Subgroup Analysis: Risk Quantification and Debiased Inference

by

Xinzhou Guo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2019

Doctoral Committee:

Professor Xuming He, Chair
Associate Professor Ben Hansen
Professor Yi Li
Professor Ji Zhu

Xinzhou Guo

xinzhoug@umich.edu

ORCID iD: 0000-0001-9161-2990

For all the people

ACKNOWLEDGEMENTS

I would like to first express my biggest thanks to my advisor Professor Xuming He for all the time he spent and all the support he gave me in the past four years. Xuming is always patient, helpful and inspirational, and it is fortunate to have him as my advisor. Beyond as a caring advisor, Xuming is also a role model for young researchers for his leadership and deep insights in the discipline. I am proud to have the opportunity to learn from him.

I would like to thank my thesis committee members, Professors Ben Hansen, Yi Li and Ji Zhu, for all the helpful suggestions they made and all the encouragement they gave me. I would also like to thank my collaborators for all the interesting discussions and projects we worked on. I look forward to continuing to work with them.

In addition, I would like to thank my friends. I will greatly miss the time we spent and the fun we had together. I am also very grateful to the department, all the students I taught, and the staffs and faculties I met in these years. A journey without them could not be as wonderful.

Finally, I would like to acknowledge my beloved parents for their endless encouragement and love.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Subgroup analysis is frequently used to account for the treatment effect heterogeneity in clinical trials. When a promising subgroup is selected from existing trial data, a decision on whether an additional confirmatory trial for the selected subgroup is worth pursuing needs to be made. Unfortunately, the usual statistical analysis applied to the selected subgroup as if the subgroup is identified independent of the data often leads to overly optimistic evaluations. Any statistical analysis that ignores how the subgroup is selected tends to suffer from subgroup selection bias. In this dissertation, we propose two new statistical tools to evaluate the selected subgroup. The first is a risk index which can be used as a simple screening tool to reduce the risk of over-optimism in naive subgroup analysis and the second is debiased inference to answer the question of how good the selected subgroup really is. The proposed tools are model-free, easy-to-implement and adjust for the subgroup selection bias appropriately. We demonstrate the merit of the proposed tools by re-analyzing the MONET1 trial. An extension of the debiased inference method is also discussed for observational studies with potentially many confounders.

CHAPTER I

Introduction

1.1 Subgroup Analysis

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. When this happens, subgroup analysis might help researchers better understand the treatment, know where the treatment would be useful and even rescue the trial. 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 while the initial trial failed to confirm the efficacy of lefitlimod for the overall study population, see the announcement from *MOLOGEN* (2018). Because of the potential benefits, subgroup analysis is widely used in clinical trials.

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. For example, *Bang et al.* (2010) studied the efficacy of trastuzumab in combination

with chemotherapy for gastric cancer in the subgroup defined by the expression level of HER2 protein and *Maemondo et al.* (2010) evaluated the efficacy of Gefitnib in the mutated epidermal growth factor receptor (EGFR) subgroup for the patient with non-small cell lung cancer. In the era of precision medicine, the evaluation of the subgroup effects is a popular research area with substantial impacts.

In practice, subgroup analysis might be conducted in many different ways but in clinical trials, it typically consists of two inter-connected steps: subgroup identification and subgroup confirmation, as in the clinical studies mentioned above. In the identification step, one looks for a promising 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 selection, and better yet, an additional clinical trial on the identified subgroup. We refer the decision to invest more in an additional clinical trial as the decision of subgroup pursuit, as shown in Fig.1.1.

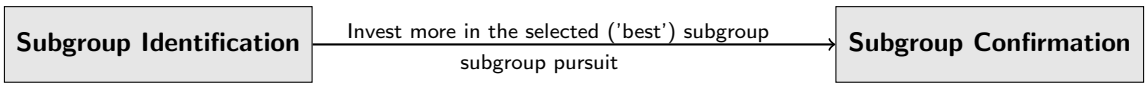

Figure 1.1: Two-step subgroup analysis

In this dissertation, we focus on subgroup pursuit where a decision on whether an additional investment should be made for a confirmatory trial on the selected subgroup needs to be made. Carefully studying the risk and bias issue in subgroup pursuit, we find that classical statistical tools which ignores how the subgroup is selected will lead to overly optimistic evaluations of the best selected subgroup effect. In response to the needs of subgroup analysis in clinical trials, we develop two new statistical analysis tools for the selected subgroup, risk quantification and debiased inference, and aim to help a better-informed decision on subgroup pursuit.

Although the statistical tools we develop in this dissertation are mainly motivated

by and focus on the current practice of subgroup analysis in clinical trials, an extension of the debiased inference tool to subgroup analysis on observational data is also introduced at the end of this dissertation. Moreover, the new statistical tools we develop are expected to have broad applications in other disciplines of data science. Besides clinical trials, subgroup analysis is also widely used in marketing and political science. For example, *Imai et al.* (2013) studied which subgroup of the disadvantaged workers benefits most from the national supported network program, a job training program, and the tools we develop here may help political scientists determine how good the most beneficial subgroup in this study really is.

1.2 Challenges in Subgroup Pursuit

In this section, we take a closer look at subgroup pursuit and carefully study the risk and bias issue there, and motivate the goal of this dissertation.

1.2.1 Risk in Subgroup Pursuit

Subgroup pursuit is referred to the additional confirmatory trial for the best selected subgroup as shown in Fig.1.1. In this dissertation, by the best selected subgroup, we refer to the subgroup that has the highest observed (or estimated) treatment effect among a pre-specified 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 algorithm or model-based method which will be discussed later. 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. A better subgroup is taken as a subgroup with a better treatment effect. If the best subgroup is non-unique, we take any one of them for the purpose of our analysis.

To confirm the best selected subgroup, subgroup pursuit is often necessary. How-

ever, there is a clear risk in subgroup pursuit that a subgroup pursuit requires additional resources but can not guarantee that the treatment would be confirmed effective in the best selected subgroup with the additional trial. In other words, while a successful subgroup pursuit might bring many potential benefits, it is also possible for a subgroup pursuit to fail and lead to a waste of resources as reported in *Naggara et al.* (2011). Therefore, greedy subgroup pursuit is not recommended in practice, and after the best subgroup is identified, the drug developers need to decide whether they should pursue further on this identified subgroup. Fig.1.2 shows a desired subgroup pursuit decision path from subgroup identification to confirmation.

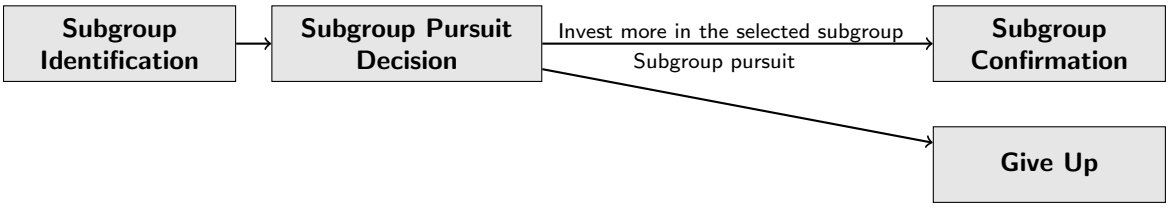

Figure 1.2: Three-step subgroup analysis

How to make scientific decisions on subgroup pursuit is a very important managerial question for the drug developer when conducting subgroup analysis. We believe that appropriate analysis of the best selected subgroup on the original trial data is needed to inform a better subgroup pursuit decision and the analysis should address the following two questions.

1. How risky is the pursuit of the selected subgroup?

2. How good is the selected subgroup?

By the analysis of the subgroup data, if the researchers can present a promising result showing the further pursuit of the selected subgroup is not very risky and the selected subgroup is good enough to meet the primary endpoint set beforehand, the managers would be led to believe that there is a very good chance that the treatment

is effective in the subgroup and it can be validated with an additional trial. Otherwise, the managers may consider giving up any further pursuit as shown in Fig.1.2.

Appropriate analysis of the selected subgroup might come from domain knowledge or statistical analysis or both. For the former, researchers will try to figure out the biological mechanism to explain why certain treatment is useful for certain patients. For example, the researchers noticed that Gefitnib is one of EGFR tyrosine kinase inhibitors and reasoned that Gefitnib might be useful for patients with EGFR mutations; see *Maemondo et al.* (2010). For the statistical analysis, researchers will try to statistically quantify the risk for subgroup pursuit and infer how good the selected subgroup really is. In this dissertation, we focus on the statistical analysis of the selected subgroup.

Many classical statistical tools are available to quantify the risk and make inference for a given subgroup. For example, the log-rank test may be used for inference on the subgroup effect for time-to-event data; see *Peto and Peto* (1972) and *Klein and Moeschberger* (2005). Unfortunately, applying the classical statistical tools to evaluate the selected subgroup often leads to disappointing results. The chance of success with confirming the treatment effect in the selected subgroup is nowhere near what it is supposed to be and overly optimistic decisions on subgroup pursuit are often made.

1.2.2 MONET1 Trial

One case study 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). It was found that this treatment is not effective for the overall population. To rescue this trial, the drug developer, Amgen, turned to subgroup analysis and 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

Amgen decided to invest additional resources and conduct 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).

Looking back at the MONET1 study, we could say that the managers actually made an overly optimistic decision on the pursuit of East Asian subgroup, which led to a waste of resources. Some might argue that the failure of the MONET1 trial may be just by chance. However, as follow-up trials to confirm a promising subgroup identified from earlier trial data failed more often than expected, question of statistical validity of classical statistical analysis becomes more acute and we have to ask whether the preplanned subgroup analysis was appropriately adjusted for.

1.2.3 Subgroup Selection Bias

It is clear that statistical analysis of the best selected subgroup identified from the same data suffers from over-optimism and is likely to lead to false discoveries, which we call subgroup selection bias. Take the MONET1 study as an example, if the drug developer used the observed effect size of the best selected subgroup from the same data to quantify the risk of the subgroup pursuit and answer the question of how good the East Asian subgroup is, the failure of the follow-up trial is not just by chance but mainly due to the subgroup selection bias.

To fix ideas, we consider a toy example consisting of two pre-specified 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 $\hat{\beta}_2 = 0.1$, then, subgroup 1 would be identified as the best selected subgroup with a promising treatment effect.

To understand the over optimism of the observed effect size of the selected subgroup, $\max(\hat{\beta}_1, \hat{\beta}_2)$, 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 1.3 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, a classical statistical analysis can not lead to a valid evaluation of the selected subgroup. The failure of the classical statistical tool for the selected subgroup is due to the fact the method is designed for the fixed population and ignores how the subgroup is selected. To make a better-informed decision, an appropriate adjustment to the subgroup selection bias is needed.

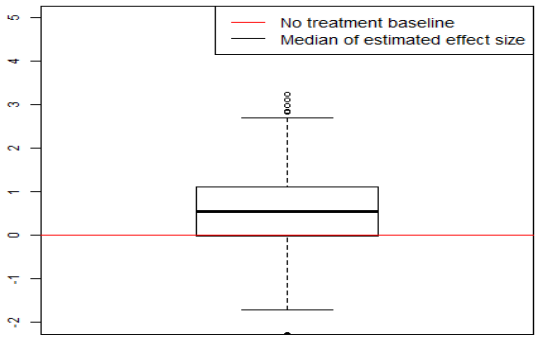

Figure 1.3: Boxplot of $\max(\hat{\beta}_1, \hat{\beta}_2)$ when $\beta_1 = \beta_2 = 0$ and $\hat{\beta}_1$, $\hat{\beta}_2 \sim^{i.i.d} \mathrm{N}(0, 1)$

1.3 The Goals

In this dissertation, we will propose two new statistical tools for the selected subgroup to help a better-informed decision on subgroup pursuit.

1. Risk Quantification: a risk measure for the pursuit of the selected subgroup;

2. Debiased Inference: a valid inference procedure for the selected subgroup.

To make sure the proposed tools provide valid analysis of the selected subgroup, we need to address subgroup selection bias appropriately. We aim to make use of

bootstrap, a resampling method, to account for the selection procedure and address the selection bias, and we call this idea bootstrapping the bias. To be specific, we will develop bootstrap-based methods to mimic the subgroup selection procedure and account for the bias. Even though the true subgroup selection procedure cannot be simulated directly without knowing the true data generating model, bootstrap has the potential to approximate this simulation procedure, and help us learn about the subgroup selection bias and develop an appropriate procedure for bias correction. In general, we aim for new statistical tools to analyze the selected subgroup, which are model-free, easy to implement and well-justified. An extension of the proposed statistical tools to observational studies will be also discussed.

Another goal of our work is to provide some practical guidance for future subgroup analysis. By revisiting the failed MONET1 trial, we wish to demonstrate that any proper statistical treatment of the selected subgroup has to depend on how the subgroup is selected. For example, if the subgroup of East Asians in the MONET1 trial were selected from only two preplanned candidate subgroups (East Asians as one subgroup and the rest of the population as the other), the selection bias would be much more limited than in the case of a broader search with a larger number of candidate subgroups. Therefore, without knowing how the subgroup is selected, no good statistical analyses exist for the selected subgroup.

1.4 Literature Review

In this section, we will review some relevant literature. We will review the literature on subgroup identification, subgroup confirmation and debiased inference in subgroup analysis together with some other relevant literature in other disciplines.

1.4.1 Subgroup Identification

How to identify the best subgroup is one important research problem in subgroup analysis. Most existing literature for subgroup identification falls into two categories: machine learning methods and model-based methods.

Machine learning methods aim to identify the best subgroup through model-free and nonparametric procedures. For example, *Lipkovich et al.* (2011) proposed a recursive partitioning-based method to identify the best subgroup. The basic idea is to split the population into two small subgroups by maximizing the value of a prespecified split criterion and retain the subgroup with the larger observed treatment effect. The process of the split will continue with the retained subgroups until certain criterion is met. *Cai et al.* (2010) considered a two-stage procedure to identify the subgroup. The authors first grouped the subjects into subgroups with certain working model and then identify the best subgroup by a nonparametric method. Some other available machine learning methods for subgroup identification include *Su et al.* (2009), *Foster et al.* (2011) and *Lipkovich and Dmitrienko* (2014).

Model-based methods aim to identify the best subgroup through parametric modelling of the subgroup effects. For example, *Shen and He* (2015) considered a mixture model for subgroup identification. The authors used the normal mixture to model the latent subgroup membership and the response in each subgroup. An EM algorithm was provided to estimate the parameters in the mixture model and help identify the subgroups. Some other available model-based methods include *Fan et al.* (2017) and *Altstein et al.* (2011).

The methods reviewed above mainly focus on subgroup identification and do not consider the inference on the subgroup selected by their algorithms in a rigorous way, while this dissertation aims to provide statistical tools for appropriate analysis of the selected subgroup. More literature review on subgroup identification can be found in *Zhang et al.* (2018) and *Loh et al.* (2019).

1.4.2 Subgroup Confirmation

After a promising subgroup is selected, how to design a new trial and confirm the selected subgroup is another important research problem in subgroup analysis. Most existing clinical trail designs in subgroup confirmation fall into two categories: fixed designs and adaptive designs.

Fixed designs aim to confirm the selected subgroup with single-stage designs. For example, *Ziegler et al.* (2012) discussed one possible strategy based on the idea of enrichment. The authors proposed to first screen the biomarker and only recruit the patient in the selected subgroup. Then, a randomized trial was conducted on the recruited patient to confirm the selected subgroup. Some other available fixed design methods include *Mandrekar and Sargent* (2009) and *Eng* (2014).

Adaptive designs aim to confirm the selected subgroup with multiple-stage designs. For example, *Friede et al.* (2012) considered a two-stages recruitment of the patients in the selected subgroup where an early stop was allowed. The authors used the conditional error function to account for the adaptation, and confirmed the selected subgroup by combining statistical evidences in all available stages. Some other available adaptive design methods include *Jenkins et al.* (2011) and *Song* (2014).

With data from additional trials, these methods are not applicable to address the problem on subgroup pursuit in which we wish to analyze the selected subgroup based on the original trial data.

1.4.3 Debiased Inference in Subgroup Analysis

As discussed before, analysis of the selected subgroup based on the original trial data suffers from subgroup selection bias, which is well-recognized in subgroup analysis as a fundamental challenge for inference; see for example *Thomas and Bornkamp* (2017) and *Magnusson and Turnbull* (2013). Several methods have been proposed to address subgroup selection bias and they mainly fall into three categories: simulta-

neous control methods, ad-hoc methods and bayesian methods.

Simultaneous control methods control subgroup selection bias by simultaneous analyzing all the subgroups. For example, *Fuentes et al.* (2018) proposed a valid confidence interval for the selected subgroup effect. Under gaussian assumption, the authors first minimized the coverage probability across all the possible subgroup effect sizes for any given critical value and then chose a critical value to guarantee the minimized coverage probability attain the nominal level. Some other available simultaneous control methods include *Venter* (1988), *Hothorn et al.* (2008) and *Hall and Miller* (2010). Although simultaneous control methods provide valid analysis of the selected subgroup, it is clearly on the conservative side.

Ad-hoc methods address subgroup selection bias by some simple statistical principles. For example, *Stallard et al.* (2008) considered the plug-in rule and proposed bias-reduced estimates for the selected subgroup effect. Under gaussian assumption, the authors first derived the form of subgroup selection bias which depends on several unknown parameters. By plugging in different possible estimates for the unknown parameters, the authors proposed several bias-reduced estimates for the selected subgroup effect for practical use. Some other ad-hoc methods include *Rosenkranz* (2016) and *Shen* (2001). Although ad-hoc methods are widely used in practice for exploratory analysis, they lack any theoretical guarantees.

Bayesian methods model and address subgroup selection bias from the bayesian view. For example, *Bornkamp et al.* (2017) provided bias-reduced estimate as well as credible interval for the selected subgroup by bayesian model averaging. The basic idea is to average the naive estimates for the subgroups with the posterior model weights and eliminate the selection bias. Some other bayesian methods include *Woody and Scott* (2018) and *Berger et al.* (2014). It is clear that bayeisan method is model-dependent and lack of frequentist interpretations.

As far as we know, asymptotically sharp, well-justified and model-free debiased

tools for the selected subgroup are still lacking. We shall propose new statistical tools to bridge this gap.

1.4.4 Relevant Methods in Other Disciplines

In other disciplines, there is some existing literature relevant to this dissertation. Here, we will focus on intersection bound and selective inference .

Intersection bound is motivated by the recent development of econometrics where the parameter of interest may be only partially identifiable and known to lie within the bounds, such as the study of the unemployment compensation reforms effect in Germany; see *Lee and Wilke* (2009). The identification and inference for such parameters is equivalent to the identification and inference on the maximum/minimum of the bounds, which is similar to the identification and inference on the best selected subgroup. *Chernozhukov et al.* (2013b) proposed a two-stage procedure to identify and infer the intersection bound. The authors first estimated the region where the maximum/minimum of the bound was achieved and then constructed a simultaneous confidence band in the estimated region. Although the method for intersection bound shares some similar properties of the proposed tools in this dissertation, the proposed tool is a direct one-stage approach based on resampling.

Selective inference is motivated by the broad applications of modern model selection tools. How to preform valid inference after model selection is a important problem, which is similar to the problem of the analysis of the selected subgroup after subgroup identification. *Lee et al.* (2016) characterized the distribution of a post-selection estimator conditional on the selection event and proposed valid inference procedure for the post-selection estimator. Although the techniques developed for selective inference may be generalized to analyze the selected subgroup, it is unclear how to do so in nonlinear models. Moreover, most selective inference methods are not model-free.

CHAPTER II

Subgroup Analysis: Risk Quantification

In this chapter, we propose a resampling-based risk index to measure the risk for subgroup pursuit. When a promising subgroup is selected, we need to address the question of how risky it is to pursue this selected subgroup in a scientific way. The proposed index naturally accounts for the subgroup selection procedure and statistically quantifies how risky it is to invest additional resources into the selected subgroup and can be used as a screening tool on subgroup pursuit. If the risk index is not small, it indicates a nontrivial risk that the selected subgroup might be an artifact and a further pursuit of the selected subgroup may not be recommended. The proposed risk index is model-free, easy to compute and transparent. We analyze the MONET1 trial with the risk index and show that the selected subgroup in the MONET1 trial is indeed risky to pursue.

2.1 A Risk Index for Subgroup Pursuit

Consider a clinical trial of n patients and the observation on the i-th subject is $(Y_i, D_i, \delta_i, Z_i)$, where Y_i is the (possibly censored) survival time with the censoring indicator δ_i, D_i is the binary treatment indicator, and $Z_i \in \{1, 2\}$ is the subgroup indicator so that $Z_i = 1, 2$ means that the subject i belongs to subgroup 1 or 2 respectively. We use the proportional hazard model, $Y \sim D$, on each subgroup and on the

combined group as the working model, and the standard partial likelihood estimates of the log-hazard ratios are denoted by $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}$, respectively for the two subgroups and for the combined group. We focus on the case of two non-overlapping subgroups for simplicity but generalizations to more than two or overlapping subgroups are quite straightforward.

