Multi-Task Adversarial Learning for Treatment Effect Estimation in Basket Trials

Zhixuan Chu CHUZHIXUAN.CZX@ALIBABA-INC.COM
Ant Group, China

Stephen L. Rathbun RATHBUN@UGA.EDU
University of Georgia, USA

Sheng Li SHENG.LI@UGA.EDU
University of Georgia, USA

Abstract

Estimating treatment effects from observational data provides insights about causality guiding many real-world applications such as different clinical study designs, which are the formulations of trials, experiments, and observational studies in medical, clinical, and other types of research. In this paper, we describe causal inference for application in a novel clinical design called basket trial that tests how well a new drug works in patients who have different types of cancer that all have the same mutation. We propose a multi-task adversarial learning (MTAL) method, which incorporates feature selection multi-task representation learning and adversarial learning to estimate potential outcomes across different tumor types for patients sharing the same genetic mutation but having different tumor types. In our paper, the basket trial is employed as an intuitive example to present this new causal inference setting. This new causal inference setting includes, but is not limited to basket trials. This setting has the same challenges as the traditional causal inference problem, i.e., missing counterfactual outcomes under different subgroups and treatment selection bias due to confounders. We present the practical advantages of our MTAL method for the analysis of synthetic basket trial data and evaluate the proposed estimator on two benchmarks, IHDP and News. The results demonstrate the superiority of our MTAL method over the competing state-of-the-art methods.

Data and Code Availability This paper uses benchmarks *IHDP* (Brooks-Gunn et al., 1992) and *News* (Schwab et al., 2018), which are available on the repositories [1] [2]. We also use one synthetic basket trial dataset and the detailed simulation procedure is provided in the Section 4.2.

1. Introduction

With the rapid development of next-generation sequencing and comprehensive genomic profiling, genomic characterization informs the treatment of a variety of cancers. Some genetic mutations have been linked to multiple cancer types; for example, BRCA1 and BRCA2 are associated with an increased risk of breast, ovarian and pancreatic cancers (Mersch et al., 2015). Traditional clinical trials focusing on patients with a single cancer are time-consuming and expensive, and frequently fail, so they are not sufficient for the development of genomic technologies. Patients are generally classified by their primary cancer and randomized controlled trials are conducted to create standard therapies for each cancer type. It is unrealistic to conduct separate clinical trials for each sub-population based on molecular subtypes or detailed classification of tumors (Hirakawa et al., 2018). Therefore, a new-style clinical trial protocol is in urgent demand in oncology.

A novel clinical design called basket trial has been developed based on the presence of a specific genomic mutation, irrespective of histology (Astsaturov, 2017; Simon, 2017; Tao et al., 2018). Unlike traditional clinical trials which test a drug against a specific cancer, the core organizing principle of basket trials is a common genomic mutation. A basket trial is usually a non-randomized, single-arm trial so

1. https://github.com/vdorie/npci
2. https://archive.ics.uci.edu/ml/datasets/bag+of+words

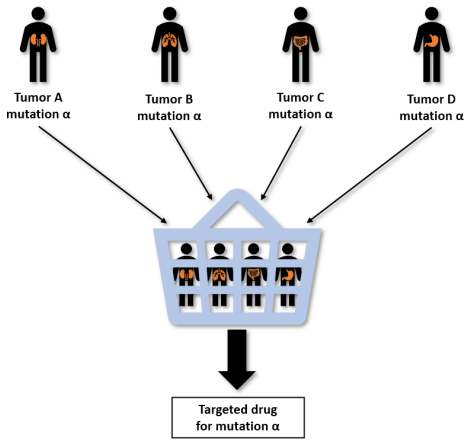

Figure 1: A basket trial is usually a non-randomized and single-arm trial so that all patients with the specified genomic mutation receive the same treatment, regardless of tumor types .

that all patients with the specified genomic mutation receive the same treatment. Treatment selection only depends on genomic mutation type, instead of tumor type. An example of a basket trial is shown in Fig. 1, where the term "basket" arises from each collection of patients sharing a particular mutation and substudies for the same drug are conducted by tumor groups within the whole "basket". Patients enrolled in a basket trial are heterogeneous with respect to tumor type, histologic type, and patient characteristics, so the treatment effects are sensitive to population heterogeneity. Therefore, the absence of a control group becomes a limitation of treatment effect evaluation (Hirakawa et al., 2018). Ignoring the heterogeneity of tumors may lead to failure to detect treatment effects and the inability to produce scientifically reliable findings (Strzebonska and Waligora, 2019). Besides, focusing only on molecular therapy targeting a single mutation without considering the complexity of tumor biology may introduce bias.

In this paper, we apply causal inference models to basket trials. Estimating causal effects from observational data has become an appealing research direction owing to the availability of data and low budget requirements compared with randomized controlled trials. This paper is the first to apply machine learning and causal inference to basket trials and explore the relationship between the traditional multiple treatments design and the basket trial design. In particular, we propose a multi-task adversarial learning (MTAL) method incorporating feature selection, multi-task representation learning, and adversarial learning to remove selection bias (tumor type heterogeneity) introduced by confounders. Our method generates all potential outcomes for each unit across all tumor types, regardless of heterogeneity from different tumor types, so that the sample size may be increased in basket trials for rare tumor types, increasing statistical power. We also define targeted group treatment effects to better describe treatment effects among sub-groups in a basket trial. We present the practicality and advantages of our MTAL method for synthetic basket trials, evaluate the proposed estimator on the IHDP and News datasets, and demonstrate its superiority over state-of-the-art methods.

2. Related Work

The landscape for oncology clinical trials is changing dramatically due to the advent of genomic characterization. Among diverse master protocols (Park et al., 2019), a basket trial evaluates the treatment effect of targeted therapy on patients with the same genomic mutation, regardless of tumor types. Bayesian hierarchical modeling has been proposed to adaptively borrow strength across cancer types to improve the statistical power of basket trials (Berry et al., 2013; Simon, 2017). To avoid inflated type I errors in Bayesian hierarchical modeling, calibrated Bayesian hierarchical modeling has been proposed to evaluate the treatment effect in basket trials (Chu and Yuan, 2018). As an alternative to Bayesian hierarchical modeling, we will apply powerful machine learning tools to basket trials.

