Received: 21 August 2022 | Revised: 23 December 2022 | Accepted: 27 February 2023

DOI: 10.1002/wics.1614

OVERVIEW

WIREs COMPUTATIONAL STATISTICS

WILEY

Subgroup analysis and adaptive experiments crave for debiasing

Jingshen Wang[1] | Xuming He[2]

[1]Division of Biostatistics, UC Berkeley, Berkeley, California, USA

[2]Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Xuming He, Department of Statistics, University of Michigan, 1085 S University, Ann Arbor, MI 48103, USA.
Email: xmhe@umich.edu

Funding information
National Science Foundation USA, Grant/Award Numbers: DMS-1914496, DMS-2015325

Edited by: James E. Gentle, Commissioning Editor and Co-Editor-in-Chief, and David W. Scott, Review Editor and Co-Editor-in-Chief

Abstract

Results obtained from reliably designed randomized experiments are often considered to be evidence of the highest grade for assessing the effectiveness of biomedical or behavioral interventions. Nevertheless, even with randomized experiments, statistical bias can arise in post hoc analysis of the data or through adaptive data collection. In this article, we discuss the need for and review some of the recent developments in statistical methodologies to address the issue of potential bias in adaptive experiments and in subgroup analysis. For adaptive experiments, we focus on adaptive treatment assignments. For subgroup analysis, we focus on post hoc subgroup selection and review several frequentist approaches for debiased inference on the best-selected subgroup effects.

This article is categorized under:
Applications of Computational Statistics > Clinical Trials

KEYWORDS

adaptive design, bootstrap, clinical trial, resampling, selection bias, sequential experiment, statistical inference

1 | INTRODUCTION

Randomized experiments have become the gold standard for assessing the effectiveness of biomedical or behavioral interventions, and the evidence collected from randomized experiments is often considered the highest grade. Even with randomized experiments, statistical bias can arise in post hoc analysis of the experimental data and through adaptive designs of data collection that aim to achieve better treatment-covariate balance or higher power of verifying the effectiveness of an intervention. This article discusses these two sources of bias.

On the one hand, while understanding and characterizing treatment effect heterogeneity from analyzing randomized controlled trial (RCT) data has become an increasingly important task in many scientific fields, statistical bias due to post doc data analyses cannot be overlooked. For example, in precision health, identifying differential treatment effects serves as an important step toward materializing the benefits of precision health. However, simply searching for the patient subpopulation that benefited most from a treatment without accounting for subpopulation selection can potentially result in false positive discoveries. In clinical settings, the study of treatment effect heterogeneity is often referred to as subgroup analysis (see, Loh et al., 2019, for a review of various approaches). Classical subgroup analysis has been traditionally divided into stages: exploratory subgroup analysis which focuses on identifying promising subgroups from existing data, and confirmatory subgroup analysis which aims to evaluate a few prospectively defined treatment effects in multiple subpopulations. There is often an intermediate stage in subgroup analysis where one would like to evaluate the treatment effect of a subgroup selected post hoc from the data. Such an evaluation needs to account for the subgroup selection process to help guide further and possibly confirmatory trials. For example, the initial clinical trial for the medicine Vectibix (panitumumab) for metastatic colorectal cancer did not win regulatory approval but a subgroup of patients without RAS gene mutation was found to be promising (Peeters et al., 2015). A fair evaluation of the selected subgroup treatment effect requires appropriate statistical treatment. To find statistically valid inference on the selected subgroup without sacrificing statistical power (from performing simultaneous inference, for instance, over all possible subgroups), we need to tailor the method to the specific selection procedure used to identify the subgroup. In this article, we shall review the selection bias issue that arises in subgroup analysis and summarize recent statistical methodological developments on debiased inference.

On the other hand, while adaptive experimental designs–that allow randomization probabilities to be adapted during the trial based on sequentially accrued data–have received much attention in the past decade (Hu & Rosenberger, 2006; Robertson et al., 2020; Rosenberger & Lachin, 2015; Thall & Wathen, 2007; Wathen and Thall, 2017), even without subgroup analysis, the statistical validity of classical treatment effect evaluation approaches can be challenged given that the collected data are no longer independently distributed. The first part of this manuscript reviews some related literature on experiment design strategies and potential bias issues of some commonly adopted estimators from a frequentist viewpoint.

The popularity of adaptive designs partiality ascribes to their potential advantages in improving statistical efficiency and in promoting precision health. This is because when subjects are sequentially enrolled in a trial, the goal of adaptive designs includes, for example, learning which treatment is more promising and/or which subgroups (if any) benefit from a treatment and subsequently changing the randomization to assign more subjects to the better treatment. Such sequential randomization strategies have equipped adaptive designs with the ability to periodically detect subgroups that respond favorably to a particular treatment, and then optimize personalized intervention assignment strategy based on the inferred context. Over time, adaptive designs can result in a robust evidence-based study not only for testing the overall treatment effect differences, but also for identifying personalized intervention strategies that are generalizable beyond the study sample.

In the rest of this article, we first review general data collection mechanisms commonly adopted in randomized experiments in Section 2. The purpose of this section is to provide a gentle introduction to some commonly used covariate-adaptive and response-adaptive designs and point to some recently developed statistical inferential tools in the literature. We also make it clear that treatment allocation plans need to be accounted for to avoid bias and ensure statistical validity of the analysis of data from such designs. We then describe and discuss the selection bias issue commonly associated with subgroup analysis in Section 3, followed by several frequentist approaches to debiased inference in Section 4.

Even though adaptive design and subgroup analysis are often considered separate topics in the literature, the need to address the potential bias remains a common thread, because in each case, the data used for effect size estimation and inference are no longer a sequence of independent and representative observations from the targeted population or subpopulations. We close this article with some concluding remarks in Section 5.

2 | DATA COLLECTION MECHANISM IN RANDOMIZED EXPERIMENTS

In this section, we review three types of data collection mechanisms in randomized experiments, including completely randomized experiments, covariate and response adaptive experiments.

To streamline the presentation and simplify the discussion, we focus on two-arm experiments, and the discussed design strategies can be extended to multi-armed settings as well. We consider a scenario where participants are sequentially enrolled in the experiment and respond to treatments without much delay. Let n be the total number of enrollments. For the ith enrolled participant, let $T_i \in \{0,1\}$ be the treatment assignment status, $Y_i(t)$ be the potential outcome if participant i receives the treatment $t \in \{0,1\}$, and $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$ be the observed outcome, for $i = 1, ..., n$. A common parameter of interest in randomized experiments is the average treatment effect (ATE),

$$\tau = \mathbb{E}[Y(1) - Y(0)],$$

which measures the mean difference between the potential outcomes in two treatment arms. Furthermore, let X_i be a vector of covariates used in statistical inference after the experiment is finished, and let Z_i denote the covariates used in adaptive experiments, which could overlap with X_i. Lastly, we use π_i to denote the probability of allocating participant i to the first treatment.

2.1 | Complete randomization

In traditional randomized experiments, a complete randomization (or fixed randomization) scheme is often adopted, and the treatment assignment probability does not change as the trial progresses. When two treatments are compared in such experiments, each enrolled participant, regardless of timing, has a fixed probability of being randomly allocated into two treatment groups, that is

$$\pi_i = \mathbb{P}(T_i = 1) = \text{constant}, \quad T_i \perp\!\!\!\perp (Y_i(1), Y_i(0)), \quad i = 1, \ldots, n,$$

where $\perp\!\!\!\perp$ denotes stochastic independence of random variables. Furthermore, when the experiment is completed, we collect a random sample $\{(Y_i, T_i, X_i, Z_i)\}_{i=1}^n$ with mutually independent observations. Such an independent structure in the sample enables valid statistical inference for many causal quantities of interest, including quantile treatment effect (Athey et al., 2021; Firpo, 2007; Giessing & Wang, 2021) and heterogeneous ATE (Guo et al., 2022; Guo & He, 2021; Imai & Ratkovic, 2013; Künzel et al., 2019).

Complete randomization is easy for both implementation and analysis but is not always the best strategy to use. When study participants arrive sequentially, it has been documented in the literature (Pocock, 1979) that the complete randomization scheme may suffer from at least two potential issues. First, the complete randomization does not really balance patients' prognostic factors such as age, gender, and disease stage that may influence the observed outcome. Second, sampling variations may lead to unequally sized treatment groups especially when the trial size is small (Efron, 1971). In finite samples, such imbalance implies higher variability in the statistical analysis.

