Conference Proceedings

CLINICAL TRIALS

Clinical Trials
2023, Vol. 20(4) 370–379
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17407745231173055
journals.sagepub.com/home/ctj
Sage

Inference on subgroups identified based on a heterogeneous treatment effect in a post hoc analysis of a clinical trial

Beibo Zhao[1], Anastasia Ivanova[1] and Jason Fine[2]

Abstract

Due to the many benefits of understanding treatment effect heterogeneity in a clinical trial, an exploratory post hoc subgroup analysis is often performed to find subpopulations of patients with conditional average treatment effect that suggests better treatment efficacy than in the overall population. A naive re-substitution approach uses all available data to identify a subgroup and then proceeds with estimation and inference using the same data set. This approach generally leads to an overly optimistic estimate of conditional average treatment effect. In this article, in a post hoc analysis, we estimate the target optimal subgroup through maximizing a utility function, from candidates systematically identified with a penalized regression. We then compare two resampling-based bias-correction methods, cross-validation and debiasing bootstrap, for obtaining approximately unbiased estimates and valid inference of conditional average treatment effect in the identified subgroup, with either an empirical or an augmented estimator. Our results show that both the cross-validation and the debiasing bootstrap methods reduce the re-substitution bias effectively. The cross-validation method appears to have less biased point estimates, smaller standard error estimates, but poorer coverages than the debiasing bootstrap method when using the empirical estimator and the sample size is moderate. Using the augmented estimator in the debiasing bootstrap method leads to less biased point estimates but poorer coverages. We conclude that bias correction should be a part of every exploratory post hoc subgroup analysis to eliminate re-substitution bias and to obtain a proper confidence interval for the estimated conditional average treatment effect in the selected subgroup.

Keywords

Subgroup analysis, treatment effect heterogeneity, bias correction, cross-validation, bootstrap

Background

The efficacy of a treatment may vary substantially among patients by their baseline biomarkers including laboratory values and socio-demographic variables, so a favorable average treatment effect for all patients from a randomized clinical trial (RCT) often does not translate to equally favorable benefits to each individual. Due to this natural heterogeneity of treatment effect within a patient population, there has been a rising need in the clinical trial community for a data-driven evaluation of patient subgroups to identify those that would benefit the most from a treatment. At the end of a RCT, in addition to the conventional objective of establishing an overall efficacy of the treatment, a post hoc subgroup discovery analysis is now more frequently performed to investigate the conditional average treatment effect (CATE) of promising patient subpopulations so that patient subgroups—clearly defined in terms of predictive covariates—with likely more pronounced and favorable responses than the population could be identified.[1] This practice has the benefits of optimizing design, sample size allocations, and the power of future trials where confirmatory subgroup analyses can be performed rigorously with these identified subgroups being prespecified while aiming at

[1]Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA

Corresponding author:
Anastasia Ivanova, Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420, USA.
Email: aivanova@bios.unc.edu

controlling the Type I error rate.[2,3] In addition, patients not belonging to the identified subgroups can be kept from exposure to potentially harmful side effects.

There are two phases in a subgroup analysis procedure: identification and confirmation. The goal of the identification phase is to select a desired subgroup from a collection of candidate subgroups. Many subgroup identification methods have been proposed.[2,4,5] The selection criteria vary across the methods, depending on the predetermined targets—maximizing CATE, a utility function that takes into account the "treatment burden," predictive power of future trials, or the value function of the optimal treatment assignment rule.[1,2,4–14]

In this article, we assume that a subgroup identification method has been previously selected for analysis and focus on the confirmatory phase. The goal of the subgroup confirmation phase is to obtain unbiased estimates and reliable inference of CATEs in the identified subgroups, also known as "honest" estimates.[2] These are typically obtained with additional test data not used for subgroup identification, but such luxury is uncommon in RCTs. A naive re-substitution approach often employed by investigators is to simply identify the subgroups from the whole sample and proceed to inference, not considering that the subgroup with the best treatment effect (or a function of the treatment effect) was selected. This practice often leads to over-optimism and false discoveries.[15,16] Since the results from subgroup analyses are often used in designing future confirmation trials and guiding practitioners to make informed treatment decisions, biased estimation of CATEs would lead to waste of resources and potential harm to patients. This is particularly true with the naive re-substitution method, where over-optimism in estimation of CATEs may lead to underpowered future studies.

When no independent test datasets are available, within the frequentist framework, resampling-based approaches—those involving cross-validation, bootstrap, and permutation—may often be the only feasible methods.[2] They often involve replicating the entire subgroup identification strategy, including estimation of any data-driven tuning parameter, in each resampled dataset. Foster et al.[17] proposed several resampling-based approaches and compared their performances with the naive re-substitution approach. They found a bias-corrected estimate based on nonparametric bootstrap to be the most promising. It was later noted, however, that using a bootstrap-based approach would introduce bias in the direction of pessimism.[2] Fuentes et al.[18] proposed the partition of sample space induced by order statistics to obtain a closed form expression for the coverage probability which allows for the construction of asymmetric intervals considering the selection procedure. However, this method requires patient subpopulations to be independent and normality distributed with a common variance, which can be a strong assumption to make. In general, most of the proposed estimators in literature are heuristic in nature and lacking in theoretical justifications. Their validity needs to be assessed by simulation under each individual setting.[19] Empirical or fully Bayesian methods provide attractive alternatives, but we limit our discussion to the frequentist framework in this article.[20–31]

In this article, we focus on the subgroup confirmation phase when no independent test dataset is available. We compare two existing and easily implementable bias-correction methods which can provide an approximately unbiased estimate of CATE and valid confidence interval (CI) for statistical inference on the selected subgroup. This work was motivated by the Precision Interventions for Severe and/or Exacerbation-prone Asthma (PrecISE) study (ClinicalTrials.gov Identifier: NCT04129931). PrecISE is a clinical trial to investigate five novel therapies for patients with severe asthma.[32] The treatments in PrecISE were selected to be likely to work in certain subgroups of patients rather than in all patients with severe asthma. One of the objectives of PrecISE is, in a post hoc analysis, to estimate and evaluate subgroups of patients who benefit the most from a given treatment. The post hoc subgroup estimation in PrecISE is performed using all available data due to relatively modest sample size. In the post hoc analysis, the optimal subgroup is defined as a subgroup which maximizes a utility function that provides a trade-off between subgroup prevalence and CATE. This optimal subgroup is estimated through selecting the subgroup that maximizes the estimated utility from a pool of promising candidates systematically identified from fitting a penalized regression that incorporates covariates of interest and thresholding the estimated linear predictors.