Embracing the rapid developments in machine learning, various treatment effect estimation methods for observational data have been proposed for causal inference(Cui et al., 2020; Li et al., 2016a; Yao et al., 2021). Balancing neural networks (BNNs) (Johansson et al., 2016) and counterfactual regression networks (CFRNET) (Shalit et al., 2017) have been proposed to balance covariate distributions across treatment and control groups by regarding the problem of counterfactual inference as a domain adaptation problem. These models may be extended to any number of treatments even with continuous parameters, as described in the perfect match (PM) approach (Schwab et al., 2018) and DRNets (Schwab et al., 2019). Li and Fu (Li and Fu, 2017) regard

counterfactual prediction as a classification problem and conduct matching based on balanced and non-linear representations. (Chu et al., 2022) utilize the mutual information to learn the infomax and domain-independent representations to solve the selection bias between treatment and control groups.

3. The Proposed Framework

3.1. Problem Statement

Clarification on New Problem Setup. In traditional causal inference for observational data, researchers consider binary or multiple treatments for a set of experimental units. For example, a person who has cancer may be offered a choice between two treatment therapies. We can observe the outcome of the chosen treatment but not the potential outcome of the treatment not selected. It is impossible to observe the potential outcomes of both therapies; one of the potential outcomes is always missing. The potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990) aims to estimate unobserved potential outcomes and then calculate the treatment effect. The basket trial tests how well a new drug works in patients who have different types of cancer with the same mutation. Patients with the same genetic mutations are put in one "basket" and are divided into different subgroups according to their cancer types. The differences in study design for potential outcome framework and basket trials are illustrated in Fig. 2. For the potential outcome framework, there is one population and several treatments, but in basket trials, there are several sub-populations and only one treatment.

Clarification on the Challenges. In the potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990), we mainly face two challenges: *missing unobserved counterfactual outcomes* for each patient under alternative treatments not received and *treatment selection bias*. In basket trials, we have the similar challenges: missing unobserved counterfactual outcomes for each patient under alternative cancers not contracted and cancer selection bias where the distributions of predictors differ among cancer types. In traditional causal inference for observational data, confounders are variables that affect both the treatment assignment and the outcome. Similarly, in basket trials, there still exist confounders that are associated with both cancer type and treatment outcome. These variables can explain why some

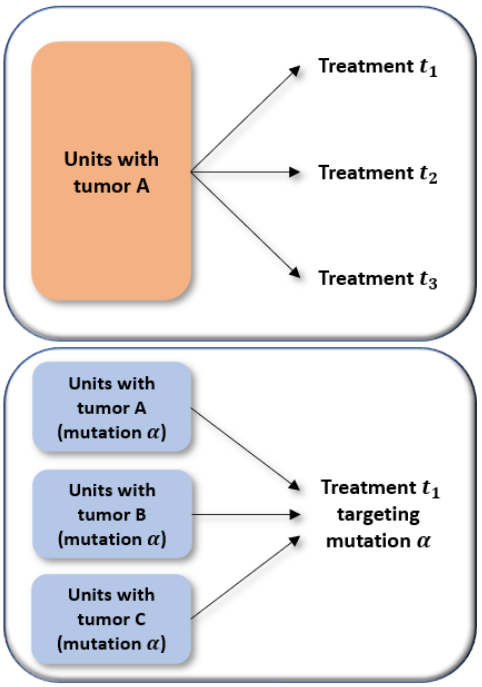

Figure 2: The relationship between conventional multiple treatment causal inference (top) and basket trial (bottom).

patients with the same mutation have different types of cancer and can also influence treatment outcomes. Due to the confounders, it is difficult in a basket trial to estimate the true treatment effects of a drug targeting the mutation of interest and the true treatment effects of a drug for a specific type of cancer. If a significant treatment effect is not found, analysis of basket trials without appropriate causal inference cannot determine that the failure is due to the uselessness of the drug, the particularity of a cancer type, or individual characteristics of patients.

Clarification on Treatment Effects Estimation. Because in this new setting, there is no control group, we do not care about the traditional treatment effects estimation between treated and control groups, e.g., average treatment effect (ATE) or individual treatment effect (ITE). We only focus on the counterfactual outcome inference problem, which is the core problem regardless of the new setting or traditional treatment effects estimation setting. Most basket trials are conducted as single-arm trials without a control group and a primary endpoint is given

by an objective response rate (ORP). We proposed a new metric named targeted group response rate (TGOR) to better describe treatment effects in basket trials. TGOR describes the overall objective response rate for a given mutation or a given tumor type. It can evaluate the treatment effect of the drug for the whole targeted population with the same mutation and the effect of the drug in the subpopulation with different specific tumors type. Our MTAL method can help remove heterogeneity across tumor types when estimating the treatment effect for targeted mutation and remove heterogeneity across patients with the same tumor when estimating the treatment effect for a targeted tumor.

Problem Setup. Suppose a basket trial is conducted as a one-arm phase II trial that tests how well a new drug works in patients who have different types of tumours but share the same genetic mutation. Data are available for n participants. Let $t_i \in \{1, ..., k\}$ denote the type of tumour for unit i; $i = 1, ..., n$. The primary endpoint is the objective response rate (ORR) (Food et al., 2007), determined by tumor assessments from radiological tests or physical examinations. Let y_t^i denote the potential outcome of the unit i $(i = 1, ..., n)$ with the tumour $t \in \{1, ..., k\}$. The observed outcome, called factual outcome is denoted by y^f and the remaining unobserved potential outcomes are called counterfactual outcomes denoted by y^{cf}. The estimated potential, factual, and counterfactual outcomes are $\hat{y}$, $\hat{y}^f$, and $\hat{y}^{cf}$, respectively. Let $X \in \mathbb{R}^d$ denote all observed covariates. We extend the potential outcome framework (Rubin, 1974) to analysis of basket trial data. The following assumptions ensure that the treatment effects can be identified: **Consistency**: The potential outcome of treatment t is equal to the observed outcome if the actual treatment received is t. **Positivity**: For any value of X, treatment assignment is not deterministic, i.e.,$P(T = t|X = x) > 0$, for all t and x. **Ignorability**: Given covariates X, treatment assignment t is independent to the potential outcomes, i.e., $(y_1, y_0) \perp\!\!\!\perp t|X$.

3.2. Model Architecture

We propose a multi-task adversarial learning (MTAL) method to analyze basket trial data or observational data in basket trials, which can remove heterogeneity across tumor types when estimating the treatment effects for a targeted mutation, remove heterogeneity among patients with one type of tumor when estimating the treatment effect for the targeted tumor, and estimating the personalized treatment effect for individual patients. Our method is also useful for studying rare cancers and cancers with rare genetic mutations by inferring the outcome of existing patients with counterfactual cancers to increase sample size and statistical power.