To avoid these potential issues in complete randomization, adaptive clinical trials (or adaptive designs in general) use some information accumulated during a trial to modify the experimental design as the trial progresses. As the patient covariate and/or outcome information accumulates during the trial, adaptive design strategies allow the trial designer to modify the treatment assignment probabilities π_i and/or to drop inferior arms based on the outcomes observed so far, often with the goal of balancing patient prognostic factors, lowering a trial's cost, reducing a trial's duration, or making more patients benefit from the better arm. The advantages of adaptive trials are widely discussed in the literature (Berry, 2012), and, in what follows, we review two common adaptive design strategies.

2.2 | Covariate adaptive design

The first category is the covariate-adaptive randomization design. Covariate-adaptive randomization has been widely used in modern clinical trials to balance treatment assignments across important prognostic factors. This design refers to a randomized treatment allocation scheme that depends on participant covariate information (e.g., prognostic factors), but is independent of the observed outcomes, conditional on the covariates used in the randomization. In other words, for the ith subject, we have

$$T_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) \mid \{Z_i\}_{i=1}^n,$$

and the allocation probability is decided by the accumulated covariate information and treatment assignment status, that is

$$\pi_i = \mathbb{P}\left(T_i = 1 \mid \left\{T_j\right\}_{j=1}^{i-1}, \left\{Z_j\right\}_{j=1}^{i}\right).$$

Rosenberger and Sverdlov (2008) provided an excellent review for covariate-adaptive randomization, some of which stem from variations of the biased coin design first introduced by Efron (1971), including minimization (instead of randomization) (Taves, 1974), stratified block adaptive randomization (Pocock & Simon, 1975), matching on the fly (Kapelner & Krieger, 2014), and designs that attempt to minimize the variance of the treatment effect (Atkinson, 1982). More recently, Qin et al. (2016) introduced a pairwise sequential randomization strategy which works with two or more participants at a time to achieve covariate balancing; see also Zhou et al. (2018) for a related approach based on stage-wise sequential re-randomization.

Upon the completion of a covariate adaptive experiment, the collected sample no longer consists of mutually independent observations, but two "treatment balance" properties, if verified to hold, enable valid statistical inference on the ATEs. The first treatment balance property is called the "global balance:"

$$\frac{1}{n}\sum_{i=1}^{n}(2T_i - 1) = o_p(1),$$

suggesting that the proportion of units in each treatment group converges to 0.5 as n increases. The second is called the "covariate balance:"

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(2T_i - 1)Z_i \rightsquigarrow \boldsymbol{\xi}, \quad \mathbb{E}[\boldsymbol{\xi}] = 0,$$

where the wiggly arrow $\rightsquigarrow$ denotes convergence in distribution, and $\boldsymbol{\xi}$ is a random vector that measures the asymptotic imbalance between the covariate distributions in the two treatment groups. The covariate balance property implies that the means of the covariates in two treatment groups approach the same as the sample size increases. These two properties together ensure that a covariate adaptive experiment achieves good balancing properties, across both treatment assignments and covariates, so that unbiased statistical inference can be drawn efficiently.

In covariate adaptive experiments, whether and how to adjust for the covariates in any follow-up data analysis has been a long-standing question. In general, bias may arise whenever we treat the data collected from covariate adaptive experiments as completely randomized experiments. Shao et al. (2010) and Shao and Yu (2013) showed that the simple t-test without using any covariate is conservative under covariate-adaptive biased coin randomization in terms of its Type I error rate. Those authors have shown that excluding covariates used for adaptive randomization from statistical analyses can lead to a distortion of the sampling distribution of the test statistics, which consequently invalidates the usual statistical inference. With discrete covariates, Ma et al. (2015) and Bugni et al. (2018) generalized the earlier results in Shao et al. (2010), and Shao and Yu (2013) to more general randomization schemes. Ma et al. (2020) and Wang et al. (2021) further studied the statistical properties of the covariate adaptive designs under more general setups incorporating continuous covariates. Recent work by Ye et al. (2022) provided a unified theoretical framework for analyzing various ATE estimators without requiring a linear model to be correctly specified, and demonstrated that the regression adjustment estimator incorporating the covariates and treatment-covariate interactions could be asymptotically optimal. In particular, Ye et al. (2022) showed that this regression adjustment estimator incorporating the covariates and treatment-covariate interactions in covariate adaptive designs has exactly the same asymptotic variance as in complete randomization (Lin, 2013). Similar conclusions have been reached in Li and Ding (2020) for rerandomization. Those recent results raised questions about whether the first-order asymptotic results are sufficient to compare covariate-adaptive designs with complete randomization in realistic sample sizes.

Before we end the section on covariate adaptive design, we found that the R package carat can be used to conduct statistical inference on the ATE. Following Ye et al. (2022), simple estimators, including the ordinary least squares estimator (which can be immediately implemented in R), can also provide unbiased estimates of the ATE.

2.3 | Response adaptive design

The second category is the response adaptive randomization (RAR) design, in which the treatment assignment probabilities can be adapted during the experiment based on the accrued evidence in the outcomes, with the goal of simultaneously achieving the experimental objectives and preserving statistical inference validity. Here, the experimental objectives include, for example, identifying the treatment assignment probability that maximizes the power of detecting the treatment effect, and minimizing the expected number of treatment failures in the trial (see Hu & Rosenberger, 2006, for a review). The RAR relies on the allocation probability of participant i as

$$\pi_i = \mathbb{P}\left(T_i = 1 | \{T_j\}_{j=1}^{i-1}, \{Y_j\}_{j=1}^{i-1}\right).$$

A recent article by Robertson et al. (2020) provided a comprehensive review of RAR designs from both frequentist and Bayesian perspectives.

Over the past decades, the development of targeted therapies and precision medicine has promoted the use of covariate information in RAR. In particular, when different participants respond to the same treatment differently, incorporating covariate information into RAR can improve the estimation efficiency of the ATE (Hahn et al., 2011; Zhang et al., 2007). Whenever the covariate information is adopted in RAR, we have the covariate-adjusted response adaptive (CARA) design (Hu et al., 2015; Sverdlov et al., 2013; Villar & Rosenberger, 2018). In such cases, the allocation probability of participant i is decided by the accumulated outcome information, the treatment assignment status, and the covariates, that is

$$\pi_i = \mathbb{P}\left(T_i = 1 | \{T_j\}_{j=1}^{i-1}, \{Y_j\}_{j=1}^{i-1}, \{Z_j\}_{j=1}^{i}\right).$$

As the sequentially collected data from RAR and CARA designs are generally not independent, conducting statistical inference on the ATE typically depends on developing martingale central limit theorems for simple difference-in-mean estimators (which measure the mean difference between the treatment and controlled arms).

For the data collected from response adaptive experiments, without restricting the observed outcome to follow any parametric distribution, we discuss some estimators of the ATE frequently used in the literature. First, take a look at the simple difference-in-mean (DiM) estimator (Bowden & Trippa, 2017; Nie et al., 2018; Shin et al., 2019a, 2019b):

$$\widehat{\tau}^{\text{DiM}} = \widehat{\tau}_1^{\text{DiM}} - \widehat{\tau}_0^{\text{DiM}} = \frac{\sum_{i=1}^{n} Y_i T_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} Y_i (1 - T_i)}{\sum_{i=1}^{n} (1 - T_i)}.$$

Its bias of the corresponding estimator of $\tau_t = E[Y(t)]$, $(t = 0, 1)$ has a closed-form expression:

$$\mathbb{E}\left[\widehat{\tau}_t^{\text{DiM}} - \tau_t\right] = -\frac{\mathbb{C}\text{ov}\left[n_t, \widehat{\tau}_t^{\text{DiM}}\right]}{n_t}, \quad n_t = \sum_{i=1}^{n} \mathbf{1}(T_i = t), \quad \text{for } t = 0, 1.$$

The downward bias occurs because the treatment in which we observe initial random upward fluctuations will be sampled more, while the treatment in which we observe initial random downward fluctuations will be sampled less. More intuitively, this negative bias can be explained heuristically following Bowden and Trippa (2017). Suppose in the early stage of a response adaptive experiment, the treatment effect of a certain arm is overestimated because a large number of participants experience a positive response by random chance. RAR (response adaptive randomization) will then assign more patients to receive the seemingly beneficial arm. This means that the larger $\widehat{\tau}_t^{\text{DiM}}$ is, the larger n_t is, which creates a positive covariance between $\widehat{\tau}_t^{\text{DiM}}$ and n_t. From a practical point of view, the magnitude of such bias in any given situation varies, and we refer to Thall et al. (2015) for some quantitative demonstration of the bias in adaptive randomization. They showed that under some adaptive randomization design, "there is a 25% chance that one will overstate the actual benefit of treatment B over A by two fold or more." Note that this bias vanishes asymptotically for the data collected from covariate adaptive designs but not so for response adaptive designs.

