MS Final Exam

Review of comparing a stratified treatment strategy with the standard treatment in randomized clinical trials

By

Yang Wang

DEPARTMENT OF BIOSTATISTICS

COLLEGE OF PUBLIC HEALTH AND HEALTH PROFESSIONS

UNIVERSITY OF FLORIDA

2017

ACKNOWLEDGMENTS

I would like to thank my committee for their generous support and valuable advice for my final project. I want to express my sincere appreciation to my adviser Dr. Baiming Zou, whose deep insight into the field of clinical trial and wise guidance are crucial for my project.

In addition, I would also like to show our gratitude to the professors in our department. I learned not only the professional theoretical knowledge from them, but also the statistical way to think about questions.

Moreover, I would also like to thank my peer classmates. Their hard work always inspires me.

Finally, I would like to show my special appreciation to my family, without their support and encouragement, would not have been possible.

1. Background

Biomarkers have been widely used for disease diagnosis in clinical trial studies. In a traditional clinical trial, in which a single maker is used and the subjects are divided into two subgroups, a new intervention is either tested in the marker positive outcome subgroup [1], or in all subjects [2]. However, as the increasing application of multiple markers in testing different treatments, more and more clinical trials had been designed to test one disease using multiple markers, especially in oncology [3]. For example, umbrella trials have been used for multiple-marker-based clinical trials, in which the trials were designed to test the efficacy of different treatments on several molecular subtype mutations in a single kind of cancer. However, the stratified treatment methods employed in these clinical studies also pose a challenge for statistical analysis.

The FOCUS4 study represents a typical umbrella trial that involves different biomarkers and different interventions along multi-stage of disease advancement [4]. In that study, patients with metastatic colorectal cancer were firstly required to take a standard chemotherapy for 16 weeks, and then were screened for 4 different biomarkers (BRAF, PIK3CA, KRAS or NRAS, EGFR). The first appearance of any of the 4 biomarkers was used to group the patients into 4 cohorts, and the fifth cohort (unclassified) consisted of patients that were negative for any of the 4 markers. Every cohort was then randomized into either intervention or placebo group. For the intervention groups in each cohort, the corresponding treatments were specific BRAF-mutated kinase inhibitor in combination with panitumab and/or MEK inhibitor, dual PI3K/mTOR inhibitor mono-therapy, dual-pathway inhibition using AKT and MEK inhibitor, HER1, HER2, and HER3 inhibitors, and capecitabine, respectively. To evaluate the treatment effects of each intervention and

the suitability of the stratification strategy, not only data from the five different cohorts need to be analyzed separately, but also there is a need to pool all the data from this study to include all intervention and placebo groups. Similar to the FOCUS4 study, the LUNG-MAP study also employed an umbrella trial approach with four biomarkers (PI3K, CDK4/6, FGFR, and HGF) [5].

Another example is the STarT Back trial-a randomized controlled trial-with the goal of testing whether using a novel risk assessment tool combined with targeted intervention in subgroups is better than the best current care at reducing long-term disability from low back pain [6]. In this trial, there were 851 patients in total, one third were in control group, while the other two thirds were assigned to the intervention. Then the patients were classified by the STarT Back Screening Tool into three risk groups: low risk (26%), medium risk (46%), and high risk (28%). The primary outcome showed significant improvement in the intervention group both at 4 months and 12 months follow-up. A secondary outcome measured both physical and emotional function. Another example is the PATHOS trial, a phase II/III trial of risk-stratified, reduced intensity adjuvant treatment in patients undergoing transoral surgery for human papillomavirus (HPV) positive oropharyngeal cancer.

The evaluation of new interventions can usually be performed in two ways. One is to compare all subjects with the new treatment to subjects in the control group, and the other is to compare each sub-intervention group to the correspondent control group. Considering these problems, this paper establishes a treatment effect in a subset of the joint subgroups. This means we don’t consider K subgroups separately; instead, there are

all 2^K-1 possible combinations subgroups to be considered. The subset selection criterion is to find out the minimal p-value when testing all possible subset.