The question of subgroup pursuit arises when one subgroup is noticeably different from the other but the overall effect size is marginal at best. For the sake of simplicity, we assume that for the given data, $\hat{\beta}_1 > \hat{\beta}_2$, and subgroup 1 is the most promising. If we pursue subgroup 1 by making additional investments in a new clinical trial on the identified sub-population, the risk materializes when the trial leads to a failure to confirm a significant treatment effect on the sub-population. It is certainly helpful if we can statistically quantify the risk before a new clinical trial is planned. There are, unarguably, many ways such a risk can be measured. However, as pointed out in Section 1.2.3, the classical statistical tool, such as the observed effect size of the selected subgroup, ignores how the subgroup is selected and does not quantify the risk appropriately.

In this dissertation, we focus on a simple risk index which naturally accounts for the subgroup selection procedure and quantifies the risk in a scientific way. The risk index is defined as the bootstrap probability of observing a selected subgroup whose estimated log-hazard ratio is as good as or better than the current observation of the best selected subgroup, $\hat{\beta}_1$, where the bootstrap sample is generated without any subgroup differences.

2.1.1 Problem Setting

Even though we focus on the time-to-event data where the proportional hazard model and the log-hazard ratio are routinely used in evaluating the subgroup effects, we would like to make a case for not relying on the strict assumption of proportional

hazard in the risk quantification. When multiple subgroups are considered and some of them might overlap, it is highly unlikely that the proportional hazard model is correctly specified for each subgroup, as well as for the whole population. In fact, the strict assumption of proportional hazard on all subgroups holds only when the population is indeed homogeneous. For this reason, any theoretical investigation we carry out should allow the proportional hazard models to be misspecified. It has been shown that even under a misspecified proportional hazard model, the observed log-hazard ratio is still consistent (and asymptotically normal) for an implicitly defined parameter, which we will continue to call the true log-hazard ratio; see *Struthers and Kalbfleisch* (1986) and *Lin and Wei* (1989). Here, we start from their results and consider a very general problem setting.

We consider a family of the distributions $\mathbb{F}_\beta$ to represent the marginal distributions of the subgroups, where $\beta \in R$ is an unknown parameter and implicitly defined by the working model of proportional hazard, $Y \sim D$; see *Lin and Wei* (1989). If $\mathbb{F}_\beta$ is the proportional hazard model with a given baseline hazard function, β is naturally taken as the log-hazard ratio of the treatment. For other models, we still perform analysis under the working model of proportional hazard, $Y \sim D$, and β is then the implicitly defined parameter, which we will continue to call it the true log-hazard ratio for the model and use it as the treatment effect.

We consider the setting where $(Y_i, \delta_i, D_i, Z_i)$ of size n is a random sample from P_1, where P_1 represents the distribution of the whole population consisting of two subgroups. In Subgroup 1 with $Z_i = 1$, the random sample, (Y_i, δ_i, D_i), of size n_1 is taken from $\mathbb{F}_{\beta_1}$, and in Subgroup 2 with $Z_i = 2$, the random sample, (Y_i, δ_i, D_i), of size n_2 is taken from $\mathbb{F}_{\beta_2}$, where β_1 and β_2 are possibly different. It is easy to see the marginal distribution of Z_i is $B(1, p)$. We assume $0 < p < 1$ and denote the implicitly defined parameter of P_1 by β_0. If $\beta_1 = \beta_2$, β_0 would take the same value. It is noteworthy that we assume the samples from two subgroups are modeled by

the same family of distributions $\mathbb{F}$ but with possibly different treatment effect/log-hazard ratio β, and without conditional on the subgroup indicator, Z_i, the marginal distribution of (Y_i, δ_i, D_i) may not fall into the family $\mathbb{F}$. We also note that the particular family $\mathbb{F}$ does not have to be known for our analysis.

In Section 2.1.2, we will show that under the general problem setting considered here, the proposed risk index is well defined and justified without imposing strict assumptions of proportional hazard, and it is in this sense that the proposed risk index is model-free.

2.1.2 Computation of the Proposed Risk Index

Although $\max(\hat{\beta}_1, \hat{\beta}_2)$ is a biased estimate of the best subgroup effect size, the statistic itself is interpretable and transparent, which is the observed effect size of the selected subgroup. Therefore, we aim to propose a risk index based on this statistic directly but with an appropriate adjustment for the subgroup selection bias. In other words, sticking to $\max(\hat{\beta}_1, \hat{\beta}_2)$, the transparent statistic, we will propose a risk index answering the question of how risky it is to pursue the selected subgroup in a scientific way, and it is in this sense the proposed risk index is transparent.

Roughly speaking, we wish to measure the risk of subgroup pursuit by calculating the probability of observing a selected subgroup whose observed effect size is as good as or better than the current observation of the selected subgroup, $\max(\hat{\beta}_1, \hat{\beta}_2)$, when the subgroups are actually homogeneous. This probability however depends on the underlying distribution of the homogeneous population, so we turn to the method of resampling. Let P^* denote the probability measure of the following bootstrap procedure and β_1^* and β_2^* be the log-hazard ratio estimates of the two subgroups from the bootstrap sample. Then, our proposed risk index is

$$RI^* = P^*(\max(\beta_1^*, \beta_2^*) \geq \max(\hat{\beta}_1, \hat{\beta}_2)),$$

and is calculated as follows in Algorithm 1.

Algorithm 1 Risk index for subgroup pursuit.

1: **for** $b = 1...B$ **do**
2: **Partial bootstrap:** Generate $\{(Y_i^*, D_i^*, \delta_i^*) : \ i = 1, \cdots, n\}$ as a bootstrap sample from the set $\{(Y_j, D_j, \delta_j), j = 1, \cdots, n\}$;
3: **Subgroup assignment:** $Z_i^* = Z_i$ for $i = 1, \cdots, n$;
4: **Estimation:** Calculate the log-hazard ratio estimate of the two groups, $\beta_{1,b}^*$ and $\beta_{2,b}^*$, based on the bootstrap sample;
5: **end for**
6: The risk index is $RI_B = B^{-1} \sum_{b=1}^{B} I(\max(\beta_{1,b}^*, \beta_{2,b}^*) \geq \max(\hat{\beta}_1, \hat{\beta}_2))$.

As $B \to \infty$, the index RI_B becomes RI^* under the bootstrap distribution. The above nonparametric bootstrap procedure is based on the pair bootstrap on (Y_i, D_i, δ_i) without subgroup labels, and the subgroup assignments, $Z_i^* = Z_i$, are made to preserve the same number of subjects in each subgroup. This resampling scheme ensures that the bootstrap distribution is homogeneous across subgroups, irrespective of whether there is a distinctive subgroup effect in the original sample or not.

2.1.3 Property and Relevance of the Risk Index

To see how the risk index can be used as a screening tool for subgroup pursuit, we need to understand the limiting behavior of RI^*. To this end, let $P_{\hat{\beta}}$ denote the probability under the data generating process that both subgroups are drawn from $\mathbb{F}_{\hat{\beta}}$ with the total sample size n and the same subgroup assignment mechanism as that of the original data. Furthermore, let $\tilde{\beta}_1$ and $\tilde{\beta}_2$ be the estimates of the log-hazard ratios for the two subgroups under $P_{\hat{\beta}}$, respectively, and define

$$RI = P_{\hat{\beta}}(\max(\tilde{\beta}_1, \tilde{\beta}_2) \geq \max(\hat{\beta}_1, \hat{\beta}_2)),$$

which is a probability depending on the original sample. We make the following modelling assumptions.

Assumption II.1. $\lim_{\beta \to \beta_0} sup_y |F_\beta(y) - F_{\beta_0}(y)| = 0$.

Assumption II.2. $\beta_0 < \max(\beta_1, \beta_2)$ *whenever* $\beta_1 \neq \beta_2$.

Assumptions II.1 requires the mapping $\beta \to F_\beta$ to be continuous at β_0 under the sup norm. Assumption II.2 removes a pathological case from consideration, that is, the log-hazard ratio of the combined group cannot be greater than the log-hazard ratios of both subgroups. The assumptions are satisfied by many models and we do not need to impose strict assumptions of proportional hazard even though we use the proportional harzard model as the working model to define and calculate the parameters.

Theorem II.3. *Under Assumptions II.1 and II.2, we have* $|RI^* - RI| \to 0$ *in probability with respect to* P_1.

Theorem II.3 ensures that the risk index RI^* is asymptotically the same as RI, the probability of observing a selected subgroup that is at least as promising as the current observation of the selected subgroup, $\max(\hat{\beta}_1, \hat{\beta}_2)$, when the two subgroups are homogeneous with $\beta_1 = \beta_2 = \hat{\beta}$. This enables us to interpret and understand the proposed risk index, and justify its use as a screening risk measure of pursuing the most promising subgroup identified from the data. When $\beta_1 \neq \beta_2$, that is, the treatment effects are indeed different for the two subgroups, we have low or no risk of pursuing the better subgroup. In this case, we note that RI approaches zero as $n \to \infty$, so our proposed risk index, RI^*, would also be close to zero. On the other hand, the risk of pursuing any subgroup becomes evident when $\beta_1 = \beta_2 = \beta_0$. In the latter setting, RI converges to a non-degenerate distribution on (0,1) as $n \to \infty$, and the proposed risk index, RI^*, will be close to zero with a small probability. In this sense, we can use the risk index as a screening measure of risk of subgroup pursuit. If the risk index is not small, we should take it as a quantitative argument against investing additional resources into the subgroup with the seemingly promising treatment effect.

How large is too large for the risk index is more of a managerial question. It depends on how much risk one is willing to take, based on the cost of an additional trial and the potential return from a successful follow-up trial. As a rough guideline, we recommend against subgroup pursuit if the risk index is 0.15 or higher.

To see an example of mis-specified working models, consider a special case where the sample (Y_i, D_i) is given by the hazard function

$$\lambda(t) = \lambda_0(t)e^{\beta D+\zeta^T W},$$

where W is a random vector independent of D, and ζ is an unknown vector. This falls into the proportional hazard model itself and satisfies Assumptions II.1 and II.2, but the working model without including W as a covariate would be mis-specified. As shown in *Lin and Wei* (1989), β remains to be the log-hazard ratio (approximately) under the working model and is interpretable.

2.1.4 The Proposed Risk Index v.s. P-value

The risk index is closely related to the concept of p-values for the null hypothesis that $\beta_1 = \beta_2 (= \beta_0)$. Since $\hat{\beta} \to \beta_0$ as the sample size increases, we may expect $P_{\hat{\beta}}(\max(\tilde{\beta}_1, \tilde{\beta}_2) \geq \max(\hat{\beta}_1, \hat{\beta}_2))$ as well as the risk index to agree with $P_{\beta_0}(\max(\tilde{\beta}_1, \tilde{\beta}_2) \geq \max(\hat{\beta}_1, \hat{\beta}_2))$ asymptotically. The latter is indeed the p-value with $\max(\hat{\beta}_1, \hat{\beta}_2)$ as the test statistic, but cannot be calculated unless β_0 is known. However, we hasten to add that this asymptotic equivalence is untrue and, indeed, the risk index is not a p-value for the null hypothesis of homogeneity itself.

Any p-value for the null hypothesis of homogeneity $\beta_1 = \beta_2 (= \beta_0)$ may serve as a risk index, but most p-values, such as that from the likelihood ratio test, are model-based. Although some p-values, such as that from the Wald test, may handle model misspecification with sandwich-type estimates, they are usually difficult to

calculate under the scenario of multiple subgroups. On the contrary, our proposed risk index is model-free and easy to compute, and has the desirable property that it converges to zero whenever $\beta_1 \neq \beta_2$ but converges to a non-degenerate distribution on (0,1) otherwise. More importantly, the risk index is directly based on $\max(\hat{\beta}_1, \hat{\beta}_2)$, a widely used and transparent quantity in the current practice of subgroup pursuit decision, and addresses its bias appropriately. Therefore, our proposed risk index is more transparent than the p-values from commonly used test statistics for the null hypothesis of homogeneity.

2.2 Synthetic Data: MONET-1 Study

In this section, we revisit the failed MONET1 trial as a case study. With our proposed risk index, we provide an appropriate guidance on subgroup pursuit decisions after the initial MONET1 trial data are available.

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 nonsquamous 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 this chapter, one may simply equate β_i in this chapter 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-value=0.0223, as reported *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 dataset that shares many of the same characteristics as the MONET1 trial for our 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 versus 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 A.

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 also 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 2.1 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, which are very close.

Table 2.1: The comparison between the synthetic data and MONET-1 study.

	Harzard Ratio	P-value
Synthetic data	0.663	0.019
MONET-1	0.669	0.022

Table 2.2: Risk index of the synthetic data. The standard errors for all the entries are less than 0.01.

No. of subgroups	2	4	6	8	10	12	14	16
Risk Index	0.02	0.08	0.14	0.15	0.15	0.16	0.17	0.18

In the MONET-1 study, a lower hazard ratio indicates a better subgroup. To make it consistent with the framework used in this chapter, we use the negative log-hazard ratio as the treatment effect in calculating the risk index. From Table 2.2, we see that as the number of candidate subgroups increases, the risk index rises. If eight candidate subgroups were considered in subgroup pursuit, the risk index is 0.15, which means that even if the population is indeed homogeneous (i.e. no subgroups), we have 15% chance to observe a subgroup that is at least as promising as the East Asians (HR=0.663). On the other hand, if only two candidate subgroups were considered (East Asians v.s. the rest) in the planning stage, the risk index would be quite low in this case. If we ask the question whether we should have recommended a follow-up trial on the East Asian population, the answer depends on how the subgroup was selected. If East Asian subgroup was selected after eight or more candidate subgroups were considered, we would have to be far more cautious.

2.3 Discussion

In this chapter, we propose a statistical quantification for the risk of subgroup pursuit. The proposed risk index aims to address the question of how risky it is to pursue the selected subgroup in a scientific way and can help a better decision on subgroup pursuit.

The risk index is model-free in the sense that it is defined and justified without imposing strict assumptions of proportional hazard. Although the proportional hazard model is routinely used in clinical trials, misspecifications of the proportional hazard model over the subgroups of interest are the norm rather than the exception.

The model-free property allows us to appropriately quantify the risk for subgroup pursuit with the proposed risk index even when the proportional hazard model is misspecified, which makes the proposed risk index practically useful.

The other property to note for the risk index is its transparency that the risk index is directly based on the observed effect size of the selected subgroup, a transparent and widely used statistic in the current practice of subgroup analysis. Although the observed effect size of the selected subgroup is biased and tends to underestimate the risk for subgroup pursuit, it is widely used by practitioners due to its simplicity and transparency. The proposed risk index adjusts for the subgroup selection bias in the observed effect size of the selected subgroup directly and is therefore transparent and easy to understand in practice.

The risk index is also easy to implement in the sense that it is based on a simple pair bootstrap. Even when the proportional hazard model is misspecified, by bootstrapping, we do not need to worry about the calculation of the sandwich-type estimate as in *Lin and Wei* (1989). Moreover, the simple algorithm enables us to easily generalize the risk index to the situations where there are multiple overlapped subgroups, and to measure the risk for subgroup pursuit there.

It is noteworthy that the risk index we have considered here is designed for the scenario that the subgroups have been predefined. This is often a recommended approach in subgroup pursuit. However, post hoc subgroup identification without prespecified subgroups is often used in practice; see *Lipkovich et al.* (2011). Therefore, it will be of interest to generalize our work to more challenging scenarios where no subgroups are predefined.

2.4 Proof of Theorem II.3

Consider the data $\{(Y_i, \delta_i, D_i)\}_{i=1}^n$ as n i.i.d realizations of (Y, δ, D), where Y is the survival time, δ is the censoring indicator, and D is the treatment indicator, respec-

tively. Let $\bigtriangledown l_n(r)$, $\bigtriangledown^2 l_n(r)$ denote the first and second derivative of the log partial likelihood, and $Y_i(t) = I_{t \le Y_i}$, $Y(t) = I_{t \le Y}$. Let E_γ denote the expectation under F_γ and for $q = 0, 1, 2$, we introduce the following quantities,

$$S^{(q)}(\beta, t) = \sum_{i=1}^{n} Y_i(t) e^{\beta D_i} D_i^q / n,$$

$$s_\gamma^{(q)}(\beta, t) = E_\gamma S^{(q)}(\beta, t),$$

$$\omega_{\gamma,i}(\beta) = \int_0^\infty \left\{ D_i - \frac{s_\gamma^{(1)}(\beta,t)}{s_\gamma^{(0)}(\beta,t)} \right\} dN_i(t) - \int_0^\infty \frac{Y_i(t) e^{\beta D_i}}{s_\gamma^{(0)}(\beta,t)} \left\{ D_i - \frac{s_\gamma^{(1)}(\beta,t)}{s_\gamma^{(0)}(\beta,t)} \right\} d\bar{F}_\gamma(t),$$

where $N_i(t) = I_{Y_t \le t, \delta_i = 1}$ and $F_n(t) = \sum_{i=1}^n N_i(t)/n$, $\bar{F}_\gamma(t) = E_\gamma(F_n(t))$. Furthermore, we let

$$\bigtriangledown \tilde{l}_{\gamma,n}(r) = \sum_{i=1}^{n} \omega_{\gamma,i}(r)/n,$$

$$\bigtriangledown^2 \tilde{l}_{\gamma,n}(r) = \sum_{i=1}^{n} \delta_i \left\{ \frac{s_\gamma^{(2)}(r, Y_i)}{s_\gamma^{(0)}(r, Y_i)} - [\frac{s_\gamma^{(1)}(r, Y_i)}{s_\gamma^{(0)}(r, Y_i)}]^2 \right\} /n.$$

The notations, $\bigtriangledown \tilde{l}$ and $\bigtriangledown^2 \tilde{l}$, do not mean that they are the first and second derivatives of some quantities, instead, we use these notations because they are approximations to $\bigtriangledown l_n$ and $\bigtriangledown^2 l_n$ respectively. In the end, we let $A_i(r) = E_{\beta_i}(\bigtriangledown^2 \tilde{l}_{\beta_i,n}(r))$ for $i = 0, 1, 2$. For any two quantities a_n and b_n, we will use $a_n \sim b_n$ to denotes $a_n - b_n \to 0$ in probability as $n \to \infty$.

As shown in *Struthers and Kalbfleisch* (1986), $\hat{\beta} \to \beta_0$ in probability w.r.t P_1. To simplify the proof, we assume that the support of Y of F_{β_0} is R^+ and the marginal distribution of Y is continuous. We also focus on a stronger version of Assumption II.1 which assumes that uniformly continuity in Assumption II.1 is true for a neighborhood B of β_0 and we still call it Assumption $II.1$ in this section. The additional assumptions are not essential and for other situations, the proof is similar. Recall As-

sumptions II.1 and II.2, we first establish some basic properties of some key quantities.

Lemma II.4. *Under Assumption II.1, we have:*

(1) For any $q = 0, 1, 2$, *both* $\sup_{\beta,t\in R} |S^{(q)}(\beta,t)/S^{(0)}(\beta,t)|$ *and* $\sup_{\beta,\gamma,t\in R} |s_\gamma^{(q)}(\beta,t)/s_\gamma^{(0)}(\beta,t)|$ *are bounded with regard to* n;

(2) For any $q = 0, 1, 2$, $\lim_{\beta\to\beta_0,\gamma\to\beta_0} \sup_{t\in R} |s_\gamma^{(q)}(\beta,t) - s_{\beta_0}^{(q)}(\beta_0,t)| = 0$ *and* $s_\gamma^{(q)}(\beta,t)$ *is continuous in* t *for any* γ *and* β;

(3) For any $q = 0, 1, 2$ *and any* $\epsilon > 0$, $\sup_{\gamma\in\mathbb{B}} P_\gamma(\sup_{\beta\in\mathbb{B},t\in R} |S^{(q)}(\beta,t) - s_\gamma^{(q)}(\beta,t)| > \epsilon) \to 0$;

(4) For all $T < \infty$, $\inf_{\gamma\in\mathbb{B},\beta\in\mathbb{B},t\in[0,T]} s_\gamma^{(0)}(\beta,t)$ *is bounded below;*

(5) $sup_{\beta,\gamma\in\mathbb{B}} E_\gamma[\int_0^\infty \frac{Y(t)}{s_\gamma^{(0)}(\beta,t)} d\bar{F}_\gamma(t)]^2 < \infty$, *and*

$\lim_{\gamma\to\beta_0} \lim_{M_1\to\infty} \sup_{\beta\in\mathbb{B}} E_\gamma[\int_{M_1}^\infty \frac{Y(t)}{s_\gamma^{(0)}(\beta,t)} d\bar{F}_\gamma(t)]^2 = 0.$

Proof. (1) We can see that $D_i^q \le D_i^0$, so, from the definition, $S^{(q)}(\beta,t)/S^{(0)}(\beta,t)$ and $s_\gamma^{(q)}(\beta,t)/s_\gamma^{(0)}(\beta,t)$ are all bounded by 1.

