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

Gustavo Arango-Argoty,[1]* Damian E. Bikiel,[1] Gerald J. Sun,[1] Elly Kipkogei,[1] Kaitlin M. Smith,[1] Etai Jacob[1]*

[1]Oncology Data Science, Oncology R&D, AstraZeneca, Waltham, MA, USA

*Corresponding authors: gustavo.arango@astrazeneca.com, etai.jacob@astrazeneca.com

ABSTRACT

Modern clinical trials can capture tens of thousands of clinicogenomic measurements per individual. Employing manual approaches to discover predictive biomarkers, as differentiated from prognostic markers, is a challenging task. To address this challenge, we present an automated neural network framework based on contrastive learning, which we have named the predictive biomarker modeling framework (PBMF). This general-purpose framework explores potential predictive biomarkers in a systematic and unbiased manner, as demonstrated in simulated "ground truth" synthetic scenarios resembling clinical trials. Applied retrospectively to real clinicogenomic data sets, particularly in the complex field of immunooncology (IO) predictive biomarker discovery, our algorithm successfully found biomarkers that identify IO-treated individuals who survive longer than those treated with chemotherapy. In a retrospective analysis, we demonstrated how our framework could have contributed to a phase 3 clinical trial (NCT02008227) by uncovering a predictive biomarker based solely on early study data. Patients identified with this predictive biomarker had a 15% improvement in survival risk, as compared

to those of the original trial. This improvement was achieved with a simple, interpretable decision tree generated via PBMF knowledge distillation. Our framework offers a rapid and robust approach to inform biomarker strategy, providing actionable outcomes for clinical decision-making.

INTRODUCTION

The promise of precision medicine is to treat patients with therapies that best target their unique disease.[1,2] To do so, we need to find a characteristic that identifies individuals who are more likely than similar individuals without that characteristic to experience a favorable effect from treatment; i.e., a predictive biomarker. The intricate interplay of genetics and environmental factors, coupled with the complexity of disease biology and treatments, makes the discovery of predictive biomarkers a daunting task. The scarcity of comprehensive data, which is often due to acquisition or technical difficulties, presents challenges to the accurate representation of diverse populations, disease subtypes, and treatment cohorts, further compounding this discovery challenge. Moreover, the presence of numerous prognostic factors often hinders the ability to pinpoint the predictive biomarker within the studied patient population. Finally, even if a putative biomarker is found, translational applicability must be assessed with independent validation cohorts, adding further complexity and cost.

Nevertheless, there are clinically validated predictive biomarkers for certain targeted therapies, exemplified by the identification of *BCR-ABL* and *EGFR* mutations guiding the use of receptor tyrosine kinase inhibitors in cancer treatment.[3] Despite these significant achievements, a considerable gap remains in the availability of predictive biomarkers, particularly for therapies such as immunotherapy (IO) that do not directly modulate the disease. Although PD-L1

expression,[4] microsatellite instability,[5] and tumor mutation burden (TMB)[6] serve as validated predictive biomarkers for IO, only a subset of responsive patients exhibit positivity for these markers.[7] With an expanding array of novel targeted therapies, immunotherapies, and their combinations under investigation in clinical trials, the development of methodologies for identifying predictive biomarkers becomes imperative to advance personalized medicine and optimize the efficacy of emerging treatments.

To address the challenge of predictive biomarker discovery, traditional regression methods such as Cox proportional hazards (PH) modeling[8] have been widely employed. However, these methods necessitate the explicit enumeration of covariates and interactions, a task that becomes impractical as the number of features increases, particularly in scenarios involving a diverse set of clinical and -omic features. More recently, algorithms have been developed that aim to discover predictive biomarkers without requiring such explicit specifications. These approaches incorporate an objective function designed to maximize the difference in target outcomes between subgroups with different treatments.[9,10] Unfortunately, even these advanced approaches encounter challenges in identifying a predictive signal in the presence of noisy data or features that uniformly influence all arms (i.e., are prognostic) and often result in overfitting.

We therefore developed a novel approach, the predictive biomarker modeling framework (PBMF), designed for end-to-end predictive biomarker discovery and evaluation (Fig. 1). This framework, now available to the research community, centers around a neural network ensemble model featuring a contrastive loss function that ensures the learning of a multivariate biomarker that is specific to a target treatment of interest but not to a control treatment. The biomarker score cutoff and sample prevalence constraints are also components of the loss function, abrogating the need for post-hoc tuning. Additionally, we provide tools for generating simulated data to

benchmark the model, along with features to distill the model into an interpretable, deployable biomarker.

Here, we provide empirical evidence showcasing the robust predictive biomarker discovery capability of the PBMF across various scenarios, including simulated biomarker discovery and randomized controlled clinical trials. Notably, the PBMF outperformed existing approaches in subgroup identification within both simulated and real data sets. Furthermore, we illustrate how the PBMF retrospectively contributed to patient selection in a phase 3 clinical trial by uncovering a predictive biomarker based solely on phase 2 trial data. This discovery led to a 15% improvement in efficacy in the original trial, achieved through a straightforward decision tree generated via PBMF knowledge distillation.

RESULTS

Predictive biomarkers, contrastive learning, and model architecture

We define a predictive biomarker, B, as a classification tool categorizing a population into positive (B+) or negative (B−) for the biomarker, specific to a given treatment. B can encompass various patient measurements (e.g., age, blood counts, RNA gene expression). The biomarker is predictive if the B+ subpopulation is selectively enriched for individuals benefitting from a treatment of interest ("treatment"), but not a comparator one ("control"; Fig. 1a). Similarly, the B− subpopulation should be selectively enriched for those not benefiting from any treatment, or perhaps benefiting instead from a comparator (Fig. 1a). In contrast, a prognostic biomarker is characterized by similar benefit irrespective of treatment (Fig. 1a, bottom).

With this definition, we formulated the PBMF as a contrastive learning task that aims to distinguish between two patient populations based on their differential response to treatments. The PBMF's loss function actively maximizes the differences in outcomes for a given treatment (similar to pushing apart dissimilar items in contrastive learning) for B+ versus B− patients. Simultaneously, it minimizes the differences in outcomes for the control arm (similar to bringing similar items closer in contrastive learning), regardless of biomarker status. By doing so, the network is trained to contrast the effects of two treatments across the biomarker-defined groups, effectively learning the distinctive features that separate patient responses. Specifically, the loss function is defined as the log difference between control and treatment log-rank test statistics (Fig. 1b; Methods). This has the effect of maximizing the separation in time-to-event data (e.g., survival) between B+ and B− in the subpopulation receiving the treatment (i.e., large log-rank test statistic) while minimizing the separation for the subpopulation receiving the control. The model therefore optimizes for predictive biomarker behavior (Fig. 1a, 1b). For applications requiring a particular biomarker prevalence, we include an additional penalization term to encourage convergence toward a predefined B+ prevalence proportion.

In the PBMF application programing interface (API), any neural network architecture is applicable, including deep, convolutional, and attention-based networks. Alongside time-to-event data, censoring, and a grouping flag denoting the treatment, the PBMF uses input features from any modality (e.g., genomics, clinical, imaging), without restriction on the number or type (e.g., categorical or continuous) of input features (Fig. 1b). The PBMF outputs a probability ("confidence") score from 0 to 1, defining the likelihood that a sample is assigned to the B+ or B– subpopulation.

Model implementation and extensions

Overfitting poses a significant challenge in biomarker discovery, due to heterogeneity in patient populations and large numbers of features, particularly when attempting to predict the efficacy of one treatment over another rather than that of a single treatment. To bolster the robustness of the model, the PBMF incorporates an ensemble of *n* independently trained neural networks (Fig. 1c, left). To align with bagging principles,[14] we provide a tunable hyperparameter that enable each model in the ensemble to use a distinct random subset of samples and features for training (Table S1). Following model training, optional ensemble pruning can be applied to further refine performance (Fig. 1c, right). In scenarios with numerous noisy or random features, certain ensemble models may predict noise, compromising overall performance. Pruning these models, which are expected to have uncorrelated predictions, enhances ensemble efficacy by retaining only those with correlated predictions.

