Numenos

A Deep Learning Framework for Causal Inference in Clinical Trial Design: The CURE AI Large Clinicogenomic Foundation Model

Authors:

Amit D Weiss[1], Vitalay Fomin[1], Di Feng[2], Zuojian Tang[2], James Cai[2], Bino John[2] and Neil T Pfister[1,3#]

[1]Numenos, New York, NY, United States.

[2]Global Computational Biology and Digital Sciences Department, Cancer Immunology and Immune Modulation, Boehringer Ingelheim Pharmaceuticals, Inc. Ridgefield, CT, United States.

[3]AI in Precision Medicine Research Group, Department of Radiation Oncology, University of Alabama at Birmingham School of Medicine, Birmingham, AL, United States.

[#]Corresponding author, to whom correspondence should be addressed: neil.pfister@numenos.ai

Abstract

Clinical research is limited by the capability to define the most important combinations of clinical features and biomarkers that predict therapeutic benefit. Here, we introduce CURE AI (Clinical trials Uncovering Real Efficacy Artificial Intelligence), a novel deep learning framework designed to predict individual patient benefit from a new therapeutic intervention compared to a standard of care. CURE AI utilizes a large clinicogenomic foundation model to understand the complex relationships between the vast clinical and multiomic features in clinical trial data. To build CURE AI, we trained a proprietary foundation model based on a deep learning architecture and training schema using a large collection of clinical and multiomics datasets. Using CURE AI, we seek to understand the complex interplay between clinicogenomic information from clinical trial arms to predict the magnitude of therapeutic benefit on the individual patient level. In this paper, we finetuned the CURE AI foundation model on lung cancer data from the OAK non-small cell lung cancer clinical trial. We observed that the trial could have been significant for progression-free survival (PFS) with fewer than half of the patients enrolled using CURE AI to guide trial enrollment ($p = 0.60$ to $p < 0.05$). The finetuned CURE AI (termed CURE Lung Cancer) demonstrated direct generalizability on a held-out independent clinical trial dataset, the POPLAR trial, by converting an insignificant PFS endpoint to significance while also including the majority of patients (88%; $p = 0.21$ to $p < 0.05$). In summary, we developed a causally-aware clinicogenomic deep learning platform that can learn to predict individualized patient benefit of investigational therapy compared to an existing standard of care. Because we use a foundation model trained on readily measurable patient characteristics, CURE AI can be applied to a variety of scientific and clinical uses including adaptive clinical trials, toxicity prediction, treatment response prediction, and understanding of drug resistance and response mechanisms.

Introduction:

Clinical research is hampered by limitations in identifying the most important combinations of clinical and biologic features that can predict therapeutic response. New biomarker discovery is crucial to help define patient populations eligible for therapies. When large scale molecular profiling of patients (e.g., transcriptomic or proteomic) are carried out, as in OAK, POPLAR and other recent oncology clinical trials, retrospective data analysis following completion of clinical trials can result in development of gene signatures that can be predictive of patient response.[1-5] However, traditional methods to identify actionable biomarkers are time consuming, expensive, mostly prognostic, and often fail to provide actionable insights.

To identify patterns within high dimensional multiomics data, artificial intelligence algorithms have been developed to assist in understanding basic patterns within the data.[6-10] At the more advanced level, complex neural networks could, in theory, identify non-linear complex relationships of clinical and biologic information that can predict treatment benefit, which cannot be achieved by classical methods but could be made feasible by the assistance of advanced AI models.[11] For this reason, we sought to develop a Large ClinicoGenomic foundation Model (LCGM) that can leverage the complexity of high dimensional data (clinical and multiomic) to predict treatment benefit. If the model can learn complex biological relationships in the foundation model, then the LCGM should be able to identify patterns that predict treatment benefit in new datasets.

Our technology utilizes a deep learning architecture and training schema to predict the benefit of a clinical trial intervention, which we define as the benefit of a therapeutic intervention (experimental arm) compared to the standard of care (control arm). The name of our LCGM is Clinical trials Uncovering Real Efficacy Artificial Intelligence, or CURE AI™. CURE AI predicts which patients benefit more from an investigational therapy than standard of care at the individual patient level based on their baseline clinicogenomic profile. Identifying patient populations more likely to benefit from treatment, a process known as predictive enrichment, can contribute to the development of robust and reliable patient enrichment strategies.[12-13] CURE AI enables predictive enrichment though enhanced classification of true responders and true non-responders to therapy. Accurately defining eligibility criteria is a key part of ensuring that patients are offered therapy that they will benefit from, which is an important consideration for new drug approval.[14-16]

CURE AI tackles two major bottlenecks in clinical drug development. The first problem is data scarcity (limited patient numbers and data missingness) combined with the high dimensionality of clinical data. Clinical trial data often presents challenges due to limited patient numbers that cannot be easily pooled into larger databases as well as a vast amount of data points per patient (>1,000,000 measurements in phase 3 trials with omics data).[17-18] Using a novel and proprietary deep learning architecture and training schema, we developed an approach to not only identify the most important clinicogenomic variables but to learn the connections and patterns between these variables on a much larger scale using our LCGM trained on a large corpus of patient data. The second challenge is sorting out the differences in prognostic and predictive markers. Current methods in clinical trials often find correlations between variables but are unable to assess all available clinical and multiomic data to pinpoint patterns within that data that explain why drugs benefit certain patient groups but not others.[19] Inability to accurately identify predictive factors can result in misinterpretation of trial outcomes, with immortal time bias being a well-known example.[20] For instance, a patient's functional status (prognostic factor) might be mistakenly interpreted to drive their health outcome to a new drug, leaving the true predictive factors undiscovered. The difficulty lies in predicting the benefit that an individual

would receive from both the investigational therapy and the standard of care, which allows for a prediction of the magnitude of excess benefit derived from the investigational therapy. We need to be able to predict response to both trial therapies to design and interpret the most personalized and effective clinical trials.

To achieve this, a perfect clinical trial dataset would contain every patient's response to both trial therapies (randomizing each patient to both trial arms), which would allow measurement of the real benefit of the investigational therapy over the existing standard of care at the individual patient level. Since it is not possible to administer different therapies to a patient in parallel, we developed CURE AI to accurately predict this difference for each patient based on accurate comparisons of their clinical and genomic characteristics to similar patients in what can be thought of as an advanced implementation of causal inference.[21-22] In this seminal manuscript, we validate the premise of CURE AI to guide patient therapy selection in two completed advanced non-small cell lung cancer randomized clinical trials using one as validation set and the other as held-out test set. We show how CURE AI can guide clinical trial interpretation and design by AI-guided prediction of therapeutic benefit.

Methods:

CURE AI Development:

CURE AI is a deep learning technology developed by Numenos which consists of two main parts: (1) a base neural architecture foundation model that was trained on a large cohort of clinicogenomic datasets in a proprietary self-supervised learning schema with multiple loss function[23] and (2) a proprietary training process to finetune CURE AI on indication-specific clinical data for the specific downstream task of predicting treatment benefit using k-fold cross-validation (CV)[24] and out-of-fold (OOF) predictions. CURE AI contains a foundation model trained on a large cohort of patients with diverse pathologies including all major cancer types.

