Econometrics and Statistics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

A Robust Quantitative Risk Screening for Subgroup Pursuit in Clinical Trials☆

Xinzhou Guo[a,*], Ruosha Li[b], Jianjun Zhou[c], Xuming He[d]

[a] Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, 999077, China
[b] Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston, 7000 Fannin Street, Suite 2250, Houston, 77030, USA
[c] Yunnan Provincial Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, 2 Cuihu North Road, Kunming, 650091, China
[d] Department of Statistics, University of Michigan, 1085 South University Ave, Ann Arbor, 48109, USA

ARTICLE INFO

Article history:
Received 5 July 2022
Revised 19 May 2023
Accepted 25 May 2023
Available online xxx

Keywords:
Bias-correction
Bootstrap
Decision risk
Model-free
Subgroup analysis
Misspecified proportional hazard model

ABSTRACT

In clinical studies, when to recommend or decide further pursuit of the most promising subgroup that has been observed from an existing trial is a very important question. It is well recognized that the working models in assessing subgroup effects might be misspecified and the observed treatment effect size of the best selected subgroup tends to be too optimistic. Therefore, a careful and robust statistical quantification of risk is useful before any decision of subgroup pursuit is made. Via the newly established bootstrap consistency for the misspecified proportional hazard model, the issue of selection bias and model misspecification in subgroup pursuit is addressed, and a robust risk quantitative measure directly based on the observed treatment effect of the selected subgroup that might be used in the decision-making of subgroup pursuit is provided. Two earlier studies are reviewed to demonstrate what can be learned from the proposed risk index.

1. Introduction

Subgroup analysis as the analysis of treatment effects in sub-populations is widely used in clinical trials (Sun et al., 2012). In some cases, a new treatment might be only marginally effective for the population of the original study, and subgroup analysis may provide useful information for the assignment of the treatment and for future study. For example, through subgroup analysis, researchers have successfully identified that lefitolimod has positive results on patients with extensive-stage small-cell lung cancer in two important subgroups (MOLOGEN, 2018).

Subgroup pursuit, the pursuit of the most promising subgroup identified from the existing data with follow-up confirmatory trials, is a natural cause of action when one subgroup stands out in terms of its effect size but the overall effect size is marginal at best. To avoid waste of resources, a decision on whether an additional investment should be made for the confirmatory trials on the seemingly responsive sub-population needs to be made. In practice, decision-makers are often tempted to rely on the observed effect size from the most promising subgroup identified from the existing trial. However, without a careful and robust risk/reward analysis of subgroup pursuit, the decisions can be of high risk for the following two

☆ The research is partly supported by the National Science Foundation (USA) Award DMS-1914496.
* Corresponding author.
E-mail address: xinzhoug@ust.hk (X. Guo).

https://doi.org/10.1016/j.ecosta.2023.05.002
2452-3062/

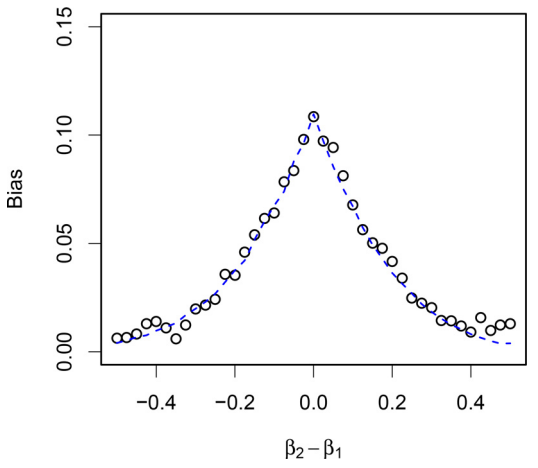

Fig. 1. Empirical bias $T - \tau$ (dots) vs. $\beta_2 - \beta_1$, where the true $\beta_1 = 0.095$ with standard error smaller than 0.01. The detailed simulation setting is given in Section 3.2.

reasons. First, the working model in assessing subgroup effects might be misspecified especially when multiple subgroups are considered. Second, even when the model is correctly specified, the observed effect size from the most promising subgroup is biased and tends to be overly optimistic due to selection bias in subgroup pursuit. Therefore, common statistical analysis applied to a post hoc-identified subgroup not only relies on model assumptions but also can be biased which together can lead to misleading decisions in subgroup pursuit. In this article, we propose a statistical risk index to measure the risk that the observed effect of the best selected subgroup is a fluke as a screening tool for subgroup pursuit. The proposed quantitative assessment of risk can help better-informed and robust decisions on subgroup pursuit.

It is well-recognized that the model misspecification and over-optimism of the observed effect size of the most promising subgroup could lead to misinformed decisions on subgroup pursuit in practice (Guo and He, 2020; Thomas and Bornkamp, 2017). Take the MONET-1 trial for example, which is the study of motesanib plus carboplatin/ paclitaxel (C/P) in patients with advanced nonsquamous nonsmall-cell lung cancer (NSCLC). Researchers identified the East Asian patients as the most promising subgroup from the failed phase III trial of MONET-1 for the overall patient population (Kubota et al., 2014). Since the observed effect size of the East Asian patients was very encouraging, the drug developer, Amgen, decided to pursue this subgroup with a new trial (AMG-706) to confirm the efficacy of the treatment for the sub-population. However, the result of the follow-up trial did not meet the primary endpoint (Kubota et al., 2017). Naturally, it is desirable to understand what may have led to the discrepancy from those trials. We argue that to make a better-informed decision on subgroup pursuit, a well-justified and robust measure of risk that the observed effect of the best selected subgroup is a fluke appropriately accounting for the selection bias can be helpful.

To fix ideas, we focus on the (possibly censored) time-to-event data and consider the log-hazard ratio as the treatment effect. Although the proportional hazard models and the log-hazard ratio are routinely used in clinical trials, the strict proportional hazard assumptions might not be satisfied in assessing the treatment effects of subgroups which we also call subgroup effects throughout the rest of the paper (Kleinbaum and Klein, 2010). 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, risk measure in subgroup pursuit should allow the proportional hazard models to be misspecified.

Besides model misspecification, selection bias is another issue in quantifying the risk in subgroup pursuit. Consider a relatively simple case where there are two subgroups from an existing trial. Let β_1, β_2 and β denote the true treatment effects for the two subgroups and the combined group respectively, and $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}$ be the corresponding estimators from proportional hazard models. In this paper, a higher value of β corresponds to a better treatment effect. If $\hat{\beta}_1 > \hat{\beta}_2$, then, subgroup 1 is the most promising subgroup identified from the data and the primary question to address here is when we should recommend the pursuit of subgroup 1 with follow-up confirmatory trials. A naive approach is to make the decision simply based on the observed effect size of the seemingly best subgroup, $T = \max(\hat{\beta}_1, \hat{\beta}_2)$. However, Figure 1 shows that $T = \max(\hat{\beta}_1, \hat{\beta}_2)$ is biased and an overly optimistic estimator for the best subgroup effect, $\tau = \max(\beta_1, \beta_2)$, especially when β_1 is equal or very close to β_2. Therefore, the naive method, simply taking $\max(\hat{\beta}_1, \hat{\beta}_2)$ as the estimated effect size of the most promising subgroup, is not recommended as a measure of risk in subgroup pursuit. This phenomenon of over-optimism is more striking as the number of candidate subgroups increases, which we refer to Section 3 for a more detailed discussion.

To make a better-informed decision on subgroup pursuit, we must address the bias of $\max(\hat{\beta}_1, \hat{\beta}_2)$ and quantify the risk that the observed effect of the best selected subgroup is a fluke in a robust way. To this end, we propose a model-free

measure of risk as the bootstrap probability of at least one estimated subgroup effect size is as good as or better than $\max(\hat{\beta}_1, \hat{\beta}_2)$ where the bootstrap sample is generated by a simple nonparametric bootstrap procedure to reflect sampling under the assumption of a homogeneous population (i.e., no subgroups). The proposed measure of risk is built on a well established result of robustness (Struthers and Kalbfleisch, 1986; Lin and Wei, 1989) 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. In this paper, we start from their results and establish the bootstrap consistency for the misspecified proportional hazard model. This theoretical result enables us to justify and interpret the proposed risk index without imposing strict assumptions of proportional hazard and it is in this sense that the risk index is model-free. Specifically, we show that, asymptotically, the measure of risk approximates the probability of observing one subgroup effect at least as good as $\max(\hat{\beta}_1, \hat{\beta}_2)$ when the population is homogeneous with $\beta_1 = \beta_2 = \hat{\beta}$, a clearly unfavorable situation for further pursuit of the subgroup. Therefore, a higher risk index indicates a higher chance of a statistical fluke for the observed effect of the best selected subgroup, and therefore argues against the subgroup pursuit without stronger reasons from biological or other scientific grounds.