Our method contains two major components: outcome generator and true or false discriminator (TF discriminator), as shown in Fig. 3. In the outcome generator, we use feature selection multi-task deep learning to estimate the potential outcomes for units across all tumor types. Because different types of the tumor may have different predictor variables, which may be components of all observed covariates, a deep feature selection model including (a) a sparse one-to-one layer between the input and the first hidden layer, and (b) an elastic net regularization term throughout the fully-connected representation layers is an essential foundation for potential outcome estimation. Our TF discriminator can tell whether the outcome given the covariates and tumor type is a factual outcome. In the beginning, the TF discriminator can easily find out which outcome is a factual outcome and which one is our inferred counterfactual outcome under alternative tumor types not contracted by those patients. The outcome generator attempts to generate counterfactual outcomes in such a way that the TF discriminator cannot easily determine which is the factual outcome. These two models are trained together in a zero-sum game and they are adversarial until the TF discriminator model is fooled by the generator. At this time, we have removed the tumor type selection bias and obtained all potential outcomes for each patient across all kinds of tumors.

3.2.1. Outcome Generator

Our goal is to correctly predict potential outcomes for each patient across all tumor types by a function $g : x \times t \to y$, which is parameterized by a feed-forward deep neural network structured by multiple hidden layers with non-linear activation functions. Deep neural networks can often dramatically increase prediction accuracy, describe complex relationships, and generate the structured high-level representation of features. The function $g : x \times t \to y$ uses features x and tumor type t as inputs to predict potential outcomes. The output of g estimates potential outcomes across k tumors including single factual outcome $\hat{y}^f$ and $k - 1$ counterfactual outcomes $\hat{y}^{cf}$. The factual

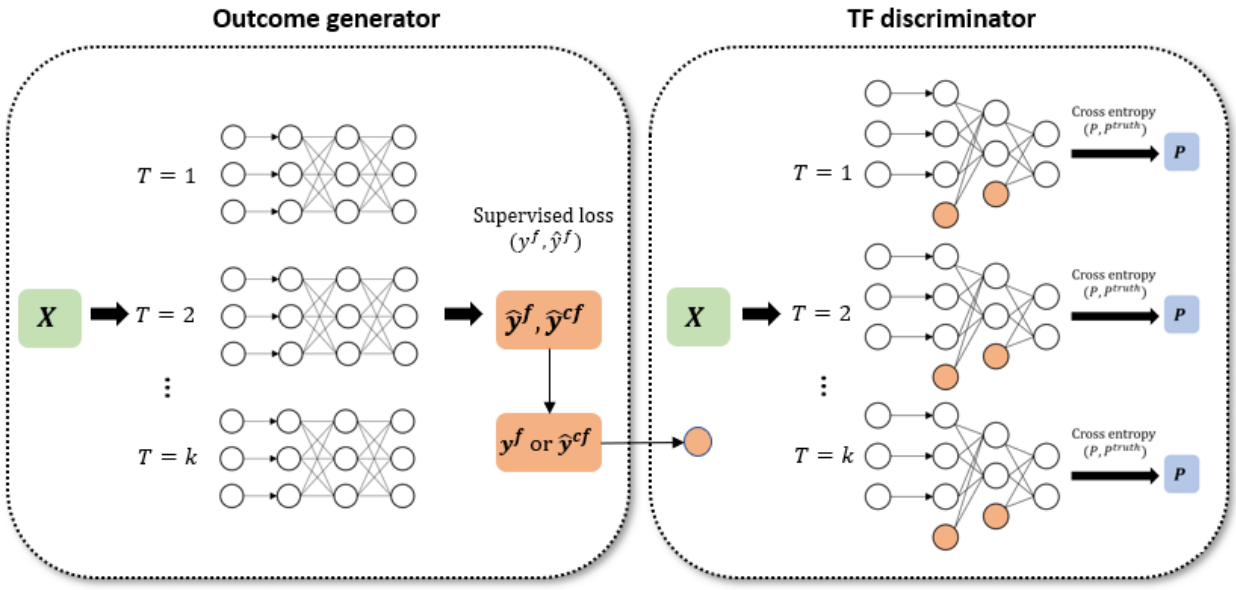

Figure 3: The framework of our multi-task adversarial learning net (MTAL). Our method contains two major components: outcome generator and true or false discriminator (TF discriminator). In the outcome generator, feature selection multi-task deep learning is utilized to estimate the potential outcomes for units across all tumor types. A deep feature selection model includes (a) a sparse one-to-one layer between the input and the first hidden layer, and (b) an elastic net regularization term throughout the fully-connected representation layers. Our TF discriminator is adopted to tell whether the outcome given the covariates and tumor type is a factual outcome. The two models are trained together in a zero-sum game and they are adversarial until the TF discriminator model is fooled by the generator.

outcomes y^f are used to minimize the loss of prediction $\hat{y}^f$.

The function $g(x,t)$ maps the features and tumor type to the corresponding potential outcomes. However, when the dimension of the observed variables is high, there is a risk of losing the influence of t on $g(x,t)$ if the concatenation of x and t is treated as input (Shalit et al., 2017). To address this problem, $g(x,t)$ is partitioned into multiple head nets $g_t(x); t = \{1, ..., k\}$ corresponding to each cancer type. For each cancer type, there is one independent head net for predicting the potential outcome under this tumor. Each unit is used to update only the head net corresponding to the observed tumor type. We aim to minimize the mean squared error in predicting factual outcomes by $g(x,t)$, i.e., $\mathcal{L}_Y = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$, where $\hat{y}_i = g(x_i, t_i)$ denotes the inferred observed outcome of unit i corresponding to factual treatment t_i.

Due to the peculiarities of different tumor types, only a subset of all observed covariates might be predictors for each tumor type. To accommodate this, we add a deep feature selection net (Li et al., 2016b; Chu et al., 2020) to each head net $g_t(x), t = \{1, 2, ..., k\}$, which enables variable selection in deep neural networks. This model takes advantage of deep structures to capture non-linearity and conveniently selects a subset of features of the data at the input level. Feature selection at the input level can help select which variables are input into the neural network and used for representing pre-treatment variables, which makes the deep neutral network more interpretable.