A popular bias corrected estimator to accommodate over-/under-sampling is the inverse propensity score weighted (IPW) estimator (Hadad et al., 2021; Rosenbaum & Rubin, 1983):

$$\widehat{\tau}^{\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_iT_i}{\pi_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(1-T_i)}{1-\pi_i}.$$

Since T_i is randomly assigned independent of Y_i given $\left\{\left(Y_j, X_j, T_j\right)\right\}_{j=1}^{i-1}$, this estimator is unbiased, or asymptotically so when π_i are consistently estimated. Nevertheless, the IPW estimator might suffer from two potential issues caused by unbalanced treatment assignments during the experiment.

On the one hand, even when the true propensity scores are bounded away from zero and one, some of the estimated propensity scores might be very close to zero or one, resulting in an inflated variance in finite samples. On the other hand, in multi-armed bandit problems (Russo et al., 2018; Xu et al., 2013) for the best arm identification, when the probability of assignment to inferior arms tends to zero, the inverse propensity score weights increase. This in turn causes the tail distributions of $\widehat{\tau}^{\mathrm{IPW}}$ to become heavier—a phenomenon known as limited overlap and positivity violation in the causal inference literature (Azevedo et al., 2020; D'Amour et al., 2021; Ma & Wang, 2020).

A generalization of the IPW estimator is the augmented inverse propensity score (AIPW) estimator

$$\widehat{\tau}^{\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi_i} + \left(1-\frac{T_i}{\pi_i}\right)\widehat{\mu}_1(Z_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i(1-T_i)}{1-\pi_i} + \left(1-\frac{1-T_i}{1-\pi_i}\right)\widehat{\mu}_0(Z_i)\right],$$

where $\mu_t(z) = \mathbb{E}[Y(t)|Z=z]$ is the conditional mean of the potential outcome given the covariates, and $\widehat{\mu}_t(z)$ is an estimator of $\mu_t(z)$. The AIPW uses a regression adjustment, but it inherits the same drawback mentioned above for the IPW estimator. When the limited overlap phenomenon is not a concern, we can carry out statistical inference on the ATE either via the R package AIPW (Zhong et al., 2021) or via the R package DoubleML (Bach et al., 2021).

To partially address the issue of limited overlap, Hadad et al. (2021) and Bibaut et al. (2021) proposed a different generalized IPW estimator with adaptive weighting, and the resulting estimators converge to normal distributions under appropriate conditions. Instead of constructing an estimator that is robust to unusually small/large propensity scores, some remedies have also been proposed to directly revise the Thompson sampling algorithm so that the propensity scores are bounded away from zero and one (see Kasy & Sautmann, 2021, for example).

Adaptive data collection beyond the clinical trial setting has received attention in the literature. The adaptive sampling problems (Bucher, 1988; Etor'e & Jourdain, 2010) and reinforcement learning algorithms (Agarwal et al., 2022; Lu et al., 2021; Russo et al., 2018; Sutton & Barto, 2018) share a similar spirit. The literature has documented that the estimation bias issues may arise in those contexts as well (Hasselt, 2010; Nikishin et al., 2022), and debiased statistical inference needs to be used to adapt to any specific design of sampling plan or treatment allocation. The rest of the review article, however will not discuss the broader issue of bias from adaptive designs and other adaptive data collection mechanisms but instead focus on statistical issues in subgroup analysis where subgroup selection bias needs to be accounted for.

3 | EXPLORATORY, DEBIASED, AND CONFIRMATORY SUBGROUP ANALYSES

After an experiment is finished, practitioners may investigate the heterogeneity of the treatment effect across multiple study sub-populations, defined based on the pretreatment covariate information. In clinical settings, such investigations, called subgroup analysis, play an essential role in interpreting clinical trial findings toward the general goal of precision medicine. Nevertheless, the limitation of subgroup analysis is well recognized in clinical trials, and post hoc inference on the best selected subgroup can lead to false positives due to subgroup selection bias. In this section, we shall provide an overview of subgroup analysis with a focus on subgroup selection bias. The existing literature on selection bias has been mainly for data with independent observations; and we will give a brief discussion at the end of this manuscript for adaptively collected data.

In clinical settings, subgroup analysis includes two types of work: the "exploratory subgroup analysis" which focuses on post hoc subgroup identification and assessments, and the "confirmatory subgroup analysis" which works with one or a few prospectively defined subsets of the overall population. The latter often evaluates the treatment effects for the subgroups with a new confirmation clinical trial. Because additional trials are costly and the seemingly promising subgroup selected from exploratory subgroup analysis suffers from "subgroup selection bias," a recent paper by Guo and He (2021) worked toward debiased evaluation and inference for the most promising subgroup selected from the same trial data. This intermediate step aims to help decision makers decide whether there is sufficient statistical evidence to move forward with a confirmatory trial on the selected subgroup. As this step aims to correct the subgroup selection bias, we refer to it as "debiased subgroup analysis."

To better illustrate different types of subgroup analysis, we adopt notations similar to those in the previous section. Suppose we have collected an *i.i.d.* sample $\{(Y_i,T_i,X_i)\}_{i=1}^n$ from a completely randomized experiments, where $Y_i \in \mathbb{R}$ is the outcome variable, $T_i \in \{0,1\}$ denotes the binary treatment indicator variable, and $X_i \in \mathcal{X}$ denotes the covariates collected from subject i. These covariates can be baseline measurements, and may include participants' demographic information, laboratory test results, or questionnaire responses.

In exploratory subgroup analysis, the goal is to find sub-populations defined by regions of the sample space $\mathcal{X}$ in which the treatment is effective or the treatment effect exceeds a prespecified threshold (Foster et al., 2011). When we measure the effectiveness of a treatment via the ATE, exploratory subgroup analysis aims to find d possibly overlapping subsets $\mathcal{A}_j \subseteq \mathcal{X}$ so that

$$\tau_j = \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i \in \mathcal{A}_j\right] \neq 0, \quad \text{or} \quad \tau_j = \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i \in \mathcal{A}_j\right] \geq c,$$

for $j = 1,...,d$, with a prespecified threshold c. Other measures of the treatment effects, including log odds ratio or log hazard ratio, can be used to quantify subgroup treatment effects as well. The candidate subgroups $\mathcal{A}_1,...,\mathcal{A}_d$ can be either prespecified or post hoc identified from the data. When subgroups are post hoc identified, Alemayehu et al. (2018) provided an overview of recent subgroup identification literature based on data-driven methods along with their operational characteristics. Liu et al. (2019) discussed tree-based approaches for subgroup identification and proposed two evaluation criteria in connection with traditional type I error and power concepts. Wei et al. (2020) discussed subgroup identification approaches for longitudinal studies.

When candidate subgroups are prespecified, the subgroup treatment effects can be either parametrically estimated by directly modeling the relationship between the observed outcome and treatment-by-covariate interactions (Brankovic et al., 2019), or nonparametrically estimated via augmented inverse propensity score weighting and targeted maximum likelihood estimators (Wei, van der Laan, et al., 2022). When subgroups need to be learned adaptively from data, there is a growing literature on identifying heterogeneous treatment effects using machine learning methods; see Loh et al. (2019) for a recent review. Shen and He (2015) proposed a latent logistic-normal mixture model where the outcome is modeled via a linear model incorporating an interaction between the treatment and a latent subgroup variable, and the subgroup membership probabilities are modeled by logistic regression with preselected biomarkers as the covariates. Imai and Ratkovic (2013); Tian et al. (2014) formulated the problem on heterogeneous treatment effect identification from a variable selection perspective, and Lasso (Tibshirani, 1996) is adopted to select significant interaction terms between the treatment and subgroup indicators. In a similar spirit, Ma and Huang (2017) adopted fusion penalized approaches for grouping subgroups. Su et al. (2009); Athey and Imbens (2016); and Wager and Athey (2018) proposed recursive partitioning tree approaches to identify treatment heterogeneity. Hill (2011) and Hahn et al. (2020) adopted Bayesian additive regression tree models for treatment heterogeneity identification. Zeldow et al. (2019) further extended Bayesian additive regression trees to a semiparametric framework.

In confirmatory subgroup analysis, the goal is to evaluate and confirm the efficacy of one or a few prespecified subgroups with a new experiment. These prespecified subgroups may come from clinical understandings of the disease and the drug or from exploratory subgroup analysis in an earlier trial. Because confirmatory subgroup analysis is less encountered in industrial or social studies, we focus on clinical settings. There, statistical methods designed for confirmatory subgroup analysis mostly aim at controlling Type I error rate or false discovery rate when multiple subgroups (or treatments) are under consideration (Dmitrienko & D'Agostino Sr, 2013).