(2) Take $q = 1$ as an example, we have the following inequality,

$$
\begin{aligned}
&\sup_{t\in R} |s_\gamma^{(q)}(\beta,t) - s_{\beta_0}^{(q)}(\beta_0,t)| \\
=&sup_{t\in R} |P_\gamma(D=1)e^\beta E_\gamma(Y(t)|D=1) - P_{\beta_0}(D=1)e^{\beta_0} E_{\beta_0}(Y(t)|D=1)| \\
=&sup_{t\in R} |P_\gamma(D=1)e^\beta P_\gamma(Y\ge t|D=1) - P_{\beta_0}(D=1)e^{\beta_0} P_{\beta_0}(Y\ge t|D=1)| \\
\le&sup_{t\in R} |P_\gamma(D=1)e^\beta P_\gamma(Y\ge t|D=1) - P_{\beta_0}(D=1)e^{\beta_0} P_\gamma(Y\ge t|D=1)| + \\
&sup_{t\in R} |P_{\beta_0}(D=1)e^{\beta_0} P_\gamma(Y\ge t|D=1) - P_{\beta_0}(D=1)e^{\beta_0} P_{\beta_0}(Y\ge t|D=1)| \\
\le&|P_\gamma(D=1)e^\beta - P_{\beta_0}(D=1)e^{\beta_0}| + \\
&P_{\beta_0}(D=1)e^{\beta_0} \sup_{t\in R} |P_\gamma(Y\ge t|D=1) - P_{\beta_0}(Y\ge t|D=1)|.
\end{aligned}
\tag{2.1}
$$

From Assumption II.1, P_γ is continuous at $\gamma = \beta_0$, so the first term on the right hand side of (2.1) goes to 0 as β and γ go to β_0. From Assumption II.1, $P_\beta(Y \geq t|D = 1)$ is continuous at $\beta = \beta_0$ uniformly in t, so the other term on the right hand side of (2.1) goes to 0 too. The proof for $q = 0, 2$ is similar. The continuity of $s_\gamma^{(q)}(\beta, t)$ in t for any γ and β is trivial from our assumptions.

(3) First, we note that $s_\gamma^{(q)}(\beta, t)$ and $S^{(q)}(\beta, t)$ are monotone in β and t. From the assumption of the continuity of the marginal distribution of Y, $s_\gamma^{(q)}(\beta, t)$ is continuous in β and t. Therefore, by Lemma II.4 (2) and Assumption II.1, given any $\epsilon > 0$ and any $\gamma \in \mathbb{B}$, there exists a constant k, which is determined by ϵ but independent of γ, and a sequence $\{(\beta_{i,\gamma}, t_{i,\gamma})\}_{i=1}^k$, which is determined by γ, such that

$$P_\gamma(\sup_{\beta\in\mathbb{B},t\in R} |S^{(q)}(\beta, t) - s_\gamma^{(q)}(\beta, t)| > \epsilon) \leq \sum_{i=1}^{i=k} P_\gamma(|S^{(q)}(\beta_{i,\gamma}, t_{i,\gamma}) - s_\gamma^{(q)}(\beta_{i,\gamma}, t_{i,\gamma})| > \epsilon/3). \tag{2.2}$$

For any γ and t,

$$\begin{aligned} &P_\gamma(|S^{(q)}(\beta, t) - s_\gamma^{(q)}(\beta, t)| > \epsilon) \\ \leq &var_\gamma(S^{(q)}(\beta, t))/\epsilon^2 = var_\gamma(Y(t)e^{D\beta}D^q)/(n\epsilon^2) \leq e^{2\beta}/(n\epsilon^2). \end{aligned} \tag{2.3}$$

Therefore, for any $\beta \in \mathbb{B}$, $\gamma \in \mathbb{B}$ and $t \in R$, $P_\gamma(|S^{(q)}(\beta, t) - s_\gamma^{(q)}(\beta, t)| > \epsilon)$ is bounded by a constant divided by n, and the constant is independent of β, γ and t. Combining it with the decomposition in (2.2), we prove for any $\gamma \in \mathbb{B}$, $P_\gamma(\sup_{\beta\in\mathbb{B},t\in R} |S^{(q)}(\beta, t) - s_\gamma^{(q)}(\beta, t)| > \epsilon)$ is bounded by a constant divided by n, and the constant is independent of γ, so the 3rd part of Lemma II.4 is proved.

(4) For any $T < \infty$, we note that $1 - F_{\beta_0}(T) > 0$, so $\inf_{t\in[0,T]} s_{\beta_0}^{(0)}(\beta_0, t)$ is bounded below. From Lemma II.4 (2), it follows naturally that $\inf_{\gamma\in\mathbb{B},\beta\in\mathbb{B},t\in[0,T]} s_\gamma^{(0)}(\beta, t)$ is bounded below.

(5) Let $N(t) = I_{Y \le t, \delta=1}$ and $G_\gamma(t) = 1 - E_\gamma Y(t)$. Notice that $G_\gamma(t) = P_\gamma(Y < t)$ and $\bar{F}_\gamma(t) = P_\gamma(Y < t, \delta = 1)$, we have $G_\gamma \ge \bar{F}_\gamma$ and

$$s_\gamma^{(0)}(\beta, t) \ge \min(e^\beta, 1) E_\gamma(Y(t)).$$

Therefore, there exists a constant $C_{\mathbb{B}} < \infty$ determined by $\mathbb{B}$ such that

$$\sup_{\beta,\gamma \in \mathbb{B}} E_\gamma[\int_0^\infty \frac{Y(t)}{s_\gamma^{(0)}(\beta, t)} d\bar{F}_\gamma(t)]^2 \le \sup_{\beta,\gamma \in \mathbb{B}} C_{\mathbb{B}} E_\gamma[\int_0^\infty \frac{Y(t)}{1 - G_\gamma(t)} dG_\gamma(t)]^2. \tag{2.4}$$

For any $\gamma \in \mathbb{B}$, $E_\gamma[\int_0^\infty \frac{Y(t)}{1 - G_\gamma(t)} dG_\gamma(t)]^2 \le \int_0^1 log^2 x dx < \infty$. Similarly, it follows that $E_\gamma[\int_{M_1}^\infty \frac{Y(t)}{1 - G_\gamma(t)} dG_\gamma(t)]^2 \le \int_0^{1 - G_\gamma(M_1)} log^2 x dx < \infty$. From Assumption II.1, we note that $\lim_{M_1 \to \infty} \inf_{\gamma \in \mathbb{B}} G_\gamma(M_1) > 1 - \epsilon_{\mathbb{B}}$ and $\epsilon_{\mathbb{B}}$ goes to 0 when $\mathbb{B}$ shrinks to the point β_0, so we prove the result. □

Lemma II.5. *Under Assumption II.1, for any $\epsilon > 0$, $P_{\hat{\beta}}(|\tilde{\beta} - \beta_0| > \epsilon) \to 0$ in probability w.r.t P_1, where $\tilde{\beta}$ is the standard partial likelihood estimate under $F_{\hat{\beta}}$.*

Proof. Since $\hat{\beta} \to \beta_0$ in probability, W.L.O.G, we assume that $\hat{\beta} \in \mathbb{B}$.

(1) First, we will show that given $r \in \mathbb{B}$, $P_{\hat{\beta}}(\sqrt{n}| \bigtriangledown l_n(r) - \bigtriangledown \tilde{l}_{\hat{\beta},n}(r)| > \epsilon) \to 0$ in probability w.r.t P_1.

Similar to the techniques used in the proof of the asymptotic normality of the partial likelihood estimator under a mis-specified Cox model in the Appendix of *Lin and Wei*

(1989), we have the following useful decomposition,

$$\begin{aligned}
&\sqrt{n}(\bigtriangledown l_n(r) - \bigtriangledown \tilde{l}_{\hat{\beta},n}(r)) \\
=& -\int_0^\infty \{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)}\} d\{\sqrt{n}(F_n(t) - \bar{F}_{\hat{\beta}}(t))\} \\
&-\int_0^\infty \{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)}\} \sqrt{n}(S^{(0)}(r,t) - s^{(0)}_{\hat{\beta}}(r,t))/s^{(0)}_{\hat{\beta}}(r,t) d\bar{F}_{\hat{\beta}}(t).
\end{aligned} \tag{2.5}$$

From Lemma II.4 (1), (3) and (4), we know that $\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)}$ is bounded and for any $\tau > 0$, we know that $P_{\hat{\beta}}(\sup_{r\in\mathbb{B},t\in[0,\tau]} |\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)}| > \epsilon) \to 0$ in probability w.r.t P_1. Notice that$\sqrt{n}(\bar{F}_n(t) - F_{\hat{\beta}}(t))$ converge to a zero-mean Gaussian process in probability w.r.t P_1, we can show that for any $\eta > 0$, there exists an appropriate partition τ_1, s.t.

$$\begin{aligned}
&\limsup P_{\hat{\beta}}(|-\int_0^\infty \left\{ \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)} \right\} d\{\sqrt{n}(F_n(t) - \bar{F}_{\hat{\beta}}(t))\}| > \epsilon/2) \\
\leq &\limsup P_{\hat{\beta}}(|-\int_0^{\tau_1} \{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)}\} d\{\sqrt{n}(F_n(t) - \bar{F}_{\hat{\beta}}(t))\}| > \epsilon/4) + \\
&\limsup P_{\hat{\beta}}(|-\int_{\tau_1}^\infty \left\{ \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)} \right\} d\{\sqrt{n}(F_n(t) - \bar{F}_{\hat{\beta}}(t))\}| > \epsilon/4) \\
<&\eta
\end{aligned} \tag{2.6}$$

in probability w.r.t P_1. Therefore, we control the 1st term on the right hand side of (2.5).

For the 2nd term on the right hand side of (2.5), notice that given $r \in \mathbb{B}$, $\sqrt{n}(S^{(0)}(r,t) - s^{(0)}_{\hat{\beta}}(r,t))$ converges to a zero-mean Gaussian process in probability w.r.t P_1, we can

decompose

$$\int_0^\infty \left\{ \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)} \right\} \sqrt{n}(S^{(0)}(r,t) - s^{(0)}_{\hat{\beta}}(r,t))/s^{(0)}_{\hat{\beta}}(r,t) d\bar{F}_{\hat{\beta}}(t)$$

into

$$\begin{aligned} &\int_0^{\tau_2} \left\{ \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)} \right\} \sqrt{n}(S^{(0)}(r,t) - s^{(0)}_{\hat{\beta}}(r,t))/s^{(0)}_{\hat{\beta}}(r,t) d\bar{F}_{\hat{\beta}}(t) \\ +&\int_{\tau_2}^\infty \left\{ \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)} \right\} \sqrt{n}(S^{(0)}(r,t) - s^{(0)}_{\hat{\beta}}(r,t))/s^{(0)}_{\hat{\beta}}(r,t) d\bar{F}_{\hat{\beta}}(t) \end{aligned} \tag{2.7}$$

with appropriate τ_2. The first term of (2.7) goes to 0 in probability due to the uniform convergence of $\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)}$ to 0, the L_∞ norm of the gaussian process and the boundness of $1/s^{(0)}_{\hat{\beta}}(r,t)$ when $t \in [0, \tau_2]$. To control the second term of (2.7), we note that $\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)}$ is bounded, so there exists a constant C such that

$$\begin{aligned} &|\int_{\tau_2}^\infty \left\{ \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)} \right\} \sqrt{n}(S^{(0)}(r,t) - s^{(0)}_{\hat{\beta}}(r,t))/s^{(0)}_{\hat{\beta}}(r,t) d\bar{F}_{\hat{\beta}}(t)| \\ \leq &|\frac{C}{\sqrt{n}} \sum_{i=1}^n \int_{\tau_2}^\infty (\frac{Y_i(t)e^{rD_i}}{s^{(0)}_{\hat{\beta}}(r,t)} - 1) d\bar{F}_{\hat{\beta}}(t)|. \end{aligned} \tag{2.8}$$

With Chebyshev's inequality, the latter one is controlled by $E_{\hat{\beta}}(\int_{\tau_2}^\infty (\frac{Y_i(t)e^{rD_i}}{s^{(0)}_{\hat{\beta}}(r,t)}) d\bar{F}_{\hat{\beta}}(t))^2$ after multiplied by a constant, which will go to 0 in probability as $\tau_2 \to \infty$ by Lemma II.4 (5), so we prove the result.

(2) Second, we will prove that for $r \in \mathbb{B}$, $P_{\hat{\beta}}(|\bigtriangledown l_{\hat{\beta},n}(r) - E_{\beta_0}\omega_{\beta_0}(r)| > \epsilon) \to 0$ in probability w.r.t P_1.

As implied in (1),

$$P_{\hat{\beta}}(|\bigtriangledown l_n(r) - E_{\beta_0}\omega_{\beta_0}(r)| > \epsilon) \sim P_{\hat{\beta}}(|\bigtriangledown \tilde{l}_{\hat{\beta},n}(r) - E_{\beta_0}\omega_{\beta_0}(r)| > \epsilon),$$

and the latter is smaller than

$$P_{\hat{\beta}}(|\bigtriangledown \tilde{l}_{\hat{\beta},n}(r) - E_{\hat{\beta}}\bigtriangledown \tilde{l}_{\hat{\beta},n}(r)| > \epsilon/2) + P_{\hat{\beta}}(|E_{\hat{\beta}}\omega_{\hat{\beta}}(r) - E_{\beta_0}\omega_{\beta_0}(r)| > \epsilon/2). \quad (2.9)$$

By Chebyshev's inequality, the first term on the right hand side of (2.9) is smaller than $E_{\hat{\beta}}\omega^2_{\hat{\beta}}(r)/n$. From Lemma II.4 (5), we see that $E_{\hat{\beta}}\omega^2_{\hat{\beta}}(r)$ is bounded in probability when n goes to infinite so the first term goes to 0 in probability.

For the second term on the right hand side of (2.9), we note that $E_\gamma\omega_\gamma(r) = E_\gamma h(\gamma, r)$, where

$$h(\gamma, r) = \int_0^\infty (D_i - \frac{s^{(1)}_\gamma(r,t)}{s^{(0)}_\gamma(r,t)})dN_i(t).$$

Therefore, the quantity in the second term can be further controlled as follows,

$$|E_{\hat{\beta}}\omega_{\hat{\beta}}(r) - E_{\beta_0}\omega_{\beta_0}(r)| \leq |E_{\hat{\beta}}h(\hat{\beta}, r) - E_{\hat{\beta}}h(\beta_0, r)| + |E_{\hat{\beta}}h(\beta_0, r) - E_{\beta_0}h(\beta_0, r)|. \quad (2.10)$$

From Lemma II.4 (2) and (4), we can show that for any $M > 0$, $|E_{\hat{\beta}}h(\hat{\beta}, r)I_{Y<M} - E_{\hat{\beta}}h(\beta_0, r)I_{Y<M}|$ goes to 0 in probability. From Lemma II.4 (1) and Assumption II.1, we can show that $|E_{\hat{\beta}}h(\hat{\beta}, r)I_{Y\geq M} - E_{\hat{\beta}}h(\beta_0, r)I_{Y\geq M}|$ will goes to 0 when M goes to infinite, so we can control the 1st term on the right hand side of (2.10). For the second part, from Assumption II.1, we note that $h(\beta_0, r)$ is continuous r.v. w.r.t Y, so the second part on the right hand side of (2.10) will go to 0 due to portmanteau lemma.

(3) Last, we will show the result, $P_{\hat{\beta}}(|\tilde{\beta} - \beta_0| > \epsilon) \to 0$ in probability w.r.t P_1.

From *Lin and Wei* (1989), we know that $r = \beta_0$ is the solution of $E_{\beta_0}\omega_{\beta_0}(r) = 0$. Since $E_\beta\omega_\beta(r) = E_\beta h(\beta, r)$ and $h(\beta, r)$ is monotone to r, the solution of $E_{\beta_0}\omega_{\beta_0}(r) = 0$ is unique and we prove the result.

□

Lemma II.6. *Under Assumption II.1, for any $\epsilon > 0$, $P_{\hat{\beta}}(|\bigtriangledown^2 l_n(\beta_n) - A_0(\beta_0)| > \epsilon) \to 0$ in probability w.r.t P_1, where β_n is between $\hat{\beta}$ and $\tilde{\beta}$. Furthermore, $A_0(\beta_0)$ is positive definite.*

Proof. W.L.O.G, we assume that $\hat{\beta} \in \mathbb{B}$. With Lemma II.4 (2) and (4), we can show that for any $\tau < \infty$,

$$P_{\hat{\beta}}(\sup_{r\in\mathbb{B},t\in[0,\tau]} |(\frac{s^{(2)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)} - (\frac{s^{(1)}_{\hat{\beta}}(r,t)}{s^{(0)}_{\hat{\beta}}(r,t)})^2) - (\frac{S^{(2)}(r,t)}{S^{(0)}(r,t)} - (\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)})^2)| > \epsilon) \to 0$$

in probability w.r.t P_1. From the definition, we note that

$$\begin{aligned}
&\bigtriangledown^2 l_n(\beta_n) - \bigtriangledown^2 \tilde{l}_{\hat{\beta},n}(\beta_n) \\
=&\sum_{i=1}^{n} \delta_i \left\{ \Big(\frac{s^{(2)}_{\hat{\beta}}(r,Y_i)}{s^{(0)}_{\hat{\beta}}(r,Y_i)} - (\frac{s^{(1)}_{\hat{\beta}}(r,Y_i)}{s^{(0)}_{\hat{\beta}}(r,Y_i)})^2 \Big) - \Big(\frac{S^{(2)}(r,Y_i)}{S^{(0)}(r,Y_i)} - (\frac{S^{(1)}(r,Y_i)}{S^{(0)}(r,Y_i)})^2 \Big) \right\} /n.
\end{aligned} \tag{2.11}$$

Thus, with appropriate partition for Y_i, $P_{\hat{\beta}}(|\bigtriangledown^2 l_n(\beta_n) - \bigtriangledown^2 \tilde{l}_{\hat{\beta},n}(\beta_n)| > \epsilon) \to 0$ in probability w.r.t P_1. Similar to the techniques we use in Lemma II.5 and notice that $\bigtriangledown^2 l_n(r)$ and $\bigtriangledown^2 \tilde{l}_{\hat{\beta},n}(r)$ are always bounded, we can show that

$$\begin{aligned}
&P_{\hat{\beta}}(|\bigtriangledown^2 l_n(\beta_n) - A_0(\beta_0)| > \epsilon) \\
\sim &P_{\hat{\beta}}(|\bigtriangledown^2 \tilde{l}_{\beta_n,n}(\beta_n) - E_{\hat{\beta}} \bigtriangledown^2 \tilde{l}_{\beta_n,n}(\beta_n)| > \epsilon/2) + P_{\hat{\beta}}(|E_{\hat{\beta}} \bigtriangledown^2 \tilde{l}_{\beta_n,n}(\beta_n) - A_0(\beta_0)| > \epsilon/2),
\end{aligned} \tag{2.12}$$

and the latter goes to 0 in probability w.r.t P_1. To be specific, the first term on the right hand side of (2.12), $P_{\hat{\beta}}(|\bigtriangledown^2 \tilde{l}_{\beta_n,n}(\beta_n) - E_{\hat{\beta}} \bigtriangledown^2 \tilde{l}_{\beta_n,n}(\beta_n)| > \epsilon/2)$, is controlled by Chebyshev's inequality and the second term, $P_{\hat{\beta}}(|E_{\hat{\beta}} \bigtriangledown^2 \tilde{l}_{\beta_n,n}(\beta_n) - A_0(\beta_0)| > \epsilon/2)$, is controlled by Lemma II.4 (1), (2), (4) and Assumption II.1. The technique is similar to what we use in the proof of Lemma II.5 and (2.10).

Since F_{β_0} is well defined with true log-hazard ratio as implied in the problem setting, from *Lin and Wei* (1989), it is not hard to see that $A_0(\beta_0)$ is positive definite.