Finally, an opaque neural network in the PBMF-generated biomarker may compromise confidence and hinder applicability in clinical settings. To address this, the PBMF incorporates an optional pipeline for distilling the model into a simple interpretable decision tree classifier. This involves deriving a high-quality subset of training data through pseudo-labeling and filtering samples based on ensemble confidence scores (Fig. 1e). By training a decision tree with this subset, one can transform the candidate predictive biomarker into a set of rules, facilitating seamless integration into the design of future clinical studies (Fig. 1d, 1e).

PBMF identification of predictive biomarkers in diverse simulated biomarker discovery scenarios

To facilitate benchmarking, we generated synthetic data sets representing realistic combinations of features and time-to-event data (i.e., survival), mirroring conditions commonly encountered in real-world scenarios (Fig. 2a). Benchmarking was performed across 100 replicates, with performance reported on held-out test data sets from each replicate. We compared performance only across PBMF and VT methods, as SIDES failed to solve the contrived scenarios.

The objective of the first benchmarking scenario was to discover a predictive signal in the presence of a prognostic signal. This scenario comprised 3 features, 2 predictive and 1 prognostic; importantly, the predictive signal was present only as a combination of the two predictive features (Fig. 2a). The PBMF yielded an area under the precision-recall curve (AUPRC) of 0.918 ± 0.047 (mean ± standard deviation) and outperformed a competing method, VT (AUPRC = 0.858 ± 0.029) (Fig. 2b, Table S2).

Real-world scenarios often involve the presence of noninformative features, complicating the extraction of the underlying predictive signal. In our second benchmarking scenario, we retained the original 3 features (2 predictive, 1 prognostic) and introduced additional varying numbers of features containing random noise (n = 7, 17, 37). Remarkably, the PBMF consistently outperformed VT with 7 (PBMF AUPRC = 0.834 ± 0.050; VT AUPRC = 0.746 ± 0.039) or 17 (PBMF AUPRC = 0.768 ± 0.044; VT AUPRC = 0.690 ± 0.040) random features (Fig. 2c). With 37 random features, both approaches exhibited similar performance (PBMF AUPRC = 0.650 ± 0.033; VT AUPRC = 0.644 ± 0.036).

We hypothesized that in noisy scenarios, the ensemble PBMF might incorporate suboptimal constituent models. Our third benchmark explored the impact of model pruning on enhancing ensemble performance. When employing only the top quartile (p75) or top decile (p90) models

within the ensemble, we observed a marked improvement in PBMF performance, particularly in the presence of some (n = 7) or many (n = 37) random features (Fig. 2d). This pruning strategy outperformed VT, but it necessitated a larger ensemble (1024 versus 128) to achieve stable performance (Fig. 2d).

Our final benchmarking scenario investigated how the performance of the PBMF scales with the size of the training data set. In the simple case of 3 total features (2 predictive and 1 prognostic; i.e., benchmark 1), both the PBMF and VT methods exhibited diminished performance when training data were reduced from 1000 to 250 samples (Fig. 2e, Table S2). Despite this reduction, the PBMF still outperformed the VT (PBMF AUPRC = 0.786 ± 0.066; VT AUPRC = 0.752 ± 0.091). In the more complex scenario of 2 predictive, 1 prognostic, and 7 random features (i.e., benchmark 2), the performance of the PBMF matched or exceeded that of VT at all training data sizes tested (n = 250, 500, 1000, 2000, 4000; Fig. 2e). Although VT performance reached a plateau at 1000–2000 samples, the PBMF demonstrated continuous improvement and superior performance; notably, at the largest training data size tested (n = 4000), the PBMF (AUPRC = 0.967 ± 0.008) significantly outperformed the VT method (AUPRC = 0.788 ± 0.027). Lastly, the introduction of model pruning further enhanced PBMF performance at training data sizes greater than 500.

Identification of predictive biomarker of hormone therapy in breast cancer

We benchmarked the PBMF against VT and SIDES for identifying a biomarker predictive of hormone therapy + tamoxifen versus chemotherapy in breast cancer across two independent data sets. Models were trained on the Rotterdam breast cancer cohort[15] and subsequently tested on the German breast cancer study cohort.[16]

On the training data set, the PBMF (B+: hazard ratio [HR] = 0.71, confidence interval [CI] = 0.54–0.94, *P* = 1.69e-2; B−: HR = 1.91 CI = 1.48–2.48, *P* = 9.37e-7) and VT (B+: HR = 0.56, CI = 0.44–0.70, *P* = 4.9e-7; B−: HR = 1.81, CI = 1.30–2.52, *P* = 4.32e-4) methods successfully identified a predictive biomarker, whereas SIDES found a prognostic biomarker (Fig. 3a, 3b, Fig. S1a). On the test data set, only the PBMF generalized as a predictive biomarker (B+: HR = 0.63, CI = 0.48–0.831, *P* = 1.13e-3; B−: HR = 1.22, CI = 0.72–2.04, *P* = 4.6e-1), whereas both VT and SIDES were prognostic. Further still, the PBMF also identified the greatest prevalence of B+ individuals who could benefit from hormone therapy + tamoxifen within the test data set (PBMF, 85%; VT, 56%; SIDES, 8.1%).

Identification of individuals with improved survival outcomes to inform phase 3 trial design with early-stage clinical trial data

One critical application of predictive biomarker discovery is to inform the patient selection strategy for phase 3 clinical trials by using data from earlier phases. Building on the promising results from real-world evidence, we evaluated the PBMF against VT and SIDES in the context of representative clinical trial decision-making. Models were trained on clinicogenomic phase 2 trial data (POPLAR,[17] NCT01903993), and tested on phase 3 trial data (OAK,[18] NCT02008227). This evaluation aimed to determine which model could effectively guide patient selection for second-line atezolizumab therapy versus chemotherapy in NSCLC (i.e., the OAK trial), relying solely on data from earlier studies.

Both PBMF (B+: HR = 0.33, CI = 0.21–0.52, *P* = 1.82e-6; B−: HR = 2.23, CI = 1.33–3.74, *P* = 2.33e-3 and VT (B+: HR = 0.38, CI = 0.24–0.60, *P* = 3.7e-5; B−: HR = 1.14, CI = 0.72–1.78, *P* = 5.7e-1) identified a predictive signal from the phase 2 POPLAR training data. SIDES identified

a mixed predictive and prognostic signal (B+: HR = 0.42, CI = 0.14–1.21, *P* = 0.1; B–: HR = 0.75, CI = 0.54–1.05, *P* = 0.09) (Fig. S1b). Importantly, when the three models trained on POPLAR study data were applied as a hypothetical patient selection biomarker for the phase 3 OAK trial test data, only the PBMF generalized as a predictive biomarker (Fig. 3c; B+: HR = 0.61, CI = 0.48–0.76, *P* = 1.3e-5; B–: HR = 0.82, CI = 0.60–1.13, *P* = 2.24e-1). Both VT (B+: HR = 0.70, CI = 0.53–0.92, *P* = 9.9e-3; B–: HR = 0.62, CI = 0.48–0.80, *P* = 2.2e-4) and SIDES (B+: HR = 0.64, CI = 0.37–1.11, *P* = 0.1; B–: HR = 0.66, CI = 0.54–0.8, *P* = 3e-5) yielded only prognostic biomarkers (Figs. 3d, S1c). The PBMF identified the highest prevalence of B+ individuals that could benefit from atezolizumab therapy (PBMF, 80%; VT, 46%; SIDES, 14%). Compared with the biomarker-evaluable population in the OAK trial, the PBMF B+ subpopulation yielded a ~7% decrease in risk of death for atezolizumab versus docetaxel treatment (PBMF, HR = 0.61; OAK trial-reported HR = 0.65). Thus, to hypothetically inform strategies for patient selection in phase 3 clinical trials, only the PBMF successfully identified a predictive, high-prevalence biomarker from phase 2 data that generalized to phase 3 results.

