Application of machine learning methods in clinical trials for precision medicine

Yizhuo Wang\textsuperscript{1}, Bing Z. Carter\textsuperscript{2}, Ziyi Li\textsuperscript{1,*}, Xuelin Huang\textsuperscript{1,*}

\textsuperscript{1}Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.
\textsuperscript{2}Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.

September 16, 2021

*To whom the correspondence should be addressed to: xilhuang@mdanderson.org and zli16@mdanderson.org.
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 nine 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 nine 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 FDA-approved treatment options and only can get new treatments through clinical trials.
1 INTRODUCTION

It is known that patients respond differently to the same treatments [2]. 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 [18]. 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 non-adaptive 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 [35].

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 (1933), who proposed employing a posterior probability estimated from the interim data to assign patients to the more effective treatment [49]. Following his idea, the application of Bayesian methods with an inherent adaptive nature has boomed in area of RAR designs [5, 60, 57, 58, 7, 30, 28, 48, 11].

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 two biomarker-based, Bayesian RAR clinical trials [26, 4]. 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 [37]. 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 [29, 2].

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 non-linear 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 [34], disease classification [13], imaging diagnosis [32], drug manufacturing [38], medication assignment [31], and genomic feature identification tasks [59].

Although several supervised ML approaches have been applied to drug response prediction [8, 47, 58, 39, 46, 40, 54], little of the work has explored incorporating ML methods into RAR trial designs. In this study, we implemented nine ML algorithms into RAR designs and further presented an ML-ensemble RAR design combining these nine 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.
2 MATERIALS AND METHODS

2.1 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 [43]. 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 Thompson (1933) [50] and Robbins (1952) [41], 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 (ARDS) [9], and the first large-scale double-blind, placebo-controlled study which tested the superiority of fluoxetine over placebo in children and adolescents with depression [12]. A general scheme of the RAR designs is shown in Figure 1.

2.2 Benchmark design: equal randomization

Randomization as a standard means for addressing the selection bias in treatment assignments has been extensively used in clinical trials [43]. 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, Furberg and DeMets (1981, 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 [16]. Here, we used the ER design as a benchmark randomization design to evaluate the performance of ML-based RAR designs.

2.3 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 $\pi_{iA}$ and $\pi_{iB}$ respectively denote the response probability of treatment $A$ and $B$ for patient $i$ predicted by the ML model.

### 2.4 Machine learning algorithms and a machine learning ensemble

We selected nine mainstream ML algorithms and implemented them in the RAR design to predict treatment response. The prediction models were built using the best parameters for each model, which were obtained by a tenfold cross-validation. These selected ML algorithms can be roughly divided into two categories:

1. Parametric models: logistic regression [10], LASSO regression [51], Ridge regression [23].

2. Non-parametric models: gradient boosting machine (GBM) [15], random forest (RF) [22], support vector machine (SVM) [6], naive bayes [25], k-nearest neighbors (KNN) [1], and artificial neural networks (NN) [21].

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 [52]. Ridge regression and LASSO regression assume that there is a linear relationship between the "dependent" variable and the explanatory variables. They are two regularization methods of GLM to prevent an over-fitting issue by adding penalties on the predictor variables that are less significant [23, 51].
KNN, which classifies data points based on the points that are most similar to it, is a typical non-parametric model such that there is no assumption for underlying data distribution, and the number of parameters grows with the size of the data \[1\]. With NNs, however, there has been some debate regarding whether they belong to parametric or non-parametric methods. NNs typically consist of three layers: input layer, hidden layer, and output layer. Here we classify NNs as a non-parametric method, as the network architecture grows adaptively to match the complexity of given data \[21\].

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 \[15, 22\].

NB and SVM can be either parametric or non-parametric depending on whether they use kernel tricks. For the NB classifier, it becomes non-parametric if using a kernel density estimation (KDE) to obtain a more realistic estimate of the probability of an observation belonging to a class \[25\]. And for SVM, the basic idea is finding a hyperplane that best divides a dataset into two classes. It is considered a non-parametric 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 \[6\].

Combing these nine 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, \(\theta\) is the threshold number of agreed models, and \(p_t\) is the threshold treatment allocation probability. Here we chose
\[ \theta = 7 \text{ and } p_t = 0.85; \text{ these threshold values can be adjusted accordingly for different application purposes.} \]

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.