In the feature selection layer, every input variable only connects to its corresponding node where the input variable is weighted. We use a 1-1 layer instead of a fully connected layer. To select input features, weights w in the feature selection layer and the fol-

lowing representation layers have to be sparse and only the features with nonzero weights are selected to enter the following layers. For the observational data with high dimensional variables, LASSO (Tibshirani, 1996) cannot remove enough variables before it saturates. To overcome this limitation, the elastic net (Zou and Hastie, 2005) is adopted in our model, which adds a quadratic term $\|w\|_2^2$ to the penalty i.e., $\Re(w) = \lambda\|w\|_2^2 + \alpha\|w\|_1$, where λ and α are tuning parameters. After combining the mean squared error and the elastic net regularization term, we minimize the objective function in the outcome generator:

$$\begin{aligned}\mathcal{L}_g = &\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \\ &+ \lambda\sum_{t=1}^{k}\sum_{s=1}^{S_t}\|w^{(s)}\|_2^2 \\ &+ \alpha\sum_{t=1}^{k}\sum_{s=1}^{S_t}\|w^{(s)}\|_1,\end{aligned} \tag{1}$$

where S_t is the number of deep feature selection layers for the t-th head net including the feature selection layer and the representation layers. The $w^{(s)}$ are the parameters of deep neural network in the s-th layer of outcome generator. The $\lambda \geq 0$ and $\alpha \geq 0$ are hyperparameters that not only control the trade-off between the regularization term and the following objective terms, but also controls the trade-off between smoothness and sparsity of the weights in the feature selection layer (Li et al., 2016b).

3.2.2. True or False Discriminator

The true or false (TF) discriminator (Chu et al., 2021) is intended to remove tumor type bias and thus improve the prediction accuracy of potential outcomes inferred in the outcome generator for each unit across all types of tumors by adversarial learning. We define one TF discriminator as $\phi : x \times t \times (y^f \text{or } \hat{y}^{cf}) \rightarrow P$, where P is the TF discriminator's judgement, i.e., probability that this outcome for unit i given x and t is factual outcome, which is defined as:

$$P = \begin{cases} P(\text{TF judges } y^f \text{as factual outcome}|x,t) \\ \qquad \text{if } t \text{ is factual type,} \\ P(\text{TF judges } \hat{y}^{cf} \text{as factual outcome}|x,t) \\ \qquad \text{if } t \text{ is not factual type.} \end{cases}$$

Similar to the outcome generator, we use multiple head nets for different tumor types in the TF discriminator. In each head net, a deep feature selection net is added to select the appropriate predictors for each type of tumor. To improve the influence of $(y^f, \hat{y}^{cf})$ in the TF discriminator, we add $(y^f \text{or } \hat{y}^{cf})$ into each layer after one on one feature selection layer and the dimension of each layer in TF discriminator decreases layer by layer.

The TF discriminator is a binary classification task, which puts one label (i.e., true or false factual outcome) on the vector concatenating the representation vector of x and potential outcome $(y^f \text{or } \hat{y}^{cf})$ under each type of tumor head net, so the loss of discrimination is measured by the cross-entropy with truth probability where $P^{\text{truth}} = 1$ if y^f is input and $P^{\text{truth}} = 0$ if $\hat{y}^{cf}$ is input. In each iteration of training, we make sure to input the same number of units in each tumor type to ensure that there exist factual units in each head net. When there are several types of tumors, we face the imbalanced classification issue. If there are k types of tumor and n units in each tumor type are input into the model training procedure, then in each head net, n units are factual outcomes and $(k-1)n$ units are inferred as counterfactual outcomes. As k increases, the imbalance of factual outcome numbers and inferred counterfactual outcome numbers in each head net will aggravate. Because inputs of TF discriminator are generated by the outcome generator $g(x,t)$, the weighted cross entropy of TF discriminator and elastic net are defined as:

$$\begin{aligned}\mathcal{L}_{\phi,g} = &-\frac{1}{n\times k}\sum_{t=1}^{k}\sum_{i=1}^{n}(w_0 p_{ti}^{\text{truth}}\log(p_{ti}) \\ &+ w_1(1-p_{ti}^{\text{truth}})\log(1-p_{ti})) \\ &+ \lambda\sum_{t=1}^{k}\sum_{r=1}^{r_t}\|w^{(r)}\|_2^2 + \alpha\sum_{t=1}^{k}\sum_{r=1}^{r_t}\|w^{(r)}\|_1,\end{aligned} \tag{2}$$

where p_{ti}^{truth} is the probability that this input outcome for unit i under tumor t is the observed factual outcome or inferred outcome from generator module, i.e., 1 or 0, separately. P_{ti} is the probability judged by TF discriminator that this input outcome for unit i under tumor t is factual outcome. The w_1 and w_0 are the proportions of factual outcomes and counterfactual outcomes in total outcomes. Because during training, the same number of units in each tumor type are input, $w_1 = \frac{1}{k}$ and $w_0 = \frac{k-1}{k}$ in each head net. The number of deep feature selection layers for

the t-th head net is denoted by r_t, and $w^{(r)}$ are the weights for the deep neural network in the r-th layer of the TF discriminator. $\lambda \geq 0$ and $\alpha \geq 0$ are the same as those in the outcome generator.

3.2.3. Adversarial Learning

We have described an outcome generator to estimate potential outcomes for each unit across all types of tumor and a TF discriminator to determine if each potential outcome given unit's features under different tumor types is factual. In the initial iterations of the training algorithm, the outcome generator may generate potential outcomes that are very different from factual outcomes as determined by the TF discriminator. As the model is trained further, the TF discriminator may no longer be able to discriminate between the generated potential outcome and the factual outcome. At this point, we have attained all potential outcomes for each unit under all tumor types. The training procedure optimizing the outcome generator and TF discriminator uses the minimax decision rule:

$$\min_g \max_\phi \ (\mathcal{L}_g - \beta \mathcal{L}_{\phi,g}), \tag{3}$$

where β is a hyper-parameter controlling the trade-off between the outcome generator and discriminator. Compared to the deep regression task in the outcome generator, the TF discriminator is a relatively simple binary classification, which is easier to optimize. In every optimization iteration, in order to get more accurate inferred potential outcomes to fool the discriminator based on the discriminator's current ability of telling which is factual outcome and which is counterfactual outcome, we can optimize $\min_g \ (\mathcal{L}_g - \beta \mathcal{L}_{\phi,g})$ several times after we optimize $\max_\phi \ (-\beta \mathcal{L}_{\phi,g})$ one time.