Our work is related to but not about the interim analysis in the adaptive design of subgroup analysis (Friede et al., 2012; Stallard et al., 2014). Much of the existing literature focused primarily on the decision of pursuing a pre-determined subgroup, whereas we consider the pursuit of a promising subgroup identified from the data. Our work is also related to the literature on subgroup confirmation (Shen and He, 2015; Fan et al., 2017). Statistical inference in subgroup analysis has mostly been model-based, and our model-free approach can provide a more robust answer in subgroup analysis. Some ad-hoc methods (Rosenkranz, 2016; Stallard et al., 2008) and Bayesian methods (Bornkamp et al., 2017) and resampling-based methods (Hall and Miller, 2010; Guo and He, 2020) have been proposed to address the bias issue in subgroup pursuit; in particular, for inference. The proposed risk index is distinguished from them by providing a well-justified risk index with simple interpretation in the frequentist sense. Our work also contributes to robust statistics by establishing the bootstrap consistency when the proportional hazard model is misspecified.

In summary, in this paper, we help address the question of when to recommend the pursuit of the most promising subgroup identified from the existing data by proposing a robust quantitative assessment of risk that the observed effect of the best selected subgroup is a fluke. The remainder of the paper is organized as follows. In Section 2, we propose a simple non-parametric bootstrap-based measure of the risk in subgroup pursuit and discuss the statistical properties and robustness of the proposed index. In Section 3, we give a quick look to the selection bias in subgroup pursuit and show how it might lead to misinformed decisions in practice. Section 4 gives concluding remarks of the methods we propose.

2. 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 event 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, $(Y, \delta) \sim D$, 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. Here, we focus on the case of two non-overlapped subgroups for simplicity but the proposed risk index can be applied to the case with more than two or overlapped subgroups.

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 one. If we pursue subgroup 1 by making additional investments in a new confirmatory 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 have a quantitative risk appreciation before a new clinical trial is planned. There are, unarguably, many ways such a risk can be measured. We focus on a simple risk index as the bootstrap probability of observing at least one subgroup whose estimated log-hazard ratio is as good as or better than $\hat{\beta}_1$ where the bootstrap sample is generated without any subgroup differences.

2.1. 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, $(Y, \delta) \sim D$, in 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. When the model assumption is subject to violation, we still perform analysis under the working model of proportional hazard, $(Y, \delta) \sim D$, and β is then the implicitly defined parameter, or, more specifically, the zero-crossing of the expectation of the Cox model score equation (Lin and Wei, 1989), 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 Bernoulli$(1, p)$, which determines n_1 and n_2.

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. 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 specified or known for our analysis, and it is in this sense that the proposed measure of risk is model-free and robust.

2.2. Computation of the proposed risk index

Although the observed effect size of the selected subgroup, $\max(\hat{\beta}_1, \hat{\beta}_2)$, is a biased estimate of the best subgroup effect size, the statistic itself is interpretable and simple even when the proportional hazard model is misspecified. Therefore, we aim to propose a risk index directly based on this statistic and naturally accounting for the selection bias. Roughly speaking, we wish to measure the risk that the observed effect of the best selected subgroup is just a fluke by calculating the probability of observing at least one subgroup whose observed effect size is as good as or better than $\max(\hat{\beta}_1, \hat{\beta}_2)$ when the subgroups are actually homogeneous in subgroup pursuit. 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 in Algorithm 2. Then, our proposed risk index is

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 treatment effect within the two groups, $\beta_{1,b}^*$ and $\beta_{2,b}^*$, based on the bootstrap sample.
5: **end for**
6: The risk index is $RI_B = B^{-1} \sum_{b=1}^{B} I\{\max(\beta_{1,b}^*, \beta_{2,b}^*) \geq \max(\hat{\beta}_1, \hat{\beta}_2)\}$.

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

and is calculated as follows.

As $B \to \infty$, the index 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^*, are made to preserve the same number of subjects in each subgroup. Here, we consider nonparametric bootstrap for the following two reasons. First, the nonparametric bootstrap allows us to justify the risk index without imposing the model assumption of proportional hazard. Second, it is unclear how to extend parametric bootstrap in the presence of overlapped subgroups. In general, the above resampling scheme ensures that the bootstrap distribution is homogeneous across subgroups, irrespective of whether there exist subgroups in the original sample and the proportional hazard model is correctly specified or not. When there are multiple or overlapped subgroups, the risk index can be calculated by the bootstrap estimate for each subgroup with the same bootstrap procedure as described in Algorithm 2. The detailed algorithm of the risk index when there are multiple

Algorithm 2 Risk index for subgroup pursuit (when they are k subgroups).

1: **for** $b = 1...B$ **do**
2: **Partial bootstrap:** Generate $\{(Y_l^*, D_l^*, \delta_l^*) : l = 1, \cdots, n\}$ as a bootstrap sample from the set $\{(Y_j, D_j, \delta_j), j = 1, \cdots, n\}$.
3: **Subgroup assignment:** $Z_l^* = Z_l$ for $l = 1, \cdots, n$.
4: **Estimation:** Calculate the log-hazard ratio estimate of treatment effect within each subgroups, $\beta_{i,b}^*$, based on the bootstrap sample for $i = 1, \ldots, k$.
5: **end for**
6: The risk index is $RI_B = B^{-1} \sum_{b=1}^{B} I\{\max_i \beta_{i,b}^* \geq \max_i \hat{\beta}_i\}$.

subgroups are provided in the Appendix D.

2.3. Robustness and relevance of the risk index

To see how the risk index might be used as a screening tool in subgroup pursuit, we need to understand the limiting behavior of RI^* without imposing the strict proportional hazard assumption. 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

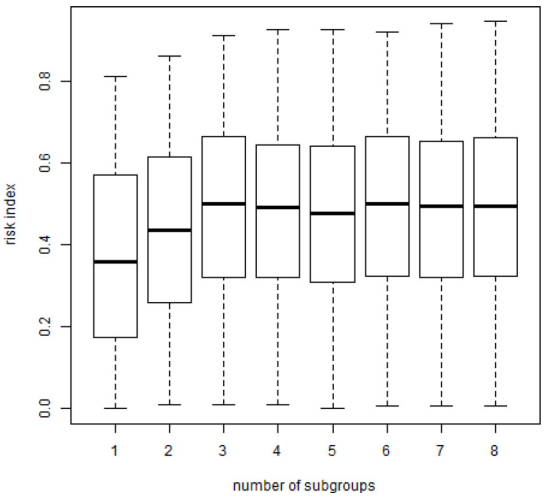

Fig. 2. Boxplots of the risk index: synthetic data generating model of MONET-1 (no subgroups).

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

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

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

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

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

Assumption 1 requires the mapping $\beta \to F_\beta$ to be continuous at β_0 under the sup norm. Assumption 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. Both assumptions are expected to be satisfied for most practical scenarios and are much more general than the strict proportional hazard assumption though we use log-hazard ratio in assessing subgroup effect.

To understand the behavior of the proposed risk index, we first need to establish the bootstrap consistency under the misspecified proportional hazard model. While previous works have shown the asymptotic normality of the log-hazard ratio estimate for the misspecified proportional hazard model (Lin and Wei, 1989), bootstrap consistency is still lacking. In the Supplementary Material, we show that the distribution of the bootstrap estimate is consistent via the approximate linear expansion from Lin and Wei (1989). With that, we have the following theory to justify the use and robustness of the proposed risk index.

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

Theorem 1 ensures that the risk index RI^* is asymptotically the same as RI, the probability of observing one subgroup that is at least as promising as $\max(\hat{\beta}_1, \hat{\beta}_2)$ 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 risk that the observed effect of the best selected subgroup is just a fluke no matter the proportional hazard model is correctly specified or not. 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, it is not difficult to see 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 $\beta_1 = \beta_2 = \beta_0$; there exist no subgroups. 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 as shown in Figure 2. In this sense, we can use the risk index as a robust measure of risk that the observed effect of the best selected subgroup is just a fluke in 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 consider a value of 0.15 and 0.30 as indication of risk and high risk that the observed effect of the best selected subgroup is just a fluke in subgroup pursuit, respectively.

The use and the properties of the risk index do not require the distribution family F_β to be known or specified and therefore the proposed risk index is robust. To see an example of mis-specified working models, consider a special case

Table 1
Risk index for MONET-1 study: synthetic data

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

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 1 and 2, but the working model without including W as a covariate would be mis-specified. As shown in the previous works in Lin and Wei (1989), β remains to be the log-hazard ratio (approximately) under the working model, and the risk index can be used without relying on the true model.

The robustness discussed here is specific to model misspecification. Other aspects of robustness such as the influence function for the proportional hazard model can be found in the literature; see, for example, Reid and Crépeau (1985). For other robust estimates of $(\hat{\beta}_1, \hat{\beta}_2)$ under different models, we refer to Hampel et al. (1986) and much of the subsequent work.

2.4. 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, 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 for any reasonable measure of risk in subgroup pursuit 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 $\max(\hat{\beta}_1, \hat{\beta}_2)$, the widely used quantity for subgroup pursuit decisions in practice, and addresses its bias appropriately. Therefore, our proposed risk index is more simple and robust than the p-values from the commonly used test statistics for the null hypothesis of homogeneity.

2.5. Applications

To illustrate how the proposed risk index might help make better-informed decisions in subgroup pursuit in practice, we apply the risk index to synthetic data from the MONET-1 study and the real data from the panitumumab study in clinical trials.