A discovery pipeline for predictive biomarker prototypes

Given the consistent ability of the PBMF to identify a predictive biomarker, particularly in clinical trial settings, we devised an end-to-end biomarker discovery pipeline that generates a human-understandable predictive biomarker prototype, poised for translation into clinical settings (Fig. 4a). Using a process similar to that described in the preceding section, we trained a PBMF ensemble model solely on phase 2 clinical trial data to identify a predictive biomarker. However, departing from our earlier approach of using all models in the ensemble, we employed ensemble pruning to select the highest 5 percentile (p95) most highly performing and consistent

models within the ensemble (Fig. S2a–c, Methods). Utilizing a consensus score across these models, we determined an optimal biomarker probability score cutoff to classify B+ and B– samples, subsequently referred to as pseudo-labels (Fig. 4d, Methods). These pseudo-labels were then used for the distillation of the complex neural network original PBMF model into a simple interpretable model—a decision tree— that could inform a strategy for a clinical study (Figs. 4d, S2a–c; Methods).

Use of knowledge distillation from the PBMF neural network to produce a simple decision tree with improved predictive value

Similar to the original PBMF from which it was derived, the distilled decision tree PBMF biomarker was predictive on both the phase 2 trial training data (B+: HR = 0.46, CI = 0.3–0.7, *P* = 2.6e-4; B–: HR = 1.34, CI = 0.8–2.2, *P* = 0.2) and phase 3 trial test (B+: HR = 0.55, CI = 0.43–0.7, *P* = 8.05e-7; B–: HR = 0.86, CI = 0.64–1.16, P = 0.3) data sets (Fig. 4f). Importantly, the HR of the distilled decision tree was improved by approximately 10% compared with the original PBMF (original PBMF HR = 0.61; distilled decision tree PBMF HR = 0.55; see Fig. 4c, 4f), owing to the reduction in prevalence from 80% to 64%. Notably, the original PBMF had a ~7% decrease in risk of death within the B+ atezolizumab versus docetaxel-treated subpopulation relative to the biomarker-evaluable population in the OAK trial, and the distilled decision tree PBMF had a ~15% decrease in risk of death (distilled PBMF HR = 0.55; original PBMF HR = 0.61; OAK trial-reported HR = 0.65).

Upon scrutinizing the decision tree of the distilled PBMF, we observed that the predictive biomarker comprises a specific subset of clinical and genomic features: the maximum circulating tumor DNA ctDNA allele frequency (MSAF), sum of longest diameter of target lesions at

baseline (blSLD), and mutation status on the *MLL2*, *TSC1*, *ATM*, *PDGFRA* and *LRP1B* genes (Fig. 4d). Collectively, all these features drive the predictive nature of the biomarker. With the exception of *ATM* mutations, which were both predictive and prognostic (POPLAR: mutation [Mut] B+ HR = 0.33, wild type [Wt] B– HR = 0.776; OAK: Mut B+ HR = 0.43, Wt B– HR = 0.68) but with a notably low prevalence (28 patients for *ATM* B+/Mut and 205 for the distilled PBMF B+), each individual feature fell short in matching the biomarker prevalence or the consistent, predictive signal of the collective (Fig. S3, Table S3). Furthermore, in comparison with a commonly described single-feature ICI biomarker, blood TMB,[19-21] the PBMF more robustly enriched for longer survival for both the training and test clinical trial data sets (Fig. 4e, 4f; Table S4).

DISCUSSION

Across diverse, challenging benchmarks spanning simulated scenarios through informing strategies for patient selection in clinical trials, the PBMF out-performed other methods for discovering predictive biomarker signals. Among comparator methods, only the PBMF found signals that were consistently predictive across training and test data sets. Along with the PBMF's ability to accurately identify known IO biomarkers from phase 2/3 trials, we also showed that the PBMF can nominate a novel composite biomarker from a set of clinicogenomic features that out-performed blood TMB.

We emphasize here the importance of the predictive constraint embedded in the PBMF. A common pitfall in biomarker discovery is to focus only on identifying populations with enhanced responses to a specific treatment.[23] In these cases, one cannot distinguish between a biomarker

that is prognostic versus one that enriches for better responses specifically in a treatment of interest. Thus, the PBMF loss function enforces the constraint that a biomarker must be considered in the context of a control treatment.

Beyond its contrastive loss function, the PBMF stands out as a unique end-to-end API for predictive biomarker discovery. The results presented here underscore the superior performance of an ensemble PBMF consisting of fully connected neural networks. At the same time, our API is versatile and compatible with any differentiable model. This flexibility makes it possible to explore predictive biomarker signals using input features from single or multiple modalities, or diverse data representations, including various combinations thereof. For instance, an attention-based transformer model could effectively model unstructured data such as clinical notes. This opens the door to leveraging pretrained models, e.g. large-language models, to imbue the PBMF with prior knowledge, potentially enabling successful predictive biomarker discovery even in situations with limited or noisy data.[24] Lastly, the PBMF provides tools to refine a biomarker toward a particular downstream application, i.e., prevalence constraints, simulations, and knowledge distillation, for clinical deployment.

In our patient selection strategy example, we successfully distilled a complex ensemble neural network model into a simple decision tree. In this regard, we can view the PBMF as a highly effective search function, as we required the complex model to discern whether a predictive signal exists and what features may drive it. Alternatively, one could model patient risk through a multivariate Cox PH model with interaction terms for treatment. Although this approach may theoretically achieve similar results, it may be impractical to implement. Whereas the gradient descent within the PBMF will implicitly traverse the vast expanse of potential feature combinations and interactions, one would have to systematically and explicitly test every single

potential case when using a Cox PH model. Further, the PBMF accounts for treatment effects simultaneously within its loss function, whereas a Cox PH model requires enumeration of each hypothesized treatment-feature interaction.

We concede that there are limitations of the PBMF, although most are common to any biomarker nomination process. First, with the known challenge of limited data sets and high heterogeneity in patient populations, the PBMF cannot be used to determine whether the data are adequate and representative of the target population and biology. Nevertheless, it is noteworthy that the PBMF demonstrated superior performance in scenarios with small data sizes. In situations with substantial data, PBMF scaled with data size, whereas the performance of the VT method reached a plateau. Second, the ensemble PBMF may be unable to maintain its magnitude of predictive power when distilled into a simple model, as there is often a tradeoff between a biomarker's predictive power and its parsimony.[26] However, the enhanced interpretability of the model may contribute to a better understanding of the biological factors underpinning the predictive signal of the biomarker. Third, while the PBMF outperformed other methods in discerning predictive signals from noisy or prognostic features, we might still find that strongly prognostic features can impede the identification of predictive signals, and therefore our method could potentially gain more from prior feature selection. Fourth, the PBMF's contrastive loss function formulation tends to attenuate the discovery of biomarkers that show a modest positive effect in the control treatment but a more substantial benefit in the treatment of interest. Finally, the PBMF is a discovery tool, and any biomarker hypothesis requires prospective clinical validation.[27-29]

Specific considerations and limitations apply when using any predictive biomarker method to inform late-stage clinical trial decision-making. As alluded to earlier, data availability is often

limiting. It would be challenging to train the PBMF with data from a phase 2 trial lacking a control arm; future work will be required to know whether non-randomized evidence, synthetic control arms, could be used (e.g., real-world data). Any such exploration would need to carefully consider the substantial heterogeneity within patient populations. A related point is that it is often difficult to ensure that cohorts are comparable across studies, as the intent-to-treat clinical trial design guarantees only within-trial comparisons. Moreover, considering the rising trend of combination therapies, it will be crucial to investigate the PBMF's performance across various arms and their pairwise combinations. Finally, future work can explore the tradeoff between data maturity, ability to extract a predictive signal, and phase 3 trial investment decision timing. Our benchmarks nonetheless demonstrate that with the availability of the appropriate data, the PBMF could nominate a predictive biomarker that is likely to outperform the original study design in selecting patients who would derive greater benefits from the new treatment in a phase 3 study. The use of the PBMF has the potential to improve strategies for patient selection over what can be achieved with conventional study designs.

METHODS

Predictive biomarker loss function

The PBMF (Fig. 1) uses as input time-to-event data with censoring, a treatment label, and a feature matrix (n patients by f features). The feature matrix $X \in \mathbb{R}^f$ is used as the input to a fully connected neural network of user-defined depth and width.