□

Lemma II.7. *Under Assumption II.1,* $\sqrt{n} \bigtriangledown l_n(\hat{\beta}) \to N(0, \sqrt{E_{\beta_0}\omega^2(\beta_0)})$ *or, in other words, for any c,* $P_{\hat{\beta}}(\sqrt{n} \bigtriangledown l_n(\hat{\beta}) > c) \to F(c)$*, where F is the survival function of* $N(0, E_{\beta_0}\omega^2(\beta_0))$*, in probability w.r.t* P_1*.*

Proof. W.L.O.G, we assume $\hat{\beta} \in \mathbb{B}$. From Lemma II.4 (1), (3) and (4), we can prove that $P_{\hat{\beta}}(\sqrt{n} \bigtriangledown l_n(\hat{\beta}) > c) \sim P_{\hat{\beta}}(\sqrt{n} \bigtriangledown \tilde{l}_{\hat{\beta},n}(\hat{\beta}) > c)$ by modifying the 1st part of the proof in Lemma II.5. and showing that $\sqrt{n}(S^{(0)}(\hat{\beta}, t) - s^{(0)}_{\hat{\beta}}(\hat{\beta}, t))$ converges to a zero-mean Gaussian process in probability w.r.t P_1. Next, we check the Lindebeger-Feller condition for CLT, $E_{\hat{\beta}}\omega^2_{\hat{\beta}}(\hat{\beta})I_{|\omega_{\hat{\beta}}(\hat{\beta})|>\sqrt{n\epsilon_1}} \to 0$, for any $\epsilon_1 > 0$. We have the following decomposition,

$$\begin{aligned} & E_{\hat{\beta}}\omega^2_{\hat{\beta}}(\hat{\beta})I_{|\omega_{\hat{\beta}}(\hat{\beta})|>\sqrt{n\epsilon_1}} \\ =& E_{\hat{\beta}}\omega^2_{\hat{\beta}}(\hat{\beta})I_{|\omega_{\hat{\beta}}(\hat{\beta})|>\sqrt{n\epsilon_1}}I_{Y<M} + E_{\hat{\beta}}\omega^2_{\hat{\beta}}(\hat{\beta})I_{|\omega_{\hat{\beta}}(\hat{\beta})|>\sqrt{n\epsilon_1}}I_{Y\geq M}. \end{aligned} \tag{2.13}$$

From Lemma II.4 (2) and (4), when $Y < M$ and $\gamma, r \in \mathbb{B}$, $\omega^2_\gamma(r)$ is bounded, so the first term of (2.13) will go to 0 in probability w.r.t P_1. The second term is smaller than $E_{\hat{\beta}}\omega^2_{\hat{\beta}}(\hat{\beta})I_{Y\geq M}$. From Lemma II.4 (5), we know that $E_{\hat{\beta}}\omega^2_{\hat{\beta}}(\hat{\beta})I_{Y\geq M} \to 0$ in probability w.r.t P_1 as $M \to \infty$.

Since F_β is well defined for $\beta \in \mathbb{B}$ as implied in the problem setting, by *Lin and Wei* (1989), $E_{\hat{\beta}}\omega_{\hat{\beta}}(\hat{\beta}) = 0$. Furthermore, with similar decomposition techniques used in the proof of Lemma II.5 and (2.10), we note that $|E_{\hat{\beta}}\omega^2_{\hat{\beta}}(\hat{\beta}) - E_{\beta_0}\omega^2_{\beta_0}(\beta_0)| \to 0$ in probability w.r.t P_1 from Lemma II.4 (5). Therefore, we show the normality as desired.

□

Theorem II.8. *Under Assumption II.1,* $\sqrt{n}(\tilde{\beta} - \hat{\beta}) \to N(0, \sigma_0^2)$ *in probability w.r.t* P_1. $\sigma_0^2 = A_0^{-2}(\beta_0)E_{\beta_0}\omega^2_{\beta_0}(\beta_0)$.

Proof. Take taylor expansion of $\bigtriangledown l_n(r)$ at $r = \hat{\beta}$ and apply Lemmas II.5–II.7, we can get the result. □

We define $\bigtriangledown l_{P_1,n}(r)$, $\bigtriangledown^2 \tilde{l}_{P_1,n}(r)$, $\omega_{P_1,i}(\beta)$ and $A_{P_1}(r)$ similar to the quantities in the 1st paragraph but replace $s^{(q)}_\gamma(\beta, Y_i)$, $\bar{F}_\gamma(t)$, E_{β_i}, $A_0(r)$ with $s^{(q)}_{P_1}(\beta, Y_i)$, $\bar{F}_{P_1}(t)$, E_{P_1} and $A_{P_1}(r)$ and the latter are with regard to P_1 instead of F_γ. Let $\{(Y_i^*, \delta_i^*, D_i^*)\}_{i=1}^n$ be n random samples (bootstrap sample) from $\{(Y_i, \delta_i, D_i)\}_{i=1}^n$. We let $\bigtriangledown l_n^*(r)$, $\bigtriangledown^2 l_n^*(r)$ be the first and second derivative of the log partial likelihood of the bootstrap sample, and β^* be the bootstrap estimator. We let $S^{(q,*)}(\beta, t) = \sum_{i=1}^n Y_i^*(t)e^{\beta D_i^*} D_i^{*q}/n$, $\bigtriangledown \bar{l}_n(r) = \sum_{i=1}^n \bar{\omega}_i(r)/n$, $\bigtriangledown \bar{l}_n^*(r) = \sum_{i=1}^n \bar{\omega}_i^*(r)/n$ and $\bigtriangledown^2 l_n^*(r) = \sum_{i=1}^n \delta_i^* \left\{ \frac{S^{(2)}(r, Y_i^*)}{S^{(0)}(r, Y_i^*)} - (\frac{S^{(1)}(r, Y_i^*)}{S^{(0)}(r, Y_i^*)})^2 \right\}/n$, where

$$\bar{\omega}_i(\beta) = \int_0^\infty \left\{ D_i - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} dN_i(t) - \int_0^\infty \frac{Y_i(t)e^{\beta D_i}}{S^{(0)}(\beta, t)} \left\{ D_i - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} dF_n(t),$$

$$\bar{\omega}_i^*(\beta) = \int_0^\infty \left\{ D_i^* - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} dN_i^*(t) - \int_0^\infty \frac{Y_i^*(t)e^{\beta D_i^*}}{S^{(0)}(\beta, t)} \left\{ D_i^* - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} dF_n^*(t),$$

and $N_i^*(t) = I_{\{Y_i^* \le t, \delta_i^* = 1\}}$ and $F_n^*(t) = \sum_{i=1}^n N_i^*(t)/n$. The following theorem shows the bootstrap consistency under misspecified cox model.

Theorem II.9. *Under Assumption II.1, the bootstrap is consistent;* $\sqrt{n}(\beta^* - \hat{\beta}) \to N(0, \sigma_{P_1}^2)$ *in probability w.r.t* P_1*, where* $\sigma_{P_1}^2 = A_{P_1}^{-2}(\beta_0) E_{P_1} \omega_{P_1}^2(\beta_0)$.

Proof. The proof is similar to the proof in Theorem A.5.

(1) We will construct similar results as Lemma II.4.

First, for $q = 0, 1, 2$, $\sup_{\beta \in \mathbb{B}, t \in R} |S^{(q,*)}(\beta, t)/S^{(0,*)}(\beta, t)|$ is bounded.

Second, for any $\epsilon > 0$, $P^*(\sup_{\beta \in \mathbb{B}, t \in R} |S^{(q,*)}(\beta, t) - S^{(q)}(\beta, t)| > \epsilon) \to 0$ in probability w.r.t P_1.

Third, there exists subsequence $S^{(0)'}(\beta, t)$ of $S^{(0)}(\beta, t)$, $P_1(\text{for any } \tau < \infty, \liminf_{\beta \in \mathbb{B}, t \in [0,\tau]} S^{(0)'}(\beta, t) > 0) = 1$.

Fourth, $sup_{\beta \in \mathbb{B}} E_{F_n}(\int_0^\infty \frac{Y(t) e^{\beta D}}{S^{(0)}(\beta, t)} dF_n(t))^2 < \infty$ and

$$P_1(\lim_{M \to \infty} \lim_{n \to \infty} \sup_{\beta \in \mathbb{B}} E_{F_n}(\int_M^\infty \frac{Y(t) e^{\beta D}}{S^{(0)}(\beta, t)} dF_n(t))^2 = 0) = 1.$$

The proof of the first is trivial. Under Assumption II.1, we note that

$$P_1(\sup_{\beta \in \mathbb{B}, t \in R} |S^{(q)}(\beta, t) - s_{P_1}^{(q)}(\beta, t)| > \epsilon) \to 0.$$

Notice that $E^*S^{(q,*)}(\beta,t) = S^{(q)}(\beta,t)$, we prove the second with similar decomposition as the proof in Lemma II.4 (2) and Chebyshev's inequality. For the third, we can show that $s_{P_1}^{(0)}(\beta,t)$ is bounded below for any $\beta \in \mathbb{B}$ and $t \in [0,\tau]$. Combining similar arguments in the second one can lead to the result. Notice that $F_n \to F_{P_1}$ a.s. w.r.t P_1, the proof of the last one is similar to the proof of Lemma II.4 (5). W.L.O.G, we assume that $P_1(\text{for any } \tau < \infty, lim\inf_{\beta\in\mathbb{B},t\in[0,\tau]} S^{(0)}(\beta,t) > 0) = 1$ and $P^*(\sup_{\beta\in\mathbb{B},t\in R}|S^{(q,*)}(\beta,t) - S^{(q)}(\beta,t)| > \epsilon) \to 0$ a.s. w.r.t P_1.

(2) Second, we will show that $P^*(\sqrt{n}|\bigtriangledown l_n^*(r) - \bigtriangledown\bar{l}_n^*(r)| > \epsilon) \to 0$ in probability.

Similar to the proof in Lemma II.5, we have the following decomposition

$$\begin{aligned}
&\sqrt{n}(\bigtriangledown l_n^*(r) - \bigtriangledown\bar{l}_n^*(r)) \\
=&-\int_0^\infty \{\frac{S^{(1,*)}(r,t)}{S^{(0,*)}(r,t)} - \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)}\} d\{\sqrt{n}(F_n^*(t) - F_n(t)\} \\
&-\int_0^\infty \{\frac{S^{(1,*)}(r,t)}{S^{(0,*)}(r,t)} - \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)}\} \sqrt{n}(S^{(0,*)}(r,t) - S^{(0)}(r,t))/S^{(0)}(r,t) dF_n(t).
\end{aligned} \tag{2.14}$$

From (1), we note that for any $\tau < \infty$, $lim\inf_{\beta\in\mathbb{B},t\in[0,\tau]} S^{(0)}(\beta,t) > 0$ a.s. w.r.t P_1. Therefore, we can decompose

$$\int_0^\infty \left\{ \frac{S^{(1,*)}(r,t)}{S^{(0,*)}(r,t)} - \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} \right\} d\{\sqrt{n}(F_n^*(t) - F_n(t)\}$$

into

$$\int_0^M \left\{ \frac{S^{(1,*)}(r,t)}{S^{(0,*)}(r,t)} - \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} \right\} d\{\sqrt{n}(F_n^*(t) - F_n(t)\} +$$

$$\int_M^\infty \left\{ \frac{S^{(1,*)}(r,t)}{S^{(0,*)}(r,t)} - \frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} \right\} d\{\sqrt{n}(F_n^*(t) - F_n(t)\}.$$

Since $P^*(\sup_{\beta\in\mathbb{B},t\in R}|S^{(r,*)}(\beta,t)-S^{(q)}(\beta,t)|>\epsilon)\to 0$ and $\sqrt{n}(F_n^*(t)-F_n(t))$ converges to a zero-mean Gaussian process in probability w.r.t P_1, we can apply similar arguments as the proof of Lemma II.5 and control the 1st term on the right hand side of (2.14). We get the results by controlling the 2nd term on the right hand side of (2.14) with similar techniques as used in the proof of Lemma II.5.

(3) Third, we will show that $P^*(|\beta^*-\beta_0|>\epsilon)\to 0$ in probability w.r.t P_1.

We can get the following decomposition

$$\begin{aligned}
&P^*(|\bigtriangledown l_n^*(r)-E_{P_1}\omega_{P_1}(r)|>\epsilon)\\
\leq&P^*(|\bigtriangledown l_n^*(r)-\bigtriangledown\bar{l}_n^*(r)|>\epsilon/3)+\\
&P^*(|\bigtriangledown \bar{l}_n^*(r)-E^*\bigtriangledown \bar{l}_n^*(r)|>\epsilon/3)+P^*(|E^*\bigtriangledown \bar{l}_n^*(r)-E_{P_1}\omega_{P_1}(r)|>\epsilon/3).
\end{aligned}\tag{2.15}$$

The first term on the right hand side of (2.15) is controlled by what we prove in part (2) of this proof. Similar to what *Lin and Wei* (1989) already showed, we note that $var^*(n\bar{l}_n^*(r))\to var_{P_1}\omega_{P_1}^2(r)$ in probability w.r.t P_1. The second term on the right hand side of (2.15) is controlled by Chebyshev's inequality. Notice that $E^*\bigtriangledown\bar{l}_n^*(r)=\bigtriangledown\bar{l}_n(r)$, the third term is controlled by the consistency of $\bigtriangledown\bar{l}_n(r)$ to $E_{P_1}\omega_{P_1}(r)$. Similar arguments as the proof of Lemma II.5 show that β_0 is the unique solution of $E_{P_1}\omega_{P_1}(r)=0$, so we prove the result.

(4) Fourth, we show that $\sqrt{n}\bigtriangledown\bar{l}_n^*(\hat{\beta})$ is asymptotically normal in probability w.r.t P_1. In other words, we show that $P^*(\bigtriangledown\bar{l}_n^*(\hat{\beta})>c)\to F(c)$ in probability w.r.t P_1, where $F\sim N(0,E_{P_1}\omega_{P_1}^2(\beta_0))$.

It is not hard to see that $E^*\bigtriangledown\bar{l}_n^*(\hat{\beta})=0$ and $\bigtriangledown\bar{l}_n^*(\hat{\beta})$ is i.i.d sum of $\bar{\omega}^*(\hat{\beta})$ w.r.t P^*.

We note that

$$\begin{aligned} & E^*(\bar{\omega}^*(\hat{\beta}))^2 I_{|\bar{\omega}^*(\hat{\beta})|>\sqrt{n\epsilon}} \\ =& E^*(\bar{\omega}^*(\hat{\beta}))^2 I_{|\bar{\omega}^*(\hat{\beta})|>\sqrt{n\epsilon}} I_{Y^*<M} + E^*(\bar{\omega}^*(\hat{\beta}))^2 I_{|\bar{\omega}^*(\hat{\beta})|>\sqrt{n\epsilon}} I_{Y^*\geq M} \\ \leq& E^*(\bar{\omega}^*(\hat{\beta}))^2 I_{|\bar{\omega}^*(\hat{\beta})|>\sqrt{n\epsilon}} I_{Y^*<M} + E^*(\bar{\omega}^*(\hat{\beta}))^2 I_{Y^*\geq M}. \end{aligned} \tag{2.16}$$

The first term on the right hand side of (2.16) will go to 0 in probability w.r.t P_1 due to the boundness $\bar{\omega}$ when $Y^* < M$. The second term is asymptotically bounded by $E_{P_1}\omega^2_{P_1}(\beta_0)I_{Y\geq M}$ and the latter goes to 0 when $M \to \infty$. As showed in *Lin and Wei* (1989), $\sum_{i=1}^n \bar{\omega}^{*2}(\hat{\beta})/n$ is consistent to $E_{P_1}\omega^2_{P_1}(\beta_0)$. Therefore, we can apply Lindeberg-Feller CLT.

In the end, similar to Lemma II.6, $P^*(|\bigtriangledown^2 l^*(\beta) - \bigtriangledown^2 \bar{l}^*(\beta)| > \epsilon) \to 0$ and $P^*(|\bigtriangledown^2 \bar{l}^*(\beta) - A_{P_1}(\beta_0)| > \epsilon) \to 0$ in probability w.r.t P_1 for $\beta \in (\hat{\beta}, \beta^*)$. Combining all the above, we show the result by taking taylor expansion of $\bigtriangledown \bar{l}^*(r)$ at $r = \hat{\beta}$. Since β_0 is well defined, $A^{-2}_{P_1}(\beta_0)$ is also positive.

□

If $\beta_1 = \beta_2$, then, it is obvious that $\sigma_0 = \sigma_{P_1}$. Let $G(\cdot|\mu_1, \mu_2, \sigma_{x_1}, \sigma_{x_2}, \rho)$ be the survival function of $max(X1, X2)$, where $(X1, X2)$ follows a joint normality with population mean μ_1 and μ_2, standard deviation σ_{x_1} and σ_{x_2} and correlation ρ respectively. We have the following lemma.

Lemma II.10. *Under Assumption II.1,*

$$P_{\hat{\beta}}(\sqrt{n}\max(\tilde{\beta}_1 - \hat{\beta}, \tilde{\beta}_2 - \hat{\beta}) \geq c) \to G(c|0, 0, \frac{\sigma_0}{\sqrt{p}}, \frac{\sigma_0}{\sqrt{1-p}}, 0)$$

in probability w.r.t P_1.

Proof. Let $(U_n, V_n) = 1/\sqrt{n}(\sum_{i=1}^n \omega_{\hat{\beta},i}(\hat{\beta})I_{Z_i=1}, \sum_{i=1}^n \omega_{\hat{\beta},i}(\hat{\beta})I_{Z_i=0})$. Similar to Lemma II.7, we can show that (U_n, V_n) are jointly normal in asymptotic sense. Therefore, by taking taylor expansion as we did in Theorem II.8, we have $\sqrt{n}(\tilde{\beta}_1 - \hat{\beta})$ and $\sqrt{n}(\tilde{\beta}_2 - \hat{\beta})$ are jointly normal in asymptotic sense. With Theorem II.8, we note that $\sqrt{n}(\tilde{\beta}_1 - \hat{\beta}) \sim N(0, \frac{\sigma_0^2}{p})$ and $\sqrt{n}(\tilde{\beta}_2 - \hat{\beta}) \sim N(0, \frac{\sigma_0^2}{1-p})$ and are asymptotically independent, which proves the lemma. □

Proof of Theorem II.3: It is easy to see that conditional on the bootstrap sample we generate in Algorithm 1, $\sqrt{n}(\beta_1^* - \hat{\beta})$ and $\sqrt{n}(\beta_2^* - \hat{\beta})$ are independent. Therefore, we can show that $\sqrt{n}\max(\beta_1^* - \hat{\beta}, \beta_2^* - \hat{\beta}) \to G(\cdot|0, 0, \frac{\sigma_{P_1}}{\sqrt{p}}, \frac{\sigma_{P_1}}{\sqrt{1-p}}, 0)$ in probability w.r.t P_1 by Theorem II.9. We note that

$$P^*(\max(\beta_1^*, \beta_2^*) \geq \max(\hat{\beta}_1, \hat{\beta}_2)) = P^*(\sqrt{n}\max(\beta_1^* - \hat{\beta}, \beta_2^* - \hat{\beta}) \geq \sqrt{n}\max(\hat{\beta}_1 - \hat{\beta}, \hat{\beta}_2 - \hat{\beta})).$$

Since $G(\cdot|0, 0, \frac{\sigma_{P_1}}{\sqrt{p}}, \frac{\sigma_{P_1}}{\sqrt{1-p}}, 0)$ is continuous, we can show that

$$P^*(\max(\beta_1^*, \beta_2^*) \geq \max(\hat{\beta}_1, \hat{\beta}_2)) \sim G(\sqrt{n}\max(\hat{\beta}_1 - \hat{\beta}, \hat{\beta}_2 - \hat{\beta})|0, 0, \frac{\sigma_{P_1}}{\sqrt{p}}, \frac{\sigma_{P_1}}{\sqrt{1-p}}, 0).$$

Similarly, by Lemma II.10, we have the following relationship

$$P_{\hat{\beta}}(\max(\tilde{\beta}_1, \tilde{\beta}_2) \geq \max(\hat{\beta}_1, \hat{\beta}_2)) \sim G(\sqrt{n}\max(\hat{\beta}_1 - \hat{\beta}, \hat{\beta}_2 - \hat{\beta})|0, 0, \frac{\sigma_0}{\sqrt{p}}, \frac{\sigma_0}{\sqrt{1-p}}, 0).$$

If $\beta_1 = \beta_2$, then,

$$\begin{aligned} &G(\sqrt{n}\max(\hat{\beta}_1 - \hat{\beta}, \hat{\beta}_2 - \hat{\beta})|0, 0, \frac{\sigma_0}{\sqrt{p}}, \frac{\sigma_0}{\sqrt{1-p}}, 0) \\ =&G(\sqrt{n}\max(\hat{\beta}_1 - \hat{\beta}, \hat{\beta}_2 - \hat{\beta})|0, 0, \frac{\sigma_{P_1}}{\sqrt{p}}, \frac{\sigma_{P_1}}{\sqrt{1-p}}, 0). \end{aligned} \tag{2.17}$$

If $\beta_1 \neq \beta_2$, then, from Assumption II.2, we note that $\beta_0 < max(\beta_1, \beta_2)$ and $\sqrt{n}max(\hat{\beta}_1 - \hat{\beta}, \hat{\beta}_2 - \hat{\beta}) \to \infty$ in probability w.r.t P_1, so the risk index will go to 0. The Theorem

II.3 is proved. □

Corollary II.11. *Under Assumptions II.1 and II.2, if $\beta_1 = \beta_2$, the risk index will converge to a non-degenerate distribution on (0,1); Otherwise, the risk index will converge to 0.*

Proof. See the proof in Theorem II.3. □

CHAPTER III

Subgroup Analysis: Debiased Inference

In this chapter, we propose a resampling-based debiased inference procedure for the best selected subgroup. When a promising subgroup is selected, we have to answer the question of how good the selected subgroup is. The proposed method addresses the subgroup selection bias appropriately by bootstrapping the bias, and provides 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 mimic the subgroup selection procedure, 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 to help a better-informed decision on subgroup pursuit. We demonstrate the merit of our proposed method by re-analyzing the MONET1 trial and answer the question of how good the East Asian subgroup really is. We show that how the subgroup is selected post hoc should play an important role in any statistical analysis.