2. Methods

2.1 Notation

There are several notations mentioned in this paper.

N denotes the total sample size for a study;

K: we assume the patients can be divided into K subgroups;

g is a union of K, g = 1,…,K, each sample size is ng and its proportions is πg = ng / N.

S denotes a subset of $\mathcal{K}$ = {1,…,K}, indicating the union of the subgroups. The corresponding sample size is $n_S = \Sigma_{g\in S} n_g$, so $\pi_S = n_S / N = \Sigma_{g\in S} \pi_g$;

θ_g denotes the treatment effect in subgroup g;

$\theta_S = \Sigma_{g\in S} \tilde{\pi}_g^S \theta_g$ denotes the treatment effect in subset S, where $\tilde{\pi}_g^S = \pi_g / \pi_S$;

$\mathcal{P}_K$ denote the set of all non-empty subsets of $\mathcal{K}$;

S* is the selected subset of $\mathcal{K}$;

$\mathcal{L}$ is any set of subsets of $\mathcal{K}$, which is $\mathcal{L} \subseteq \mathcal{P}_K$.

2.2 Analytical framework

This paper mainly uses the model $Y_{gi} = \mu_g + \theta_g T_{gi} + \varepsilon_{gi}$, where Y_{gi} denotes the outcome of patient i in subgroup g and T_{gi} denotes the indicator with 0 indicating the standard

treatment and 1 the new stratified strategy. In addition, ε_{gi} are iid with mean 0 and variance σ^2.

The null hypothesis is shown as follows for any $S \in \mathcal{P}K$.

$H_0^S : \theta_S \leqslant 0$ versus $H_1^S : \theta_S > 0$

The stated linear model with the null hypothesis indicates that it is a one-sided test at a significant level α.

2.3 Five approaches for subset selection

This paper considers five different methods to select a subset S* which indicates the selected subgroup that has a positive treatment effect. Details of all five approaches are given in table I.

Table I. Overview of the five subset selection approaches.

Set of subsets	Selected subset S^*
$M1$ $\mathscr{L}1 = \{\mathcal{K}\}$	$\mathcal{K}$ if $p_{\mathcal{K}} \leqslant \alpha$; $\emptyset$ otherwise
$M2$ $\mathscr{L}2 = \{\{1\}, \{2\},\ldots, \{K\}\}$	$U_{S\in\mathscr{L}2, p_S \leqslant \alpha} S$
$M3$ $\mathscr{L}3 = \{\{1\}, \{2\},\ldots, \{K\},\mathcal{K}\}$	$U_{S\in\mathscr{L}3, p_S \leqslant \alpha} S$
$M4$ $\mathscr{L}4 = \mathcal{P}K$	$argmin_{S\in\mathscr{L}4}\, p_S$ if $min\{p_S \mid S \in \mathscr{L}4\} \leqslant \alpha$; $\emptyset$ otherwise
$M5$ $\mathscr{L}5 = \{S \in \mathcal{P}K \mid \pi S \geqslant \gamma\}$	$argmin_{S\in\mathscr{L}5}\, p_S$ if $min\{p_S \mid S \in \mathscr{L}5\} \leqslant \alpha$; $\emptyset$ otherwise

The first approach M1 is the overall test. The subset is the overall population, denoted as $\mathscr{L}1 = \{\mathcal{K}\}$, with the overall effect $\theta_{\text{overall}} = \theta_{\mathcal{K}}$. If the null hypothesis$H_0^{\mathcal{K}}$: $\theta_{\text{overall}} \leqslant 0$ is rejected, we select all patients, and otherwise, no patient is selected.