3.3. Targeted Group Analysis

The proposed MTAL method can generate all potential outcomes for each unit across all tumor types, which can help basket trials increase sample size and thus increase statistical power, and remove the influence of heterogeneity among different tumor types.

In basket trials, we must consider different configurations of effectiveness. For example, the drug may truly work for only one type of tumor due to the heterogeneity of tumors. Alternatively, it may actually work for all types of tumors, which means it works for the mutation regardless of the tumor types. Each of these configurations can lead to markedly different statistical properties (Cunanan et al., 2017). Therefore, we not only want to evaluate the treatment effect of the drug for the mutation (the whole population in the study) but also want to evaluate the effect of the drug for specific tumors (the sup-population in the study). In addition, most basket trials are conducted as single-arm trials without a control group and a primary endpoint is given by an objective response rate (ORP). We propose a new metric named targeted group response rate (TGOR) to better describe treatment effects in basket trials. TGOR describes the overall objective response rate for a given mutation or a given tumor type, which is defined as:

$$\text{TGOR}_{\text{mu}} = \frac{1}{n \times k} \sum_{t=1}^{k} \sum_{i=1}^{n} y^{ti}$$

and

$$\text{TGOR}_{\text{tu}} = \frac{1}{n_c} \sum_{i=1}^{n_c} y^{i},$$

where n is the number of patients with that mutation and n_c is the sub-sample who have that mutation and that specific cancer i.e., a subset of mutation sample n.

Our MTAL method can help remove heterogeneity across tumor types in basket trials when estimating the treatment effect for targeted mutation TGOR_{mu}, remove heterogeneity across patients with the same type of tumor when estimating the treatment effect for a targeted tumor TGOR_{tu}, and estimate the individualized treatment effects for an individual patient. Our method is also useful for studying rare cancers and cancers with rare genetic mutations by borrowing strength from more common cancers sharing the same mutation to infer the potential outcomes of existing patients under counterfactual cancer to increase sample size and statistical power.

4. Experiments and Analysis

Because our method is the first model for estimating treatment effects for basket trials, no other baseline methods are available. To evaluate our model's estimation performance, we modify our model (by removing the deep feature selection module) to coordinate the settings in traditional treatment effect estimation (binary and multiple treatments) and use benchmarks (*IHDP* and *News*) to demonstrate our

estimation performance on the counterfactual outcomes. We also use one synthetic basket trial dataset to demonstrate the method's stability in basket trials.

4.1. Performance Evaluation on Estimating the Counterfactual Outcomes

We coordinate our model to be compatible with the settings in traditional treatment effect estimation and conduct experiments on binary treatment benchmark *IHDP* and multiple treatment benchmark *News* with $2, 4, 8,$ and 16 treatment options.

Datasets. *IHDP.* The IHDP data set is a commonly adopted benchmark collected by the Infant Health and Development Program (Brooks-Gunn et al., 1992). These data are generated based on a randomized controlled trial where intensive high-quality care and specialist home visits were provided to low-birth-weight and premature infants. There are a total of 25 pre-treatment covariates and 747 units, including 608 control units and 139 treatment units. The outcome is the infants' cognitive test scores which can be simulated using the pre-treatment covariates and the treatment assignment information through the NPCI package [3]. In the IHDP data set, a biased subset of the treatment group is removed to simulate the selection bias (Shalit et al., 2017). We repeat these procedures 1000 times so as to conduct evaluations of the uncertainty of estimates. *News.* The News data set was first introduced for binary treatments counterfactual estimation by (Johansson et al., 2016) and extended to multiple treatment benchmarks by (Schwab et al., 2018). The News benchmark includes 5000 randomly sampled news articles from the NY Times corpus and the opinions of a media consumer exposed to multiple news items. Each unit is a news item and the features are word counts. The outcome represents the reader's opinion of the news item. The treatment options are various devices used for viewing news items, e.g. smartphones, tablets, desktops, televisions, or others. We use the extended multiple treatment data set according to the specification in (Schwab et al., 2018). A topic model is trained on the whole NY Times corpus to model consumers' preferences towards reading given news items on specific devices, where $k+1$ centroids are randomly picked in the topic space where k represents the number of treatment options and the remaining is the control group. We use four different variants of this data set with 5000 units, 2870 features and $k = 2, 4, 8,$ and 16 treatment options.

Table 1: Hyperparameters and ranges.

	IHDP	News
β	0, $\{10^k\}_{k=-6}^{2}$	0, $\{10^k\}_{k=-6}^{2}$
λ, α	0, $\{10^k\}_{k=-6}^{-1}$	0, $\{10^k\}_{k=-6}^{-1}$
No. layers	2, 3, 4, 5	2, 3, 4, 5
Dim. layer	50, 100, 150	50, 100, 150
Batch size	50×2, 75×2, 100×2	$30k$, $40k$, $50k$

Baselines. To evaluate the accuracy of our model's treatment effect estimation, we compare our multi-task adversarial learning net model with the following methods: k-nearest neighbor (kNN) (Ho et al., 2007), Causal forests (CF) (Wager and Athey, 2018),Random forests (RF) (Breiman, 2001), Bayesian additive regression trees (BART) (Chipman et al., 2010), Generative adversarial nets for inference of ITE (GANITE) (Yoon et al., 2018), Propensity score matching with logistic regression (PSM) (Ho et al., 2011), Treatment-agnostic representation network (TARNET) (Shalit et al., 2017), Counterfactual regression network ($\text{CFRNET}_{\text{wass}}$) (Shalit et al., 2017), local similarity preserved individual treatment effect estimation method (SITE) (Yao et al., 2018), and Perfect match (PM) (Schwab et al., 2018).

Parameter Settings. To ensure a fair comparison, we follow a standardized implementation [4] to realize hyperparameter optimization for IHDP and News data sets and extend the original binary treatment effect estimation methods to multiple treatments according to specifications in (Schwab et al., 2018). The hyperparameters of our method are chosen based on performance on the validation data set, and the searching range as shown in Table 1. MTAL is implemented using feed-forward neural networks with Dropout (Srivastava et al., 2014) and the ReLU activation function. Adam (Kingma and Ba, 2014) is adopted to optimize the objective function.