In this article, we aim to address the following question: what method can be used to give an unbiased estimation of CATE and to perform inference in the selected subgroup? To answer this question, we compare two methods: (1) a cross-validation (CV) approach to correct the re-substitution bias, with a bootstrap procedure to estimate the standard error for inference,[14,15] and (2) a model-free and asymptotically sharp debiasing bootstrap inference procedure proposed by Guo and He who provide theoretical justifications.[33] To improve the efficiency of estimation without compromising the consistency, we consider not only the standard empirical estimator but also an augmented estimator with the doubly robust property established in the field of causal inference.[34–36]

The rest of this article is organized as follows. In the "Methods" section, we formulate the problem, explain the definition of the optimal subgroup, describe the subgroup identification method, provide the forms of the estimators, and give a brief overview of the two bias-correction methods. In the "Results" section, we evaluate the finite-sample performance of both empirical and augmented estimators in these two bias-correction methods through a Monte Carlo simulation study. In the "Discussion" section, we summarize our work and give concluding remarks.

Methods

Problem setting

Consider a parallel group clinical trial, where n patients are randomized 1:1 to either the treatment arm or the control arm. Let T denotes the treatment indicator, 0 for placebo and 1 for treatment; let $\boldsymbol{X}$ denote a $M-$dimensional vector of covariates of interest measured at baseline prior to treatment. Let Y be a normally distributed outcome; let higher values of Y indicate more beneficial outcomes. Under the potential outcome framework, $Y = (Y^1, Y^0)$ indicates the possible outcomes that patients will have after receiving either treatment $(T = 1)$ or control $(T = 0)$. We assume that the observed data consist of n independent and identically distributed copies of $(Y, T, \boldsymbol{X})$, $\{(Y_i, T_i, \boldsymbol{X}_i), i = 1, \dots, n\}$. Let $\boldsymbol{H}(\boldsymbol{X}, T)$ be a $p-$dimensional function of baseline covariate vector $\boldsymbol{X}$ and treatment T, including the intercept.

The fundamental goal of causal inference is the comparison between the observed outcome under one regime and the counterfactual outcome that would have been observed under the other regime.[37,38] Within this framework, the average treatment effect is formulated as $E[Y^1 - Y^0]$, indicting the expected treatment effect among all patients in the trial. Let χ denote the support of $\boldsymbol{X}$, and let $S \subset \mathcal{X}$ denote a subgroup defined by a set of covariate values. The CATE is formulated as $\Delta(S) = E[Y^1 - Y^0 \mid S]$, indicating the expected treatment effect in the subgroup S.

Defining the optimal subgroup

In PrecISE, the goal is to find a subgroup with a good balance between subgroup size and CATE, as we do not want to expose patients unnecessarily to treatment unless there is a sufficient benefit. We define the optimal subgroup S_{opt} as a subgroup that maximizes the utility function $\pi_S^w \Delta(S)$ for a prespecified value of w, where π_S is the subgroup prevalence in the sample. The parameter w is the power adjustment term to π_S and reflects a trade-off between subgroup size and CATE. When w is 0, the optimal subgroup maximizes CATE and is typically very small. Increasing w leads to selection of larger subgroups with smaller CATEs, and the whole population will become the optimal subgroup under large values of w. We choose $w = 0.5$, as then $\pi_S^{0.5} \Delta(S)$ is proportional to the power of treatment comparison or, equivalently, the noncentrality parameter in the test for the treatment effect.[13] The specific choice of w is somewhat arbitrary and should be left to the discretion of the investigators.

Subgroup identification method

We identify the optimal subgroup using the following steps:

1. In the original dataset, with a slight abuse of notation, propose a set of $H \geqslant 1$ possibly overlapping candidate subgroups $\bar{S} = \{S_h\}_{h=1}^H$ each associated with true utility $\pi_{S_h}^{0.5} \Delta(S_h)$ and estimated utility $\pi_{S_h}^{0.5} \hat{\Delta}(S_h)$, where $\Delta(S_h)$ and $\hat{\Delta}(S_h)$ are CATE and estimated treatment effect in the h th candidate, respectively. By our definition, the optimal subgroup is $S_{opt} = \underset{h \in [H]}{\arg\max} \left\{ \pi_{S_h}^{0.5} \Delta(S_h) \right\}$.
2. Estimate S_{opt} through selecting from $\bar{S}$ the subgroup $\hat{S}_{opt} = \underset{h \in [H]}{\arg\max} \left\{ \pi_{S_h}^{0.5} \hat{\Delta}(S_h) \right\}$.

In this section, we provide a description of a subgroup identification method designed to propose the set of candidates $\bar{S} = \{S_h\}_{h=1}^H$. Many subgroup identification methods have been proposed and comparatively reviewed through a variety of performance metrics under the frequentist framework to identify the optimal subgroup defined by a specified rule on biomarkers, for a given treatment under the randomized trial setting.[2,4,5,39–41] It is not the focus of this article to add to the comparisons among these methods. Hence, we choose the penalized regression for its wide utility, ease of implementation, and built-in complexity control. Penalized regression is a data-driven approach that explicitly incorporates model selection and provides a comprehensive characterization of treatment effect heterogeneity. This global outcome modeling method models the response as a function of prognostic effects of baseline covariates as the main effects, and their predictive effects as interactions with treatment, with proper constraints (i.e. penalty functions) to perform parameter estimation simultaneously with automatic variable selection. When lacking knowledge of how potentially predictive covariates interact with each other to influence the outcome, the complexity control from the penalty terms supports complex fits that expand the covariate space.[2]

