


Research and Applications

Application of machine learning methods in clinical trials for precision medicine

Yizhuo Wang ¹, Bing Z. Carter², Ziyi Li¹, and Xuelin Huang ¹

¹Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA, and ²Section of Molecular Hematology and Therapy, Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Corresponding Author: Xuelin Huang, Department of Biostatistics, The University of Texas MD Anderson Cancer, Center, Houston, TX 77030, USA; xluang@mdanderson.org; Ziyi Li, Department of Biostatistics, The University of Texas MD Anderson Cancer, Center, Houston, TX 77030, USA; zli16@mdanderson.org

Received 16 September 2021; Revised 15 November 2021; Editorial Decision 18 November 2021; Accepted 1 December 2021

ABSTRACT

Objective: A key component for precision medicine is a good prediction algorithm for patients' response to treatments. We aim to implement machine learning (ML) algorithms into the response-adaptive randomization (RAR) design and improve the treatment outcomes.

Materials and Methods: We incorporated 9 ML algorithms to model the relationship of patient responses and biomarkers in clinical trial design. Such a model predicted the response rate of each treatment for each new patient and provide guidance for treatment assignment. Realizing that no single method may fit all trials well, we also built an ensemble of these 9 methods. We evaluated their performance through quantifying the benefits for trial participants, such as the overall response rate and the percentage of patients who receive their optimal treatments.

Results: Simulation studies showed that the adoption of ML methods resulted in more personalized optimal treatment assignments and higher overall response rates among trial participants. Compared with each individual ML method, the ensemble approach achieved the highest response rate and assigned the largest percentage of patients to their optimal treatments. For the real-world study, we successfully showed the potential improvements if the proposed design had been implemented in the study.

Conclusion: In summary, the ML-based RAR design is a promising approach for assigning more patients to their personalized effective treatments, which makes the clinical trial more ethical and appealing. These features are especially desirable for late-stage cancer patients who have failed all the Food and Drug Administration (FDA)-approved treatment options and only can get new treatments through clinical trials.

Key words: clinical trial, adaptive design, machine learning, precision medicine

Lay Summary

In a typical controlled clinical trial, patients are equally randomized to receive different treatments. However, it is possible that one treatment demonstrates advantages over others during the trial. Utilizing that information can benefit subsequent patients. This is why response-adaptive randomization (RAR), which allows uneven treatment assignment probabilities based on existing knowledge, has become popular recently. A key component of RAR is a good prediction algorithm for patients' response to treatments. Previous works have explored using machine learning (ML) to predict treatment response, but few incorporated ML methods into RAR. This study implements 9 commonly used ML methods into RAR trial designs. We further present an ML-ensemble RAR design that builds upon the majority consensus of the 9 ML methods' predictions. Extensive simulation studies and real-world applications show that using ML methods in RAR leads to the assignment of more patients to their optimal treatments, increasing the overall response rate. The proposed method will become a useful tool for future clinical trial design in the era of precision medicine.

INTRODUCTION

It is known that patients respond differently to the same treatments.¹ The demand for selecting the optimal treatment for each and every patient has resulted in a rapidly developing field called precision medicine, also known as personalized medicine.² This field aims to provide guidance to select the most effective treatment based on distinctive patient biomarkers. As clinical trials also evolve in the age of precision medicine, there is a substantial need for novel trial designs to deliver more ethical and precise care. Compared with classical nonadaptive trials, adaptive trials have become popular among clinicians as they integrate accumulating patient data to modify the parameters of the trial protocol, provide personalized treatment assignment, and ultimately optimize patients' outcomes. For example, the adaptive designs in phase 2/3 clinical trials take advantage of the interim treatment response data during the course of the trial and allocate more patients to the presumably more effective treatments.³

Among different adaptive designs, one common adaptation is response-adaptive randomization (RAR). It refers to the adjustments of treatment allocations based on intermediate patient responses and new patients' characteristics collected during the clinical trials. This RAR design is useful when the interaction between biomarkers and treatments are only putative or not known at the beginning of a trial, and it is also practical when there are multiple treatments to be considered. Its ultimate objective is to provide more patients with their personalized optimal therapies according to their biomarker profiles. The starting point of RAR can be traced back to Thompson,⁴ who proposed employing a posterior probability estimated from the interim data to assign patients to the more effective treatment. Following his idea, the application of Bayesian methods with an inherent adaptive nature has boomed in area of RAR designs.^{5–13}

Currently, there are several major successes in applying Bayesian RAR concepts in clinical trials, from protocol development through legitimate registration. The BATTLE-1 trial (Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination) for patients with advanced non-small cell lung cancer (NSCLC) and the I-SPY 2 trial (Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis) for patients with breast cancer in the setting of neoadjuvant chemotherapy are 2 biomarker-based, Bayesian RAR clinical trials.^{14,15} However, Bayesian RAR designs have a number of challenges and limitations. Due to the modeling restrictions, Bayesian RAR methods usually consider only a very small number of biomarkers. With complex diseases or symptoms, hundreds or even thousands of biomarkers may need to be considered at the same time for treatment assignment.¹⁶ Also, some Bayesian RAR methods adjust the design by separating

the cohort based on the existence of biomarker(s), and thus these methods rely heavily on how well the biomarker(s) interact(s) with the treatments. If the biomarker is chosen incorrectly, it is possible to make wrong adjustments afterwards.^{1,17}

As the development of modern sequencing technology, clinicians have faced a massive volume of high dimensional data with a complex, nonlinear structure. How to build an effective and scalable algorithm for randomization becomes a fundamental question for the research of RAR trial designs. Machine learning (ML) methods have been applied to solve many real-world problems and have successfully demonstrated their strengths in processing large data sets, as well as capturing nonlinear data structures. With the expectations and resources to analyze this large amount of complex healthcare data, ML methods have established their supremacy in disease prediction,¹⁸ disease classification,¹⁹ imaging diagnosis,²⁰ drug manufacturing,²¹ medication assignment,²² and genomic feature identification tasks.²³

Although several supervised ML approaches have been applied to drug response prediction,^{10,24–34} little of the work has explored incorporating ML methods into RAR trial designs. In this study, we implemented 9 ML algorithms into RAR designs and further presented an ML-ensemble RAR design combining these 9 ML algorithms. Specifically, ML methods help to match patient biomarker profiles with prediction of treatment outcomes and, in turn, have determined treatment allocation for future patients. These ML methods are able to address large data and complex structures. We have successfully demonstrated, in both simulation study and a real-world example, that ML-based RAR designs have higher response rates as there are more patients receiving effective treatments. The ensemble method outperformed all other single ML methods.