3.1 Inference with Predefined Subgroups

In this section, we start from a relatively simple scenario in subgroup analysis where the candidate subgroups are predefined. We propose bootstrap-based asymp-

totically sharp inference and a bias-reduced estimator on the effect size of the best selected subgroup.

3.1.1 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 ith subgroup, respectively, for $i = 1, \ldots, k$. 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 ith 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. Indeed, our proposed method works for any measure of the treatment effect as long as the treatment effect measure satisifes two very mild assumptions as specified in Section 3.1.3. Moreover, we even do not need to assume the model we use for defining and calculating the treatment effect measure is correctly specified and the proposed method works for many misspecified models, such as the misspecified proportional hazard model as discussed in Section 2.1.1. Without sticking to a specific model to measure the treatment effect and assuming the model is correctly specified, we say the proposed method is model-free. Without loss of generality, we assume that a larger value of β_i means a better treatment effect.

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

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

Note that $\beta_{\max}$ 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 $\beta_{\max}$ and show that the same procedure works for the inference 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 the root-n rate of convergence, any inference based on the estimated skew-normal distribution is no longer valid. Of course, the problem does not become less challenging when $k > 2$, which calls for a new inferential method to be developed.

3.1.2 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 $j = 1, \cdots, n$ 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 $\{D_j^*, Z_j^*\}, j = 1, \cdots, n$, by drawing n subjects with replacements. The subgroup treatment effects for the bootstrapped sample are then denoted by $\hat{\beta}_i^*$ for $i = 1, \cdots, k$. 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 Algorithm 2.

Algorithm 2 Lower confidence limit for β_{max}.

1: For $i = 1, \ldots, k$, set $d_i = (1 - n^{r-0.5})(\hat{\beta}_{\max} - \hat{\beta}_i)$;
2: **for** $b = 1, \ldots, B$ **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 $c_\alpha = \text{quantile}(T^*_b, 1-\alpha)$. The level $1-\alpha$ lower confidence limit is $\hat{\beta}_{\max} - c_\alpha/\sqrt{n}$.

3.1.3 Asymptotic Validity

Just as $\hat{\beta}_{\max} = \max_{i\in[k]} \hat{\beta}_i$ is a biased estimator of $\beta_{\max}$, the bootstrap estimate $\beta^*_{\max} = \max_{i\in[k]} \beta^*_i$ 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 ith 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 smaller. The modified bootstrap estimate of β_{max} is

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

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 sugbroup. We use P and P^* to denote the probability under the sampling distribution and the bootstrap-induced distribution, respectively.

Assumption III.1 (Asymptotic normality)**.** $\sqrt{n}(\hat{\beta}_1 - \beta_1, \hat{\beta}_2 - \beta_2, \ldots, \hat{\beta}_k - \beta_k)$ *is asymptotically normal.*

Assumption III.2 (Bootstrap consistency)**.** $\sqrt{n}(\beta^*_1 - \hat{\beta}_1, \beta^*_2 - \hat{\beta}_2, \ldots, \beta^*_k - \hat{\beta}_k)$ *is bootstrap consistent, that is, conditional on the data, the asymptotic distribution of* $\sqrt{n}(\beta^*_1 - \hat{\beta}_1, \beta^*_2 - \hat{\beta}_2, \ldots, \beta^*_k - \hat{\beta}_k)$ *is the same as the limiting distribution in Assumption III.1. in probability.*

In typical parametric and semi-parametric models, Assumption III.1 is satisfied

for a wide range of estimators $\hat{\beta}_i$. Assumption III.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 III.3. *Under Assumptions III.1 and III.2, and for any* $0 < r < 0.5$*, we have,*

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

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

Theorem III.3 confirms that under very mild assumptions, the proposed inference for $\beta_{\max}$ is asymptotically sharp in the sense that the confidence interval we propose in Algorithm 2 will cover $\beta_{\max}$ exactly at the nominal level we choose as the sample size goes to infinite. The following corollary facilitates inference on β_s.

Corollary III.4. *Under Assumptions III.1 and III.2, and for any* $0 < r < 0.5$*, we have*

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

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

Corollary III.4 indicates that the proposed bootstrap-based confidence interval for $\beta_{\max}$ can also serve as an asymptotically sharp prediction interval for β_s. Therefore, we can use the same procedure to infer 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 3.1.5.

3.1.4 Bias-reduced Estimator

Following the results in Theorem III.3, we propose a bias-reduced estimator for $\beta_{\max}$, and a biased-reduced predictor for β_s. Note that the bias $E[\hat{\beta}_{max} - \beta_{max}]$ is $o(1/\sqrt{n})$ when the number of the best subgroups is 1 (e.g., $\beta_1 > \beta_2$ in the case of two subgroups), but the bias is in the order of $O(1/\sqrt{n})$ and non-negligible when the number of the best subgroups is more than 1 (e.g., $\beta_1 = \beta_2$). To be more specific, let $H = \{i : \beta_i = \beta_{\max}\}$. Then, the number of the best subgroups is just the size of H, $|H|$.

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

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

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

Assumption III.5 (2nd moment bound)**.** $\limsup_{n\to\infty} E[\sqrt{n}(\hat{\beta}_i - \beta_i)]^2 < \infty$*, for* $i = 1, \ldots, k$*.*

Assumption III.6 (2nd bootstrap moment)**.** $\limsup_{n\to\infty} E^*[\sqrt{n}(\beta^*_i - \hat{\beta}_i)]^2 < \infty$*, in probability, for* $i = 1, \ldots, k$*.*

Theorem III.7. *Under Assumptions III.1, III.2, III.5 and III.6, and for any* $0 < r < 0.5$*, we have*

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

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

Theorem III.7 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 III.8. $\limsup_{n\to\infty} E\{E^*[\sqrt{n}(\beta_i^* - \hat{\beta}_i)]^2\} < \infty$, *for* $i = 1, \ldots, k$,

we can have the following result.

Corollary III.9. *Under Assumptions III.1, III.2, III.5 and III.8, and for any* $0 < r < 0.5$, *we have*

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

Corollary III.9 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 $o(1/\sqrt{n})$.

3.1.5 Choice of the Tuning Parameter

A smaller value of the tuning parameter r in Algorithm 2 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 $\beta_{\max}$. To make this possible without knowing the true value of $\beta_{\max}$, 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 $A = \{r_1, \ldots, r_m\}$ denote a set of possible tuning parameters in the range of $(0, 0.5)$ with $r_1 < \cdots < r_m$ and m is a finite integer. The following algorithm can be used to choose $r \in A$.

Algorithm 3 Cross-validated choice of tuning parameter r.

1: Randomly partition the data into v (approximately) equal-sized subsamples;
2: **for** $l = 1, \ldots, m$ **do**
3: **for** $j = 1, \ldots, v$ **do**
4: **Basic setup**: use the jth 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** $i = 1, \ldots, k$ **do**
7: **Calculations on the testing data**: use the reference data to estimate the effect size of the ith subgroup, $\hat{\beta}_{i,j}$, and its standard error $\hat{\sigma}_{i,j}$;
8: **Evaluation of accuracy**: calculate $h_{i,j}(r_l) = (\hat{\beta}_{\max,\text{reduced},j}(r_l) - \hat{\beta}_{i,j})^2 - \hat{\sigma}^2_{i,j}$;
9: **end for**
10: **end for**
11: **end for**
12: The tuning parameter is chosen to be $argmin_{r_l}\{\min_{i\in[k]}[\sum_{j=1}^{j=v} h_{i,j}(r_l)/v]\}$.

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 III.10. *Under the assumptions of Corollary III.9 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 III.10 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 $\hat{\sigma}^2_{i,j}$ in the calculation of $h_{i,j}(r_l)$ in Step 8 of the above algorithm is to account for the variation in $\hat{\beta}_{i,j}$ used there.

3.2 Inference with Post-hoc Identified Subgroups

In this section, we generalize the procedure to 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; i.e., $\max_{c \in D} \hat{\beta}(c)$ exists almost surely.

3.2.1 Asymptotically Sharp Inference

We generalize the inference procedure for the predefined subgroups in Section 3.1 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 4 Lower confidence limit for $\gamma_{\max}$.

1: For $c \in D$, let $d(c) = (1 - n^{r-0.5})(\hat{\gamma}_{\max} - \hat{\beta}(c))$;
2: **for** $b = 1, \ldots, B$ **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 III.11 (Asymptotically Gaussian process)**.** $\sqrt{n}(\hat{\beta}(\cdot) - \beta(\cdot)) \to^d G(\cdot)$ *in* $l_\infty(D)$*, where* $G(\cdot)$ *is a Gaussian process with continuous sample path in probability.*

Assumption III.12 (Bootstrap consistency)**.** $\sqrt{n}(\beta^*(\cdot) - \hat{\beta}(\cdot)) \to^d G(\cdot)$ *in* $l_\infty(D)$ *in probability.*

Assumption III.13 (Continuous mapping)**.** $c \to \beta(c)$ *is a continuous mapping in* D*.*

Theorem III.14. *Under Assumptions III.11, III.12 and III.13 and for any* $0 < r < 0.5$*, we have as* $n \to \infty$*,*

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

in probability w.r.t P*.*

Theorem III.14 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 versions of Assumptions III.1 and III.2. If the (bootstrap) estimated effect size can be written in a form of an empirical process, then, Assumptions III.11 and III.12 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 3.1.4, we can have a bias-reduced estimator of $\gamma_{\max}$ as

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

3.2.2 Selected Subgroup Inference

Previously in the case of predefined subgroups, the inference procedure in Section 3.1.2 works for both $\beta_{\max}$ 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 $\sqrt{n}(\beta_s - \beta_{\max}) \to 0$ 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 γ_s.

From the proof of Theorem III.14, we see that, asymptotically, the one-sided confidence interval for $\gamma_{\max}$ 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 3.2 locally simultaneous confidence intervals, which is in contrasts to any inference based on a (globally) simultaneous confidence band of $\beta(c)$ for all $c \in D$ and make the proposed procedure asymptotically sharp when inferring $\gamma_{\max}$. Because $K \subset D$, the resulting inference is more efficient than the methods based on simultaneous confidence bands (interval) such as that of *Fuentes et al.* (2018) and the latter is clearly not asymptotically sharp and on the conservative side for the inference on $\gamma_{\max}$.

Furthermore, although γ_s may not equal $\gamma_{\max}$ with probability one, it falls into a local neighborhood of K, which shrinks to K as the sample size increases. This enables us to establish the following result, analogous to Theorem III.7 for β_s.

Theorem III.15. *Under the assumptions of Theorem III.14, we have, as $n \to \infty$,*

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

in probability w.r.t P.

3.3 Synthetic Data: MONET1 Continued

In this section, as a follow-up to Section 2.2, we revisit the failed MONET1 trial again. We demonstrate the merit of our proposed debiased inference tool by re-analyzing the failed MONET1 trial. With our proposed method, we provide an

appropriate guidance on subgroup pursuit decisions after the initial MONET1 trial data are available and answer the question of how good the East Asian subgroup really is in a more scientific way.

Similar to what we did in Section 2.2, we apply the proposed debiased inference procedure on the best subgroup effect to the synthetic data of MONET-1. From the arguments in Section 2.2, we again equate β_i to the negative log-hazard ratio to make it consistent to the framework in the chapter. We compare the proposed debiased inference procedure with the naive method which assumes that the subgroup of East Asians is not selected from the same data; see Table 3.1.

Table 3.1: 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

With the naive method for the East Asian subgroup, the hazard ratio of 0.663 is statistically significant. If only two pre-defined 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.

3.4 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 3.4.1 and 3.4.2, we evaluate the empirical coverage and bias for the predefined subgroups and the post-hoc identified subgroups, respectively. In Section 3.4.3, we compare the empirical coverage based on the synthetic data generating model used in Section 3.3.

3.4.1 Proportional Hazard Model: 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. We also consider a simultaneous inference procedure for comparison, where a lower bound for the selected subgroup effect is constructed based on the max-type statistic, $\max_{i\in[k]} \sqrt{n}(\hat{\beta}_i - \beta_i)$, as in *Hothorn et al.* (2008). For convenience, the critical value for $\max_{i\in[k]} \sqrt{n}(\hat{\beta}_i - \beta_i)$ is also estimated by bootstrap. The performance of the naive method and simultaneous 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 3.1.1, as well as the empirical bias; see Tables 3.2, 3.3 and 3.4, respectively.

Table 3.2: Empirical coverage of the 95% lower confidence bound of β_s: two predefined subgroups. The standard errors for all the entries are around 0.005. The columns correspond to different smoothing parameters r, and the column under "adaptive" corresponds to the data-dependent choice of r with 5 folds cross-validation ($v = 5$).

	$r = 1/3$	1/12	1/21	1/30	naive	adaptive	simultaneous
$\beta_2 = 0$	0.933	0.950	0.952	0.952	0.896	0.943	0.952
1/10	0.926	0.945	0.947	0.947	0.912	0.936	0.952
2/10	0.928	0.949	0.951	0.951	0.910	0.939	0.959
3/10	0.941	0.957	0.959	0.959	0.919	0.947	0.960
4/10	0.939	0.955	0.956	0.957	0.927	0.945	0.964
5/10	0.952	0.965	0.965	0.966	0.934	0.953	0.972

Table 3.3: Average distance between the 95% lower bound and β_s: two predefined subgroups.

	$r = 1/3$	1/12	1/21	1/30	naive	adaptive	simultaneous
$\beta_2 = 0$	0.248	0.265	0.266	0.266	0.213	0.258	0.269
1/10	0.252	0.269	0.270	0.270	0.218	0.262	0.274
2/10	0.267	0.285	0.288	0.287	0.233	0.277	0.295
3/10	0.290	0.311	0.313	0.314	0.258	0.302	0.323
4/10	0.301	0.326	0.328	0.329	0.273	0.313	0.344
5/10	0.310	0.339	0.342	0.343	0.286	0.323	0.350

The results show clearly that the naive method falls short in coverage probability, especially when $\beta_2 - \beta_1$ is smaller than 1/5. Although the simultaneous method can preserve the coverage probability, it is clearly on the conservative side, especially when $\beta_2 - \beta_1$ is larger than 2/5. In comparison, the proposed method preserves the coverage probability well across a broad range of choices for the tuning parameter

Table 3.4: 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

r and at the same time remains efficient. The data-adaptive choice of the tuning parameter can further improve the performance of the proposed method; it achieves better coverage and at the same time the distance between the lower confidence limit and the true value is much lower on average compared with that of the simultaneous 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 $i = 1, \cdots, k,$, and each subgroup has the same sample size 200 and the total sample size is $n = 200k$. We use the same treatment assignment and the same censoring scheme as before. To assess how much the subgroup selection bias might be, we focus on the most challenging scenario with $\beta_1 = \cdots = \beta_k = 0$, and calculate the empirical coverage of the proposed method and the naive method based on 2000 Monte Carlo repetitions. The results are summarized in Table 3.5.

From Table 3.5, 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

Table 3.5: Empirical coverage of the 95% lower bound of β_s: multiple predefined subgroups (naive).

	r=1/3	r=1/12	r=1/21	r=1/30	naive	adaptive
$k = 2$	0.929	0.952	0.953	0.953	0.900	0.939
4	0.911	0.943	0.946	0.947	0.824	0.932
6	0.891	0.941	0.943	0.945	0.739	0.930
8	0.877	0.946	0.949	0.949	0.680	0.932
10	0.866	0.944	0.949	0.950	0.594	0.927
12	0.860	0.946	0.950	0.950	0.543	0.925

generally work better.

In the end, to assess how much the asymptotically sharp property of the proposed method might help, we focus on the scenario where the best subgroup is singled out and the subgroup selection bias is at the minimums with $\beta_1 = 1$ and $\beta_2 = \cdots = \beta_k = 0$, and calculate the empirical coverage of the proposed method and the simultaneous method based on 2000 Monte Carlo repetitions. The results are summarized in Tables 3.6 and 3.7.

From Tables 3.6 and 3.7, we see that the coverage probability for the simultaneous method is clearly on the conservative side and the coverage probability is above 0.99 when there are more than 6 subgroups. On the contrary, the proposed method is not as conservative as the simultaneous method with a lower coverage probability and, more importantly, much shorter average distance. The results are somewhat more sensitive to the choice of r when the number of subgroups increases and a larger tuning parameter aims for the nominal level better. With data-adaptive choice of the tuning parameter, the proposed method preserves the coverage well and is more efficient than the simultaneous method with a 15% shorter average distance when there are more than 6 subgroups.

Table 3.6: Empirical coverage of the 95% lower bound of β_s: multiple predefined subgroups (simultaneous).

	r=1/3	r=1/12	r=1/21	r=1/30	simultaneous	adaptive
$k = 2$	0.958	0.965	0.966	0.966	0.977	0.959
4	0.955	0.978	0.974	0.980	0.989	0.964
6	0.957	0.981	0.972	0.983	0.991	0.965
8	0.963	0.991	0.993	0.993	0.997	0.973
10	0.964	0.993	0.993	0.993	0.997	0.972

Table 3.7: Average distance between the 95% lower bound and β_s: multiple predefined subgroups (simultaneous).

	r=1/3	r=1/12	r=1/21	r=1/30	simultaneous	adaptive
$k = 2$	0.311	0.336	0.340	0.342	0.365	0.323
4	0.323	0.386	0.393	0.395	0.424	0.354
6	0.330	0.416	0.424	0.426	0.453	0.377
8	0.331	0.434	0.442	0.445	0.471	0.396
10	0.340	0.454	0.462	0.465	0.489	0.413

3.4.2 Proportional Hazard Model: Post-hoc Identified Case

To continue, we consider a post-hoc identified subgroup 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 of Weibull$(1, 1)$, and the function $b(\cdot)$ is to be specified. We assume that 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 way as that in Section 3.4.1. 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 $c \in [30, 60]$. It is 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. As pointed out

in Sections 3.1.4, 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 varing 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. 3.8.

Table 3.8: 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

From Table 3.8, 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.

3.4.3 Synthetic Data Generating Model

We consider a simulation setting based on the synthetic data generating model of MONET1 in Section 3.3. We focus on the scenario of eight subgroups by the coding of the following varialbes: 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 3.4.1. To make it consistent to the convention used in MONET1, we use 2000 Monte Carlo samples in evaluating the empirical coverage of the 95% upper bound for the log hazard ratio of the best selected subgroup in Table 3.9.

Table 3.9: 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
empirical coverage	0.917	0.946	0.950	0.805

From Table 3.9, 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 accommodation for in any serious subgroup analysis.

3.5 Discussion

In this chapter, we propose a debiased inference tool for the selected subgroup. The proposed inference tool removes the subgroup selection bias and answers the question of how good the selected subgroup really is in a scientific way regardless of whether the candidate subgroups are pre-defined or identified post hoc from the data. The proposed method can lead to a better decision on subgroup pursuit.

The debiased inference tool we propose is model-free in the sense that we do not need to stick to a specific model to measure the treatment effect and the model we use for defining and calculating the treatment effect can be even misspecified. Indeed, the debiased inference tool only requires very mild assumptions as discussed in 3.1.3 and works for many treatment effect measures, including log-hazard ratio, log-odds

ratio and sample mean, which makes the proposed inference tool widely applicable.