The CURE AI LCGM was finetuned on progression-free survival (PFS)-outcome stratified random selection of four-fifths of the patients on the OAK trial for whom RNA sequencing was available. The model was then used to predict benefit on the held-out one-fifth of the validation patients. This process was iterated five total times (to define the features that predict treatment benefit on the entire included OAK trial dataset) to create the finetuned CURE Lung Cancer™ model. A schematic of the CURE AI training model is shown in Figure 1.

CURE Score™ and CURE Index™ Evaluation of Individual Patient Benefit

CURE AI calculates a counterfactual outcome for each patient to predict how each patient would have responded to the trial therapy that they did not receive.[25] This individualized patient benefit prediction is called the CURE Score. The estimate of benefit of investigational therapy over standard of care in a specific trial is based on finetuning CURE AI with Conditional Average Treatment Effect (CATE) modeling.[26] The CURE Score is calculated by assessing the difference of the predicted treatment outcomes. A positive score favors investigational therapy, and a negative score favors standard of care.

The ordered ranking of each patient within a clinical trial by their CURE Score is called the CURE Index. The CURE Index is used to define subgroups of patients. We create a Cox-proportional hazard model, with the treatment arm being the only covariate for each subgroup. Because of randomization, benefit prediction is equally distributed between the trial arms, allowing a balanced dataset to train the Cox-proportional hazard model and measure treatment effect using the hazard ratio for the treatment covariate for that subgroup. The log-rank test is used to generate the p-value for therapeutic benefit. Thus, treatment effect is based on the hazard ratio and significance is based on the p-value.

The CURE Curve

We developed the CURE Curve as the aggregated p-value plot of all patients included in the order of the CURE Index. The CURE Curve presents CURE AI results and allows for straightforward identification of the patients who benefit more from either the investigational therapy or the standard of care. We enroll each patient based on their order in the CURE Index, starting from the highest CURE Score. For each additional patient enrolled (who received either atezolizumab or docetaxel), a p-value is calculated to assess whether we can reject the null hypothesis that there are no differences in outcome between the treatment groups. If we observe significance in CURE Index-defined subsets ($p < 0.05$), then CURE AI successfully identified patient characteristics that predict better response to atezolizumab.

In order to compare the CURE Curve to a random baseline, we simulated trial enrollment of patients based on random enrollment (rather than CURE index-defined enrollment, we used random ranking). A p-value is calculated for the random enrollment, which is performed 1000 times. The p-value line for random patient enrollment is presented with its mean value as well as 1 standard deviation on each side of the mean. At every point where the CURE Curve p-value is above the upper confidence interval of the random selection line (ie: the p-value has a lower absolute value for CURE Indexed patients) CURE AI is significantly outperforming random trial enrollment.

Because the CURE Score is based on baseline information, CURE Scores are directly comparable for patients on similar clinical trials. We can therefore compare groups of patients across different trials (comparing OAK patients to POPLAR patients), based on each patient's CURE Score.

The CURE Curve generates "natural subgroups" of patients with similar predicted benefit values. We selected three representative points along the OAK and POPLAR trial CURE Curves that represent two possible inclusion thresholds generated by CURE AI (2 thresholds plus the entire dataset for each trial). For the OAK trial, we selected the top 35% and 52% of CURE Score-defined patients. The trial percentiles for POPLAR trial patients with the same CURE Scores corresponded to the top 44% and 88% of patients.

CURE Permutation Test™

We developed a permutation test as an additional assessment of confidence in CURE AI. The permutation test involves randomizing the outcome of each patient so that the treatment outcome is no longer correctly associated with the same patient. This is done by randomizing PFS independently within each trial arm, so that the distribution of outcomes per trial arm remains the same. Permutated data from OAK and POPLAR were independently analyzed by CURE AI as described above. If the permutated data leads to loss of significance (loss of relationship between clinical variables and outcome), then we are assured of more confidence in our model.

Kaplan-Meier Estimation of Progression Free Survival

We plot the Kaplan-Meier estimates for the probability of progression-free survival for three representative points along the CURE Curve for both the OAK and POPLAR trials. We utilized PFS as the outcome because the OAK and POPLAR trials both failed to meet this endpoint. Additionally, PFS is most directly related to cancer treatment effect and therefore less prone for bias, whereas overall survival can be affected by factors unrelated to cancer treatment or progression. Kaplan-Meier plots were generated using Python. The number of patients at risk (alive and evaluable; non-censored) at each timepoint are shown for the non-permutated plots.

Clinical Trial Datasets

The phase 3 OAK and phase 2 POPLAR advanced non-small cell lung cancer randomized prospective clinical trials both tested the investigational therapy atezolizumab compared to docetaxel in previously treated advanced non-small cell lung cancer.[3,4] We restricted our analysis to patients who had available RNA-sequencing data, which led to the inclusion of 699 patients on the OAK trial and 193 patients on the POPLAR trial. Tabular clinical trial data were integrated into CURE AI as training and testing variables.

Pre-processing of Bulk RNA Sequencing data:
Raw read counts and transcripts per million (TPM) were utilized for sample normalization. These data integrate into CURE AI as training and testing variables.

Results:

CURE AI predicts a benefit score for each patient, which allows enrolled patients to be ranked from high to low predicted benefit to investigational therapy compared to standard of care (Figure 1). This order of patients is called the CURE Index. We then enroll each patient by the CURE Index order and calculate a p-value for each additional enrolled patient, which generates the CURE Curve. We developed the CURE Curve as a visual representation of the performance of CURE AI in predicting clinical trial treatment benefit.

The CURE Curves and associated Kaplan-Meier estimates for the OAK trial[3] are demonstrated in Figure 2. The OAK trial represents a validation of the performance of the CURE AI foundation model because the benefit prediction was finetuned on OAK using out-of-fold benefit prediction derived only from baseline OAK patient profiles (see Figure 1). Practically, this means that CURE AI is a high-performing foundation model as patients who benefit from atezolizumab more than docetaxel could be identified through analysis of a single clinical trial.

Two OAK trial CURE Curve thresholds were selected for Kaplan-Meier estimation. These two thresholds represent the first point that the CURE Curve reaches significance (top 35% of OAK patients, corresponding to the top 44% of POPLAR patients) and the point on the POPLAR trial that is the most inclusive for patients who experience more benefit from atezolizumab relative to docetaxel (the top 88% of patients on POPLAR, which corresponds to the top 52% of the OAK CURE Index). These two thresholds represent the same CURE Score, meaning that they have the same predicted benefit to atezolizumab relative to docetaxel.

The point where the CURE Curve p-value first reaches significance is upon enrollment of the top 35% of the CURE Index patients (Figure 2, top panel). The Kaplan-Meier estimates for progression-free survival calculation for the 35% threshold is shown in Figure 2, top right panel ($p = 0.045$ indicating significant benefit to atezolizumab in this group). Figure 2 (middle panels) demonstrate the CURE Curve and associated Kaplan-Meier estimates for the CURE Curve threshold of 52% ($p = 0.047$). Both of the CURE Curve thresholds significantly outperform the entire training set (which is shown in Figure 2, bottom panels) which has a p-value of 0.595. At every point along the CURE Curve up to when the p-value line begins to descend, that patient cohort (all patients enrolled from the highest CURE Index and to that point) has more benefit to atezolizumab than docetaxel. The point where the CURE index p-value begins to consistently decrease (around 65% on the OAK CURE Curve) represents the threshold where patients benefit more from docetaxel than from atezolizumab.