MATERIALS AND METHODS

Adaptive design: response-adaptive randomization

In clinical trial design, adaptive design means making changes to the trial protocol after the trial has started and some data have been collected. These changes are based on the information from the collected data, including (1) the total sample size, (2) interim analyses, (3) patient allocation to different treatment arms, and more.³⁵ For (3), it refers to the RAR design in which the treatment allocation probability varies in order to favor the treatment estimated to be more effective and to increase the response rate in patients. The initial concept can be traced back to Thompson⁴ and Robbins,³⁶ and led to others. Some famous RAR trials include the extracorporeal membrane oxygenation (ECMO) trial, which tested the efficacy of ECMO in patients with severe acute respiratory distress syndrome

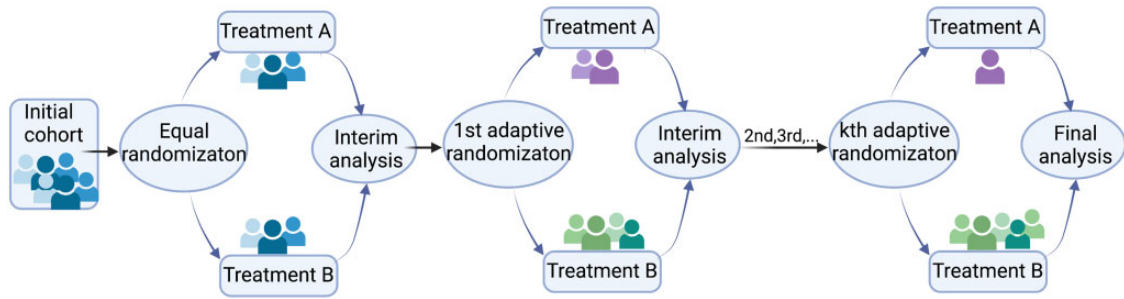


Figure 1. Response-adaptive randomization (RAR) design. The number of adaptive randomization is adjustable per application.

(ARDS),³⁷ and the first large-scale double-blind, placebo-controlled study which tested the superiority of fluoxetine over placebo in children and adolescents with depression.³⁸ A general scheme of the RAR designs is shown in Figure 1.

Benchmark design: equal randomization

Randomization as a standard means for addressing the selection bias in treatment assignments has been extensively used in clinical trials.³⁵ It helps to achieve balance among treatment groups and accounts for the genuine uncertainty about which treatment is better at the beginning of the trial. Randomly assigning patients to treatment arms on a 1:1 basis is known as equal randomization (ER). Friedman et al³⁹ (p. 41) presented that equal allocation in principle maximizes statistical power and is consistent with the concept of equipoise that should exist before the trial starts. Here, we used the ER design as a benchmark randomization design to evaluate the performance of ML-based RAR designs.

Allocation rule

The key of the proposed method is to model the relationship of patient responses and biomarkers. Such a model will then predict the response rate of each treatment for each new patient and provide guidance for treatment assignment. In detail, current enrolled patients' biomarker profiles and treatment response data were used to train ML models, which later were used to predict future patients' treatment responses based on their biomarker profiles. Given treatment A and B, the probability of allocating each treatment for patient i is shown as below:

$$p_{iA} = \frac{\pi_{iA}}{\pi_{iA} + \pi_{iB}}$$

$$p_{iB} = \frac{\pi_{iB}}{\pi_{iA} + \pi_{iB}}$$

where π_{iA} and π_{iB} , respectively, denote the response probability of treatment A and B for patient i predicted by the ML model.

ML algorithms and a ML ensemble

We selected 9 mainstream ML algorithms and implemented them in the RAR design to predict treatment response. The prediction models were built using the best-fitting parameters for each model, which were obtained by the grid search method with a 10-fold cross-validation.^{40,41} Grid search is a standard method which allows us to try a variety of tuning parameter combinations for the model within a reasonable amount of time. The 10-fold cross-validation performs the fitting process for a total of 10 times with randomly selected nine-tenths of the data (90%) to train the model in each fit and the rest of

the data to validate. By doing this, we avoid bias from using a random single split. The selected model will generalize better to all of the samples in the dataset. Combining grid search with cross-validation, we evaluate the performance of each parameter combination and select the best parameters for each ML model. Here we conducted this hyperparameter tuning procedure in R using the "Caret" package⁴²; similar techniques are available in the scikit-learn Python ML library.⁴³ These selected ML algorithms can be roughly divided into 2 categories:

1. Parametric models: logistic regression,⁴⁴ LASSO regression,⁴⁵ and Ridge regression.⁴⁶
2. Nonparametric models: gradient boosting machine (GBM),⁴⁷ random forest (RF),⁴⁸ support vector machine (SVM),⁴⁹ Naive Bayes,⁵⁰ k-nearest neighbors (KNN),⁵¹ and artificial neural networks (NNs).⁵²

For logistic regression, Ridge regression and Lasso regression, they are all considered parametric models. In detail, logistic regression assumes the linearity of independent variables and log odds. It is a particular form of GLM.²⁴ Ridge regression and LASSO regression assume that there is a linear relationship between the "dependent" variable and the explanatory variables. They are 2 regularization methods of GLM to prevent an over-fitting issue by adding penalties on the predictor variables that are less significant.^{46,53}

KNN, which classifies data points based on the points that are most similar to it, is a typical nonparametric model such that there is no assumption for underlying data distribution, and the number of parameters grows with the size of the data.⁵¹ With NNs, however, there has been some debate regarding whether they belong to parametric or nonparametric methods. NNs typically consist of 3 layers: input layer, hidden layer, and output layer. Here we classify NNs as a nonparametric method, as the network architecture grows adaptively to match the complexity of given data.⁵²

Both GBM and RF are nonparametric methods that consist of sets of decision trees. Specifically, GBM builds one tree at a time and each new tree helps to correct errors made by previously trained tree by adding weights to the observations with the worst prediction from the previous iteration; RF trains each tree independently using a random sample of the data, and the results are aggregated in the end.^{47,48}

NB and SVM can be either parametric or nonparametric depending on whether they use kernel tricks. For the NB classifier, it becomes nonparametric if using a kernel density estimation (KDE) to obtain a more realistic estimate of the probability of an observation belonging to a class.⁵⁰ And for SVM, the basic idea is finding a hyperplane that best divides a dataset into 2 classes. It is considered a nonparametric when using the kernel trick to find this hyperplane. This is because the kernel is constructed by computing the pair-wise

distances between the training points, and the complexity of the model grows with the size of the dataset.⁴⁹

