MULTI-STAGE ADAPTIVE ENRICHMENT TRIALS

Neha Joshi

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2018

Approved by:

Anastasia Ivanova

Chirayath Suchindran

David Couper

Jason Fine

John Baron

ABSTRACT

Neha Joshi: Multi-Stage Adaptive Enrichment Trials
(Under the direction of Anastasia Ivanova)

We first consider the problem of estimating a biomarker-based subgroup and testing for treatment effect in the overall population and in the subgroup after the trial. We define the best subgroup as the subgroup that maximizes the power for comparing the experimental treatment with the control. In the case of continuous outcome and a single biomarker, both a non-parametric method of estimating the subgroup and a method based on fitting a linear model with treatment by biomarker interaction to the data perform well. Several procedures for testing for treatment effect in all and in the subgroup are discussed. Cross-validation with two cohorts is used to estimate the biomarker cut-off to determine the best subgroup and to test for treatment effect. An approach that combines the tests in all patients and in the subgroup using Hochberg's method is recommended. This test performs well in the case when there is a subgroup with sizable treatment effect and in the case when the treatment is beneficial to everyone.

We also consider the problem of estimating the best subgroup and testing for treatment effect prospectively in a clinical trial. We define the best subgroup as the subgroup that maximizes a utility function that reflects the trade-off between the subgroup size and the treatment effect. For subgroup estimation in trials with moderate effects sizes and sample sizes, simpler methods, such as linear regression, work better than more complex tree-based approaches. We propose a three-stage enrichment design, where the subgroup is estimated at the first interim analysis and then

refined in the second interim analysis, along with a futility analysis. A weighted inverse normal combination test is used to test the hypothesis of no treatment effect across the three stages.

Additionally, we consider a problem of subgroup estimation based on a multivariate outcome in both parallel group and crossover trials. We compare three methods of defining and estimating the best subgroup: a method based on the average and the maximum of the outcomes and the method based on the p-value for the treatment comparison.

ACKNOWLEDGEMENTS

I would like to begin by thanking my advisor Dr Anastasia Ivanova for her unwavering patience, guidance and support throughout - I could not have imagined having a better mentor. I would also like to thank my committee members - Dr Jason Fine, Dr David Couper, Dr Chirayath Suchindran and Dr John Baron for their invaluable and constructive feedback; Rong Chu from Amgen for providing the dataset used in Chapter 2 and anonymous reviewers for their helpful comments on Chapter 2 in this dissertation. I am indebted to Dr David Couper and Dr Kelly Evenson for allowing me to work on their projects as a GRA and their excellent supervision.

My sincere thanks are due to friends, acquaintances, old and new, who have made this process a little less painful. In particular, I am grateful to Jahnavi Punekar, Mandakini Singh, and Bithika Jain for half-a-lifetime of friendship and to Poulami Maitra for being the best classmate one could ask for. A million thanks to my parents, Ashok and Neelam Joshi and my brother, Aditya Joshi, for their constant encouragement. Gratitude is also due to Sanjay Purohit for his conversations and to everyone else who sent good wishes from miles away.

Lastly, my rock of Gibraltar, Raul Vyas, thank you for everything.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES xi

CHAPTER 1: LITERATURE REVIEW 1

1.1 Subgroup Estimation 1

1.2 Adaptive Designs 6

CHAPTER 2: POST-HOC SUBGROUP ESTIMATION AND TESTING 14

2.1 Introduction 14

2.2 Defining and estimating the subgroup 16

2.2.1 Defining the subgroup based on utility 16

2.2.2 Estimating the subgroup 17

2.3. Testing for the treatment effect 21

2.4. Simulation study 23

2.5. Example 25

2.6. Conclusions 26

CHAPTER 3: MULTI-STAGE ADAPTIVE ENRICHMENT DESIGN 28

3.1 Introduction 28

3.2 Subgroup Estimation Methods 30

3.3 Adaptive design with enrichment 35

3.4. Simulation Study 38

3.4.1 Comparison of subgroup estimation methods 38

3.4.2 Comparison of designs 41

3.5. Discussion 43

CHAPTER 4: FINDING A SUBGROUP WITH DIFFERENTIAL TREATMENT EFFECT WITH MUTIVARIATE OUTCOME 45

4.1 Introduction 45

4.2 Subgroup Estimation for Multiple Outcomes in a parallel group trial 47

4.2.1 Setup 47

4.2.2 Methodology 47

4.3 Subgroup Estimation with Multiple Outcomes in a crossover trial 50

4.4 Simulations 51

4.5 Discussion 54

CHAPTER 5: FUTURE RESEARCH 55

5.1 Limitations 55

5.2 Future Work 56

APPENDIX: FIGURES AND TABLES 58

REFERENCES 86

LIST OF TABLES

Table 1: Effect size in all (ES_{all}), effect size in the best subgroup (ES_S) and the prevalence of the best subgroup, π^*, corresponding to U_1 and U_2 for Model 3, $E[Y] = X^aT$ 63

Table 2: Type I error rate where the best subgroup is estimated by maximizing utilities U_1 and U_2 with estimation by the non-parametric (NP) method and parametric (P) method based on linear model with interaction. The type I error rate is evaluated for tests Z_{All} , $Z_{All,S}$, $\tilde{Z}_S$, Z_S, and for the Hochberg (HC) procedure applied to Z_{All} and Z_S. $\tilde{Z}_S$ is a naïve test for type I error rate and Z_S is a permutation test in subgroup. The total sample size in the trial is 500. 64

Table 3: Change-point model with parameters δ, θ and π_0. Best subgroup is estimated by maximizing utilities U_1 and U_2 with estimation by the non-parametric (NP) method and parametric (P) method based on linear model with interaction. Column π^* shows the median, 25% and 75% for the prevalence of the estimated subgroup. Power is for tests Z_{All} , $Z_{All,S}$, Z_S, which is a permutation-based test of the treatment effect in the subgroup, and for the Hochberg (HC) procedure applied to Z_{All} and Z_S. The best power for each test $Z_{All,S}$, Z_S, and HC in each scenario is in bold. 65

Table 4: Bivariate normal model with parameters δ, ρ_T , ρ_C , σ^2_T . Best subgroup is estimated by maximizing utilities U_1 and U_2 with estimation by the non-parametric (NP) method and parametric (P) method based on linear model with interaction. Column π^* shows the median, 25% and 75% for the prevalence of the estimated subgroup. Power is for tests Z_{All} , $Z_{All,S}$, Z_S, which is a permutation based test of the treatment effect in the subgroup, and for the Hochberg (HC) procedure applied to Z_{All} and Z_S. The best power for each test $Z_{All,S}$, Z_S, and HC in each scenario is in bold. 67

Table 5: A linear model with interaction $E[Y] = X^a T$, total sample size of N. Best subgroup is estimated by maximizing utilities U_1 and U_2 with estimation by the non-parametric (NP) method and parametric (P) method based on linear model with interaction. Column π^* shows the median, 25% and 75% for the prevalence of the estimated subgroup. Power is for tests Z_{All}, $Z_{All,S}$, Z_S, which is a permutation based test of the treatment effect in the subgroup, and for the Hochberg (HC) procedure applied to Z_{All} and Z_S. The best power for each test $Z_{All,S}$, Z_S, and HC in each scenario is in bold. 69

Table 6: Data analysis of a phase 2 study of 1C4D4 in patients with metastatic pancreatic cancer. Best subgroup is selected based on utilities U_1 and U_2 with estimation by the non-parametric (NP) approach using the logrank test and parametric approach (P) by fitting a Cox model with interaction. The adjusted Hochberg p-value is to test the intersection hypothesis of no treatment effect in all and in the subgroup. 71

Table 7 : Proportion of times covariates were selected using the overlapping group LASSO method for Models 1-6 using 2-biomarkers. Predictive column is the proportion of trials where the right set of predictive biomarkers were selected. Subset is the proportion of trials when an exact subset of the right set of predictive biomarkers were selected. Noise is the proportion to trials when at least one noise biomarker was selected. No Biomarker is the proportion of trials where no biomarker (predictive or noise) was selected. 72

Table 8 : Results for comparing design U_2U_1 using LM and OGLASSO for models D1-D5. Column π shows the true subgroup prevalence in the first row, and the median, 25% and 75% for the prevalence of the estimated subgroup at the second interim analysis for each of the designs. Column $\%U$ shows the median, 25% and 75% of the percentage of the true utility estimated in the design at the second interim analysis. Power is for tests Z_{All}, based on all subjects enrolled and Z'. The power incorporating the single futility look at the second interim analysis is given by Z'_f. The proportion of trials stopped for futility is given by p_f. Total sample size used is n = 360. The best power and $\%U$ for the best method for each model is in bold. 74

Table 9 : Type I error rate for the enrichment design for tests Z' and Z'_f. The proportion of trials stopped for futility, p_f are also presented. Total sample size in the trial is n = 360. 76

Table 10 : Comparison for four methods: LM_{max}, LM_{avg}, hyperplane and hyperplane with penalty in terms of %U, and estimated prevalence $\hat{\pi}^*$ in a parallel trial. Median values are reported. $\%X$ denotes the proportion of trials for which the right set of biomarkers were used to draw the hyperplane. The %X values for scenarios where the best subgroup is based both biomarkers is in (). 77

Table 11 : Comparison for four methods: LM_{max}, LM_{avg}, hyperplane and hyperplane with penalty in terms of %U, and estimated prevalence $\hat{\pi}^*$ in a crossover trial. Median values are reported. $\%X$ denotes the proportion of trials for which the right set of biomarkers were used to draw the hyperplane. The %X values for scenarios where the best subgroup is based both biomarkers is in ( ). 79

Supplemental Table 1: Models used to generate data to compare the subgroup estimations methods. #Outcome represents the number of outcomes associated with the treatment and $\%U_{all}$ is the $\%U$ value including all subjects and. S^* is the true subgroup, δ^* is the true treatment difference and π^* is the underlying prevalence of the subgroup. U is the utility value in the model. 81

Supplemental Table 2: Comparison for three methods LM_{avg} , hyperplane and hyperplane with penalty in terms of %U, estimated prevalence $\hat{\pi}^*$ and mean treatment difference $\hat{\delta}^*$ in a parallel group trial. Median, 25th and 75th percentile reported. 82

Supplemental Table 3: Comparison for four methods: LM_{max}, LM_{avg}, the hyperplane method and the hyperplane with penalty in terms of %U, estimated prevalence $\hat{\pi}^*$ and mean treatment difference $\hat{\delta}^*$ in a crossover trial. Median, 25th and 75th percentile reported. 84

LIST OF FIGURES

Figure 1: Best Subgroup for models 1-6 58

Figure 2: Large sample subgroup estimation for models 1-6 by linear model method with two biomarkers for LM and OGLASSO 59

Figure 3 : Comparison of the Linear Model (LM), Overlapping Group LASSO (OGLASSO), Support Vector Machines (SVM), Classification and Regression Trees (CART) and Random Forests (RF) for subgroup estimation in a clinical trial with 400 patients using box plots for the distribution of %U. Horizontal line represents %U in all subjects. 61

Figure 4: Candidate lines for subgroup estimation using the hyperplane method. 62

CHAPTER 1: LITERATURE REVIEW

We consider the problem of estimating the subset of subjects that benefit the most from a treatment and testing for treatment effect in this subgroup and overall in a randomized clinical trial. This can be done as a post hoc analysis or prospectively with adaptive enrichment. We define the best subgroup as the subgroup that maximizes a utility function that reflects the trade-off between the subgroup size and the treatment effect. As a result, methods incorporating adaptively determining the subgroup are gaining traction. Subgroup identification and population enrichment can increase the odds of showing that a new therapy is more effective than a control.

1.1 Subgroup Estimation

In many published methods (Song and Chi, 2007; Alosh and Huque, 2009; Jenkins, Stone and Jennison, 2011) the subgroup of interest is pre-specified before the trial. The goal is to have an efficient procedure for testing the treatment effect that controls the overall type I error rate, when testing overall and in the subgroup. Some of the papers propose a method that tests a subgroup based on the significance of the overall test at a reduced level or use closed testing methods incorporating variable alpha allocation and correlation between all subjects and subgroup to achieve type I error control. A common design considered (Wang et al, 2007) enrolls subjects stratified on the subgroup status in stage 1 and performs an interim analysis to choose whether to enroll in all or in the subgroup in stage 2. A group sequential design was proposed by Magnusson and Turnbull (2013) with pre-defined subgroups. The selection of subgroups at the end of first stage is carried out in two ways: First, if the statistic for testing treatment effect is below a specified

threshold for any subgroup, it is dropped from the trial. Second, if the subgroups are ordered and a specific subgroup exceeds the threshold, all subgroups below it in hierarchy are dropped. The rest of the subgroups are pooled and treated as one population in the rest of the trial.

Lipkovich, Dmitrienko and D'Agostino (2017) extensively reviewed commonly used subgroup identification methods. Broadly speaking, they classified subgroup analysis into two categories based on the interaction between biomarkers and treatment effect. The first is called a qualitative interaction where the goal is to identify the right subject for a given treatment. It leads to the conclusion that the treatment is better than the control in a subgroup (identified based on biomarkers) and it is not better for those not in this subgroup (could be equivalent to control or worse). This is useful when the treatment is not effective overall but could help save the trial by determining a subgroup post-hoc where the treatment is effective. The second is a quantitative interaction where the goal is to identify the right treatment for the subject. Here, the treatment is better than the control within and outside a subgroup of subjects but with different benefits. This is useful when trying to find an optimal treatment regime for a given set of subjects. We are interested in the first classification. Methods that deal with it can be described as: modeling underlying outcome (numerous interactions terms between number of biomarkers and covariates are fit in a complex model - 'black box' - to determine treatment homogeneity based on potential outcomes), modeling underlying treatment effect (focuses on directly estimating the predictive/treatment effect or treatment contrast), and direct search for subgroups that benefit from the treatment (search for treatment-biomarker interactions and select specific regions with higher treatment effect). Methods for the second classification focus on predictive biomarkers with qualitative interactions including treatment assignment variable being considered as the outcome variable. For adaptive or post-hoc estimation of subgroups, we assume there are one or more

biomarkers available that can be used to identify a subset with greater effect in treatment vs. control.

In the literature reviewed, there are three approaches to subgroup estimation. The first, a classical approach to identifying subgroups, uses the interaction terms between treatment and biomarkers, as described briefly in Kehl and Ulm (2006). This can be implemented when there is no significant difference between the treatments in a trial including all subjects. If the interaction term is significant and the coefficient is such that a predictive factor results in higher response in the treated group, then subjects with this factor are positive responders to the experimental treatment. If the coefficient shows that the predictive factor lowers the response in the treated group as compared to control, then subject with this factor are negative responders to the experimental treatment. Even though this is a simple method to implement, it has its disadvantages. Factors or mixture of factors have to be included in the interaction term to be considered predictive and necessary for defining subgroup; if the number of covariates is large, we could we potentially be looking at a lot on multi-way interaction terms and thus require larger sample size to detect all of them.

The second approach involves choosing a subgroup based on a minimum treatment effect (Freidlin and Simon, 2005; Zhang 2018). We can either use pre-specified values or compute using the data available (Freidlin and Simon, (2005, 2010)). A subgroup defined this way can be claimed to be the one with the maximum treatment effect and can be easily interpreted. This subgroup does not take into account the subgroup size and thus the estimated subgroup could have a rather small prevalence.

The third approach requires optimizing a utility that balances the trade-off between treatment effect and subgroup size (Lai, Lavori and Liao, 2014; Zhang, 2017). This ensures that

the estimated subgroup is not either too small or too large in size. A disadvantage would be that since we do not control the magnitude of the treatment effect, we could potentially end up with a subgroup that has very small (~ 0) treatment effect.