For a continuous outcome Y, we use a generalized linear model as the working model to model the expected outcome as the p-dimensional function $\boldsymbol{H}(\boldsymbol{X}, T)$—including the intercept, the treatment indicator, main effects of baseline biomarkers, biomarkers' two-way interactions (including quadratic terms), biomarkers' interactions with treatment, and biomarkers' two-way interactions with treatment—with the p-dimensional vector of model parameters $\boldsymbol{\beta}$. We estimated $\hat{\boldsymbol{\beta}}$ through minimizing an objective function defined as a sum of the loss function and penalty function:

$$\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \sum\nolimits_{i=1}^{n} L(y_i, \boldsymbol{H}(\boldsymbol{x}_i, T_i|\boldsymbol{\beta})) + J_\lambda(\boldsymbol{\beta}),$$

where $\boldsymbol{H}(\boldsymbol{X}, T)$ is the specified linear regression function; $L(Y, \boldsymbol{H}(\boldsymbol{X}, T)) = (Y - \boldsymbol{H}(\boldsymbol{X}, T))^2$ is the squared-error loss function; $J_\lambda(\boldsymbol{\beta})$ is a selected penalty function for variable selection with λ as the penalty parameter.

Since $\boldsymbol{H}(\boldsymbol{X}, T)$ has the overlapped grouping of main effects and their interactions, the selected penalty should incorporate the information in said structure for variable selection purpose. Hence, we choose the overlapping group exponential lasso penalty function which was developed as an extension of its nonoverlapping version, group exponential lasso, to allow for overlapping groups for the interaction terms in the model. This penalty is a bi-level selection method that can select not only the important groups but also the important elements within groups. In addition, this penalty has been shown to have superior performance than conventional penalties in terms of estimation accuracy, group selection, and individual variable selection.[42]

Supposing that the total p predictors in $\boldsymbol{H}$ are assigned into J possibly overlapping groups, the overlapping group exponential lasso penalty is the overlapping group extension by decomposing the original coefficient vectors in the group exponential lasso penalty, which has the penalization applied to a coefficient decay exponentially as the group j grows in importance. This penalty is sparse at both the group and the individual levels, so it is a bi-level selection method that can select not only the important groups but also the important coefficients within those groups:

$$\sum\nolimits_{j=1}^{J} f_{\lambda,\zeta}(\boldsymbol{\gamma}^{\boldsymbol{j}}),$$

where $f(*)$ denotes the exponential penalty[43]

$$f_{\lambda,\zeta}(\theta) = \frac{\lambda^2}{\zeta}\left\{1 - \exp\left(-\frac{\zeta\theta}{\lambda}\right)\right\},$$

with regularization parameter λ and coupling parameter ζ. Coupling parameter $\zeta = 1/3$ which was found to be broadly successful as a default value. Optimal λ is determined from the following CV procedure. Suppose that the estimator from the fitted model $\hat{f} = \hat{f}_\lambda$ depending on λ, the predictive accuracy of $\hat{f}_\lambda$ can be optimized via a CV procedure. Randomly divide the training data into K folds, the CV error is

$$CV(\lambda) = \frac{1}{n}\sum\nolimits_{k=1}^{K} \sum\nolimits_{i \in F_k} \left(y_i - \hat{f}_\lambda^{-k}(x_i)\right)^2.$$

For our analysis, we obtain the optimal λ via a 10-fold CV through the minimal rule: selecting the model that minimizes the CV error curve.

After fitting the penalized regression, a set of estimated linear predictors, $\hat{\eta}_1, \ldots, \hat{\eta}_n$, is obtained for the set of covariate vectors $\boldsymbol{X}$, through plugging in estimated coefficients $\hat{\boldsymbol{\beta}}$ while fixing $T = 1$. For each estimated $\hat{\eta}_i$, we may define two subgroups through thresholding: $\left\{S_{\boldsymbol{s},\boldsymbol{g}} : \hat{\eta}_s > \hat{\eta}_i\right\}$ and $\{S_{\boldsymbol{s},\boldsymbol{l}} : \hat{\eta}_s \leqslant \hat{\eta}_i\}$, resulting in a set of total $2n$ candidate subgroups where each of them is defined as a set of inequalities between estimated linear predictors, denoted with $\bar{S} = \{S_h\}_{h=1}^{H}$. Note that the whole sample space is also considered as a candidate subgroup by the second inequality. We choose to define candidate subgroups this way as it is easily implementable and provides a large pool of candidates. We then estimate S_{opt} through selecting from candidates the subgroup $\hat{S}_{opt} = \underset{h \in [H]}{\arg\max}\left\{\pi_{S_h}^{0.5}\hat{\Delta}(S_h)\right\}$.

Types of estimators

Once we have the estimated optimal subgroup $\hat{S}_{opt}$ from the subgroup identification phase, we consider two types of estimators for CATE $\Delta(\hat{S}_{opt})$ in the following subgroup confirmation phase: empirical $\hat{\Delta}(\hat{S}_{opt})^{emp}$ and augmented $\hat{\Delta}(\hat{S}_{opt})^{aug}$. The empirical estimator is the inverse probability weighting estimator that is based on propensity score modeling in causal inference.[44] Since under the RCT setting, the propensity score is always correctly specified due to randomization with a set ratio (i.e. not necessarily 1:1), and unconfoundedness to be safely assumed, the empirical estimator is asymptotically consistent for $\Delta(\hat{S}_{opt})$, and can be simply formulated as the difference of sample means:

$$\hat{\Delta}(S)^{emp} = \frac{\sum_{i=1}^{n} Y_i I[\boldsymbol{X}_i \in S,\ T_i = 1]}{\sum_{i=1}^{n} I[\boldsymbol{X}_i \in S,\ T_i = 1]} - \frac{\sum_{i=1}^{n} Y_i I[\boldsymbol{X}_i \in S,\ T_i = 0]}{\sum_{i=1}^{n} I[\boldsymbol{X}_i \in S,\ T_i = 0]}.$$

