



Inference on Selected Subgroups in Clinical Trials

Xinzhou Guo & Xuming He

To cite this article: Xinzhou Guo & Xuming He (2021) Inference on Selected Subgroups in Clinical Trials, *Journal of the American Statistical Association*, 116:535, 1498-1506, DOI: [10.1080/01621459.2020.1740096](https://doi.org/10.1080/01621459.2020.1740096)

To link to this article: <https://doi.org/10.1080/01621459.2020.1740096>



[View supplementary material](#)



Published online: 17 Apr 2020.



[Submit your article to this journal](#)



Article views: 1908



[View related articles](#)



[View Crossmark data](#)



Citing articles: 4 [View citing articles](#)



Inference on Selected Subgroups in Clinical Trials

Xinzhou Guo and Xuming He

Department of Statistics, University of Michigan, Ann Arbor, MI

ABSTRACT

When existing clinical trial data suggest a promising subgroup, we must address the question of how good the selected subgroup really is. The usual statistical inference applied to the selected subgroup, assuming that the subgroup is chosen independent of the data, may lead to an overly optimistic evaluation of the selected subgroup. In this article, we address the issue of selection bias and develop a de-biasing bootstrap inference procedure for the best selected subgroup effect. The proposed inference procedure is model-free, easy to compute, and asymptotically sharp. We demonstrate the merit of our proposed method by reanalyzing the MONET1 trial and show that how the subgroup is selected post hoc should play an important role in any statistical analysis. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received June 2019

Accepted February 2020

KEYWORDS

Bias correction; Bootstrap;
Sharp inference; Subgroup analysis.

1. Introduction

Subgroup analysis aims to uncover and confirm heterogeneity of treatment effects within a population. In clinical trials, a new treatment might turn out to be marginally effective with the overall study population, but it is often the case that the treatment appears very promising for a subgroup. For example, isosorbide dinitrate and hydralazine hydrochloride (BiDil) was approved by the FDA as an effective treatment for heart failure for African Americans, a subgroup previously noted to have a favorable response (see Brody and Hunt 2006). It was recently found through subgroup analysis that lefitolimod appears effective on patients with extensive-stage small-cell lung cancer in two important subgroups, see the announcement from MOLOGEN (2018). Sun et al. (2012) showed that among the published randomized trials in core medical journals in 2007, 207 of them (44%) contained subgroup analysis results.

How to evaluate the subgroup effect in view of a data-dependent search used to find the subgroup is an interesting statistical question with substantial impacts on the managerial decisions and regulatory deliberations on clinical trials. The question of statistical validity of post-hoc subgroup analysis has become more acute as follow-up trials to confirm a promising subgroup identified from earlier trial data failed frequently. One example to note is the MONET1 study, a study of motesanib plus carboplatin/paclitaxel (C/P) in patients with advanced nonsquamous nonsmall-cell lung cancer (NSCLC). Based on MONET1, East Asian patients were found to be responsive to the treatment (see Kubota et al. 2014). The observed effect size of this subgroup was promising and the drug developer, Amgen, decided to invest additional resources and designed a new trial for this subgroup. However, the follow-up trial (AMG-706) failed to confirm the efficacy of the treatment for the East

Asian subgroup (see Kubota et al. 2017). Therefore, we ask a natural question whether the earlier subgroup analysis was appropriately adjusted for.

In practice, subgroup analysis might be conducted in many different ways, but, as in the MONET1 study, it typically consists of two inter-connected steps: subgroup identification and subgroup confirmation. In the identification step, one looks for the best selected subgroup in the population. The candidate subgroups might come from biological or clinical considerations, expert opinions, or simply a form of data mining applied to the available data. The confirmation step often requires a rigorous statistical inference procedure that accounts for the subgroup identification, and better yet, an additional clinical trial on the identified subgroup. In this article, we focus on statistical inference on the best selected subgroup and propose an approximately de-biased estimate of the subgroup treatment effect as well as a valid confidence bound.

By the best selected subgroup we refer to the subgroup that has the highest observed (or estimated) treatment effect among a predefined set of candidate subgroups under consideration. The best subgroup may be identified through a known subgroup identification method/algorithm. Available methods include machine learning-based algorithms as in Lipkovich et al. (2011) and Su et al. (2009), or model-based methods as in Shen and He (2015) and Fan, Song, and Lu (2017). Whatever the case, the best selected subgroup is associated with a set of competing subgroups, and this set must be specified explicitly or implicitly by the subgroup identification method. If the best subgroup is nonunique, we take any one of them for the purpose of our analysis. After the best subgroup is identified post hoc, one needs to decide how good it really is and whether one should invest additional resources to conduct a clinical trial on the subpopula-

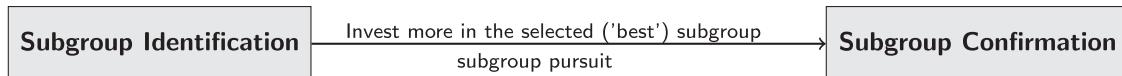


Figure 1. Two-step subgroup analysis.

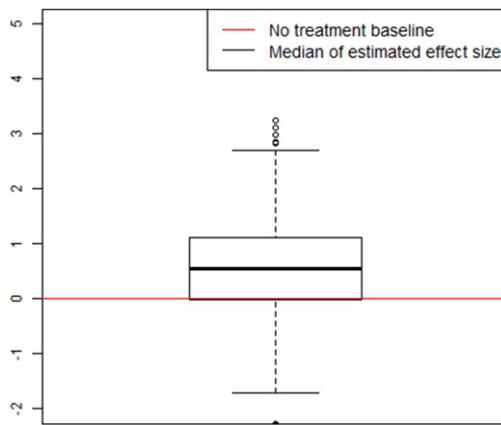


Figure 2. Boxplot of $\max(\hat{\beta}_1, \hat{\beta}_2)$ when $\beta_1 = \beta_2 = 0$ and $\hat{\beta}_1, \hat{\beta}_2 \sim \text{iid } N(0, 1)$.

tion. We refer the decision to invest more in additional clinical trials as the decision of subgroup pursuit, as shown in Figure 1.