### 2.5 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 [3]. 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 [42]. 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:

\[
\begin{align*}
    w_i &= \frac{Z_i}{e_i} + \frac{(1 - Z_i)}{1 - e_i} \\
\end{align*}
\]

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 [24]. Incorporated with the IPTW idea, the ATE estimator is defined as follows:

\[
\begin{align*}
    \frac{1}{n} \sum_{i=1}^{n} \frac{Z_i Y_i}{e_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - Z_i) Y_i}{1 - e_i} \\
\end{align*}
\]

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 [33].
2.6 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 (2.5). Additionally, we proposed a new criteria, 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 \(\hat{P}_i(Y = 1 \mid T, x)\) denote the probability of responding to the received treatment \(T\). Let \(\hat{P}_i(Y = 1 \mid T = \text{OPT}, x)\) denote the probability of responding to the optimal treatment determined by the true model. Then we define the personalized loss function as follows:

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

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.
3 RESULTS

3.1 Simulation

We used simulation studies to evaluate the proposed methods.

3.1.1 Setting

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

$$\text{logit}(\eta_i) = \alpha T_i + \sum_{j=1}^{10} \beta_j f_j(X_{ji}) + \sum_{j=1}^{10} \gamma_j T_i X_{ji} + \epsilon_i, i = 1, \cdots, 300; j = 1, \cdots, 10$$

where $T_i$ is the treatment indicator (either treatment 0 or treatment 1), $\alpha$ 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. $\gamma_j$ 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 $\epsilon_i$. 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 equal randomization design.

3.1.2 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 equal randomization (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 nine 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 nine 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.

3.1.3 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 individual 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.

3.1.4 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 [36, 44]. 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.
3.2 Real-world example

We analyzed a publicly available acute myeloid leukemia (AML) dataset from Kornblau(2009) where most of the clinical biomarkers are expression levels of cellular proteins [27]. 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 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) [55]. 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 three of them were found to have the most significant interactions with the treatment, i.e., 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 three 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 [17]. 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 [14, 45]. The 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 [20, 56]. Hence, we chose these three proteins to build ML models in our proposal.

The whole dataset (256 observations) was randomly shuffled and divided into two 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 two 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 two 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%, i.e., 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.

4 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 pre-defined. 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 nine 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 nine-ML methods to improve the prediction and decision making. Our proposed ML-ensemble RAR design builds on the predictive ability of nine 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 \( p_{i,A} = \frac{m}{9} \) for \( 2 \leq m \leq 7 \), let \( p_{i,A} = 0.85 \) for \( m \geq 7 \), and \( p_{i,A} = 0.15 \) 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 the trial. These settings can be tuned based on prior knowledge of the treatment selections.

While we only considered settings of two 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 \( p_{il} = \frac{\pi_{il}}{\sum_{l=1}^{L} \pi_{il}} \), where \( \pi_{il} \) denotes the response probability of \( l^{th} \) treatment for patient \( i \) predicted by the ML algorithm. For example, NN can naturally adapt to a multi-class classification problem by replacing the binary cross-entropy loss to a categorical cross-entropy loss [19].

### 5 Conclusion

Machine learning 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.
6 Data availability

The AML dataset used for real-world illustration was downloaded from https://bioinformatics.mdanderson.org/public-datasets/supplements/ under "RPPA Data in AML" [27].

7 Funding

The research of Huang 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.

8 CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

References

[1] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[2] M. Antoniou, R. Kolamunnage-Dona, and A. L. Jorgensen. Biomarker-guided non-adaptive trial designs in phase ii and phase iii: A methodological review. *J Pers Med*, 7(1), 2017.

[3] P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*, 46(3):399–424, 2011.

[4] A. D. Barker, C. C. Sigman, G. J. Kelloff, N. M. Hylton, D. A. Berry, and L. J. Esserman. I-spy 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther*, 86(1):97–100, 2009.

[5] G. A. Bello and R. T. Sabo. Outcome-adaptive allocation with natural lead-in for three-group trials with binary outcomes. *Journal of Statistical Computation and Simulation*, 86(12):2441–2449, 2015.
[6] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. ACM Press, 1992.

[7] W. Brannath, E. Zuber, M. Branson, F. Bretz, P. Gallo, M. Posch, and A. Racine-Poon. Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Stat Med*, 28(10):1445–63, 2009.

[8] A. M. Chekroud, R. J. Zotti, Z. Shehzad, R. Gueorguieva, M. K. Johnson, M. H. Trivedi, T. D. Cannon, J. H. Krystal, and P. R. Corlett. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*, 3(3):243–50, 2016.

[9] A. Combes, D. Hajage, G. Capellier, A. Demoule, S. Lavoué, C. Guervilly, D. Da Silva, L. Zafrani, P. Tirot, B. Veber, E. Maury, B. Levy, Y. Cohen, C. Richard, P. Kalfon, L. Bouadma, H. Mehdaoui, G. Beduneau, G. Lebreton, L. Brochard, N. D. Ferguson, E. Fan, A. S. Slutsky, D. Brodie, and A. Mercat. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome. *New England Journal of Medicine*, 378(21):1965–1975, 2018. PMID: 29791822.

[10] J. Cramer. The Origins of Logistic Regression. Technical report, Tinbergen Institute, Dec. 2002.