Nevertheless, we have been frequently reminded of the failure of Phase III trials to confirm a seemingly promising subgroup identified from exploratory subgroup analysis (Kubota et al., 2014; Petticrew et al., 2012). The subgroup selection bias is a major culprit (Cook et al., 2014; Guo & He, 2021), because such bias may lead to an overly optimistic evaluation of the selected subgroup obtained from the exploratory subgroup analysis. In what follows, we review the subgroup selection bias within a statistical framework and introduce the concept of debiased subgroup analysis when candidate subgroups in exploratory analyses are prespecified.

Without loss of generality, we assume that a larger value of τ_j means a better treatment effect and write the ordered values of $\tau_1, \ldots, \tau_d$ as $\tau_{(1)} \geq \ldots \geq \tau_{(d)}$. Suppose we have constructed a consistent estimator $\widehat{\tau}_j$ of τ_j, for $j = 1, \ldots, d$ and write the order statistics of $\widehat{\tau}_1, \ldots, \widehat{\tau}_d$ as $\widehat{\tau}_{(1)} \geq \ldots \geq \widehat{\tau}_{(d)}$. For practical reasons, analysts often need to select the most promising subgroups with the highest observed treatment effects among a set of candidate subgroups $\mathcal{A}_1, \ldots, \mathcal{A}_d$ in exploratory subgroup analysis. The subgroup selection bias refers to the fact that the subgroups with the highest observed treatment effects have the tendency to overestimate the effects. Take the subgroup with the observed highest treatment effect for example, we would expect that $\widehat{\tau}_{(1)}$ is an overly optimistic estimate of $\tau_{(1)}$, and the resulting bias $\mathbb{E}\left[\widehat{\tau}_{(1)} - \tau_{(1)}\right]$ represents the subgroup selection bias.

To demonstrate the selection bias issue, we consider a toy example in which the treatments are randomly assigned. We take 1000 Monte Carlo samples from

$$Y_i = 0.5 + \tau_j \sum_{j=1}^{d} \mathbf{1}\left(X_i \in \mathcal{A}_j\right) T_i + X_{i,1} + X_{i,2} + \varepsilon_i,$$

$$T_i \sim \text{Bernoulli}(0.5), \quad \varepsilon_i \sim \mathcal{N}(0,1), \quad i = 1, \ldots, n = 200$$

where $X_i = (X_{i,1}, X_{i,2}, X_{i,3})$, with $(X_{i,1}, X_{i,2})$ distributed as bivariate normal with $X_{i,1} \sim \mathcal{N}(0,1)$, $X_{i,2} \sim \mathcal{N}(0,1)$, $\text{Corr}(X_{i,1}, X_{i,2}) = 0.5$, and $X_{i,3} \sim \text{Bernoulli}(0.5)$, independent of $(X_{i,1}, X_{i,2})$. We consider cases with $d \in \{4, 8\}$. When $d = 4$, the sample space of the covariates $\mathcal{X} = \mathbb{R} \times \mathbb{R} \times \{0,1\}$ contains the following subsets: $\mathcal{A}_1 = \{(0,\infty) \times \mathbb{R} \times \{1\}\}$, $\mathcal{A}_2 = \{(-\infty,0) \times \mathbb{R} \times \{1\}\}$, $\mathcal{A}_3 = \{(0,\infty) \times \mathbb{R} \times \{0\}\}$, and $\mathcal{A}_4 = \{(-\infty,0) \times \mathbb{R} \times \{0\}\}$. When $d = 8$, we consider $\mathcal{A}_1 = \{(0,\infty) \times (0,\infty) \times \{1\}\}$, $\mathcal{A}_2 = \{(0,\infty) \times (0,\infty) \times \{0\}\}$, $\mathcal{A}_3 = \{(-\infty,0] \times (0,\infty) \times \{1\}\}$, $\mathcal{A}_4 = \{(-\infty,0] \times (0,\infty) \times \{0\}\}$, $\mathcal{A}_5 = \{(0,\infty) \times (-\infty,0] \times \{1\}\}$, $\mathcal{A}_6 = \{(0,\infty) \times (-\infty,0] \times \{0\}\}$, $\mathcal{A}_7 = \{(-\infty,0] \times (-\infty,0] \times \{1\}\}$, and $\mathcal{A}_8 = \{(-\infty,0] \times (-\infty,0] \times \{0\}\}$. Under this data generating process, we estimate the subgroup treatment effects by the ordinary least squares. When we are interested in the top two subgroups with the highest treatment effects, to understand the over-optimism of $\widehat{\tau}_{(1)}$ and $\widehat{\tau}_{(2)}$, we assume that the treatment has zero effects in all considered subgroups, that is $\tau_j = 0$, for $j = 1, \ldots, d$.

Figure 1 provides the boxplots of $\widehat{\tau}_{(1)}$ and $\widehat{\tau}_{(2)}$ for $d = 4$ and $d = 8$. The results clearly show that both $\widehat{\tau}_{(1)}$ and $\widehat{\tau}_{(2)}$ are inflated estimators of $\tau_{(1)}$ and $\tau_{(2)}$ in this example. The selection bias increases with the number of candidate subgroups under comparison. In fact, the biases (when multiplied by $\sqrt{n}$) are so large that they easily lead to false discoveries in statistical analysis without bias adjustments.

In the above example, we generated nonoverlapping subgroups to demonstrate the selection bias issue. In cases with overlapping subgroups, the magnitude of the selection bias will be affected by the correlation between subgroups, and the subgroup selection bias reduces as the correlation increases. To see this point, we may consider a simple scenario when we have only two subgroups with correlated estimated treatment effects $\text{Corr}\left(\widehat{\beta}_1, \widehat{\beta}_2\right) = \rho$ and their true treatment effects are tied $\beta_1 = \beta_2 = 0$. When ρ is close to one, the two estimated subgroup treatment effects are likely to be very similar (i.e., $\widehat{\beta}_1 \approx \widehat{\beta}_2$), and then there is small bias in using $\max\left\{\widehat{\beta}_1, \widehat{\beta}_2\right\}$.

The selection bias issue has been widely recognized in economics, statistics, and data science at large. Several attempts have been made to address the subgroup selection bias. Some existing procedures based on a plug-in correction of the selection bias are not well grounded (Lee & Shen, 2018), Bayesian methods tend to lack frequentist interpretations (Woody et al., 2022), and simultaneous inference-based approaches tend to be conservative as they aim to control the family-wise error rate for all candidate subgroups (Dezeure et al., 2017; Hall & Miller, 2010). Those conservative simultaneous inference procedures are statistically valid but can be costly in subgroup analysis in that they may have inadequate power to confirm the most promising or most vulnerable subgroup.

If the goal is to perform a significance test on the selected subgroup or to construct an interval estimate of the subgroup treatment effect, it is better to focus on the selected subgroup to ensure that a probability statement can be ensured given the selection without the need to control error rates for all the subgroups. In most applications, the