For subgroup estimation, when dealing with multiple biomarkers, regression models may not be ideal to deal with higher order interactions or unknown forms of relationship between covariates and response. Loh, He and Man (2015) discussed tree-based regression that partition the biomarker space and thus can be used identify subgroups of subjects with higher treatment effect, which are essentially the terminal nodes of the trees. These subgroups identified are patients who are most likely to benefit from the treatment. These tree-based methods primarily use the principal of Classification and regression trees (CART) approach (Breiman et al., 1984) to develop and control the size of the decision tree. CART recursively splits the data into two disjoint parts or subsets by minimizing the heterogeneity of the outcome in each partition. The resulting model can be illustrated by a single decision tree and the terminal nodes of the tree can be interpreted as subgroups of heterogeneous outcome. CART tends to overfit and there is heavy dependence on the initial set used to train or develop the model. CART is also greedy and as a result is biased towards selecting variables that have more splits. To overcome the deficiencies of CART, Random Forests (RF) can be used (Breiman, 2001). RF is made up of several decision trees and results are averaged over this collection of trees. This aggregating of results leads to more reliable prediction using RF as compared to CART. At each node, a random subset of covariates is used for splitting. Unlike a single decision tree, random forests cannot be interpreted and are considered as a 'black box'. RFs also have the option of incorporating variable importance to avoid splitting on noise variables. A method called Interaction Trees (Su et al, 2009) builds a tree by using binary splits on all covariates and values assumed by the covariate (continuous variables are

converted to binary based on a pre-specified threshold). For each split, a regression model, with interaction between the treatment and the indicator variable associated with that split is fit and the split that minimizes the interaction p-value is chosen. Subgroup estimation is performed to determine how treatment effect varies across subgroups, after significance test using all is done. This method inherits the disadvantages associated with CART. The Virtual Twins (VT) (Foster, Taylor and Ruberg, 2011) method used RF to build a model to estimate treatment effect for each subject. A regression/classification tree is then built using these treatment effect and associated covariates associated to identify the subgroup. This method is developed for situations where the new treatment does not show improvement over control in all, but there are subgroups that show promise. Lipkovich et al. (2011) introduced a recursive partitioning method to identify subgroup(s) of interest, called SIDES (Subgroup Identification based on Differential Effect Search). The method considers multiple possible subgroups by identifying the five best splits at a node that optimize a pre-specified measure (e.g. split-by-treatment interaction). This procedure is then repeated for the child node. The method results in subgroups with large treatment effects and splits at each node on a variable not previously considered. A 2-stage extension of the SIDES method was developed in Lipkovich and Dmitrienko (2014). The first stage was used to reduce the number of biomarkers/covariates using a variable importance index and then the SIDES method was applied on the chosen few in the second stage. Another approach called Qualitative Interaction Tree (QUINT) by Dusseldorp and Van Mechelen (2013) splits each node to optimize a utility that allow simultaneous control of effect size and subgroup size. As in CART, QUINT finds the subgroups by searching over all covariates and all possible splits and thus suffers from selection bias.

Chen et al (2018) proposed another method for subgroup estimation for a continuous outcome when dealing with multiple biomarkers. This method aims to select a set of subjects that respond to treatment and treatment doesn't have the same effect on all subjects. Before identifying the subgroup, a set of predictive biomarkers are identified by fitting a linear model with interaction between the biomarker and treatment indicator for each available biomarker. Those with significant interaction terms are considered predictive. The subgroup identification method is divided into two steps: First, a scoring model is used to convert a subjects' biomarker values into a univariate score and second, a cutoff for the score is identified to divide subjects into biomarker positive and negative groups. Suppose there are r predictive biomarkers. A scoring model is fit for the i^{th} subject for every predictive biomarker as $y_i = \alpha + \beta t_i + \delta_j x_{ij} + \lambda_j t_i x_{ij}$ where x_{ij} are the observed values of the j^{th} predictive biomarker for the i^{th} subject. Let z_j be the standardized test statistic for the test of interaction for the j^{th} predictor. A composite score for each subject is defined as $s_i = \sum_{j=i}^{r} z_j x_{ij}$. A grid of candidate scores are considered and the best cutoff is chosen as the one that maximizes the log-likelihood of a changepoint model, based on this cutoff.

1.2 Adaptive Designs

Some papers outlined designs that include both the identification and validation of the subgroup in the same trial (Freidlin and Simon, 2005; Jiang, Freidlin and Simon, 2007; Freidlin, Jiang and Simon, 2010, Renfro et. al 2014, Zhang et. al, 2017,2018; Diao et. al, 2018).

In their seminal paper, Freidlin and Simon (2005) introduced *Adaptive Signature Designs* (*ASD*), an all-comers single-stage design analyzed in 2-stages, developed for a binary outcome. The method is applicable to a phase III randomized trial comparing the experimental treatment with a control. If there is no predictive biomarker already identified or a large number of available

biomarkers that could determine a subgroup, then this method will both identify and validate the predictive biomarkers while defining the subgroup. The aim was to identify a subgroup with significant treatment effect, which is then tested along with a test in all to perform an overall test of significance in the trial. A total of *n* subjects are enrolled with stage 1 data (n_1) used as the training set to develop a classifier, which is then applied to stage 2 data (n_2) to classify subjects as belonging to the subgroup or not. Step 1 fits a logistic model with interaction terms between all available biomarkers and treatment to identify a subset of biomarkers provided that are significant (interaction coefficient) at a predetermined level. In step 2, a subject in stage 2 is classified as belonging to the subgroup if the odds ratio for experimental treatment vs. control exceeds a pre-specified threshold for at least *G* of the biomarkers selected in step 1. The simulation studies used n=400, suggested $n_1=n_2=n/2$ as the size for each of the stages and assumed choosing 3, 10 or 20 biomarkers in step 1. The test for treatment effect are carried out in all *n* subjects at level=0.04 and in the subgroup at level = 0.01, in order to preserve overall $\alpha = 0.05$ using Bonferroni. If either of the test is significant, then the trial is considered successful. An extension of *ASD*, called Cross-Validated *ASD* (*CVASD*) was proposed by Freidlin, Jiang and Simon (2010). The trial population is split into *K* cohorts of equal size with proposed $K = 10$. At the k^{th} step, $k = 1\ldots, K$, cohort *k* is used as a validation cohort and the rest of the subjects are part of the development cohort. Since each subject appears exactly in one of the validation cohorts, at the end of the cross-validation procedure, each subject is classified as being in the subgroup or not. The subgroup development is implemented in the same way as in Freidlin and Simon (2005). The cross-validated design is more powerful since all subjects are being used to develop the classifier. But testing for treatment effect in the subgroup will suffer from re-substitution bias and as a result permutation p-values were used (Jiang, Freidlin and Simon, 2007) to control type-I error. Subgroup selection, estimating

and testing treatment effect in ASDs was further discussed by Zhang et. al (2017). The goal was to estimate *a* subgroup with positive treatment effect, which could also include everyone. Subgroup of interest is identified by maximizing a utility function, based on multiple baseline covariates, instead of using a minimum threshold for treatment effect (Freidlin and Simon, 2005). The utility is defined as the product of the prevalence of the subgroup and the power of the subgroup such that the best subgroup could also include everyone. As a result, the test of treatment effect is proposed to be carried out only in the subgroup. Since the same set of subjects is being used to estimate the subgroup and test for treatment effect, there is re-substitution A cross-validation method is used to estimate the treatment effect.

In Jiang, Freidlin and Simon (2007), the subgroup estimation method is developed by determining a threshold for a single continuous predictive biomarker, that results in strong evidence of treatment. A changepoint model is assumed to be true that is, it is assumed that there is no treatment effect below a cutoff. Two approaches for testing the treatment effect are considered. In the first, if a test of no treatment effect carried out in all at a reduced level α_1 is significant, the testing ends. Otherwise, the test is carried out in a subgroup identified at level α_2 where the overall type-I error $\alpha = \alpha_1 + \alpha_2$ is controlled. The test statistic in the first scenario is calculated as $T = \max_{0.5<c<1} L(c)$ where c is the candidate cutoff for the biomarker and L is the log-likelihood ratio statistic based on the changepoint model for that cutoff. In the second approach, the test statistic is $T = \max(L(0) + R, \max_{0<c<1}\{L(c)\}),$ where $L(0)$ is the log-likelihood ratio statistic for all and R is a pre-specified constant. To test for treatment effect, a permutation test statistic is used. The test-statistic of interest, *T,* is computed for the original dataset and then permuted datasets are created by randomly permuting the treatment labels *B* times and the entire procedure

of cross-validation is repeated to get the permuted test-statistic T^*. The p-value is computed as $\frac{1+\#T^* > T}{1+B}$. A threshold for the biomarker is estimated as the cutoff that maximizes the partial log-likelihood function based on the model only if the test for treatment effect in the subgroup is significant.

Adaptive enrichment trials adapt the entry criteria based on data observed in the trial so far to restrict enrollment to subjects in whom the experimental treatment is believed to work. Enriching the subject population in the trial can increase statistical power to detect a treatment effect if a new therapy only works in a subgroup of subjects. Additionally, it reduces the number of subjects in the trial who have no apparent benefit from the drug therefore not exposing them to potentially harmful side-effects. Simon and Simon (2013) proposed an adaptive enrichment method where the enrollment is modified at the interim to recruit only those in the subgroup. A testing methodology that controls type-I error is also discussed. A possible implementation as a 2-stage design, with a single continuous biomarker is described as follows. A changepoint model for the outcome based on the biomarker, X, is considered such that there is treatment effect only above the true cut-off, x^*. X is assumed to take values in the interval [0,1]. At interim, the loglikelihood of the data is computed for a grid of candidate cutoffs. The estimate for the cutoff, $\hat{x}^*$ is the one that maximizes the log-likelihood. If the corresponding log-likelihood, $L(\hat{x}^*) \geq L(1) + 0.25,$ then the enrollment at stage 2 is restricted to subjects with biomarkers values greater than $\hat{x}^*$, otherwise the trial is stopped. They however do not discuss how the treatment effect in the subgroup can be estimated. Simon and Simon (2017) covers the issues not touched upon earlier that is to determine the target population intending to use the drug and estimation of treatment effect in this subgroup. A group sequential method with multiple updates to the subgroup

criteria is implemented with the rule used for the last stage or one based on the entire trial data is considered as appropriate for future use. This was also explored in a 2-stage adaptive enrichment design by Zhang et al. (2018). The subgroup is estimated by applying a pre-specified function, Ψ (e.g. OR for experimental treatment vs. control being greater than a specified value), on stage 1 data (n_1 subjects) to yield a subgroup rule based on one or more biomarkers. The second stage enrolls n_2 subjects that satisfy this rule. The treatment effect is estimated by using a weighted inverse normal statistic from the both the stages. Since stage 1 data is used to determine the subgroup, there is re-substitution bias. Bootstrap based methods were found to be more precise and less biased than cross-validation used by Freidlin, Jiang and Simon, (2010) or a naïve approach of using treatment difference. A 2-stage adaptive design, with an interim analysis, was proposed recently by Kimani et al (2018) to estimate the threshold of a single continuous biomarker. Monotonicity is assumed such that a higher biomarker value leads to a larger (smaller) treatment effect. Stage 1 data recruits from the full population and is partitioned into multiple candidate subgroups based on several pre-defined cut-off values. If the treatment effect in all these candidate subgroups are all below a specified futility threshold, the trial stops for futility, else the candidate subgroup with the largest treatment effect is chosen for enrollment in stage 2. A number of estimators for the treatment effect in the selected subgroup are compared and a hybrid estimator, conditional on the subgroup is recommended.

Renfro et. al (2014) described a 2-stage adaptive design incorporating subgroup estimation, interim analysis with possible enrichment and testing for treatment effect at the end. A single continuous predictive biomarker is considered and higher values of the biomarker are assumed to correspond to better response. The biomarker is dichotomized based on a series of cut-points and for each of these cut-points, a regression model is fit with an interaction term between the

dichotomized biomarker and treatment. The best cutoff is the one with the smallest p-value for the interaction term which is below a certain pre-specified p-value. If the treatment effect is higher in the subjects with values above this cutoff (than those below), then the biomarker is considered promising. For a promising biomarker, at interim, futility is tested in each of the biomarker low and high groups separately. The trial can stop for futility in both groups, or (a) continue enrollment in stage 2 only in biomarker high group (if futile in biomarker low group) or (b) continue unrestricted enrollment if futile in neither. If there is no promising biomarker, the futility is tested in all and the trial stops for futility or (c) enrollment in stage 2 is unrestricted. The test for treatment effect at the end is either only in (a) stage 2 biomarker high subjects, (b) union of stage 1 and stage 2 biomarker high subjects or (c) in all. Sample sizes and pre-specified with caps on total and biomarker low enrollment. A recent 2-stage adaptive enrichment design was suggested by Diao et al. (2018) with two different methods for threshold estimation of a single biomarker based on the difference of treatment effects in the subgroup and outside the subgroup ($X > c$ or $X < c$). The paper compares three scenarios for testing for treatment effect – only the stage 2 enriched data, biomarker positive subgroup from both stages and all data in both stages and conclude that using the biomarker positive subgroup, though the most powerful, leads to biased results of the treatment effect. They also briefly discuss subgroup estimation in case of non-monotonic relationship between biomarkers and treatment effect.

Lai, Lavori and Liao (2014) proposed a three-stage group sequential design, with two planned interims, for an adaptive enrichment trial with a set of pre-defined subgroups based on two or more baseline biomarkers. Here, it is assumed that the treatment works only in a subgroup of patients, that is unknown at the outset. A utility function, equal to the Kullback–Leibler information number, is used to identify the best subgroup. Under the assumption of equal variances

of the treatment effect in all subgroups, maximizing this utility is the same as maximizing the square root of the prevalence of the subgroup multiplied by the treatment effect in the subgroup, and is the same as maximizing the power of the treatment comparison. This utility incorporates both the treatment effect and the subgroup size and is likely to estimate a more balanced subgroup such that we can avoid choosing a very small set with very high treatment effect. Recruitment is unrestricted in stage 1 and in case of evidence of efficacy at interim 1, the best subgroup is chosen as everyone and the trial is stopped. If there is no evidence of either efficacy or futility, the unrestricted enrollment continues in Stage 2. In case of futility in stage 1 population, the candidate subgroup, say *S*, that maximizes a specified utility is tested. If *S* exhibits futility as well, the trial is stopped no subgroup with enhanced treatment effect is found. In case of efficacy, the trial is stopped and *S* is concluded to be the best subgroup. But, if there is neither efficacy nor futility in *S*, enrollment is continued in it at the second stage. At the second interim, we again test for futility or efficacy in the chosen subgroup. If the subgroup enrolled in the second stage is everyone, we repeat earlier steps and either stop trial, continue unrestricted enrollment in everyone (in stage 3) or choose to continue enrolling in a subgroup S' that maximizes the utility. If the enrollment in stage 2 was continued in *S*, we either stop for efficacy, futility or continue enrollment in *S* (in stage 3) if neither is observed at interim 2. At the end, we stop for either futility of the trial and efficacy in the enrolled subgroup in stage 3.

Chapter 2 considers the problem of estimating a biomarker-based subgroup in a post-hoc analysis of a single stage clinical trial for the case of continuous outcome and a single biomarker. The best subgroup is defined as the subgroup that maximizes a utility function that reflects the trade-off between the subgroup size and the treatment effect (Lai, Lavori, Liao, 2014; Zhang et al, 2017). Several procedures for testing for treatment effect in all and in the subgroup are presented.

Both non-parametric and parametric methods are implemented to estimate the biomarker cut-off to determine the best subgroup. Cross-validation with two cohorts is used to improve the estimation of the subgroup and to test for treatment effect (Freidlin, Jiang and Simon, 2010). A permutation test (Jiang, Freidlin and Simon, 2007) is considered to control re-substitution bias. Simulation show that an approach that combines the tests in all patients and in the subgroup using Hochberg's method is recommended. In Chapter 3, multiple methods for subgroup estimation are discussed for 2 or more biomarkers. We propose a three-stage enrichment design (Lai, Lavori, Liao, 2014) where the subgroup is estimated at the first interim analysis and refined in the second interim analysis, for the case of a continuous outcome and multiple continuous biomarkers. A weighted inverse normal combination test is used to test the hypothesis of no treatment effect across the three stages. Simulations compare different designs using the best method for subgroup estimation. Chapter 4 considers subgroup estimation is based on a multivariate outcome. We compared three methods of defining and estimating the best subgroup: a method based on the average and the maximum of the outcomes and the method based on the p-value for the treatment comparison. These methods are compared in the setting of both parallel and crossover trials. Chapter 5 discusses the limitations of methods discussed in Chapters 2-4 and outlines future direction of the work.