MONET-1 is a study of motesanib plus carboplatin/ paclitaxel (P/C) in patients with advanced nonsquamous nonsmall-cell lung cancer. Subgroups were considered after the trial failed to show clear significance in the overall population. Analysts identified the subgroup of East Asians as the most promising subgroup with the hazard ratio HR=0.669 and P-value=0.0223 (Kubota et al., 2014). Encouraged by the subgroup analysis, a confirmatory trial was conducted on the East Asian subpopulation. However, the results of the follow-up trial failed to confirm the efficacy of the treatment with hazard ratio HR=0.81 and P-value=0.0825 (Kubota et al., 2017). We use this example to demonstrate how the proposed risk index can help measure the risk that the observed effect of the East Asian subgroup is just a fluke with the MONET-1 data.

Since the MONET-1 data are not publicly available and how the East Asian subgroup is selected is unclear, we use synthetic data in this paper and consider different number of candidate subgroups. The data (with 1090 patients as in the MONET1 study) are generated from a model that mirrors the main features of the estimated survival functions in Figure 1.A of Kubota et al. (2014) and assumes the subgroups are homogeneous without treatment effect. The details of the synthetic data is given in the Appendix B. We consider the number of candidate subgroups ranging from 2 to 16 based on the binary coding (yes or no) of each of the following variables: East Asian patient, stage IIIB, received radiotherapy, male, Age greater than 65, never smoked, ECOG PS status 0 and Adenocarcinoma histology. Because each binary variable splits the sample into two subgroups, we have a total 16 candidate subgroups to consider if all eight variables are considered in the planning stage and the candidate subgroups are clearly overlapping. To make it consistent with the framework used in this paper, we use the negative log-hazard ratio as the treatment effect in calculating the risk index. For the MONET-1 study, we do not know how many candidate subgroups were actually considered to single out the East Asian subgroup, so we calculate the risk index for different number of candidate subgroups with $B = 1000$ as summarized in Table 1 to show that the risk indeed increases with the number of candidate subgroups used in the analysis. Ignoring how the subgroup is selected would disallow us to measure the risk in subgroup pursuit appropriately.

From Table 1, we see that if 12 or more candidate subgroups were considered in subgroup identification, the risk index is over 0.15, which means that even if the population is indeed homogeneous (i.e. no subgroups), we have higher than 15% chance to observe a subgroup that is at least as promising as the result we saw for 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 many candidate subgroups were considered, we would have to be far more cautious as the risk that HR=0.663 for East Asian subgroup is just a fluke is relatively large.

The panitumumab study is a study of panitumumab in patients with metastatic colorectal cancer (mCRC). The initial trial failed to confirm the efficacy of panitumumab in patients with mCRC (Van Cutsem et al. (2007)), and the European Medicines Agency (EMA) declined to approve the treatment. Amado et al. (2008) identified a subgroup of the patients with wild-type KRAS, an oncogene existing in about 50% of the patients with mCRC, as the most promising subgroup with $HR = 0.45$. A follow-up phase III trial for the wild-type KRAS subgroup was reported in Peeters et al. (2014). The pursuit of the wild-type KRAS subgroup succeeded, and panitumumab has been approved by both EMA and FDA.

To see what the proposed risk index would have conveyed in such studies, we consider the problem of two or four candidate subgroups but vary the total number of subjects in the samples from 341 (as in the original study reported in Amado et al. (2008)) to thrice as many so it matches the sample size in MONTE1 study; more details can be found in the Appendix B. The two subgroups are wild-type KRAS versus the rest, and the additional two subgroups are male or female. The use of a larger number of subjects in the samples does not address the actual panitumumab study but aims to demonstrate how the risk index would decease with the sample size in the trial in the cases where a meaningful subgroup does exist. As with the MONET1 study, we use $B = 1000$ in the calculation of the risk index.

At the original sample size of 341 subjects, the risk index stood at 0.12 for two candidate subgroups and 0.24 for four candidate subgroups, which indicates a modest level of risk in subgroup pursuit. If the sample size increases (but without changing the empirical distributions of the patient responses), the risk index decreases to 0.01 and 0.02, respectively, as the sample size reaches 1000. This is consistent with the asymptotic result that the risk index would converge to zero as the sample size increases unless we have a homogeneous population. This is also consistent with our general brief that more caution is needed when the most promising subgroup is identified from a smaller trial. The comparison of the risk index in the MONET1 study and in this study at a comparable sample size indicates that the risk in the MONET1 study was much higher as confirmed by the follow-up studies.

2.6. Simulation

To further understand the behavior of the proposed risk index, we consider a simulation setting based on the synthetic data generating model of MONET-1 used in Section 2.5. Again, we consider the number of candidate subgroups ranging from 2 to 16 based on the binary coding (yes or no) of the same variables used in Section 2.5 and the negative log-hazard ratio as the subgroup effect to make it consistent with the framework used in the paper.

Figure 2 displays the boxplots of the risk index based on 2000 Monte Carlo samples given the East Asian was selected as the most promising subgroup with $B = 200$. It is clear that the risk index increases but levels off quickly as the number of candidate subgroups increases. This is also consistent with the practical experience that we should be more cautious in the pursuit of the selected subgroup if it is identified from many candidate subgroups. More simulation results with various sample sizes, various numbers of subgroups and various differences between subgroups are included in the Appendix C where the behavior of the risk index is consistent to Theorem 1 and demonstrates the relevance of the risk index.

3. Selection bias in subgroup pursuit

To better understand the selection bias on subgroup pursuit, we consider quantifying the bias of the observed best selected subgroup effect in this section.

3.1. Bias quantification

For the sake of simplicity, we consider two predefined subgroups and let $\tau = \max\{\beta_1, \beta_2\}$ denote the best subgroup effect. We start with the examination of the bias for using the naive estimate, the observed effect size of the most promising subgroup identified from the trial, $T = \max(\hat{\beta}_1, \hat{\beta}_2)$, as an estimator for τ. Consider the case where $\sqrt{n}(\hat{\beta}_1 - \beta_1, \hat{\beta}_2 - \beta_2)$ is asymptotically normal with mean zero and variance-covariance matrix $((\sigma_1^2, \rho\sigma_1\sigma_2)', (\rho\sigma_1\sigma_2, \sigma_2^2)')$. When the two subgroups do not overlap, the correlation $\rho = 0$. Write $\sigma_{jn} = n^{-1/2}\sigma_j$ as the standard error of $\hat{\beta}_j$ for $j = 1, 2$. Let $\theta_n = \sqrt{\sigma_{1n}^2 + \sigma_{2n}^2 - 2\rho\sigma_{1n}\sigma_{2n}}$. If $(\hat{\beta}_1, \hat{\beta}_2)$ is exactly normal, following the study of the skew normal distribution in Nadarajah and Kotz (2008), we have

$$E(T) - \tau \doteq \eta_n = -\Psi(-\delta_n)\theta_n\delta_n + \theta_n\phi(\delta_n), \tag{3.1}$$

where $\delta_n = |\beta_1 - \beta_2|/\theta_n$, and Ψ and ϕ are the distribution function and the density function of the standard normal variable, respectively. Similar quantification of bias for multiple subgroups depending on the distances between and the standard

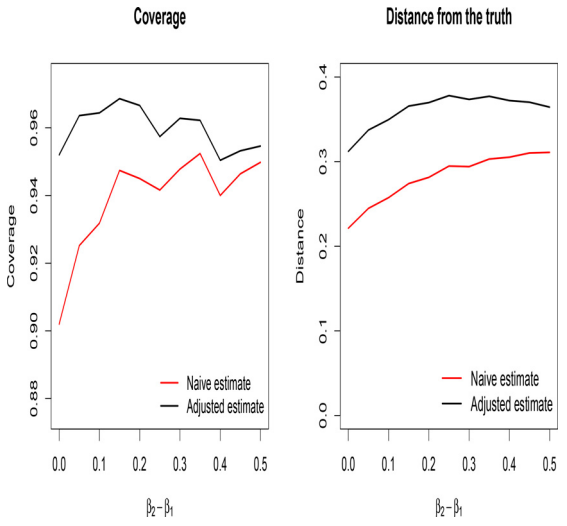

Fig. 3. Coverage rates and distance from the truth for two estimates of the 95% lower bound of τ with standard error smaller than 0.01.

deviations of candidate subgroups can be derived (Nadarajah and Kotz, 2008). When $(\hat{\beta}_1, \hat{\beta}_2)$ is asymptotically normal, it is not difficult to see that η_n is the approximate bias. For small to moderate sample sizes, the bias may not be negligible and actually can be in the order of $1/\sqrt{n}$. For example, when $\beta_1 = \beta_2$, $\eta_n = \theta_n \phi(0) = O(1/\sqrt{n})$. Therefore, simply using the observed best subgroup effect to measure the risk in subgroup pursuit can be misleading, because the bias could be in the same magnitude as the standard error, invalidating any resulting inference.