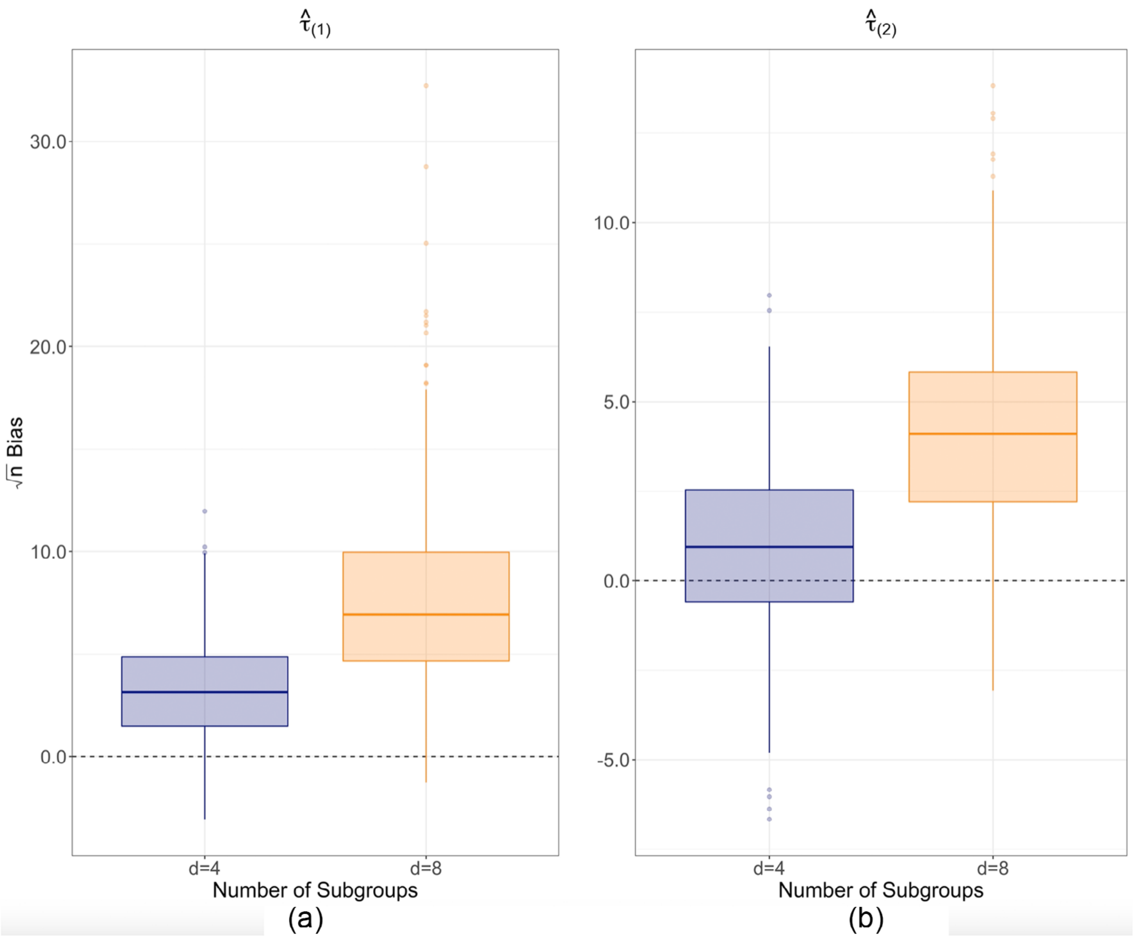

FIGURE 1 Illustration of the selection bias issue. Panel (a) shows the boxplots of $\hat{\tau}_{(1)}$ when $d=4$ and $d=8$. issue. Panel (b) shows the boxplots of $\hat{\tau}_{(2)}$ when $d=4$ and $d=8$.

selected subgroup tends to be the most promising subgroup based on the data, for example, the subgroup with the highest observed treatment effect or odds ratio.

To address the subgroup selection bias issue, we refer to the evaluation of the best-selected subgroups as debiased subgroup analysis, in which the goal is to provide bias-corrected point estimates and valid statistical inference for those subgroups. To be specific, we focus on $d_0 < d$ best subgroups based on observed or estimated treatment effects, and the quantities of interest in debiased subgroup analysis are:

1. the best subgroup treatment effects in the population: $\tau_{(1)}, \ldots, \tau_{(d_0)}$;
2. the observed best subgroup treatment effect sizes: $\tau_{\hat{j}}$, where $\hat{j} = \sum_{k=1}^{d} k \cdot 1\left(\hat{\tau}_k = \hat{\tau}_{(j)}\right)$, for $j = 1, \ldots, d_0$.

In the following section, we shall review some recent methodological advancements that deliver valid statistical inference on the above quantities.

4 | STATISTICAL METHODOLOGIES FOR DEBIASED SUBGROUP ANALYSIS

In what follows, we review four approaches to making valid statistical inference on selected subgroups from a frequentist point of view. At the end of this section, we will provide a discussion about their performance and potential.

4.1 | Calibrated bootstrap

Guo and He (2021) proposed an inference method for $\tau_{(1)}$, the best subgroup treatment effect, based on a calibrated bootstrap approach under two assumptions. The first assumption is that all subgroup treatment effect estimators jointly converge to a normal distribution as the sample size n goes to infinity

$$\sqrt{n}(\hat{\tau}_1-\tau_1,\hat{\tau}_2-\tau_2,\ldots,\hat{\tau}_d-\tau_d) \rightsquigarrow \mathcal{N}(0,\Sigma),$$

for some covariance matrix Σ. The second assumption states that the bootstrap consistently replicates the sampling distribution of the subgroup treatment effect estimators, that is, conditional on the data, the asymptotic distribution of $\sqrt{n}\left(\hat{\tau}_1^*-\hat{\tau}_1,\hat{\tau}_2^*-\hat{\tau}_2,\ldots,\hat{\tau}_d^*-\hat{\tau}_d\right)$ is the same $\mathcal{N}(0,\Sigma)$. These two conditions are readily satisfied under a wide range of commonly used models and parameters.

The procedure starts with calculating a calibration term $d_j=(1-n^{r-0.5})\left(\hat{\tau}_{(1)}-\hat{\tau}_j\right)$ for each subgroup $j=1,\ldots,d$, where $r\in(0,0.5)$ is either a user-specified constant or chosen adaptively by some form of cross-validation (see Algorithm 2 in Guo and He (2021)). Then, one generates bootstrap replicates $\left(\hat{\tau}_1^*,\ldots,\hat{\tau}_d^*\right)$ of $(\hat{\tau}_1,\ldots,\hat{\tau}_d)$ based on an appropriate bootstrap procedure (e.g., by sampling from all the subjects with replacement), and calculates the centered bootstrap statistic

$$T^*=\sqrt{n}\left(\max_{j=1,\ldots,d}\left(\hat{\tau}_j^*+d_j\right)-\hat{\tau}_{(1)}\right). \tag{1}$$

Based on B bootstrap samples (say, $B=1000$), one obtains B independent centered bootstrap statistics T_b^* $(b=1,2,\cdots,B)$. The level $(1-\alpha)$ one-sided confidence interval for $\tau_{(1)}$ is given as

$$\left[\hat{\tau}_{(1)}-\hat{q}_\alpha/\sqrt{n},\ +\infty\right). \tag{2}$$

where $\hat{q}_\alpha$ is the $(1-\alpha)$-quantile of T_b^* $(b=1,2,\cdots,B)$. A consistent bias-reduced estimator of $\tau_{(1)}$ can be obtained via

$$\hat{\tau}_{(1),\text{reduced}}=\hat{\tau}_{(1)}-\frac{1}{B}\sum_{b=1}^{B}T_b^*/\sqrt{n}.$$

Theorem 1 of Guo and He (2021) confirms that, for any choice of $r\in(0,0.5)$ the above inference procedure for $\tau_{(1)}$ is asymptotically sharp, that is, the proposed confidence interval in Equation (2) achieves the exact nominal level of coverage, $1-\alpha$, as the sample size goes to infinity. This property makes the method more attractive than simultaneous inference or uniform inference methods over all subgroups, because the latter tends to be conservative leading to a loss of power in detecting treatment effects of the selected subgroup.

The constant r in the procedure can be viewed as tuning. In fact, the procedure is closer to uniform inference over all subgroups if r approaches zero, and the case of $r=0.5$ corresponds to direct (but invalid) inference on the selected subgroup without accounting for the selection. Further studies on the effect of the tuning may help develop better data-adaptive choices of r. For an example R code of using the method of Guo and He (2021), see https://github.com/xinzhoug/Data/tree/MONET1-Trial.

4.2 | Resampling and tie sets identification

In the context of meta-analysis, Claggett et al. (2014) proposed a method that delivers valid inference on ordered study parameters when independent estimates of those parameters are available. Wei, Zhou, et al. (2022) extended the approach to more general cases. The generalization enables this approach to be used for debiased subgroup analysis.

Suppose that all the subgroup treatment effect estimators jointly converge to a multivariate normal distribution

$$\sqrt{n}(\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau}) = \sqrt{n}(\widehat{\tau}_1 - \tau_1, \widehat{\tau}_2 - \tau_2, \ldots, \widehat{\tau}_d - \tau_d) \rightsquigarrow \mathcal{N}(0, \Sigma),$$

and the asymptotic covariance matrix can be consistently estimated as $\widehat{\Sigma} = \Sigma + o_p(1)$, so that $\widehat{\Sigma}^{-1/2} \times \sqrt{n}(\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \rightsquigarrow \mathcal{N}(0, I)$, where I is a d-dimensional identity matrix.