Results and Analysis. We adopt two commonly used evaluation metrics. The first one is the error in average treatment effect (ATE) estimation defined as $\epsilon_{\text{ATE}} = |\text{ATE} - \widehat{\text{ATE}}|$, where $\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}(Y_1^i - Y_0^i)$ and $\widehat{\text{ATE}}$ is an estimated ATE. The second one is the error of expected precision in estimation of heterogeneous effect (PEHE) (Hill, 2011), which is

3. https://github.com/vdorie/npci

4. https://github.com/d909b/perfect_match

Table 2: Performance on IHDP and News data sets. We present mean ± standard deviation of $\sqrt{\epsilon_{\text{PEHE}}}$ and $\sqrt{\epsilon_{\text{mPEHE}}}$ on the test sets.

	IHDP	News-2	News-4	News-8	News-16
Method	$\sqrt{\epsilon_{\text{PEHE}}}$	$\sqrt{\epsilon_{\text{PEHE}}}$	$\sqrt{\epsilon_{\text{mPEHE}}}$	$\sqrt{\epsilon_{\text{mPEHE}}}$	$\sqrt{\epsilon_{\text{mPEHE}}}$
kNN	6.66 ± 6.89	18.14 ± 1.64	27.92 ± 2.44	26.20 ± 2.18	27.64 ± 2.40
PSM	2.70 ± 3.85	17.40 ± 1.30	37.26 ± 2.28	30.50 ± 1.70	28.17 ± 2.02
RF	4.54 ± 7.09	17.39 ± 1.24	26.59 ± 3.02	23.77 ± 2.14	26.13 ± 2.48
CF	4.47 ± 6.55	17.59 ± 1.63	23.86 ± 2.50	22.56 ± 2.32	21.45 ± 2.23
BART	2.57 ± 3.97	18.53 ± 2.02	26.41 ± 3.10	25.78 ± 2.66	27.45 ± 2.84
GANITE	5.79 ± 8.35	18.28 ± 1.66	24.50 ± 2.27	23.58 ± 2.48	25.12 ± 3.53
PD	5.14 ± 6.55	17.52 ± 1.62	20.88 ± 3.24	21.19 ± 2.29	22.28 ± 2.25
TARNET	1.32 ± 1.61	17.17 ± 1.25	23.40 ± 2.20	22.39 ± 2.32	21.19 ± 2.01
$\text{CFRNET}_{\text{wass}}$	0.88 ± 1.25	16.93 ± 1.12	22.65 ± 1.97	21.64 ± 1.82	20.87 ± 1.46
PM	0.84 ± 0.61	16.76 ± 1.26	21.58 ± 2.58	20.76 ± 1.86	20.24 ± 1.46
SITE	$\mathbf{0.81 \pm 1.22}$	16.87 ± 1.34	22.33 ± 2.08	21.84 ± 2.21	20.88 ± 1.52
MTAL	1.06 ± 1.28	$\mathbf{16.58 \pm 1.20}$	$\mathbf{20.42 \pm 1.88}$	$\mathbf{19.98 \pm 2.01}$	$\mathbf{19.32 \pm 1.76}$

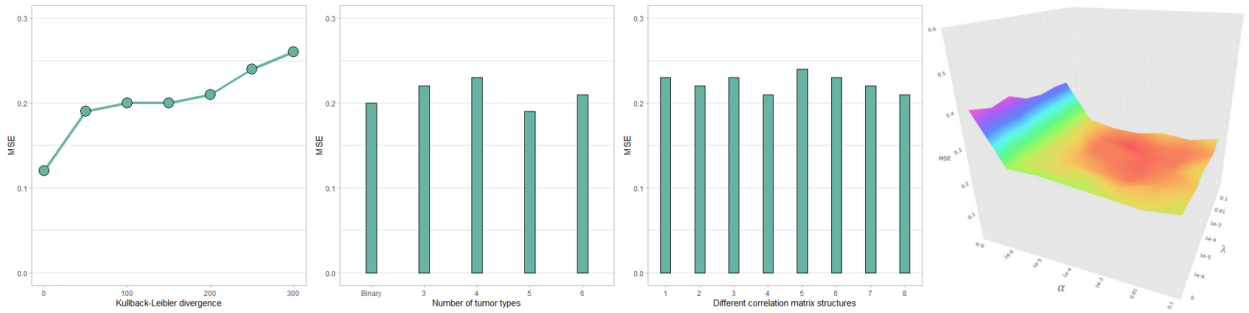

Figure 4: The results for synthetic basket trial data sets.

defined as $\epsilon_{\text{PEHE}} = \frac{1}{n}\sum_{i=1}^{n}(\text{ITE}_i - \widehat{\text{ITE}}_i)^2$, where $\text{ITE}_i = Y_1^i - Y_0^i$ and $\widehat{\text{ITE}}_i$ is an estimated ITE for unit i. In addition, for multiple treatment evaluations, we follow the definitions in (Schwab et al., 2018), where both ϵ_{PEHE} and ϵ_{ATE} can be extended to multiple treatments by averaging PEHE and ATE between every possible pair of treatments. They are defined as $\epsilon_{\text{mPEHE}} = \frac{1}{\binom{k}{2}}\sum_{t=1}^{k}\sum_{j=1}^{t}\epsilon_{\text{PEHE},t,j}$ and $\epsilon_{\text{mATE}} = \frac{1}{\binom{k}{2}}\sum_{t=1}^{k}\sum_{j=1}^{t}\epsilon_{\text{ATE},t,j}$. Table 2 and table 3 show the performance of our method and baseline methods on the IHDP and News data sets. MTAL achieves the best performance with respect to PEHE and ATE for News data sets with different numbers of treatment options. For the IHDP data set, MTAL still has competitive performance when compared to the best baseline methods with respect to PEHE and ATE. The results on these two benchmarks for conventional binary and multiple treatments effects estimation can demonstrate that our method is capable of precisely estimating the treatment effects.

4.2. Synthetic Basket Trial Data Set

Dataset. To evaluate our model's performance for basket trials, we simulate one synthetic data set which mimics the characteristics of a basket trial. Because different types of tumors may have different predictor variables, which may be a subset of all observed covariates, we use different subsets of the observable covariates to generate the outcomes for different tumor types. To mimic the real situation fur-

Table 3: Performance on IHDP and News data sets. We present mean ± standard deviation of ϵ_{ATE} and ϵ_{mATE} on the test sets.