The goal of the neural network is to assign patients to either the B+ or B− group. To refine this categorization, we employed a contrastive learning approach in which patients in the B+ group, when under treatment, show an improvement in survival times compared with those in the B−

group. Conversely, in the control arm, the model aims to minimize the differences in survival times between the two biomarker groups according to the principle of contrastive learning.[30-32]

The distinction or similarity in survival times is quantified using log-rank test statistics[33] within each treatment arm as follows:

$$TLogRank(a) = \frac{\left(E_a^+ - O_a^+\right)^2}{E_a^+} + \frac{(E_a^- - O_a^-)^2}{E_a^-},$$

where the E_a^+, E_a^- pair represents the expected number of events for the treatment a, under B+ and B−, respectively. The O_a^+, O_a^- pair depicts the observed events within the treatment a for B+ and B−, respectively.

Formally, the expected and observed events are defined as follows:

$$E_a^b = \sum_i^N B_i^b * I(A_i = a) * \lambda_i$$

$$O_a^b = \sum_i^N B_i^b * I(A_i = a) * I(C_i = 1)$$

$$\lambda_i = \sum_t \frac{\Omega_t}{N_t} I(T_i > t)$$

where the treatment arm is defined by $a \in \{Treatment\ (Tr), Control\ (CR)\}$ and the indicator function $I(A_i = a)$ determines whether the patient i is under treatment a or not. The biomarker group is defined by the output of the neural network where $b \in$ {positive (+), negative (−)}. Therefore, each patient i has a probability of being labeled as being in the positive (B_i^+) or negative (B_i^-) group. C_i represents the censoring status of patient I, and λ_i is a scalar independent

on the parameters of the neural network and can be precalculated (see Meier et al.[34]). Ω_t is the number of observed events at time t, and N_t is the number of subjects at risk at time t.

The log-rank test for the treatment and control is then defined as:

$$LR(Tr) = \frac{(\sum_i^N B_i^+ * I(A_i = Tr)[\lambda_i - I(C_i = 1)])^2}{\sum_i^N B_i^+ * I(A_i = Tr) * \lambda_i} + \frac{(\sum_i^N B_i^- * I(A_i = Tr)[\lambda_i - I(C_i = 1)])^2}{\sum_i^N B_i^- * I(A_i = Tr) * \lambda_i}$$

$$LR(Cr) = \frac{(\sum_i^N B_i^+ * I(A_i = Cr)[\lambda_i - I(C_i = 1)])^2}{\sum_i^N B_i^+ * I(A_i = Cr) * \lambda_i} + \frac{(\sum_i^N B_i^- * I(A_i = Cr)[\lambda_i - I(C_i = 1)])^2}{\sum_i^N B_i^- * I(A_i = Cr) * \lambda_i}$$

The contrastive nature of the loss function is evident in its formulation as follows:

- Treatment arm optimization: For patients receiving the actual treatment, the model maximizes the survival time difference between B+ and B− groups. This is quantified by the treatment log rank test score, $LR(Tr)$.
- Control arm optimization: For the control group, the model minimizes the survival time difference between the two biomarker groups. This is quantified by the control log rank test score, $LR(Cr)$.

The contrastive loss for the predictive biomarker is then defined as the ratio between the control log rank test score by the treatment log-rank test score:

$$loss_b = \frac{LR(Cr)}{LR(Tr)}.$$

The custom contrastive loss is the ratio of two log-rank tests computed over the time-to-event data, grouped by the treatment label, and stratified by the neural network output score. During optimization, the neural network learns a set of parameters that outputs scores to maximize the

separation (i.e., larger log-rank test statistic) for the treatment while minimizing the separation (i.e., smaller log-rank test statistic) for the control. This ensures that the neural network will learn to generate a predictive biomarker score, since it will only stratify patients for a specific treatment.

We also integrated a population prevalence term to the loss to enable the model to identify a predictive biomarker given a specific desired minimal population (*minP*) such that:

$$prev(B^{+}) = \frac{\sum_{i}^{N} B_{i}^{+}}{\sum_{i}^{N}(B_{i}^{+} + B_{i}^{-})}$$

$$loss_{p} = \left(\frac{prev(B^{+})}{minP} - 1\right)^{2}$$

The *loss_p* will have a minimum value of 0 when *minP* is equal to the population of B^{+}. Finally, the composite PBMF loss function takes the following form:

$$Loss \ = \ \omega_{1} * loss_{b} + \ \omega_{2} * loss_{p},$$

where ω_1 and ω_2 dictate the contribution of each loss component. For example, when $\omega_2 = 0$, the PBMF finds a population with the best predictive power independent of the number of patients, and when $\omega_2 = 0.5$ the PBMF identifies a predictive biomarker of the treatment at a 50% patient prevalence.

Biomarker scoring

The output of the neural network ($B \in \mathbb{R}^2$) is composed of two units representing the B+ and B− scores $\{b^+, b^-\}$. Scores are then passed through a SoftMax activation to convert the network scores into probabilities. Thus, the biomarker scores for a given patient i can be expressed as:

$$B_i^+ = \frac{e^{b_i^+}}{e^{b_i^+} + e^{b_i^-}},\ B_i^- = \frac{e^{b_i^-}}{e^{b_i^+} + e^{b_i^-}}.$$

The probability of the negative biomarker can be written as $B^- = (1 - B^+)$. In this way, B^+ values close to 0 indicate B− and values close to 1 indicate B+. We assume the B+ to be contained within the neuron at index 0 from the output of the neural network. However, because the loss function does not have control of the directionality of the assignments, it is possible that the B+ is placed in the neuron at the index 1. Therefore, after training and when making predictions, we corrected the B+ by computing the HR between the B+ and B− within the treatment arm as $HR^{Treatment} = \frac{\sum O^+ / \sum O^-}{\sum E^+ / \sum E^-}$. Thus, an $HR^{Treatment} < 1$ defines the B+ in the neuron 0, whereas an $HR^{Treatment} > 1$ defines the biomarker positive the neuron 1.

With ensemble of neural networks, for a given patient i and a total of M neural network models, we generated a set of scores $\{B_{i,1}^+, \dots, B_{i,M}^+\}$ and computed a consensus score defined by the average score over all the models in the patient i such that $B_i^+ = \frac{1}{M}\sum_{m=1}^{M} B_{i,m}^+$.

Feature and patient subsetting during model training

A random subset of patients and features can be specified (Table S1). Patient subsetting is performed before model loss computation, and a different subset of patients will be excluded at each gradient update. Feature subsetting is performed before model training, and the given model

will only train on the feature subset; when training an ensemble, each model will utilize its own unique random subset. During ensemble model evaluation, no patients or features are excluded.

PBMF ensemble model pruning

Under the assumption that some models in the ensemble perform poorly and damage the entire ensemble's performance, we implemented the following model pruning approach. We first binarized the set of scores, $\{B_{i,1}^{+}, \ldots, B_{i,M}^{+}\}$, generated from the trained ensemble, using the default 0.5 score threshold for the PBMF. Using this N patients by M models binary matrix, R, we then compute an $N \times N$ patient agreement matrix, A, by calculating the proportion of models that assigned two different patients to the same class[35]:

$$A_{ij} = \frac{1}{M} \sum_{k=1}^{M} I(R_{ik} = R_{jk})$$

A contains 1 along its diagonal, is symmetric, and contains values $\in [0,1]$. Patients with similar scores across each model in the ensemble will tend to have higher values; those with dissimilar scores will have lower values. Each column or row of A represents how consistently patients were assigned to a particular class by the models in the ensemble, from the reference point of one patient.

We then computed the Pearson correlation between each column in A with each column in R to generate an $N \times M$ matrix, C, of correlation coefficients that represents how well the patient scores from an individual model in the ensemble correlate with the patient agreement matrix. We assumed that only a minority of models have poor performance, such that we should keep models that agree on how patients should be scored and discard models that disagree. This was

done by selecting a percentile, e.g., the 90th percentile of all the correlations. By thresholding on the value in C associated with this percentile, the models were sorted by the number of times that each model exceeded the threshold, to generate a $1 \times M$ vector of counts. We then thresholded on the value associated with our percentile in this vector to return the final subset of models, M_S, that exceed this threshold. A new consensus score was then computed as the average score across the reduced set of models in the ensemble.

Model distillation: pseudo-labeling

The distribution of scores generated from the ensemble is used to identify patients with "high-quality" predictions, i.e., those whose distributions are heavily skewed toward 0 (strongly B−) or 1 (strongly B+).