To address concerns of overfitting, we compared CURE AI performance to a random recruitment of patients. The random selection line on the CURE Curve includes an average of 1000 iterations of simulated randomized trial enrollment of the same number of included patients to that point along the CURE Curve. A p-value is calculated for this patient group (the cohort receiving atezolizumab compared to the cohort receiving docetaxel) for each enrollment iteration, and the p-value is averaged. The standard deviation for random patient enrollment p-value is shown on the CURE Curve. At every point where the CURE Curve p-value line is above the upper confidence interval of the random selection line, CURE AI significantly outperforms random trial enrollment.

We hypothesized that if a trial is well balanced, CURE AI should not generate an imbalance in trial arms. Indeed, we found that the CURE Curve-defined groups remain well balanced in both patient number and clinical features. Table 1 demonstrates that some of the major variables associated with clinical response are not imbalanced due to CURE patient selection for the OAK trial including gender, histology, TMB[27], and PD-L1 status. These data demonstrate that CURE AI can effectively predict benefit to therapy on the OAK trial while keeping baseline characteristics balanced and despite significant data missingness (PD-L1 was only defined for about half of patients, for example).

We deliberately only included the OAK trial in the finetuning process to showcase two points: (1) how CURE AI performs on a real clinical trial when used for both training and cross-validation and (2) to assess the performance on a smaller and completely held-out trial (POPLAR), which we describe next. CURE AI, now finetuned on OAK non-small cell lung cancer data, is termed CURE Lung Cancer.

CURE Lung Cancer demonstrated direct generalizability on a held-out independent clinical trial dataset, the POPLAR trial.[4] The POPLAR trial has a similar patient population and treatment arms (atezolizumab vs. docetaxel) as the OAK trial. We utilized CURE Lung Cancer to calculate the CURE Curve for the POPLAR clinical trial. The CURE Curves and Kaplan-Meier estimates for the POPLAR trial are demonstrated in Figure 3. The top panel demonstrates the top 44% of the CURE Index patients (corresponding to the top 35% of patients on the OAK trial, which represents the same CURE Score). We find that the PFS is significant with a p-value of 0.021 (compared to 0.211 on the entire patient cohort; Figure 3, bottom panel).

Interestingly, the top 52% of patients on the OAK trial corresponded to the top 88% of patients on the POPLAR trial, based on the same CURE Score cutoffs. This means that the patient group on the POPLAR trial contained more patients who we predict would have more benefit to atezolizumab. We find that on POPLAR, if the trial included the top 88% of patients by CURE Index ranking, the trial would have had a positive PFS endpoint ($p = 0.035$).

Since the POPLAR trial was well balanced, CURE AI should not lead to an imbalance in trial arms. Indeed, we found that the CURE Curve-defined groups remain well balanced in both patient number and clinical features. Table 2 demonstrates that major variables associated with clinical response are not imbalanced due to CURE patient selection for the POPLAR trial. These data demonstrate that CURE Lung Cancer can effectively predict benefit to therapy for the POPLAR trial, which is a completely independent test of CURE AI training and finetuning process.

Since the design of CURE AI intrinsically limits the risk of confounding by validation testing on held-out trial data, and we saw no signs of data overfitting or bias within our models, we have high confidence in the performance of CURE AI. Indeed, our finetuned model generalized to the POPLAR trial. However, as an additional test to affirm our confidence in the model, we developed the CURE Permutation test to assess for intrinsic bias within the model finetuning schema. The permutation test involves randomizing the outcome of each patient so that treatment outcomes are no longer correctly associated with the same patient's clinicogenomic information. Within each trial arm, PFS is randomized so that the distribution of outcomes per trial arm remains the same, but the association of PFS for a particular patient is incorrect. This should result in loss of signal for the model training, as the clinicogenomic relationships that are

predictive for patient benefit are no longer accurate, so the model should not be able to learn an association since the correct association no longer exists.

Permutated patient outcome data from OAK and POPLAR were independently assessed by CURE AI, as described in the schema from Figure 1. Figure 4 demonstrates the CURE Permutation test on the OAK Trial. We find that there is a loss of significance across the entire dataset, as the CURE Curve no longer outperforms the randomized patient selection. At the top 35% of the CURE Index (Figure 4, top panel), the p-value is insignificant at $p = 0.391$, compared to $p = 0.045$ on the CURE AI model for OAK. At the top 52% of the CURE Index (Figure 4, bottom panel), the p-value is insignificant at $p = 0.703$, compared to $p = 0.047$ on the CURE AI model for OAK. Figure 5 demonstrates the CURE Permutation test on the POPLAR Trial. We find that there is a loss of significance across the entire dataset, as the CURE Curve no longer outperforms the randomized patient selection. At the top 44% of the CURE Index (Figure 5, top panel), the p-value is insignificant at $p = 0.629$, compared to $p = 0.021$ on the CURE AI model for POPLAR. At the top 88% of the CURE Index (Figure 5, bottom panel), the p-value is insignificant at $p = 0.371$, compared to $p = 0.035$ on the CURE AI model for POPLAR. The results demonstrate that CURE AI is not prone to generating false associations between data and outcomes.

Discussion:

The CURE AI foundation model is an innovative example of how complex neural network architectures can be applied to clinical trials to increase the likelihood that patients will receive a drug for which they will benefit. We independently validated the performance of CURE AI on two clinical trials and showed that CURE AI can predict patient benefit while being blinded to patient outcome. This is the first demonstration that an LCGM can proactively characterize patients by predicted benefit to an investigational therapy.

While further validation is required, based on the presented evidence it is attractive to speculate that LCGMs will cause a shift in the way clinical trials are approached through enhancement of enrollment criteria that enhances patient selection for treatment benefit. With improved enrollment criteria, we should observe a Will Rogers effect, where the clinical outcomes of all patients screened for a trial will improve. This is because patients who are eligible for enrollment will be enriched for benefit to investigational therapy, and patients who are screened ineligible will avoid receiving an investigational therapy that is predicted to be inferior to the existing therapeutic option. For example, on POPLAR, the top 88% of patients should be enrolled and randomized. However, the patients who are predicted to benefit more from standard of care could instead be offered standard of care alone or an alternative clinical trial. In practice, CURE AI could be used to understand which actionable clinical and genetic factors define benefit from a phase 2 trial, which would refine eligibility criteria for a phase 3 trial. Another exciting use case is to adapt ongoing clinical trials based on interim analyses, in which patient eligibility criteria could be modified during the trial to optimize for therapeutic benefit.

CURE AI has the potential to improve clinical trial design, analysis, and interpretation. In the future, it is possible that LCGMs will enable ‘virtual control arms’ for clinical trials, as standard control arms could raise concerns for not offering treatment with high predicted benefit over standard of care, based on existing data. Regulatory bodies will need to collaborate with academic, biotechnology, and patient advocacy leaders to ensure safe, transparent, and straightforward implementation of technologies like CURE AI, as the clinical translation of transformative patient treatments could be rapidly expedited with the assistance of these technologies.