[11] G. Dong, W. J. Shih, D. Moore, H. Quan, and S. Marcella. A bayesian-frequentist two-stage single-arm phase ii clinical trial design. *Stat Med*, 31(19):2055–67, 2012.

[12] G. Emslie, A. Rush, W. Weinberg, R. Kowatch, C. Hughes, T. Carmody, and J. Rintelmann. A double-blind, randomized, placebo-controlled trial of fluoxetine in children and adolescents with depression. *Archives of general psychiatry*, 54 11:1031–7, 1997.

[13] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber. Using deep learning to enhance cancer diagnosis and classification. *Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare*, 06 2013.

[14] Y. Feng, X. Chen, K. Cassady, Z. Zou, S. Yang, Z. Wang, and X. Zhang. The role of mtor inhibitors in hematologic disease: From bench to bedside. *Frontiers in Oncology*, 10:3043, 2021.
[15] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

[16] L. M. Friedman, C. Furberg, and D. L. DeMets. *Fundamentals of clinical trials*. J. Wright, PSG Inc., Boston, 1981.

[17] N. Gallay, C. D. Santos, L. Cuzin, M. Bousquet, V. S. Gouy, C. Chaussade, M. Attal, B. Payrastre, C. Demur, and C. Recher. 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*, 112(11):1503–1503, 11 2008.

[18] G. S. Ginsburg and K. A. Phillips. Precision medicine: From science to value. *Health Aff (Millwood)*, 37(5):694–701, 2018.

[19] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[20] V. Gouilleux-Gruart, F. Gouilleux, C. Desaint, J.-F. Claisse, J.-C. Capiod, J. Delobel, R. Weber-Nordt, I. Dusanter-Fourt, F. Dreyfus, B. Groner, and L. Prin. Stat-related transcription factors are constitutively activated in peripheral blood cells from acute leukemia patients. *Blood*, 87(5):1692–1697, 1996.

[21] S. Grossberg. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, 1(1):17–61, 1988.

[22] T. K. Ho. *Random Decision Forests*. ICDAR ’95. IEEE Computer Society, 1995.

[23] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[24] G. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 2004.

[25] G. H. John and P. Langley. *Estimating Continuous Distributions in Bayesian Classifiers*. UAI’95. Morgan Kaufmann Publishers Inc., 1995.
[26] E. S. Kim, R. S. Herbst, I. I. Wistuba, J. J. Lee, G. R. Blumenschein, A. Tsao, D. J. Stewart, M. E. Hicks, J. Erasmus, S. Gupta, C. M. Alden, S. Liu, X. Tang, F. R. Khuri, H. T. Tran, B. E. Johnson, J. V. Heymach, L. Mao, F. Fossella, M. S. Kies, V. Papadimitrakopoulou, S. E. Davis, S. M. Lippman, and W. K. Hong. The battle trial: Personalizing therapy for lung cancer. *Cancer Discovery*, 1(1):44–53, 2011.

[27] S. M. Kornblau, R. Tibes, Y. H. Qiu, W. Chen, H. M. Kantarjian, M. Andreeff, K. R. Coombes, and G. B. Mills. Functional proteomic profiling of AML predicts response and survival. *Blood*, 113(1):154–164, 01 2009.

[28] J. J. Lee and C. T. Chu. Bayesian clinical trials in action. *Stat Med*, 31(25):2955–72, 2012.

[29] J. J. Lee, S. Liu, and X. Gu. Bayesian adaptive randomization designs for targeted agent development. *Clinical Cancer Research*, 14(19 Supplement):PL06–PL06, 2008.

[30] J. J. Lee, G. Xuemin, and L. Suyu. Bayesian adaptive randomization designs for targeted agent development. *Clin Trials*, 7(5):584–96, 2010.

[31] Z. Liang, G. Zhang, X. Huang, and Q. V. Hu. Deep learning for healthcare decision making with emrs. *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 556–559, 2014.

[32] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng. Early diagnosis of alzheimer’s disease with deep learning. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 1015–1018, 2014.

[33] J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

[34] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang. *Risk Prediction on Electronic Health Records with Prior Medical Knowledge*. KDD ’18. Association for Computing Machinery, New York, NY, USA, 2018.
[35] J. Maca, S. Bhattacharya, V. Dragalin, P. Gallo, and M. Krams. Adaptive seamless phase ii/iii designs—background, operational aspects, and examples. *Drug Information Journal*, 40(4):463–473, 2006.

[36] V. F. Melfi and C. Page. Variability in adaptive designs for estimation of success probabilities. *Project Euclid*, 1998.

[37] D. Nevozhay, R. M. Adams, K. F. Murphy, K. Josić, and G. Balázsi. Negative autoregulation linearizes the dose–response and suppresses the heterogeneity of gene expression. *Proceedings of the National Academy of Sciences*, 106(13):5123–5128, 2009.