Inference on the best selected subgroup identified from the same data suffers from over-optimism and is likely to lead to false discoveries, due to what we shall call subgroup selection bias. For example, a naive approach is to simply use the estimated effect size of the selected subgroup and perform statistical inference conditional on the subgroup. To fix ideas, we consider a toy example consisting of two prespecified subgroups with true treatment effect sizes (e.g., log odds ratio) β_1 and β_2 , respectively. Suppose that the estimated effect sizes are $\hat{\beta}_1 = 0.6$ and $\hat{\beta}_2 = 0.1$, and that a treatment effect above 0.3 is considered statistically and clinically significant. Then, subgroup 1 would be identified as the best selected subgroup with a significant treatment effect. To understand the over optimism of $\max(\hat{\beta}_1, \hat{\beta}_2)$, let us assume both subgroups have no treatment effects, $\beta_1 = \beta_2 = 0$, and $\hat{\beta}_1$ and $\hat{\beta}_2$ are independent and follow the standard normal distribution. Then, Figure 2 gives the boxplot of $\max(\hat{\beta}_1, \hat{\beta}_2)$ based on 2000 random samples. It shows clearly that $\max(\hat{\beta}_1, \hat{\beta}_2)$ is an inflated estimate of $\max(\beta_1, \beta_2)$ in this case. In fact simple calculations show $E[\max(\hat{\beta}_1, \hat{\beta}_2)] \approx 0.6$. It means that even under this very unfavorable situation for subgroup pursuit where both subgroups have no treatment effects, we can still observe the best subgroup effect size of 0.6 on average. Therefore, the naive approach is very risky for subgroup pursuit. To make a better-informed decision, an appropriate adjustment to the subgroup selection bias is needed.

In this article, we propose a resampling-based method to address subgroup selection bias. To be specific, we develop a bias-reduced estimator and a valid one-sided confidence bound on the selected subgroup effect size as measured by log-odds ratio, for instance. Even though the standard bootstrap method does not estimate the bias correctly, we use the bootstrap to learn about the bias and develop an appropriate procedure for

bias correction. Our proposed method is model-free, easy to compute and provides asymptotically sharp inference.

Subgroup selection bias is quite well-recognized in subgroup analysis as a fundamental challenge for inference on the selected subgroup effect (see, e.g., Magnusson and Turnbull 2013; Thomas and Bornkamp 2017). Some attempts have been made to address the issue. Fuentes, Casella, and Wells (2018) and Hall and Miller (2010) proposed valid inference based on simultaneous controls so the resulting inference procedures tend to be conservative. Some ad hoc methods to correct for the bias have been suggested in Rosenkranz (2016) and Stallard, Todd, and Whitehead (2008), among others, but those methods lack theoretical justifications. Bornkamp et al. (2017) and Woody and Scott (2018) considered Bayesian inference, which is clearly model-dependent. As far as we know, model-free and asymptotically sharp inference on the best selected subgroup has been lacking, and the purpose of our work is to bridge this gap and help users make a better-informed decision on subgroup pursuit.

The remainder of this article is organized as follows. In Section 2, we propose a bootstrap-based confidence bound and a bias-reduced estimator for the best selected subgroup effect when the subgroups are predefined. In Section 3, we generalize the proposed inference procedure to accommodate subgroups that are identified post hoc. In Section 4, we analyze synthetic data that mimic the MONET1 study to show how the proposed method can make a better-informed decision on subgroup pursuit in such a case study. In Section 5, we study the finite sample performance of the proposed method by simulation. In Section 6, we give a summary of our work with some concluding remarks.

2. Inference With Predefined Subgroups

In this section, we focus on a relatively simple scenario in subgroup analysis where a small number of candidate subgroups are predefined. We propose bootstrap-based asymptotically sharp inference and a bias-reduced estimator on the effect size of the best selected subgroup.

2.1. Problem Setting

We consider the problem of k (possibly overlapped) subgroups with β_i and $\hat{\beta}_i$ as the effect size and the observed effect size of the i th subgroup, respectively, for $i = 1, \dots, k$. The subgroups are usually defined by baseline characteristics of the subjects. We assume k is a fixed constant, but the total sample size for the trial is n . We also assume that the data include n_i subjects in the i th subgroup, and $\sum_{i=1}^k n_i \geq n$, where equality occurs only when the k subgroups are mutually exclusive. In any subsequent asymptotic analysis, we assume that n_i/n is bounded away from 0 and 1, as the sample size n increases. At this point, we leave

the specification of the treatment effect to each individual study. It could be a log odds ratio, log hazard ratio, or a simple mean or a regression coefficient, with $\hat{\beta}_i$ estimated from a sample of n_i subjects. Without loss of generality, we assume that a larger value of β_i means a better treatment effect.

Let $[k] = \{1, \dots, k\}$ be the index set. Two quantities of interest in the subgroup analysis are

1. the best selected subgroup effect: β_s , where $s = \operatorname{argmax}_{i \in [k]} \hat{\beta}_i$;
2. the best subgroup effect: $\beta_{\max} = \max_{i \in [k]} \beta_i$.

Note that β_{\max} is a fixed parameter, whereas β_s is the true effect size of the selected subgroup. One may debate which quantity should be used for subgroup pursuit decisions, and our proposed inference method works for both quantities. We will start from inference on β_{\max} and show that the same procedure works for inferring on β_s .

In the cases with $k = 2$ and when $\hat{\beta}_i, i = 1, 2$, are jointly normally distributed, the statistic $\hat{\beta}_{\max} = \max_{i \in [k]} \hat{\beta}_i$ has a skew-normal distribution (see Nadarajah and Kotz 2008). However, the skew-normal distribution has unknown parameters, and if those parameters are replaced by their best possible estimates with root- n rate of convergence, any inference based on the estimated skew-normal distribution does not lead to valid inference. Of course, the problem does not become less challenging when $k > 2$, which calls for a new inferential method to be developed.

2.2. Proposed Method

We propose the following bootstrap-based method to construct a lower confidence limit for β_{\max} for any $k \geq 2$. The method has a tuning parameter $r \in (0, 0.5)$, and uses the estimated subgroup effects $\hat{\beta}_i$ and their maximum value $\hat{\beta}_{\max}$.

Suppose that the data consist of independent observations $\{D_j, Z_j\}$ from $j = 1, \dots, n$ subjects, where D_j represents treatment and outcome measures, and $Z_j \subset [k]$ indicates which subgroup or subgroups subject j belongs to. We may use the bootstrap sample $\{D_j^*, Z_j^*\}, j = 1, \dots, n$, by drawing n subjects with replacements. The subgroup treatment effects for the bootstrapped sample are then denoted by $\hat{\beta}_i^*$ for $i = 1, \dots, k$. Depending on the specific model being used to calculate the treatment effects, other bootstrap methods might be used, so long as some bootstrap consistency results are satisfied as specified in the next subsection. With the bootstrap samples at hand, the proposed method proceeds with the following algorithm.