In summary, we have developed a causal LCGM that can predict and accurately measure individualized patient benefit of investigational therapy compared to an existing standard of care. The development of CURE AI enables a shift from simple outcomes (such as treatment response) to individualized continuous benefit measurement. We are optimistic that special implementations of AI, such as CURE AI for clinical trials, will demonstrate consistent successes and lead to a revolution in healthcare including a major step toward more individualized medicine. CURE AI can be applied to a variety of scientific and clinical uses including adaptive clinical trials, toxicity prediction, treatment response prediction, and understanding of drug resistance and response mechanisms.

Funding

This study was supported by Numenos and Boehringer Ingelheim Pharmaceuticals, Inc.

Availability of data and materials

All data generated or analyzed during this study that are available are included in this published article. The code underlying all CURE AI models is a proprietary trade secret. CURE AI models are currently available to access by commercial entities and academic collaborations or for non-profit organizations for clinical trial design and analysis through Numenos. The models are not currently available for independent public research use.

Authors' contributions

AW and VF conceived and developed CURE AI. AW, NP and VF designed the trial analyses in this study. AW, VF, and NP analyzed and interpreted the trial analyses in this study. DF, ZT, JC, and BJ contributed to the supervision and project strategy. All authors contributed to writing and reviewing the manuscript. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Competing interests

All intellectual property relevant to this work has been developed and is owned by exclusively by Numenos. AW, VF, and NP have equity in Numenos. Authors DF, ZT, JC, and BJ are employees of Boehringer Ingelheim.

Figures

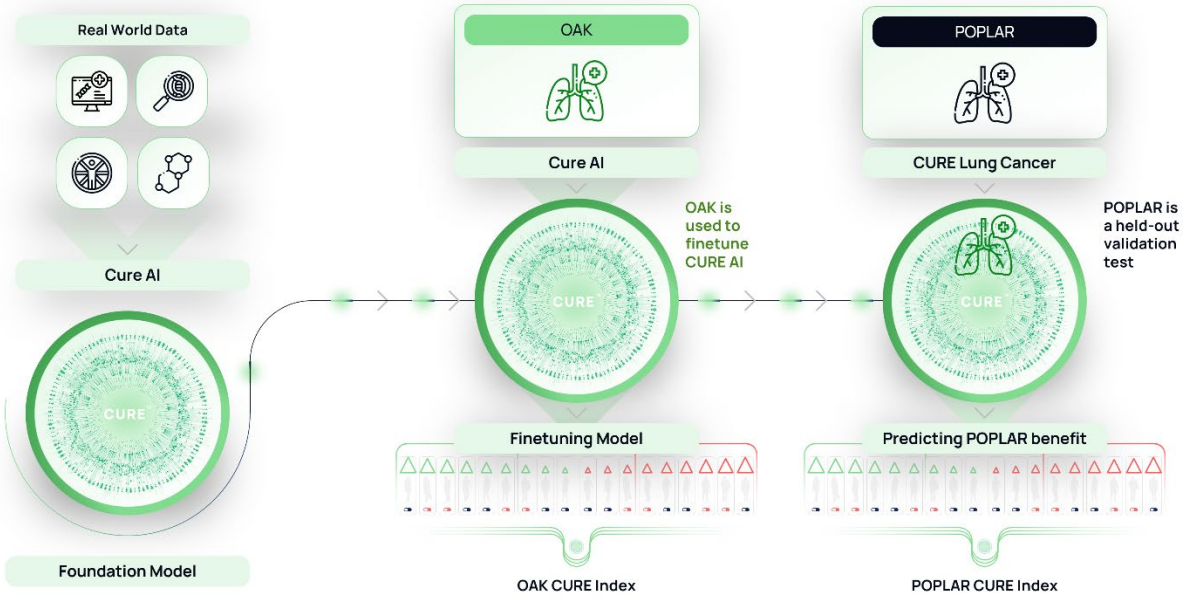

Figure 1. Outline of CURE AI Development, Finetuning, and Validation Testing. The CURE AI foundation model was trained on vast clinical and multiomic datasets to learn and understand the interplay between clinical and multiomic features. CURE AI was applied to the OAK clinical trial to finetune the model for training (finetuning CURE AI into CURE Lung Cancer using conditional average treatment effect modeling) and cross-validation (testing how well CURE AI predicts outcomes on OAK). CURE Lung Cancer was then validation tested on the completely held-out POPLAR clinical trial dataset. The CURE Index is created as an output from the models, from which we define CURE AI-informed simulated trial enrollment.