3.2. Simulation

To illustrate the bias in practical settings, we consider a simulation study with $n_1 = n_2 = 200$ for a total sample size of 400 based on 2000 Monte Carlo samples. Let $D_i \sim$ Bernoulli(0.5) be the treatment indicator. For observations in subgroup j (i.e., with $Z_i = j$), $j = 1, 2$, we generate the event time from the following Weibull regression model

$$\log(T_i) = 0.5\varepsilon_i - 0.5\beta_j D_i,$$

where ε_i follows a standard extreme value distribution. Thus β_j is the true log-hazard ratio of treatment for subgroup j which we also call subgroup effect for the j-th subgroup. We let $\beta_1 = \log(1.1)$ in subgroup 1 and let β_2 in subgroup 2 stay in the range of $[\beta_1, \beta_1 + 0.5]$. We generate the censoring time C such that $\log(C) \sim$ Uniform (-1.25,1), leading to around 40% censoring. The follow-up time is $Y = \min(T, C)$, and the censoring indicator is $\delta = I(T \le C)$.

We first vary the true value of β_2 in subgroup 2 and examine the bias of the observed best subgroup effect as displayed in Figure 1. It is clear that the naive estimator tends to over-estimate the truth, especially when $d = |\beta_2 - \beta_1|$ is small and the bias is monotonically decreasing with d. When $\beta_1 = \beta_2$ the bias is the largest and is about 0.1, which indicates a non-negligible 10% inflation of the hazard ratio estimate.

To address the selection bias, a natural idea is to estimate the bias η_n and subtract it from the naive estimate. Unfortunately, η_n cannot be estimated up to the order of $1/\sqrt{n}$, and plugging the estimated counterparts into (3.1) does not lead to satisfactory results in our simulation studies and tends to underestimate the bias when the distance between two subgroups $d = |\beta_2 - \beta_1|$ is small. Noting the decreasing trend in Figure 1, we consider a bias-adjustment procedure by first deriving a level $1 - \alpha_1$ lower confidence bound for d, denoted by $\hat{d}_L$, as a conservative estimate for d, and plug it into Eq. (3.1) to obtain a conservative bias estimate for T, denoted by $\hat{\eta}(\alpha_1)$. We consider $T - \hat{\eta}(\alpha_1)$ as an adjusted estimate of τ and proceed the inference with the adjusted estimate.

Figure 3 displays the coverage rates of the 95% one-sided lower confidence bound based on the naive estimate and the adjusted estimate described above. For the latter method, we used $\alpha_1 = 0.2$ in estimating d, and the coverage rates for the latter are clearly better. This simple simulation experiment shows that the naive approach is clearly anti-conservative, and we need an appropriate adjustment to selection bias to overcome any over-optimism. For an asymptotically sharp inference procedure on the best selected subgroup effect, we refer to Guo and He (2020).

3.3. Synthetic data: MONET-1 continued

To show how selection bias might lead to misinformed decisions on subgroup pursuit in practice, we apply the naive and the adjusted inference procedure on the best subgroup effect to the synthetic data of MONET-1 used in Section 2.5. Here we focus on the case of two candidate subgroups, East Asians and the rest, and take $\alpha_1 = 0.2$ in the bias adjustment calculation.

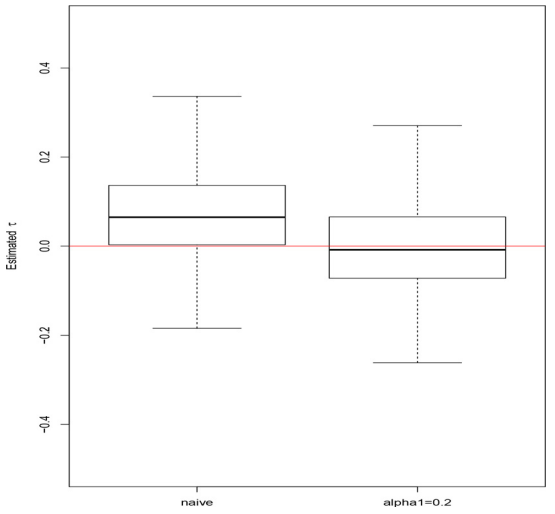

Fig. 4. Boxplots of the naive estimator and the adjusted estimator. The red line gives the true τ.

From the arguments in Section 2.5, we again use the negative log-hazard ratio as the treatment effect here to make it consistent to the framework in the paper. The naive estimate of the negative log-hazard ratio is 0.411, and the naive 95% lower confidence bound is 0.124, while the adjusted estimate is 0.369 with 95% lower confidence bound 0.101. From the result, we see that the adjusted estimate of the best subgroup effect is slightly weaker than that indicated by the naive estimate, but the 95% lower bound is still above zero. The result is consistent with the low risk index obtained in Section 2.5 in the scenario of two candidate subgroups in the planning stage.

By generating the synthetic data 2000 times without subgroup effects, we assess the empirical coverage of the 95% lower bound and the average distance between the estimate and the true parameter. The empirical coverage of the naive method is 0.904 with the average distance of 0.145 from the true value, but the empirical coverage of the adjusted method is 0.963 with the average distance 0.210. In Figure 4, we compare the boxplots of the naive estimate and the adjusted estimate. It is clear that the naive method failed to achieve the desired coverage probability and is overly optimistic. In summary, the selection bias can lead to significant over-estimation of the effect size of the selected subgroup and misinformed decisions on subgroup pursuit.

4. Conclusions

In this paper, we propose a robust statistical quantification for the risk that the observed effect of the best selected subgroup is just a fluke and give a quick look to the selection bias in subgroup pursuit. The proposed quantitative assessment appropriately handles the model misspecification issue and accounts for the bias in the observed treatment effect for a subgroup identified from outcome-based selection, and can help us make a better-informed decision of subgroup pursuit.

Our proposed risk index measures the risk from a reasonable and robust angle and can be easily generalized to the situations where there are multiple overlapped subgroups. The risk index is model-free, easy-to-compute, and simple. The behavior of the risk index is well understood and the index might be used as a quantitative screening tool in subgroup pursuit. The theoretical investigation of the risk index is built on the newly established bootstrap consistency of misspecified proportional hazard model and demonstrates the advantages of the proposed risk index over p-value. We also take a quick look at the selection bias in subgroup pursuit, and argue that the selection bias could lead to a misinformed decision on subgroup pursuit especially when subgroups are close.

We have examined the risk index to two earlier clinical trials, the MONET-1 study and the panitumumab study. The results show that the proposed method aims to address the bias issue appropriately and may be used as a screening tool and help make a more informative decision on subgroup pursuit. We emphasize that the risk that the observed effect of the best selected subgroup is just a fluke generally increases with the number of candidate subgroups considered in the identification of the best subgroup, and the details of how a subgroup is selected need to be known before any decision is made on subgroup pursuit. The proposed method might be also used to measure the risk of adverse effect as the observed adverse effect in the most vulnerable subgroup might be also a fluke due to selection bias (Guo et al., 2022).

The proposed risk index measures the risk that the observed effect of the best selected subgroup is just a fluke. In practice, we might need to take other risk factors, such as budget constraints, into account in subgroup pursuit, which we would like to leave to future works. We note that the methods we have considered here are designed for the scenario of having a set of pre-specified subgroup candidates. This is often a recommended approach in subgroup pursuit. However, post hoc subgroup identification without pre-specified subgroups is often used in practice (Cai et al., 2010; Lipkovich et al., 2011), therefore, it will be of interest to generalize our work and propose appropriate robust statistical quantification of

risk to more challenging scenarios where the subgroups are post hoc identified. In the end, besides the best subgroup, the second best subgroup might be of interest in practice. One natural idea to quantify the risk that the observed effect of the second best selected subgroup is just a fluke in subgroup pursuit is to apply the proposed risk index to the subgroups with the best selected subgroup removed. We would like to leave the development of the risk quantification for the other promising subgroups to future work.

Appendix A

The Appendix contains the technical proof of Theorem 1 in *Appendix A*, the data processing for the MONET-1 study and the panitumumab study in *Appendix B*, additional simulation in *Appendix C* and the detailed algorithm when there are multiple subgroups in *Appendix D*.

A1. Proof of Theorem 1

Let $\triangledown l_n(r)$, $\triangledown^2 l_n(r)$ denote the first and second derivative of the log partial likelihood evaluated at r based on $\{(Y_i, \delta_i, D_i)\}_{i=1}^n$, 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, $S^{(q)}(\beta, t) = \sum_{i=1}^n Y_i(t)e^{\beta D_i}D_i^q/n$, $s_\gamma^{(q)}(\beta, t) = E_\gamma S^{(q)}(\beta, t)$, and

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

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

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

The notations, $\triangledown\tilde{l}$ and $\triangledown^2\tilde{l}$, do not mean that they are the first and second derivatives of some quantities, instead, we use these notations because they are approximations to $\triangledown l_n$ and $\triangledown^2 l_n$ respectively. In the end, we let $A_i(r) = E_{\beta_i}\{\triangledown^2\tilde{l}_{\beta_i,n}(r)\}$ 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 1 which assumes that uniformly continuity in Assumption 1 is true for a neighborhood B of β_0 and we still call it Assumption 1 in this supplementary material. The additional assumptions are not essential and for other situations, the proof is similar. Recall Assumptions 1 and 2, we first establish some basic properties of some key quantities.