Algorithm 1 Lower confidence limit for β_{\max}

- 1: For $i = 1, \dots, k$, set $d_i = (1 - n^{r-0.5})(\hat{\beta}_{\max} - \hat{\beta}_i)$;
 - 2: **for** $b = 1, \dots, B$ **do**
 - 3: For bootstrap sample b ; calculate the subgroup effect sizes $\beta_{i,b}^*$, and then $T_b^* = \sqrt{n}(\max_{i \in [k]} (\beta_{i,b}^* + d_i) - \hat{\beta}_{\max})$;
 - 4: **end for**
 - 5: Let $c_\alpha = \operatorname{quantile}(T_b^*, 1 - \alpha)$. The level $1 - \alpha$ lower confidence limit is $\hat{\beta}_{\max} - c_\alpha / \sqrt{n}$.
-

2.3. Asymptotic Validity

Just as $\hat{\beta}_{\max} = \max_{i \in [k]} \hat{\beta}_i$ is a biased estimator of β_{\max} , the bootstrap estimate $\hat{\beta}_{\max}^* = \max_{i \in [k]} \beta_i^*$ for each bootstrap sample is not centered at $\hat{\beta}_{\max}$. The proposed method makes an adjustment to each subgroup effect estimate in the bootstrap sample by the amount d_i , which measures how far the i th subgroup is from the best selected subgroup based on the estimated subgroup effect sizes. The amount of adjustment is greater if $\hat{\beta}_i$ is further away from $\hat{\beta}_{\max}$, and this adjustment enables T_b^* to correct the subgroup selection bias while the usual bootstrap method fails. The modified bootstrap estimate of β_{\max} is

$$\beta_{\max, \text{modified}}^* = \max_{i \in [k]} (\beta_i^* + d_i).$$

To establish the validity of the proposed method, we require asymptotic normality of the subgroup effect estimates as well as their bootstrap estimates at each subgroup. We use P and P^* to denote the probability under the sampling distribution and the bootstrap-induced distribution, respectively.

Assumption 1.1 (Asymptotic normality). $\sqrt{n}(\hat{\beta}_1 - \beta_1, \hat{\beta}_2 - \beta_2, \dots, \hat{\beta}_k - \beta_k)$ is asymptotically normal.

Assumption 1.2 (Bootstrap consistency). $\sqrt{n}(\beta_1^* - \hat{\beta}_1, \beta_2^* - \hat{\beta}_2, \dots, \beta_k^* - \hat{\beta}_k)$ is bootstrap consistent, that is, conditional on the data, the asymptotic distribution of $\sqrt{n}(\beta_1^* - \hat{\beta}_1, \beta_2^* - \hat{\beta}_2, \dots, \beta_k^* - \hat{\beta}_k)$ is the same as the limiting distribution in Assumption 1.1. in probability.

In typical parametric and semiparametric models, Assumption 1.1 is satisfied for a wide range of estimators $\hat{\beta}_i$. Assumption 1.2 is satisfied for most smooth estimators, including the parameter estimates from the proportional hazard models (see Efron and Tibshirani 1994). Our main result is given as follows.

Theorem 1. Under Assumptions 1.1 and 1.2, and for any $0 < r < 0.5$, we have,

$$\begin{aligned} \sup_{x \in R} |P^*(\sqrt{n}(\beta_{\max, \text{modified}}^* - \hat{\beta}_{\max}) \leq x) \\ - P(\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max}) \leq x)| \rightarrow 0 \end{aligned}$$

as $n \rightarrow \infty$, in probability w.r.t. P .

Theorem 1 confirms that the proposed inference for β_{\max} is asymptotically sharp in the sense that the proposed confidence bound in Algorithm 1 will achieve the exact nominal level as the sample size goes to infinite under very mild assumptions, which distinguishes the proposed inference from conservative methods. The following corollary facilitates inference on β_s .

Corollary 1. Under Assumptions 1.1 and 1.2, and for any $0 < r < 0.5$, we have

$$\begin{aligned} \sup_{x \in R} |P^*(\sqrt{n}(\beta_{\max, \text{modified}}^* - \hat{\beta}_{\max}) \leq x) \\ - P(\sqrt{n}(\hat{\beta}_{\max} - \beta_s) \leq x)| \rightarrow 0, \end{aligned}$$

as $n \rightarrow \infty$, in probability w.r.t P .

Corollary 1 indicates that the proposed bootstrap-based confidence interval for β_{\max} can also serve as an asymptotically sharp prediction interval for β_s . Therefore, we can use the same procedure to infer on the best and the best selected subgroup effect in subgroup pursuit, without having to choose which quantity to focus on. The remaining issue with the proposed method is the tuning parameter r . In theory, it can be any positive value less than 1/2 but we defer the discussion on the practical choices of the tuning parameter to Section 2.5.

2.4. Bias-Reduced Estimator

Following the results in Theorem 1, we propose a bias-reduced estimator for β_{\max} , and a biased-reduced predictor for β_s . From Lemma 2.1 in Appendix A.1 of the supplementary materials, we see that under regularity conditions, $E[\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max})]$ is asymptotically equivalent to $E[\max_{i \in H} \sqrt{n}(\hat{\beta}_i - \beta_i)]$ where $H = \{i : \beta_i = \beta_{\max}\}$. Therefore, there is a bias $E[\hat{\beta}_{\max} - \beta_{\max}]$ in the order of $O(1/\sqrt{n})$ when the number of the best subgroups is greater than 1 (e.g., $\beta_1 = \beta_2$ in the case of two subgroups), because $E[\max_{i \in H} \sqrt{n}(\hat{\beta}_i - \beta_i)]$ converges to the mean of an asymmetric distribution (e.g., skew normal in the case of $K = 2$ as studied in Nadarajah and Kotz (2008)). To be more specific, the bias is nonnegligible for inference if the size of H , $|H|$, is greater than 1.

We propose a bias-reduced estimator $\hat{\beta}_{\max,\text{reduced}}$ as follows.

$$\hat{\beta}_{\max,\text{reduced}} = \hat{\beta}_{\max} - E^*[\beta_{\max,\text{modified}}^* - \hat{\beta}_{\max}],$$

where E^* denotes the expectation under the bootstrap distribution. For a rigorous justification, we need the following two mild assumptions.

Assumption 2.1 (2nd moment bound). $\limsup_{n \rightarrow \infty} E[\sqrt{n}(\hat{\beta}_i - \beta_i)]^2 < \infty$, for $i = 1, \dots, k$.

Assumption 2.2 (2nd bootstrap moment). $\limsup_{n \rightarrow \infty} E^*[\sqrt{n}(\beta_i^* - \hat{\beta}_i)]^2 < \infty$, in probability, for $i = 1, \dots, k$.