The other property to note for the debiased inference tool is that it is asymptotically sharp in the sense that the proposed confidence interval for the best selected subgroup achieves the exact asymptotic coverage probability as desired as the sample size goes to infinite. In other words, the debiased inference tool will only remove the bias as much as is needed. Although an adjustment in the subgroup selection bias is usually needed as we search candidate subgroups to find the best subgroup, an over adjustment can lead to an inference on the conservative side for the best selected subgroup. For example, the confidence interval from simultaneous inference aims for a simultaneous coverage for all the subgroups, and is therefore too conservative and not asymptotically sharp for the subgroup we select as reviewed in Section 1.4.3 and demonstrated by simulation in Section 3.4.1. The asymptotically sharp property helps protect the proposed inference tool against over adjustments to the subgroup selection bias even as the number of candidate subgroups increases. This makes our proposed inference tool not so conservative, and thus practically useful, even if all possible candidate subgroups are taken into account.

It is worth noting that the proposed methods implicitly rely on the assumption that the difference between any two subgroups can be viewed as a constant. In applications to the studies where among some subgroups, the differences are very small, it may not be sensible to treat the difference as a constant any more and we need to investigate further how the proposed method adapts.

No matter the candidate subgroups are predefined or post-hoc identified, they are clearly specified by the biomarker. However, in practice, the subgroup might come from a clustering result and implicitly defined by the biomarker, which we call the grouping subgroup; see *Cai et al.* (2010). How to generalize the statistical tools developed here to the grouping subgroup is worth further investigation.

3.6 Proofs of Theorems III.3-III.10

To simplify the notations, let $T_i = \sqrt{n}(\hat{\beta}_i - \beta_i)$, $T_i^* = \sqrt{n}(\beta_i^* - \hat{\beta}_i)$, $d_{\max} = \beta_{\max} - \max_{i \notin H} \beta_i$ and $V = (V_1, \ldots, V_k)$ where $V_i = \hat{\beta}_{\max} - \beta_{\max} - \hat{\beta}_i + \beta_i$. We start with two lemmas that are essential for the proof of Theorem III.3.

Lemma III.16. *Under Assumption III.1, we have*

$$\sup_{x \in R} |P(\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max}) \leq x) - P(\max_{i \in H} T_i \leq x)| \to 0$$

as $n \to \infty$.

Proof. For any $x \in R$, we have

$$\begin{aligned} & |P(\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max}) \leq x) - P(\max_{i \in H} T_i \leq x)| \\ =& |P(\max_{i \in H} T_i \leq x, T_j \leq x + \sqrt{n}(\beta_{\max} - \beta_j), j \notin H) - P(\max_{i \in H} T_i \leq x)| \qquad (3.1) \\ \leq& 1 - P(\max_{i \notin H} T_i \leq x + \sqrt{n} d_{\max}). \end{aligned}$$

For any $x \in R$, by Assumption III.1, we have $1 - P(\max_{i \notin H} T_i \leq x + \sqrt{n} d_{\max}) \to 0$, so

$$|P(\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max}) \leq x) - P(\max_{i \in H} T_i \leq x)| \to 0$$

as $n \to \infty$. The result follows naturally from the property of the cdf. □

Lemma III.17. *Under Assumptions III.1 and III.2, for* $0 < r < 0.5$*, we have*

$$\sup_{x \in R} |P^*(\sqrt{n}(\beta^*_{\max,\text{modified}} - \hat{\beta}_{\max}) \leq x) - P(\max_{i \in H} T_i \leq x)| \to 0$$

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

Proof. Similar to the proof in Lemma III.16, we have

$$\begin{aligned}
&|P^*(\sqrt{n}(\beta^*_{\max,\text{modified}} - \hat{\beta}_{\max}) \leq x) - P^*(T^*_i \leq x + n^r V_i, i \in H)| \\
=&|P^*(T^*_i \leq x + n^r V_i, i \in H, T^*_j \leq x + n^r V_j + n^r(\beta_{\max} - \beta_j), j \notin H) - \\
&P^*(T^*_i \leq x + n^r V_i, i \in H)| \\
\leq&1 - P^*(T^*_j \leq x + n^r d_{\max} - n^r ||V||_\infty, j \notin H).
\end{aligned} \tag{3.2}$$

When $0 < r < 0.5$, by Assumptions III.1 and III.2, we have $n^r||V||_\infty \to 0$ in probability and

$$1 - P^*(T^*_j \leq x + n^r d_{\max} - n^r ||V||_\infty, j \notin H) \to 0$$

in probability. By Assumptions III.1 and III.2, we have

$$|P^*(T^*_i \leq x + n^r V_i, i \in H) - P(\max_{i \in H} T_i \leq x)| \to 0$$

in probability so

$$|P^*(\sqrt{n}(\beta^*_{\max,\text{modified}} - \hat{\beta}_{\max}) \leq x) - P(\max_{i \in H} T_i \leq x)| \to 0$$

in probability w.r.t P. The result is naturally followed by the property of cdf.

□

Proof of Theorem III.3: It follows from Lemmas III.16 and III.17. □

Proof of Corollary III.4: The result is true, because $\sqrt{n}(\beta_s - \beta_{\max}) \to 0$ in probability. Otherwise, by the definition of β_s, the probability of the event, $\max_{i \notin H} \hat{\beta}_i \geq \max_{i \in H} \hat{\beta}_i$, will not go to 0 asymptotically, which violates the consistency implied in Assumption III.1. □

Lemma III.18. *Under Assumptions III.1 and III.5, we have*

$$|E[\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max})] - E[\max_{i\in H} T_i]| \to 0$$

as $n \to \infty$.

Proof. By Assumptions III.5, we have

$$E[\max_{i\in H} T_i]^2 \leq E[\max_{i\in H} T_i^2] \leq \sum_{i\in H} ET_i^2 < \infty$$

and similarly $E[\max_{i\notin H} T_i]^2 < \infty$ uniformly in n, so

$$\begin{aligned} & E[\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max})]^2 \\ =& E[\max_{i\in H} T_i + \max(0, \max_{j\notin H} \sqrt{n}(\hat{\beta}_j - \max_{i\in H} \hat{\beta}_i))]^2 \\ \leq& 2\{E[\max_{i\in H} T_i]^2 + E[\max(0, \max_{j\notin H}(T_j - \max_{i\in H} T_i))]^2\} \\ <& \infty \end{aligned} \tag{3.3}$$

uniformly in n. By Assumption III.1 and Lemma III.16, the lemma follows from the uniform integrability for $\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max})$; i.e.

$$\begin{aligned} & E|\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max})| I_{|\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max})| > c} \\ \leq& [E[\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max})]^2 P(\sqrt{n}(\hat{\beta}_{\max} - \beta_{\max}) > c)]^{1/2} \to 0 \end{aligned} \tag{3.4}$$

uniformly in n as $c \to \infty$. □

Lemma III.19. *Under Assumptions III.1, III.2, and III.6, and for any* $0 < r < 0.5$*, we have*

$$|E^*[\sqrt{n}(\beta^*_{\max,\text{modified}} - \hat{\beta}_{\max})] - E \max_{i\in H} T_i| \to 0$$

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

Proof. Similar to the proof in Lemma III.18, by Assumption III.6, we have $E^*[\max_{i\in H}(T_i^* - n^r V_i)]^2 < \infty$ and $E^*[\max_{i\notin H}(T_i^* - n^r V_i)]^2 < \infty$ uniformly in n in probability, and then

$$\begin{aligned}
&E^*[\sqrt{n}(\beta^*_{\max,\text{modified}} - \hat{\beta}_{\max})]^2 \\
=&E^*[\max_{i\in H}(T_i^* - n^r V_i) + \max(0, \max_{j\notin H}(T_j^* - n^r(\beta_{\max} - \beta_j) - n^r V_j - \max_{i\in H}(T_i^* - n^r V_i)))]^2 \\
\leq&2\{E^*|\max_{i\in H}(T_i^* - n^r V_i)|^2 + E^*|\max(0, \max_{j\notin H}(T_j^* - n^r V_i - \max_{i\in H}(T_i^* - n^r V_i)))|^2\} \\
<&\infty
\end{aligned} \tag{3.5}$$

uniformly in n. The lemma then follows from Assumptions III.1, III.2 and Lemma III.17 with the similar argument of uniform integrability in Lemma III.18 but for $\sqrt{n}(\beta^*_{\max,\text{modified}} - \hat{\beta}_{\max})$. □

Proof of Theorem III.7: By Lemmas III.18 and III.19, the result follows. □

Proof of Corollary III.9: It follows from Theorem III.7 and similar arguments of the uniform integrability in Lemma III.19 but for $E^*[\sqrt{n}(\beta^*_{\max,\text{modified}} - \hat{\beta}_{\max})]$. □

Proof of Theorem III.10: By definition, we have

$$\begin{aligned}
&E[\hat{\beta}_{\max,\text{reduced},1}(r) - \beta_i]^2 \\
=&E[\hat{\beta}_{\max,\text{reduced},1}(r) - \beta_{\max}]^2 + (\beta_{\max} - \beta_i)^2 + 2E[\hat{\beta}_{\max,\text{reduced},1}(r) - \beta_{\max}](\beta_{\max} - \beta_i).
\end{aligned} \tag{3.6}$$

By Assumptions in Corollary III.9, we have $E[\hat{\beta}_{\max,\text{reduced},1}(r) - \beta_{\max}] = o(1)$, which implies the desired result. □

3.7 Proofs of Theorems III.14-III.15

To simplify the notation, we let $K+d = \{c : \text{distance}(c, K) \le d\} \bigcap D$, $\overline{K+d}$ denote the complement of the set $K+d$, $(K+d)-K = (K+d) \bigcap \overline{K}$, $G_n(c) = \sqrt{n}(\hat{\beta}(c)-\beta(c))$ and $G_n^*(c) = \sqrt{n}(\beta^*(c) - \hat{\beta}(c))$. By Assumption III.13, we know $\gamma_{\max}$ is achievable and K is a compact set, so all the above notations are well defined.

Lemma III.20. *Under Assumptions III.11 and III.13, for any $x \in R$,*

$$\lim_{n\to\infty} P(\sqrt{n}(\hat{\gamma}_{\max} - \gamma_{\max}) \le x) = P(\sup_{c\in K} G(c) \le x).$$

Proof. By definition, we have

$$\begin{aligned}
&P(\sqrt{n}(\hat{\gamma}_{\max} - \gamma_{\max}) \le x) \\
=&P(\sqrt{n}\sup_{c\in D}(\hat{\beta}(c) - \beta(c) + \beta(c) - \gamma_{\max}) \le x) \\
=&P(\max(\sup_{c\in K+d}(G_n(c) + \sqrt{n}(\beta(c) - \gamma_{\max})), \sup_{c\in\overline{K+d}}(G_n(c) + \sqrt{n}(\beta(c) - \gamma_{\max}))) \le x).
\end{aligned} \tag{3.7}$$

From Assumption III.13, $\sqrt{n}(\beta(c) - \gamma_{\max})$ converges to negative infinity uniformly in $\overline{K+d}$, so by Assumption III.11, we have

$$\sup_{c\in\overline{K+d}}(G_n(c) + \sqrt{n}(\beta(c) - \gamma_{\max})) \to -\infty$$

in probability. Therefore, the right hand side of (3.7) is asymptotically equivalent to

$$P(\sup_{c\in K+d}(G_n(c) + \sqrt{n}(\beta(c) - \gamma_{\max})) \le x) \tag{3.8}$$

for any given d. Since for $c \in K$, $\beta(c) = \gamma_{\max}$, and for $c \in (K+d) - K$, $\beta(c) < \gamma_{\max}$,

we have the following inequality.

$$P(\sup_{c\in K+d} G_n(c) \le x) \le (3.8) \le P(\sup_{c\in K} G_n(c) \le x). \tag{3.9}$$

Let $n \to \infty$, under Assumption III.11, we have

$$P(\sup_{c\in K+d} G(c) \le x) \le \liminf (3.8) \le \limsup (3.8) \le P(\sup_{c\in K} G(c) \le x). \tag{3.10}$$

Let $L_n(x) = P(\sqrt{n}(\hat{\gamma}_{\max} - \gamma_{\max}) \le x)$ and recall that $L_n(x)$ is asymptotically equivalent to (3.8). Therefore, for any $d > 0$, we have

$$P(\sup_{c\in K+d} G(c) \le x) \le \liminf L_n(x) \le \limsup L_n(x) \le P(\sup_{c\in K} G(c) \le x). \tag{3.11}$$

Under Assumptions III.11 and III.13, $\lim_{d\to 0} \sup_{c\in K+d} G(c) = \sup_{c\in K} G(c)$, in probability. Let $d \to 0$ in (3.11), we prove the desired result.

□

Lemma III.21. *Under Assumptions III.11, III.12 and III.13 and* $0 < r < 0.5$*, for any* $x \in R$*,*

$$\lim_{n\to\infty} P^*(\sqrt{n}(\gamma^*_{\max,\text{modified}} - \hat{\gamma}_{\max}) \le x) = P(\sup_{c\in K} G(c) \le x)$$

in probability.

Proof. With similar arguments as those in the proof of Lemma III.20, we have

$$\begin{aligned}
&P^*(\sqrt{n}(\gamma^*_{\max,\text{modified}} - \hat{\gamma}_{\max}) \le x) \\
=&P^*(\sup_{c\in D}(G^*_n(c) - n^r \sup_{d\in D}[\hat{\beta}(d) - \hat{\beta}(c)]) \le x) \\
=&P^*(\sqrt{n}\max(\sup_{c\in K+d}(G^*_n(c) - n^r \sup_{d\in D}(\hat{\beta}(d) - \hat{\beta}(c))), \sup_{c\in\overline{K+d}}(G^*_n(c) - n^r \sup_{d\in D}(\hat{\beta}(d) - \hat{\beta}(c)))) \le x).
\end{aligned} \tag{3.12}$$

Let $L_n^*(x) = P^*(\sqrt{n}(\gamma^*_{\max,\text{modified}} - \hat{\gamma}_{\max}) \leq x)$. We notice that

$$n^r \sup_{c\in\overline{K+d}} \sup_{d\in D}(\hat{\beta}(d)-\hat{\beta}(c)) \leq n^r \sup_{c\in\overline{K+d}} \sup_{d\in D}(\hat{\beta}(d)-\beta(d)-(\hat{\beta}(c)-\beta(c))-n^r \inf_{c\in\overline{K+d}}(\gamma_{\max}-\beta(c)). \tag{3.13}$$

From Assumptions III.11, III.12 and III.13, the 1st term of the right hand side of (3.13) converges to 0 in probability and the second term converges to negative infinite. Therefore, $n^r \sup_{d\in D}[\hat{\beta}(d) - \hat{\beta}(c)] \to -\infty$ uniformly for $c \in \overline{K+d}$ in probability and $L_n^*(x)$ is asymptotically equivalent to

$$P^*\left(\sup_{c\in K+d}(G_n^*(c) - n^r \sup_{d\in D}[\hat{\beta}(d) - \hat{\beta}(c)]) \leq x \right)$$

in probability. Similar to (3.13), we show that

$$\sup_{c\in K+d} |n^r \sup_{d\in D}[\hat{\beta}(d) - \hat{\beta}(c)] - n^r(\gamma_{\max} - \beta(c))| \to 0$$

in probability. Therefore, we have $L_n^*(x)$ is asymptotically equivalent to

$$P^*\left(\sup_{c\in K+d}(G_n^*(c) - n^r(\gamma_{\max} - \beta(c))) \leq x \right)$$

in probability. Now, we have the following inequality in probability.

$$P^*(\sup_{c\in K+d} G_n^*(c) \leq x) \leq \liminf L_n^*(x) \leq \limsup L_n^*(x) \leq P^*(\sup_{c\in K} G_n^*(c) \leq x).$$

With similar arguments used in the proof of Lemma III.20 and Assumption III.13, we prove the desired result. □

Proof of Theorem III.14: It naturally follows from Lemmas III.20 and III.21 with the property of the cdf. □

With $\gamma_{\max} = \sup_c \beta(c)$, we let $S_n = \{c : \beta(c) \geq \gamma_{\max} - \log(n)/n\}$ and $\gamma_{ss} = \beta(\tilde{c})$, where $\tilde{c}$ is a random variable as the value of c that achieves the minimum of $\beta(c)$ among all possible values of $argmax_{c\in D}\hat{\beta}(c)$. We further denote the smallest value of c that achieves the maximum of $\beta(c)$, $\gamma_{\max}$, by c_0, which is a well-defined fixed value from our continuous and compactness assumptions. Then, we have the following lemma to characterize the distribution of γ_{ss}.

Lemma III.22. *Under Assumption III.11, we have*

$$P(\gamma_{\max} - \gamma_{ss} < \log(n)/n) \to 1.$$

In other words, $P(\tilde{c} \in S_n) \to 1$, *as* $n \to \infty$.

Proof. If $\tilde{c} \notin S_n$, then, $\hat{\beta}(c_0) < \sup_{c\in\bar{S_n}} \hat{\beta}(c)$. Therefore, we have,

$$\begin{aligned}
&P(\gamma_{\max} - \gamma_{ss} > \log(n)/n) \\
\leq &P\Big(\hat{\beta}(c_0) < \sup_{c\in\bar{S_n}} \hat{\beta}(c)\Big) \\
= &P\Big(\sqrt{n}(\hat{\beta}(c_0) - \gamma_{\max}) < \sqrt{n}\sup_{c\in\bar{S_n}}(\hat{\beta}(c) - \beta(c) + \beta(c) - \gamma_{\max})\Big) \\
\leq &P\Big(\sqrt{n}(\hat{\beta}(c_0) - \gamma_{\max}) < \sqrt{n}\sup_{c\in\bar{S_n}}(\hat{\beta}(c) - \beta(c)) - \sqrt{n}\inf_{c\in\bar{S_n}}(\gamma_{\max} - \beta(c))\Big).
\end{aligned} \tag{3.14}$$

Since $\sqrt{n}\sup_{c\in\bar{S_n}}(\hat{\beta}(c)-\beta(c)) \leq \sqrt{n}\sup_{c\in D}(\hat{\beta}(c)-\beta(c))$ and $\sqrt{n}\inf_{c\in\bar{S_n}}(\gamma_{\max}-\beta(c)) > \log(n)$, the right hand side of (3.14) converges to 0 and we finish the proof. □

Lemma III.23. *Under Assumption III.11 and for any fixed* x, *we have*

$$P\Big(\sqrt{n}\sup_{c\in S_n}(\hat{\beta}(c) - \beta(c)) \leq x\Big) \to P(\sup_{c\in K} G(c) \leq x).$$

Proof. Since $S \subseteq S_n \subseteq S_2$, we can use similar techniques as those in the proof of Lemma III.20. □

Lemma III.24. *Let q_α be the $1-\alpha$ quantile of $\sup_{c\in K} G(c)$, then, under Assumption III.11, we have*

$$\lim_{n\to\infty} P\Big(\sqrt{n}\sup_{c\in D}(\hat{\beta}(c)-\gamma_{ss})\le q_\alpha\Big)=1-\alpha.$$

Proof. From the definition, we know $\hat{\beta}(\tilde{c})=\sup_{c\in D}\hat{\beta}(c)$ and $\gamma_{ss}=\beta(\tilde{c})$ so we have

$$\begin{aligned}
&P\Big(\sqrt{n}\sup_{c\in D}(\hat{\beta}(c)-\gamma_{ss})\le q_\alpha\Big)\\
=&P\Big(\sqrt{n}[\hat{\beta}(\tilde{c})-\beta(\tilde{c})](I_{\tilde{c}\in S_n}+I_{\tilde{c}\notin S_n})\le q_\alpha\Big)\\
=&P\Big(\sqrt{n}(\hat{\beta}(\tilde{c})-\beta(\tilde{c}))\le q_\alpha,\tilde{c}\in S_n\Big)+P\Big(\sqrt{n}(\hat{\beta}(\tilde{c})-\beta(\tilde{c}))\le q_\alpha,\tilde{c}\notin S_n\Big).
\end{aligned}\tag{3.15}$$

From Lemma III.22, we know the second part of the right hand side of (3.15) converges to 0. Notice that

$$\{\sqrt{n}(\hat{\beta}(\tilde{c})-\beta(\tilde{c}))\le q_\alpha,\tilde{c}\in S_n\}\supseteq\{\sqrt{n}\sup_{c\in S_n}((\hat{\beta}(c)-\beta(c))\le q_\alpha,\tilde{c}\in S_n\},$$

from Lemmas III.22 and III.23, we have

$$\begin{aligned}
&\liminf_{n\to\infty} P\Big(\sqrt{n}(\hat{\beta}(\tilde{c})-\beta(\tilde{c}))\le q_\alpha,\tilde{c}\in S_n\Big)\\
\ge&\liminf_{n\to\infty} P\Big(\sqrt{n}\sup_{c\in S_n}((\hat{\beta}(c)-\beta(c))\le q_\alpha,\tilde{c}\in S_n\Big)\\
=&\liminf_{n\to\infty} P\Big(\sqrt{n}\sup_{c\in S_n}((\hat{\beta}(c)-\beta(c))\le q_\alpha\Big)\to 1-\alpha.
\end{aligned}\tag{3.16}$$

Since $\gamma_{ss}\le\gamma_{\max}=\sup_{c\in D}\beta(c)$, from Lemma III.20, we have

$$\limsup_{n\to\infty} P\Big(\sqrt{n}\sup_{c\in D}(\hat{\beta}(c)-\gamma_{ss})\le q_\alpha\Big)\le\limsup_{n\to\infty} P\Big(\sqrt{n}\sup_{c\in D}(\hat{\beta}(c)-\max_{c\in D}\beta(c))\le q_\alpha\Big)=1-\alpha$$

Therefore, we complete the proof. □

Proof of Theorem III.15: It naturally follows from Lemma III.24 and Theorem III.14. □

CHAPTER IV

Inference for High Dimensional Subgroup Analysis

By high dimensional subgroup analysis, we refer to the subgroup analysis in the presence of (possibly) high dimensional confounders. In this chapter, we consider a model-based approach for high dimensional subgroup analysis and generalize the proposed debiased inference tool in Section 3.1 to analyze the best selected subgroup in high dimensions.