The second approach M2 is a subgroup analysis. The subset is each single subgroup where $\mathscr{L}2 = \{\{1\}, \{2\},\ldots, \{K\}\}$. In this method, we perform K tests on the null hypotheses$H_0^{\{g\}}$: $\theta g \leqslant 0$ for g = 1,…, K and select all the subgroups with significant treatment effects.

The third approach M3 is a combination of M1 and M2.We perform tests for the overall population and for each subgroup, that is, $\mathscr{L}3 = \{\{1\}, \{2\},\ldots, \{K\},\mathcal{K}\}$. The difference is that we perform M1 first, if the $H_0^{\mathcal{K}}$ is rejected, all patients are selected; otherwise, we perform M2 method, and select all significant subgroups.

The fourth method M4 is the subset analysis. This indicates that we have $2^K - 1$ combinations of subset, that is $\mathscr{L}4 = \mathcal{P}_K$. with the idea that to select the subset with the smallest p-value. The reason why we do not consider all of the significant subsets is that some negative treatment effect estimates could be selected. While this situation cannot happen to the subset with minimal p-values.

The last approach M5 is similar to M4, but the sample size of the subsets should meet a requirement as $\mathscr{L}5 = \{S \in \mathcal{P}_K | \pi_S \geqslant \gamma\}$, which means the sample size of subset must be greater than a fraction pre-specified γ. The paper considers γ = 50%. From the explanation above, it indicates that M5 drops the subgroups with small sample size, which will increases the power compared to M4.

This paper mainly focuses on one-sided tests to avoid undesirable results compared to the current clinical trial studies that use two-sided tests. However, this may overlook negative effects in some subgroups. To address this, their corresponding confidence intervals are recommended to be inspected.

2.4 Quality and performance measures

For the quality control, this paper uses the impact and inferior rate to estimate how good the selected subset S* is. The impact is stated as follows:

$$I(S^*) = \sum_{g \in S^*} \pi g \theta g = \pi_{S^*} \theta_{S^*}$$

which equals to the proportion of the selected subset times its treatment effect.

The other criteria inferior rate has the following expression:

$$R(S^*) = \sum_{g \in S^*} \pi g \, \Phi\left(-\frac{\theta_g}{\sigma_I}\right)$$

Where $\Phi\left(-\frac{\theta_g}{\sigma_I}\right)$ is the proportion of patients who cannot benefit from the new treatment, instead got negative treatment effect in subgroup g. σ_I is the individual effects deviation if we assume they are normally distributed. Thus the optimum outcome is to have higher impact and lower inferior rate.

3. Illustrative example

To better understand the five approaches, the paper provides an example with six subgroups corresponding to a new treatment for each subgroup. Patients were randomized to either treatment or control within each subgroup. The six subgroups have different sample size, they are 145, 167, 63, 292, 86, and 153. In all subgroups, outcomes Y were measured; the treatment effect, which is differences of the mean values between

the treatment and control groups in each subgroup, their corresponding 95% confidence intervals are shown in Figure 1.

The process for implementing the five approaches is, first we fit the model $Y_{gi} = \mu_g + \theta_g T_{gi} + \varepsilon_{gi,}$ construct contrast matrices within each approach and apply the ghlt procedure.