	IHDP	News-2	News-4	News-8	News-16
Method	ϵ_{ATE}	ϵ_{ATE}	ϵ_{mATE}	ϵ_{mATE}	ϵ_{mATE}
kNN	3.19 ± 1.49	7.83 ± 2.55	19.40 ± 3.12	15.11 ± 2.34	17.27 ± 2.10
PSM	0.49 ± 0.81	4.89 ± 2.39	30.19 ± 2.47	22.09 ± 1.98	18.81 ± 1.74
RF	0.64 ± 1.25	5.50 ± 1.20	18.03 ± 3.18	12.40 ± 2.29	15.91 ± 2.00
CF	0.65 ± 1.24	4.02 ± 1.33	13.54 ± 2.48	9.70 ± 1.91	8.37 ± 1.76
BART	0.53 ± 1.02	5.40 ± 1.53	17.14 ± 3.51	14.80 ± 2.56	17.50 ± 2.49
GANITE	0.98 ± 1.90	4.65 ± 2.12	13.84 ± 2.69	11.20 ± 2.84	13.20 ± 3.28
PD	1.37 ± 1.65	4.69 ± 3.17	8.47 ± 4.51	7.29 ± 2.97	10.65 ± 2.22
TARNET	0.24 ± 0.29	4.58 ± 1.29	13.63 ± 2.18	9.38 ± 1.92	8.30 ± 1.66
$\text{CFRNET}_{\text{wass}}$	0.20 ± 0.24	4.54 ± 1.48	12.96 ± 1.69	8.79 ± 1.68	8.05 ± 1.40
PM	0.24 ± 0.20	3.99 ± 1.01	10.04 ± 2.71	6.51 ± 1.66	5.76 ± 1.33
SITE	$\mathbf{0.18 \pm 0.23}$	4.53 ± 1.32	12.75 ± 1.88	9.01 ± 1.86	8.63 ± 1.41
MTAL	0.34 ± 0.28	$\mathbf{3.88 \pm 1.11}$	$\mathbf{8.01 \pm 1.43}$	$\mathbf{5.97 \pm 1.58}$	$\mathbf{5.12 \pm 1.31}$

ther, we consider different covariance matrices in the covariates simulation. For example, the covariates predicting outcomes in each tumor type are taken to have stronger correlations than covariates predicting outcomes for other tumor types.

We generate a set of synthetic data sets which reflects the complexity of observational medical records data. The sample size for tumor type k is n_k, where $k = 1, 2, ..., K$. Therefore, the total sample size is $n = \sum_{k=1}^{K} n_k$ units. The predictor variables for tumor type k are $x_k \in \mathbb{R}^d$. The potential outcomes y_k for tumor type k are generated as $y_k|x_k \sim \cos\left((w_k^\intercal x_k)^2\right)$, where $w_k \sim Uniform((-1,1)^{d\times 1})$. The vector of all observed covariates $x = (x_1^\intercal, x_2^\intercal, ..., x_K^\intercal)^\intercal$ is sampled from a multivariate normal distribution with mean μ_k and different random positive definite covariance matrices Σ. By varying the value of μ_k, data with different levels of selection bias are generated (Yoon et al., 2018; Yao et al., 2018). Let D be the diagonal matrix with the square roots of the diagonal entries of Σ on its diagonal, i.e., $D = \sqrt{diag(\sigma)}$, then the correlation matrix is given by $R = D^{-1}\Sigma D^{-1}$. We simulate correlation matrix to better explain the relationship of covariates among and within different tumor types, instead of directly simulating covariates matrix. We use algorithm 3 in (Hardin et al., 2013) to simulate positive definite correlation matrices consisting of different within tumor type correlations and between tumor type correlations. Our correlation matrices are based on the hub correlation structure which has a known correlation between a hub variable and each of remaining variables (Zhang and Horvath, 2005; Langfelder et al., 2008). Each variable in a tumor type is correlated with the hub-variable with decreasing strength from specified maximum correlation to minimum correlation and different tumor types are generated independently or with weaker correlation among tumor types. Defining the first variable as the hub, for the ith variable ($i = 2, 3, ..., d$), the correlation between it and the hub-variable in one tumor type is given by $R_{i,1} = \rho_{\max} - \left(\frac{i-2}{d-2}\right)^{\gamma}(\rho_{\max} - \rho_{\min})$, where $\rho_{\max}$ and $\rho_{\min}$ are specified maximum and minimum correlations, and the rate γ controls rate at which correlations decay.

After specifying the relationship between the hub variable and the remaining variables in one tumor type, we use the Toeplitz structure to fill out the remainder of the hub correlation matrix and get the hub-Toeplitz correlation matrix R_k for tumor type k. Here, R is the $d \times d$ matrix having the blocks $R_1, R_2, ..., R_K$ along the diagonal and zeros at off-diagonal elements. This yields a correlation matrix with nonzero correlations within tumor types and zero correlation among tumor types. The amount of correlation among tumor types that can be added to the positive-definite correlation matrix R is determined by its smallest eigenvalue.

Results and Analysis. The mean squared error is used as the performance metric to eval-

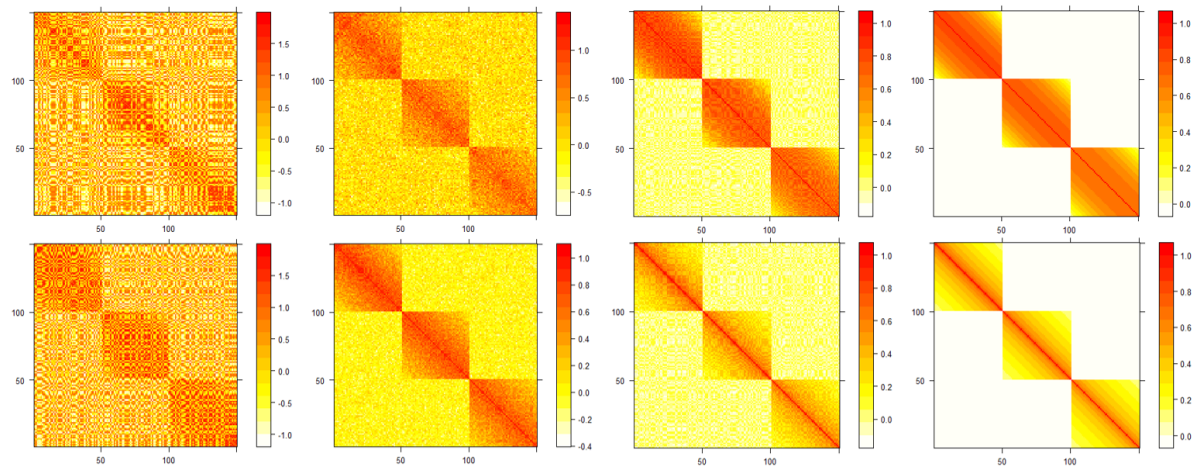

Figure 5: Different covariates correlation structures.

uate our model under the settings of binary or multiple tumor types, different selection bias, and different correlation matrix for observed covariates. The mean squared error is defined as $\text{MSE} = \frac{1}{n \times K} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(y_k(x_i) - \hat{y}_k(x_i)\right)^2$, where $y_k(x_i)$ and $\hat{y}_k(x_i)$ are factual and estimated outcomes for unit i with features x_i and tumor type k, respectively.