Theorem 2. Under Assumptions 1.1, 1.2, 2.1, and 2.2, and for any $0 < r < 0.5$, we have

$$|E^*[\sqrt{n}(\beta_{\max,\text{modified}}^* - \hat{\beta}_{\max})] - E[\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max})]| \rightarrow 0$$

as $n \rightarrow \infty$, in probability w.r.t P .

Theorem 2 confirms that we can use the bootstrap to approximate the bias, $E[\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max})]$, and asymptotically the accuracy of the approximation is $o_P(1/\sqrt{n})$, even when $|H| > 1$. Under a slightly stronger bootstrap 2nd moment condition

Assumption 2.3. $\limsup_{n \rightarrow \infty} E\{E^*[\sqrt{n}(\beta_i^* - \hat{\beta}_i)]^2\} < \infty$, for $i = 1, \dots, k$.

We can have the following result.

Corollary 2. Under Assumptions 1.1, 1.2, 2.1, and 2.3, and for any $0 < r < 0.5$, we have

$$\begin{aligned} |E\{E^*[\sqrt{n}(\beta_{\max,\text{modified}}^* - \hat{\beta}_{\max})]\} - E[\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max})]| \\ \rightarrow 0, \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Corollary 2 implies the following comparisons between $\hat{\beta}_{\max}$ and $\hat{\beta}_{\max,\text{reduced}}$ in terms of bias. If there is only one best subgroup ($|H| = 1$), the biases of $\hat{\beta}_{\max}$ and $\hat{\beta}_{\max,\text{reduced}}$ are both $o(1/\sqrt{n})$. However, if there is more than one best subgroup ($|H| > 1$), the bias of $\hat{\beta}_{\max}$ is $O(1/\sqrt{n})$ while the bias of $\hat{\beta}_{\max,\text{reduced}}$ is reduced to $o(1/\sqrt{n})$.

2.5. Choice of the Tuning Parameter

A small value of the tuning parameter r tends to preserve the coverage probability better in finite samples at the cost of possibly conservative confidence bounds. We suggest a data-adaptive cross-validated choice of r to help practitioners. The basic idea is to choose r to minimize the mean square error between $\hat{\beta}_{\max,\text{reduced}}(r)$ and β_{\max} . To make this possible without knowing the true value of β_{\max} , we provide an approximation to the mean square error that can be computed from the data, and use cross-validation to choose the tuning parameter.

Let $A = \{r_1, \dots, r_m\}$ denote a set of possible tuning parameters in the range of $(0, 0.5)$ with $r_1 < \dots < r_m$ and m is a finite integer. The following algorithm can be used to choose $r \in A$.

Algorithm 2 Cross-validated choice of tuning parameter r

```

1: Randomly partition the data into  $v$  (approximately) equal-sized subsamples;
2: for  $l = 1, \dots, m$  do
3:   for  $j = 1, \dots, v$  do
4:     Basic setup: use the  $j$ th subsample as the reference data and the rest as the training data;
5:     Bias-reduced estimator: use the training data to obtain the bias-reduced estimator of the best subgroup,  $\hat{\beta}_{\max,\text{reduced},j}(r_l)$ , with  $r_l$  as the tuning parameter;
6:     for  $i = 1, \dots, k$  do
7:       Calculations on the reference data: use the reference data to estimate the effect size of the  $i$ th subgroup,  $\hat{\beta}_{i,j}$ , and its standard error  $\hat{\sigma}_{i,j}$ ;
8:       Evaluation of accuracy: calculate  $h_{i,j}(r_l) = (\hat{\beta}_{\max,\text{reduced},j}(r_l) - \hat{\beta}_{i,j})^2 - \hat{\sigma}_{i,j}^2$ ;
9:     end for
10:   end for
11: end for
12: The tuning parameter is chosen to be  $\operatorname{argmin}_{r_l} \{\min_{i \in [k]} [\sum_{j=1}^{j=v} h_{i,j}(r_l)/v]\}$ .

```

To motivate the use of $\min_{i \in [k]} [\sum_{j=1}^{j=v} h_{i,j}(r_l)/v]$ as an approximate objective function for cross-validation, we state the following result.

Theorem 3. Under the assumptions of Corollary 2 and given the set A , there exists an integer, N_A , such that for any $n > N_A$ and $r \in A$, we have

$$E[\hat{\beta}_{\max,\text{reduced},1}(r) - \beta_{\max}]^2 = \min_{i \in [k]} E[\hat{\beta}_{\max,\text{reduced},1}(r) - \beta_i]^2.$$

Theorem 3 implies that minimizing the mean square error of the bias-reduced estimator is asymptotically equivalent to minimizing $\min_{i \in [k]} E[\hat{\beta}_{\max,\text{reduced},1}(r) - \beta_i]^2$ as a function of

$r \in A$. The inclusion of $\hat{\sigma}_{ij}^2$ in the calculation of $h_{ij}(r_l)$ in Step 8 of the above algorithm is to account for the variation in $\hat{\beta}_{ij}$ used there.

3. Inference With Post-Hoc Identified Subgroups

In this section, we consider the cases where the best subgroup is post-hoc identified by searching over many (possibly infinitely many) subgroups. To be more specific, let $\{S(c) : c \in D\}$ denote the family of subgroups, where $S(c)$ is a subgroup indexed by $c \in D$ and D is a compact set in a Euclidean space. Let $\beta(c)$ and $\hat{\beta}(c)$ represent the effect size and the estimated effect size of subgroup $S(c)$, respectively.

To distinguish from the best subgroup effect size defined in the previous section, we use $\gamma_{\max} = \sup_{c \in D} \beta(c)$ as the best subgroup effect and γ_s as the best selected subgroup effect, which is the true effect size of the subgroup that has the highest $\hat{\beta}(c)$ among $c \in D$. We further assume the best selected subgroup is achievable; that is, $\max_{c \in D} \hat{\beta}(c)$ exists almost surely.

3.1. Asymptotically Sharp Inference

We generalize the inference procedure for the predefined subgroups in Section 2.2 to the following algorithm, where $\hat{\gamma}_{\max} = \sup_{c \in D} \hat{\beta}(c)$, and $\beta^*(c)$, $c \in D$, are the estimated effect sizes of the subgroups for a bootstrap sample. As before, we take the tuning parameter as any value $r \in (0, 1/2)$.