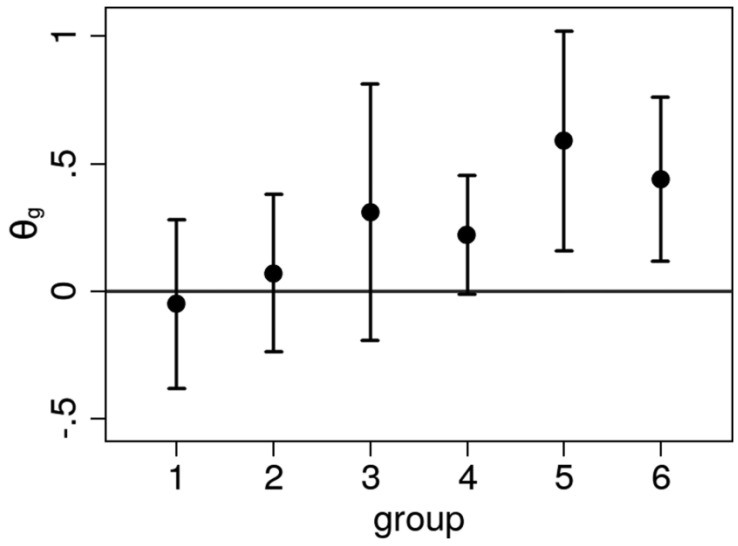

Figure 1. Subgroup-specific treatment effect estimates with pointwise 95% confidence intervals in a hypothetical study with six subgroups. Estimates and confidence intervals are based on the model described in Section 2.2.

The significant level is $\alpha = 0.025$. For the first approach M1, the p-value is 0.0007, so the all of the subjects are selected. For M2-the Subgroup analysis, the results show that only the subgroups 5 and 6 with rather large effects and adjusted p-values of p = 0.011 and p = 0.014, 5 and 6 subgroups are selected. M3 selects all of the subjects since the adjusted p-value of the overall analysis is 0.0046. For M4, a minimal adjusted p-value of 0.0016 was found for the subset {3, 4, 5, 6}, which is below 0.025, therefore, this subset was selected. As for M5, based on the result of M4, subset {3, 4, 5, 6} contains more than 50% of the

whole population, so it is also selected. The following table shows the result of which subgroups are selected according to five different approaches.

Table 2. The subsets S∗ selected by the five approaches in the example of a hypothetical study with results as shown in Figure 1.

	Group					
Methods	1	2	3	4	5	6
M1	×	×	×	×	×	×
M2					×	×
M3	×	×	×	×	×	×
M4			×	×	×	×
M5			×	×	×	×

The subgroups included in S* are marked by a cross.

4. Simulation study

4.1 Design of simulation study

For the simulation study part, my implementation is not exactly the same as the original paper. The paper has many different settings regarding to different parameters, I choose some of the settings and implemented the simulation study due to limited time. But the basic ideas are the same. I will introduce the relative parameters first.

The basic setting is the same; let the overall treatment effect be 0.2, that is $\theta_A = \theta_{overall} = 0.2$. P is the fraction that splits the subgroups into two parts, one part with effects $\theta_1,\ldots,\theta_K$ are set to 0, the other subgroups are selected so that the effect equals to θ_A. Another parameter is τ, which equals to the one half of the difference between the maximum and

the minimum divided by the average of the maximum and minimum within the non-zero effect subgroup. For the subgroup sample size, two scenarios are considered, one is equal, and the other is not.

In my simulation, I set the following parameters: assume the true value of the overall treatment effect is known, when K = 2, $\theta_A = \theta_{overall} = 0.2$. When K = 3, $\theta_1 = 0.1$, $\theta_2 = 0.2$, $\theta_3 = 0.3$(θ_1, θ_2 and θ_3 are the mean treatment effect for each subgroup.) About the total sample size, unlike the paper that sets N = 1056, I simulate the sample size with 300, 600 and 1200 with equal allocation for each method. The significant level $\alpha = 0.05$, the other two parameters P = 0 and $\tau = 0.5$.

The K values that I set are both relatively small, so it's not really useful to implement Method 5 with the limitation that each subgroup covers at least 50% of the total sample size. Therefore, I simulate four methods altogether.

Another problem which needs to be addressed is that with these constrictions for subgroup determination, there may exist selection bias problem, which implies that other factors such as age, gender, sites, disease history or health condition and other situations of the subjects were not taken into consideration. Therefore, I added a parameter X to represents other situation.

4.2 Results

The following figure 2 and figure 3 are the results of this paper, summarized by using the performance measures impact and inferiority rate. Figure 2 compares success rate changes with regards to different methods. Figure 3 shows how impact and inferiority rates are related to each other in different settings when comparing the five methods.