We simulate 5 data sets with $2, 3, 4, 5,$ and 6 tumor types, separately. From the second figure in Fig. 4, our MTAL performs relatively steadily for binary and multiple tumor types. To evaluate the performance with respect to selection bias, Kullback-Leibler divergence is adopted to quantify selection bias among different tumor types. Here, we use the data sets with binary tumor types. All observed covariates in two tumor types are generated by a multivariate normal distribution with mean 0 and different mean μ_1 for the remaining tumor types, so different values of μ_1 represent different Kullback-Leibler divergences; i.e., selection bias between two tumor types. From the first figure in Fig. 4, for the MTAL method, MSE increases modestly with increasing selection bias. To evaluate the sensitivity of the MTAL method to the correlation structure of the covariates, we generate 8 different correlation matrices with different levels of correlation for variables within each tumor type and among different tumor types in Fig. 5. From the third figure in Fig. 4, we find that the MSE in our MTAL method is not sensitive to the structure of the correlation matrices. In addition, from the fourth figure in Fig. 4, the performance of our model, with respect to MSE, is significantly improved compared to the models without L_1 or L_2 penalties. Also, the overall performance on different combinations of hyperparameters of L_1 and L_2 penalties is stable over a large range of tuning parameter values, which confirms the effectiveness and robustness of deep feature selection in our MTAL method.

5. Conclusion

In this paper, we propose a multi-task adversarial learning (MTAL) method by incorporating feature selection multi-task deep learning and adversarial learning to remove heterogeneity of tumor types in basket trials. To the best of our knowledge, our model is the first work introducing machine learning and causal inference to the task of analyzing basket trial data. It not only improves the basket trial analysis but also has its superiority over state-of-the-art methods in estimating multiple treatment effects for observational data. In future work, we will follow this direction to apply causal inference models and machine learning methods into more medical practical applications, such as umbrella, platform trials, and so on.

Institutional Review Board (IRB)

Our research does not require IRB approval.

References

Igor Astsaturov. Future clinical trials: genetically driven trials. *Surgical Oncology Clinics*, 26(4):791–797, 2017.

Scott M Berry, Kristine R Broglio, Susan Groshen, and Donald A Berry. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase ii oncology clinical trials. *Clinical Trials*, 10 (5):720–734, 2013.

Leo Breiman. Random forests. *Machine learning*, 45 (1):5–32, 2001.

Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359, 1992.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Yiyi Chu and Ying Yuan. A bayesian basket trial design using a calibrated bayesian hierarchical model. *Clinical Trials*, 15(2):149–158, 2018.

Zhixuan Chu, Stephen L Rathbun, and Sheng Li. Matching in selective and balanced representation space for treatment effects estimation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 205–214, 2020.

Zhixuan Chu, Stephen L Rathbun, and Sheng Li. Graph infomax adversarial learning for treatment effect estimation with networked observational data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 176–184, 2021.

Zhixuan Chu, Stephen Rathbun, and Sheng Li. Learning infomax and domain-independent representations for causal effect inference with real-world data. In *SDM*, 2022.

Peng Cui, Zheyan Shen, Sheng Li, Liuyi Yao, Yaliang Li, Zhixuan Chu, and Jing Gao. Causal inference meets machine learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3527–3528, 2020.

Kristen M Cunanan, Mithat Gonen, Ronglai Shen, David M Hyman, Gregory J Riely, Colin B Begg, and Alexia Iasonos. Basket trials in oncology: a trade-off between complexity and efficiency. *Journal of Clinical Oncology*, 35(3):271, 2017.

US Food, Drug Administration, et al. Guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics. *Federal Register*, 72, 2007.

Johanna Hardin, Stephan Ramon Garcia, and David Golan. A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, pages 1733–1762, 2013.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Akihiro Hirakawa, Junichi Asano, Hiroyuki Sato, and Satoshi Teramukai. Master protocol trials in oncology: review and new trial designs. *Contemporary clinical trials communications*, 12:1–8, 2018.

Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3): 199–236, 2007.

Daniel E Ho, Kosuke Imai, Gary King, Elizabeth A Stuart, et al. Matchit: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, http://gking. harvard. edu/matchit*, 2011.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720, 2008.

Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*, pages 929–939, 2017.

Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *IJCAI*, pages 3768–3774, 2016a.

Yifeng Li, Chih-Yu Chen, and Wyeth W Wasserman. Deep feature selection: theory and application to identify enhancers and promoters. *Journal of Computational Biology*, 23(5):322–336, 2016b.

Jacqueline Mersch, Michelle A Jackson, Minjeong Park, Denise Nebgen, Susan K Peterson, Claire Singletary, Banu K Arun, and Jennifer K Litton. Cancers associated with brca 1 and brca 2 mutations other than breast and ovarian. *Cancer*, 121 (2):269–275, 2015.

Jay JH Park, Ellie Siden, Michael J Zoratti, Louis Dron, Ofir Harari, Joel Singer, Richard T Lester, Kristian Thorlund, and Edward J Mills. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials*, 20(1):1–10, 2019.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. *arXiv preprint arXiv:1902.00981*, 2019.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085, 2017.

Richard Simon. Critical review of umbrella, basket, and platform designs for oncology clinical trials. *Clinical Pharmacology & Therapeutics*, 102(6): 934–941, 2017.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Karolina Strzebonska and Marcin Waligora. Umbrella and basket trials in oncology: ethical challenges. *BMC medical ethics*, 20(1):58, 2019.

Jessica J Tao, Alison M Schram, and David M Hyman. Basket studies: redefining clinical trials in the era of genome-driven oncology. *Annual review of medicine*, 69:319–331, 2018.

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

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: estimation of individualized treatment effects using generative adversarial nets. In *6th International Conference on Learning Representations*, 2018.

Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.