To identify the patients with the best high-quality scores, we choose a 0.5 cut point and add an offset value ε, such that the biomarker label for a patient ii is defined as:

$$L_i = \begin{cases} B^+ \; if \; Cs > 0.5 + \varepsilon \\ B^- \; if \; cs < 0.5 + \varepsilon \\ No \; biomarker \; other \; case \end{cases}$$

We set $\boldsymbol{\varepsilon} \in \{0, 0.1, 0.2, 0.3, 0.4\}$ and then fitted a Cox PH model to compute the hazard ratios between the treatment and the control arms for both the B+ and B−. The optimal ε score is extracted by determining the maximum difference between the absolute log of the B+ and B− hazard ratios.

$$optimal \; \varepsilon = \mathrm{Max}_{\varepsilon_i \in \boldsymbol{\varepsilon}} \{ | log\left(HR^+_{\varepsilon_i}\right) - log\left(HR^-_{\varepsilon_i}\right) | \}$$

We then applied the optimal ε to compute a reduced set of patients with high-quality scores.

Model distillation: tree-based model explainability

Once the high-quality population is defined, a tree classifier (python sklearn[36] tree classifier package, max_depth = 3, random_seed = 0) is fit, using the input features and the B+ and B− as the labels. The goal of the tree classifier is to define a simple rule that approximates the neural network−derived predictive biomarker. The tree model was then applied to the validation data sets. (An example of this framework is shown in Fig. 4f).

VT implementation

We implemented the VT approach proposed by Foster et al.[11] as follows. We used a random survival forest model[37] to predict time to event based on the log-rank test loss (pySurvival[38]). We built two survival models $\{M_T, M_C\}$, where T and C refer to the population under treatment and under the control, respectively. Each model was trained using only its respective population. We then computed the difference in risk score between the treatment and control models to define the contrafactual risk score $r_i = M_T(i) - M_C(i)$ for any given patient i.

To stratify patients into B+ and B−, we computed the median value of the contrafactual risk score distribution across all patients and assigned to B+ those patients below the median score (low risk) and to B−those with a contrafactual risk score above the median. Consequently, this design choice intrinsically classified patients evenly, 50% being assigned to B+ and the remaining 50% to B–. This can potentially lead to an overestimation of favorable results in data sets where the predictive biomarker prevalence is 50%.

Model hyperparameters were tuned as described in Supplemental Information and Table S5.

SIDES implementation

The SIDES algorithm was set for survival analysis using the time and event variables as the targets and the treatment versus control setting. The variables used were the same as those used for PBMF and VT and depended on the analyzed data set. We used the R implementation of SIDES provided by the SIDES authors (sides.dylib, CSIDES.r, and stochSIDES_util.R). The following parameters were used: min_subgroup_size = 10, criterion_type = 1, depth = 3, width = 5, gamma = c(NA,NA,NA) local_mult_adj = 1, n_perms_mult_adjust = 200, subgroup_search_algorithm = "SIDES procedure", n_top_biomarkers = 2. We then selected the best biomarker sorted by the adjusted *P* value and assigned it as B+. The discovered predictive biomarker rule was then validated in the independent test set.

Synthetic data generation

We generated 10,000 patients for each data set. For a given replicate, 2000 patients (20%) were randomly selected, without replacement. Among those selected, a 50-50 training/test split was performed. Evaluation metrics are reported only from the test set. Proportional hazard assumptions were imposed to induce each one of the behaviors (Fig. 2a). The ability of each methodology to correctly call the biomarker was measured by recording the AUPRC, precision, recall, and F1 score of a holdout test data set (2000 patients for each data set).

The generation of synthetic data sets involves three stages. Initially, a set of covariates with predetermined level of correlation and prevalence is defined (Fig. 2a). These covariates establish subgroups for which desired hazard ratios will be generated. For the parametric model, the cumulative hazard is

$$H_i(t) = \lambda(t^{\gamma}) \exp(X_i^T \beta)$$

Where X_i is a vector of covariates associated to the parameters β. The β parameters used to sample survival times can be estimated after setting the HR requirements between groups. For example, assuming a treatment variable and a predictive biomarker, we can define the following hazard ratios:

$$HR^{Control,B+\ vs\ B-} = HR_1$$

$$HR^{Treatment,B+\ vs\ B-} = HR_2$$

$$HR^{B+,Treatment\ vs\ Control} = HR_3$$

$$HR^{B-,Treatment\ vs\ Control} = HR_4.$$

The time-independent part of $H_i(t)$ can be expanded as:

$$H_i \sim \exp\left(\beta_{trt} trt_i + \beta_{x1} x1_i + \beta_{trt-x1} trt_i x1_i\right).$$

Replacing for each one of the cases in equation 1, we obtain the following equations:

$$\log(HR_1) = \beta_{x1}$$

$$\log(HR_2) = \beta_{x1} + \beta_{trt-x1}$$

$$\log(HR_3) = \beta_{trt} + \beta_{trt-x1}$$

$$\log(HR_4) = \beta_{trt}.$$

Random survival times are then obtained using the technique outlined in Crowther and Lambert (2013),[39]

$$t_i = \left(\frac{-\log(u)}{\lambda exp(X_i^T \beta)}\right)^{\frac{1}{\gamma}}$$

where λ and γ and are the scale and shape parameters, and u is a random variable sampled from the uniform distribution $U(0, 1)$. Note that additional censoring, not covered in this work, can also be introduced.

Clinical data sets

The Rotterdam breast cancer cohort[40] (863 patients) was used as a training data set, and the German breast cancer study cohort[16] (686 patients) was used as a test data set. We selected only patients treated with hormone-based treatments and chemotherapy. The 7 features used for training the PBMF are age, menopause, tumor size, tumor grade, number of nodes, pr (progesterone receptor status), and er (estrogen receptor status). We trained the model using overall survival and death. Minimum population (minp) was empirically selected from the training data set by screening different minimum populations [0, 0.25, 0.5, 0.75]. The model with the best separation on the training set was selected (0.75).

The POPLAR and OAK clinical trials were used to represent phases 2 and 3, respectively, to evaluate the efficacy of atezolizumab as a second-line therapy for patients unresponsive to first-line platinum-based chemotherapy in the NSCLC population. The therapeutic potential of atezolizumab was compared against that of docetaxel. The data set, sourced from Gandara et al.,[19] encompasses ctDNA from blood samples in addition to patient demographics and clinical biomarkers, as detailed in Table S6. We conducted a prevalence-based ranking of ctDNA genes

from patients in the POPLAR trial, identifying the top 20 genes that exhibit a minimum prevalence of 20% across the combined data set from both atezolizumab and docetaxel cohorts. The PBMF was not trained by using progression-free survival, and this outcome was used for testing only. POPLAR trial data were used for training the PBMF, and OAK was used for independent evaluation. We used the overall survival time and event as endpoints. The PBMF ensemble model performance is depicted in Fig. 4c.

Competing interests

G.A.-A., D.B., G.J.S., E.K., K.M.S., and E.J. are employees of AstraZeneca with stock ownership, interests, and/or options in the company.

Additional information

Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to Gustavo Arango-Argoty and Etai Jacob.

REFERENCES

1. Ciardiello, F.*, et al.* Delivering precision medicine in oncology today and in future-the promise and challenges of personalised cancer medicine: a position paper by the European Society for Medical Oncology (ESMO). *Ann Oncol* **25**, 1673-1678 (2014).

2. Schwartzberg, L., Kim, E.S., Liu, D. & Schrag, D. Precision oncology: who, how, what, when, and when not? *Am Soc Clin Oncol Educ Book* **37**, 160-169 (2017).

3. Goossens, N., Nakagawa, S., Sun, X. & Hoshida, Y. Cancer biomarker discovery and validation. *Translational Cancer Research* **4**, 256-269 (2015).

4. Herbst, R.S.*, et al.* Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet* **387**, 1540-1550 (2016).

5. Chung, H.C.*, et al.* Efficacy and safety of pembrolizumab in previously treated advanced cervical cancer: results from the phase II KEYNOTE-158 study. *J Clin Oncol* **37**, 1470-1478 (2019).