Combing these 9 models, an ML ensemble method was built and implemented in the RAR design to obtain a better treatment allocation rule. We defined the treatment allocation probability function for patient i as follows:

$$p_i = \begin{cases} p_t, & \text{if } m \geq \theta \\ m/M, & \text{otherwise} \end{cases}$$

where m is the number of agreed models, $M = 9$ is the total number of models, θ is the threshold number of agreed models, and p_t is the threshold treatment allocation probability. Here we chose $\theta = 7$ and $p_t = 0.85$ for $m \geq \theta$; these threshold values can be adjusted accordingly for different application purposes. To further understand the impacts of selecting different parameter values, we did simulations using different threshold number of agreed models ($\theta = 5/6/7/8$) and different threshold allocation probabilities ($p_t = 0.7/0.75/0.8/0.85/0.9$). The results are shown in Supplementary Figures S1–S4. In Supplementary Figures S1 and S3 right-sided figures, when the treatment main effect is high, as the threshold parameter increases, the response rate of the ensemble method increases and the individual loss decreases. This intuitively makes sense because when the consensus method reaches the correct decision, increasing p_t will increase the probability of patients receiving their optimal treatments (Supplementary Figure S2). When the treatment main effect is low, the increments of the response rate are not very significant (Supplementary Figure S1, left), but we still observe obvious differences regarding the optimal treatment percentage and the individual loss (Supplementary Figures S2 and S3, left). Meanwhile, it is still desirable to maintain some randomness of treatment assignment in a clinical trial and thus an allocation probability of 1 for the optimal treatment is not recommended. For Supplementary Figure S4, we show the results of the response rate and the optimal treatment percentage, and we can see that the difference of using different threshold values of agreed ML models is minor.

Apart from comparing with the ER “benchmark” design, the current study also examined whether the ML ensemble could assign more patients to the best available treatment beyond other ML methods in adaptive design with the same assessment of individuals.

Inverse probability of treatment weighting

Similar to the observational study in which certain outcomes are measured without attempting to change the outcome, the treatment selection of future patients in RAR trials is often influenced by individual characteristics of the initial block of patients.²⁵ As a result, when estimating the impact of treatment on responses, systemic variations in baseline characteristics between differently treated individuals must be taken into account. Here we applied the inverse probability of treatment weighted (IPTW) method to decrease or remove the effects of confounding when using the observational data to estimate the treatment effects. The idea of IPTW is to use weights based on the propensity score to create a synthetic sample in which the distribution of baseline characteristics is independent of treatment.⁵⁴ The propensity score refers to the probability of treatment allocation tied to the observed individual characteristics. And the weight based on it is defined as follows:

$$w_i = \frac{Z_i}{e_i} + \frac{(1 - Z_i)}{1 - e_i}$$

where Z_i is the treatment indicator and e_i is the propensity score for the i th subject. Different estimators for treatment effects based on

IPTW have been developed; here we used an estimator of the average treatment effect (ATE), which is defined as $E[Z_i(1) - Z_i(0)]$ where $Z_i(1) - Z_i(0)$ is the effect of treatment.⁵³ Incorporated with the IPTW idea, the ATE estimator is defined as follows:

$$\frac{1}{n} \sum_{i=1}^n \frac{Z_i Y_i}{e_i} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - Z_i) Y_i}{1 - e_i}$$

where Y_i denotes the response variable of the i th subject, n denotes the total number of subjects, and e_i still denotes the propensity score.⁵⁵

Evaluation metrics

Two commonly used criterion in the field of precision medicine, namely the overall response rate and the percentage of individuals receiving optimal treatments, are our primary evaluation metrics. The formulas of the response rate and the optimal treatment percentage are as follows:

$$\text{Response rate} = \frac{\text{No. of patients who responded}}{\text{No. of patients in the trial}}$$

$$\text{Optimal treatment percentage} = \frac{\text{No. of patients receiving their personalized optimal treatments}}{\text{No. of patients in the trial}}$$

The power and the average treatment effect (ATE) adjusted by the IPTW method were also reported to thoroughly evaluate each methods’ performance. The power of a clinical trial refers to the probability of detecting a difference between different treatment groups when it truly exists; ATE was defined in the previous section. Additionally, we proposed a new criterion, the individual loss to quantify the loss for each patient due to receiving suboptimal treatments. For the individual loss, we first define a match for the enrolled patients. A match occurs when the patient’s actual treatment received is the same as the best treatment from the true model. For an enrolled patient i with signature x , let $\hat{P}_i(Y = 1 | T, x)$ denote the probability of responding to the received treatment T . Let $\hat{P}_i(Y = 1 | T = OPT, x)$ denote the probability of responding to the optimal treatment determined by the true model. Then we define the personalized loss function as follows:

$$\text{Loss}(i) = \begin{cases} 0 & \text{if it is a match} \\ \hat{P}_i(Y = 1 | T, x) - \hat{P}_i(Y = 1 | T = OPT, x) & \text{if there is not a match} \end{cases}$$

A low individual loss value suggests that the majority of patients have received the treatment and will respond at least as well as the real model’s optimal therapy.

RESULTS

Simulation

We used simulation studies to evaluate the proposed methods.

Setting

We generated the i th patient’s response from a logistic regression model with 10 biomarkers:

$$\begin{aligned} \text{logit}(\eta_i) &= \alpha T_i + \sum_{j=1}^{10} \beta_{kj} f_j(X_{ji}) + \sum_{j=1}^{10} \gamma_j T_i X_{ji} + \epsilon_i, \quad i = 1, \dots, 300; j \\ &= 1, \dots, 10 \end{aligned}$$

where T_i is the treatment indicator (either treatment 0 or treatment 1), α is the treatment main effect coefficient, X_{ji} is the j th biomarker

for patient i , and $f_j(X_{ji})$ incorporates some polynomial and step function terms. γ_j is the biomarker-treatment interaction coefficient and $X_5 \sim X_6$ were assumed to interact with the treatment. A random noise for each subject is denoted as ϵ_i . In detail, each biomarker X_{ji} was assumed to follow a normal distribution with a mean of 0 and a standard deviation of 1. Among these 10 biomarkers, $X_2 \sim X_4$ contributed to the true model as third-degree polynomials, while $X_7 \sim X_{10}$ contributed to the true model as step functions:

$$s(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

Seven scenarios of different treatment main effects ($\alpha=0, 0.5, 0.7, 1, 1.3, 1.5, 1.7$) and a fixed treatment-biomarker interaction ($\gamma = 0.5$) were considered. We conducted 1000 Monte Carlo simulations for each scenario and compared the results obtained by the ML-based and ML-ensemble RAR designs with the results from the ER design.