Lemma 1.1. *Under Assumption 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,

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

From Assumption 1, P_γ is continuous at $\gamma = \beta_0$, so the first term on the right hand side of (A.1) goes to 0 as β and γ go to β_0. From Assumption 1, $P_\beta(Y \ge t|D = 1)$ is continuous at $\beta = \beta_0$ uniformly in t, so the other term on the right hand side

of (A.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 1.1 (2) and Assumption 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

$$P_\gamma\left(\sup_{\beta\in\mathbb{B},t\in R}|S^{(q)}(\beta,t)-s_\gamma^{(q)}(\beta,t)|>\epsilon\right)\le\sum_{i=1}^{i=k}P_\gamma\left(|S^{(q)}(\beta_{i,\gamma},t_{i,\gamma})-s_\gamma^{(q)}(\beta_{i,\gamma},t_{i,\gamma})|>\epsilon/3\right). \tag{A.2}$$

For any γ and t,

$$\begin{aligned}&P_\gamma\left(|S^{(q)}(\beta,t)-s_\gamma^{(q)}(\beta,t)|>\epsilon\right)\\&\quad\le var_\gamma\left\{S^{(q)}(\beta,t)\right\}/\epsilon^2=var_\gamma\left\{Y(t)e^{D\beta}D^q\right\}/(n\epsilon^2)\le e^{2\beta}/(n\epsilon^2).\end{aligned} \tag{A.3}$$

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 (A.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 1.1 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 1.1 (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

$$\sup_{\beta,\gamma\in\mathbb{B}}E_\gamma\left\{\int_0^\infty\frac{Y(t)}{s_\gamma^{(0)}(\beta,t)}d\bar{F}_\gamma(t)\right\}^2\le\sup_{\beta,\gamma\in\mathbb{B}}C_{\mathbb{B}}E_\gamma\left\{\int_0^\infty\frac{Y(t)}{1-G_\gamma(t)}dG_\gamma(t)\right\}^2. \tag{A.4}$$

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

$$E_\gamma\left\{\int_{M_1}^\infty\frac{Y(t)}{1-G_\gamma(t)}dG_\gamma(t)\right\}^2\le\int_0^{1-G_\gamma(M_1)}\log^2 xdx<\infty.$$

From Assumption 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 1.2. *Under Assumption 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 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 misspecified Cox model in the Appendix of Lin and Wei (1989), we have the following useful decomposition,

$$\begin{aligned}&\sqrt{n}\left\{\bigtriangledown l_n(r)-\bigtriangledown\tilde{l}_{\hat\beta,n}(r)\right\}\\&\quad=-\int_0^\infty\left\{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)}-\frac{s_{\hat\beta}^{(1)}(r,t)}{s_{\hat\beta}^{(0)}(r,t)}\right\}d\left[\sqrt{n}\left\{F_n(t)-\bar{F}_{\hat\beta}(t)\right\}\right]\\&\quad\quad-\int_0^\infty\left\{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)}-\frac{s_{\hat\beta}^{(1)}(r,t)}{s_{\hat\beta}^{(0)}(r,t)}\right\}\sqrt{n}\left\{S^{(0)}(r,t)-s_{\hat\beta}^{(0)}(r,t)\right\}/s_{\hat\beta}^{(0)}(r,t)d\bar{F}_{\hat\beta}(t).\end{aligned} \tag{A.5}$$

From Lemma 1.1 (1), (3) and (4), we know that $\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s_{\hat\beta}^{(1)}(r,t)}{s_{\hat\beta}^{(0)}(r,t)}$ is bounded and for any $\tau > 0$, we know that $P_{\hat\beta}(\sup_{r\in\mathbb{B},t\in[0,\tau]}|\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s_{\hat\beta}^{(1)}(r,t)}{s_{\hat\beta}^{(0)}(r,t)}| > \epsilon) \to 0$ 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.

$$\begin{aligned}
&\limsup P_{\hat{\beta}}\left(|-\int_0^{\infty}\left\{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)}-\frac{s_{\hat{\beta}}^{(1)}(r,t)}{s_{\hat{\beta}}^{(0)}(r,t)}\right\}d\left[\sqrt{n}\left\{F_n(t)-\bar{F}_{\hat{\beta}}(t)\right\}\right]| > \epsilon/2\right)\\
&\le \limsup P_{\hat{\beta}}\left(|-\int_0^{\tau_1}\left\{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)}-\frac{s_{\hat{\beta}}^{(1)}(r,t)}{s_{\hat{\beta}}^{(0)}(r,t)}\right\}d\left[\sqrt{n}\left\{F_n(t)-\bar{F}_{\hat{\beta}}(t)\right\}\right]| > \epsilon/4\right)\\
&+ \limsup P_{\hat{\beta}}\left(|-\int_{\tau_1}^{\infty}\left\{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)}-\frac{s_{\hat{\beta}}^{(1)}(r,t)}{s_{\hat{\beta}}^{(0)}(r,t)}\right\}d\left[\sqrt{n}\left\{F_n(t)-\bar{F}_{\hat{\beta}}(t)\right\}\right]| > \epsilon/4\right) \ < \eta
\end{aligned} \tag{A.6}$$

in probability w.r.t P_1. Therefore, we control the 1st term on the right hand side of (A.5). For the 2nd term on the right hand side of (A.5), notice that given $r \in \mathbb{B}$, $\sqrt{n}\{S^{(0)}(r,t) - s_{\hat{\beta}}^{(0)}(r,t)\}$ converges to a zero-mean Gaussian process in probability w.r.t P_1, we can decompose

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

into

$$\begin{aligned}
&\int_0^{\tau_2}\left\{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)}-\frac{s_{\hat{\beta}}^{(1)}(r,t)}{s_{\hat{\beta}}^{(0)}(r,t)}\right\}\sqrt{n}\left\{S^{(0)}(r,t)-s_{\hat{\beta}}^{(0)}(r,t)\right\}/s_{\hat{\beta}}^{(0)}(r,t)d\bar{F}_{\hat{\beta}}(t)\\
&+\int_{\tau_2}^{\infty}\left\{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)}-\frac{s_{\hat{\beta}}^{(1)}(r,t)}{s_{\hat{\beta}}^{(0)}(r,t)}\right\}\sqrt{n}\left\{S^{(0)}(r,t)-s_{\hat{\beta}}^{(0)}(r,t)\right\}/s_{\hat{\beta}}^{(0)}(r,t)d\bar{F}_{\hat{\beta}}(t)
\end{aligned} \tag{A.7}$$

with appropriate τ_2. The first term of (A.7) goes to 0 in probability due to the uniform convergence of $\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s_{\hat{\beta}}^{(1)}(r,t)}{s_{\hat{\beta}}^{(0)}(r,t)}$ to 0, the L_∞ norm of the gaussian process and the boundness of $1/s_{\hat{\beta}}^{(0)}(r,t)$ when $t \in [0, \tau_2]$. To control the second term of (A.7), we note that $\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)} - \frac{s_{\hat{\beta}}^{(1)}(r,t)}{s_{\hat{\beta}}^{(0)}(r,t)}$ is bounded, so there exists a constant C such that

$$\begin{aligned}
&|\int_{\tau_2}^{\infty}\left\{\frac{S^{(1)}(r,t)}{S^{(0)}(r,t)}-\frac{s_{\hat{\beta}}^{(1)}(r,t)}{s_{\hat{\beta}}^{(0)}(r,t)}\right\}\sqrt{n}\left\{S^{(0)}(r,t)-s_{\hat{\beta}}^{(0)}(r,t)\right\}/s_{\hat{\beta}}^{(0)}(r,t)d\bar{F}_{\hat{\beta}}(t)|\\
&\le |\frac{C}{\sqrt{n}}\sum_{i=1}^{n}\int_{\tau_2}^{\infty}\left\{\frac{Y_i(t)e^{rD_i}}{s_{\hat{\beta}}^{(0)}(r,t)}-1\right\}d\bar{F}_{\hat{\beta}}(t)|.
\end{aligned} \tag{A.8}$$

With Chebyshev's inequality, the latter one is controlled by $E_{\hat{\beta}}\{\int_{\tau_2}^{\infty}(\frac{Y_i(t)e^{rD_i}}{s_{\hat{\beta}}^{(0)}(r,t)})d\bar{F}_{\hat{\beta}}(t)\}^2$ after multiplied by a constant, which will go to 0 in probability as $\tau_2 \to \infty$ by Lemma 1.1 (5), so we prove the result.