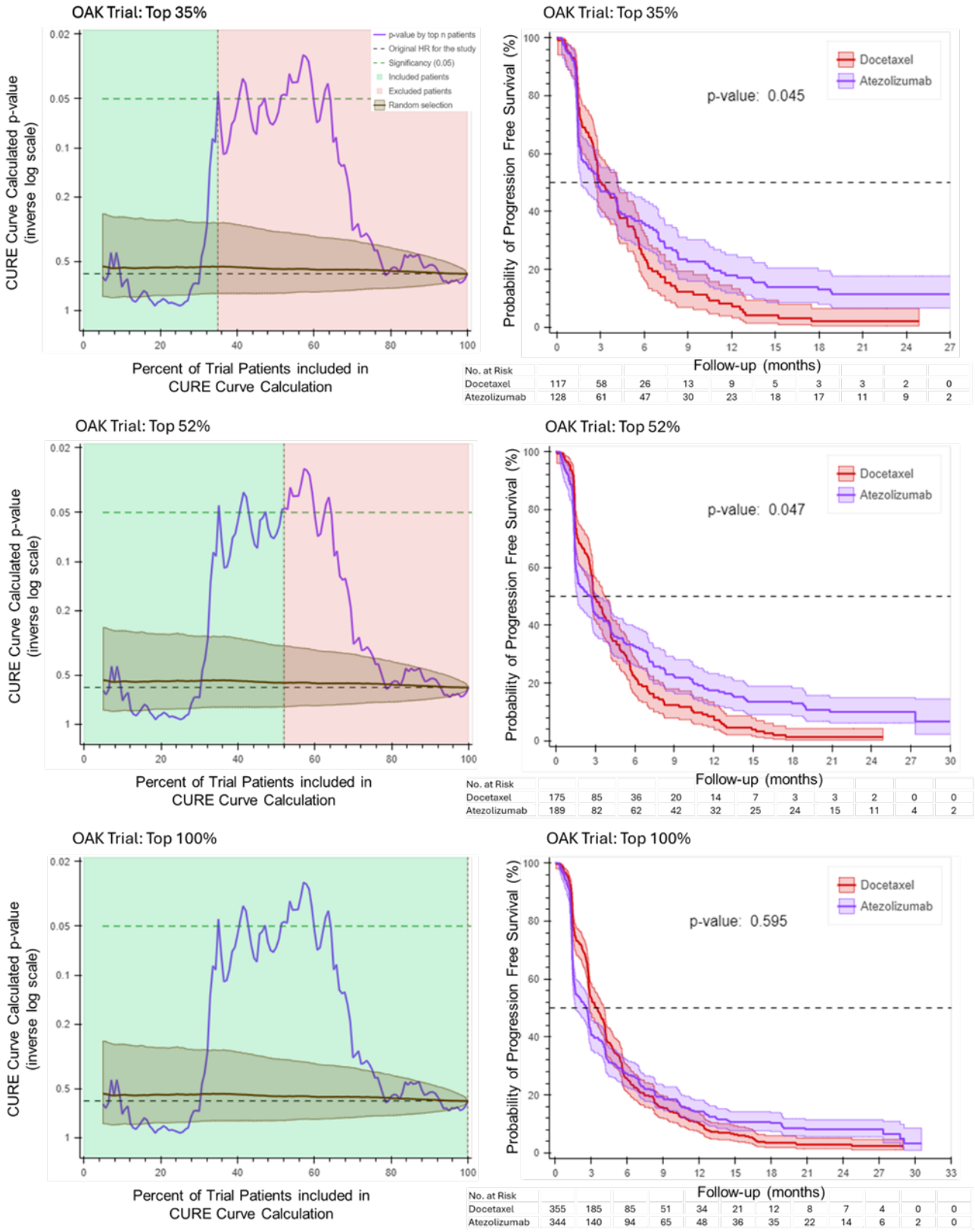

Figure 2. CURE Curves and Kaplan-Meier Estimates for the OAK Trial

In the OAK trial, the CURE Curve shows that patients enrolled up to around the 65th percentile (where the p-value line begins to descend) are more likely to benefit from atezolizumab than docetaxel. When the CURE Curve begins to descend, patients who have more predicted benefit to docetaxel are included, which results in loss of trial significance in the total trial population as the standard of care arm outperforms atezolizumab in the >65th percentile subgroup.

Top Panel. The top 35% of CURE Index defined patients were selected. At this threshold, the CURE Curve reaches significance ($p < 0.05$), with the corresponding Kaplan-Meier estimation shown on the right panel. This CURE Index is equivalent to the top 44% of patients on the POPLAR trial.
Middle Panel. The top 52% of CURE Index defined patients were selected. At this threshold, the CURE Curve remains above significance ($p < 0.05$), with the corresponding Kaplan-Meier estimation shown on the right panel. This CURE Index is identical to the top 88% of patients on the POPLAR trial.
Bottom Panel. The entire included trial population is shown with the corresponding Kaplan-Meier estimation shown on the right panel ($p = 0.595$).
For each Kaplan-Meier estimate, the number of patients at risk (alive and evaluable) at each timepoint is shown.

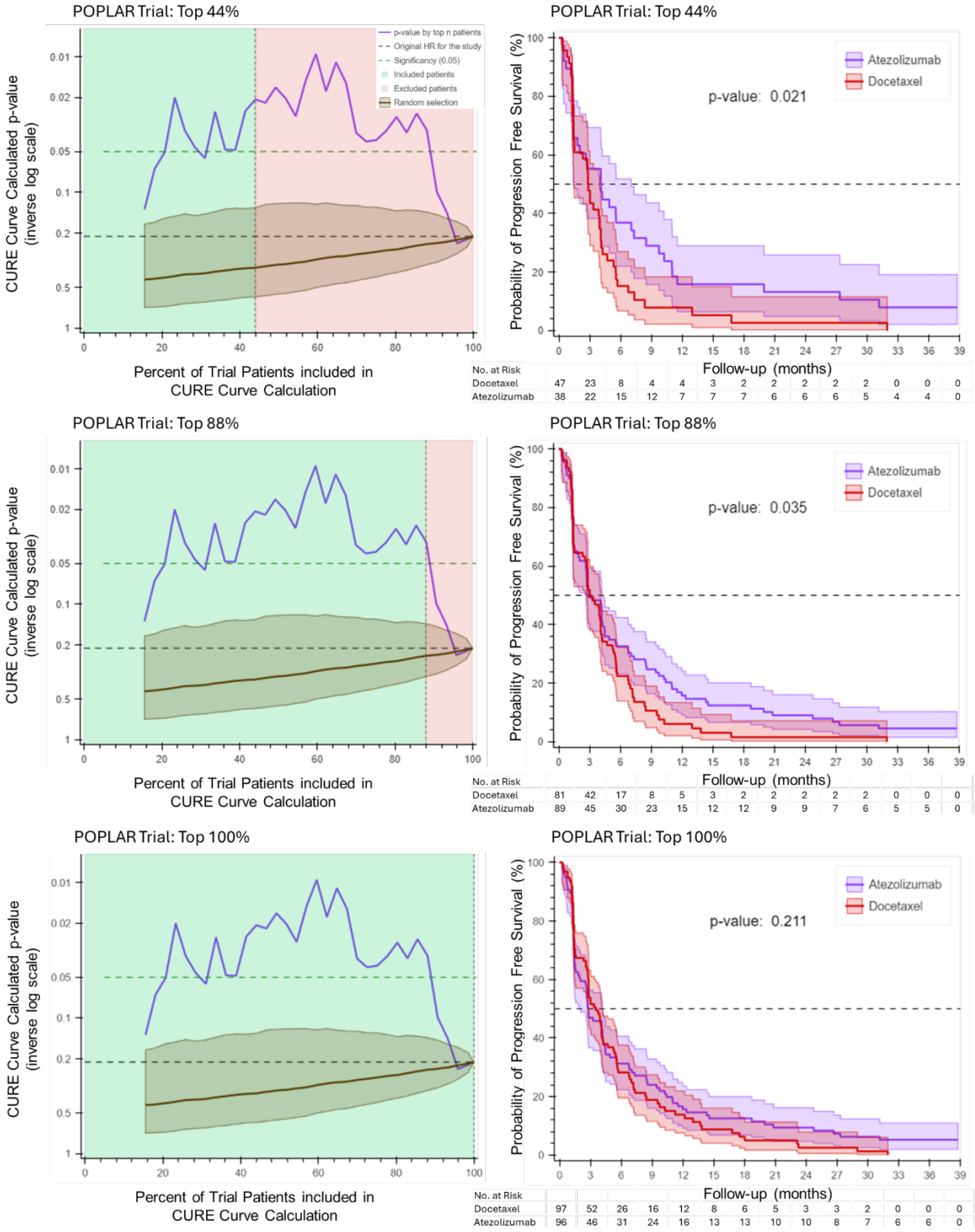

Figure 3. CURE Lung Cancer Curves and Kaplan-Meier Estimates for the POPLAR Trial

In the POPLAR trial, the CURE Curve shows that patients enrolled up to around the 88th percentile (where the p-value line begins to descend) are more likely to benefit from atezolizumab than docetaxel. When the CURE Curve begins to descend, patients who have more predicted benefit to docetaxel are included, which results in loss of trial significance in the total trial population as the standard of care arm outperforms atezolizumab in the >88th percentile subgroup. Note that the POPLAR trial was analyzed on CURE Lung Cancer (CURE AI that has been finetuned on the OAK trial).