Response rate and optimal treatment percentage

The response rate results and the percentage of receiving the optimal treatment are shown in Figure 2. Overall, the performance of ML-based RAR designs is better than the performance of the ER design. When the treatment main effect is zero, the differences for both response rate and the optimal treatment percentage between ML-based RAR designs and the ER design are not significant. As the treatment effect increases, these differences become more obvious. Among these 9 ML algorithms, the neural network has the highest response rate and the highest proportion of patients receiving their optimal treatments. Additionally, the ensemble method combining these 9 ML methods outperforms all other methods and achieves an approximate 5% higher response rate and a more than 20% larger optimal treatment percentage compared to the ER design.

Individual loss, ATE, and power

The individual loss and the ATE results are shown in Figure 3. The interpretation of the individual loss results coincides with the previous response rate results and the optimal treatment percentage results such that the ML-ensemble RAR design has the lowest indi-

vidual loss value among all scenarios, which is preferred in the trial. The ATE has been adjusted by the IPTW method to account for confounding effect of using observational data. The logistic regression method now has the highest ATE, followed by the NN method. The ensemble method has a relatively low ATE, but it is higher than the ER method when the treatment main effect becomes larger. This shows that the average effect of changing the entire population from untreated to treated using RAR designs is better than that of using the ER design.

Power

The power results are shown in Figure 4. The power is also weighted by the IPTW method to address potential bias. For the power analysis, the Type I error is controlled at 0.05. Several papers have shown in their simulation studies that the correlation among treatment assignments was inevitable when performing inference on the data from RAR design-implemented studies.^{56,57} This correlation can increase the binomial variability and lower the power. In our simulation, the RAR design using the NN method has the lowest power, followed by using the logistic regression method. However, other ML-based RAR designs have comparable or even higher power than that of the ER design. The ensemble method has a relatively low power, but it is still better than the NN method.

Real-world example

We analyzed a publicly available acute myeloid leukemia (AML) dataset from Kornblau et al⁵⁸ where most of the clinical biomarkers are expression levels of cellular proteins. Kornblau et al sequenced protein expressions in leukemia-enriched cells from 256 newly diagnosed AML patients with a primary goal of eventually establishing a proteomic-based categorization of AML. The treatment and the response variables were carefully adjusted to binary variables. Specifically, the treatments were binarized to high-dose ara-C (HDAC)-based treatments and non-HDAC treatments; the responses were binarized to complete response (CR) and non-CR.

We first performed a feature selection to decide what interaction terms should be included in the model. We used each protein-treatment interaction term to build the generalized linear model (GLM) model and reported the p-value for each interaction to assess

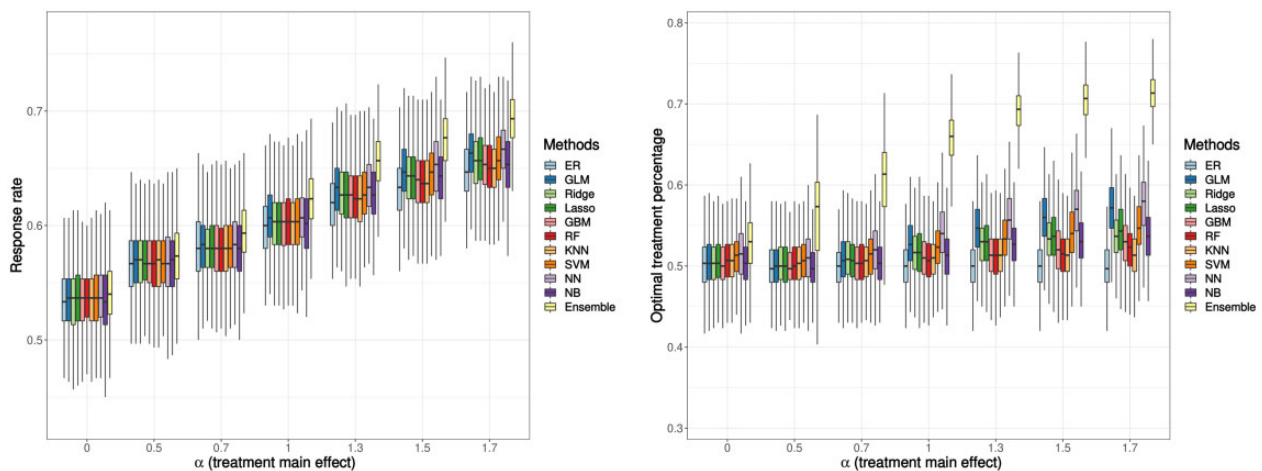


Figure 2. Simulation result: response rate (left), percentage of patients receiving their optimal treatments (right). The treatment-biomarker interaction, γ is fixed at 0.5. Boxplots display the median (middle line), the interquartile range (hinges), and 1.5 times the interquartile range (lower and upper whiskers) based on 1000 times simulation. The mean (over 1000 simulations) response rate ranges from 0.53 to 0.69, and the mean of optimal treatment percentages ranges from 0.50 to 0.71.