Algorithm 3 Lower confidence limit for γ_{\max}

- 1: For $c \in D$, let $d(c) = (1 - n^{r-0.5})(\hat{\gamma}_{\max} - \hat{\beta}(c))$;
 - 2: **for** $b = 1, \dots, B$ **do**
 - 3: For bootstrap sample b ; calculate effect sizes $\beta_b^*(c)$ for $c \in D$, and then $T_b^* = \sqrt{n}(\sup_{c \in D} (\beta_b^*(c) + d(c)) - \hat{\gamma}_{\max})$;
 - 4: **end for**
 - 5: Let $c_\alpha = \text{quantile}(T_b^*, 1 - \alpha)$, the level α lower confidence limit is $\hat{\gamma}_{\max} - c_\alpha/\sqrt{n}$.
-

The bootstrap procedure is based on the modified bootstrap estimator, $\gamma_{\max, \text{modified}}^* = \sup_{c \in D} (\beta^*(c) + d(c))$, where $d(c)$ does not depend on the bootstrap sample. The justification of the above procedure needs the following assumptions.

Assumption 4.1 (Asymptotically Gaussian process). $\sqrt{n}(\hat{\beta}(\cdot) - \beta(\cdot)) \rightarrow^d G(\cdot)$ in $l_\infty(D)$, where $G(\cdot)$ is a Gaussian process with continuous sample path in probability.

Assumption 4.2 (Bootstrap consistency). $\sqrt{n}(\beta^*(\cdot) - \hat{\beta}(\cdot)) \rightarrow^d G(\cdot)$ in $l_\infty(D)$ in probability.

Assumption 4.3 (Continuous mapping). $c \rightarrow \beta(c)$ is a continuous mapping in D .

Theorem 4. Under Assumptions 4.1–4.3 and for any $0 < r < 0.5$, we have as $n \rightarrow \infty$,

$$\begin{aligned} \sup_{x \in R} |P^*(\sqrt{n}(\gamma_{\max, \text{modified}}^* - \hat{\gamma}_{\max})) \leq x) \\ - P(\sqrt{n}(\hat{\gamma}_{\max} - \gamma_{\max}) \leq x)| \rightarrow 0 \end{aligned}$$

in probability w.r.t P .

Theorem 4 implies that the proposed inference is asymptotically sharp. Except the continuous path assumptions for $\beta(c)$ and for $G(c)$, the assumptions required here are the stochastic process version of Assumptions 1.1 and 1.2. If the (bootstrap) estimated effect size can be written in a form of an empirical process, then, Assumptions 4.1 and 4.2 can be often verified by the use of the Donsker class (see Van Der Vaart and Wellner 1996). In other words, these assumptions can be expected to hold in many applications.

Similar to Section 2.4, we can have a bias-reduced estimator of γ_{\max} as

$$\hat{\gamma}_{\max, \text{reduced}} = \hat{\gamma}_{\max} - E^*[\gamma_{\max, \text{modified}}^* - \hat{\gamma}_{\max}].$$

3.2. Selected Subgroup Inference

Previously in the case of predefined subgroups, the inference procedure in Section 2.2 works for both β_{\max} and β_s . This is true because as the sample size goes to infinity, the probability that we select the best subgroup converges to one, which implies $\sqrt{n}(\beta_s - \beta_{\max}) \rightarrow 0$ in probability. However, the almost sure selection cannot be expected for post-hoc identified subgroups in general and we have to take a critical look how we can infer on γ_s .

From the proof of Theorem 4, we see that, asymptotically, the one-sided confidence interval for γ_{\max} is actually based on the one-sided confidence band for $\beta(c)$ on $c \in K$, where $K = \{c : \beta(c) = \sup_{d \in D} \beta(d)\}$ is the set of c values corresponding to the best subgroup effect. More specifically, the critical value, c_α is the $1 - \alpha$ quantile of $\sup_{c \in K} G(c)$ asymptotically. In this sense, we call the interval estimates constructed in Section 3.1 locally simultaneous confidence intervals, in contrasts to any inference based on a (globally) simultaneous confidence band of $\beta(c)$ for all $c \in D$. Because $K \subset D$, the resulting inference is more efficient than the methods based on simultaneous confidence bands such as that of Fuentes, Casella, and Wells (2018).

Furthermore, although γ_s may not equal γ_{\max} with probability one, it falls into a local neighborhood of K , which shrinks to K as sample size increases. This enables us to establish the following result, analogous to Corollary 1 for β_s .

Theorem 5. Under the assumptions of Theorem 4, we have, as $n \rightarrow \infty$,

$$\begin{aligned} \sup_{x \in R} |P^*(\sqrt{n}(\gamma_{\max, \text{modified}}^* - \hat{\gamma}_{\max})) \leq x) \\ - P(\sqrt{n}(\hat{\gamma}_{\max} - \gamma_s) \leq x)| \rightarrow 0 \end{aligned}$$

in probability w.r.t P .

4. Example

In this section, we demonstrate the merit of our proposed method by reanalyzing the failed MONET1 trial. With our proposed method, we can provide an appropriate guidance on subgroup pursuit decisions based on the initial MONET1 trial data.

The purpose of the phase III of MONET1 trial was to confirm the efficacy of an experimental treatment of motesanib plus

carboplatin/ paclitaxel (C/P) in patients with advanced non-squamous nonsmall-cell lung cancer (NSCLC). The trial failed to confirm the overall efficacy, but the East Asian subgroup was found to be highly promising, as reported in Kubota et al. (2014). The MONET1 study reported the hazard ratio, where a hazard ratio of less than 1 is in favor of the treatment. To make this convention consistent with the general treatment earlier in the article, one may simply equate β_i in this article to the negative log-hazard ratio.

The MONET1 trial data showed that for the East Asian subgroup the treatment has the hazard ratio of $HR = 0.669$ and $p\text{-value} = 0.0223$, as reported in Kubota et al. (2014). Predefined subgroups were used in the identification of this subgroup, but we could not find any information on how many and which candidate subgroups were actually considered. The earlier investigation and the existing literature did not pay attention to this question, and consequently ignored the subgroup selection bias in the analysis.

Because the original data from the MONET1 trial were proprietary, we turn to synthetic data that share many of the same characteristics as the MONET1 trial for the case study. To that end, we consider the situations where the number of candidate subgroups ranges from 2 to 16 based on binary coding of some or all of the following variables in the data: East Asian patient, stage IIIB, received radiotherapy, male, age greater than 65, never smoked, ECOG PS status 0, and adenocarcinoma histology. If the first indicator variable of East Asian patient is used, we have two candidate subgroups only (East Asian vs. the others). If each of the eight indicator variables are used, we have a total of 16 subgroups, and they are clearly overlapping. Suppose that the best subgroup is selected from the candidates based on the estimated hazard ratios.