(2) Second, we 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), we have

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

and the latter is smaller than

$$P_{\hat{\beta}}\left(|\bigtriangledown \tilde{l}_{\hat{\beta},n}(r) - E_{\hat{\beta}}\bigtriangledown \tilde{l}_{\hat{\beta},n}(r)| > \epsilon/2\right) + P_{\hat{\beta}}\left(|E_{\hat{\beta}}\omega_{\hat{\beta}}(r) - E_{\beta_0}\omega_{\beta_0}(r)| > \epsilon/2\right). \tag{A.9}$$

By Chebyshev's inequality, the first term on the right hand side of (A.9) is smaller than $E_{\hat{\beta}}\omega_{\hat{\beta}}^2(r)/n$. From Lemma 1.1 (5), we see that $E_{\hat{\beta}}\omega_{\hat{\beta}}^2(r)$ 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 (A.9), we note that $E_\gamma\omega_\gamma(r) = E_\gamma h(\gamma, r)$, where

$$h(\gamma, r) = \int_0^{\infty}\left\{D - \frac{s_\gamma^{(1)}(r,t)}{s_\gamma^{(0)}(r,t)}\right\}dN(t)$$

and $N(t) = I_{Y\le t,\delta=1}$. Therefore, the quantity in the second term can be further controlled as follows,

$$|E_{\hat\beta}\omega_{\hat\beta}(r) - E_{\beta_0}\omega_{\beta_0}(r)| \le |E_{\hat\beta}h(\hat\beta, r) - E_{\hat\beta}h(\beta_0, r)| + |E_{\hat\beta}h(\beta_0, r) - E_{\beta_0}h(\beta_0, r)|. \tag{A.10}$$

From Lemma 1.1 (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 1.1 (1) and Assumption 1, we can show that $|E_{\hat\beta}h(\hat\beta, r)I_{Y\ge M} - E_{\hat\beta}h(\beta_0, r)I_{Y\ge M}|$ will goes to 0 when M goes to infinite, so we can control the 1st term on the right hand side of (A.10). For the second part, from Assumption 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 (A.10) will go to 0 due to portmanteau lemma.

(3) Last, we 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 1.3. *Under Assumption 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 1.1 (2) and (4), we can show that for any $\tau < \infty$,

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

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

$$\begin{aligned}&\bigtriangledown^2 l_n(\beta_n) - \bigtriangledown^2\tilde l_{\hat\beta,n}(\beta_n)\\ &\quad= \sum_{i=1}^n \delta_i\left[\left\{\frac{s^{(2)}_{\hat\beta}(r,Y_i)}{s^{(0)}_{\hat\beta}(r,Y_i)} - \left(\frac{s^{(1)}_{\hat\beta}(r,Y_i)}{s^{(0)}_{\hat\beta}(r,Y_i)}\right)^2\right\} - \left\{\frac{S^{(2)}(r,Y_i)}{S^{(0)}(r,Y_i)} - \left(\frac{S^{(1)}(r,Y_i)}{S^{(0)}(r,Y_i)}\right)^2\right\}\right]/n.\end{aligned} \tag{A.11}$$

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 1.2 and notice that $\bigtriangledown^2 l_n(r)$ and $\bigtriangledown^2\tilde l_{\hat\beta,n}(r)$ are always bounded, we can show that

$$\begin{aligned}&P_{\hat\beta}\left(|\bigtriangledown^2 l_n(\beta_n) - A_0(\beta_0)| > \epsilon\right)\\ &\quad\sim P_{\hat\beta}\left(|\bigtriangledown^2\tilde l_{\beta_n,n}(\beta_n) - E_{\hat\beta}\bigtriangledown^2\tilde l_{\beta_n,n}(\beta_n)| > \epsilon/2\right) + P_{\hat\beta}(|E_{\hat\beta}\bigtriangledown^2\tilde l_{\beta_n,n}(\beta_n) - A_0(\beta_0)| > \epsilon/2),\end{aligned} \tag{A.12}$$

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 (A.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 1.1 (1), (2), (4) and Assumption 1. The technique is similar to what we use in the proof of Lemma 1.2 and (A.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 1.4. *Under Assumption 1, $\sqrt{n}\bigtriangledown l_n(\hat\beta) \to N(0, E_{\beta_0}\omega^2(\beta_0))$ 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 1.1 (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 1.2. 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,

$$\begin{aligned}&E_{\hat\beta}\left\{\omega^2_{\hat\beta}(\hat\beta)I_{|\omega_{\hat\beta}(\hat\beta)|>\sqrt{n\epsilon_1}}\right\}\\ &\quad= E_{\hat\beta}\left\{\omega^2_{\hat\beta}(\hat\beta)I_{|\omega_{\hat\beta}(\hat\beta)|>\sqrt{n\epsilon_1}}I_{Y<M}\right\} + E_{\hat\beta}\left\{\omega^2_{\hat\beta}(\hat\beta)I_{|\omega_{\hat\beta}(\hat\beta)|>\sqrt{n\epsilon_1}}I_{Y\ge M}\right\}.\end{aligned} \tag{A.13}$$

From Lemma 1.1 (2) and (4), when $Y < M$ and $\gamma, r \in \mathbb{B}$, $\omega^2_\gamma(r)$ is bounded, so the first term of (A.13) will go to 0 in probability w.r.t P_1. The second term is smaller than $E_{\hat\beta}\{\omega^2_{\hat\beta}(\hat\beta)I_{Y\ge M}\}$. From Lemma 1.1 (5), we know that $E_{\hat\beta}\omega^2_{\hat\beta}(\hat\beta)I_{Y\ge M} \to 0$ 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 1.2 and (A.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 1.1 (5). Therefore, we show the normality as desired. □

Lemma 1.5. *Under Assumption 1, $\sqrt{n}(\tilde{\beta} - \hat{\beta}) \to N(0, \sigma_0^2)$ in probability w.r.t P_1 where $\sigma_0^2 = A_0^{-2}(\beta_0)E_{\beta_0}\omega^2_{\beta_0}(\beta_0)$.*

Proof. Take taylor expansion of $\nabla l_n(r)$ at $r = \hat{\beta}$ and apply Lemmas 1.2–1.4, we can get the result. □

To prove the bootstrap counterpart of Lemma 1.5, We let $\nabla l_{P_1,n}(r)$, $\nabla^2 \tilde{l}_{P_1,n}(r)$, $\omega_{P_1,i}(\beta)$ and $A_{P_1}(r)$ similar to the quantities in the 1st paragraph but replace $s_\gamma^{(q)}(\beta, Y_i)$, $\bar{F}_\gamma(t)$, E_{β_i}, $A_0(r)$ with $s_{P_1}^{(q)}(\beta, Y_i)$, $\bar{F}_{P_1}(t)$, E_{P_1} and $A_{P_1}(r)$ and the latter are with regard to P_1 instead of F_γ. We let $\nabla l_n^*(r)$, $\nabla^2 l_n^*(r)$ be the first and second derivative of the log partial likelihood evaluated at r based on the bootstrap sample $\{(Y_i^*, \delta_i^*, D_i^*)\}_{i=1}^n$, and β^* be the bootstrap estimator. We let $S^{(q,*)}(\beta, t) = \sum_{i=1}^n Y_i^*(t)e^{\beta D_i^*}(D_i^*)^q/n$, $\nabla \bar{l}_n(r) = \sum_{i=1}^n \bar{\omega}_i(r)/n$, $\nabla \bar{l}_n^*(r) = \sum_{i=1}^n \bar{\omega}_i^*(r)/n$ and $\nabla^2 l_n^*(r) = \sum_{i=1}^n \delta_i^* \left\{ \frac{S^{(2)}(r, Y_i^*)}{S^{(0)}(r, Y_i^*)} - (\frac{S^{(1)}(r, Y_i^*)}{S^{(0)}(r, Y_i^*)})^2 \right\}/n$, 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 lemma shows the bootstrap consistency under misspecified cox model.

Lemma 1.6. *Under Assumption 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 $\sigma_{P_1}^2 = A_{P_1}^{-2}(\beta_0)E_{P_1}\omega^2_{P_1}(\beta_0)$.*

Proof. The proof is similar to the proof in Lemma 1.5.

(1) First, we construct similar results as Lemma 1.1.

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

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

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

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

Notice that $E^* S^{(q,*)}(\beta, t) = S^{(q)}(\beta, t)$, we prove the second with similar decomposition as the proof in Lemma 1.1 (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 1.1 (5). W.L.O.G, we assume that $P_1(\text{for any } \tau < \infty, \liminf_{\beta \in \mathbb{B}, t \in [0,\tau]} S^{(0)}(\beta, t) > 0) = 1$ 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 show that $P^*(\sqrt{n}|\nabla l_n^*(r) - \nabla \bar{l}_n^*(r)| > \epsilon) \to 0$ in probability w.r.t P_1.

Similar to the proof in Lemma 1.2, we have the following decomposition

$$\begin{aligned}
&\sqrt{n}\left\{ \nabla l_n^*(r) - \nabla \bar{l}_n^*(r) \right\} \\
&= -\int_0^\infty \left\{ \frac{S^{(1,*)}(r, t)}{S^{(0,*)}(r, t)} - \frac{S^{(1)}(r, t)}{S^{(0)}(r, t)} \right\} d\left[\sqrt{n}\{F_n^*(t) - F_n(t)\} \right] \\
&- \int_0^\infty \left\{ \frac{S^{(1,*)}(r, t)}{S^{(0,*)}(r, t)} - \frac{S^{(1)}(r, t)}{S^{(0)}(r, t)} \right\} \sqrt{n}\left\{ S^{(0,*)}(r, t) - S^{(0)}(r, t) \right\}/S^{(0)}(r, t) dF_n(t).
\end{aligned} \tag{A.14}$$

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

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

into

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

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

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 1.2 and control the 1st term on the right hand side of (A.14). We get the results by controlling the 2nd term on the right hand side of (A.14) with similar techniques as used in the proof of Lemma 1.2.