6. Marabelle, A.*, et al.* Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. *The Lancet Oncology* **21**, 1353-1365 (2020).

7. Luchini, C.*, et al.* ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour

mutational burden: a systematic review-based approach. *Annals of Oncology* **30**, 1232-1243 (2019).

8. Cox, D.R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**, 187-220 (1972).

9. Loh, W.-Y., Cao, L. & Zhou, P. Subgroup identification for precision medicine: A comparative review of 13 methods. *WIREs Data Mining and Knowledge Discovery* **9**, e1326 (2019).

10. Alemayehu, D., Chen, Y. & Markatou, M. A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. *Statistical Methods in Medical Research* **27**, 3658-3678 (2018).

11. Foster, J.C., Taylor, J.M. & Ruberg, S.J. Subgroup identification from randomized clinical trial data. *Statistics in medicine* **30**, 2867-2880 (2011).

12. Lipkovich, I. & Dmitrienko, A. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *Journal of biopharmaceutical statistics* **24**, 130-153 (2014).

13. Lipkovich, I., Dmitrienko, A., Patra, K., Ratitch, B. & Pulkstenis, E. Subgroup identification in clinical trials by stochastic SIDEScreen methods. *Statistics in Biopharmaceutical Research* **9**, 368-378 (2017).

14. Breiman, L. Bagging predictors. *Machine Learning* **24**, 123-140 (1996).

15. Royston, P. & Altman, D.G. External validation of a Cox prognostic model: principles and methods. *BMC medical research methodology* **13**, 1-15 (2013).

16. Sauerbrei, W. & Royston, P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162**, 71-94 (1999).

17. Fehrenbacher, L.*, et al.* Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *The Lancet* **387**, 1837-1846 (2016).

18. Rittmeyer, A.*, et al.* Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *The Lancet* **389**, 255-265 (2017).

19. Gandara, D.R.*, et al.* Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nature medicine* **24**, 1441-1448 (2018).

20. Wang, Z.*, et al.* Assessment of blood tumor mutational burden as a potential biomarker for immunotherapy in patients with non–small cell lung cancer with use of a next-generation sequencing cancer gene panel. *JAMA oncology* **5**, 696-702 (2019).

21. Kim, E.S.*, et al.* Blood-based tumor mutational burden as a biomarker for atezolizumab in non-small cell lung cancer: the phase 2 B-F1RST trial. *Nature medicine* **28**, 939-945 (2022).

22. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

23. Italiano, A. Prognostic or predictive? It's time to get back to definitions! *J Clin Oncol* **29**, 4718; author reply 4718-4719 (2011).

24. Arango-Argoty, G.*, et al.* Pretrained transformers applied to clinical studies improve predictions of treatment efficacy and associated biomarkers. *medRxiv*, 2023.2009.2012.23295357 (2023).

25. Frankle, J. & Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635* (2018).

26. Harrell, F.E.J. *Biostatistics for Biomedical Research*, (Vanderbilt University, Nashville, TN, 2023).

27. Sun, X., Briel, M., Walter, S.D. & Guyatt, G.H. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* **340**, c117 (2010).

28. Dmitrienko, A., Muysers, C., Fritsch, A. & Lipkovich, I. General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *J Biopharm Stat* **26**, 71-98 (2016).

29. Ondra, T.*, et al.* Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *J Biopharm Stat* **26**, 99-119 (2016).

30. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. in *International conference on machine learning* 1597-1607 (PMLR, 2020).

31. van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

32. Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A. & Jegelka, S. Debiased contrastive learning. *Advances in neural information processing systems* **33**, 8765-8775 (2020).

33. Woolson, R.F. Rank tests and a one-sample logrank test for comparing observed survival data to a standard population. *Biometrics*, 687-696 (1981).

34. Meier, A.*, et al.* Hypothesis-free deep survival learning applied to the tumour microenvironment in gastric cancer. *The Journal of Pathology: Clinical Research* **6**, 273-282 (2020).

35. Monti, S., Tamayo, P., Mesirov, J.P. & Golub, T.R. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* **52**, 91-118 (2003).

36. Pedregosa, F.*, et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).

37. Ishwaran, H., Kogalur, U.B., Blackstone, E.H. & Lauer, M.S. Random survival forests. in *Wiley StatsRef: Statistics Reference Online* (Wiley, 2008).

38. Fotso, S. PySurvival: open source package for survival analysis modeling. https://www.pysurvival.io (2019).

39. Crowther, M.J. & Lambert, P.C. Simulating biologically plausible complex survival data. *Statistics in Medicine* **32**, 4118-4134 (2013).

40. Sauerbrei, W., Royston, P. & Look, M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biom J* **49**, 453-473 (2007).

41. Fernandes, L.E.*, et al.* Real-world evidence of diagnostic testing and treatment patterns in US patients with breast cancer with implications for treatment biomarkers from RNA sequencing data. *Clinical Breast Cancer* **21**, e340-e361 (2021).

42. Subramanian, A.*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

43. Liberzon, A.*, et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).

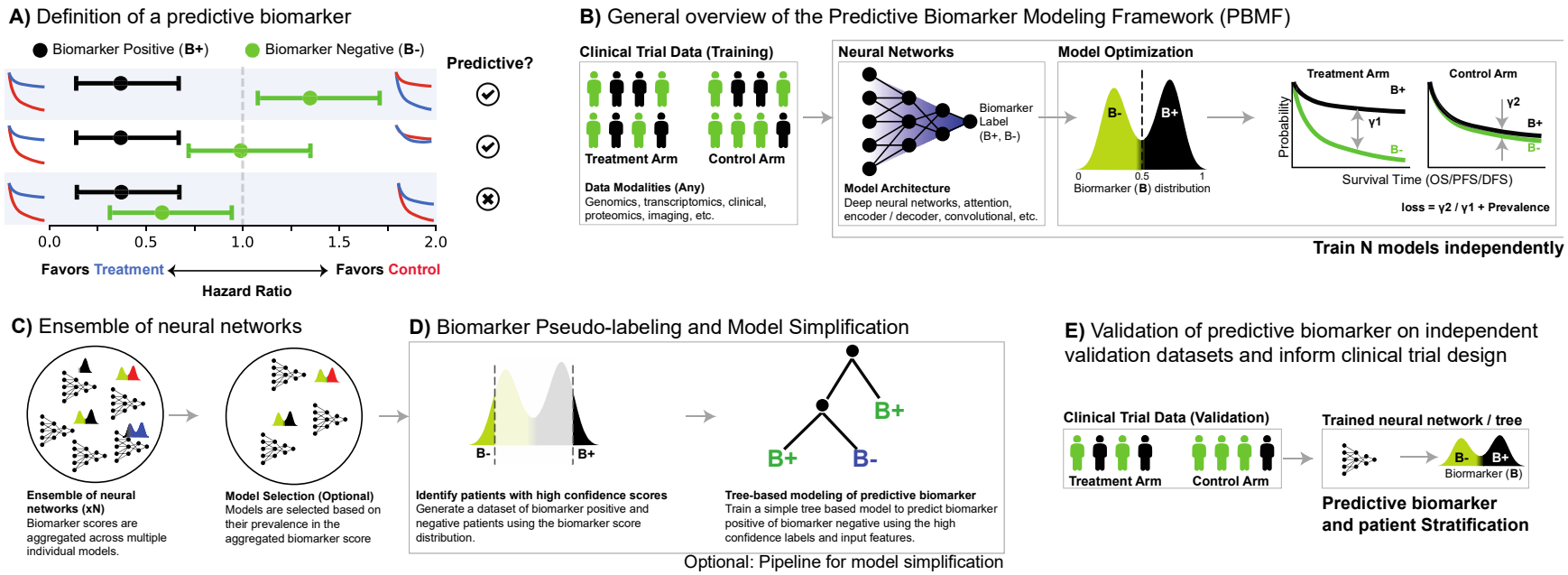


Fig. 1 | Detailed schematic of the PBMF. a, Discrimination between predictive and prognostic biomarkers, with the subdivision into B+ and B– cohorts. B+ is indicative of patients who benefit from treatment as opposed to the control, and B– signifies a lack of superiority of treatment or an advantage in the control group. **b**, The PBMF trained a set (*N*) of neural networks, each independently trained on clinical trial data with a contrastive loss function. The loss is designed to enhance the differential impact of B+ versus B– in the treatment group and concurrently minimize B+ influence over B– in the control arm. **c**, The ensemble of PBMF models synthesizes into a consolidated predictive score, refining the model collection by filtering out non-contributory models to retain only those with significant impact. **d**, High-confidence patient samples are identified through biomarker pseudo-labeling, which then serve to construct an interpretable, simplified decision tree model, categorizing patients as B+ or B–. **e**, External dataset validation of the PBMF model affirms the biomarker's predictive capacity, demonstrating the model's reliability from ensemble to simplified tree representation, thus reinforcing its utility in clinical trial stratification.