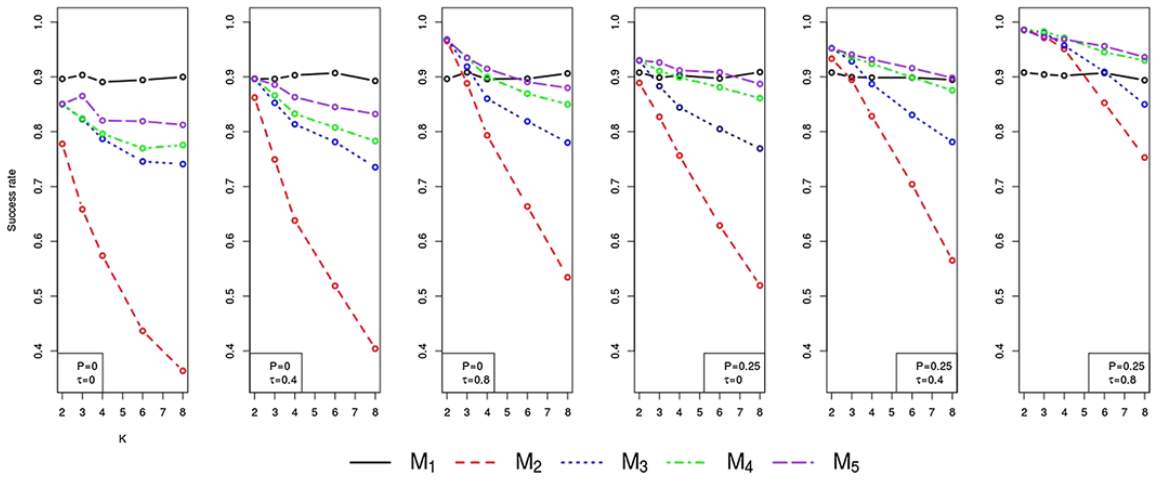

Figure2. Success rate depending on number of subgroups K for different choices of P and τ for equal subgroup sizes.