4.1 High Dimensional Subgroup Analysis

The rise of high dimensional subgroup analysis is in response to the needs to evaluate the subgroup effects in observational studies. Although randomized trials remain the gold standard for subgroup analysis as discussed in *Alosh et al.* (2017), the vast resources that randomized trials require to achieve good power and generability in evaluating the subgroup effects is a big concern to drug developers, which could make the subgroup analysis on randomized trials infeasible in practice; see *Hébert et al.* (2002). Sometimes randomized trials may be even considered unethical due to the use of placebo control, especially in pediatric studies; see *McMahon and Dal Pan* (2018). In the era of big data, numerous observational data, such as electronic health records, are available. An observational study is much less costly and has the advantages of studying heterogeneity, and therefore may be an alternate to randomized trials for

subgroup analysis.

In practice, several attempts on the subgroup analysis for observational studies have been made. For example, in the study of efficacy of insulin detemir on patients with type II diabetes, *Echtay et al.* (2017) analyzed the efficacy within subpopulations, such as Lebanese subgroup, based on an observational study. In general, there is an increasing interest in subgroup analysis for observational studies due to the potential benefits.

To conduct valid subgroup analysis on observational data, we must address the confounder bias and model the subgroup effects appropriately. Unfortunately, most existing literature in subgroup analysis is based on randomized trials and does not account for the (potential) confounder bias, and thus would be invalid for observational studies as reviewed in Section 1.4 and *Ondra et al.* (2016). Some recent literature in subgroup analysis considers the (potential) pre-treatment information and may be generalized to subgroup analysis on observational data. For example, *Imai et al.* (2013) adapted the Support Vector Machine classifier and proposed a framework to model the subgroup effects in the presence of pre-treatment variables. The authors also proposed a Lasso-type penalization method to estimate the subgroup effects and to select the best subgroup. Some other methods potentially applicable for subgroup analysis on observational data include *Fan et al.* (2017) and *Wager and Athey* (2018). However, these methods mainly focus on subgroup identification or the inference of a given subgroup. As far as we know, a careful treatment of subgroup analysis on observational data, especially the inference on the selected subgroup, is still lacking and we aim to bridge this gap.

In this chapter, we consider a simple linear model to model the subgroup effects in the presence of (possibly) high dimensional confounders and to assist high dimensional subgroup analysis. We propose a new statistical tool to select and infer the best subgroup in a high dimensional setting. Our work aims to facilitate a better subgroup

analysis on observational data and advance the use of real-world evidence in evaluating the subgroup effects.

4.2 Inference with Subgroups in High Dimensions

In this section, we focus on a simple linear model and propose a bootstrap-based asymptotically sharp inference for the best selected subgroup in the presence of (possibly) high dimensional confounders.

4.2.1 Problem Setting

Suppose we have a random sample of n observations, $\{(Y_i, Z_i, X_i)\}_{i=1}^n$, where Y_i is the response variable, Z_i is a p_1-dimensional vector of variables representing interactions between the treatment variable and the p_1 non-overlapping predefined subgroups of interest, and X_i is a p_2-dimensional vector of the potential confounders for the i-th subject. It is clear that in the presence of high dimensional confounders, we focus on the regime where $p_2 > n$. As for the number of the predefined subgroups p_1, it is usually considered as a fixed constant as we did in Section 3.1, but in Section 4.2.3, we will show that our proposed method in this chapter is indeed not relying on the assumption of fixed p_1, which can help the proposed method more robust to the relative size between p_1 and n in practice.

To model the subgroup effects in the presence of (possibly) high dimensional confounders, we generalize the model used in *Wang et al.* (2019) and consider the following model.

$$Y_i = Z_i'\beta + X_i'\gamma + \epsilon_i, \quad i = 1, \ldots, n, \tag{4.1}$$

where $\epsilon = (\epsilon_i, \ldots, \epsilon_n)'$ are the error terms and the coefficients $\beta = (\beta_1, \ldots, \beta_{p_1})'$ capture the subgroup effects. More discussions about the validity of using (4.1) in adjusting for the confounder bias and modelling the treatment effect can be found

in *Wang et al.* (2019) and *Belloni et al.* (2014). In the presence of high dimensional confounders where $p = p_1 + p_2 > n$, a frequently employed approach to estimate the subgroup effects is the Lasso estimate proposed in *Tibshirani* (1996) and defined by:

$$(\hat{\beta}'_{\text{Lasso}}, \hat{\gamma}'_{\text{Lasso}})' = \underset{\beta \in \mathbb{R}^{p_1}, \gamma \in \mathbb{R}^{p_2}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - Z_i'\beta - X_i'\gamma)^2 + + \lambda_n (||\beta||_1 + ||\gamma||_1) \right\},$$

where $\hat{\beta}_{\text{Lasso}} = (\hat{\beta}_{1,\text{Lasso}}, \ldots, \hat{\beta}_{p_1,\text{Lasso}})$ are the Lasso estimates for the subgroup effects and λ_n is a tuning parameter and may be selected by cross-validation in practice; see *Bühlmann and Van De Geer* (2011).

Similar to Section 3.1 and *Imai et al.* (2013), we propose to select the subgroup based on the Lasso estimate $\hat{\beta}_{\text{Lasso}}$. By letting $[p_1] = \{1, \ldots, p_1\}$, we focus on the inference of the following two quantities in high dimensional subgroup analysis:

1. *the best selected subgroup effect*: β_s, where $s = \text{argmax}_{i \in [p_1]} \hat{\beta}_{i,\text{Lasso}}$;
2. *the best subgroup effect*: $\beta_{\max} = \max_{i \in [p_1]} \beta_i$.

To facilitate the analysis later, we now introduce the notations that we will use for the remainder of the section. Denote the maximum of a q-dimensional vecotr a as $a_{\max} = \max\{a_j, j = 1, \cdots q\}$, and a collection of integers from 1 to q as $[q]$. Let M be a subset of $[q]$ and $M^c = [q] - M$. For a matrix $\mathbf{W}$ with q columns, let $\mathbf{W}_{ij}$ be the entry located in the i-th row and the j-th column, $\mathbf{W}_M$ be the submatrix of W consisting of the vectors in $\{\mathbf{W}_j, j \in M\}$ and $\mathbf{W}_{-M}$ be the submatrix of W consisting of the vectors in $\{\mathbf{W}_j, j \in M^c\}$, where $\mathbf{W}_j$ is the j-th column of $\mathbf{W}$, for $j = 1, \cdots, q$.

4.2.2 Proposed Method

One big challenge for the inference in high dimensional subgroup analysis is that it suffers from not only the subgroup selection bias but also the bias induced by the penalization routinely used in high dimensional estimation, which we call penalization

bias. It is well known in the literature that any high dimensional statistical inference must address the penalization bias appropriately and the Lasso estimate, $\hat{\beta}_{\text{Lasso}}$, is biased for inference; see *Van de Geer et al.* (2014).

Several procedures have been proposed to address the penalization bias for Lasso estimate and we will consider one procedure, desparsified Lasso in particular, based on estimating the inverse covariance matrix to remove this bias in $\hat{\beta}_{\text{Lasso}}$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ and $\mathbf{X} = (X_1, \ldots, X_n)'$, and we assume $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are all appropriately centered. Following *Zhang and Zhang* (2014), the desparsified Lasso provides an estimate $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_{p_1})$ of β as follows

$$\hat{b}_j = \hat{\beta}_{j,\text{Lasso}} + \frac{V_j'(\mathbf{Y} - \mathbf{Z}\hat{\beta}_{\text{Lasso}} - \mathbf{X}\hat{\gamma}_{\text{Lasso}})}{V_j'\mathbf{Z}_j}, \quad j = 1, \ldots, p_1, \tag{4.2}$$

where the second part on the right hand side of the above equation is an estimate of the penalization bias in Lasso, and

$$V_j = \mathbf{Z}_j - (\mathbf{Z}, \mathbf{X})_{-j}\hat{\zeta}_j, \quad j = 1, \ldots, p_1, \tag{4.3}$$

with

$$\hat{\zeta}_j = \underset{\zeta_j \in \mathbb{R}^{p_1+p_2-1}}{\arg\min} \left\{ ||\mathbf{Z}_j - (\mathbf{Z}, \mathbf{X})_{-j}\zeta_j||_2^2/n + \lambda_{\mathbf{X}}||\zeta_j||_1 \right\}, \quad j = 1, \ldots, p_1,$$

where $\lambda_{\mathbf{X}}$ is a tuning parameter for the above Lasso procedure and may be different from λ_n used in estimating $\hat{\beta}_{\text{Lasso}}$. Under certain regularity assumptions, the desparsified Lasso estimate removes the penalization bias.

It is clear that while removing the penalization bias, the desparsified Lasso estimate can not be used to infer the best selected subgroup directly due to the subgroup selection bias. Here, to address the subgroup selection bias, we again apply the idea

of bootstrapping the bias.

In high dimensions, several bootstrap procedures are available; see for example *Chernozhukov et al.* (2013a) and *Zhang and Cheng* (2017). In particular, *Dezeure et al.* (2017) proposed the following wild bootstrap procedure for the desparsified Lasso estimate,

$$Y_i^* = Z_i'\hat{\beta}_{\text{Lasso}} + X_i'\hat{\gamma}_{\text{Lasso}} + \varepsilon_i^*, \qquad i = 1, \ldots, n, \tag{4.4}$$

where $\varepsilon_i^* = u_i\hat{\varepsilon}_i$ and u_i is i.i.d and independent of the data with $E[u_i] = 0$, $E[u_i^2] = 1$ and $E[u_i^4] < \infty$, and $\hat{\varepsilon}_i = Y_i^* - Z_i'\hat{\beta}_{\text{Lasso}} - X_i'\hat{\gamma}_{\text{Lasso}}$, for $i = 1, \ldots, n$. The bootstrap desparsified Lasso $b^* = (b_1^*, \ldots, b_{p_1}^*)$ is defined as the desparsified Lasso estimate for β similar to (4.2) but based on the bootstrap sample $\{(Y_i^*, Z_i, X_i)\}_{i=1}^n$. *Dezeure et al.* (2017) showed that under some regularity conditions, the bootstrap procedure in (4.4) is consistent in the sense that, conditional on the data, the asymptotic distribution of $\sqrt{n}(b^* - \hat{\beta}_{\text{Lasso}})$ is equivalent to the limiting distribution of the desparsified Lasso estimate $\sqrt{n}(\hat{b} - \beta)$.

With the bootstrap procedure for the desparsified Lasso estimate at hand, we extend the proposed debiased tool in Section 3.1 to infer the best selected subgroup in our high dimensional subgroup analysis. We first propose to modify $b_{\max}^*$ as follows,

$$b_{\text{modified;max}}^* = \max_{j \in [p_1]}(b_j^* + c_j(r)), \tag{4.5}$$

where $c_j(r) = (1 - n^{r-0.5})(\hat{\beta}_{\text{max;Lasso}} - \hat{\beta}_{j,\text{Lasso}})$ and r is a tuning parameter. With the modified bootstrap estimate $b_{\text{modified;max}}^*$, the proposed method proceeds with Algorithm 5.

Algorithm 5 One-sided confidence interval for $\beta_{\max}$ in high dimensions.

1: **Set-up:** For $j \in [p_1]$, calculate $c_j(r)$;
2: **for** l=1,...,B **do**
3: **Wild bootstrap:** Generate the bootstrap sample $\{(Y_i^*, Z_i, X_i)\}_{i=1}^n$ by (4.4);
4: **Calculation:** Calculate $T_l^* = b^*_{\text{modified;max}} - \hat{\beta}_{\max}$ based on $\{(Y_i^*, Z_i, X_i)\}_{i=1}^n$ by (4.5);
5: **end for**
6: The α-level one-sided confidence interval for $\beta_{\max}$ is $\hat{b}_{\max} - \text{quantile}(T_l^*, \alpha)$.

4.2.3 Asymptotic Validity

Even though in our high dimensional subgroup analysis, we focus on the scenario where the number of the predefined subgroups p_1 is often viewed as a constant, we would like to justify Algorithm 5 without relying on the assumption of fixed p_1. In fact, when many subgroups are searched, p_1 may be very close to n and may not be viewed as a constant any more. In this case, any asymptotic validity assuming p_1 is fixed may not be able to guarantee a good finite sample approximation. For this reason, we would carry out a theoretical investigation while allowing p_1 increases with n.

To establish the asymptotic validity of the proposed inference in Algorithm 5, we need to make the following assumptions which fall into two categories: (1) assumptions for (bootstrap) desparsified Lasso, Assumptions IV.1–IV.6, which are classical and directly from *Dezeure et al.* (2017), and ensure the validity of the desparsified lasso and the consistency of the wild bootstrap counterpart; (2) assumptions for the tuning parameter, Assumptions IV.7–IV.8, which require the tuning parameter to be neither too large (Assumption IV.7) nor too small (Assumption IV.8) to ensure that the modified bootstrap can learn the bias correctly. Let $H = \{j \in [p_1] : \beta_j = \beta_{\max}\}$, $d_s = \beta_{\max} - \max_{j \in [p_1]-H} \beta_j$ and β^* and γ^* represent the Lasso estimate on bootstrap sample, the assumptions are given below.

Assumption IV.1. $||\hat{\beta}_{Lasso} - \beta||_1 + ||\hat{\gamma}_{Lasso} - \gamma||_1 = o_P(1/\sqrt{\log(p)\log(p_1)})$.

Assumption IV.2. $\lambda_{\mathbf{X}} \propto \sqrt{\log(p)/n}$, $||V_j||_2^2/n \geq L_V$, $||V_j||_4 = o(||V_j||_2)$, $j \in [p_1]$.

Assumption IV.3. $\epsilon_1, \ldots, \epsilon_n$ *independent,* $E[\epsilon] = 0$, $E||\epsilon||_2^2/n = \sigma_\epsilon^2$, $E|\epsilon_i|^2 \geq L$, $E|\epsilon_i|^4 \leq C$.

Assumption IV.4. $||\beta^* - \hat{\beta}_{Lasso}||_1 + ||\gamma^* - \hat{\gamma}_{Lasso}||_1 = o_{P^*}(1/\sqrt{\log(p)\log(p_1)})$ *in probability.*

Assumption IV.5. $\max_{ij} |\mathbf{X}_{ij}| \leq C_{\mathbf{X}}$ *and* $\max_{ij} |\mathbf{Z}_{ij}| \leq C_{\mathbf{Z}}$.

Assumption IV.6. $\max_{j \in [p_1]} ||V_j||_\infty \leq K$, $\log(p_1) = o(n^{1/7})$.

Assumption IV.7. $n^r ||\hat{\beta}_{Lasso} - \beta||_\infty = o_P(1/\sqrt{\log(p_1)})$.

Assumption IV.8. $\liminf_{n\to\infty}[n^r d_s - \log(p_1)] = \infty$.

Here σ_ϵ, L_V, C, L, $C_{\mathbf{X}}$, $C_{\mathbf{Z}}$ and K are positive constants uniformly bounded away from 0 and ∞. We note that Assumption IV.6 allows p_1 increases with n as we wish. Relative to the classical assumptions, Assumptions IV.1–IV.6, used in *Dezeure et al.* (2017), the extra assumptions, Assumptions IV.7–IV.8, are mild. If $n^r = O(\sqrt{\log(p)})$, then, Assumption IV.1 directly implies Assumption IV.7. If d_s is a constant with regard to n, then, $r \geq 1/7$ and Assumption IV.6 directly implies Assumption IV.8. In the case of predefined subgroups when p_1 is viewed as a fixed constant, $0 < r < 0.5$ can easily satisfy Assumptions IV.7–IV.8. In fact, we only require a little more than Assumptions IV.1–IV.6 to justify the inference on $\beta_{\max}$ in high dimensions. The justification is as follows.

Theorem IV.9. *Under Assumptions IV.1-IV.8, we have*

$$\sup_c |P(\sqrt{n}(\hat{b}_{\max} - \beta_{\max}) \leq c) - P^*(\sqrt{n}(b^*_{\text{modified;max}} - \hat{\beta}_{\max}) \leq c)| = o_P(1)$$

Theorem IV.9 confirms the proposed one-sided confidence interval in Algorithm 5 is asymptotically sharp for $\beta_{\max}$ under the classic assumptions for the (bootstrap)

desparsified Lasso and two mild assumptions for the tuning parameter. Any choice of the tuning parameter satisfying Assumptions IV.7–IV.8 can guarantee the asymptotic validity. However, we suggest to use the data-adaptive cross-validated method in Section 3.1.5 to choose the tuning parameter in practice to improve the finite sample performance. The following corollary confirms that the procedure also works for the inference on the best selected subgroup in high dimensional subgroup analysis.

Corollary IV.10. *Under Assumptions IV.1-IV.8, we have*

$$\sup_c |P(\sqrt{n}(\hat{b}_{\max} - \beta_s) \leq c) - P^*(\sqrt{n}(b^*_{\text{modified;max}} - \hat{\beta}_{\max}) \leq c)| = o_p(1).$$

Although we focus on the subgroup selected by the Lasso estimate in Corollary IV.10, it is not essential. Our proposed method indeed works for the subgroup selected by any subgroup selection method as long as it can select the best subgroup with high probability in high dimensions, which many modern model selection tools can easily achieve. In other words, Corollary IV.10 can be further generalized to other subgroup selection methods in high dimensions, besides the Lasso-type selection we consider here, and our proposed method is widely applicable in practice.

4.3 Discussion

In this chapter, we discuss an extension of the debiased inference tool to high dimensional subgroup analysis. The proposed tool enables an appropriate modelling, selection and inference on the best subgroup in high dimensions. We anticipate that the proposed tool can help advance the use of real-world data in subgroup analysis.

One property to note is that the validity of the proposed method is not relying on the assumption of fixed p_1. Although we focus on the scenario where the subgroups are predefined, it is clear that this property is useful and can help the proposed method more robust to the relative size between p_1 and n. Moreover, without relying

on a fixed p_1, the proposed method might be useful in other research areas, such as the study of multiple treatments.

The proposed tool provides valid subgroup analysis even in the presence of high dimensional confounders. Although the proposed tool relies on the linear model assumption, the validity of the proposed method in high dimensions is not limited to linear models because more flexible models, such as the semi-parametric model, can be well approximated by the high dimensional linear model we assume here under some regularity conditions as discussed in *Wang et al.* (2019).

One limitation of the proposed tool is that it requires the candidate subgroups to be non-overlapping in order to make sure the modelling of the subgroup effects is appropriate. Although non-overlapping subgroups are used in many applications, overlapping subgroups also arise in practice, such as the male subgroup and East Asian subgroup considered in the synthetic data study of the MONET1 trial as discussed in Seciont 2.2. Further research is needed to address the high dimensional inference question for overlapped subgroups.