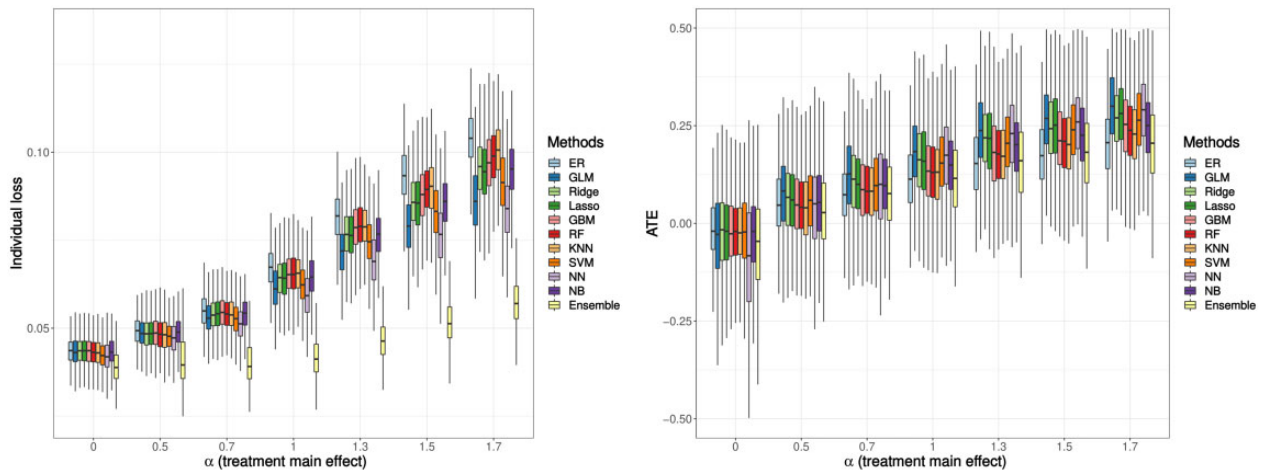


Figure 3. Simulation result: individual loss (left), average treatment effect (ATE, right). The treatment-biomarker interaction, γ is fixed at 0.5. Boxplots display the median (middle line), the interquartile range (hinges), and 1.5 times the interquartile range (lower and upper whiskers) based on 1000 times simulation. The mean (over 1000 simulations) individual loss ranges from 0.04 to 0.10, and the mean ATE ranges from -0.12 to 0.30.

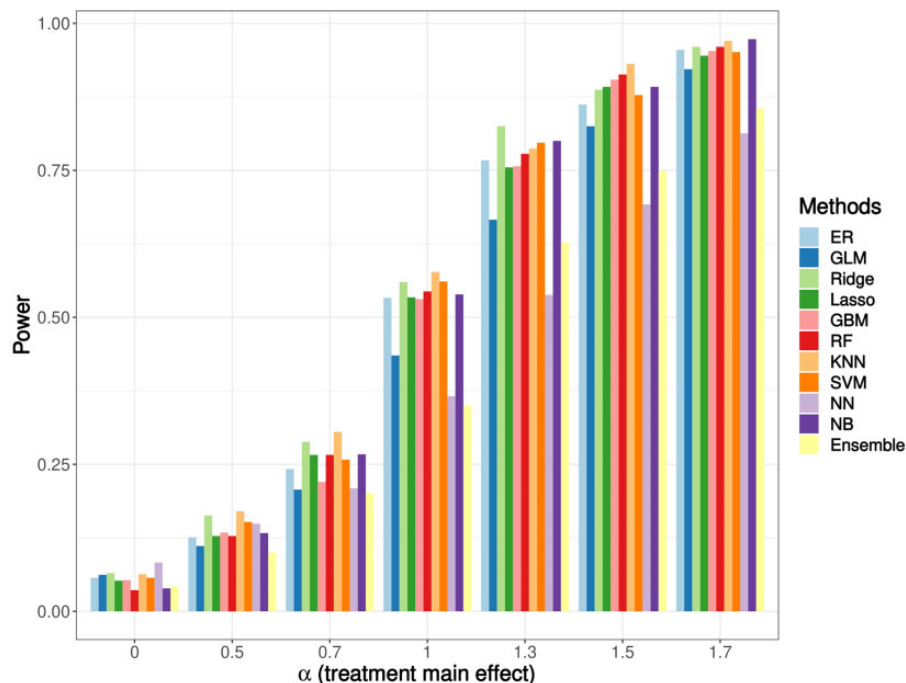


Figure 4. Simulation result: power. The treatment-biomarker interaction, γ is fixed at 0.5. The Type I error is controlled at 0.05. The power ranges from 0.04 to 0.97.

whether it has strong correlation with the dependent variable/the treatment response. The top 10 proteins whose interaction variables have the smallest P -values were selected. We then performed a gene network analysis on the genes that code for these proteins using GeneMANIA (<http://genemania.org>).²⁶ This analysis helps to illustrate the hidden interaction and network of the corresponding genes. Additionally, it shows other genes that have been reported to associate with the input 10 genes, using extensive existing knowledge such as protein and genetic interactions, pathways, co-expression, co-localization, and protein domain similarity. The results are presented in Figure 5. The top 10 genes corresponding to the biomarkers identified in our study are highlighted with red circles.

Using a cut-off P -value of 0.1 among 71 proteins, the expression levels of 3 of them were found to have the most significant interactions with the treatment, that is, the strongest correlation with the treatment outcomes: phosphothreonine 308 of Akt (Thr 308 p-Akt), the mechanistic target of rapamycin (mTOR), and signal transducer and activator of transcription 1 (STAT1). Studies have shown that these 3 proteins play critical roles in human AML. The level of Thr 308 p-Akt is associated with high-risk cytogenetics and predicts poor overall survival for AML patients.²⁷ In AML, the mTOR signaling pathway is deregulated and activated as a consequence of genetic and cytogenetic abnormalities. The mTOR inhibitors are often used to target aberrant mTOR activation and signaling.^{59,60} The