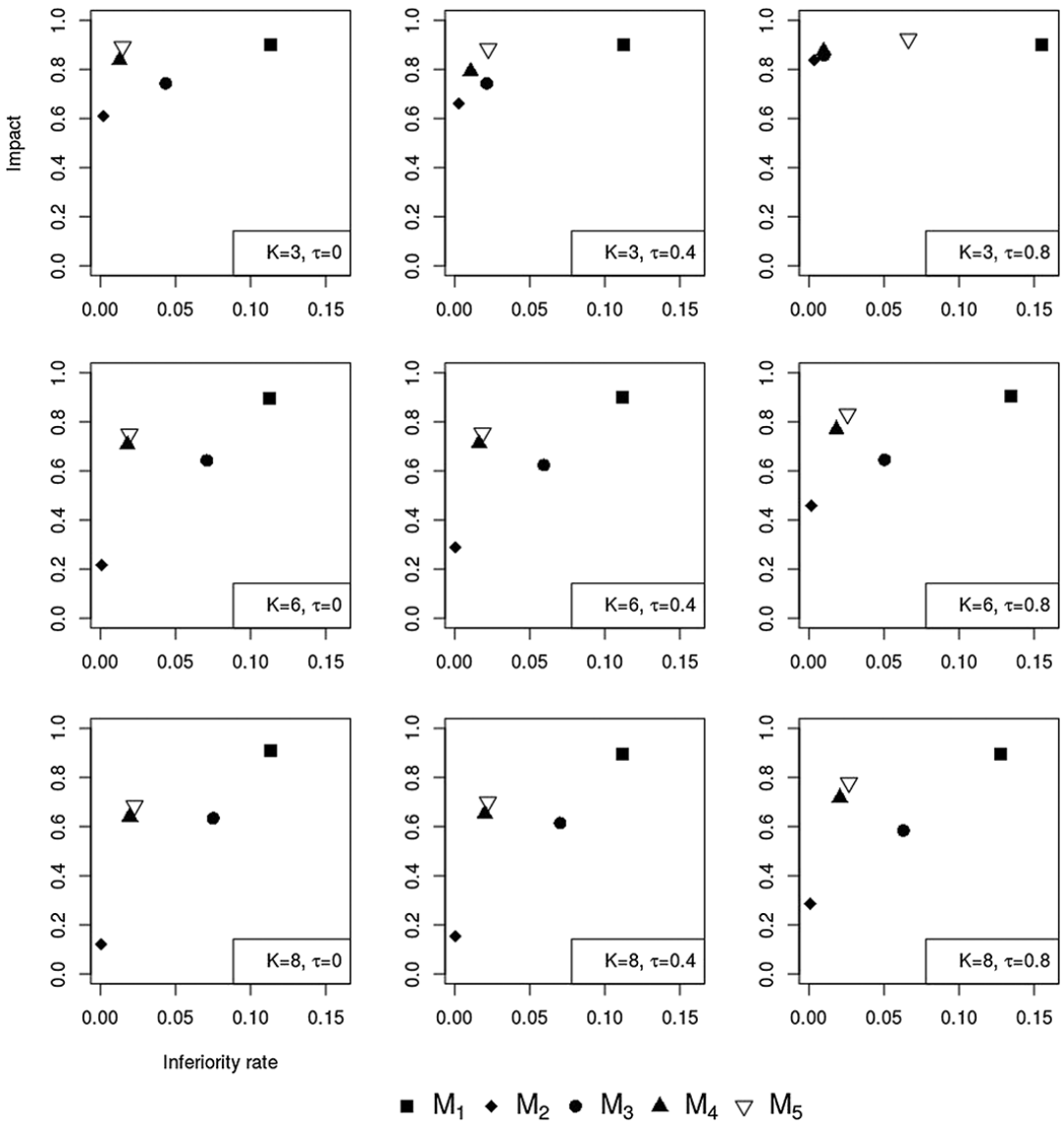

Figure3. Impact versus inferiority rate for P = 0.25 and different choices of τ = 0, 0.4, 0.8 and K = 3, 6,8 for equal subgroup sizes. The five methods are distinguished by different symbols.

Figure 2 shows how success rates changes as the number of subgroups increases under different settings. M1 shows relatively stable success rates, because it analyzes all of the subjects no matter how many subgroups are there. For other four approaches, it's

observed that success rates were decreasing as K increases. For the other four methods, the order of success rate is M2 < M3 < M4 < M5.

Figure 3 indicates how impact and inferiority rates are related under different settings when comparing the five methods. It shows that M1 always has the highest impact, but in the meantime it also has larger inferiority rates. M2 has the lowest impact and least inferiority. M3 is in the middle of M1 and M2. M4 and M5 always appear around the left top corner – the optimal area with high impact and low inferiority rate. M4 and M5 don't differ too much.

While after investigation of this study, we found there exists selection bias by using the subset selection, besides, the paper didn't look into power, so we conducted extended study. Figure 4 shows the results of power performance with respect to sample size without other covariates considered when K = 2 and 3, P = 0 and $\tau = 0.5$. Figure 5 has the same setting with Figure 4, but with the difference of with other covariates X taken into account.