Note that for brevity, we use the generic subgroup notation S in the formulation, but actual estimation is conducted for $\hat{S}_{opt}$. This estimator does not incorporate any covariate information, leading to potential inefficiency. To improve the efficiency of this estimator without compromising its consistency, we also consider the following augmented estimator, which is the augmented inverse probability weighting estimator that has the doubly robust property:[35,36,45]

$$\hat{\Delta}(S)^{aug} = \hat{\Delta}(S)^{emp} - \left\{\frac{\sum_{i=1}^{n} I[\boldsymbol{X}_i \in S]\{I[T_i = 1] - \hat{\pi}_{1,S}\}\ \widehat{\mu}_1(\boldsymbol{X}_i)}{\sum_{i=1}^{n} I[\boldsymbol{X}_i \in S,\ T_i = 1]} - \frac{\sum_{i=1}^{n} I[\boldsymbol{X}_i \in S]\{I[T_i = 0] - \hat{\pi}_{0,S}\}\ \widehat{\mu}_0(\boldsymbol{X}_i)}{\sum_{i=1}^{n} I[\boldsymbol{X}_i \in S,\ T_i = 0]}\right\},$$

where $\hat{\pi}_{t,S}$ is the proportion of patients in S that receive treatment regimen t; $\hat{\mu}_t(\boldsymbol{X_i})$ is a regression-based estimator of the expected outcome for a patient with covariates vector $\boldsymbol{X_i}$ while receiving treatment regimen t. In this article, we obtain the regression estimator from fitting a penalized linear regression as the working model in the subgroup identification phase. We do not consider this estimator separately because its consistency relies on the working model being correctly specified. Under a randomized setting, $\hat{\Delta}(S)^{aug}$ is guaranteed to be asymptotically consistent for $\Delta(\hat{S}_{opt})$ and it will achieve the semiparametric efficiency bound if $\hat{\mu}_t(\boldsymbol{X})$ is consistent for $\mu_t(\boldsymbol{X})$.[46] However, there is the potential issue of efficiency loss in $\hat{\Delta}(S)^{aug}$ if $\hat{\mu}_t(\boldsymbol{X}_i)$ is heavily biased.

Bias-correction methods

In the subgroup confirmation phase, we apply the following two bias-correction methods to debias either the empirical estimator $\pi^{0.5}_{\hat{S}_{opt}}\hat{\Delta}(\hat{S}_{opt})^{emp}$ or the augmented $\pi^{0.5}_{\hat{S}_{opt}}\hat{\Delta}(\hat{S}_{opt})^{aug}$ estimator of true utility in $\hat{S}_{opt}$, denoted by $\pi^{0.5}_{\hat{S}_{opt}}\Delta(\hat{S}_{opt})$.

CV method. Proposed by Freidlin and colleagues[15,47,48] this method is a CV extension of their adaptive signature design and has been shown to considerably improve its performance. They propose to use k-fold CV to debias an estimator and conduct inference using a permutation test which permutes treatment labels. Zhang et al.[14] argue that it is more desirable to use a bootstrap procedure for inference as it will test the specific null hypothesis of no treatment effect heterogeneity in the selected subgroup, instead of a test of the sharp null hypothesis of no treatment effect heterogeneity in any subgroup. Therefore, we use a nonparametric bootstrap procedure to obtain the bootstrap standard error and construct a level α CI of the debiased estimator accordingly.

We execute the CV method in the subgroup confirmation phase with the following procedure:

1. Randomly partition the original dataset into roughly equal-sized K folds.
2. For each $k \in \{1,\dots,K\}$, use the kth fold as the testing fold and combine the rest into the training fold. Apply the entire subgroup identification strategy described in the subgroup identification phase to the training fold to propose a set of $H_{\{-k\}} \geqslant 1$ possibly overlapping candidate subgroups $\bar{S}_{\{-k\}} = \left\{S_{h_{\{-k\}}}\right\}_{h_{\{-k\}}=1}^{H_{\{-k\}}}$ and select from $\bar{S}_{\{-k\}}$ the subgroup $\hat{S}_{\{-k\},opt} = \underset{h_{\{-k\}}\in[H_{\{-k\}}]}{\operatorname{argmax}}\left\{\pi^{0.5}_{S_{h_{\{-k\}}}}\hat{\Delta}\left(S_{h_{\{-k\}}}\right)\right\}$, with $\{-k\}$ emphasizes these objects are based on the training sample. Next, apply $\hat{S}_{\{-k\},opt}$ to the kth testing fold to obtain the k-fold estimate of subgroup proportion $\pi_{k,\hat{S}_{opt}}$ and for a given estimator of true utility $\pi^{0.5}_{\hat{S}_{opt}}\Delta(\hat{S}_{opt})$ (either empirical or augmented in our case; if augmented, outcome model is re-estimated based on the training fold and then applied to the testing fold), the k-fold CV result of the estimator. Calculate the average of CV results across all K folds as the final debiased CV estimator of true utility in $\hat{S}_{opt}$.
3. Create a nonparametric bootstrap sample b by sampling n patients with replacement from the original dataset. Next, apply the process described in step 2 to obtain a bootstrap estimate. Repeat this process for a total B times to produce a collection of bootstrap estimates, whose sample standard deviation is the bootstrap standard error for construction of a level α CI for the final debiased CV estimator.

Debiasing bootstrap method. Guo and He proposed a resampling-based bias-correction method that takes the selection process into account to address subgroup selection bias, but without the need of repeating the entire subgroup identification strategy in each resampled dataset.[33] The proposed method uses a bootstrap procedure to learn about the bias and to construct a one-sided level α CI for any specification of the subgroup effect as the target of maximization.[33] We specify our defined utility function as the target and construct a two-sided level α CI for inference, which we note is different from the scheme considered in Guo and He. The theoretical investigation is beyond the scope of this article. However, Guo and He establish the validity of the proposed method requiring asymptotic normality of the subgroup effect estimates and their bootstrap estimates at each subgroup, in addition to some mild assumptions—we may safely assume that our utility estimators satisfy these conditions.