Top Panel. The top 44% of CURE Index defined patients were selected. At this threshold, the CURE Curve passes significance ($p < 0.05$), with the corresponding Kaplan-Meier estimation shown on the right panel.
Middle Panel. The top 88% of CURE Index defined patients were selected. At this threshold, the CURE Curve remains above significance ($p < 0.05$), with the corresponding Kaplan-Meier estimation shown on the right panel.
Bottom Panel. The entire included trial population is shown with the corresponding Kaplan-Meier estimation shown on the right panel ($p = 0.211$).
For each Kaplan-Meier estimate, the number of patients at risk (alive and evaluable) at each timepoint is shown.

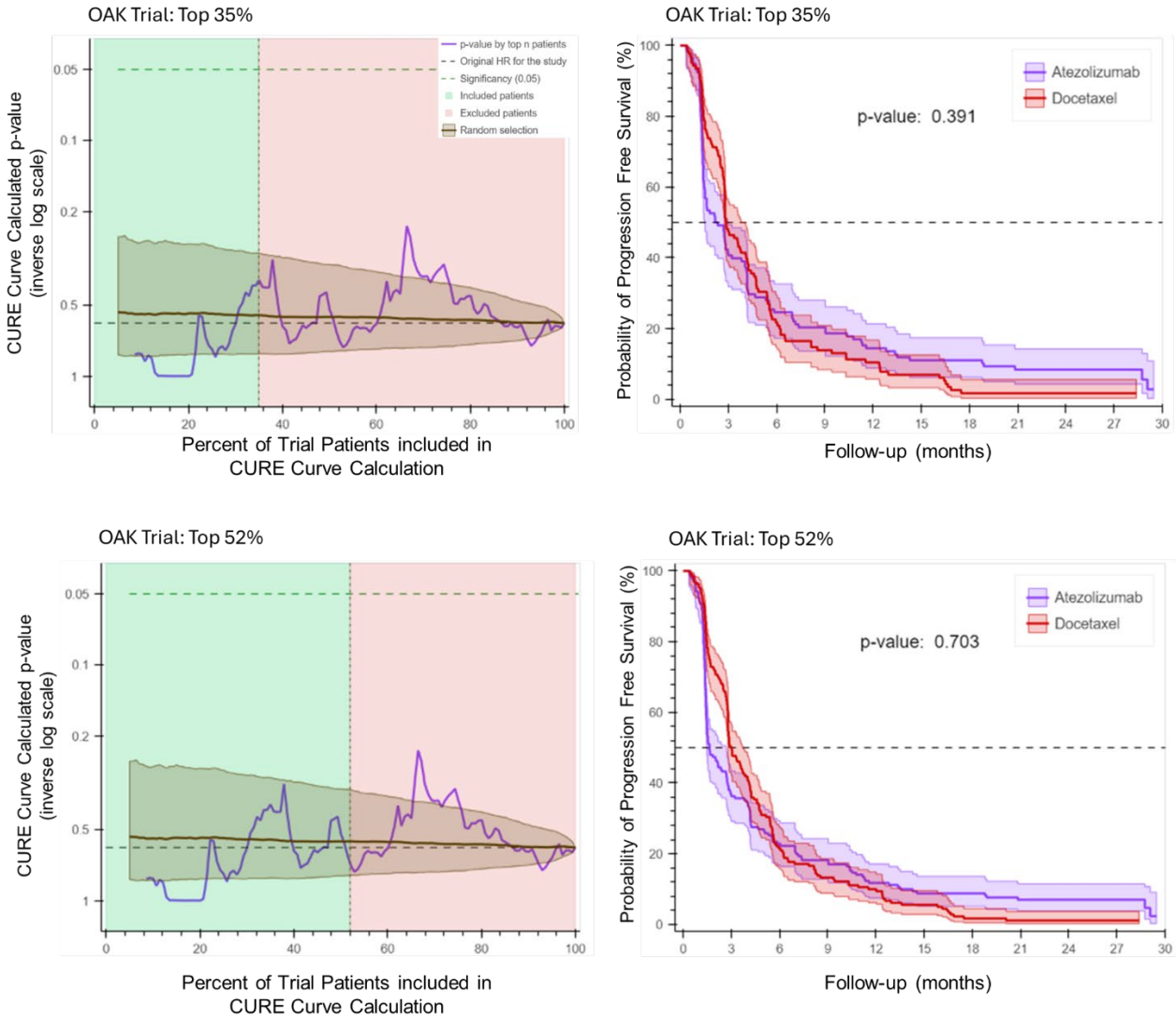

Figure 4. CURE Permutation Test and Kaplan-Meier Estimates for the OAK Trial

Using the CURE Permutation test, when PFS is randomized within each arm of the OAK trial, CURE AI does not identify any significant associations between clinicogenomic variables and patient outcome. The top 35% and top 52% of CURE Index defined patients during permutation testing are demonstrated. The CURE Curve is not significantly different than the random selection enrollment ($p > 0.05$ at all points), with the corresponding Kaplan-Meier estimates shown on the right panel. This loss of significance is a test of confidence in the CURE AI model.