Assuming the subgroups are homogeneous and no treatment effect exists in any subgroup, we generate the synthetic data with the estimated survival function and censoring distribution based on Figure 1.A in Kubota et al. (2014). Additional details for the generation of the synthetic data are given in the Appendix of the supplementary materials.

Now, we have a data-generating model, which enables us to generate a lot of datasets. To mimic MONET1, we focus on one realization with which the East Asian subgroup is selected as the best subgroup among the subgroups we consider and the estimated effect size and p -value of the East Asian subgroup are similar to those reported in Kubota et al. (2014). Table 1 shows the estimated effect size and p -value of the East Asian subgroup from MONET1 reported in Kubota et al. (2014) and the synthetic dataset we use.

With this synthetic dataset, we apply the proposed inference procedure and compare it with the naive method which assumes that the subgroup of East Asians is not selected from the same data (see Table 2). With the naive method for the East Asian subgroup, the hazard ratio of 0.663 is statistically

Table 1. Comparison of hazard ratio and p -value of the best selected subgroup, the East Asian subgroup, between MONET1 study and the synthetic data.

	Hazard ratio	p -value
Synthetic data	0.663	0.019
MONET1	0.669	0.022

Table 2. The bias-reduced estimate and the 95% upper bound of the hazard ratio of the best selected subgroup ($r = 0.03$).

No. of subgroups	2	4	8	10	16	Naive
Upper bound	0.894	0.947	1.012	1.013	1.024	0.883
Hazard ratio	0.711	0.747	0.781	0.790	0.818	0.663

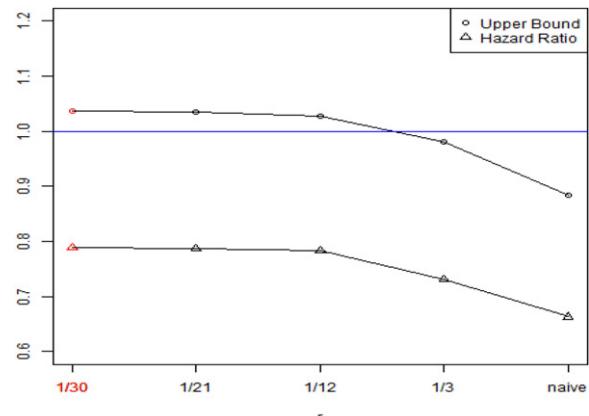


Figure 3. Impact of r : the bias-reduced estimate and the 95% upper bound of the hazard ratio of the best selected subgroup in the case of $k = 8$. The tuning parameter marked with red color indicates the value chosen by the adaptive procedure.

significant. From Table 2, we note that if only two predefined subgroups are considered in the subgroup selection, the 95% upper confidence limit on the hazard ratio is below 1.0, and the subgroup treatment effect is still significant. However, if eight or more candidate subgroups are considered in the selection process, the 95% upper confidence limit on the hazard ratio exceeds 1, implying that the selected subgroup effect is no longer significant. If that is how the East Asian subgroup was identified, our analysis would reach a different conclusion from that of Kubota et al. (2014). Ignoring how the East Asian subgroup was identified would disallow us to evaluate statistical evidence for the selected subgroup.

To see how sensitive the results would be as we choose different values for the tuning parameter r , we refer to Figure 3. When $k = 8$ candidate subgroups were considered, any value of $r \leq 1/12$ or the proposed adaptive method led to the conclusion that the selected subgroup effect is statistically insignificant. As r gets larger, the result would be closer to the naive method, but over a reasonable range of choices for r around the proposed adaptive method we are comforted by the evidence of stability in the analysis.

5. Simulation Study

In this section, we use Monte Carlo simulations to evaluate the finite-sample performance of the proposed method in terms of bias and coverage probabilities. We focus on censored outcomes where the treatment effect is measured by the log hazard ratio from the proportional hazard model. In Sections 5.1 and 5.2, we evaluate the empirical coverage and bias for the predefined subgroups and the post-hoc identified subgroups, respectively. In Section 5.3, we compare the empirical coverage based on the synthetic data generating model used in Section 4.

5.1. Simulation With Predefined Subgroups

To start with, we consider a simple setting consisting of two predefined subgroups. Let D denote the treatment indicator, and random samples of size $n = 400$ are generated from the proportional hazard model with the hazard function $\lambda(t) = \lambda_0(t)e^{\beta_i D}$ for subgroup $i = 1, 2$, respectively, where $\lambda_0(t)$ is the baseline hazard function of Weibull(1, 1), and the parameters β_i are to be specified. The subjects fall into one of the two subgroups with probability 0.5, and the treatment assignment is also random with equal probability. The response generated from the above model is then censored randomly from the right by a censoring variable C , where $\log(C)$ follows the uniform distribution on $(-1.25, 1.00)$. The censoring rate is about 40% across different choices of β_i considered in this study.

In the comparison, we include what we call the naive method, with which we simply select the better subgroup from $\hat{\beta}_i$ and proceed as if the subgroup were selected independent of the data. The performance of the naive method versus the proposed method is affected by the distance between subgroup effects, $|\beta_1 - \beta_2|$. In the study, we fix the effect of subgroup 1 by setting $\beta_1 = 0$ while varying the value of β_2 in $[0, 0.5]$. We use 2000 Monte Carlo samples in evaluating the empirical coverage and average distance from the true value for the 95% lower confidence bound for the selected subgroup effect, β_s , defined in Section 2.1, as well as the empirical bias; see Tables 3, 4, and 5, respectively.

The results show clearly that the naive method falls short in coverage probability, especially when $\beta_2 - \beta_1$ is smaller

than 1/5, and the proposed method preserves the coverage probability much better across a broad range of choices for the tuning parameter r . The data-adaptive choice of the tuning parameter performs quite well; it achieves better coverage and at the same time the distance between the lower confidence limit and the true value does not significantly increase on average compared with that of the naive method. The bias-reduced estimate reduces the bias from around 0.1 for the naive method to around 0.01 for the proposed method. A bias of 0.1 in this case means a roughly 10% relative bias for the hazard ratio estimation.

Next, we evaluate the performance of the proposed method with different numbers of candidate subgroups. Here, we assume there are k subgroups. Following the model used earlier with only two subgroups, the survival time is generated by the proportional hazard model, $\lambda(t) = \lambda_0(t)e^{\beta_i D}$ for $i = 1, \dots, k$, and a subject has equal probability to fall into each subgroup. We use the same treatment assignment and the same censoring scheme as before, but keep the total sample size $n = 200k$. To assess how much the subgroup selection bias might be, we focus on the most challenging scenario with $\beta_1 = \dots = \beta_k = 0$, and calculate the empirical coverage and the empirical bias of the proposed method and the naive method based on 2000 Monte Carlo repetitions. The results are summarized in Table 6.