The procedure starts with generating replicates $\widehat{\boldsymbol{\tau}}^* = \left(\widehat{\tau}_1^*, \ldots, \widehat{\tau}_d^*\right)'$ of $\widehat{\boldsymbol{\tau}}$ from the multivariate normal distribution

$$\widehat{\boldsymbol{\tau}}^* \mid \text{Data} \sim \mathcal{N}\left(\widehat{\boldsymbol{\tau}}, \widehat{\Sigma}/n\right), \tag{3}$$

and sort the elements in $\widehat{\boldsymbol{\tau}}^*$ as $\widehat{\tau}_{(1)}^* \geq \ldots \geq \widehat{\tau}_{(d)}^*$. Next, given properly chosen b_L and b_R so that $b_L - b_R = O(n^{-\delta})$ with $\delta \in \left(0, \frac{1}{2}\right)$, one can estimate a "near tie" set that captures subgroups that have similar effect sizes to subgroup j:

$$\widehat{\mathcal{H}}_{(j)} = \left\{k : \widehat{\tau}_{(j)}^* - b_L \leq \widehat{\tau}_k^* \leq \widehat{\tau}_{(j)}^* + b_R,\ k = 1, \ldots, d\right\}.$$

We refer to supporting information section C.1 of Wei, Zhou, et al. (2022) for data-adaptive choices of b_L and b_R. Then, one calculates the averages of $\widehat{\tau}_1^*, \ldots, \widehat{\tau}_d^*$ and of $\widehat{\tau}_1, \ldots, \widehat{\tau}_d$ in the estimated tie set $\widehat{\mathcal{H}}_{(j)}$ as

$$\widetilde{\tau}_{(j)}^* = \frac{\sum\limits_{k \in \widehat{\mathcal{H}}_{(j)}} \widehat{\tau}_k^*}{\mid \widehat{\mathcal{H}}_{(j)} \mid}, \text{and} \widetilde{\tau}_{(j)} = \frac{\sum\limits_{k \in \widehat{\mathcal{H}}_{(j)}} \widehat{\tau}_k}{\mid \widehat{\mathcal{H}}_{(j)} \mid}, \tag{4}$$

where $\mid \widehat{\mathcal{H}}_{(j)} \mid$ denotes the cardinality of the set $\widehat{\mathcal{H}}_{(j)}$.

For confidence interval construction, one generates a sequence of B independent samples of $\widetilde{\tau}_{(j)}^*$ as in Equation (4), and then define $\widehat{q}_{(j)}(u)$ to be the empirical u-quantile of the sequence. The level $1 - \alpha$ confidence interval for $\tau_{(j)}$ is given by

$$\left[\widehat{q}_{(j)}(\alpha/2),\ \widehat{q}_{(j)}(1 - \alpha/2)\right],\quad j = 1, \ldots, d_0.$$

Wei, Zhou, et al. (2022) show that the above confidence interval serves as an asymptotically exact level-α interval estimate for $\tau_{(j)}$ and $\tau_{\widehat{j}}$. For debiased point estimates, one can use either $\widetilde{\tau}_{(j)}$ or the average of resampled statistics $\widetilde{\tau}_{(j)}^*$ for $\tau_{(j)}$ and $\tau_{\widehat{j}}$. The implementation code of the above discussed approach is publicly available on the authors' website at https://github.com/zyuqing1125/desubgb.

We note that a sufficient condition for the discussed approaches in Sections 4.1 and 4.2 is that subgroup treatment effect estimates are asymptotically linear (and Gaussian). This assumption holds in a wide range of settings when the covariate dimension is small relative to the sample size. The assumptions are not to be tested in practice, because they are about the large-sample behavior of the estimators of effect size being used.

4.3 | Testing many moment inequalities

Some work in the econometrics literature can be adapted for debiased inference for selected subgroup effects. Chernozhukov et al. (2013, 2019) consider the problem of testing the null hypothesis

$$H_0 : \tau_j \leq 0, \quad \text{for all}\, j = 1, \ldots, d. \tag{5}$$

The inequalities $\tau_j \leq 0$, for $j = 1, \ldots, d$ can be called moment inequalities, and they have been studied for inference on causal and structural parameters in partially identified models in econometrics.

Note that the moment inequalities literature focuses on deriving powerful tests for the null hypotheses defined in Equation (5) and does not aim to provide direct point estimates of $\tau_{(j)}$; such null hypotheses, however, can be of interest for debiased subgroup analysis. For example, suppose a positive value of τ_j suggests the treatment is effective for subgroup j. Then the rejection of the null hypothesis in Equation (5) confirms (statistically) that the best subgroup treatment is effective. For this reason, we will briefly review the test proposed by Chernozhukov et al. (2019) to broaden the toolset of debiased subgroup analysis.

Consider the maximum of the studentized statistics:

$$T = \max_{1 \le j \le d} \frac{\sqrt{n}\hat{\tau}_j}{\hat{\sigma}_j},$$

where $\hat{\sigma}_j^2$ is the estimated variance of $\sqrt{n}\hat{\tau}_j$. Larger values of T indicate that H_0 is likely to be violated, so the critical region of the test takes the form $T > c_n$, where c_n is a critical value. Various approaches for computing the critical value c_n have been proposed. Here we describe the two-step bootstrap approach which tends to be less conservative than the one-step approach discussed in their manuscript when the moment inequalities are strict (see comment 4.3 in Chernozhukov et al. (2019)).

Let $\beta_n \in (0, \alpha/2)$ be a tuning parameter, and define the set

$$\hat{J}_B := \left\{ j \in \{1, \ldots, d\} : \ \sqrt{n}\hat{\tau}_j / \hat{\sigma}_j > -2c(\beta_n) \right\},$$

where $c(\beta_n)$ is obtained data-adaptively from Algorithm EB or MB in section 4.2.1. of Chernozhukov et al. (2019). Intuitively, the set $\hat{J}_B$ aims to find subgroups whose τ_j are greater than some threshold on which the critical value is based. Next, one generates bootstrap replicates $(\hat{\tau}_1^*, \ldots, \hat{\tau}_d^*)$ of $(\hat{\tau}_1, \ldots, \hat{\tau}_d)$ and obtains

$$W^* = \begin{cases} \max\limits_{j \in \hat{J}_B} \dfrac{\sqrt{n}\left(\hat{\tau}_j^* - \hat{\tau}_j\right)}{\hat{\sigma}_j} & \text{if } \hat{J}_B \text{ is not empty} \\ 0 & \text{if } \hat{J}_B \text{ is empty.} \end{cases}$$

Finally, the critical value c_n is calculated as the $(1 - \alpha + 2\beta_n)$-quantile of W^* based on a sequence of bootstrapped samples.

4.4 | Conditional (or selective) inference

Recent work by Andrews et al. (2019, 2022) considered performing conditional and unconditional inference on observed best policies. Their approach focuses on delivering sharp inference on the selected subgroup treatment effects conditional on these subgroups having been selected. Following the selective inference literature (Lee et al., 2016; Taylor & Tibshirani, 2015), we write the goal as finding a confidence interval C_1 so that

$$\lim_{n \to \infty} \mathbb{P}\left(\tau_{\hat{\theta}} \in C_1 \mid \text{subgroup } j \text{ is selected as the best one} \right) = 1 - \alpha, \quad \text{for} \quad j = 1, \ldots, d,$$

where $\hat{\theta}$ is the index of the selected best subgroup.

To provide some heuristics of the conditional inference approach, we describe this approach in a simplified scenario with two nonoverlapping subgroups being compared. To this end, suppose that their treatment effect estimators follow the bivariate normal distribution with a known covariance matrix:

$$\begin{pmatrix} \hat{\tau}_1 \\ \hat{\tau}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right).$$

Suppose further that the first subgroup is selected and the observed subgroup treatment effect in the second subgroup is $\widehat{\tau}_2 = a_2$. Then, conditional on the selection event that subgroup 1 is selected, $\widehat{\tau}_1$ (which is also $\tau_{\widehat{\theta}}$) follows a normal distribution truncated below at a_2, that is

$$\widehat{\tau}_1 \mid \text{subgroup 1 is selected as the best} \overset{\text{d.}}{=} \widehat{\tau}_1 \mid \widehat{\tau}_1 > a_2 \sim \mathcal{TN}\left(\tau_1, \sigma_1^2; a_2, +\infty\right).$$

Therefore, to construct the confidence interval C_1, we only need to find the quantiles of the truncated normal distribution $\mathcal{TN}\left(\tau_1, \sigma_1^2; a_2, +\infty\right)$. When subgroup 2 is selected, the confidence interval C_1 can be constructed in a similar fashion. In this setting, the confidence interval C_1 will have an exact asymptotic conditional coverage probability $1-\alpha$. By the law of the iterated expectations, we conclude that C_1 also has unconditional coverage probability $1-\alpha$, that is

$$\lim_{n\to\infty} \mathbb{P}\left(\tau_{\widehat{\theta}} \in C_1\right) = 1-\alpha.$$

When the subgroup treatment effect estimators are correlated across subgroups and when the covariance matrix of the estimators is unknown, the conditional distribution of $\tau_{\widehat{\theta}}$ is no longer as simple. At this point, we do not have much experience in how the selective inference approach compares with the others.