We execute the debiasing bootstrap method in the subgroup confirmation phase with the following procedure:

1. Assume the observed data consist of n independent and identically distributed copies of $\{(Y_i, T_i, Z_i),\ i = 1,\dots,n\}$ where $Z_i \subset \bar{S} = \{S_h\}_{h=1}^{H}$ is an indicator denotes which candidate subgroup patient i is part of. Note that the set of candidate subgroups $\bar{S} = \{S_h\}_{h=1}^{H}$ are generated in the subgroup identification phase.
2. Create a nonparametric bootstrap sample b by sampling n patients with replacement from $\{(Y_i, T_i, Z_i),\ i = 1,\dots,n\}$. The estimated utility in the hth candidate subgroup in sample b is denoted by $\pi^{0.5}_{S_h}\hat{\Delta}_b(S_h)$, for $h = 1,\dots,H$. Given the assumption that utility estimates are asymptotically normal and the assumption of bootstrap consistency, the bias between true utility $\pi^{0.5}_{\hat{S}_{opt}}\Delta(\hat{S}_{opt})$ and estimated utility $\pi^{0.5}_{\hat{S}_{opt}}\Delta(\hat{S}_{opt})$—based on a given estimator (either empirical or augmented in our case; if augmented, outcome model is re-estimated based on the

bootstrap sample)—in $\hat{S}_{opt}$ can be learned through the bias between estimated utility $\pi^{0.5}_{\hat{S}_{opt}}\hat{\Delta}(\hat{S}_{opt})$ and the bootstrap-adjusted estimate $\max_{h\in[H]}(\pi^{0.5}_{S_h}\hat{\Delta}_b(S_h) + d_h)$ in which the adjustment value d_h is computed as $d_h = (1 - n^{r-0.5})[\pi^{0.5}_{\hat{S}_{opt}}\hat{\Delta}(\hat{S}_{opt}) - \pi^{0.5}_{S_h}\hat{\Delta}(S_h)]$, where $r \in (0, 0.5)$ is a tuning parameter.

3. Obtain the debiased estimator of $\pi^{0.5}_{\hat{S}_{opt}}\Delta(\hat{S}_{opt})$ by removing from $\pi^{0.5}_{\hat{S}_{opt}}\hat{\Delta}(\hat{S}_{opt})$ the learned bias from bootstrap estimates $\pi^{0.5}_{\hat{S}_{opt}}\hat{\Delta}(\hat{S}_{opt}) - E[\max_{h\in[H]}(\pi^{0.5}_{S_h}\hat{\Delta}_b(S_h) + d_h) - \pi^{0.5}_{\hat{S}_{opt}}\hat{\Delta}(\hat{S}_{opt})]$, where E is the expectation under the bootstrap distribution. The corresponding two-sided level α CI is $\pi^{0.5}_{\hat{S}_{opt}}\hat{\Delta}(\hat{S}_{opt}) \pm c_{\alpha/2}/\sqrt{n}$, where $c_{\alpha/2} = \text{quantile}(\sqrt{n}(\max_{h\in[H]}(\pi^{0.5}_{S_h}\hat{\Delta}_b(S_h) + d_h) - \pi^{0.5}_{\hat{S}_{opt}}\hat{\Delta}(\hat{S}_{opt})), 1-\alpha/2)$.

Guo and He also proposed a CV procedure to adaptively select the tuning parameter $r \in (0, 0.5)$ through minimizing the mean square error of the debiased estimator.[33] We implement this adaptive procedure in our analysis. The detailed steps are not listed in this article for brevity as the modification are similar to how we adapt their debiasing procedure for our defined utility function. This bias-correction method, as originally proposed, uses the empirical estimator that is model-free, that is, no modeling in each resampled dataset. We further implement the augmented estimator to improve efficiency. This method is no longer model-free when using an augmented estimator because the specification of the augmented estimator requires a regression-based estimator to incorporate covariate information, that is, re-modeling is now required in each resampled dataset. Since the augmented estimator is robust to misspecification of the regression model in the setting of RCTs, the debiasing methods are also robust to such misspecification when using this estimator.

We do note that the proposed method in Guo and He[33] is developed to be theoretically valid for predefined candidate subgroups and post hoc identified candidate subgroups defined over a given space, while in our implementation the candidate subgroups are identified from the same data which might induce additional bias. To fully investigate the impact of this difference, we considered two additional settings for the empirical estimator: (1) evaluate the set of candidate subgroups $\bar{S} = \{S_h\}_{h=1}^{H}$ in separate data following the same procedure and (2) re-estimate the outcome model in each bootstrap sample (making bootstrap no longer model-free).

Results

We use Monte Carlo simulations to evaluate the finite-sample performance of the two bias-correction methods in terms of bias and empirical coverage of the estimated CATE. In each simulation trial, we generate patients' covariates matrix $\boldsymbol{X}$ which consists of five standard uniform variables $X_1, ..., X_5$ and five Bernoulli variables $X_6, ..., X_{10}$ with $p = 0.5$, all independent from each other. The patients are randomized with a ratio of 1:1 to receive either treatment ($T = 1$) or control ($T = 0$). We generate continuous outcome Y from the following change-point model:

$$Y = \mu(\boldsymbol{X}) + \beta T I(\boldsymbol{X} \in S_{opt}) + \epsilon\varepsilon,$$

Where $\mu(\boldsymbol{X})$ is an unknown baseline mean function for patients in control; β is a fixed scalar-value parameter; ε is a standard normal random variable.

We choose $\mu(\boldsymbol{X})$ to be a constant value of 0.1 and use $\beta = 0.6$ (true CATE Δ_{opt}) for all scenarios. We consider the following optimal subgroup scenarios for data generation

Scenario 1: $S_{opt} = \{\varnothing\}, \pi_{opt} = 1$;
Scenario 2: $S_{opt} = \{X_1 + X_2 > 1\}, \quad \pi_{opt} = 0.5$;
Scenario 3: $S_{opt} = \{X_1 + X_2 > 1 X_6 + X_7 \geqslant 1\}, \pi_{opt} = 0.375$;

where $\varnothing$ denotes the empty set—indicating the sharp null hypothesis that there is no treatment effect heterogeneity in any subgroup that is defined by $\boldsymbol{X}$.