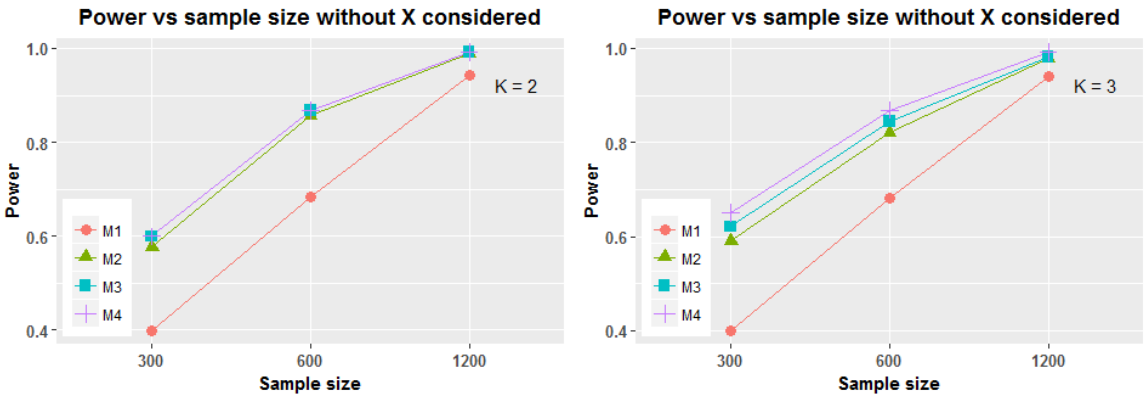

Figure4. Power performance with respect to sample size without other covariates considered.

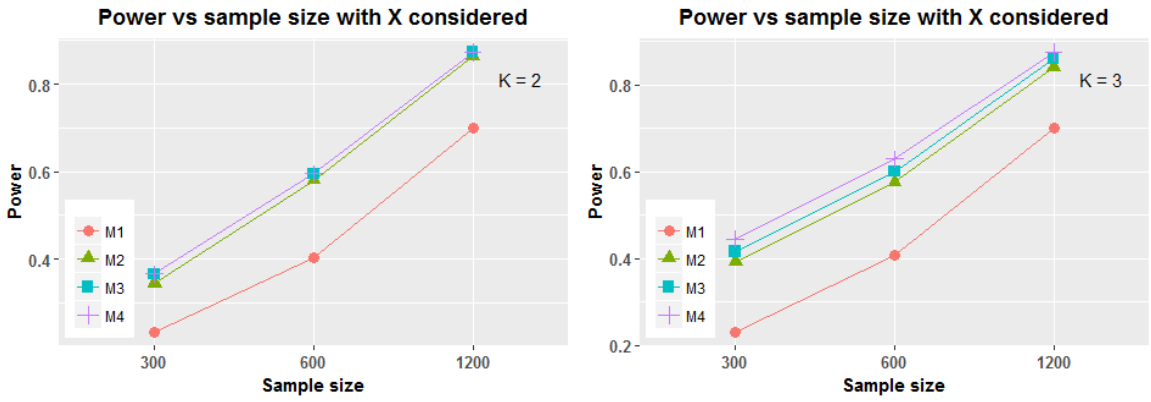

Figure5. Power performance with respect to sample size with other covariates considered.

From the plots above, we can see the obvious trend that as the sample size increases, power increases dramatically from average 0.3 to more than 0.99.

For the power performance, the power seems do not change too much with the change of K values. In Figure 4, power can be as low as 0.4 when the sample size is 300, but can reach 0.99 when sample size increases to 1200. But in Figure 5, when add other covariates considered, even the sample size is high, the power is still low, at most 0.87.

In general, M1 has relatively lower power compared to other three methods. For the power performance, with the sample size increasing, the power increases dramatically. Overall, the methods are ordered as M1 < M2 < M3 < M4 with respect to power.

For more detailed information, the four tables- 3, 4, 5 and 6 are present as follows.