5 | DISCUSSION AND OPEN PROBLEMS

In this article, we have discussed two sources of statistical bias in the analysis of adaptive experiments and in post hoc subgroup analysis, respectively. Awareness of possible bias due to adaptive data collection and data snooping has led to ongoing research in debiased inference, some of which has been briefly reviewed here.

For adaptive experiments, we have reviewed three data collection mechanisms including complete randomization, covariate adaptive randomization, and response adaptive randomization. Complete randomization removes bias by design but unbalanced covariate distributions between the treated and controlled arms may lead to increased variability. Covariate adaptive randomization can reduce such variability by adaptively balancing the covariate distributions when participants are sequentially enrolled. Response-adaptive designs can be used to help accelerate the adoption of better treatments but require appropriate debiased inference methods.

For debiased subgroup analysis, we have reviewed four methods from a frequentist point of view. For the calibrated bootstrap, the tie set identification, and the moment inequality approach, the selection of a tuning parameter is important for achieving good performance in finite samples. Direct comparisons of those methods need to take the tuning parameter into account. In practice, the candidate set of tuning parameters can be prefixed before carrying out the debiased subgroup analysis. In addition to the methods reviewed here for conducting debiased subgroup analysis on the selected best subgroups, we certainly have simultaneous inference procedures in the statistics literature. Those methods aim to cover all subgroups and therefore tend to be conservative and consequently may have inadequate power to confirm the most promising subgroups. Testing the null hypothesis that there is no subgroup treatment effect at all is one way to avoid simultaneous inference. We refer to Shen and He (2015) and Sun et al. (2022) for some work in this direction. Bayesian methods can be adopted to remove the selection bias as well. For example, from an empirical Bayes point of view, Efron (2011) proposed a method to handle the winner's curse bias with Tweedie's formula. More recent work by Woody et al. (2022) and Woody (2020) studied the selection bias issue and postselection inference from a Bayesian perspective.

Our review of adaptive experiments is certainly not comprehensive. For example, subgroup analysis in adaptive experiments is implicitly performed in some clinical trials, such as adaptive enrichment designs that allow practitioners to use interim data to identify treatment-sensitive patient subgroups by sequentially changing patient enrollment criteria; see some frequentist enrichment designs in Simon and Simon (2013); Lai et al. (2019); Götte et al. (2015); Stallard (2022) for example.

This article reviews a number of debiased inference methods that are recently developed for subgroup analysis and for adaptive designs. Going forward, we expect more research in those areas. For example, the validity of inference on selected subgroup effect size has been proven with predefined candidates of subgroups but more needs to be done for more flexible and possibly more greedy search of subgroups from the data. Data-adaptive tuning parameters also need better development and more testing for the inferential methods discussed in this article. If adaptive designs are used, whether it is covariate adaptive or response adaptive or both, subgroup identification and inference needs to account for the characteristics of the design. Debiased inference for selected subgroups with data from adaptive designs is little developed so far and calls for future research.

AUTHOR CONTRIBUTIONS

Jingshen Wang: Investigation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Xuming He:** Conceptualization (equal); methodology (equal); project administration (equal); writing – review and editing (equal).

ACKNOWLEDGMENTS

We thank Dr. Feifang Hu for helpful discussions on adaptive clinical trials. The research was partially supported by the NSF Award DMS-1914496 and DMS-2015325.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Xuming He https://orcid.org/0000-0002-3442-4173

RELATED WIREs ARTICLES

Subgroup identification for precision medicine: A comparative review of 13 methods

REFERENCES

Agarwal, A., Jiang, N., Kakade, S. M., & Sun, W. (2022). Reinforcement learning: Theory and algorithms. *Preprint*, pages 1–193.

Alemayehu, D., Chen, Y., & Markatou, M. (2018). A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. *Statistical Methods in Medical Research*, *27*(12), 3658–3678.

Andrews, I., Bowen, D., Kitagawa, T., & McCloskey, A. (2022). Inference for losers. *AEA Papers and Proceedings*, *112*, 635–642.

Andrews, I., Kitagawa, T., & McCloskey, A. (2019). *Inference on winners*. Technical report, National Bureau of Economic Research.

Athey, S., Bickel, P. J., Chen, A., Imbens, G., & Pollmann, M. (2021). *Semiparametric estimation of treatment effects in randomized experiments*. Technical report, National Bureau of Economic Research.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113* (27), 7353–7360.

Atkinson, A. C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, *69*(1), 61–67.

Azevedo, E. M., Deng, A., Montiel Olea, J.'e. L., Rao, J., & Weyl, E. G. (2020). A/B testing with fat tails. *Journal of Political Economy*, *128*(12), 4614–000.

Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2021). Doubleml–an objectoriented implementation of double machine learning in r. *arXiv preprint arXiv:2103.09603*.

Berry, D. A. (2012). Adaptive clinical trials in oncology. *Nature Reviews Clinical Oncology*, *9*(4), 199–207.

Bibaut, A., Dimakopoulou, M., Kallus, N., Chambaz, A., & van der Laan, M. (2021). Post-contextual-bandit inference. *Advances in Neural Information Processing Systems*, *34*, 28548–28559.

Bowden, J., & Trippa, L. (2017). Unbiased estimation for response adaptive clinical trials. *Statistical Methods in Medical Research*, *26*(5), 2376–2388.

Brankovic, M., Kardys, I., Steyerberg, E. W., Lemeshow, S., Markovic, M., Rizopoulos, D., & Boersma, E. (2019). Understanding of interaction (subgroup) analysis in clinical trials. *European Journal of Clinical Investigation*, *49*(8), e13145.

Bucher, C. G. (1988). Adaptive sampling—An iterative fast Monte Carlo procedure. *Structural Safety*, *5*(2), 119–126.

Bugni, F. A., Canay, I. A., & Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, *113*(524), 1784–1796.

Chernozhukov, V., Chetverikov, D., & Kato, K. (2019). Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies*, *86*(5), 1867–1900.

Chernozhukov, V., Lee, S., & Rosen, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica*, *81*(2), 667–737.

Claggett, B., Xie, M., & Tian, L. (2014). Meta-analysis with fixed, unknown, study-specific parameters. *Journal of the American Statistical Association*, *109*(508), 1660–1671.

Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G., & Pangalos, M. N. (2014). Lessons learned from the fate of AstraZeneca's drug pipeline: A five dimensional framework. *Nature Reviews Drug Discovery*, *13*(6), 419–431.

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, *221*(2), 644–654.

Dezeure, R., Bühlmann, P., & Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, *26*(4), 685–719.

Dmitrienko, A., & D'Agostino, R., Sr. (2013). Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*, *32*(29), 5172–5218.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, *58*(3), 403–417.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, *106*(496), 1602–1614.

Etor'e, P., & Jourdain, B. (2010). Adaptive optimal allocation in stratified sampling methods. *Methodology and Computing in Applied Probability*, *12*(3), 335–360.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, *75*(1), 259–276.

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30* (24), 2867–2880.

Giessing, A., & Wang, J. (2021). Inference on heterogeneous quantile treatment effects via rank-score balancing. *arXiv preprint arXiv: 2102.01753*.

Götte, H., Donica, M., & Mordenti, G. (2015). Improving probabilities of correct interim decision in population enrichment designs. *Journal of Biopharmaceutical Statistics*, *25*(5), 1020–1038.

Guo, X., & He, X. (2021). Inference on selected subgroups in clinical trials. *Journal of the American Statistical Association*, *116*(535), 1498–1506.

Guo, X., Wei, W., Liu, M., Cai, T., Wu, C., & Wang, J. (2022). Assessing heterogeneous risk of type ii diabetes associated with statin usage: Evidence from electronic health record data. *arXiv preprint arXiv:2205.06960*.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., & Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, *118*(15), e2014602118.

Hahn, J., Hirano, K., & Karlan, D. (2011). Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, *29*(1), 96–108.

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, *15*(3), 965–1056.

Hall, P., & Miller, H. (2010). Bootstrap confidence intervals and hypothesis tests for extrema of parameters. *Biometrika*, *97*(4), 881–892.

Hasselt, H. (2010). Double q-learning. *Advances in Neural Information Processing Systems*, 2613–2621.

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

Hu, F., & Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials* (Vol. 525). John Wiley & Sons.

Hu, J., Zhu, H., & Hu, F. (2015). A unified family of covariate-adjusted response-adaptive designs based on efficiency and ethics. *Journal of the American Statistical Association*, *110*(509), 357–367.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, *7*(1), 443–470.

Kapelner, A., & Krieger, A. (2014). Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*, *70*(2), 378–388.

Kasy, M., & Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, *89*(1), 113–132.