CHAPTER 2: POST-HOC SUBGROUP ESTIMATION AND TESTING

2.1 Introduction

Subgroup identification and population enrichment can increase the odds of showing that a new therapy is more effective than a control. In many published methods (Song and Chi, 2007; Alosh and Huque, 2009; Jenkins, Stone and Jennison, 2011) the subgroup of interest is already known before the trial. At an interim analysis, the decision to continue enrolling the same patient population or restricting enrollment to a subgroup is made. The goal is to have an efficient procedure for testing if the treatment effect is zero that controls the overall type I error rate. Other published methods are focused on estimating the subgroup with a high treatment effect. In the majority of the methods for identifying a subgroup based on multiple biomarkers (Kelh and Ulm, 2006; Renfro et al., 2014) the biomarker cut-off is selected based on the interaction in a linear model between treatment and the biomarker. Recursive partitioning tree methods consider treatment-biomarker interaction when splitting the population of subjects (Su et al., 2009; Lipkovich et al., 2011; Foster, Taylor and Ruberg, 2011).

Several authors proposed identifying and validating the subgroup within the same clinical trial (Freidlin and Simon, 2005; Jiang, Freidlin and Simon, 2007; Freidlin, Jiang and Simon, 2010). In these methods, treatment effect is often tested in all patients as well as in the subgroup. In the adaptive signature design (Freidlin and Simon, 2005), the subgroup is identified using the first stage data while the treatment effect is tested in the subgroup using the second stage data. The treatment effect is also tested in all patients based on data from both stages. A biomarker is

considered promising if the interaction with treatment is significant at a threshold based on stage 1. Patients are included in the subgroup if the predicted treatment effect, by a model that includes all promising biomarkers, is higher than a certain value. Freidlin, Jiang and Simon (2010) extended the method in Freidlin and Simon (2005) by using cross validation for estimating the subgroup and testing the treatment effect. The trial population is split into K cohorts of equal size with $K = 10$. At the kth step, $k = 1,\ldots,K$, cohort k is removed from the data set and the subgroup is estimated from the remaining data, called the development cohort. Then patients in cohort k, that serves as a validation cohort in the kth step, are classified as being in the subgroup or not using the results of the estimation in the development cohort. Since each subject appears exactly in one of the validation cohorts, at the end of the cross-validation procedure, each subject is classified as being in the subgroup or not. The subgroup development is implemented in the same way as in Freidlin and Simon (2005). As in Jiang, Freidlin and Simon (2007), a permutation p-value is computed to test the treatment effect in the subgroup.

Simon and Simon (2013) described a trial with adaptive enrichment with a single continuous biomarker and a binary outcome. The best subgroup is defined as a subgroup where the treatment effect is larger than a given value. Renfro et al. (2014) selected the biomarker cut-off as the one with the smallest p-value for the interaction term between a single biomarker and treatment. Diao et al. (2019) defined the subgroup based on the difference of treatment effects in the subgroup and outside the subgroup. Lai, Lavori and Liao (2014) defined the best cut-off for a single biomarker based subgroup as the one that maximizes a utility function. They proposed a utility function equal to the Kullback–Leibler information number. Under the assumption of equal variances of the treatment effect in all subgroups, maximizing this utility is the same as maximizing the squared root of the prevalence of the subgroup multiplied by the treatment effect in the

subgroup, and is the same as maximizing the power of the treatment comparison. Zhang et al. (2017) proposed a utility function where the power is additionally multiplied by the prevalence of the subgroup. Defining the best subgroup based on a utility function allows for a trade-off between the size of the subgroup and the treatment effect in the subgroup. For example, if a single biomarker and the treatment effect follow a change-point model, selecting the subgroup with the higher treatment effect, as considered in Jiang, Freidlin and Simon (2007), might not be the best choice. When the difference between the treatment effects below and above the change point is small, the whole population should be considered and not only the subgroup above the change point. The trade-off between the treatment effect and the subgroup size should be taken into account when selecting the best subgroup.

2.2 Defining and estimating the subgroup

2.2.1 Defining the subgroup based on utility

Consider the case of a single continuous biomarker X, where a subgroup is defined as subjects with $X > c$, where c is a biomarker cut-off. Let Y be the response to treatment. Patients are randomized between treatment ($T = 1$) and control ($T = 0$), where T is the treatment indicator. Let $\mu_T(c) = E[Y \mid X > c, T = 1]$ be the treatment response in subjects with $X > c$ receiving treatment and $\mu_C(c) = E[Y \mid X > c, T = 0]$ the treatment response in subjects with $X > c$ receiving control. Let μ_T be the mean treatment response in subjects randomized to treatment, and μ_C be the mean treatment response in subjects randomized to control. The prevalence of the subgroup $\{X > c\}$ is $\pi(c) = \mathrm{P}[X > c]$. One way to define the best subgroup is through the minimum value of the treatment effect (Freidlin and Simon, 2005; Jiang, Freidlin and Simon, 2007). When the value of the minimum treatment effect is not available, one can define the best subgroup based on

a utility function that reflects the trade-off between the prevalence of the subgroup and the treatment effect in the subgroup. A natural form of utility is

$$U(c,\gamma) = \pi(c)^{\gamma}\left[\mu_T(c) - \mu_C(c)\right],$$

as it provides a trade-off between the size of the subgroup and the magnitude of the treatment effect. The best subgroup is then defined as $X > c^*$, where $c^* = \arg\max\left\{\pi(c)^{\gamma}\left[\mu_T(c) - \mu_C(c)\right]\right\}$. Lai, Lavori and Liao (2014) considered

$$U_1(c) = U(c,\gamma = 0.5) = \pi(c)^{0.5}\left[\mu_T(c) - \mu_C(c)\right].$$

It is proportional to the power of treatment comparison or, equivalently, the non-centrality parameter in the test for the treatment effect. Zhang et al. (2017) considered $U(c,\gamma = 1.5)$. This utility gives more weight to larger subgroups. As we show in the Appendix, $U(c,\gamma)$ with $\gamma \geq 1$ is not a good choice. In the change-point model the best subgroup corresponding to the utility $U(c,\gamma)$ with $\gamma \geq 1$ has the prevalence of 1 regardless of the model parameters. Since it can be advantageous to select a subgroup of larger size, here, in addition to $U(c,\gamma = 0.5)$ we consider $U_2(c) = U(c,\gamma = 0.75)$.

2.2.2 Estimating the subgroup

A number of methods can be used to estimate the cut-off that maximizes the utility. We propose a non-parametric approach to estimate the subgroup. In this non-parametric method, the only assumption regarding the biomarker-response relationship is that the treatment response is non-decreasing with biomarker. To estimate the cut-off, for each possible candidate cut-off *c*, we compute the test statistic for treatment effect in patients with $X > c$, then select the cut-off that maximizes the test statistic. When estimating the subgroup, it is helpful to consider subgroups with at least, say, 0.20 estimated prevalence to avoid estimated subgroups with very few subjects.

A frequently used method to estimate the subgroup from data is to fit a linear model with treatment and treatment by biomarker interaction. Then, the cutoff for the subgroup that maximizes the utility U_1 or U_2 can be obtained from the estimated coefficients in the linear model.

Below, we describe several possible models for the data where the outcome Y is continuous and give formulas for c^* for the best subgroup according to U_1 and U_2. The biomarker is distributed $X \sim N(0, 1)$ in Models 1 and 2 and $X \sim$ Uniform (0, 1) in Model 3.

Model 1. The change-point model for a continuous outcome is defined as

$$Y|T, X \sim N(E[Y|T, X],\ \sigma^2),$$

$$E[Y \mid T, X] = \beta_0 + \theta T + \delta I(X > c_0)T,$$

with $\delta > 0$ and $\pi_0 = P[X > c_0]$. The cut-off c^* that maximizes power for treatment comparison might not coincide with c_0. The best subgroup can include all subjects in which case $c^* = -\infty$. To determine if the best subgroup is defined as $X > c_0$ or includes everyone, we compare the value of $U_1(c_0)$ with $U_1(-\infty)$. Similarly, for U_2. For utility $U(c,\gamma)$ with $\gamma \geq 1$, the best subgroup for the change-point model always includes everyone (see Appendix), e.g. $c^* = -\infty$, and therefore, we do not believe it gives a good trade-off between the prevalence of the subgroup and the treatment effect.

For this change-point model, let $\pi_0 = P[X > c_0]$ and $\pi^* = P[X > c^*]$, where c^* is the cut-off that maximizes a utility function. It is clear that $U(c, \gamma)$ is maximized either at $c^* = -\infty$ or at $c^* = c_0$. Utility U_1 is maximized at $c^* = -\infty$ (corresponding subgroup prevalence is $\pi^* = 1$) when $\theta - \sqrt{\pi_0}\delta > 0$, otherwise it is maximized at $c^* = c_0$ (corresponding prevalence $\pi^* = \pi_0$). For U_2, the best cut-off is $c^* = -\infty$ when $\theta + \delta\pi_0 > (\theta + \delta)\pi_0^{3/4}$, otherwise $c^* = c_0$.

If $\gamma \geq 1$, the best cut off for the utility function $U(c,\gamma)$ defined in Section 2.2.1 is $c^* = -\infty$ because

$$U(-\infty, \gamma = 1) = \theta + \delta\pi_0 > (\theta + \delta)\pi_0^{\ 1} = U(c_0, \gamma = 1).$$

Model 2. Here we assume a bivariate normal distribution for the biomarker and outcome in the treatment group and in the control group:

$$\begin{pmatrix} Y_T \\ X \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_T \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_T^2 & \rho_T\sigma_T \\ \rho_T\sigma_T & 1 \end{pmatrix}\right], \begin{pmatrix} Y_C \\ X \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_C \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_C^2 & \rho_C\sigma_C \\ \rho_C\sigma_C & 1 \end{pmatrix}\right].$$

Clearly, if there is no correlation between *Y* and *X*, $\rho_C = \rho_T = 0$, selecting subjects based on *X* does not change the treatment effect. Same is true if $\sigma_C\rho_C = \sigma_T\rho_T$ (see Appendix for details). Otherwise, for U_1 and U_2 , if $\sigma_C\rho_C < \sigma_T\rho_T$, there is always a subgroup with the power higher than in the overall population.

In this bivariate normal model, using standard formulas (Arnold et al., 1993)

$$E\left[Y_T - Y_C \mid X > c\right] = \mu_T - \mu_C + (\sigma_T\rho_T - \sigma_C\rho_C)\left(\frac{\phi(c)}{1-\Phi(c)}\right),$$

$$Var(Y_T - Y_C \mid X > c) = \sigma_T^2 + \sigma_C^2 - (\sigma_T{}^2\rho_T{}^2 - \sigma_C{}^2\rho_C{}^2)\left[\left(\frac{\phi(c)}{1-\Phi(c)}\right)^2 - c\left(\frac{\phi(c)}{1-\Phi(c)}\right)\right],$$

where $\phi(c)$ is a normal density and Φ is normal cumulative distribution function. We maximize $\frac{E(Y_T - Y_C \mid X > c)\pi^\gamma}{\sqrt{Var(Y_T - Y_C \mid X > c)}}$ with γ = 1/2 if U_1 is used or γ = 3/4 if U_2 is used. There is no closed form formulae for the optimal cut-off.

Model 3. The model with treatment by biomarker interaction is

$$Y \mid T, X \sim N\left(E[Y], \sigma^2\right),$$

$$E[Y] = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 g(X)T, \; g(X) = X^a \quad \text{with } a > 0.$$

Without loss of generality we set $\beta_0 = \beta_1 = \beta_2 = 0$, and $\beta_3 > 0$. Refer to the Appendix for how to find the best cutoff, c^*, for U_1 and U_2. Table 1 shows effect sizes and subgroup prevalence for several values of a in $E[Y] = X^a T$. As expected, maximizing U_2 yields a larger subgroup with a smaller effect size. For $a \geq 1$, the effect size in the subgroup is much larger than the effect size in all patients.

The mean and the variance in the overall population with size N are

$$Var(Y) = \sigma^2 + Var(\beta_3 T X^a) = \sigma^2 + \beta_3^2 T \frac{a^2}{(2a+1)(a+1)^2},$$

$$E(\bar{Y}_T - \bar{Y}_C) = \frac{\beta_3}{a+1},$$

$$Var_{H_1}(\bar{Y}_T - \bar{Y}_C) = \frac{\sigma^2 + \sigma^2}{N/2} + \frac{\beta_3^2}{N/2} \frac{a^2}{(2a+1)(a+1)^2},$$

$$Var_{H_0}(\bar{Y}_T - \bar{Y}_C) = \frac{\sigma^2 + \sigma^2}{N/2}.$$

For the subgroup $X > c$ with prevalence π we have

$$E(\bar{Y}_T - \bar{Y}_C) = \frac{\beta_3(1-c^{a+1})}{(a+1)(1-c)},$$

$$Var_{H_1}(\bar{Y}_T - \bar{Y}_C) = \frac{\sigma^2 + \sigma^2}{(N/4)\pi} + \beta_3^2 \frac{1-c^{2a+1}}{(2a+1)(1-c)},$$

$$Var_{H_0}(\bar{Y}_T - \bar{Y}_C) = \frac{\sigma^2 + \sigma^2}{(N/4)\pi}.$$

Then we maximize

$$\frac{E_{H_1}(\bar{Y}_T - \bar{Y}_C \mid X > c)}{\sqrt{V_{H_1}(\bar{Y}_T - \bar{Y}_C \mid, X > c)}} \propto \frac{\beta_3 \frac{(1-c^{a+1})}{(1-c)(a+1)}(1-c)^{\gamma}}{\sqrt{2\sigma^2 + \beta_3{}^2 \left\{ \frac{(1-c^{2a+1})}{(1-c)(2a+1)} - \left(\frac{(1-c^{a+1})}{(1-c)(a+1)} \right)^2 \right\}}},$$

with $\gamma = 1/2$ if U_1 is used and $\gamma = 3/4$ if U_2 is used.

2.3. Testing for the treatment effect

In section 2.2, we discussed the estimation of the cut-off c^* for the best subgroup. In this section we are interested in testing for the treatment effect in the estimated subgroup $X > c^*$. Patients are randomized between treatment and control. The trial is run as a single stage trial, however, to estimate the cut-off, the study subjects are divided into two cohorts. The biomarker data from each cohort is then used to estimate the cut-off for the biomarker in the other cohort. We are interested in testing the equality of treatment effects in all subjects, $H_{0,All}: \mu_T = \mu_C$, as well as the equality of treatment effects in the subgroup $H_{0,S}: \mu_T(c^*) = \mu_C(c^*)$. We are also interested in testing the intersection hypothesis $H_{\cap}: H_{0,All} \cap H_{0,S}$ against the alternative hypothesis that there is a treatment effect in everyone or in the subgroup.