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

We can get the following decomposition

$$\begin{aligned} &P^*(|\bigtriangledown l_n^*(r) - E_{P_1}\omega_{P_1}(r)| > \epsilon) \\ &\quad \le P^*\left(|\bigtriangledown l_n^*(r) - \bigtriangledown \bar{l}_n^*(r)| > \epsilon/3\right) \\ &\quad + P^*\left(|\bigtriangledown \bar{l}_n^*(r) - E^* \bigtriangledown \bar{l}_n^*(r)| > \epsilon/3\right) + P^*\left(|E^* \bigtriangledown \bar{l}_n^*(r) - E_{P_1}\omega_{P_1}(r)| > \epsilon/3\right). \end{aligned} \tag{A.15}$$

The first term on the right hand side of (A.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 (A.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 1.2 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

$$\begin{aligned} &E^*\left\{ \bar{\omega}^*(\hat{\beta})^2 I_{|\bar{\omega}^*(\hat{\beta})| > \sqrt{n}\epsilon} \right\} \\ &\quad = E^*\left\{ \bar{\omega}^*(\hat{\beta})^2 I_{|\bar{\omega}^*(\hat{\beta})| > \sqrt{n}\epsilon} I_{Y^* < M} \right\} + E^*\left\{ \bar{\omega}^*(\hat{\beta})^2 I_{|\bar{\omega}^*(\hat{\beta})| > \sqrt{n}\epsilon} I_{Y^* \ge M} \right\} \\ &\quad \le E^*\left\{ \bar{\omega}^*(\hat{\beta})^2 I_{|\bar{\omega}^*(\hat{\beta})| > \sqrt{n}\epsilon} I_{Y^* < M} \right\} + E^*\left\{ \bar{\omega}^*(\hat{\beta})^2 I_{Y^* \ge M} \right\}. \end{aligned} \tag{A.16}$$

The first term on the right hand side of (A.16) will go to 0 in probability w.r.t P_1 due to the boundness of $\bar{\omega}$ when $Y^* < M$. The second term is asymptotically bounded by $E_{P_1}\omega_{P_1}^2(\beta_0) I_{Y \ge M}$ 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 $E_{P_1}\omega_{P_1}^2(\beta_0)$. Therefore, we can apply Lindeberg-Feller CLT.

In the end, similar to Lemma 1.3, $P^*(|\bigtriangledown^2 l^*(\beta) - \bigtriangledown^2 \bar{l}^*(\beta)| > \epsilon) \to 0$ 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, $A_{P_1}^{-2}(\beta_0)$ 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 1.7. *Under Assumption 1,*

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

in probability w.r.t P_1.

Proof. Let $(U_n, V_n) = (\sum_{i=1}^n \omega_{\hat{\beta},i}(\hat{\beta}) I_{Z_i=1}/\sqrt{n}, \sum_{i=1}^n \omega_{\hat{\beta},i}(\hat{\beta}) I_{Z_i=0}/\sqrt{n})$. Similar to Lemma 1.4, we can show that (U_n, V_n) are jointly normal in asymptotic sense. Therefore, by taking taylor expansion as we did in Lemma 1.5, we have $\sqrt{n}(\tilde{\beta}_1 - \hat{\beta})$

and $\sqrt{n}(\tilde{\beta}_2 - \hat{\beta})$ are jointly normal in asymptotic sense. With Lemma 1.5, 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 1: 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 Lemma 1.5. We note that

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

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

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

Similarly, by Lemma 1.7, we have the following relationship

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

If $\beta_1 = \beta_2$, then, $\sigma_0 = \sigma_{P_1}$ and

$$\begin{aligned} &G\left(\sqrt{n}\max\left(\hat{\beta}_1 - \hat{\beta}, \hat{\beta}_2 - \hat{\beta}\right)|0, 0, \frac{\sigma_0}{\sqrt{p}}, \frac{\sigma_0}{\sqrt{1-p}}, 0\right) \\ &= G\left(\sqrt{n}\max\left(\hat{\beta}_1 - \hat{\beta}, \hat{\beta}_2 - \hat{\beta}\right)|0, 0, \frac{\sigma_{P_1}}{\sqrt{p}}, \frac{\sigma_{P_1}}{\sqrt{1-p}}, 0\right). \end{aligned} \tag{A.17}$$

If $\beta_1 \neq \beta_2$, then, from Assumption 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 1 is proved.

From the proof of Theorem 1, We see that under Assumption 1 and 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.

A2. Data Processing

We give details about the data processing of the MONET-1 study and the panitumumab study in *B1* and *B2* respectively.

B1. MONET-1 Study

We revisit the MONET-1 study by using the synthetic data mimicking the real data reported in Kubota et al. (2014), where the East Asian subgroup was initially identified.

Following the MONET-1 trial, we consider a simple setting of 1090 patients with $(Y_i, D_i, \delta_i, Z_i)$ as the observation for the i-th subject, where Y_i is the (possibly censored) survival time, 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. Specifically, we consider the following subgroups in sequence: 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), with 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).

To generate the synthetic data, we consider an estimated distribution for $(Y_i, D_i, \delta_i, Z_i)$ based on Figure 1.A in Kubota et al. (2014) under the assumptions of no treatment effect. The details of the distribution can be found in the Appendix of Guo and He (2020). With the distribution at hand, we take one realization where the East Asian is identified as the best subgroup and its overall hazard ratio and associated p-value are very close to what were reported in the MONET-1 study as summarized in Table 2 as the synthetic dataset for the MONET-1 study. The synthetic dataset can be found in https://github.com/xinzhoug/Data.

B2. Panibumumab Study

We revisit the panitumumab study by using the data reported in Amado et al. (2008), where the wild-type KRAS subgroup was initially identified.

To facilitate our analysis, we focus on a publicly available dataset, which includes 370 patients in the trial (Amado et al., 2008) with about 7% censoring rate. We deleted 29 patients whose KRAS statuses are missing, and the total number of patients included in our analysis is 341. We consider the candidate subgroups defined by the following two binary biomarkers,

Table 2
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

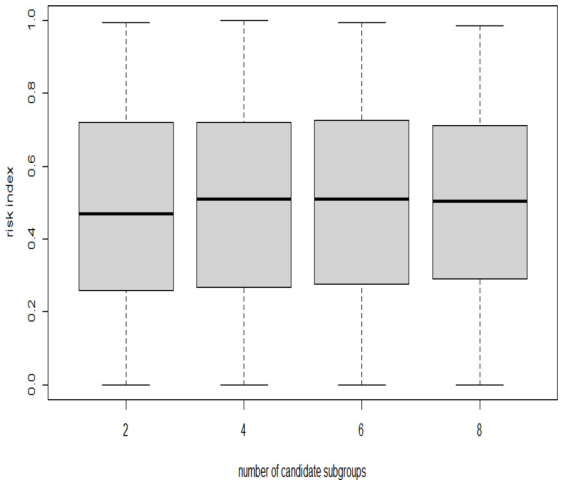

Fig. 5. Boxplots of the risk index: no subgroup exists.

KRAS (wild-type or mutant) and gender (male or female), and it is obvious that the candidate subgroups can overlap. The dataset can be found in https://data.projectdatasphere.org/projectdatasphere/html/content/310.

We also consider a dataset consisting of patients thrice as many as those in the original dataset we describe above. To create such an expanded dataset, we use the patients from the original dataset and let each patient appear three times in the expanded dataset. This expanded dataset is used only for the purpose of illustration.

A3. Additional Simulations

We conduct additional simulation with various sample sizes, various numbers of subgroups and various differences between subgroups to better study the behavior of the risk index. Here, we consider the following proportional hazard model to generate the survival time

$$\lambda(t) = \lambda_0(t)e^{-\beta_i D},$$

where β_i is the treatment effect of the i-th subgroup which we also call the subgroup effect of the i-th subgroup and D is the randomly assigned treatment indicator. We consider the censoring scheme the same as the previous work (Guo and He, 2020) which leads to about 40% censoring. We vary the number of subjects in each nonoverlapped subgroup from 50 to 200 and the number of nonoverlapped candidate subgroups from 2 to 8 and consider $\beta_1 = 1$ (one subgroup stands out) or $\beta_1 = 0$ (no subgroup exists) while keeping other subgroup effects equal 0 to study the behaviors of the risk index. The boxplots of risk index based on 1000 Monte Carlo samples are summarized in Figures 5, 6, 7 and 8. From these figures, we can see that when one subgroup stands out, the risk index will become smaller and go to 0 as sample size increases or the number of candidate subgroups increases, and when no subgroup exists, the risk index will center around 0.5, which is consistent to our theoretical investigation.