4.4 Proof of Theorem IV.9

Following the proof in *Dezeure et al.* (2017), we let $(\zeta_1, \ldots, \zeta_{p_1})$ denote a Gaussian vector with $E\zeta_j = 0$ and

$$E\zeta_j\zeta_k = \frac{E(\epsilon \circ V_j)^T(\epsilon \circ V_k)/n}{(EV_j^T\mathbf{Z}_j/n)(EV_k^T\mathbf{Z}_k/n)},$$

where $\circ$ denote the elementwise multiplication for vectors. Let P^* denote the bootstrap probability and $T_i = \sqrt{n}(\hat{b}_i - \beta_i)$ and $T_i^* = \sqrt{n}(b_i^* - \hat{\beta}_{i,\text{Lasso}})$. We have the following approximations.

Lemma IV.11. *Under Assumptions IV.1-IV.6. For any subset M of $[p_1]$, we have*

$$\sup_c |P(\max_{i\in M} T_i \leq c) - P(\max_{i\in M} \zeta_i \leq c)| = o(1);$$

$$\sup_c |P^*(\max_{i\in M} T_i^* \leq c) - P(\max_{i\in M} \zeta_i \leq c)| = o_P(1).$$

Proof. Let $\xi_j = \dfrac{V_j^T \epsilon/\sqrt{n}}{V_j^T \mathbf{Z}_j/n}$ and

$$\Delta_j = \frac{(\sum_{k\neq j} V_j^T \mathbf{Z}_k(\beta_k - \hat{\beta}_{k,\text{Lasso}}) + \sum_k V_j^T \mathbf{X}_k(\gamma_k - \hat{\gamma}_{k,\text{Lasso}}))/\sqrt{n}}{V_j^T \mathbf{Z}_j/n},$$

we have

$$T_j = \sqrt{n}(\hat{b}_j - \beta_j) = \xi_j + \Delta_j.$$

By Assumptions IV.2, IV.3 and IV.6 and with similar arguments as those in Proposition 1 in *Dezeure et al.* (2017), we have

$$\sup_c |P(\max_{i\in M} \xi_i \leq c) - P(\max_{i\in M} \zeta_i \leq c)| = o(1).$$

Moreover, the anti-concentration inequality asserts that if $B = \dfrac{o(1)}{\log^{1/2}(2|M|)}$,

$$\sup_c P(c \leq \max_{i\in M} \zeta_i \leq c + B) = o(1).$$

By Assumptions IV.1 and IV.5, we have

$$||\Delta||_\infty = O_P(\lambda_X \sqrt{n}(||\hat{\beta}_{\text{Lasso}} - \beta|| + ||\hat{\gamma}_{\text{Lasso}} - \gamma||)) = o_P(1/\sqrt{\log(p_1)}).$$

Therefore, we prove the first statement. For the bootstrap part, similar arguments lead to the proof. □

Lemma IV.12. *Under Assumptions IV.1-IV.8. We have*

$$\sup_c |P(\sqrt{n}(\max_{i\in[p_1]} \hat{b}_i - \max_{i\in[p_1]} \beta_i) \le c) - P(\max_{i\in H} T_i \le c)| = o(1);$$

$$\sup_c |P^*(\sqrt{n}(\max_{i\in[p_1]} b_i^* - \max_{i\in[p_1]} \hat{\beta}_{i,Lasso}) \le c) - P(\max_{i\in H} T_i \le c)| = o_P(1).$$

Proof. (1) We have

$$P(\sqrt{n}(\max_{i\in[p_1]} \hat{b}_i - \max_{i\in[p_1]} \beta_i) \le c) = P(\max_{i\in H} T_i \le c, T_j \le c + \sqrt{n}(\max_{i\in H} \beta_i - \beta_j), j \notin H).$$

Given a fixed value c_0, we have the following bound:

$$\sup_{c>c_0} |P(\sqrt{n}(\max_{i\in[p_1]} \hat{b}_i - \max_{i\in[p_1]} \beta_i) \le c) - P(\max_{i\in H} T_i \le c)| \le 1 - P(\max_{j\in[p_1]-H} T_j \le c_0 + \sqrt{n}d_s).$$

From Lemma IV.11, the latter is equivalent to

$$1 - P(\max_{j\in[p_1]-H} \zeta_i \le c_0 + \sqrt{n}d_s) \le Cp_1 e^{-d_s\sqrt{n}/2},$$

where C is a constant independent of n. By Assumption IV.8, $Cp_1 e^{-d_s\sqrt{n}/2} = o(1)$. Notice that

$$\limsup_{c\to-\infty} P(\max_{i\in[p_1]} T_i \le c) = 0,$$

we prove the 1st statement.

(2) Let $K_i = n^r(\max_{j\in[p_1]} \hat{\beta}_{j,\text{Lasso}} - \max_{j\in[p_1]} \beta_j - \hat{\beta}_{i,\text{Lasso}} + \beta_i)$ and $\tilde{d}_i = n^r(\max_{j\in[p_1]} \beta_j - \beta_i)$, we have

$$P(\sqrt{n}(\max_{i\in[p_1]}(b_i^*+c_i) - \max_{i\in[p_1]} \hat{\beta}_{i,\text{Lasso}}) \le c) = P(T_i^* \le c + K_i, i \in H, T_j^* \le c + n^r\tilde{d}_j + K_j, j \notin H).$$

Similar to (1), we have

$$\begin{aligned}&\sup_{c>c_0}|P^*(\sqrt{n}(\max_{i\in[p_1]}(b_i^*+c_i)-\max_{i\in[p_1]}\hat{\beta}_{i,\text{Lasso}})\leq c)-P^*(T_i^*\leq c+K_i,i\in H)|\\ \leq&1-P^*(T_j^*\leq c+n^r\tilde{d}_i+K_i,i\notin H).\end{aligned}\tag{4.6}$$

The latter is bounded by

$$1-P^*(T_j^*\leq c+n^r d_s-||K||_\infty,j\notin H).\tag{4.7}$$

By Assumption IV.7, we have $||K||_\infty=o_P(1/\sqrt{\log(p_1)})$. From Lemma IV.11, we can asymptotically bound (4.7) by

$$1-P(\max_{j\in[p_1]-H}T_j\leq c_0+n^r d_s),$$

which is $o(1)$ by Assumption IV.8 and Lemma IV.11. Therefore, the right hand side of (4.6) goes to 0 in probability. Since $||K||_\infty=o_P(1/\log(p_1))$, by Lemma IV.11, we have

$$\sup_{c>c_0}|P^*(T_i^*\leq c+K_i,i\in H)-P(\max_{i\in H}T_i\leq c)|=o_P(1).$$

With similar arguments in (1), we prove the 2nd statement. □

Proof of Theorem IV.9: It is directly from Lemma IV.12.

Proof of Corollary IV.10: It is directly from Lemma IV.12 and Assumptions IV.7-IV.8.

CHAPTER V

Summary

When the best subgroup is selected from the data over a set of candidate subgroups, appropriate statistical analysis of the selected subgroup is needed to inform a scientific decision on subgroup pursuit. However, 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 this dissertation is that appropriate statistical analysis of the selected subgroup must take the selection process into account. We propose two new statistical tools, which account for the selectiong procedure and appropriately address subgroup selection bias, to analyze the selected subgroup and help a better-informed decision on subgroup pursuit.

The first tool measures the risk of the pursuit of the selected subgroup from a reasonable angle and can be easily generalized to the situations where there are multiple overlapping subgroups. The risk index is model-free, easy-to-compute, and transparent. The behavior of the risk index is well understood and the index can be used as a quantitative screening tool in subgroup pursuit.

The second tool provides a debiased inference for the selected subgroup. The proposed tools are model-free, easy to implement, and the resulting statistical analysis is asymptotically sharp, regardless of whether the subgroups are pre-defined or identified post hoc from the data. An extension to observational studies is also discussed,

which can help promote the use of real-world evidence in subgroup analysis.

Risk quantification and 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. Through a case study of the MONET1 trial and its failed follow-up trial, we show that lessons can be learned for future subgroup analysis in clinical work. Our analysis shows that 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 methods aim at statistical analysis; i.e. risk quantification and debiased inference, for the best selected subgroup. In practice, other considerations, such as domain knowledge and budge issue, might be taken into consideration in subgroup identification. The proposed methods would then serve as a conservative approach to those identified subgroups.

For now, the proposed methods aim for a setting where the difference between any two subgroups can be viewed as a constant and the candidate subgroups are clearly defined by biomarkers. However, in practice, more complicated scenarios, such as the scenario where the difference between two subgroups can be in a root-n order and the scenario where the candidate subgroups are from a clustering result, may be considered in subgroup analysis. To help a better subgroup analysis in the future, further investigations into these more challenge scenarios are needed.

Besides subgroup analysis, this dissertation may be of interests for statisticians in other research areas. In our study, we notice subgroup selection bias is closely related to the bias issue in post selection inference and the inflation of the family wise error rate in multiple testing. We may transform some post selection inference problems and multiple testing problems into the study of the correction of the subgroup selection bias and provide alternative solutions from subgroup analysis views.

This dissertation may even have potential impacts beyond statistics and promote

the use of subgroup analysis in the broad areas of data science. In this dissertation, we have demonstrated the methods we develop can help more scientific decisions on subgroup pursuit in clinical trials. We anticipate that the proposed methods can also benefit the researchers in other disciplines where heterogeneity often arises, such as sociology and finance, and lead to more important scientific discoveries.

In summary, we hope that this dissertation argues convincingly that valid statistical analysis of post hoc identified subgroups needs to be and can be performed effectively with the new risk quantification and debiased inference tools and believe the proposed methods can promote a better and broader use of subgroup analysis.

APPENDIX

APPENDIX A

Synthetic Data for the MONET1 Study

To generate the synthetic data of the MONET1 trial, we consider a simple setting of n observations, $(Y_i, D_i, \delta_i, Z_i)$, $i = 1, \cdots, n$, where Y_i is the (possibly censored) survival time of the i-th subject, D_i is the treatment indicator, δ_i is the censoring indicator, and $Z_i = (Z_{i,1}, \ldots, Z_{i,K})$ is the subgroup indicator indicating whether the subject belongs to any of the $2K$ subgroups we consider.

Following the MONET1 trial in *Kubota et al.* (2014), we let $n = 1090$ and $K = 8$ with the following subgroups: East Asian patient or not ($Z_{i,1} = 1$ or 0), received radiotherapy or not ($Z_{i,2} = 1$ or 0), stage IIIB or not ($Z_{i,3} = 1$ or 0), Age greater than 65 or not ($Z_{i,4} = 1$ or 0), ECOG PS equal to 0 or not ($Z_{i,5} = 1$ or 0), Adenocarcinoma histology or not ($Z_{i,6} = 1$ or 0), male or female ($Z_{i,7} = 1$ or 0), and never smoked or not ($Z_{i,8} = 1$ or 0). We independently let $Z_{i,1} \sim \text{Bernoulli}(1, p_i)$, and $D_i \sim \text{Bernoulli}(1, p)$, where the parameters are estimated by the sample proportion in Table 1 and Figure 1.A in *Kubota et al.* (2014) and given in Table A.1. We independently generate the event time T_i by the distribution F and the censoring time C_i by the distribution G as defined in Table A.2, both of which are estimated based on Figure 1.A in *Kubota et al.* (2014) under the assumptions of no treatment effect.

Finally, we obtain $Y_i = \min(T_i, C_i)$ and $\delta_i = I_{T_i \leq C_i}$.

Table A.1: Proportion of the subgroups p_i and the proportion p of subjects with treatment.

p	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
0.5	0.208	0.143	0.139	0.339	0.362	0.817	0.615	0.281

Table A.2: Distribution for the event time and the censoring time: $F(x) = P(T = x)$ and $G(x) = P(C = x)$.

x	8	16	24	32	40	48	56	64	72	80
F	0.05	0.05	0.08	0.06	0.06	0.06	0.04	0.05	0.03	0.04
G	0	0	0	0	0	1/15	1/15	1/15	1/15	1/15
x	88	96	104	112	120	128	136	144	152	160
F	0.04	0.04	0.02	0.02	0.02	0.02	0.02	0.00	0.00	0.3
G	0	0	0	0	2/15	2/15	2/15	2/15	2/15	0

In the end, we focus on one synthetic dataset, which is similar to the real data in the MONET1 trial as discussed in Section 2.2 and can be found in `https://github.com/xinzhoug/Data`.

BIBLIOGRAPHY

BIBLIOGRAPHY

Alosh, M., M. F. Huque, F. Bretz, and R. B. D'Agostino Sr (2017), Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials, *Statistics in medicine*, *36*(8), 1334–1360.

Altstein, L. L., G. Li, and R. M. Elashoff (2011), A method to estimate treatment efficacy among latent subgroups of a randomized clinical trial, *Statistics in medicine*, *30*(7), 709–717.

Bang, Y.-J., et al. (2010), Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of her2-positive advanced gastric or gastro-oesophageal junction cancer (toga): a phase 3, open-label, randomised controlled trial, *The Lancet*, *376*(9742), 687–697.

Belloni, A., V. Chernozhukov, and C. Hansen (2014), Inference on treatment effects after selection among high-dimensional controls, *The Review of Economic Studies*, *81*(2), 608–650.

Berger, J. O., X. Wang, and L. Shen (2014), A bayesian approach to subgroup identification, *Journal of biopharmaceutical statistics*, *24*(1), 110–129.

Bornkamp, B., D. Ohlssen, B. P. Magnusson, and H. Schmidli (2017), Model averaging for treatment effect estimation in subgroups, *Pharmaceutical statistics*, *16*(2), 133–142.

Brody, H., and L. M. Hunt (2006), Bidil: assessing a race-based pharmaceutical, *The Annals of Family Medicine*, *4*(6), 556–560.

Bühlmann, P., and S. Van De Geer (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Cai, T., L. Tian, P. H. Wong, and L. Wei (2010), Analysis of randomized comparative clinical trial data for personalized treatment selections, *Biostatistics*, *12*(2), 270–282.

Chernozhukov, V., D. Chetverikov, K. Kato, et al. (2013a), Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors, *The Annals of Statistics*, *41*(6), 2786–2819.

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

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

Echtay, A., E. Andari, P. Atallah, R. Moufarrege, and R. Nemr (2017), Insulin detemir in combination with oral antidiabetic drugs improves glycemic control in persons with type 2 diabetes in near east countries: Results from the lebanese subgroup, *Ethnicity & disease*, *27*(1), 45.

Efron, B., and R. J. Tibshirani (1994), *An introduction to the bootstrap*, CRC press.

Eng, K. H. (2014), Randomized reverse marker strategy design for prospective biomarker validation, *Statistics in medicine*, *33*(18), 3089–3099.

Fan, A., R. Song, and W. Lu (2017), Change-plane analysis for subgroup detection and sample size calculation, *Journal of the American Statistical Association*, *112*(518), 769–778.

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

Friede, T., N. Parsons, and N. Stallard (2012), A conditional error function approach for subgroup selection in adaptive clinical trials, *Statistics in Medicine*, *31*(30), 4309–4320.

Fuentes, C., G. Casella, M. T. Wells, et al. (2018), Confidence intervals for the means of the selected populations, *Electronic Journal of Statistics*, *12*(1), 58–79.

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

Hébert, P. C., D. J. Cook, G. Wells, and J. Marshall (2002), The design of randomized clinical trials in critically ill patients, *Chest*, *121*(4), 1290–1300.

Hothorn, T., F. Bretz, and P. Westfall (2008), Simultaneous inference in general parametric models, *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*(3), 346–363.

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

Jenkins, M., A. Stone, and C. Jennison (2011), An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints, *Pharmaceutical statistics*, *10*(4), 347–356.

Klein, J. P., and M. L. Moeschberger (2005), *Survival analysis: techniques for censored and truncated data*, Springer Science & Business Media.

Kubota, K., et al. (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.

Kubota, K., et al. (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, *J Clin Oncol*, *35*(32), 3662–3670.

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

Lee, S., and R. A. Wilke (2009), Reform of unemployment compensation in germany: A nonparametric bounds analysis using register data, *Journal of Business & Economic Statistics*, *27*(2), 193–205.

Lin, D. Y., and L.-J. Wei (1989), The robust inference for the cox proportional hazards model, *Journal of the American statistical Association*, *84*(408), 1074–1078.

Lipkovich, I., and A. Dmitrienko (2014), Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides, *Journal of biopharmaceutical statistics*, *24*(1), 130–153.

Lipkovich, I., A. Dmitrienko, J. Denne, and G. Enas (2011), Subgroup identification based on differential effect searcha recursive partitioning method for establishing response to treatment in patient subpopulations, *Statistics in medicine*, *30*(21), 2601–2621.

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

Maemondo, M., et al. (2010), Gefitinib or chemotherapy for non–small-cell lung cancer with mutated egfr, *New England Journal of Medicine*, *362*(25), 2380–2388.

Magnusson, B. P., and B. W. Turnbull (2013), Group sequential enrichment design incorporating subgroup selection, *Statistics in medicine*, *32*(16), 2695–2714.

Mandrekar, S. J., and D. J. Sargent (2009), Clinical trial designs for predictive biomarker validation: one size does not fit all, *Journal of biopharmaceutical statistics*, *19*(3), 530–542.

McMahon, A. W., and G. Dal Pan (2018), Assessing drug safety in childrenthe role of real-world data, *The New England journal of medicine*, *378*(23), 2155.

MOLOGEN (2018), Final analysis of impulse study confirms topline data with positive subgroup results, `www.businesswire.com/news/home/20180406005228/en/MOLOGEN-Final-Analysis-IMPULSE-Study-Confirms-Topline`.

Nadarajah, S., and S. Kotz (2008), Exact distribution of the max/min of two gaussian random variables, *IEEE Transactions on very large scale integration (VLSI) systems*, *16*(2), 210–212.

Naggara, O., J. Raymond, F. Guilbert, and D. Altman (2011), The problem of subgroup analyses: An example from a trial on ruptured intracranial aneurysms, *American Journal of Neuroradiology*, *32*(4), 633–636.

Ondra, T., A. Dmitrienko, T. Friede, A. Graf, F. Miller, N. Stallard, and M. Posch (2016), Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review, *Journal of biopharmaceutical statistics*, *26*(1), 99–119.

Peto, R., and J. Peto (1972), Asymptotically efficient rank invariant test procedures, *Journal of the Royal Statistical Society: Series A (General)*, *135*(2), 185–198.

Rosenkranz, G. K. (2016), Exploratory subgroup analysis in clinical trials by model selection, *Biometrical Journal*, *58*(5), 1217–1228.

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

Shen, L. (2001), An improved method of evaluating drug effect in a multiple dose clinical trial, *Statistics in Medicine*, *20*(13), 1913–1929.

Song, J. X. (2014), A two-stage patient enrichment adaptive design in phase ii oncology trials, *Contemporary clinical trials*, *37*(1), 148–154.

Stallard, N., S. Todd, and J. Whitehead (2008), Estimation following selection of the largest of two normal means, *Journal of Statistical Planning and Inference*, *138*(6), 1629–1638.

Struthers, C. A., and J. D. Kalbfleisch (1986), Misspecified proportional hazard models, *Biometrika*, *73*(2), 363–369.

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

Sun, X., et al. (2012), Credibility of claims of subgroup effects in randomised controlled trials: systematic review, *Bmj*, *344*, e1553.

Thomas, M., and B. Bornkamp (2017), Comparing approaches to treatment effect estimation for subgroups in clinical trials, *Statistics in Biopharmaceutical Research*, *9*(2), 160–171.

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

Van de Geer, S., P. Bühlmann, Y. Ritov, R. Dezeure, et al. (2014), On asymptotically optimal confidence regions and tests for high-dimensional models, *The Annals of Statistics*, *42*(3), 1166–1202.

Van Der Vaart, A. W., and J. A. Wellner (1996), Weak convergence, in *Weak convergence and empirical processes*, pp. 16–28, Springer.

Venter, J. (1988), Confidence bounds based on the largest treatment mean, *South African Journal of Science*, *84*(5), 340.

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

Wang, J., X. He, and G. Xu (2019), Debiased inference on treatment effect in a high-dimensional model, *Journal of the American Statistical Association*, doi:10.1080/01621459.2018.1558062.

Woody, S., and J. G. Scott (2018), Optimal post-selection inference for sparse signals: a nonparametric empirical-bayes approach, *arXiv preprint arXiv:1810.11042*.

Zhang, C.-H., and S. S. Zhang (2014), Confidence intervals for low dimensional parameters in high dimensional linear models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 217–242.

Zhang, X., and G. Cheng (2017), Simultaneous inference for high-dimensional linear models, *Journal of the American Statistical Association*, *112*(518), 757–768.

Zhang, Z., H. Seibold, M. V. Vettore, W.-J. Song, and V. François (2018), Subgroup identification in clinical trials: an overview of available methods and their implementations with r, *Annals of translational medicine*, *6*(7).

Ziegler, A., A. Koch, K. Krockenberger, and A. Großhennig (2012), Personalized medicine using dna biomarkers: a review, *Human genetics*, *131*(10), 1627–1638.