To estimate the cut-off, we use the cross-validation approach from Freidlin, Jiang and Simon (2010). While Freidlin, Jiang and Simon (2010) used $K = 10$, we use $K = 2$ because we did not see any difference in performance among the values of K between 2 and 10. With $K = 2$ the sample is split into two cohorts. We estimate the cut-off from cohort 1 and use this estimated cut-off to define the subgroup in cohort 2. Then, we estimate the cut-off from cohort 2 and use this estimated cut-off to define the subgroup in cohort 1. A non-parametric and a parametric method we used to estimate the cut-offs are described in section 2.4. Let $\hat{c}_1$ be the cut-off estimated from cohort 2 data to define the subgroup in cohort 1. Similarly, let $\hat{c}_2$ be the cut-off estimated from

cohort 1 data to define the subgroup in cohort 2. Denote $Z_{i,All}$ and $Z_{i,S}$ to be the test statistics to test $H_{0,All}$ and $H_{0,S}$ based on data from cohort *i*. The test $Z_{i,S}$ is based on cohort *i* data where the subgroup is defined as $X > \hat{c}_i$. Consider the following test statistics:

$$Z_{All} = \sqrt{0.5}Z_{1,All} + \sqrt{0.5}Z_{2,All},$$

$$Z_{All,S} = \sqrt{0.5}Z_{1,All} + \sqrt{0.5}Z_{2,S},$$

$$\tilde{Z}_S = \sqrt{0.5}Z_{1,S} + \sqrt{0.5}Z_{2,S}.$$

When the number of subjects in each cohort is equal, test Z_{All} is equivalent to testing the treatment effect in the overall population in combined cohorts 1 and 2. The test $Z_{All,S}$ uses data from all subjects in cohort 1 and subjects in the subgroup in cohort 2, and is a test of the null hypothesis that there is no treatment effect in all patients and in the subgroup. This test preserves the type I error rate since $Z_{1,All}$ and $Z_{2,S}$ are independent. This test can be viewed as a test of any treatment effect, as it combines the test of the treatment effect in everyone with testing for treatment effect in a more promising subset of patients, the estimated subgroup. If the best subgroup does not coincide with the overall population, the power of this test is lower than when testing in the subgroup only. The test based on $\tilde{Z}_S$ does not control type I error rate because the biomarker cut-off that defines the subgroup in cohort 1 is based on the estimate from cohort 2 data and vice versa. Our simulations show that the type I error rate can be as high as 0.062 for $\tilde{Z}_S$ (refer Table 2). Instead one can use a permutation-based test (Jiang, Freidlin and Simon, 2007; Freidlin, Jiang and Simon, 2010) to test $H_{0,S}$ based on $\tilde{Z}_S$. We refer to this test as Z_S. The p-value for the permutation based test was defined as the proportion of permutations of treatment assignments

where the resulting test statistic is higher in the absolute value than the test corresponding to the original data. We used the Hochberg method to test $H_{\cap}$, rejecting both hypotheses if the larger of the two p-values is less than α or rejecting the intersection hypothesis with p-value smaller than $\alpha/2$.

2.4. Simulation study

The goals of the simulation study were to compare the non-parametric and parametric methods for cut-off estimation, to illustrate subgroup selection based on utilities U_1 and U_2 and to see if the power of testing for treatment effect can be increased through finding the best subgroup in retrospective analysis of data. Data were generated from the three models described in section 2.2.2. The total sample size was 500 in trials for Models 1 and 2 and between 50 and 250 for Model 3. Simulations were performed in R with 5000 simulation runs in each scenario under alternative hypothesis, with 10000 simulations runs under the null hypothesis. When reporting the prevalence of the estimated subgroup, we computed the true prevalence corresponding to estimated cut-offs, $\hat{c}_1$ and $\hat{c}_2$, and reported the average $0.5P\left[X > \hat{c}_1\right] + 0.5P\left[X > \hat{c}_2\right]$.

We performed simulations under the null (Table 2) and alternative hypotheses (Table 3). The type I error rate was as high as 0.063 for testing using the naïve approach with $\tilde{Z}_S$ (Table 2). After applying the permutation method to test the treatment effect in the subgroup, the type I error rate was well controlled for all models using the non-parametric methods, with slight inflation using the parametric method (Table 2).

Table 3 contains simulation results where treatment outcomes were simulated from the change-point model, Model 1 in section 2.2.2 with $\sigma^2 = 1$. Trial data were split into two cohorts of 250 subjects each to estimate the biomarker cut-off. We show the true cut-off c^* and the theoretical power corresponding to c^* to illustrate the amount of power loss when the cut-off was

not estimated precisely. As can be seen from Table 3, when the true model is a change-point model, neither the non-parametric approach nor the parametric approach of estimating c^* yielded good estimates. This is unfortunate because we were expecting for the non-parametric method to do well and better compared to a linear model in this scenario. Improving the performance of a linear model in change-point model scenarios was the reason of investigating the non-parametric method for subgroup estimation. Both the non-parametric and parametric methods yielded lower power in the estimated subgroup compared to the true theoretical power when the best subgroup is known. As expected, U_2 yields a larger subgroup than U_1. In the setting of re-analysis of data considered here, defining the subgroup based on U_1 is theoretically optimal. Despite U_1 being optimal for power, the power was comparable to that in the subgroup that optimized for U_2 compared to U_1. Non-parametric method yielded slightly better power than the parametric method.

Table 4 shows results for the bivariate normal model, Model 2 in section 2.2.2 with $\sigma^2_C = 1$. Overall the parametric method is better than the non-parametric method for both the estimation of the subgroup and power. As in the change-point model, U_2 yields a larger subgroup than U_1. Both U_1 and U_2 subgroups yielded similar power in the first scenario. When the best true subgroup had a prevalence of 1, U_2 yielded higher power than U_1, as expected, as it yields a larger estimated subgroup.

Table 5 shows simulations for the linear model with treatment by biomarker interaction, Model 3 in section 2.2.2 with $\sigma^2 = 1$. The model we fit in the parametric method coincides with the model we used to generate the data when $a = 1$. Therefore, the parametric approach is expected to perform well in that scenario. Interestingly, parametric and non-parametric approach performed similarly in this scenario. Overall both methods performed similarly with a slight advantage of the parametric method.

We compare the proposed non-parametric method with a method where the biomarker cut-off for the subgroup is selected based on minimizing the p-value for testing the interaction between the continuous biomarker and treatment, in a linear model. For model 1, for example, the interaction method selects subgroups that are much smaller than expected and the power in the subgroup does not exceed the power in all (and is lower than the corresponding subgroup power using non-parametric method).

The Hochberg approach that combines the tests of the subgroup and overall population is a robust test to detect any treatment effect (Tables 3, 4 and 5). It maintains good power in cases where the subgroup is estimated poorly, for example, when the parametric method is applied with U_1 in scenario 2 (Table 3), or when the subgroup coincides with the overall population (scenario 3, Table 3). Therefore, we recommend using this test instead of relying on the test of the subgroup only or using $Z_{All,S}$.

2.5. Example

We applied our methods to data from a phase 2 study of a novel treatment 1C4D4 to treat patients with metastatic pancreatic cancer (Wolpin et al., 2013). A total of 205 subjects were randomized in the ratio of 2:1 to Gemcitabine plus 1C4D4 and Gemcitabine alone. Among the randomized patients, 123 had adequate tumor tissue for immunohistochemistry (IHC) analysis of prostate stem cell antigen (PSCA). This was used as the analysis set. The primary outcome was overall survival. The median survival in the Gemcitabine+1C4D4 arm was 7.92 months and in the Gemcitabine alone arm was 5.52 months, yielding the logrank test p-value of 0.20. A continuous biomarker, prostate stem cell antigen expression measured by IHC, H-SCORE, with values from 0 to 290, was believed to be a possible effect modifier for 1C4D4. We applied our non-parametric and parametric approaches to find the best subgroup based on H-SCORE by maximizing U_1 and

U_2 (Table 6). In the parametric approach, we fit the Cox-model with biomarker by treatment interaction. In the non-parametric approach, we used the logrank test. Table 6 shows the sizes of the estimated best subgroups and p-values. Selecting patients with higher values of H-SCORE did not result in a smaller p-value. Both non-parametric and parametric methods yielded similar results indicating that there might not be a subgroup defined by H-SCORE with better treatment effect than in the overall population. In fact, H-SCORE appears to have more of a prognostic rather than predictive effect. In a Cox model with H-SCORE dichotomized at the median H-SCORE = 120, the coefficient for H-SCORE is significant (p-value = 0.03) while the interaction term is close to 0 with the p-value of 0.96.

2.6. Conclusions

For several true models of response to treatment and biomarker, such as a change-point model, a bivariate normal model and a linear model with interaction, we compared two methods of estimation of the best subgroup, non-parametric and model-based. In a model-based approach, we used the linear model with treatment by biomarker interaction, the model that is used frequently for subgroup estimation (Freidlin and Simon, 2005; Jiang, Freidlin and Simon, 2007; Freidlin, Jiang and Simon, 2010). Our conclusion is that the non-parametric method performed very similarly to fitting a linear model with interaction with slight advantage of a linear model. It is no surprise that fitting a linear model with interactions is a preferred method for subgroup estimation.

We illustrated the use of a utility function to choose the best subgroup in a clinical trial. The best subgroup was defined through maximizing the non-centrality parameter, utility U_1, or through maximizing utility U_2 that gives more weight to larger subgroups. In the retrospective data analysis setting we considered, U_1 is the optimal choice because it maximizes the power of treatment comparison. In our simulations both two approaches performed equally well. There is

no obvious method for selecting the best subgroup in adaptive enrichment trials where further patient enrollment is restricted to the selected subgroup. The class of utilities $U(c,\gamma) = \pi(c)^{\gamma}\left[\mu_T(c) - \mu_C(c)\right]$ with $0 < \gamma < 1$ can be useful for selecting a subgroup in adaptive enrichment trials.

Using cross-validation as in Freidlin, Jiang and Simon (2010), we gain the advantage of utilizing all observations for both estimating the cut-off and testing for the treatment effect. Permutation test used after cross-validation controls type I error rate for the test in the subgroup well. To test for any treatment effect, the Hochberg method is a robust method to test the intersection hypothesis of the treatment effect in all and in the subgroup. It yields good power in both cases, when power is high in all subjects, but not in a subgroup and when power is only high in the subgroup. There might be more powerful alternatives to the Hochberg method that make a better use of the correlation between the tests.

Our investigation shows that the subgroup can be estimated after the clinical trial with subsequent computation of a valid p-value for treatment effect in the subgroup. Power in some clinical trials can be increased by estimating the subgroup from collected data and testing for treatment effect in it if there is a subgroup of patients with a higher treatment effect.

CHAPTER 3: MULTI-STAGE ADAPTIVE ENRICHMENT DESIGN

3.1 Introduction

Adaptive enrichment trials adapt the entry criteria based on data observed in the trial so far to restrict enrollment to subjects in whom the experimental treatment is believed to work. Enriching the subject population in the trial can increase statistical power to detect a treatment effect if a new therapy only works in a subgroup of subjects. Additionally, it reduces the number of subjects in the trial who have no apparent benefit from the drug therefore not exposing them to potentially harmful side-effects. Adaptive enrichment literature can be divided into two categories: methods that adaptively enrich based on already pre-specified subgroups and methods where the subgroup is estimated during the trial. Most methods with predefined subgroups, specify a single subgroup (Jenkins, Stone and Jennison, 2011; Ondra et. al, 2016), while some allow predefining several candidate subgroups, usually up to three (Wang, Hung and O'Neill, 2009; Lai, Lavori and Liao, 2014; Lui et. al, 2010; Wassmer and Dragalin, 2015). Only a handful of publications describe clinical trial designs where the subgroup is estimated and the treatment effect is tested in the same trial (Friedlin and Simon, 2005; Jiang, Freidlin and Simon, 2007, 2010; Simon and Simon, 2013, 2017; Zhang et. al, 2017, 2018; Diao et. al, 2019).

Lipkovitch, Dmitrienko and D'Agostino (2017) recently reviewed methods for subgroup estimation. They categorized trials with subgroup estimation into two classes based on the objectives in enrichment. The first class is the enrichment trials where the goal is to find the 'best subject' for a given treatment. Subgroup in these methods is usually estimated by fitting a model

with interaction. The second class is trials with the goal of finding the optimal treatment rules for a given subject. Here our goal is the former, that is, to find the best subjects for a given treatment and to demonstrate that the treatment is better than control. For subgroup estimation, when dealing with multiple biomarkers, non-parametric tree-based regression methods may be more suited to deal with higher order interactions, or unknown forms of relationship between covariates and response (Loh, He and Man, 2014).

In publications with methods where the subgroup is being estimated during the trial, several of these (Simon and Simon, 2013, 2017; Zhang et al.,2017; Diao et. al, 2019) considered a trial with adaptive enrichment. Simon and Simon (2013) described a multi-stage design with a single biomarker. The best cut-off for the single biomarker was defined as the one maximizing the interaction term. At each stage the best subgroup is estimated and the next stage enrolls patients from the best subgroup only. Zhang et al. (2018) considered a two-stage adaptive enrichment design with up to two predictive biomarkers where the second stage enrolls patients in the subgroup estimated using the first stage data. Diao et al. (2019) described a two-stage adaptive enrichment design with a single continuous predictive biomarker and time to event endpoint.

The best subgroup is generally defined in one of two ways. The first way is to define the best subgroup as subjects within the biomarker subset where the treatment effect is equal to or higher than a minimally clinically relevant treatment effect (Friedlin and Simon, 2005; Renfro et. al, 2014). The second way to define the best subgroup is through a utility function (Lai, Lavori and Liao, 2014; Zhang et. al, 2017; Joshi et al., 2018). The utility function specifies the trade-off between the size of the subgroup and the treatment effect in the subgroup. An example of a utility function is the function equal to the square root of subgroup prevalence multiplied by the treatment effect in the subgroup (Lai, Lavori and Liao, 2014). Under the assumption of equal variances of

the treatment effect in all subgroups, maximizing this utility is the same as maximizing the square root of the prevalence of the subgroup multiplied by the treatment effect in the subgroup, and is the same as maximizing the power of the treatment comparison. The subgroup defined this way yields good power of treatment comparison in a post-hoc analysis on unenriched population when the subgroup is estimated and tested (Joshi et al., 2018). The advantage of the utility function approach is that there is no need to pre-specify the minimum treatment effect.

In this chapter, we evaluate a number of methods to estimate the subgroup in an adaptive enrichment trial with the goal of establishing an initial efficacy of a new treatment in any subgroup. We propose a three-stage design where the best subgroup is estimated after stage 1 and refined after stage 2. Lai, Lavori and Liao (2014) also considered a three-stage design but they worked with a pre-specified set of candidate subgroups rather than estimating the subgroup during the trial. In Section 3.2, we give the definition of the best subgroup and illustrate it on several true models for response to treatment as a function of biomarker and treatment. Testing for the treatment effect and the design of the trial is discussed in Section 3.3. In Section 3.4 we compare several methods of subgroup estimation via simulations. Conclusions are presented in Section 3.5.

3.2 Subgroup Estimation Methods

Let $\boldsymbol{X} = (X_1, X_2, ..., X_M)$ be a vector of continuous biomarkers measured at baseline. We work with X_m in [0,1], as biomarkers can be always rescaled. Subjects are randomized between active treatment (T = "Active") and control (T = "Control"), where T is the treatment indicator. Let Y be a continuous response variable such that higher values indicate improvement in the well-being of the subject. Let, $\mu_T(\boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x}, T = \text{"Active"})$ and $\mu_C(\boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x}, T = \text{"Control"})$ be the expected responses of a subject at observed biomarker values, $\boldsymbol{x}$, randomized to the active treatment and control respectively. For the i^{th} randomized subject with i

= 1…, $n, (x_{i1},...,x_{iM}, y_i, t_i)$ represents the observed data, where $x_{i1}...x_{iM}$ are the observed values of continuous biomarkers some or none of which are associated with response to treatment.

To identify the best subgroup, i.e., a subset of the full population that is not too small and shows response to the treatment, we use a utility function to quantify the trade-off between the size of the subgroup and the magnitude of the treatment effect (Lai, Lavori and Liao, 2014; Zhang et. al, 2017). Let $S \equiv S(\boldsymbol{X})$ be a subgroup based on the biomarker vector $\boldsymbol{X}$. A natural form of the utility is,

$$U(S,\gamma) = \pi(S)^{\gamma}\left[\mu_T(S) - \mu_C(S)\right],$$

where $\pi(S) = P(\boldsymbol{X} \in S)$ is the prevalence of the subgroup and $\gamma \in [0,1]$ denotes the corresponding weight. Here $\mu_T(S)$ and $\mu_C(S)$ are the expected responses to the treatment and control in the subgroup. For a given value of γ, the best subgroup is then defined as S^* where $S^* = \arg\max\left\{\pi(S)^{\gamma}[\mu_T(S) - \mu_C(S)]\right\}$.