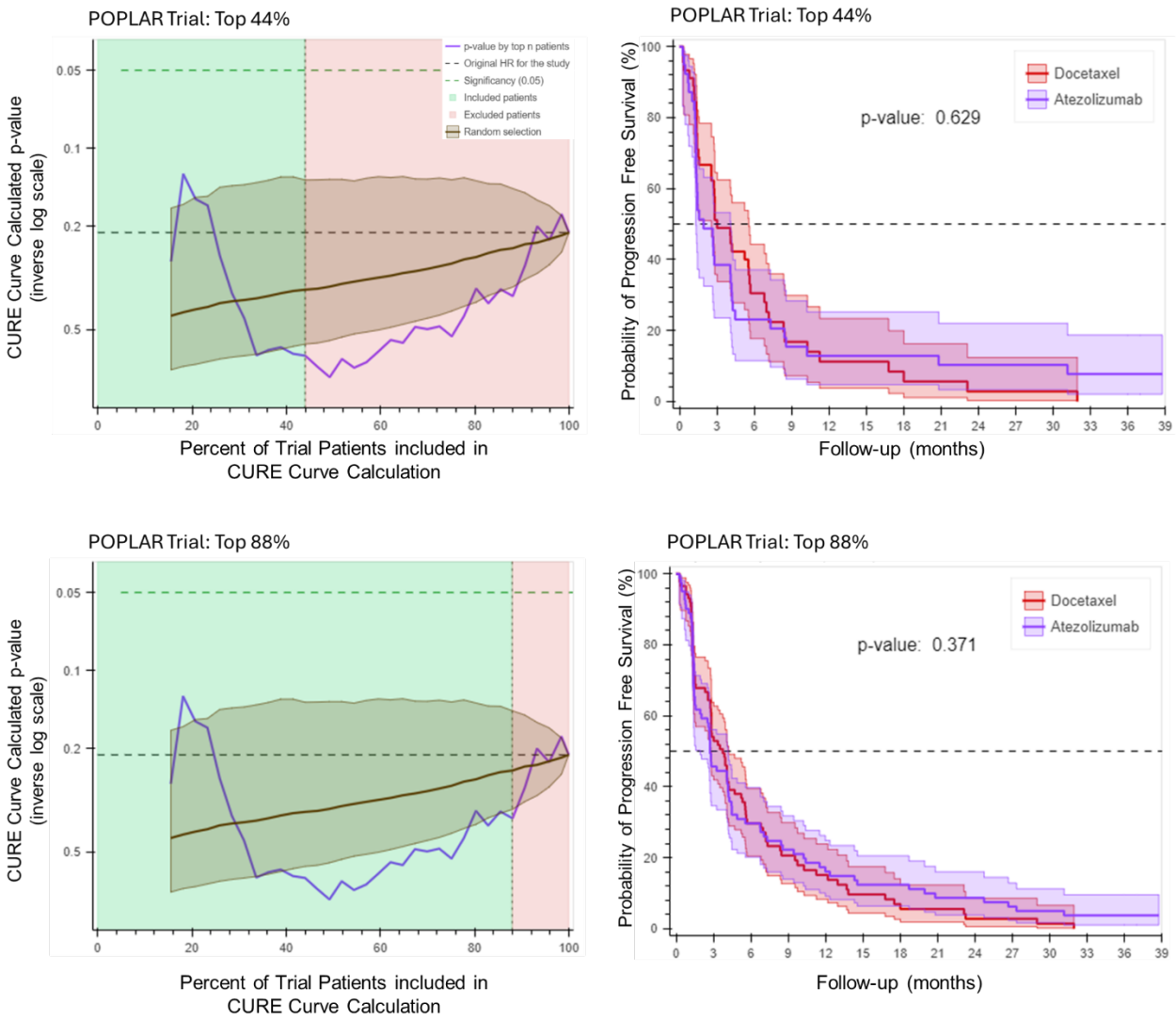

Figure 5. CURE Permutation Test and Kaplan-Meier Estimates for the POPLAR Trial

Using the CURE Permutation test, when PFS is randomized within each arm of the POPLAR trial, CURE AI does not identify any significant associations between clinicogenomic variables and patient outcome. The top 44% and top 88% of CURE Index defined patients during permutation testing are demonstrated. The CURE Curve is not significantly different than the random selection enrollment ($p > 0.05$ at all points), with the corresponding Kaplan-Meier estimates shown on the right panel. This loss of significance is a test of confidence in the CURE AI model.

Tables

OAK 35

Missing	Variable		35% OAK	Atezolizumab	Docetaxel	P-value
		Patient Number	245 (100)	128 (52.2)	117 (47.8)	
0	HIST, n (%)	Non-Squamous	164 (66.9)	88 (68.8)	76 (65.0)	0.621
		Squamous	81 (33.1)	40 (31.2)	41 (35.0)	
91	PD-L1, n (%)	[1,50)	79 (51.3)	42 (51.9)	37 (50.7)	1.000
91 unknown		[50,100]	75 (48.7)	39 (48.1)	36 (49.3)	
0	Gender, n (%)	Female	57 (23.3)	34 (26.6)	23 (19.7)	0.260
		Male	188 (76.7)	94 (73.4)	94 (80.3)	
113	TMB, n (%) 113	< 16 (low)	84 (63.6)	37 (61.7)	47 (65.3)	0.804
unknown		> 16 (high)	48 (36.4)	23 (38.3)	25 (34.7)	

OAK 52

Missing	Variable		52% OAK	Atezolizumab	Docetaxel	P-value
		Patient Number (n)	364 (100)	189 (51.9)	175 (48.1)	
0	HIST, n (%)	Non-Squamous	253 (69.5)	136 (72.0)	117 (66.9)	0.346
		Squamous	111 (30.5)	53 (28.0)	58 (33.1)	
93	PD-L1, n (%) 93	[0,1)	87 (32.1)	45 (32.1)	42 (32.1)	0.973
unknown		[1,50)	105 (38.7)	55 (39.3)	50 (38.2)	
		[50,100]	79 (29.2)	40 (28.6)	39 (29.8)	
0	Gender, n (%)	Female	96 (26.4)	49 (25.9)	47 (26.9)	0.934
		Male	268 (73.6)	140 (74.1)	128 (73.1)	
143	TMB, n (%) 143	< 16 (low)	149 (67.4)	67 (62.6)	82 (71.9)	0.183
unknown		> 16 (high)	72 (32.6)	40 (37.4)	32 (28.1)	

OAK 100

Missing	Variable		100% OAK	Atezolizumab	Docetaxel	P-value
		Patient Number (n)	699 (100)	344 (49.2)	355 (50.8)	
0	HIST, n (%)	Non-Squamous	509 (72.8)	257 (74.7)	252 (71.0)	0.307
		Squamous	190 (27.2)	87 (25.3)	103 (29.0)	
338	PD-L1, n (%) 338	[0,1)	177 (49.0)	85 (47.2)	92 (50.8)	0.769
unknown		[1,50)	105 (29.1)	55 (30.6)	50 (27.6)	
		[50,100]	79 (21.9)	40 (22.2)	39 (21.5)	
0	Gender, n (%)	Female	258 (36.9)	125 (36.3)	133 (37.5)	0.818
		Male	441 (63.1)	219 (63.7)	222 (62.5)	
274	TMB, n (%) 274	< 16 (low)	323 (76.0)	157 (75.5)	166 (76.5)	0.895
unknown		> 16 (high)	102 (24.0)	51 (24.5)	51 (23.5)	

Table 1. OAK Trial Patient Demographics by CURE Curve Threshold (top 35% of patients, top table; top 52% of patients, middle table; all included patients, bottom table). Number of patients with an unknown variable is listed in the left column. These data demonstrate that CURE Lung Cancer patient group demographics are not significantly altered compared to all included patients for the OAK trial.

POLAR 54

Missing	Variable		44% POLAR	Atezolizumab	Docetaxel	P-value
		Patient Number (n)	85 (100)	38 (44.7)	47 (55.3)	
0	HIST, n (%)	Non-Squamous	48 (56.5)	21 (55.3)	27 (57.4)	1.000
		Squamous	37 (43.5)	17 (44.7)	20 (42.6)	
21	PD-L1, n (%) 21 unknown	[1,50)	44 (68.8)	17 (65.4)	27 (71.1)	0.837
		[50,100]	20 (31.2)	9 (34.6)	11 (28.9)	
0	Gender, n (%)	Female	24 (28.2)	6 (15.8)	18 (38.3)	0.040
		Male	61 (71.8)	32 (84.2)	29 (61.7)	
0	TMB, n (%)	< 16 (low)	20 (23.5)	9 (23.7)	11 (23.4)	1.000
		≥ 16 (high)	65 (76.5)	29 (76.3)	36 (76.6)	

POLAR 88

Missing	Variable		88% POLAR	Atezolizumab	Docetaxel	P-value
		Patient Number (n)	170 (100)	89 (52.4)	81 (47.6)	
1	HIST, n (%) 1 unknown	Non-Squamous	101 (59.8)	54 (61.4)	47 (58.0)	0.775
		Squamous	68 (40.2)	34 (38.6)	34 (42.0)	
21	PD-L1, n (%) 21 unknown	[0,1)	85 (57.0)	51 (66.2)	34 (47.2)	0.058
		[1,50)	44 (29.5)	17 (22.1)	27 (37.5)	
		[50,100]	20 (13.4)	9 (11.7)	11 (15.3)	
0	Gender, n (%)	Female	56 (32.9)	25 (28.1)	31 (38.3)	0.212
		Male	114 (67.1)	64 (71.9)	50 (61.7)	
0	TMB, n (%)	< 16 (low)	51 (30.0)	25 (28.1)	26 (32.1)	0.688
		≥ 16 (high)	119 (70.0)	64 (71.9)	55 (67.9)	