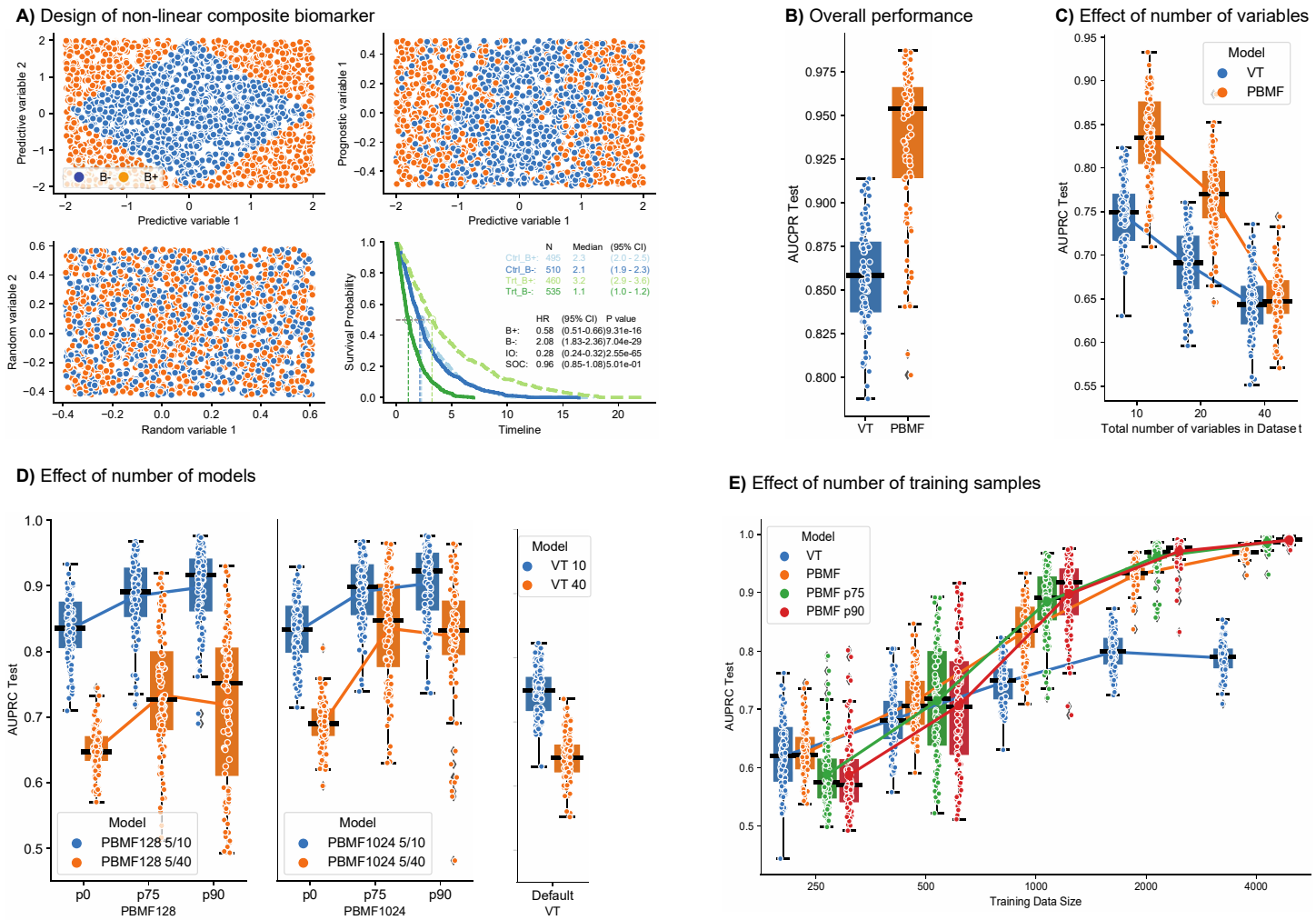


Fig. 2 | Simulated and benchmark tests. a, Synthetic data set generation and behavior. A predictive biomarker is generated by var1 and var2 (top left), which creates a particular Kaplan-Meier plot, showing the differential effect in the treatment and control arms (top right). A prognostic variable is added as var3, which has a different effect when added to var1 (bottom left). Random variables with no

structure can be added (var4 and var5). **b**, AUPRC for the test set comparing the PBMF model developed for a data set containing 3 variables (2 predictive,1 prognostic) in orange against virtual twins in blue. The training was performed on 1000 data points, with 100 training-test split replicates **c**, Effect of the number of random variables in the AUCPR for PBMF and VT. The PBMF model contains 128 ensembles of 5 variables chosen from data sets with 10, 20, and 40 total variables, in which only 2 are predictive and 1 is prognostic. Models are trained with 1000 data points, with 100 training-test split replicas. **d**, Effect of the number of models in the ensemble for PBMF (128 vs 1024) against VT at two different levels of noise (10 and 40 total variables). Models are trained with 1000 data points, with 100 training-test split replicas. **e**, Effect of the training size on AUCPR for VT (blue), PBMF (orange), and two different levels of post-pruning (top quartile [p75, green] and top decile [p90, red] percentile of models). The data set contains 10 total variables (2 predictive, 1 prognostic, and 7 random). PBMF ensemble models comprised 128 models containing only 5 variables from the 10.