Lai, Lavori and Liao (2014) considered $U_1 = U(S, \gamma = 0.5) = \pi(S)^{0.5}\left[\mu_T(S) - \mu_C(S)\right]$. This function is proportional to the power of treatment comparison or, equivalently, the non-centrality parameter in the test for the treatment effect. Joshi et al (2018) showed that for weights larger than 1 the best subgroup always coincides with the whole population. They considered the utility $U_2 = U(S, \gamma = 0.75) = \pi(S)^{0.75}[\mu_T(S) - \mu_C(S)]$ that favors larger subgroups.

Defining the subgroup through maximizing a utility allows comparing methods of estimation of the best subgroup using a single measure. We introduce a measure we refer to as $\%U$. It is computed by taking the ratio of the value of the utility corresponding to the estimated subgroup to the value of the utility of the best theoretical subgroup for a given true model and then multiplying by 100%. If the best subgroup is estimated perfectly, the $\%U$ is equal to 100%.

Methods for subgroup estimation are usually compared using the sensitivity and specificity. Since methods with higher sensitivity usually have lower specificity, it is hard to compare estimation methods based on these two values. Using the new measure, %*U*, we evaluate several methods for subgroup estimation that have been discussed in the literature (Lipkovitch, Dmitrienko and D'Agostino, 2017).

A common parametric approach to estimating the subgroup for continuous outcome is to consider a linear model (LM) with first order main effects for biomarker and all pairwise interaction terms between all available biomarkers and treatment. Let T^* be the treatment indicator with $T^* = 1$ for subjects randomized to treatment and $T^* = -1$ for subjects on control, in the estimation models. We use T^* here to distinguish it from the treatment indicator T = 1 or 0 more commonly used. We will use T = 1 or 0 later to define true models. Consider the model,

$$E(Y \mid \boldsymbol{X}, T^*) = \alpha + \sum_{m=1}^{M} \gamma_m X_m + \sum_{m=1}^{M} \delta_m X_m T^* + \sum_{l=1}^{M-1} \sum_{m=l+1}^{M} \lambda_{lm} X_l X_m T^* .$$

Following Tian et. al (2014), consider also the model where the outcome is modified by multiplying the responses of subjects on control are by -1

$$E(T^* Y \mid \boldsymbol{X}) = \alpha T^* + \sum_{m=1}^{M} \gamma_m X_m T^* + \sum_{m=1}^{M} \delta_m X_m T^{*2} + \sum_{l=1}^{M-1} \sum_{m=l+1}^{M} \lambda_{lm} X_l X_m T^{*2} .$$

Since T^* = {-1, 1} with probability 0.5, we have

$$E(T^* Y \mid \boldsymbol{X}) = \sum_{m=1}^{M} \delta_m X_m + \sum_{l=1}^{M-1} \sum_{m=l+1}^{M} \lambda_{lm} X_l X_m . \qquad (1)$$

This is the model we fit to the data. After the coefficients are estimated from the data, we can compute the expected treatment response for i^{th}, i = 1,…,n, subject with biomarker vector $\boldsymbol{X}_i$. Let vector $(\hat{y}_1, ..., \hat{y}_n)$ be the estimated treatment responses for the set of biomarker vectors

$\{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n\}$ in the data set. For each predicted response $\hat{y}_i$, we define the corresponding subgroup as the set of subjects with predicted response larger than $\hat{y}_i$ or $\{\boldsymbol{X}_j : \hat{y}_j > \hat{y}_i\}$. Then, we compute the estimated prevalence, estimated treatment response and estimated utility of this subgroup. Denote the estimated treatment response for which the estimated utility is maximized by y^*, $y^* = \hat{y}_i$, for some *i*, *i* = 1,…,*n*. The best subgroup $\{\boldsymbol{X}_j : \hat{y}_j > y^*\}$ includes subjects with the biomarker vector such that the predicted treatment response for that vector is higher than y^*. We use this approach for all subgroup estimation methods we consider.

Least Absolute Shrinkage and Selection Operator or LASSO (Tibshirani, 1996) is a shrinkage and variable selection method based on linear regression. The coefficients in a linear model (1) with covariates $\{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_M\}$ are estimated by minimizing sum of squared residuals, with a bound on the sum of absolute value of the coefficients. Alternatively, we minimize the sum of squared residuals plus a penalty term equal to the sum of with absolute values of the coefficients multiplied by $\gamma > 0$

$$min \sum_{i=1}^{n} \left(T_i^* Y_i - \sum_{m=1}^{M} \delta_m X_{im} + \sum_{l=1}^{M-1} \sum_{m=l+1}^{M} \lambda_{lm} X_{il} X_{im} \right)^2 + \tau \left(\sum_{m=1}^{M} | \delta_m | + \sum_{l=1}^{M-1} \sum_{m=l+1}^{M} | \lambda_{lm} | \right),$$

Large values of γ puts a higher penalty and shrinks most coefficients to zero and lead to underfitting. Smaller values of γ results in LASSO shrinking coefficients of some covariates (if not considered important) to zero and can thus can help with variable selection by removing covariates that are not associated with the outcome. Yuan and Lin (2006) developed a group LASSO method where the covariates are considered together in non-overlapping groups. If a specific group is selected, then the coefficient estimates of all those in the group will be non-zero and zero if they belong to a group not selected. An advantage of forming groups is that we avoid choosing the interaction term if the corresponding main effects are not selected. A disadvantage

of using just grouped lasso is that it prevents model variables from belonging to multiple groups. Zeng and Breheny (2016) improved upon this by adding an overlap condition allowing for a model variable to belong to more than one group and thus have non-zero coefficient if any of the groups it belongs to is selected. Suppose, if a covariate, X_1 belongs to two groups $G_1 = \{X_1, X_2, X_1X_2\}$ and $G_2 = \{X_1, X_3, X_1X_3\}$ such that only G_1 is truly related to the outcome. If we do not use overlapping group LASSO, then if G_2 is not selected, the coefficient for X_1 is set to 0, even though it is present in G_1.We used the overlapping group LASSO (OGLASSO) of Zeng and Breheny (2016) in the simulations. The linear model is re-formulated in terms of the group coefficients that are obtained by minimizing,

$$min\left(T^*Y - \tilde{\boldsymbol{X}}\boldsymbol{\theta}\right)^T \left(T^*Y - \tilde{\boldsymbol{X}}\boldsymbol{\theta}\right) + \tau\left(\sum_{g=1}^{G}\sqrt{K_g}\,\|\,\theta^g\,\|\right).$$

where, $\boldsymbol{\theta}^g = (\theta_1^g, ..., \theta_M^g)$ is the Mx1 vector of coefficients corresponding to each original predictor in the g^{th} group, $\boldsymbol{\theta}$ is the vector of all $\boldsymbol{\theta}^g$, $\tilde{\boldsymbol{X}}$ is the design matrix, G is the number of groups and K_g is the number of elements in the g^{th} group. When $\boldsymbol{\theta}^g$ is selected, all model variables in this group are selected, irrespective of whether they are present in another group. A tree-based method, Classification And Regression Trees (CART) (Breiman et al., 1984), recursively partitions the data into two disjoint subsets by minimizing the heterogeneity of the outcome in each partition. The resulting prediction model can be illustrated by a single decision tree and the terminal nodes of the tree can be interpreted as subgroups. In a tree-based method, Random Forests (RF) (Breiman, 2001) the predicted value is an average over a collection of trees rather than a single tree as in CART. Unlike a single decision tree in CART, random forests prediction model cannot be described as a set of rules as the CART model making it a 'black box' type prediction

model. Support Vector Machine (SVM) introduced by Cortes and Vapnik (1995) is used for both classification and regression problems. In case of a continuous outcome, it fits a hyperplane or a function such that all points on either side of this function are within a certain pre-defined distance from the function and there is a penalty for points falling outside the range. The regression method seeks to find a linear function which can be used to predict the outcome for each subject. We compare these four methods in the simulation study in Section 3.4 to give recommendation on the method to be used for subgroup estimation in a clinical trial.

3.3 Adaptive design with enrichment

We propose a three-stage enrichment design for a randomized trial comparing a new treatment with control. We enroll a total of n subjects, $n_1 + n_2 + n_3 = n$, with n_k subjects enrolled in the k^{th} stage, $k = 1, 2, 3$. At each stage, the subjects are equally likely to be randomized to the experimental treatment or control. The dual objective of the trial is to demonstrate the efficacy of a new therapy in any subgroup and to estimate the best subgroup. The best subgroup is defined as the subgroup that maximizes the utility U_1. We propose the following three-stage design:

(1) In stage 1, n_1 subjects are enrolled from the full population. At the first interim analysis, using data from stage 1, the best subgroup is estimated based on maximizing utility U_2.

(2) In stage 2, subject population is fully enriched that is only subjects from the subgroup estimated at the end of stage 1 are enrolled. At the second interim analysis, we use data from $n_1 + n_2$ subjects enrolled so far to estimate the subgroup based on maximizing a utility U_1.

(3) In stage 3, only subjects from the subgroup estimated at the end of stage 2 are enrolled.

(4) At the end of the trial, let Z_k be the test statistic to test H_0 based on stage *k* data, defined as

$$Z_k = \frac{\hat{\mu}_{T,k} - \hat{\mu}_{C,k}}{\hat{\sigma}_k \sqrt{\frac{1}{0.5n_k} + \frac{1}{0.5n_k}}},$$

where $\hat{\mu}_{T,k}$ and $\hat{\mu}_{C,k}$ are the estimated mean responses in treatment and control arms respectively and $\hat{\sigma}_k^2$ is the estimated common variance for treatment and control groups at stage *k*, *k* = 1,2,3. Consider the test statistic:

$$\tilde{Z} = \sqrt{n_1 / n} Z_1 + \sqrt{n_2 / n} Z_2 + \sqrt{n_3 / n} Z_3 .$$

A test based on $\tilde{Z}$ preserves the type I error rate since, conditional on the enrollment decision taken at the end of stages 1 and 2, the components Z_k are independent. A similar approach for testing the hypothesis of no treatment effect was used in Simon and Simon (2013). Assuming that the response to the new treatment is not worse than control for any set of biomarkers, the test $\tilde{Z}$ is the test for any treatment effect $H_0: \mu_T - \mu_C = 0$.

In Stage 2, only patients from the estimated best subgroup are enrolled. This leads to oversampling of subjects in the best estimated subgroup in combined stage 1 and stage 2 sample. Hence, inverse probability weighting needs to be used when working with the combined stage1 and 2 sample. Each subject in the estimated best subgroup, stage 2 population, regardless of the stage of enrollment, is assigned a weight of $\hat{\pi}_1 / (1 + \hat{\pi}_1)$ where $\hat{\pi}_1$ is the estimated prevalence of the best estimated subgroup Stage 2 population was sampled from. Otherwise, the weight is 1. The weights are used after the model is fitted at the utility maximization step described in Section 2. To adjust the mean response, the set of subjects with predicted response less and greater than the predicted cutoff after stage 1 is weighed by the corresponding 'true' proportion.

In Section 3.4, we run simulations to optimize design parameter: choosing the utility to define the subgroup at the end of stage 1. The goal is to find a design that results in a good quality of estimation of the best subgroup by the end of stage 2 and, at the same time, yields a good power of treatment comparison. Though the best subgroup is defined based on U_1, identifying a larger subgroup at the end of stage 1 might improve the subgroup estimation at the end of stage 2.

As a new therapy might not work, we consider the possibility of stopping for futility at the end of stage 2. If a new therapy only works in a small subgroup of subjects, the observed treatment effect might be low after stage 1 when the new therapy is investigated in the full population. Therefore, stopping for futility is not considered at the end of stage 1. We use the method of He, Lai and Liao (2012) to test for futility with a one-sided test $H_{1,f}: \mu_{T,1-2} - \mu_{C,1-2} \geq \delta_f$. We set the value δ_f equal to the alternative used for power calculations for the trial. Let $\mu_{T,1-2}$ and $\mu_{C,1-2}$ be denote the expected responses in treatment and control groups for subjects enrolled up to the second interim. analysis. We compute $Z'_f = (\hat{\mu}_{T,1-2} - \hat{\mu}_{C,1-2} - \delta_f) \Big/ \sqrt{\hat{\sigma}^2_{1-2}\left[1/(0.5n_1) + 1/(0.5n_2)\right]}$ where $\hat{\mu}_{T,1-2}$ and $\hat{\mu}_{C,1-2}$ are the estimated responses in treatment and control groups and $\hat{\sigma}^2_{1-2}$ is the estimated pooled sample variance for data collected up to the second interim analysis. Values of $Z'_f < \tilde{b}$ indicate that $H_{1,f}$ is not true and the trial is stopped for futility. The futility stopping cut-off, $\tilde{b}$, is computed to have $P_{\delta_f}(\text{Reject } H_{1,f}) = 0.05$, that is, the probability of stopping for futility when the treatment effect is equal to δ_f is equal to 0.05.

3.4. Simulation Study

3.4.1 Comparison of subgroup estimation methods

First, we compare the performance of the five subgroup estimation methods described in Section 3.2 for several true models via simulations. The best subgroup is defined as the subgroup that maximizes the utility U_1. In the change-point models below, the best subgroup either includes everyone or is defined by the function of biomarkers in the indicator function of the model. To find whether it is the former or the latter, we compute U_1 on both of these candidate subgroups and select the one with the larger value of U_1. In more complex cases, for example in model 5, one can find the best subgroup by generating a large set of biomarker vectors, e.g., 100,000. The mean treatment response values are then computed for each of the vectors, the values are ordered to find the value of response y^* such that the utility U_1 is maximized for all the biomarker vectors that yield the values of treatment response larger than y^*.

In the models below $X_m \sim U(0,1)$, m = 1 and 2, and $Y \sim N(E[Y \mid \boldsymbol{X}, T], 1)$. The treatment indicator variable T is defined as $T = 1$ for an active treatment and $T = 0$ for control. Figure 1 shows the best subgroup for each of the examples.

Model 1. Change-point model with a single biomarker with

$$E[Y \mid \boldsymbol{X}, T] = \mu T + \theta I(X_1 > c_1) T .$$

When $\mu = 0$, $\theta = 0.4$ and c_1 = 0.5, the subgroup that maximizes U_1 for this model is $S^* = \{X_1 > 0.5\}$. For this best subgroup, the treatment effect is $\delta = 0.4$, the prevalence is $\pi_{U_1} = P[\boldsymbol{X} \in S^*] = 0.5$ and the utility is $U_1 = \delta\sqrt{\pi_{U_1}} = 0.28$. Note that another candidate for the best subgroup is the subgroup that includes all subjects. The corresponding utility is 0.2, lower than the utility for S^*.

Model 2. Change-point model with two biomarkers with

$$E[Y \mid \boldsymbol{X}, T] = \mu T + \theta\, I(X_1 > c_1 \text{ and } X_2 > c_2\ )T\,.$$

When $\mu = 0$, $\theta = 0.4$, c_1 = c_2 = 0.5, the subgroup that maximizes U_1 for this model is $S^* = \{X_1 > 0.5 \text{ and } X_2 > 0.5\}$. For this best subgroup, the treatment effect is $\delta = 0.3$, the prevalence is $\pi_{U_1} = P\left[\boldsymbol{X} \in S^*\right] = 0.25$ and the utility is $U_1 = \delta\sqrt{\pi_{U_1}} = 0.20$.