POLAR 100

Missing	Variable		100% POLAR	Atezolizumab	Docetaxel	P-value
		Patient Number (n)	193 (100)	96 (49.7)	97 (50.3)	
1	HIST, n (%) 1 unknown	Non-Squamous	120 (62.5)	59 (62.1)	61 (62.9)	1.000
		Squamous	72 (37.5)	36 (37.9)	36 (37.1)	
37	PD-L1, n (%) 37 unknown	[0,1)	92 (59.0)	55 (67.9)	37 (49.3)	0.056
		[1,50)	44 (28.2)	17 (21.0)	27 (36.0)	
		[50,100]	20 (12.8)	9 (11.1)	11 (14.7)	
0	Gender, n (%)	Female	72 (37.3)	28 (29.2)	44 (45.4)	0.029
		Male	121 (62.7)	68 (70.8)	53 (54.6)	
0	TMB, n (%)	< 16 (low)	51 (26.4)	25 (26.0)	26 (26.8)	1.000
		≥ 16 (high)	142 (73.6)	71 (74.0)	71 (73.2)	

Table 2. POPLAR Trial Patient Demographics by CURE Curve Threshold (top 44% of patients, top table; top 88% of patients, middle table; all included patients, bottom table). Number of patients with an unknown variable is listed in the left column. These data demonstrate that the CURE Lung Cancer patient group demographics are not significantly altered compared to all included patients for the POPLAR trial.

References

1. Hwang S, Kwon AY, Jeong JY, et al. **Immune gene signatures for predicting durable clinical benefit of anti-PD-1 immunotherapy in patients with non-small cell lung cancer**. *Sci Rep*. 2020;10:643.

2. Motzer RJ, Robbins PB, Powles T, et al. **Avelumab plus axitinib versus sunitinib in advanced renal cell carcinoma: biomarker analysis of the phase 3 JAVELIN Renal 101 trial**. *Nat Med*. 2020;26(11):1733-1741.

3. Rittmeyer A, Barlesi F, Waterkamp D, et al.; OAK Study Group. **Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial**. *Lancet*. 2017;389(10066):255-265.

4. Fehrenbacher L, Spira A, Ballinger M, et al.; POPLAR Study Group. **Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial**. *Lancet*. 2016;387(10030):1837-1846.

5. Mazieres J, Rittmeyer A, Gadgeel S, et al. **Atezolizumab Versus Docetaxel in Pretreated Patients With NSCLC: Final Results From the Randomized Phase 2 POPLAR and Phase 3 OAK Clinical Trials**. *J Thorac Oncol*. 2021;16(1):140-150.

6. Harrer S, Shah P, Antony B, Hu J. **Artificial Intelligence for Clinical Trial Design**. *Trends Pharmacol Sci*. 2019;40(8):577-591.

7. Gupta R, Srivastava D, Sahu M, et al. **Artificial intelligence to deep learning: machine intelligence approach for drug discovery**. *Mol Divers*. 2021;25:1315-1360.

8. Askin S, Burkhalter D, Calado G, El Dakrouni S. **Artificial Intelligence Applied to clinical trials: opportunities and challenges**. *Health Technol*. 2023;13:203-213.

9. Azuaje F. **Artificial intelligence for precision oncology: beyond patient stratification**. *npj Precision Oncology*. 2019;3:6. doi:10.1038/s41698-019-0078-1. Published February 25, 2019.

10. Bhalla S, Laganà A. **Artificial Intelligence for Precision Oncology**. In: Bhalla S, Laganà A, eds. *Computational Methods for Precision Oncology*. 1st ed. Springer; 2022:249-268. Published March 1, 2022.

11. Singh S, Kumar R, Payra S, Singh SK. **Artificial Intelligence and Machine Learning in Pharmacological Research: Bridging the Gap Between Data and Drug Discovery**. *Cureus*. 2023;15(8):e44359.

12. Harrer S, Shah P, Antony B, Hu J. **Artificial Intelligence for Clinical Trial Design**. *Trends Pharmacol Sci*. 2019;40(8):577-591.

13. Temple R. **Enrichment of Clinical Study Populations**. *Clin Pharmacol Ther*. 2010;88(6):774-778.

14. U.S. Food and Drug Administration. **Enrichment Strategies for Clinical Trials to Support Determination of Effectiveness of Human Drugs and Biological Products:**

Guidance for Industry. Silver Spring, MD: FDA; 2019. Available at: https://www.fda.gov/media/121320/download

15. Magnuson A, Bruinooge SS, Singh H, et al. **Modernizing Clinical Trial Eligibility Criteria: Recommendations of the ASCO-Friends of Cancer Research Performance Status Work Group**. *Clin Cancer Res*. 2021;27(9):2424-2429.

16. Beaver JA, Ison G, Pazdur R. **Reevaluating Eligibility Criteria - Balancing Patient Protection and Participation in Oncology Trials**. *N Engl J Med*. 2017;376(16):1504-1505. doi: 10.1056/NEJMp1615879.

17. Tufts CSDD. **Rising protocol design complexity is driving rapid growth in clinical trial data volume**. *Tufts CSDD Impact Report*. 2021;23(1):1-6. January/February.

18. Mitra R, McGough SF, Chakraborti T, et al. **Learning from data with structured missingness**. *Nat Mach Intell*. 2023;5:13-23. Published January 25, 2023.

19. Simon R. **Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology**. *Per Med*. 2010;7(1):33–47. doi: 10.2217/pme.09.49.

20. Yadav K, Lewis RJ. **Immortal Time Bias in Observational Studies**. JAMA. 2021;325(7):686–687. doi:10.1001/jama.2020.9151

21. Imbens GW, Rubin DB. **Rubin Causal Model**. In: *Microeconometrics*. Springer; 2010:229-241.

22. Chiba Y. **Causal Measures for Prognostic and Predictive Biomarkers**. *Open J Stat*. 2018;8(2):214-221. doi: 10.4236/ojs.2018.82014.

23. Yadav S, Shukla S. **Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification**. In: *2016 IEEE 6th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. IEEE; 2016:1-6.

24. Wang Q, Ma Y, Zhao K, Tian Y. **A Comprehensive Survey of Loss Functions in Machine Learning**. *Ann Data Sci*. 2022;9:187-212. Published April 12, 2020.

25. Melnychuk V, Frauen D, Feuerriegel S. **Causal Transformer for Estimating Counterfactual Outcomes**. In: *Proceedings of the 39th International Conference on Machine Learning*; 2022:15293-15329. PMLR 162.

26. Curth A, Svensson D, Weatherall J, van der Schaar M. **Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices in Treatment Effect Estimation**. In: *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*; 2021.

27. Stenzinger A, Allen JD, Maas J, et al. **Tumor mutational burden standardization initiatives: Recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions**. *Genes Chromosomes Cancer*. 2019;58(8):578–588. doi: 10.1002/gcc.22733.