For each scenario, we run 1000 simulation trials, each with a moderate sample size of 400. We run an additional set of 1000 trials for scenario 3, with a much larger sample size of 2000. In each trial, we analyze the data using the procedure described in the subgroup identification section: fit the working model, obtain a set of candidate subgroups, and then proceed to estimate S_{opt} with $\hat{S}_{opt}$ through selecting the candidate that maximizes the estimated utility.

To obtain naive estimate of utility, we apply this procedure to the whole sample and proceed to point estimation and inference without any adjustment. To obtain CV estimate of utility, we apply the CV bias-correction method with 10-fold to the original sample for fold-average estimation and its 100 bootstrap samples for bootstrap standard error. To obtain debiasing bootstrap estimate of utility, we apply the debiasing bootstrap method to the whole sample and its 100 bootstrap samples for debiased estimation and confidence limits. For the adaptive selection of the tuning parameter $r \in (0, 0.5)$ in this method, we use the CV procedure proposed by Guo and He[33] with 10-fold.

To compare these three sets of estimates and to assess the finite-sample performance of the methods, we create an external validation sample of size 20,000 from the same distribution, independent from the original sample, and on which the same process to obtain naive estimates is repeated, except for treating $\hat{S}_{opt}$ estimated from the original sample as predefined. We refer to the obtained estimates as reference estimates. Since

Table 1. Simulation results for naive, CV, and debiasing bootstrap methods.

CATE	True	Reference	Naive, empirical	CV, empirical	Debiasing bootstrap			
					Empirical	Augmented	Empirical, separate data	Empirical, re-fitting
Scenario 1 (n = 400)								
Point estimate (bias)	0.00	0.00 (—)	0.19 (0.19)	0.00 (0.00)	0.06 (0.06)	0.03 (0.03)	0.05 (0.05)	0.07 (0.07)
Estimated SE	—	0.03	0.17	0.14	0.23	0.26	0.21	0.23
95% CI coverage	—	—	0.83	0.95	0.94	0.94	0.93	0.94
Scenario 2 (n = 400)								
Point estimate (bias)	0.60	0.47 (—)	0.71 (0.23)	0.44 (−0.03)	0.54 (0.07)	0.48 (0.01)	0.45 (−0.03)	0.53 (0.06)
Estimated SE	—	0.10	0.16	0.17	0.21	0.25	0.27	0.24
95% CI coverage	—	—	0.71	0.92	0.95	0.84	0.94	0.94
Scenario 3 (n = 400)								
Point estimate (bias)	0.60	0.37 (—)	0.68 (0.31)	0.34 (−0.03)	0.49 (0.13)	0.43 (0.07)	0.40 (0.03)	0.48 (0.12)
Estimated SE	—	0.11	0.18	0.17	0.24	0.28	0.30	0.28
95% CI coverage	—	—	0.62	0.89	0.93	0.84	0.93	0.92
Scenario 3 (n = 2000)								
Point estimate (bias)	0.60	0.54 (—)	0.59 (0.05)	0.53 (−0.01)	0.52 (−0.01)	0.53 (−0.01)	0.53 (−0.02)	0.51 (−0.02)
Estimated SE	—	0.06	0.06	0.07	0.08	0.13	0.14	0.10
95% CI coverage	—	—	0.87	0.90	0.94	0.89	0.96	0.96

CATE: conditional average treatment effect; CI: confidence interval; CV: cross-validation; SE: standard error.

the focus of this article is not on the performance of the subgroup identification method, we use the reference estimate as the target for bias and coverage calculation, instead of the true value. We report the results with estimated CATE instead of estimated utility—which is the subgroup effect estimate targeted in both subgroup identification and confirmation phases—because CATE is more interpretable in terms of clinical meaningfulness. Note that both bias-correction methods can target either the estimated utility or the estimated CATE. We conduct simulations comparing these two approaches and find that debiasing the estimated CATE leads to larger bias, larger estimated standard errors, and under-coverage. This is likely due to the inconsistency between how the optimal subgroup is estimated (maximizing utility) and the debiasing target (CATE). We also conduct simulations comparing the number of bootstrap samples and find that increasing to up to 500 samples hardly changes the results, indicating that our current choice of 100 bootstrap samples is proper. These additional simulation results are available from the first author.

Table 1 summarizes the simulation results in terms of the empirical means of point estimation, bias, estimated standard errors, and the empirical 95% CI coverage. The true CATE is shown under "True." The reference estimates are shown under "Reference." We present results of the naive empirical estimates, the CV empirical estimates, the debiasing bootstrap empirical estimates, and the debiasing bootstrap augmented estimates. In addition, we present results from two debiasing bootstrap variations: empirical estimates with separate data and empirical estimates with re-fitting. We do not include augmented estimates for naive and CV methods because they do not show improvement in performance compared with their empirical counterparts. The biases in point estimates and empirical 95% CI coverages for these methods are calculated using the reference point estimate as the truth.

First, we note that the reference estimates are close to the true values and with very small estimated standard errors in scenario 1 when there is no treatment effect heterogeneity. As scenarios increase in complexity (scenarios 2 and 3), the reference estimates gradually deviate from the true values as $\hat{S}_{opt}$ estimated from the subgroup identification method departs from the global optimum S_{opt}. In scenario 3 with large sample size (n = 2000), $\hat{S}_{opt}$ estimates S_{opt} quite well again. Even though the performance of the subgroup identification method is not the focus of this article, this observation indicates that penalized regression is adequate and supports the use of reference estimates as our targets. Second, we note that the naive method leads to overly optimistic estimation of CATE (the re-substitution bias), resulting in under-coverage even with large sample size.