Model 3. Change-point model with two biomarkers with

$$E[Y \mid \boldsymbol{X}, T] = \theta\ [I(X_1 > c_1)(X_2 < c_2) + I(X_1 < c_1)(X_2 > c_2) + I(X_1 > c_1)(X_2 > c_2)]T\,.$$

When $\theta = 0.4$, c_1 = 0.8 and c_2 = 0.75, the best subgroup that maximizes U_1 is $S^* = S_1 \bigcup S_2$ where $S_1 = \{X_1 > 0.8 \text{ and } X_2 < 0.75\}$, $S_2 = \{X_1 < 0.8 \text{ and } X_2 > 0.75\}$. For the best subgroup, the treatment effect is $\delta = 0.4$, the prevalence is $\pi_{U_1} = P[\boldsymbol{X} \in S^*] = 0.40$ and the utility is $U_1 = \delta\sqrt{\pi_{U_1}} = 0.25$.

Model 4. Change-point model with two biomarkers with

$$E[Y \mid \boldsymbol{X}, T] = \mu T + \theta\, I((X_1 + X_2) > c)T\,.$$

When $\mu = 0$, $\theta = 0.38$, and c = 0.1, the best subgroup that maximizes U_1 for this model is $S^* = \{(X_1 + X_2) > 0.85\}$. For the best subgroup, the treatment difference is $\delta = 0.38$, the prevalence is $\pi_{U_1} = P\left[\boldsymbol{X} \in S^*\right] = 0.5$ and the utility is $U_1 = \delta\sqrt{\pi_{U_1}} = 0.3$.

Model 5. A model with two biomarkers with

$$E[Y \mid \boldsymbol{X}, T] = \mu e^{-\theta(|X_1 - c_1| + |X_2 - c_2|)}T\,.$$

The best subgroup S^* when $\mu = 1.5$, $\theta = 6.5$, c_1 = 0.8 and c_2 = 0.75 is shown in Figure 1. The average treatment effect in the best subgroup is $\delta = 0.5$, the prevalence is $\pi_{U_1} = 0.15$ and the utility is $U_1 = \delta\sqrt{\pi_{U_1}} = 0.19$.

Model 6. Change-point model with two biomarkers with

$$E[Y \mid \boldsymbol{X}, T] = \mu T + \theta I(X_1 > c_1 \text{ and } X_2 > c_2\ )T.$$

When $\mu = 0.25$, $\theta = 0.35$, c_1 = c_2 = 0.5, the subgroup includes everyone, $\pi_{U_1} = 1$, with the treatment effect of $\delta = 0.34$ and utility of $U_1 = \delta\sqrt{\pi_{U_1}} = 0.34$. First, we compare the performance of the five methods, LM, OGLASSO, CART, RF and SVM, by fitting the models using biomarkers X_1 and X_2. Predictions for LM, OGLASSO, CART, RF and SVM were obtained by using functions lm, grpregOverlap, rpart, randomForest, and svm in R with default parameters for all, except OGLASSO where $\tau = 0.025$.

Figure 2 shows the subgroups estimated using LM with unlimited sample size. Unlike CART, RF and SVM, with unlimited number of subjects, the linear model-based methods do not yield the perfect subgroup estimation. The corresponding %U in the six models for LM are 89, 79, 78, 75, 69 and 99 while for OGLASSO with $\tau = 0.025$ they are 84, 64, 68, 70, 69 and 99. Interestingly, for large sample sizes, as we increase τ , %U decreases, with values of 79, 52, 63, 66, 69 and 99 for $\tau = 0.05$. On the other hand, for small to moderate sample sizes, a linear model-based method yields good parameter estimates since the number of parameters to estimate is small. These can lead to better overall performance of the LM and OGLASSO for small and moderate sample sizes compared to more complex CART, RF and SVM. We simulated data based on the six models above for n = 400 subjects, 200 in treatment and 200 in control arm. Figure 3 shows the box plots for the distribution of %U based on 3000 simulation runs for methods LM,

OGLASSO, SVM, CART and RF. For models 2 and 4, the linear model performs the best followed by OGLASSO, SVM, CART and RF. In model 5, the approximation of the best subgroup by a linear model is poor, resulting in poor relative performance of the LM The OGLASSO method performs better than LM in models 5 and 6, and does similar to in models 1 and 3. Table 7 compares the variable selection capability of the methods LM and OGLASSO with 2 biomarkers present in the prediction model. We see that LM always chooses all the biomarkers present in the model. Thus, adding a penalty in the OGLASSO method, ensures a better prediction in the presence of noise biomarkers.

We conclude that linear model-based methods are the best to use for estimation of the subgroup based on two biomarkers both of which are associated with treatment response. The %U in all subjects generated from Models 1-6 are 71, 50, 63, 71, 67 and 100 respectively.

3.4.2 Comparison of designs

We consider a design that estimates the subgroup by maximizing U_2 after stage 1, such that assignment in stage 2 is inside the subgroup. Stage 3 is enriched based on subgroup estimated at the second interim analysis that maximizes U_1. The total sample size in the trial is n = 360 subjects with stage-wise sample sizes of $n_1 = n_2 = n_3 = 360/3 = 120$.

We used some of the models from Section 3.4.1 to simulate the design results, models D1-D5 below. In these change-point models, we use π_0 to denote the prevalence of the biomarker space where the treatment effect is higher, than the average treatment effect in all comers, Δ, can be computed as $\Delta = \mu + \theta\pi_0$. The prevalence of the best subgroup for U_1 and U_2 is shown in Table 8.

Model D1: Model 1 with $\mu = 0.05$, $\theta = 0.4$, c_1 = 0.4, yielding $\pi_0 = 0.6$ and Δ = 0.29.

Model D2: Model 4 with $\mu = 0.05$, $\theta = 0.35$, c = 0.85, yielding $\pi_0 = 0.64$ and Δ = 0.28.

Model D3: Model 2 with $\mu = 0.10$, $\theta = 0.55$, c_1 = 0.65, c_2 = 0.4, yielding $\pi_0 = 0.21$ and Δ = 0.21.

Model D4: Model 2 with $\mu = 0$, $\theta = 0.55$, c_1 = 0.32, c_2 = 0.32, yielding $\pi_0 = 0.46$ and Δ = 0.25.

Model D5: Model 2 with $\mu = 0$, $\theta = 0.30$, c_1 = 0 and c_2 = 0, yielding $\pi_0 = 1.00$ and Δ = 0.30.

We use the linear model-based methods from Section 3.2 to estimate the subgroup as they performed better than other methods (Section 3.4.1). We used two biomarkers such that either one or both of them were effect modifying in all true models considered. The total sample size, n = 360, were chosen to yield 80% - 90% power in the simulated trials. This corresponds to the average effect size of about 0.3. Because of enrichment, the average effect size in the patient population in our trial is higher than the average effect size in all patients. In the design, we consider futility stopping at the second interim analysis. The futility boundary $\tilde{b} = -1.64$ was computed to yield the probability of stopping for futility of 0.05 if the true effect size is 0.3.

Table 8 show the results for $\%U$ and power for testing in an enriched population for the design using linear model-based methods, LM and OGLASSO. For OGLASSO, we used $\tau = 0.05$ at interim 1 and $\tau = 0.05$ at interim 2.

The design using LM as a method of subgroup estimation at interims 1 and 2 vs OGLASSO performs similarly in terms of $\%U$ and power for models D1-D3, with LM performing better in D4 and OGLASSO performing the best in D5. OGLASSO tends to choose a much larger subgroup as compared to LM. To understand what $\%U$ metric means, we computed this measure for the subgroup that includes all patients. For scenario D1-D5, $\%U$ in all subjects is equal to 83%, 85%, 72%, 68% and 100% correspondingly. The median of the $\%U$ of the estimated subgroup is often not higher than the $\%U$ corresponding to the subgroup that includes all patients. However, the power in the estimated subgroup is much higher resulting in significant increase in power for

scenarios D1-D4 compared to a non-enriched trial. The estimation of the subgroup is satisfactory except for scenario D5, where there is no subgroup and the treatment effect is the same everywhere. $\%U$ for the estimated subgroup in this scenario for LM and OGLASSO is 87% and 95% implying that the median prevalence of the estimated subgroup is 75% and 91% correspondingly, lower than the true prevalence of 100%. The proposed design offers a reasonable balance between quality of subgroup estimation and power.

With a futility look added at the second interim, the power of the design decreases about 1%. The probability of stopping for futility, is in the range 4-11% under the alternative, while about 74% of trials are stopped for futility under the null hypothesis. Table 9 contains the type I error rates and the probability of stopping for futility under the null hypothesis.

Two slight variations of the design were also considered – first that maximizes U_1 at interim 1 and second that does not estimate a subgroup at interim 1, and enrolls everyone in stage 2. These designs did not perform very well compared our design (results are available from the authors).

3.5. Discussion

We considered the problem of estimating the best subgroup and testing for treatment effect in a clinical trial. The best subgroup was defined through maximizing the non-centrality parameter, utility $U(c,\gamma) = \pi(c)^{\gamma}\left[\mu_T(c) - \mu_C(c)\right]$. We introduced a metric $\%U$ to measure the quality of estimation of the subgroup. It is the % of the ratio of the utility in the estimated subgroup to the true utility of the underlying model. For several true models of response as a function of treatment and biomarkers, we compared four methods of estimation of the best subgroup, linear model, RF, CART and SVM. For moderate sample sizes, fitting a linear model-based method with main effect and first order pairwise interaction terms performed better that more complex methods such as RF,

CART and SVM. Using $\tau = 0.05$ in the OGLASSO method at interim 1 chooses a large subgroup at stage 1, but provides variable selection. As a result, the subject enrolled at stage 2 are based only on the important variables. We $\tau = 0.025$ at interim 2 to zero in on to the correct prevalence.

We propose a multi-stage enrichment design where subgroup is estimated at both interims 1 and 2. At the first interim analysis the subgroup is estimated by maximizing the utility U_2. The three-stage design we proposed can be used for initial assessment of efficacy of treatment that is not believed to be efficacious in all patients but might be efficacious in a subgroup of patients. If such a treatment is investigated in a trial with all comers, the efficacy signal will be diluted and might be missed. Adaptive enrichment allows the signal detection, even when the subgroups of patients for whom the treatment is working has rather small prevalence.

CHAPTER 4: FINDING A SUBGROUP WITH DIFFERENTIAL TREATMENT EFFECT WITH MUTIVARIATE OUTCOME

4.1 Introduction

Various methods to identify subject subpopulations with better outcomes compared to other subjects have been proposed (Song and Chi, 2007; Lipkovich et al, 2011; Loh 2011; Renfro 2016; Zhang et al, 2017, 2018). Most of the methods are for studies with a single outcome, binary, continuous or time-to-event. In some clinical trials, multivariate outcome is considered to evaluate the efficacy of a treatment, for example, treatment effect is evaluated using a subject reported outcome as well as a clinical endpoint. This helps to maximize the information to answer the question is the treatment works. If a subgroup of subjects exists that yields better outcomes, considering multiple outcomes, or a multivariate outcome, can be more powerful than estimating the subgroup based on a single outcome.

There are a handful of publications addressing the problem of finding the best subgroup based on multivariate outcome. Loh et al. (2016) extended a previously published method Generalized, Unbiased Interaction Detection and Estimation (GUIDE) (Loh et al., 2013) to multiple outcomes. GUIDE is a tree-based classification and regression method for subgroup estimation. It first tests each covariate for interaction with treatment and chooses the most significant covariate to split on by maximizing chi-squared test statistic. The spit that minimizes the sum of squares of residuals in the child node is chosen. In case of multiple outcomes GUIDE

is applied to each outcome at a time and the biomarker that maximized the sum of chi-squared values over all outcomes was chosen. In order to utilize potential correlation between the outcomes, Loh et al. (2016) suggested two approaches. In the first, the responses were replaced by their principal components or discriminant coordinates while choosing the biomarker to split on. The second method considered the fact that each outcome maybe on a different scale or are not all as equally important. In the former scenario, they were normalized and in the latter scenario, a weighted total sum of squares was used to search for best split of a covariate with user defined weights. Igor, Pu and Faltings (2018) considered a problem of finding the subgroup that reflects the tradeoff between favorable and unfavorable effects of a treatment. For example, two outcomes where the treatment improves the first outcome but worsens the second outcome. The best subgroup is defined as a set of subjects set who respond favorably as far as the first outcome without responding too poorly as far as the second outcome. The outcomes were assumed to follow a multivariate normal, and modeled given the baseline covariates, treatment and the cluster affiliation of a subject and incorporating the priors defined for the parameters involved.

Most of the subgroup estimation methods have been developed for parallel group trials where subjects are randomized to either treatment or control. We consider estimation the subgroup in a parallel group trial and, additionally in a setting of a two-period two-sequence crossover.

In this chapter, we define the best subgroup is the subgroup that maximizes the power to detect the treatment effect with respect to the outcome(s). We propose several approaches to estimating the best subgroup in case of a multivariate outcome. The approaches are compared via simulations in both parallel group and crossover design settings.

4.2 Subgroup Estimation for Multiple Outcomes in a parallel group trial

4.2.1 Setup

In a parallel trial setting, each subject is equally randomized either to the active treatment (T = "Active Treatment") or control (T = "Control"). Let $\boldsymbol{X} = (X_1, X_2, ..., X_M)$ be a vector of continuous biomarkers measured at baseline, scaled to [0,1]. We further assume that, as before, K continuous responses are measured on each subject, such that higher values indicate improvement in the well-being of a subject. Let $\mathbf{Y} = (Y_1, ..., Y_K)$ be the outcome vector. For the i^{th} randomized subject, the vector $(x_{i1}...x_{iM}, y_{i1}...y_{iK}, t_i)$, $i = 1...n$, represents the observed data, where $x_{i1}...x_{iM}$ are baseline biomarker values, $y_{i1}...y_{iK}$ are K outcome values and t_i is the treatment indicator. Subject's response might depend on treatment as well as subject's biomarkers with Lx1 mean vectors $\boldsymbol{\mu}_T(\boldsymbol{x}) = E(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x}, T = \text{"Active"})$ and $\boldsymbol{\mu}_C(\boldsymbol{x}) = E(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x}, T = \text{"Control"})$.

4.2.2 Methodology

Let $S \equiv S(\boldsymbol{X})$ is any subgroup defined using the values of the biomarker vector $\boldsymbol{X}$. We define the best subgroup, S^*, as the one that maximizes the utility $U(S, \gamma) = \pi(S)^{\gamma} \left[\mu_T(S) - \mu_C(S) \right]$ (defined earlier) over all possible subgroups S (Lai, Lavori and Liao, 2014; Joshi, Fine, Chu and Ivanova, 2019). Here, $\pi(S) = P(\boldsymbol{X} \in S)$ is the prevalence of the subgroup and $\mu_T(S)$ and $\mu_C(S)$ are the expected responses to the treatment and control in the subgroup.. In this Chapter, we will assume $\gamma = 0.5$ and maximize $U(S, 0.5) = \sqrt{\pi(S)} \left[\mu_T(S) - \mu_C(S) \right]$ and denoted it by U. This function is proportional to the power of treatment comparison or, equivalently, the non-centrality parameter in the test for the treatment effect.