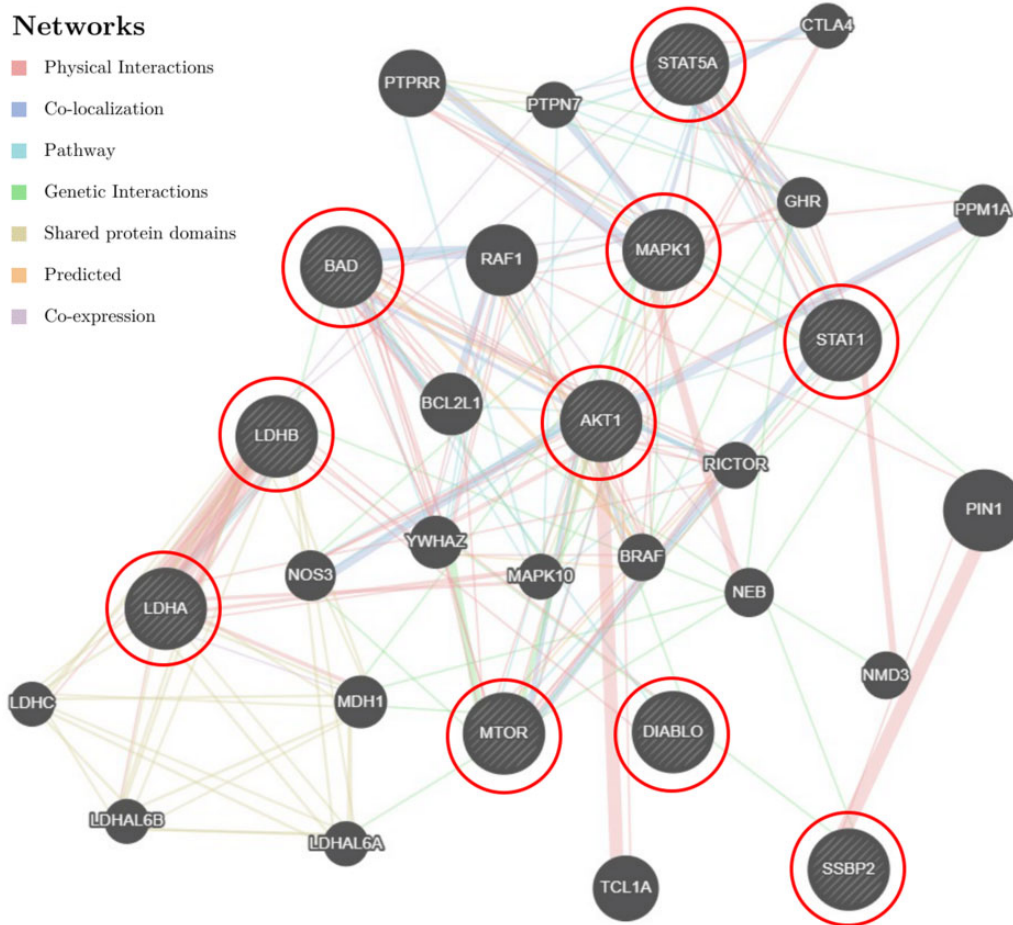


Figure 5. AML data: the gene network analysis. The input 10 genes, namely the genes coding for top 10 proteins that significantly interacted with the treatment, were highlighted using red circles. Other genes that were presumably involved in AML were returned by GeneMANIA.

STAT1 transcription factor is constitutively activated in human AML cell lines and might contribute to the autonomous proliferation of AML blasts. The inhibition of this pathway can be of great interest for AML treatments.^{61,62} Hence, we chose these 3 proteins to build ML models in our proposal.

The whole dataset (256 observations) was randomly shuffled and divided into 2 equal-sized blocks: block 1 and block 2. Each block was taken in turn as either the training set or the testing set. The results were aggregated after 100 repetitions. Since this clinical trial is already completed and it is not possible to get actual treatment responses using our methods, we separated the enrolled patients into 2 groups: a consistent group whose real treatments are the same as the treatments using the ML-based RAR designs and an inconsistent group whose real treatments are different from the treatments using the ML-based RAR designs. We compared the response rates in these 2 groups to elucidate the potential gain if the proposed RAR had been implemented. The results of each method are shown in Figure 6. In the consistent group, the response patient percentages are at least 10% higher than 50%; while the response patient percentages in the inconsistent group are all lower than 50%, that is, we observe higher response rates in the consistent group. This means that patients in the inconsistent group may likely benefit from the RAR method we developed.

DISCUSSION

Patients are accrued in groups sequentially. RAR designs determine the treatment allocation for new groups of patients based on the accrued information of how previous groups of patients responded to their treatments. The number of RAR implementations, k , should be predefined. The choice of k may depend on the total sample size, trial length, and other logistics and practical considerations. Our simulation study used $k = 2$ for a total sample size of 300. In the real data analysis with a smaller sample size of 256 subjects, we used $k = 1$.

We developed novel methods for RAR designs by incorporating 9 ML methods to predict treatment response and assign treatments accordingly. We showed that our ML-based RAR designs can effectively improve treatment response rates among patients. We further proposed an ensemble approach based on the consensus of the 9 ML methods to improve the prediction and decision making. Our proposed ML-ensemble RAR design builds on the predictive ability of 9 ML methods and can further improve prediction accuracy and patient outcome. Specifically, suppose m out of 9 models indicate that treatment A is better than treatment B for patient i , then we let $p_{iA} = m/9$ for $2 \leq m \leq 7$, let $p_{iA} = 0.85$ for $m \geq 7$, and $p_{iA} = 0.15$ for $m \leq 2$. For $m \geq 7$, we keep the assignment probability as a constant of 0.85 because we still want to reserve some randomness in

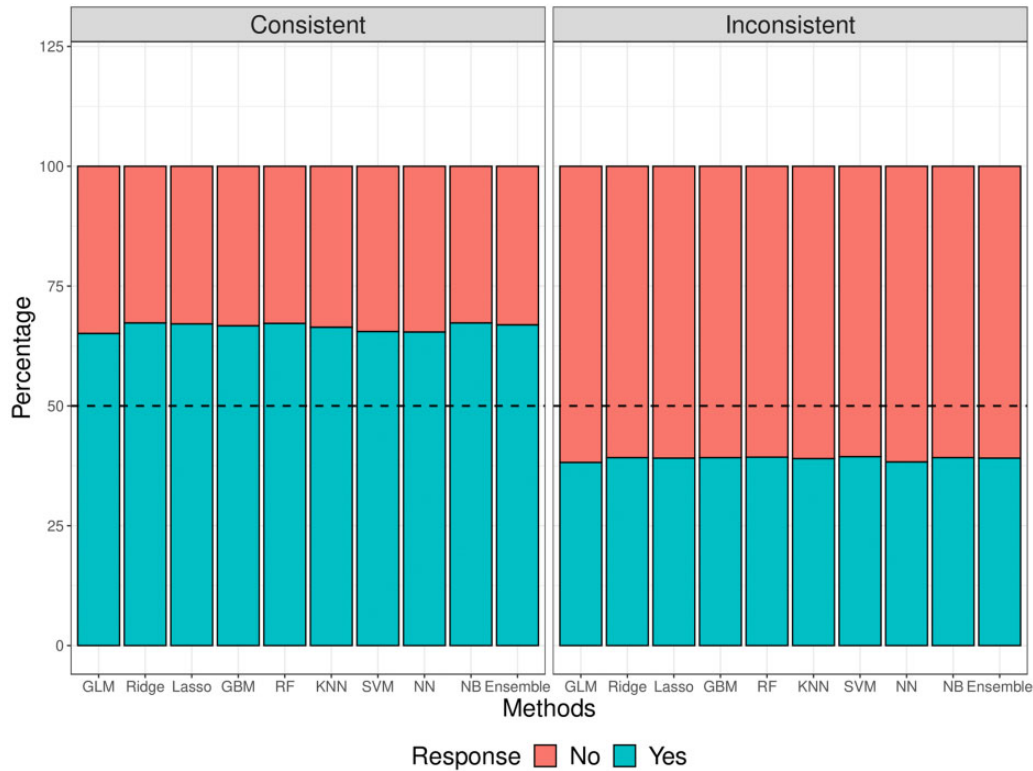


Figure 6. AML data result: the response percentage. Patients in the consistent group (left) were assigned to the same treatments using our ML-based RAR designs, while patients in the inconsistent group (right) were assigned to different treatments using our ML-based RAR designs. The 50% response percentage is marked with a black dashed line.

the trial. These settings can be tuned based on prior knowledge of the treatment selections.

