.035 |
| 2 | 50 | 3 | 0.896 ± 0.030 | 0.882 ± 0.030 |
| 2 | 50 | 5 | 0.922 ± 0.021 | 0.907 ± 0.018 |
| 2 | 50 | 7 | 0.920 ± 0.019 | 0.904 ± 0.017 |
| 2 | 100 | 3 | 0.910 ± 0.023 | 0.894 ± 0.022 |
| 2 | 100 | 5 | 0.932 ± 0.017 | 0.914 ± 0.015 |
| 2 | 100 | 7 | 0.929 ± 0.018 | 0.911 ± 0.014 |
| 2 | 200 | 3 | 0.913 ± 0.024 | 0.897 ± 0.022 |
| 2 | 200 | 5 | 0.936 ± 0.017 | 0.918 ± 0.016 |
| 2 | 200 | 7 | 0.932 ± 0.017 | 0.915 ± 0.014 |
| 4 | 10 | 3 | 0.898 ± 0.025 | 0.881 ± 0.028 |
| 4 | 10 | 5 | 0.909 ± 0.022 | 0.893 ± 0.023 |
| 4 | 10 | 7 | 0.899 ± 0.024 | 0.884 ± 0.027 |
| 4 | 50 | 3 | 0.921 ± 0.019 | 0.908 ± 0.016 |
| 4 | 50 | 5 | 0.938 ± 0.012 | 0.928 ± 0.013 |
| 4 | 50 | 7 | 0.933 ± 0.014 | 0.923 ± 0.017 |
| 4 | 100 | 3 | 0.929 ± 0.020 | 0.917 ± 0.019 |
| 4 | 100 | 5 | 0.942 ± 0.013 | 0.932 ± 0.013 |
| 4 | 100 | 7 | 0.938 ± 0.014 | 0.927 ± 0.015 |
| 4 | 200 | 3 | 0.934 ± 0.02 | 0.92 ± 0.019 |
| **4** | **200** | **5** | **0.946 ± 0.014** | **0.933 ± 0.015** |
| 4 | 200 | 7 | 0.94 ± 0.015 | 0.927 ± 0.016 |
| 5 | 10 | 3 | 0.891 ± 0.030 | 0.873 ± 0.027 |
| 5 | 10 | 5 | 0.908 ± 0.023 | 0.89 ± 0.023 |
| 5 | 10 | 7 | 0.899 ± 0.024 | 0.878 ± 0.027 |
| 5 | 50 | 3 | 0.923 ± 0.019 | 0.912 ± 0.019 |
| 5 | 50 | 5 | 0.938 ± 0.014 | 0.929 ± 0.016 |
| 5 | 50 | 7 | 0.933 ± 0.017 | 0.923 ± 0.018 |
| 5 | 100 | 3 | 0.929 ± 0.020 | 0.918 ± 0.020 |
| 5 | 100 | 5 | 0.942 ± 0.015 | 0.931 ± 0.016 |
| 5 | 100 | 7 | 0.937 ± 0.016 | 0.926 ± 0.017 |
| 5 | 200 | 3 | 0.934 ± 0.022 | 0.921 ± 0.022 |
| 5 | 200 | 5 | 0.944 ± 0.015 | 0.932 ± 0.018 |
| 5 | 200 | 7 | 0.939 ± 0.016 | 0.926 ± 0.019 |

**Table S6** Clinical, genomic, and demographic features in the OAK and POPLAR studies

| **Feature name** | **Description** |
| --- | --- |
| BAGE | Baseline age |
| blSLD | Sum of longest diameter of target lesions at baseline |
| HIST | Histology |
| METSITES | Number of metastatic sites at enrollment |
| TOBHX | Smoking history |
| ECOGGR | ECOG group |
| MSAF | Maximum somatic allele frequency |
| bTMB | Blood TMB |
| SEX | Patient sex |
| Top 20 prevalent genes | *TP53*, *LRP1B*, *DNMT3A*, *SPTA1*, *NF1*, *FAT3*, *KEAP1*, *STAG2*, *MLL2*, *MLL3*, *ATM*, *TSC1*, *FAT1*, *EGFR*, *STK11*, *NFE2L2*, *PREX2*, *EPHA3*, *CHEK2*, *PDGFRA* |

ECOG, Eastern Cooperative Oncology Group; TMB, tumor mutation burden.