Kubota, K., Ichinose, Y., Scagliotti, G., Spigel, D., Kim, J. H., Shinkai, T., Takeda, K., Kim, S.-W., Hsia, T.-C., & Li, R. K. (2014). Phase III study (MONET1) of motesanib plus carboplatin/paclitaxel in patients with advanced nonsquamous nonsmall-cell lung cancer (NSCLC): Asian subgroup analysis. *Annals of Oncology*, *25*(2), 529–536.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Meta-learners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.

Lai, T. L., Lavori, P. W., & Tsang, K. W. (2019). Adaptive enrichment designs for confirmatory trials. *Statistics in Medicine*, *38*(4), 613–624.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44* (3), 907–927.

Lee, M. R., & Shen, M. (2018). Winner's curse: Bias estimation for total effects of features in online controlled experiments. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 491–499.

Li, X., & Ding, P. (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(1), 241–268.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, *7*(1), 295–318.

Liu, Y., Ma, X., Zhang, D., Geng, L., Wang, X., Zheng, W., & Chen, M.-H. (2019). Look before you leap: Systematic evaluation of tree-based statistical methods in subgroup identification. *Journal of Biopharmaceutical Statistics*, *29*(6), 1082–1102.

Loh, W.-Y., Cao, L., & Zhou, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(5), e1326.

Lu, X., Van Roy, B., Dwaracherla, V., Ibrahimi, M., Osband, I., & Wen, Z. (2021). Reinforcement learning, bit by bit. *arXiv preprint arXiv: 2103.04047*.

Ma, S., & Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, *112* (517), 410–423.

Ma, W., Feifang, H., & Zhang, L. (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association*, *110*(510), 669–680.

Ma, W., Qin, Y., Li, Y., & Feifang, H. (2020). Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association*, *115*(531), 1488–1497.

Ma, X., & Wang, J. (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, *115*(532), 1851–1860.

Nie, X., Tian, X., Taylor, J., & Zou, J. (2018). Why adaptively collected data have negative bias and how to correct for it. In International conference on artificial intelligence and statistics, PMLR, pp. 1261–1269.

Nikishin, E., Schwarzer, M., D'Oro, P., Bacon, P.-L., & Courville, A. (2022). The primacy bias in deep reinforcement learning. In International conference on machine learning, PMLR, pp. 16828–16847.

Peeters, M., Oliner, K. S., Price, T. J., Cervantes, A., Sobrero, A. F., Ducreux, M., Hotko, Y., André, T., Chan, E., Lordick, F., Punt, C. J. A., Strickland, A. H., Wilson, G., Ciuleanu, T. E., Roman, L., van Cutsem, E., He, P., Yu, H., Koukakis, R., … Patterson, S. D. (2015). Analysis of KRAS/NRAS mutations in a phase III study of panitumumab with FOLFIRI compared with FOLFIRI alone as second-line treatment for metastatic colorectal cancer panitumumab plus FOLFIRI and RAS mutations in colorectal cancer. *Clinical Cancer Research*, *21*(24), 5469–5479.

Petticrew, M., Tugwell, P., Kristjansson, E., Oliver, S., Ueffing, E., & Welch, V. (2012). Damned if you do, damned if you don't: Subgroup analysis and equity. *Journal of Epidemiology and Community Health*, *66*(1), 95–98.

Pocock, S. J. (1979). Allocation of patients to treatment in clinical trials. *Biometrics*, *35*, 183–197.

Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, *31*, 103–115.

Qin, Y., Li, Y., Ma, W., & Hu, F. (2016). Pairwise sequential randomization and its properties. *arXiv preprint arXiv:1611.02802*.

Robertson, D. S., Lee, K. M., Lopez-Kolkovska, B. C., & Villar, S. S. (2020). Responseadaptive randomization in clinical trials: From myths to practical considerations. *arXiv Preprint arXiv:2005.00564*.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenberger, W. F., & Lachin, J. M. (2015). *Randomization in clinical trials: Theory and practice*. John Wiley & Sons.

Rosenberger, W. F., & Sverdlov, O. (2008). Handling covariates in the design of clinical trials. *Statistical Science*, *23*(3), 404–419.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, *11*(1), 1–96.

Shao, J., Xinxin, Y., & Zhong, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, *97*(2), 347–360.

Shao, J., & Yu, X. (2013). Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics*, *69* (4), 960–969.

Shen, J., & He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association*, *110*(509), 303–312.

Shin, J., Ramdas, A., & Rinaldo, A. (2019a). On the bias, risk and consistency of sample means in multi-armed bandits. *arXiv preprint arXiv: 1902.00746*.

Shin, J., Ramdas, A., & Rinaldo, A. (2019b). Are sample means in multi-armed bandits positively or negatively biased? *Advances in Neural Information Processing Systems*, *32*, 7102–7111.

Simon, N., & Simon, R. (2013). Adaptive enrichment designs for clinical trials. *Biostatistics*, *14*(4), 613–625.

Stallard, N. (2022). Adaptive enrichment designs with a continuous biomarker. *Biometrics*, *79*, 9–19. https://doi.org/10.1111/biom.13644

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, *10*(Feb), 141–158.

Sun, Y., He, X., & Jianhua, H. (2022). An omnibus test for detection of subgroup treatment effects via data partitioning. *Annals of Applied Statistics*, *16*(4), 2266–2278.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Sverdlov, O., Rosenberger, W. F., & Ryeznik, Y. (2013). Utility of covariate-adjusted response-adaptive randomization in survival trials. *Statistics in Biopharmaceutical Research*, *5*(1), 38–53.

Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics*, *15*(5), 443–453.

Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, *112*(25), 7629–7634.

Thall, P. F., & Wathen, J. K. (2007). Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, *43*(5), 859–866.

Thall, P., Fox, P., & Wathen, J. (2015). Statistical controversies in clinical research: Scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of Oncology*, *26*(8), 1621–1628.

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532.

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

Villar, S. S., & Rosenberger, W. F. (2018). Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking Gittins index rule. *Biometrics*, *74*(1), 49–57.

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

Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M., & Rosenblum, M. (2021). Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association*, 1–12. https://doi.org/10.1080/01621459.2021.1981338

Wathen, J. K., & Thall, P. F. (2017). A simulation study of outcome adaptive randomization in multi-arm clinical trials. *Clinical Trials*, *14*(5), 432–440.

Wei, W., van der Laan, M., Zheng, Z., Wu, C., & Wang, J. (2022). Efficient targeted learning of heterogeneous treatment effects for multiple subgroups in observational studies. *Bometrics*, https://doi.org/10.1111/biom.13800

Wei, W., Zhou, Y., Zheng, Z., & Wang, J. (2022). Inference on the best policies with many covariates. *arXiv preprint arXiv:2206.11868*.

Wei, Y., Liu, L., Xiaogang, S., Zhao, L., & Jiang, H. (2020). Precision medicine: Subgroup identification in longitudinal trajectories. *Statistical Methods in Medical Research*, *29*(9), 2603–2616.

Woody, S. A. *Bayesian approaches for inference after selection and model fitting*. [PhD thesis].

Woody, S., Padilla, O. H. M., & Scott, J. G. (2022). Optimal post-selection inference for sparse signals: A nonparametric empirical Bayes approach. *Biometrika*, *109*(1), 1–16.

Xu, M., Qin, T., & Liu, T.-Y. (2013). Estimation bias in multi-armed bandit algorithms for search advertising. *Advances in Neural Information Processing Systems*, 2400–2408.

Ye, T., Shao, J., Yi, Y., & Zhao, Q. (2022). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association*, 1–13. https://doi.org/10.1080/01621459.2022.2049278

Zeldow, B., Vincent Lo Re, I. I. I., & Roy, J. (2019). A semiparametric modeling approach using bayesian additive regression trees with an application to evaluate heterogeneous treatment effects. *The Annals of Applied Statistics*, *13*(3), 1989.

Zhang, L.-X., Feifang, H., Cheung, S. H., & Chan, W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics*, *35*(3), 1166–1182.

Zhong, Y., Kennedy, E. H., Bodnar, L. M., & Naimi, A. I. (2021). AIPW: An R package for augmented inverse probability–weighted estimation of average causal effects. *American Journal of Epidemiology*, *190*(12), 2690–2699.

Zhou, Q., Ernst, P. A., Morgan, K. L., Rubin, D. B., & Zhang, A. (2018). Sequential rerandomization. *Biometrika*, *105*(3), 745–752. https://doi.org/10.1093/biomet/asy031

How to cite this article: Wang, J., & He, X. (2023). Subgroup analysis and adaptive experiments crave for debiasing. *WIREs Computational Statistics*, *15*(6), e1614. https://doi.org/10.1002/wics.1614