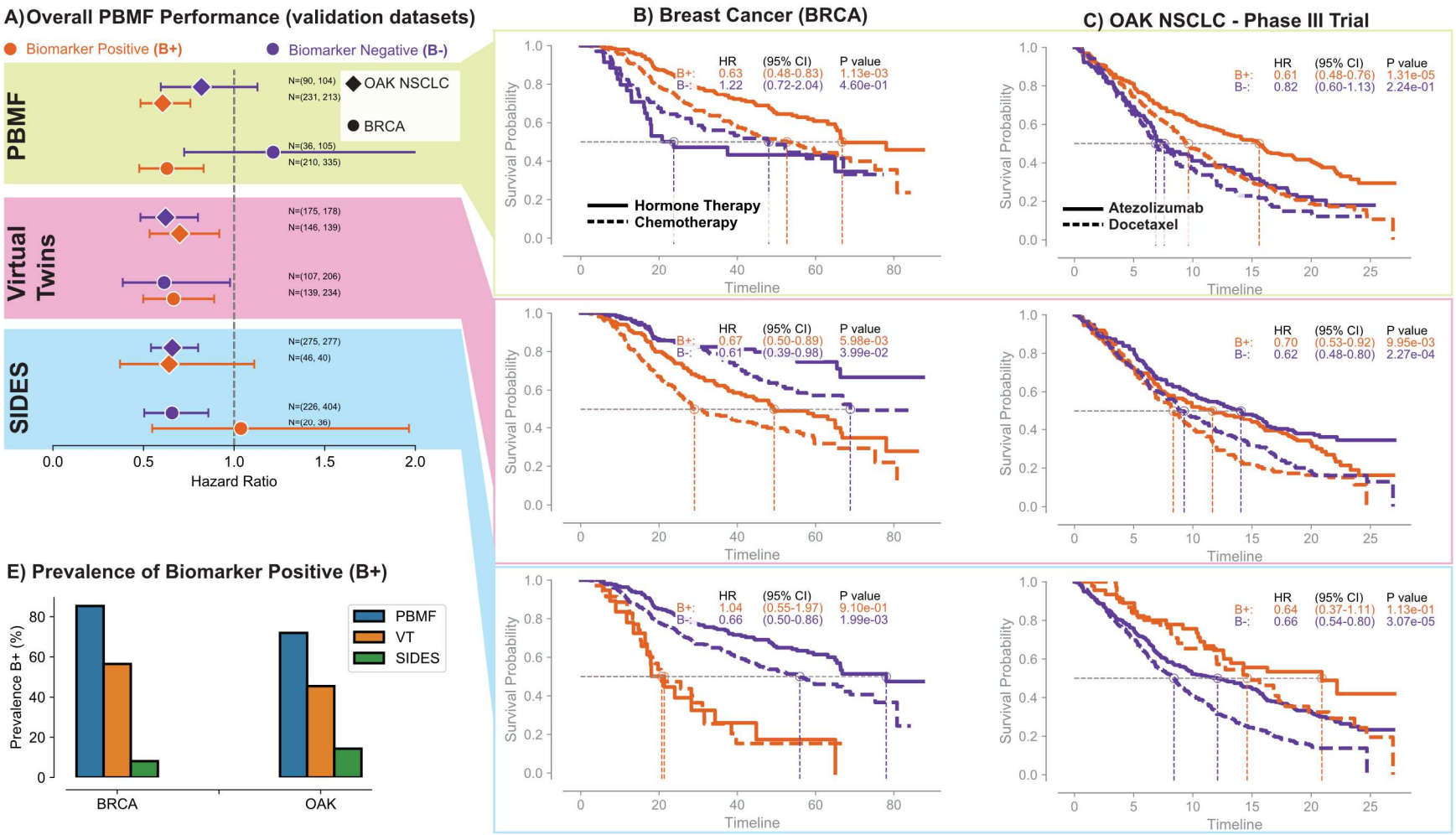


Fig. 3 | Evaluation of PBMF for predictive biomarker identification on real data sets against other methods. a, Forest plot illustrating the performance comparison of PBMF with VT and SIDES methodologies, applied to validation data sets encompassing breast cancer hormone versus chemotherapy and the OAK phase 3 clinical trial. **b**, Kaplan-Meier survival estimates over the

independent breast cancer data set, showcasing the comparative predictive accuracy of PBMF, VT, and SIDES. **c.**, Cross-trial validation of predictive biomarkers identified in the POPLAR phase 2 clinical trial by evaluating their performance in the subsequent OAK phase 3 clinical trial, using PBMF, VT, and SIDES approaches.

A) Biomarker discovery pipeline

1 Clinical Trial Data Collection
Discovery: POPLAR Phase II trial
Validation: OAK Phase III trial

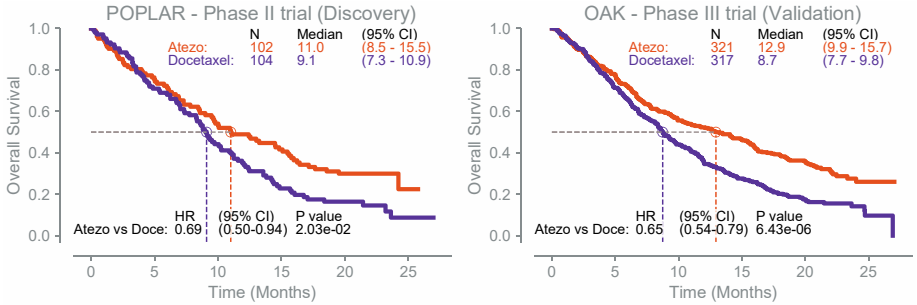
Clinical + ctDNA biomarkers

2 Biomarker Analysis
Composite biomarker using the **PBMF**. Probability assignment of biomarker positive and negative

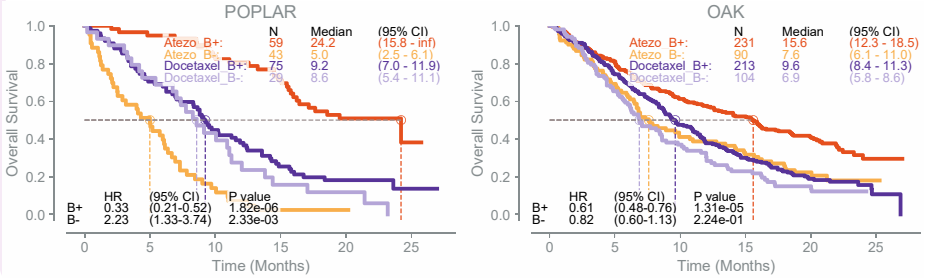
3 Biomarker Refinement
Selection of high-confidence biomarker predictions.

4 Translational Modeling
Interpretable model (decision tree) for predictive biomarker model

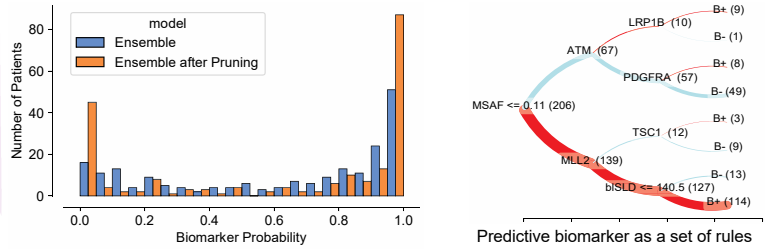
B) Survival Analysis by Biomarker Status: Phase II vs. Phase III



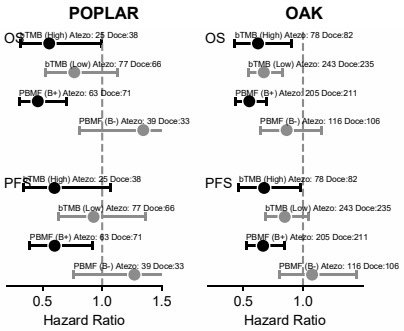
C) PBMF Biomarker Performance: Discovery and Validation Stages



D) Predictive biomarker pseudo-labeling on phase II patient population and biomarker approximation using tree-based model distilation



E) Comparison of decision tree predictive biomarker against bTMB



F) Patient Stratification Based on Decision Tree-Derived Biomarker

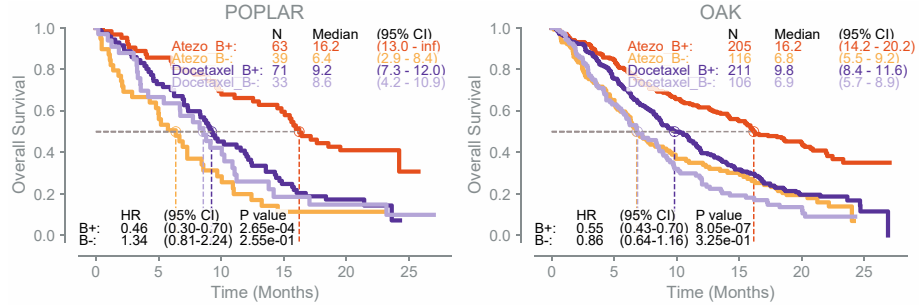


Fig. 4 | Application of PBMF in the design of biomarker-driven clinical trials. a, Overview of the proposed integrative framework for the discovery of predictive biomarkers in phase II trials to enhance phase III trial design, incorporating initial data acquisition from early-phase trials, PBMF analysis, biomarker optimization through interpretable models, and subsequent application in clinical trial planning. **b**, Clinical trial data and endpoints collection: Kaplan-Meier curves for the discovery (POPLAR phase II clinical trial) and the validation (OAK Phase III) data sets. **c**, Identification of predictive biomarker: Using the discovery data set (POPLAR trial) the PBMF successfully identified a biomarker with a positive predictive value for atezolizumab over docetaxel, demonstrating consistency in the OAK trial for validation. **d**, Refinement of predictive biomarker: The enhancement of the predictive biomarker involves pruning to eliminate spurious models from the ensemble (left) and the subsequent derivation of a rule set that encapsulates the biomarker's predictive power (right). **e**, Patient stratification using the simplified predictive biomarker identified in the POPLAR trial and subsequently applied to the OAK trial. **f**, Comparison of the predictive biomarker against blood TMB in the discovery (POPLAR) and validation (OAK) data sets, with an additional evaluation of the biomarker on progression-free survival (PFS), despite the PBMF's initial training on overall survival (OS).