Table 3 K = 2 without other covariates considered

	Methods	Sample size	Power
K=2	M1	300	0.399
		600	0.683
		1200	0.943
	M2	300	0.577
		600	0.857
		1200	0.989
	M3	300	0.6
		600	0.868
		1200	0.992
	M4	300	0.600
		600	0.868
		1200	0.992

Table 4. K = 3 without other covariates considered

	Methods	Sample size	Power
K =3	M1	300	0.400
		600	0.682
		1200	0.941
	M2	300	0.590
		600	0.821
		1200	0.979
	M3	300	0.621
		600	0.844
		1200	0.982
	M4	300	0.651
		600	0.869
		1200	0.992

Table 5. K = 2 with other covariates X considered

	Methods	Sample size	Power
K=2	M1	300	0.233
		600	0.402
		1200	0.701
	M2	300	0.345
		600	0.58
		1200	0.865
	M3	300	0.366
		600	0.595
		1200	0.873
	M4	300	0.366
		600	0.595
		1200	0.873

Table 6. K = 3 with other covariates X considered

	Methods	Sample size	Power
K =3	M1	300	0.23
		600	0.408
		1200	0.7
	M2	300	0.391
		600	0.576
		1200	0.841
	M3	300	0.415
		600	0.600
		1200	0.86
	M4	300	0.445
		600	0.629
		1200	0.875

5. Discussion

Because of the availability of testing multiple markers from subjects, stratified treatment methods will be wildly developed. The umbrella trials and adaptive methods are popular for addressing the above problem, but they have their drawbacks. Consequently, new analyses are needed. From section 4.2, it indicates that when sample size is small, the subgroup analyses are not so attractive based on power performance unless we recruit a large number of subjects. On the contrary, M1-the overall comparison is always valid based on the results above, but it overlooked some subgroups that have no effect. That's one reason why this paper sets the parameter P-the fraction split effect and non-effect subgroups.

To deal with the discrepancy, the paper proposed subset analysis that can in some way combine the overall analysis and subgroup analysis. In agreement with that idea, our results showed that the power in M3 and M4 are higher than that of first two methods.

Another problem is the selection criteria for M4 and M5, in the simulation study the minimal p-value rule is applied. For further study, other methods can also be considered. But the minimal p-value has the advantage that it excludes the subgroups with negative treatment effect. Or the union for significant subgroups can be selected together. Furthermore, other clinical knowledge can be taken into consideration as well. For example, similar markers that have similar effect could be combined.

As shown in our simulation study, the power decreases significantly when other factors exist in addition to treatment effects. For M4 and M5, type 1 error was controlled from

the proof of the paper-the proof of FWER control for M4 and M5. But further studies on similar situation with new methods are still necessary.

Reference

[1] Temple RJ. Special study designs: early escape, enrichment, studies in non-responders. Communications in Statistics-Theory and Methods 1994; 23(2):499–531.

[2] Freidlin B, McShane LM, Korn EL. Phase III clinical trials that integrate treatment and biomarker evaluation. Journal of Clinical Oncology 2013; 31:3158–3161.

[3] Kaplan R, Maughan T, Crook A, Fisher D,Wilson R, Brown L, Parmar M. Evaluating many treatments and biomarkers in oncology: a new design. Journal of Clinical Oncology 2013; 31:4562–4568.

[4]Menis J, Hasan B, Besse B. New clinical research strategies in thoracic oncology: clinical trial design, adaptive, basket and umbrella trials, new end-points and new evaluations of response. European Respiratory Review 2014; 23:367–378.

[5] Steuer CE, Papadimitrakopoulou V, Herbst RS, Redman MW, Hirsch FR, Mack PC, Ramalingam SS, Gandara DR.Innovative clinical trials: the LUNG-MAP study. Clinical Pharmacology & Therapeutics 2015; 97(5):488–491.

[6] Hay EM, Dunn KM, Hill JC, Lewis M, Mason EE, Konstantinou K, Sowden G, Somerville S, Vohora K, Whitehurst D, Main CJ. A randomised clinical trial of subgrouping and targeted treatment for low back pain compared with best current care. The STarT Back trial study protocol. BioMed Central 2008; 9:58. DOI: 10.1186/1471-2474-9-58.