From Table 6, we see that the coverage probability for the naive method drops below 0.60 when there are 10 subgroups, and the proposed method has slightly lower coverage than the nominal level of 0.95. The results are somewhat more sensitive to the choice of r when the number of subgroups increases, and smaller values of r generally work better. Table 6 also shows that the naive method suffers from the subgroup selection bias and the bias becomes more severe as the number of subgroups increases, while the proposed method can reduce the bias significantly. The adaptive method for choosing the tuning parameter r led to more under-coverage when the number of subgroups is higher, which suggests that additional research is needed to find a more reliable adaptive method across a wide range of problem settings.

5.2. Proportional Hazard Model: Post-Hoc Identified Case

To continue, we consider a post-hoc identified case based on the proportional hazard model. Let D and W denote the treatment indicator and a continuous variable used to define the post-hoc subgroups, respectively, and random samples of size $n = 400$ are generated from the proportional hazard model with the hazard function $\lambda_0(t)e^{b(W)D}$, where $\lambda_0(t)$ is the hazard function

Table 6. Results include “cover”: empirical coverage of the 95% lower bound, and “bias”: empirical bias for β_s : multiple predefined subgroups.

		$r = 1/3$	1/12	1/21	1/30	Naive	Adaptive
$k = 2$	Cover	0.929	0.952	0.953	0.953	0.900	0.939
	Bias	0.029	0.006	0.004	0.004	0.105	0.014
6	Cover	0.891	0.941	0.943	0.945	0.739	0.930
	Bias	0.060	0.011	0.009	0.008	0.240	0.029
10	Cover	0.866	0.944	0.949	0.950	0.594	0.927
	Bias	0.066	0.009	0.006	0.005	0.290	0.031
12	Cover	0.860	0.946	0.950	0.950	0.543	0.925
	Bias	0.062	0.003	0.001	0.001	0.302	0.026

Table 5. Empirical bias for β_s : two predefined subgroups.

	$r = 1/3$	1/12	1/21	1/30	Naive	Adaptive
$\beta_2 = 0$	0.028	0.008	0.007	0.006	0.107	0.018
1/10	0.024	0.002	0.000	-0.001	0.100	0.012
2/10	0.005	-0.021	-0.022	-0.023	0.077	-0.008
3/10	-0.003	-0.045	-0.036	-0.037	0.061	-0.018
4/10	-0.018	-0.063	-0.065	-0.066	0.029	-0.042
5/10	-0.027	-0.067	-0.070	-0.071	0.022	-0.040

Table 7. Empirical coverage of the 95% lower bound of γ_s : post-hoc identified case.

	$r = 1/3$	1/12	1/21	1/30	Naive
$\beta_2 = 0$	0.947	0.961	0.962	0.962	0.872
1/10	0.960	0.972	0.972	0.972	0.879
2/10	0.958	0.966	0.967	0.967	0.890
3/10	0.959	0.969	0.970	0.970	0.895
4/10	0.962	0.968	0.968	0.968	0.906
5/10	0.964	0.972	0.973	0.973	0.901

of Weibull(1, 1), and the function $b(\cdot)$ is to be specified. We assume D and W are independent, D follows Bernoulli(1, 0.5) and W follows Unif[0, 80]. The response generated from the above model is then censored the same as that in Section 5.1. The censoring rate is about 40% across different choices of $b(\cdot)$ considered in this study. We consider the following post-hoc identified subgroups: $S(c) = \{W \leq c\}$, and let $\beta(c)$ denote the subgroup effect of $S(c)$ for $c \in [30, 60]$. From Lin and Wei (1989), we note though given the subgroup, $S(c)$, the event time may not follow the proportional hazard model, $\beta(c)$ is still well-defined. It is also noteworthy that $\beta(c)$ is usually not equal to $b(c)$ but, instead, $\beta(c)$ can be viewed as a weighted average of $b(\cdot)$ in the range [0, c].

In the comparison, we include what we call the naive method where the inference on the best selected subgroup is conducted as if the selection were independent of the data. As pointed out in Sections 1 and 2.4, the performance of the naive method versus the proposed method is affected by whether the subgroups are homogeneous. To change the homogeneity for post-hoc identified subgroups, we consider a simple setting where $b(w) = \begin{cases} \beta_1 & w > 30 \\ \beta_2 & w \leq 30 \end{cases}$. In the study, we fix $\beta_1 = 0$ while varying β_2 in [0, 0.5]. When $\beta_2 = \beta_1$, the post-hoc identified subgroups are homogeneous and the subgroup selection bias is most severe. As β_2 increases, the subgroups are farther away from homogeneity, and the best subgroup, $S(30)$, is more distinctive from the others. We use 2000 Monte Carlo samples in evaluating the empirical coverage for the best selected subgroup effect, γ_s (see Table 7).

From Table 7, we see that for post-hoc identified subgroups, the naive method falls short in coverage probability especially when β_2 is small, and the proposed method preserves the coverage probability much better across a broad range of choices of the tuning parameter. In summary, the proposed method provides trustable inference for the post-hoc identified case in finite samples.

5.3. Synthetic Data Generating Model

We consider a simulation setting based on the synthetic data generating model of MONET1 in Section 4. We focus on the scenario of eight subgroups by the coding of the following variables: East Asian patient, stage IIIB, received radiotherapy, and male. We note that the negative log-hazard ratio of the best selected subgroup, β_s , equals 0 because the synthetic data generating model assumes that the subgroups are homogeneous with no treatment effect. In the comparison, we include the naive method used in Section 5.1. To make it consistent to the convention used in MONET1, we use 2000 Monte Carlo samples

Table 8. Empirical coverage of the 95% upper bound of the log hazard ratio of the best selected subgroup: the synthetic data model.

r	1/3	1/9	1/30	Naive	Adaptive
Coverage	0.917	0.946	0.950	0.805	0.935

in evaluating the empirical coverage of the 95% upper bound for the log hazard ratio of the best selected subgroup in Table 8.

From Table 8, we see that the naive method is once again unable to provide the desired confidence, but the proposed method does well. These results explain the over-optimism in the original study of Kubota et al. (2014); the failure of the subgroup pursuit in MONET1 trial is not just by chance, and the subgroup selection bias deserves accounting for in any serious subgroup analysis.