We also conduct additional simulation to study how the overlapped subgroups might affect the behavior of risk index. Here, we focus on the no subgroups case and consider overlapped subgroups by assuming each subject fall into each subgroup or not with probability 0.5 with total sample size 400. The boxplots of risk index when we set $\max_i \hat{\beta}_i = 0.3$ based on 1000 Monte Carlo samples are summarized in Figures 9 and 10. From these figures, we can see that in the presence of overlapping subgroups, the proposed risk index tends to be smaller which is consistent with the belief the selection bias is most severe with independent subgroups.

In general, the simulation results are consistent to the theoretical investigation and the usual belief and suggest the relevance of using risk index in measuring the risk that the observed effect of the best selected subgroup is just a fluke.

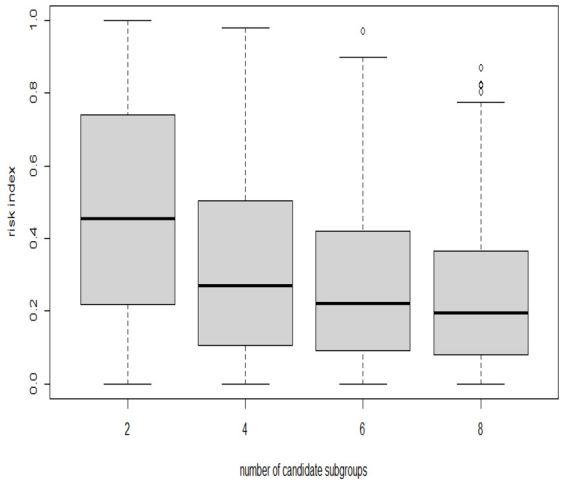

Fig. 6. Boxplots of the risk index: one subgroup stands out with 200 subjects in each subgroup.

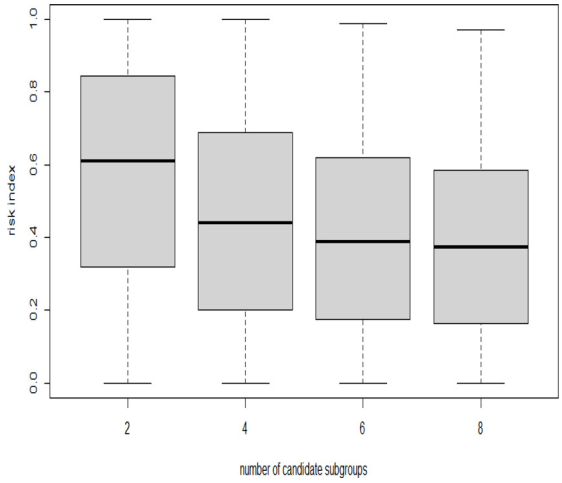

Fig. 7. Boxplots of the risk index: one subgroup stands out with 50 subjects in each subgroup.

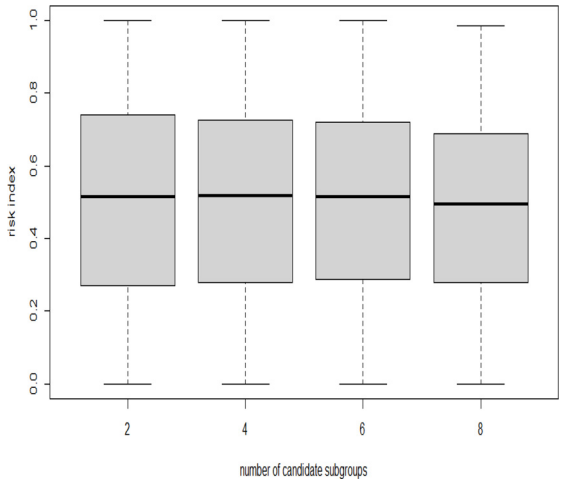

Fig. 8. Boxplots of the risk index: no subgroup exists with 50 subjects in each subgroup.

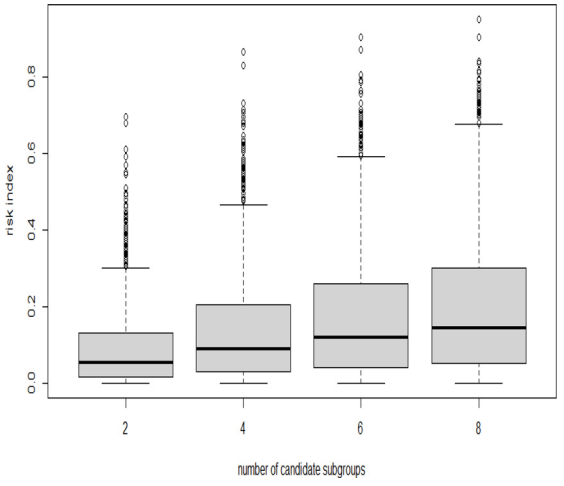

Fig. 9. Boxplots of the risk index: overlapped subgroups.

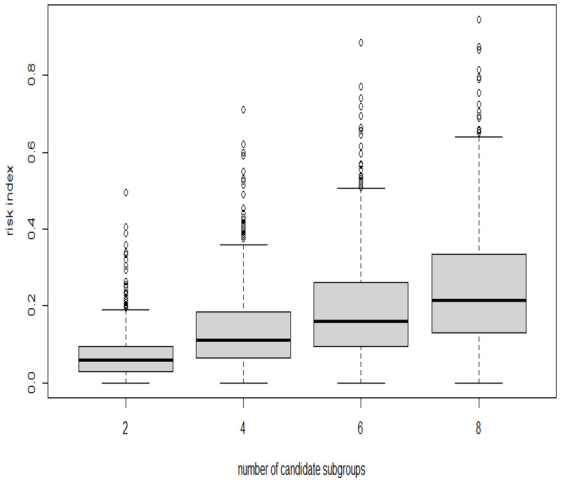

Fig. 10. Boxplots of the risk index: nonoverlapped subgroups.

A4. Algorithm

Here, we present the detailed algorithm to calculate the risk index when there are k (possibly overlapped) subgroups. Let $\hat{\beta}_i$ and $\beta^*_{i,b}$ denote the observed log-hazard ratio and the bootstrap log-hazard ratio of the i-th subgroup respectively.

References

Amado, R. G., Wolf, M., Peeters, M., Van Cutsem, E., Siena, S., Freeman, D. J., Juan, T., Sikorski, R., Suggs, S., Radinsky, R., et al., 2008. Wild-type kras is required for panitumumab efficacy in patients with metastatic colorectal cancer.

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

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

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

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

Guo, X., He, X., 2020. Inference on selected subgroups in clinical trials. Journal of the American Statistical Association 1–19.

Guo, X., Wei, W., Liu, M., Cai, T., Wu, C., Wang, J., 2022. Assessing the most vulnerable subgroup to type ii diabetes associated with statin usage: Evidence from electronic health record data. Journal of the American Statistical Association (just-accepted) 1–26.

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

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley.

Kleinbaum, D.G., Klein, M., 2010. Survival analysis, Vol. 3. Springer.

Kubota, K., Ichinose, Y., Scagliotti, G., Spigel, D., Kim, J., Shinkai, T., Takeda, K., Kim, S.-W., Hsia, T.-C., Li, R., 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., Yoshioka, H., Oshita, F., Hida, T., Yoh, K., Hayashi, H., Kato, T., Kaneda, H., Yamada, K., Tanaka, H., 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. Journal of Clinical Oncology 35 (32), 3662–3670.

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

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

MOLOGEN, 2018. Final analysis of impulse study confirms topline data with positive subgroup results. MOLOGEN Press Releases. https://www.mologen.com/uploads/media/20180406_Press_Release_N_6_MOLOGEN_IMPULSE_final_results.pdf

Nadarajah, S., Kotz, S., 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.

Peeters, M., Price, T., Cervantes, A., Sobrero, A., Ducreux, M., Hotko, Y., André, T., Chan, E., Lordick, F., Punt, C., et al., 2014. Final results from a randomized phase 3 study of folfiri±panitumumab for second-line treatment of metastatic colorectal cancer. Annals of Oncology 25 (1), 107–116.

Reid, N., Crépeau, H., 1985. Influence functions for proportional hazards regression. Biometrika 72 (1), 1–9.

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

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

Stallard, N., Hamborg, T., Parsons, N., Friede, T., 2014. Adaptive designs for confirmatory clinical trials with subgroup selection. Journal of biopharmaceutical statistics 24 (1), 168–187.

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

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

Sun, X., Briel, M., Busse, J.W., You, J.J., Akl, E.A., Mejza, F., Bala, M.M., Bassler, D., Mertz, D., Diaz-Granados, N., et al., 2012. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. Bmj 344, e1553.

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

Van Cutsem, E., Peeters, M., Siena, S., Humblet, Y., Hendlisz, A., Neyns, B., Canon, J.-L., Van Laethem, J.-L., Maurel, J., Richardson, G., et al., 2007. Open-label phase iii trial of panitumumab plus best supportive care compared with best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer. Journal of clinical oncology 25 (13), 1658–1664.