We also tried the combination of NN and GLM algorithms as another binary-combination method and conducted additional simulations. Since these 2 models may not always be in consensus regarding optimal treatment selection for each individual, we took the average of the treatment assigning probabilities of the NN and GLM methods. Similar as we have done previously for the 9 ML methods and the ensemble approach, we evaluated its performance through the overall response rate in simulated trials, the percentage of patients receiving their individually optimal treatment, and the average individual loss for all trial participants. We have provided the results in the [Supplementary Figure S5](#). Although the performance of this combination is slightly better than using either NN or GLM alone, it is still substantially worse than that of the ensemble using 9 ML algorithms, especially when the treatment main effect is high.

While we only considered settings of 2 treatment options in this work, ML-based RAR design can extend to multiple targeted treatments. Given L treatments, the l th treatment allocating probability of patient i is shown as $p_{il} = \pi_{il} / \sum_{l=1}^L \pi_{il}$, $l = 1, \dots, L$, where π_{il}

denotes the response probability of l th treatment for patient i predicted by the ML algorithm. For example, NN can naturally adapt to a multiclass classification problem by replacing the binary cross-entropy loss to a categorical cross-entropy loss.⁶³

Although our work can effectively improve the treatment outcomes in the clinical trial, there are a few limitations that we would like to point out as directions for further research. First, equal weight was given to each of the 9 ML algorithms in the ensemble

method. However, it is likely that different ML methods have distinct prediction accuracy at different scenarios. Incorporating such information by attaching different weights for different ML algorithms in the ensemble method could potentially lead to better adaptation to the data and may provide more precise treatment suggestions for personalized medicine. Second, although our method has been extensively evaluated using simulated and real data, we did not consider the setting with high-dimensional data, for example, the data from omics experiments. With the development of modern sequencing technology, more clinical trials seek to include such information in clinical decision making and trial design. With high-dimensional data, there are more challenges, such as adding appropriate feature selection steps, etc. Moreover, our current model did not consider the situation when complex interactions between treatment and individualized biomarkers exist in the dataset. When this problem is of interest, we might resort to other models that are specifically designed to address the heterogeneous treatment effect caused by these interactions, such as the honest causal forest model,⁶⁴ that are specifically designed to address the heterogeneous treatment effect caused by these interactions.

CONCLUSION

ML methods have successfully demonstrated their superior prediction performance in many applications, but have not been applied to conduct RAR in clinical trials. In this study, we developed novel methods for RAR designs by incorporating ML algorithms to predict treatment response and assign treatments accordingly. We showed that the ML-based RAR designs have better performance than that of the traditional ER design. And the ensemble approach

demonstrated better results than the ER design at the greatest extent. As the ML field is getting mature and abundant packages are available on different programming software, our method is easy to implement in current clinical trial systems.

FUNDING

The research of XH was partially supported by the US National Institutes of Health grants U54CA096300, U01CA253911, and 5P50CA100632, and the Dr. Mien-Chie Hung and Mrs. Kinglan Hung Endowed Professorship.

AUTHOR CONTRIBUTIONS

XH, ZL, and YW conceived the concept of the study and designed the method. YW implemented the method, performed the experiments, and drafted the initial manuscript. BC interpreted the real-world data for the work. All authors edited and approved the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

Simulation data can be reproduced by the R script that has been deposited in the online Dryad repositior.⁶⁵ The AML dataset used for real-world illustration can be downloaded from <https://bioinformatics.mdanderson.org/public-datasets/supplements/> under “RPPA Data in AML”.⁵⁸

REFERENCES

- Antoniou M, Kolamunnage-Dona R, Jorgensen AL. Biomarker-guided non-adaptive trial designs in phase II and Phase III: a methodological review. *JPM* 2017; 7 (1): 1.
- Ginsburg GS, Phillips KA. Precision medicine: from science to value. *Health Aff (Millwood)* 2018; 37 (5): 694–701.
- Maca J, Bhattacharya S, Dragalin V, Gallo P, Krams M. Adaptive seamless phase II/III designs—background, operational aspects, and examples. *Drug Information J* 2006; 40 (4): 463–73.
- Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 1933; 25 (3-4): 285–94.
- Bello GA, Sabo RT. Outcome-adaptive allocation with natural lead-in for three-group trials with binary outcomes. *J Stat Comput Simul* 2016; 86 (12): 2441–9.
- Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine. *Clin Trials* 2008; 5 (3): 181–93.
- Williamson SF, Jacko P, Villar SS, Jaki T. A Bayesian adaptive design for clinical trials in rare diseases. *Comput Stat Data Anal* 2017; 113: 136–53.
- Trippa L, Lee EQ, Wen PY, et al. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *J Clin Oncol* 2012; 30 (26): 3258–63.
- Brannath W, Zuber E, Branson M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Stat Med* 2009; 28 (10): 1445–63.
- Lee JJ, Xuemin G, Suyu L. Bayesian adaptive randomization designs for targeted agent development. *Clin Trials* 2010; 7 (5): 584–96.
- Lee JJ, Chu CT. Bayesian clinical trials in action. *Stat Med* 2012; 31 (25): 2955–72.
- Stallard N, Whitehead J, Cleall S. Decision-making in a phase II clinical trial: a new approach combining Bayesian and frequentist concepts. *Pharm Stat* 2005; 4 (2): 119–28.
- Dong G, Shih WJ, Moore D, Quan H, Marcella S. A Bayesian-frequentist two-stage single-arm phase II clinical trial design. *Stat Med* 2012; 31 (19): 2055–67.
- Kim ES, Herbst RS, Wistuba II, et al. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discov* 2011; 1 (1): 44–53.
- Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA, Esserman LJ. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther* 2009; 86 (1): 97–100.
- Nevozhay D, Adams RM, Murphy KF, Josi K, Balzsi G. Negative autoregulation linearizes the dose-response and suppresses the heterogeneity of gene expression. *Proc Natl Acad Sci U S A* 2009; 106 (13): 5123–8.
- Lee JJ, Liu S, Gu X. Bayesian adaptive randomization designs for targeted agent development. *Clin Cancer Res* 2008; 14 (19 Suppl): PL06.
- Ma F, Gao J, Suo Q, You Q, Zhou J, Zhang A. Risk prediction on electronic health records with prior medical knowledge. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK: Association for Computing Machinery; 2018: 1910–9.
- Fakoor R, Ladhak F, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. In: Proceedings of the International Conference on Machine Learning; 2013: 3937–49.
- Liu S, Liu S, Cai W, Pujol S, Kikinis R, Feng D. Early diagnosis of Alzheimer’s disease with deep learning. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI); 2014: 1015–8.
- Pantuck A, Lee D-K, Kee T, et al. Modulating BET bromodomain inhibitor ZEN-3694 and enzalutamide combination dosing in a metastatic prostate cancer patient using CURATE.AI, an artificial intelligence platform. *Adv Ther* 2018; 1 (6): 1800104.
- Liang Z, Zhang G, Huang X, Hu QV. Deep learning for healthcare decision making with EMRs. *IEEE Int Conf Bioinform Biomed* 2014: 556–9. doi:10.1109/BIBM.2014.6999219.
- Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015; 12 (10): 931–4.
- Tibshirani R, Hastie T, Friedman J. Regularized paths for generalized linear models via coordinate descent. *J Stat Soft* 2010; 33 (1): 1–22.
- Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011; 46 (3): 399–424.
- Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 2010; 38 (Web Server issue): W214–20.
- Gallay N, Santos CD, Cuzin L, et al. The level of AKT phosphorylation on threonine 308 but not on serine 473 is associated with high-risk cytogenetics and predicts poor overall survival in acute myeloid leukemia. *Blood* 2008; 112 (11): 1503.
- Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* 2016; 3 (3): 243–50.
- Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res* 2017; 4 (1): e000234.
- Zazzi M, Incardona F, Rosen-Zvi M, et al. Predicting response to antiretroviral treatment by machine learning: the EuResist project. *Intervirology* 2012; 55 (2): 123–7.
- Patel MJ, Andreescu C, Price JC, Edelman KL, Reynolds CF, Aizenstein HJ. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. *Int J Geriatr Psychiatry* 2015; 30 (10): 1056–67.
- Schmidt-Erfurth U, Bogunovic H, Sadeghipour A, et al. Machine learning to analyze the prognostic value of current imaging biomarkers in neovas-