6. Conclusions

When the best subgroup is selected from the data over a set of candidate subgroups, naive estimation and inference for the treatment effect on the selected subgroup that ignores the selection process leads to bias and over-optimism. The salient point of the present article is that appropriate statistical analysis of the selected subgroup effect size must take the selection process into account. We propose a bias-adjusting bootstrap procedure to infer the best selected subgroup effect. The proposed method is model-free, easy to implement, and the resulting statistical inference is asymptotically sharp, regardless of whether the subgroups are predefined or identified post hoc from the data.

De-biased inference for the best selected subgroup is critical to inform better decision making and help reduce false discoveries in subgroup pursuit in clinical trials. By revisiting the MONET1 trial and its failed follow-up trial, we show that lessons can be learned for future subgroup analysis in clinical work and demonstrate the merit of our proposed method. Our analysis shows that the proposed method can appropriately adjust for the subgroup selection bias and if eight or more subgroups were considered as candidates in the subgroup identification stage in the MONET1 study, we would not have found statistical significance in the East Asian subgroup.

The proposed method aims at inference for the best selected subgroup based on estimated treatment effects for candidate subgroups. In practice, other considerations might be taken into consideration in subgroup identification. The proposed inference method would then serve as a conservative approach to those identified subgroups.

A larger adjustment in the estimation of the subgroup treatment effect is usually needed as we search over more candidate subgroups to find the best subgroup. However, as the number of candidate subgroups increases, there are more overlaps (and thus statistical correlations) across subgroups, so the adjustment size tends to level off quickly when the number of candidate subgroups reaches a threshold. This makes our proposed method not so conservative, and thus practically useful, even if all possible candidate subgroups are taken into account.

Our proposed method relies on subgroup effect estimates that are asymptotically Gaussian. In applications to observational studies where covariate-adjusted estimates are needed

and the potential covariates are of high dimensions, we may need to investigate further how the proposed method adapts. We hope that the present work argues convincingly that valid inference on post hoc identified subgroups needs to be and can be performed effectively with appropriate de-biasing tools in statistics.

Supplementary Materials

The supplement contains the proofs of all the theoretical results and the data generating procedure for the synthetic dataset used in the paper.

Funding

The authors gratefully acknowledge the NSF award DMS-1607840 and National Natural Science Foundation of China grant 11129101.

References

- Bornkamp, B., Ohlssen, D., Magnusson, B. P., and Schmidli, H. (2017), “Model Averaging for Treatment Effect Estimation in Subgroups,” *Pharmaceutical Statistics*, 16, 133–142. [1499]
- Brody, H., and Hunt, L. M. (2006), “BiDil: Assessing a Race-Based Pharmaceutical,” *The Annals of Family Medicine*, 4, 556–560. [1498]
- Efron, B., and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press. [1500]
- Fan, A., Song, R., and Lu, W. (2017), “Change-Plane Analysis for Subgroup Detection and Sample Size Calculation,” *Journal of the American Statistical Association*, 112, 769–778. [1498]
- Fuentes, C., Casella, G., and Wells, M. T. (2018), “Confidence Intervals for the Means of the Selected Populations,” *Electronic Journal of Statistics*, 12, 58–79. [1499,1502]
- Hall, P., and Miller, H. (2010), “Bootstrap Confidence Intervals and Hypothesis Tests for Extrema of Parameters,” *Biometrika*, 97, 881–892. [1499]
- Kubota, K., Ichinose, Y., Scagliotti, G., Spigel, D., Kim, J., Shinkai, T., Takeda, K., Kim, S.-W., Hsia, T.-C., Li, R., and Tiangco, B. J. (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, 529–536. [1498,1503,1505]
- Kubota, K., Yoshioka, H., Oshita, F., Hida, T., Yoh, K., Hayashi, H., Kato, T., Kaneda, H., Yamada, K., Tanaka, H., and Ichinose, Y. (2017), “Phase III, Randomized, Placebo-Controlled, Double-Blind Trial of Motesanib (AMG-706) in Combination With Paclitaxel and Carboplatin in East Asian Patients With Advanced Nonsquamous Non-Small-Cell Lung Cancer,” *Journal of Clinical Oncology*, 35, 3662–3670. [1498]
- Lin, D. Y., and Wei, L.-J. (1989), “The Robust Inference for the Cox Proportional Hazards Model,” *Journal of the American Statistical Association*, 84, 1074–1078. [1505]
- Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011), “Subgroup Identification Based on Differential Effect Search—A Recursive Partitioning Method for Establishing Response to Treatment in Patient Subpopulations,” *Statistics in Medicine*, 30, 2601–2621. [1498]
- Magnusson, B. P., and Turnbull, B. W. (2013), “Group Sequential Enrichment Design Incorporating Subgroup Selection,” *Statistics in Medicine*, 32, 2695–2714. [1499]
- MOLOGEN (2018), “Final Analysis of Impulse Study Confirms Topline Data With Positive Subgroup Results,” *MOLOGEN Press Releases*. [1498]
- Nadarajah, S., and Kotz, S. (2008), “Exact Distribution of the Max/Min of Two Gaussian Random Variables,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 16, 210–212. [1500,1501]
- Rosenkranz, G. K. (2016), “Exploratory Subgroup Analysis in Clinical Trials by Model Selection,” *Biometrical Journal*, 58, 1217–1228. [1499]
- Shen, J., and He, X. (2015), “Inference for Subgroup Analysis With a Structured Logistic-Normal Mixture Model,” *Journal of the American Statistical Association*, 110, 303–312. [1498]
- Stallard, N., Todd, S., and Whitehead, J. (2008), “Estimation Following Selection of the Largest of Two Normal Means,” *Journal of Statistical Planning and Inference*, 138, 1629–1638. [1499]
- Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009), “Subgroup Analysis via Recursive Partitioning,” *Journal of Machine Learning Research*, 10, 141–158. [1498]
- Sun, X., Briel, M., Busse, J. W., You, J. J., Akl, E. A., Mejza, F., Bala, M. M., Bassler, D., Mertz, D., Diaz-Granados, N., and Vandvik, P. O. (2012), “Credibility of Claims of Subgroup Effects in Randomised Controlled Trials: Systematic Review,” *BMJ*, 344, e1553. [1498]
- Thomas, M., and Bornkamp, B. (2017), “Comparing Approaches to Treatment Effect Estimation for Subgroups in Clinical Trials,” *Statistics in Biopharmaceutical Research*, 9, 160–171. [1499]
- Van Der Vaart, A. W., and Wellner, J. A. (1996), “Weak Convergence,” in *Weak Convergence and Empirical Processes*, New York: Springer, pp. 16–28. [1502]
- Woody, S., and Scott, J. G. (2018), “Optimal Post-Selection Inference for Sparse Signals: A Nonparametric Empirical-Bayes Approach,” arXiv no. 1810.11042. [1499]