We consider three methods of handling multiple outcomes in subgroup estimation for parallel groups. In the first method, for each subject, we define a single outcome $W_{avg} = (Y_1 + ... + Y_K)/K$; as the average of all outcomes. In the second method, a single outcome for the subject is defined as $W_{\max} = \max(Y_1, ..., Y_K)$. Here we assume that all the outcomes are of the same type and have the same scale, e. g. the vector of endpoints is a multivariate normal and all outcomes have the same variance. After the problem with multiple outcomes is converted to the problem with a single outcome, *W*, the subgroup is estimated by one of the methods developed for subgroup estimation based on a single outcome. We will omit the subscript of *W* in the reminder of this section as the method described below can be applied to W_{avg} or $W_{\max}$. Specifically, we fit a linear model with first order main effects for biomarker and all pairwise interaction terms between all available biomarkers. Let T^* be the treatment indicator with $T^* = 1$ for subjects randomized to treatment and $T^* = -1$ for subjects on control. We use the observation by Tian et al (2014) that treatment does not have to be in the model if the model is fit to *WT*, that is, outcomes of subjects who received the control treatment are multiplied by -1. Multiplying the responses of subjects on control are by -1,

$$E(T^*W \mid \boldsymbol{X}) = \alpha T^* + \sum_{m=1}^{M} \gamma_m X_m T^* + \sum_{m=1}^{M} \delta_m X_m T^{*2} + \sum_{l=1}^{M-1} \sum_{m=l+1}^{M} \lambda_{lm} X_l X_m T^{*2} \ .$$

Since T^* = {-1, 1} with probability 0.5, we can fit a model without treatment:

$$E(T^*W \mid \boldsymbol{X}) = \sum_{m=1}^{M} \delta_m X_m + \sum_{l=1}^{M-1} \sum_{m=l+1}^{M} \lambda_{lm} X_l X_m \ . \qquad (1)$$

After the coefficients are estimated from the data, we can compute the expected treatment response for the i^{th} subject, $i = 1...n$, with biomarker vector $\boldsymbol{X}_i$. Let vector $(\hat{w}_1, ..., \hat{w}_n)$ be the estimated treatment effects for the set of biomarker vectors $\{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n\}$ in the data set with

maximum responses and average response respectively. The definition of the best subgroup is the same as in earlier chapters. For each predicted $\hat{w}_i$, we define the corresponding subgroup as the set of subjects with predicted response larger than $\hat{w}_i$ that is, $\{\boldsymbol{X}_j : \hat{w}_j > \hat{w}_i\}$ and $\{\boldsymbol{X}_j : \hat{w}_j > \hat{w}_i\}$. Then, we compute the estimated prevalence, estimated treatment response and estimated utility of this subgroup. Denote the estimated treatment response for which the estimated utility is maximized by $w^*_{\max}$, $w^* = \hat{w}_i$, for some *i*, *i* = 1,…,*n*. The best subgroup using the maximum outcome $\{\boldsymbol{X}_j : \hat{w}_j > w^*\}$, includes subjects with the biomarker vector such that the predicted treatment response for that vector is higher than w^*. We refer to the first method of subgroup estimation as LM_{avg} and to the second as $LM_{\max}$.

The third method of subgroup estimation is based on the idea of covariate partitioning introduced by Guo, Ji and Catenacci (2016). A set of hyperplanes of the form: $a_1 X_1 + ... + a_M X_M = c$ with $a_1^2 + ... + a_M^2 = 1$divide the covariate space into two non-overlapping partitions. A set of pairs of partitions generated by the set of hyperplanes and the entire covariate space act as a candidate set of subgroups. We choose the one that minimizes the p-value for testing for the treatment effect in the subgroup based on *K* outcomes, or, equivalently, the one that maximizes power or maximizes *U*. The p-value is obtained using the method of choice to adjust for multiple comparisons. We refer to this method as the hyperplane method. A variation on the hyperplane method is also considered, incorporating variable selection in order to reduce the number of biomarkers used. This is based on least absolute shrinkage and selection operator method or LASSO (Tibshirani, 1996).

Consider a linear model of the form defined earlier in (1). In LASSO, the parameter coefficients are estimated by minimizing the residual sum of squares with an additive penalty term applied to sum of absolute values of the coefficients i.e.

$$\sum_{i=1}^{n}(T_i^{*}W_i - \sum_{m=1}^{M}\delta_m X_m - \sum_{l=1}^{M-1}\sum_{m=l+1}^{M}\lambda_{lm}X_l X_m)^2 + \beta\left(\sum_{m=1}^{M}|\delta_m| + \sum_{l=1}^{M-1}\sum_{m=l+1}^{M}|\lambda_{lm}|\right).$$

We modify this slightly by putting a penalty based on whether a biomarker is included or not, rather than on its coefficient. That is, we propose minimizing p-value + $\beta[I(a_1 \neq 0) + ... + I(a_M \neq 0)]$, such that $\beta \in (0,1)$. This small positive term adds a penalty if the hyperplane chosen to partition the covariate space includes more biomarkers. We refer to this method as the hyperplane with penalty method.

4.3 Subgroup Estimation with Multiple Outcomes in a crossover trial

Crossover designs are more efficient compared to parallel group trials with respect to power of the treatment comparison and subgroup estimation. Consider a two-period two-sequence crossover trial with possible sequences "TC" or "CT". Here we assume that there is no carry over effect and that we are dealing with a chronic disease so that a subject cannot be cured in period 1. As before $\boldsymbol{X} = (X_1, X_2, ..., X_M)$ is a vector of continuous biomarkers. In order to simplify notation, we will use Y_{ij} for the difference of the outcomes from the i^{th} subject on treatment and control $i = 1...n,\ j = 1...K$. The only difference from the methods in section 4.2 is that we fit a model to outcome *W* defined above instead of multiplying the outcome by *T*. Alternatively, one can use the same model as in section 4.2 and additionally model the correlation between the outcomes from the same subject. The hyperplane methods described in section 4.2.2 is applied similarly in a crossover setting.

4.4 Simulations

We compared the four subgroup estimation methods discussed in sections 2 and 3 in terms of $\%U$ for parallel and crossover trials. The metric $\%U$ was introduced in Joshi, Nguyen and Ivanova (2019) and is defined as the ratio of the value of the utility function in the estimated subgroup over the utility for the best subgroup. We use a sample size of n = 200 and n = 400 for crossover and parallel trials respectively and run 3000 simulated trials. We considered scenarios with three correlated continuous outcomes, K = 3. We consider two independent biomarkers X_1 and X_2 with uniform (0,1) distribution. The means of all three outcomes in control group are set to 0, $\mu_{kC} = 0$, $k = 1,2,3$. The $2K$x$2K$ variance-covariance matrix is defined with $corr(y_{kt}, y_{kC}) = \rho_1$, $corr(y_{kT}, y_{k'T}) = corr(y_{kC}, y_{k'C}) = \rho_2$, $k \neq k'$ and $corr(y_{lt}, y_{l'C}) = \rho_3$, $k \neq k'$ and σ^2 is the common, known variance. For the parallel group only non-zero entries for $corr(y_{kT}, y_{k'T}) = corr(y_{kC}, y_{k'C}) = \rho_2$. In the crossover trial, the $2K$x$2K$ variance-covariance matrix is defined with all non-zero correlations, $corr(y_{kT}, y_{k'T}) = corr(y_{kC}, y_{k'C}) = \rho_2$. We assume $\rho_1 = 0.5$, $\rho_2 = 0.3$, $\rho_3 = 0.15$ and $\sigma^2 = 1$. In order to implement the hyperplane method, we consider a set of pre-specified candidate lines (Figure 4). We use a Hochberg step-up method to get the adjusted p-values and set the overall p-value for treatment effect to equal to the minimum adjusted p-value.

The methods were compared using the following true models. In models 1, 4 and 7 all outcomes are associated with the treatment and have the same mean response. In models 2, 5 and 8 only two of the outcomes are associated with the treatment with the same mean response. In models 3, 6 and 9, only one outcome is associated with the treatment. For each model, the best

true subgroup S^* is specified along with its prevalence, π^*, mean treatment difference, δ^* and the true utility, $U = \delta^* \sqrt{\pi^*}$. The scenarios for data generation are listed below and also described in Supplemental Table 1, along with the corresponding true subgroup definitions. The scenarios for data generation are listed below.

Model 1: $\mu_{kT} = 0.5I(X_1 > 0.35)$, $k = 1,2,3$ with $S^* = \{X_1 > 0.35\}$, $\pi^* = 0.65$, $\delta^* = 0.5$ and $U = 0.4$.

Model 2: $\mu_{1T} = \mu_{2T} = 0.5I(X_1 > 0.35)$, $\mu_{3T} = 0$. True subgroup values same as Model 1.

Model 3: $\mu_{1T} = 0.5I(X_1 > 0.35)$, $\mu_{2T} = \mu_{3T} = 0$. True subgroup values same as Model 1.

Model 4: $\mu_{kT} = 0.5I(X_1 > 0.57)$, $k = 1,2,3$ with $S^* = \{X_1 > 0.57\}$, $\pi^* = 0.43$, $\delta^* = 0.5$ and $U = 0.33$.

Model 5: $\mu_{1T} = \mu_{2T} = 0.5I(X_1 > 0.57)$, $\mu_{3T} = 0$. True subgroup values same as Model 4.

Model 6: $\mu_{1T} = 0.5I(X_1 > 0.57)$, $\mu_{2T} = \mu_{3T} = 0$. True subgroup values same as Model 4.

Model 7 $\mu_{kT} = 0.5I(X_1 > 0.5 \,\&\, X_2 > 0.5)$, $k = 1,2,3$ with $S^* = \{X_1 > 0.5 \,\&\, X_2 > 0.5\}$, $\pi^* = 0.25$, $\delta^* = 0.5$, and $U = 0.25$.

Model 8: $\mu_{1T} = \mu_{2T} = 0.5I(X_1 > 0.5 \,\&\, X_2 > 0.5)$, $\mu_{3T} = 0$. True subgroup values same as Model 7.

Model 9: $\mu_{1T} = 0.5I(X_1 > 0.5 \,\&\, X_2 > 0.5)$, $\mu_{2T} = \mu_{3T} = 0$. True subgroup values same as Model 7.

Model 10: $\mu_{kT} = 0.3$, $k = 1,2,3$ with $S^* = \{X_1 > 0 \;\&\; X_2 > 0\}$, $\pi^* = 1.00$, $\delta^* = 0.3$ and $U = 0.3$.

Model 11: $\mu_{1T} = 0.3$, $\mu_{2T} = \mu_{3T} = 0$. True subgroup values same as Model 10.

Model 12: $\mu_{1T} = 0.3$. True subgroup values same as Model 10.

For models 1, 2 and 3, the utility computed in all subjects, U_{all} is 0.35. For models 4,5 and 6, U_{all} is 0.12, for models 7,8 and 9, U_{all} is 0.21 while for models 10,11 and 12 it is 0.3. The corresponding $\%U_{all}$ values are 87%, 48%, 64% and 100%.

Table 10 (more detailed results are present in Supplemental table 2) compares the three methods of subgroup estimation, LM_{avg} ,hyperplane and hyperplane with penalty based on $\%U$ for models 1-10 for a parallel trial with $n = 400$. For parallel trial, we do not run simulations for $LM_{\max}$, but only compare LM_{avg} and the two hyperplane methods. Table 11 (more detailed results are present in Supplemental table 3) compares $LM_{\max}$, LM_{avg} and hyperplane methods for a crossover trial setting with $n = 100$. We assume $\lambda = 10^{-5}$ as the penalty.

In Table 10, for models 1, 2 and 3, if the candidate lines coincide with the underlying subgroup, then the hyperplane methods do well, and are better compared to LM_{avg} . When the candidate lines do not coincide with the best subgroup boundary, the LM_{avg} does better as seen in models 4,5 and 6. Within the hyperplane and LM_{avg} methods, as expected, the performance becomes worse as the number of outcomes associated with treatment reduce. That is, $\%U$ in model 1 is better than that in model 2, which is better than $\%U$ in model 3; same trend continues for models 4,5 and 6 and models 7, 8, and 9. This decrease is larger in LM_{avg} than in hyperplane. When the best subgroup includes everyone, LM_{avg} does better than the hyperplane method. The $\%X$ column shows the percentage of trials where the hyperplane chosen uses the exact number of biomarkers used to define the true subgroup. For models 1-6, it is the % of times only X_1 was chosen to draw the hyperplane that minimizes the p-value; for models 7-9, $\%X$ has the % of

times both X_1 and X_2 are used to draw the hyperplane chosen. For models 10-12, it reports the % of times the entire biomarker space is chosen. Between the two hyperplane methods, the one with the penalty does almost as good or better than the one without penalty in terms of $\%U$ for all models. The penalty method either is equivalent or improves the performance of the hyperplane method with respect to variable selection, except for models 1-2.

We notice, in Table 11, the same trend when comparing LM_{avg} and hyperplane methods, as well as within each method as in Table 10. For models 3, 7 and 9, we expect $LM_{\max}$ to do better than LM_{avg}, but similar to the hyperplane method. But interestingly, the $LM_{\max}$ always estimates the best subgroup as everyone, which is unexpected and requires more investigation.

4.5 Discussion

In this chapter, we aim to establish a method for estimating a subgroup with higher treatment effect as compared to all, when there are multiple continuous correlated outcomes and multiple biomarkers available. We consider both a crossover and a parallel group trial setting. Based on our results, we conclude that in the case of multiple outcomes, drawing a single hyperplane to divide the covariate space performs better than prediction based on a linear model. The addition of penalty on the number of biomarkers used to define the hyperplane leads to better biomarker selection compared to the hyperplane method without penalty while keeping %U similar. Future work would involve using multiple outcomes of different data types. We could also consider more than one hyperplane for partitioning the subgroup in order to account for more complex underlying subgroup. These estimation methods can also be integrated into a design in order to develop an adaptive enrichment design.

CHAPTER 5: FUTURE RESEARCH

In this dissertation, we considered subgroup estimation in a randomized clinical trial designs, performed both post-hoc and prospectively. We also considered enrichment designs to increase power of the trial by prospectively enriching patient population during the trial.

In Chapter 2, we looked at how to estimate the subgroup and test for treatment effect in a post-hoc analysis of data from a clinical trial with a single predictive biomarker in all-comers. In Chapter 3, we compared several methods for subgroup estimation in the presence of multiple biomarkers and developed a three-stage design that enriches in the subgroup estimated at interim. Results showed that such a trial design will have higher power for testing for treatment effect, as compared to an all-comers trial. Chapter 4 considered the problem of subgroup estimation with multiple outcomes and multiple biomarkers available. Several approaches to handle multiple outcomes were compared for both parallel and crossover trial settings.

5.1 Limitations

We define the subgroup as the subgroup that maximizes a utility function U . This approach provides for a trade-off between the size of the subgroup and the treatment effect in the subgroup. A limitation of this approach and, hence, the methods we have developed for this approach in Chapters 2-4, is that the treatment effect in the estimated subgroup can be lower than the clinically meaningful treatment effect. As illustrated in Chapter 3, subgroup estimation methods, including variable selection methods, require sample sizes much larger than those used in clinical trials. Therefore, subgroup estimation performed during the trial might not yield an accurate subgroup. This is a major limitation of prospective enrichment in a clinical trial. In Chapter 4, we proposed a subgroup estimation method that is based on the smallest p-value

across multiple outcomes. Since we are essentially looking for the best outcome among all available outcomes, the method is not appropriate in cases if one of the treatments is harmful as measured by one or more of the outcomes and, at the same time, the treatment is rather efficacious as measured by one of the outcomes. That is, this method is not recommended in the case where multiple outcomes on a patient are negatively correlated.

5.2 Future Work

We illustrated proposed methods by simulations with continuous biomarkers and continuous outcomes. For Chapter 2, simulations assuming a binary or time-to-event outcome could be run to confirm the conclusions seen for continuous outcome. In Chapter 3, we could consider binary outcome to evaluate the methods (logistic regression instead of linear regression model etc.). Time-to-event data is challenging in trial with prospective enrichment as in Chapter 3 since the trial data may not have independent stage-wise data if the subjects have long follow-up times. In addition, we can determine the sample size while planning a trial based on the enrichment design described in section 3.3. The test-statistic defined to test the hypothesis of interest is $\tilde{Z} = w_1 Z_1 + w_2 Z_2 + w_3 Z_3$, where $w_k = \frac{n_1}{n_1 + n_2 + n_3}$, $k = 1, 2, 3$. For each stagewise statistic, $Z_k \sim N\left(\frac{n_k(\mu_{T,k} - \mu_{C,k})}{\sigma^2 / n_k + \sigma^2 / n_k}, 1 \right)$, where $\mu_{T,k}$ and $\mu_{C,k}$ are the mean response in the treated and control group respectively, at the k^{th} stage, n_k is the stagewise sample size and σ^2 is the common variance across treatment groups and stages. Since Z_k's are conditionally independent,

we get $\tilde{Z} \sim N\left(\sum_{k=1}^{K} \frac{w_k(\mu_{T,k} - \mu_{C,k})}{\sigma^2/n_k + \sigma^2/n_k}, 1\right)$, asymptotically. This can be used in the standard sample size formula.