Next, we compare the bias-correction methods. In terms of point estimation, all methods do correct the re-substitution bias, as seen in scenario 3 ($n = 2000$) where the biases are close to 0. However, with smaller sample size ($n = 400$), there are noticeable differences in their finite-sample performances. The CV method appears to be less biased than the debiasing bootstrap method, and its estimates are slightly biased in the negative direction, suggesting that the CV method has the tendency to over-correct. The estimates from the debiasing bootstrap method, on the contrary, are slightly biased in the positive direction. It appears that the bias in CV method is less affected by the complexity of scenarios—as seen in the same −0.03 bias in both scenarios 2 and 3. The bias in debiasing bootstrap method greatly increases in complex scenarios as indicated by the increase in bias from 0.07 to 0.13 with the empirical estimator and from 0.01 to 0.07 with the augmented estimator. Using separate data or using the augmented estimator instead of the empirical estimator in the debiasing bootstrap method reduces the bias substantially. Re-fitting the outcome model in the bootstrap sample does not appear to reduce bias in the empirical estimator. In terms of efficiency, the CV method appears to be the most efficient (smallest estimated standard errors) across all scenarios. The debiasing bootstrap method is less efficient than the CV method, with the augmented estimator having slightly larger estimated standard errors than the empirical estimator. Re-fitting the outcome model appears to lead to slightly worse efficiency with more complex scenarios. This is likely due to the mis-specified working model. Our working model is very different from the model used to generate the data. This results in unreliable regression-based estimator, which appears to have a negative impact on the augmented estimator.

In terms of inference, the empirical coverages for the debiasing bootstrap empirical estimates are close to the nominal level in all scenarios. The empirical coverages for the CV empirical estimates drop to around 0.90 in scenario 3—increasing the sample size to 2000 does reduce estimated standard errors by half and causes a small increase in coverage from 0.89 to 0.90. The empirical coverages for the debiasing bootstrap and augmented estimates drop to around 0.85 in scenario 2 and scenario 3—increasing the sample size to 2000 does reduce estimated standard errors by half and improve the coverage from 0.84 to 0.89. Inference with CV empirical estimates and with debiasing bootstrap empirical estimates are affected by the complexity of scenarios—they both experience a slight drop in coverages comparing scenario 2 and scenario 3 ($n = 400$). The coverages for debiasing bootstrap augmented estimates are not affected. Increasing the sample size does improve the performance of all bias-correction methods. These methods appear to provide unbiased point estimation and correct inference asymptotically.

Discussion

Subgroup identification and subsequent subgroup confirmation are the two integral parts in any post hoc subgroup analyses of clinical trials. Consistently, identifying subgroups with a heterogeneous treatment effect and accurately estimating CATE can be very valuable to the design of future confirmatory clinical trials, and to the practice of clinicians. The naive estimation and inference of CATEs in selected subgroups result in over-optimism (large bias) and under-coverage. In this article, we compare the performance of two bias-correction methods to inform us on addressing a key objective in our motivating study PrecISE: to estimate and evaluate subgroups of patients who benefit the most from a given treatment in a post hoc analysis. In the identification phase, we fit a penalized regression as the working model to the data to generate candidates from which to estimate the optimal subgroup through maximizing a specified utility function that is equivalent to maximizing the power for demonstrating treatment efficacy. In the confirmation phase, we apply the bias-correction methods to the estimated utility in the estimated optimal subgroup to obtain a debiased estimate of CATE and to construct the corresponding 95% CI. We consider not only the empirical estimator but also the augmented estimator. Simulation results show that both the CV method and the debiasing bootstrap method reduce re-substitution bias effectively, especially when the sample size is large. When using the empirical estimator, the CV method with bootstrap standard error appears to have less biased point estimates, smaller standard errors, but poorer coverages than the debiasing bootstrap method under moderate sample size. In addition, the CV method requires more computational time and resources because the calculation of bootstrap standard error requires applying the entire subgroup identification strategy to each bootstrap sample, effectively multiplying the time needed by the number of bootstrap samples. The debiasing bootstrap method, on the contrary, is quick and easy to implement when using the empirical estimator because it is model-free, that is, no modeling in each resampled dataset. Using separate data to infer the utility function reduces the biases in point estimates and has good coverages, corresponding to the results shown in the theoretical framework laid out by Guo and He.[33] Re-fitting the outcome model, however, does not appear to lead to any improvement. When using the augmented estimator, the debiasing bootstrap method has less biased point estimates but poorer coverages comparing to using the empirical estimator, with a moderate sample size. In addition, it is no longer model-free because the outcome model needs to be re-estimated based for each bootstrap sample to obtain the regression estimates, leading to a similar computational cost as the CV method.

In summary, our simulation study shows that both the CV method and the debiasing bootstrap method can be successfully used to obtain bias-corrected estimates of CATE in post hoc subgroup estimation. The debiasing bootstrap method appears to be more efficient but more biased than the CV method with the empirical estimator and a moderate sample size. However, it has the advantage of being model-free, thus not computationally intensive. Using the augmented estimator reduces the finite-sample bias but leads to efficiency loss, possibly caused by poor regression estimates. These results are helpful in informing our analysis for the PrecISE trial in the future, as bias correction will be an important component of a post hoc subgroup analysis which aims to identify the optimal subgroup and make inference on the selected subgroup.

Acknowledgement

The authors thank anonymous reviewers and the associate editor for their helpful comments.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Work of Anastasia Ivanova is supported in part by the National Heart, Lung, and Blood Institute, National Institutes of Health (grant no. U24 HL138998).