[38] A. Pantuck, D.-K. Lee, T. Kee, P. Wang, S. Lakhotia, M. Silverman, C. Mathis, A. Drakaki, A. Beldegrun, C.-M. Ho, and D. Ho. Modulating bet bromodomain inhibitor zen-3694 and enzalutamide combination dosing in a metastatic prostate cancer patient using curate.ai, an artificial intelligence platform. *Advanced Therapeutics*, 1, 2018.

[39] M. J. Patel, C. Andreescu, J. C. Price, K. L. Edelman, R. Reynolds, C. F., and H. J. Aizenstein. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. *Int J Geriatr Psychiatry*, 30(10):1056–67, 2015.

[40] A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019. PMID: 30943338.

[41] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, pages 527–535, 1952.

[42] P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

[43] W. Rosenberger and J. Lachin. *Randomization in Clinical Trials: Theory and Practice*. Wiley Series in Probability and Statistics. Wiley, 2015.

[44] W. F. Rosenberger, N. Stallard, A. Ivanova, C. N. Harper, and M. L. Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–13, 2001.
[45] N. Sandhöfer, K. Metzeler, M. Rothenberg, T. Herold, S. Tiedt, V. Groiß, M. Carlet, G. Walter, T. Hinrichsen, O. Wachter, M. Grunert, S. Schneider, M. Subklewe, A. Dufour, S. Fröhling, H.-G. Klein, W. Hiddemann, I. Jeremias, and K. Spiekermann. Dual pi3k/mtor inhibition shows antileukemic activity in mll-rearranged acute myeloid leukemia. *Leukemia*, 29(4):828—838, April 2015.

[46] U. Schmidt-Erfurth, H. Bogunovic, A. Sadeghipour, T. Schlegl, G. Langs, B. S. Gerendas, A. Osborne, and S. M. Waldstein. Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration. *Ophthalmol Retina*, 2(1):24–30, 2018.

[47] D. W. Shimabukuro, C. W. Barton, M. D. Feldman, S. J. Mataraso, and R. Das. 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*, 4(1):e000234, 2017.

[48] N. Stallard, J. Whitehead, and S. Cleall. Decision-making in a phase ii clinical trial: A new approach combining bayesian and frequentist concepts. *Pharmaceutical Statistics*, 4:119 – 128, 04 2005.

[49] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[50] W. R. THOMPSON. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 12 1933.

[51] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[52] R. Tibshirani, T. Hastie, and J. Friedman. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 02 2010.

[53] L. Trippa, E. Q. Lee, P. Y. Wen, T. T. Batchelor, T. Cloughesy, G. Parmigiani, and B. M. Alexander. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *J Clin Oncol*, 30(26):3258–63, 2012.
[54] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*, 18(6):463–477, 2019.

[55] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(suppl2):W214–W220, 06 2010.

[56] R. Weber-Nordt, C. Egen, J. Wehinger, W. Ludwig, V. Gouilleux-Gruart, R. Mertelsmann, and J. Finke. Constitutive activation of stat proteins in primary lymphoid and myeloid leukemia cells and in epstein-barr virus (ebv)–related lymphoma cell lines. *Blood*, 88(3):809–816, 1996.

[57] S. F. Williamson, P. Jacko, S. S. Villar, and T. Jaki. A bayesian adaptive design for clinical trials in rare diseases. *Comput Stat Data Anal*, 113:136–153, 2017.

[58] M. Zazzi, F. Incardona, M. Rosen-Zvi, M. Prosperi, T. Lengauer, A. Altmann, A. Sonnerborg, T. Lavee, E. Schulter, and R. Kaiser. Predicting response to antiretroviral treatment by machine learning: the euresist project. *Intervirology*, 55(2):123–7, 2012.

[59] J. Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*, 12:931–4, 2015 Oct 2015.

[60] X. Zhou, S. Liu, E. S. Kim, R. S. Herbst, and J. J. Lee. Bayesian adaptive design for targeted therapy development in lung cancer–a step toward personalized medicine. *Clin Trials*, 5(3):181–93, 2008.
Figure 1: Response-adaptive randomization (RAR) design. The number of adaptive randomization is adjustable per application.
Figure 2: Simulation result: response rate (top), percentage of patients receiving their optimal treatments (bottom). The treatment-biomarker interaction, $\gamma$ is fixed at 0.5. Boxplots display the median (middle line), the inter-quartile range (hinges), and 1.5 times the inter-quartile 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 (top), average treatment effect (ATE, bottom). The treatment-biomarker interaction, $\gamma$ is fixed at 0.5. Boxplots display the median (middle line), the inter-quartile range (hinges), and 1.5 times the inter-quartile 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, $\gamma$ is fixed at 0.5. The Type I error is controlled at 0.05. The power ranges from 0.04 to 0.97.
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.
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.