- cular age-related macular degeneration. *Ophthalmol Retina* 2018; 2 (1): 24–30.
33. Rajkumar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019; 380 (14): 1347–58.
 34. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* 2019; 18 (6): 463–77.
 35. Rosenberger WF, Lachin JM. *Randomization in Clinical Trials: Theory and Practice*. Hoboken, NJ: John Wiley & Sons; 2015.
 36. Robbins H. Some aspects of the sequential design of experiments. *Bull Amer Math Soc* 1952; 58 (5): 527–35.
 37. Combes A, Hajage D, Capellier G; EOLIA Trial Group, REVA, and ECMONet, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome. *N Engl J Med* 2018; 378 (21): 1965–75.
 38. Emslie G, Rush A, Weinberg W, et al. A double-blind, randomized, placebo-controlled trial of fluoxetine in children and adolescents with depression. *Arch Gen Psychiatry* 1997; 54 (11): 1031–7.
 39. Friedman LM, Furberg C, DeMets DL. *Fundamentals of Clinical Trials*. Boston, MA: J. Wright, PSG Inc.; 1981.
 40. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Ser B (Methodol)* 1974; 36 (2): 111–33.
 41. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res* 2012; 13 (10): 281–305.
 42. Kuhn M. caret: Classification and Regression Training; 2015. [Database] <https://github.com/topepo/caret/> Accessed October 9, 2021.
 43. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–30.
 44. Cramer JS. *The Origins of Logistic Regression*. Tinbergen Institute Discussion Papers 02-119/4. Amsterdam, Netherlands: Tinbergen Institute; 2002.
 45. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B (Methodol)* 1996; 58 (1): 267–88.
 46. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; 12 (1): 55–67.
 47. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001; 29 (5): 1189–232.
 48. Ho TK. *Random Decision Forests*. In: Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. Manhattan, NY: IEEE; 1995: 278–82.
 49. Boser BE, Guyon IM, Vapnik VN. *A Training Algorithm for Optimal Margin Classifiers*. In: Proceedings of the fifth annual workshop on Computational learning theory (COLT '92). New York, NY: Association for Computing Machinery (ACM); 1992: 338–45.
 50. John GH, Langley P. *Estimating Continuous Distributions in Bayesian Classifiers*. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (UAI'95). San Francisco, CA: Morgan Kaufmann Publishers; 1995: 338–45.
 51. Altman NS. An introduction to Kernel and nearest-neighbor nonparametric regression. *Am Statist* 1992; 46 (3): 175–85.
 52. Grossberg S. Nonlinear neural networks: principles, mechanisms, and architectures. *Neural Networks* 1988; 1 (1): 17–61.
 53. Imbens G. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat* 2004; 86 (1): 4–29.
 54. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc* 1987; 82 (398): 387–94.
 55. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004; 23 (19): 2937–60.
 56. Melfi VF, Page C. Variability in adaptive designs for estimation of success probabilities. *Project Euclid* 1998; 34: 106–14.
 57. Rosenberger WF, Stallard N, Ivanova A, Harper CN, Ricks ML. Optimal adaptive designs for binary response trials. *Biometrics* 2001; 57 (3): 909–13.
 58. Kornblau SM, Tibes R, Qiu YH, et al. Functional proteomic profiling of AML predicts response and survival. *Blood* 2009; 113 (1): 154–64.
 59. Feng Y, Chen X, Cassady K, et al. The role of mTOR inhibitors in hematologic disease: from bench to bedside. *Front Oncol* 2020; 10: 611690.
 60. Sandhöfer N, Metzeler KH, Rothenberg M, et al. Dual PI3K/mTOR inhibition shows antileukemic activity in MLL-rearranged acute myeloid leukemia. *Leukemia* 2015; 29 (4): 828–38.
 61. Gouilleux-Gruart V, Gouilleux F, Desaint C, et al. STAT-related transcription factors are constitutively activated in peripheral blood cells from acute leukemia patients. *Blood* 1996; 87 (5): 1692–97.
 62. Weber-Nordt RM, Egen C, Wehinger J, et al. Constitutive activation of STAT proteins in primary lymphoid and myeloid leukemia cells and in Epstein-Barr Virus (EBV)-related lymphoma cell lines. *Blood* 1996; 88 (3): 809–16.
 63. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
 64. Athey S, Tibshirani J, Wager S. Generalized Random Forests; 2018. <https://grf-labs.github.io/grf/> Accessed July 14, 2021.
 65. Wang Y, Huang XH, Li Z. Application of machine learning methods in clinical trials for precision medicine. *Dryad* 2021.