ORCID iD

Anastasia Ivanova https://orcid.org/0000-0003-4321-2073

References

1. Talisa VB and Chang CH. Learning and confirming a class of treatment responders in clinical trials. *Stat Med* 2021; 40: 4872–4889.
2. Lipkovich I, Dmitrienko A and D' Agostino RB. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med* 2017; 36: 136–196.
3. Tanniou J, Van Der Tweel I, Teerenstra S, et al. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. *BMC Med Res Methodol* 2016; 16: 20.
4. Alemayehu D, Chen Y and Markatou M. A comparative study of subgroup identification methods for differential treatment effect: performance metrics and recommendations. *Stat Methods Med Res* 2018; 27(12): 3658–3678.
5. Loh WY, Cao L and Zhou P. Subgroup identification for precision medicine: a comparative review of 13 methods. *Wires Data Mining Knowl Discovery* 2019; 9: e1326.
6. Zhang P, Ma J, Chen X, et al. A nonparametric method for value function guided subgroup identification via gradient tree boosting for censored survival data. *Stat Med* 2020; 39: 4133–4146.
7. Dusseldorp E and Van Mechelen I. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Stat Med* 2014; 33: 219–237.
8. Lipkovich I, Dmitrienko A, Denne J, et al. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat Med* 2011; 30: 2601–2621.
9. Huber C, Benda N and Friede T. A comparison of subgroup identification methods in clinical drug development: simulation study and regulatory considerations. *Pharm Stat* 2019; 18(5): 600–626.
10. Chen S, Tian L, Cai T, et al. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* 2017; 73(4): 1199–1209.
11. Zhao L, Tian L, Uno H, et al. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clin Trials* 2012; 9(5): 570–577.
12. Joshi N, Fine J, Chu R, et al. Estimating the subgroup and testing for treatment effect in a post-hoc analysis of a clinical trial with a biomarker. *J Biopharm Stat* 2019; 29(4): 685–695.
13. Lai TL, Lavori PW and Liao OY. Adaptive choice of patient subgroup for comparing two treatments. *Contemp Clin Trials* 2014; 39(2): 191–200.
14. Zhang Z, Li M, Lin M, et al. Subgroup selection in adaptive signature designs of confirmatory clinical trials. *J R Stat Soc Ser C Appl Stat* 2017; 66: 345–361.
15. Freidlin B, Jiang W and Simon R. The cross-validated adaptive signature design. *Clin Cancer Res* 2010; 16: 691–698.
16. Simon R. Development and validation of biomarker classifiers for treatment selection. *J Stat Plan Inference* 2008; 138: 308–320.
17. Foster JC, Taylor JMG and Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med* 2011; 30: 2867–2880.
18. Fuentes C, Casella G and Wells MT. Confidence intervals for the means of the selected populations. *Electr J Stat* 2018; 12: 58–79.
19. Dusseldorp E, Conversano C and Van Os BJ. Combining an additive and tree-based regression model simultaneously: STIMA. *J Comput Graph Stat* 2010; 19: 514–530.
20. Dixon DO and Simon R. Bayesian subset analysis. *Biometrics* 1991; 47: 871–881.
21. Hodges JS, Cui Y, Sargent DJ, et al. Smoothing balanced single-error-term analysis of variance. *Technometrics* 2007; 49: 12–25.
22. Andrew G, Aleks J, Maria Grazia P, et al. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat* 2008; 2: 1360–1383.
23. Park T and Casella G. The Bayesian Lasso. *J Am Stat Assoc* 2008; 103: 681–686.
24. Gu X, Yin G and Lee JJ. Bayesian two-step Lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemp Clin Trials* 2013; 36(2): 642–650.

25. Chipman HA, George EI and McCulloch RE. Bayesian CART Model Search. *J Am Stat Assoc* 1998; 93: 935–948.
26. Ohwada S and Morita S. Bayesian adaptive patient enrollment restriction to identify a sensitive subpopulation using a continuous biomarker in a randomized phase 2 trial. *Pharm Stat* 2016; 15(5): 420–429.
27. Krisam J and Kieser M. Performance of biomarker-based subgroup selection rules in adaptive enrichment designs. *Stat Biosci* 2016; 8: 8–27.
28. Ondra T, Jobjörnsson S, Beckman RA, et al. Optimized adaptive enrichment designs. *Stat Methods Med Res* 2019; 28: 2096–2111.
29. Krisam J and Kieser M. Decision rules for subgroup selection based on a predictive biomarker. *J Biopharm Stat* 2014; 24(1): 188–202.
30. Bornkamp B, Ohlssen D, Magnusson BP, et al. Model averaging for treatment effect estimation in subgroups. *Pharm Stat* 2017; 16(2): 133–142.
31. Woody S and Scott J. Optimal post-selection inference for sparse signals: a nonparametric empirical-Bayes approach. arXiv, 2018, https://arxiv.org/abs/1810.11042
32. Ivanova A, Israel E, Lavange LM, et al. The precision interventions for severe and/or exacerbation-prone asthma (PrecISE) adaptive platform trial: statistical considerations. *J Biopharm Stat* 2020; 30: 1026–1037.
33. Guo X and He X. Inference on selected subgroups in clinical trials. *J Am Stat Assoc* 2021; 116: 1498–1506.
34. Bang H and Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; 61(4): 962–973.
35. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* 1994; 89: 846–866.
36. Qin J. *Causal inference and missing data problems*. New York: Springer, 2017, pp. 353–408.
37. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974; 66: 688–701.
38. Imbens GW and Rubin DB. *Causal inference for statistics, social, and biomedical sciences: an introduction*. Cambridge: Cambridge University Press, 2015.
39. Doove LL, Dusseldorp E, Van Deun K, et al. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Adv Data Anal Classificat* 2014; 8: 403–425.
40. Antoniou M, Jorgensen AL and Kolamunnage-Dona R. Biomarker-guided adaptive trial designs in phase II and phase III: a methodological review. *PLoS ONE* 2016; 11: e0149803.
41. Ondra T, Dmitrienko A, Friede T, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *J Biopharm Stat* 2016; 26(1): 99–119.
42. Breheny P. The group exponential lasso for bi-level variable selection. *Biometrics* 2015; 71(3): 731–740.
43. Breheny P and Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat Comput* 2015; 25(2): 173–187.
44. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70: 41–55.
45. Scharfstein DO, Rotnitzky A and Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Am Stat Assoc* 1999; 94: 1096–1120.
46. Tsiatis AA. *Semiparametric theory and missing data*. New York: Springer, 2006.
47. Freidlin B and Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res* 2005; 11: 7872–7878.
48. Jiang W, Freidlin B and Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst* 2007; 99: 1036–1043.