For Chapter 4, one can develop a modification of the hyperplane method that estimates the best subgroup by maximizing the class of utilities U with various values of γ and not only $\gamma = 0.5$.

APPENDIX: FIGURES AND TABLES

Figure 1: Best Subgroup for models 1-6

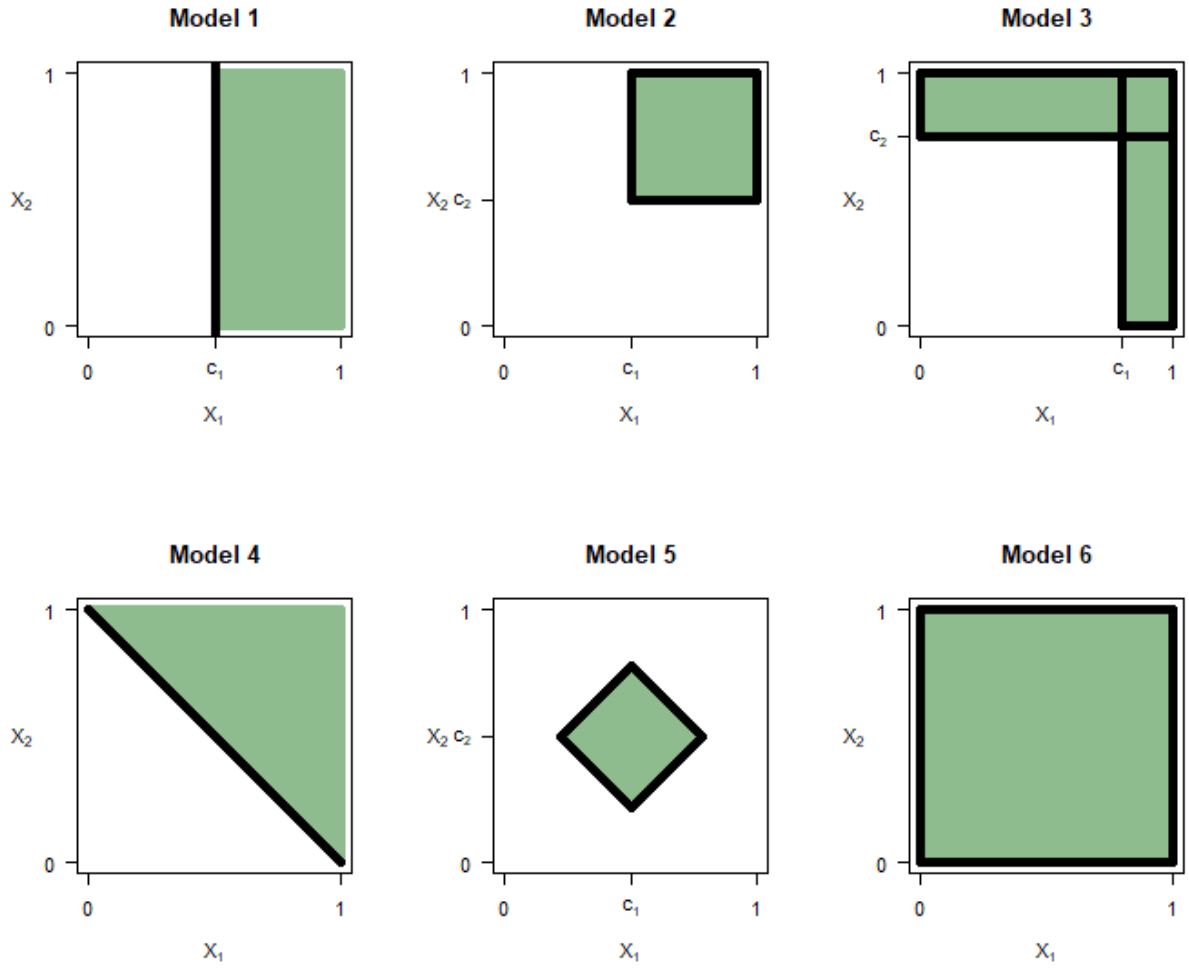
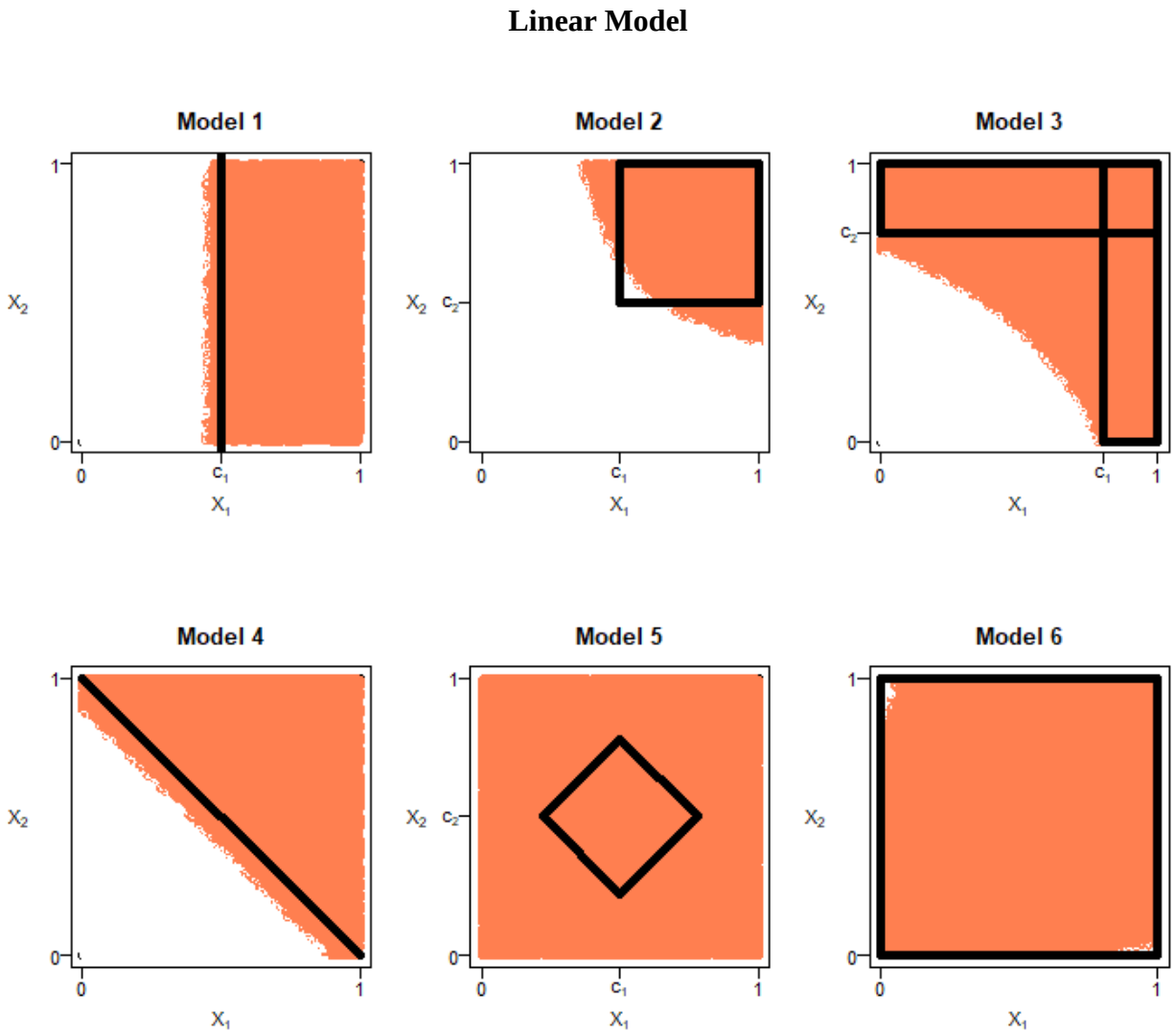


Figure 2: Large sample subgroup estimation for models 1-6 by linear model method with two biomarkers for LM and OGLASSO



OGLASSO

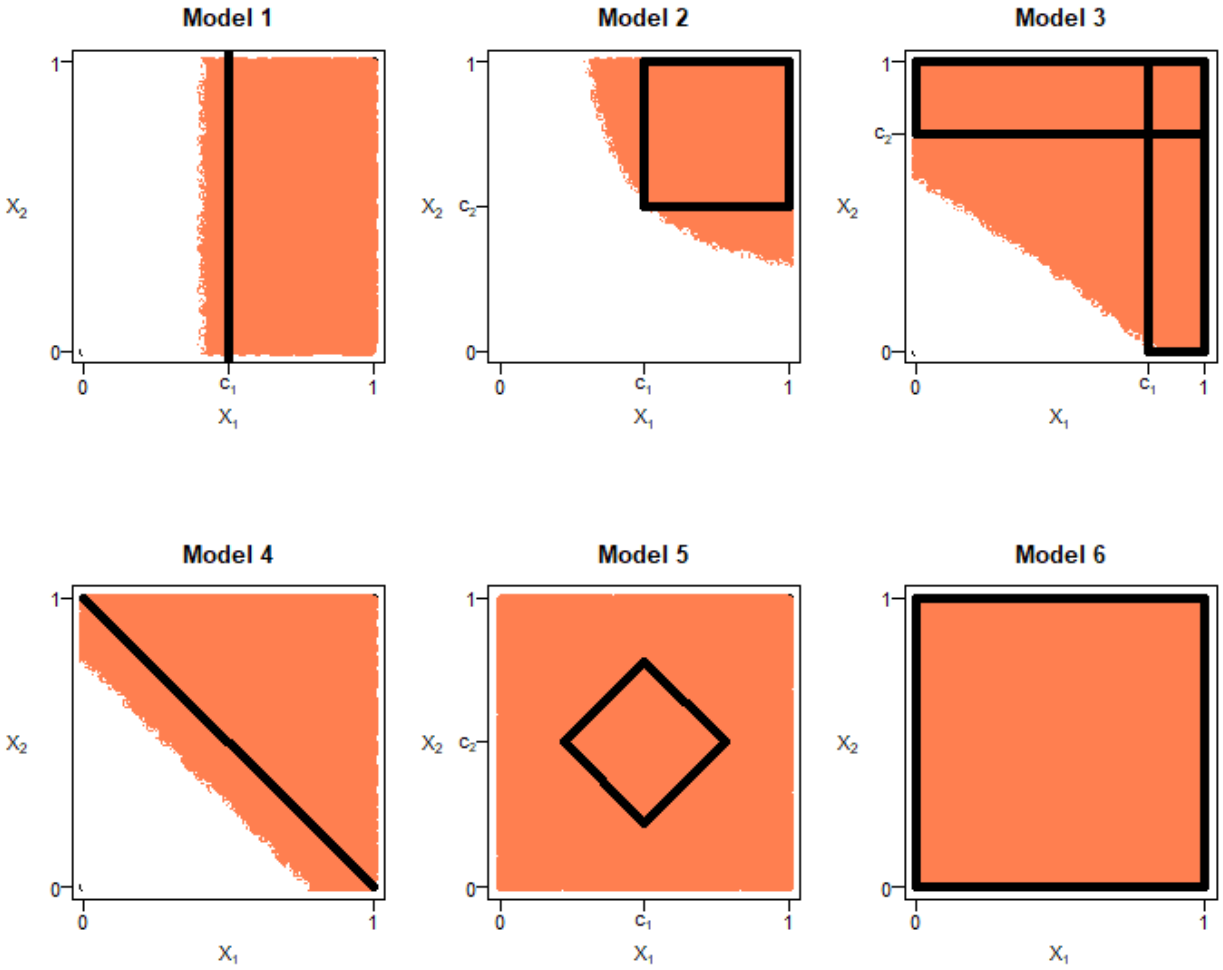


Figure 3 : Comparison of the Linear Model (LM), Overlapping Group LASSO (OGLASSO), Support Vector Machines (SVM), Classification and Regression Trees (CART) and Random Forests (RF) for subgroup estimation in a clinical trial with 400 patients using box plots for the distribution of %U. Horizontal line represents %U in all subjects.

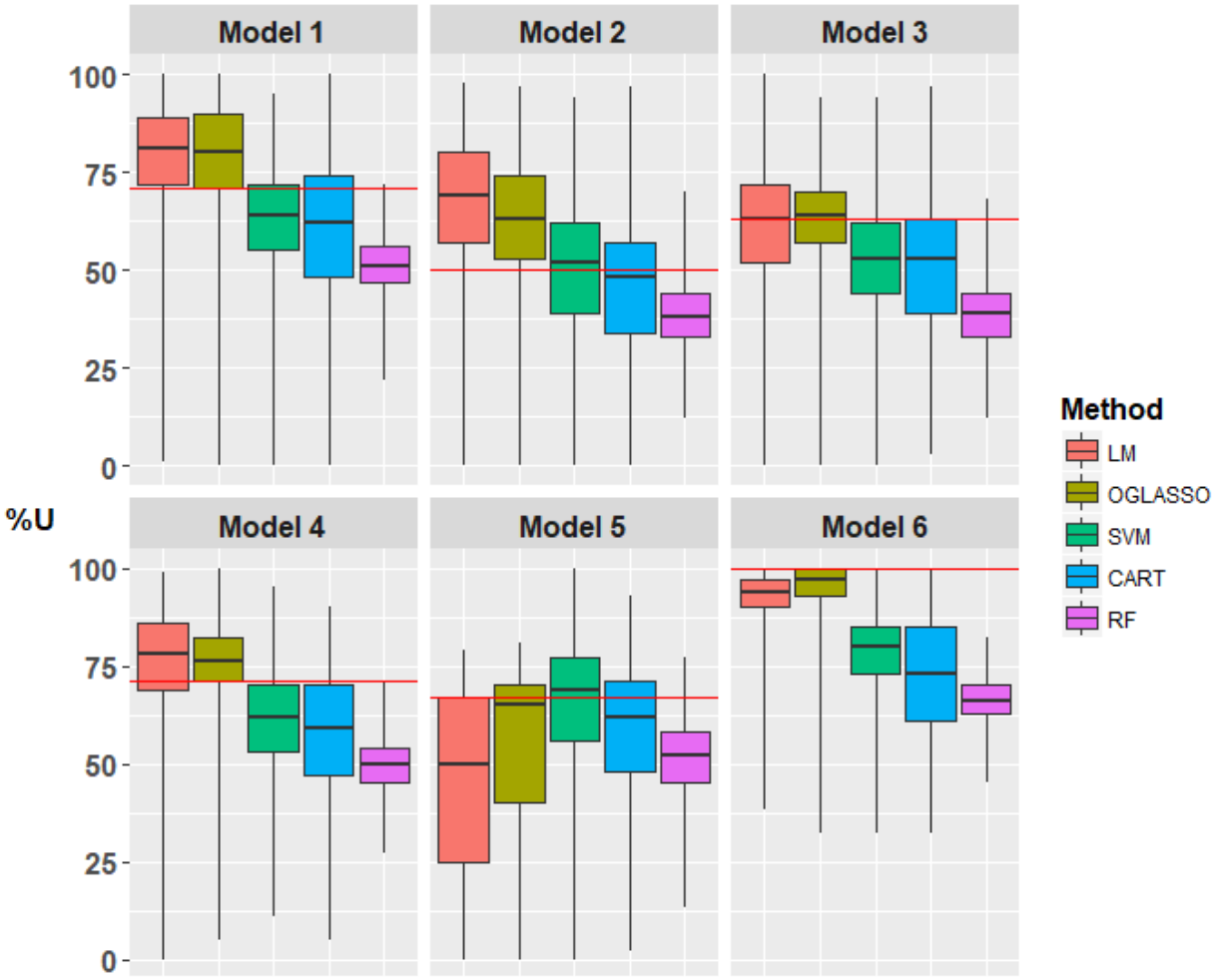


Figure 4: Candidate lines for subgroup estimation using the hyperplane method.

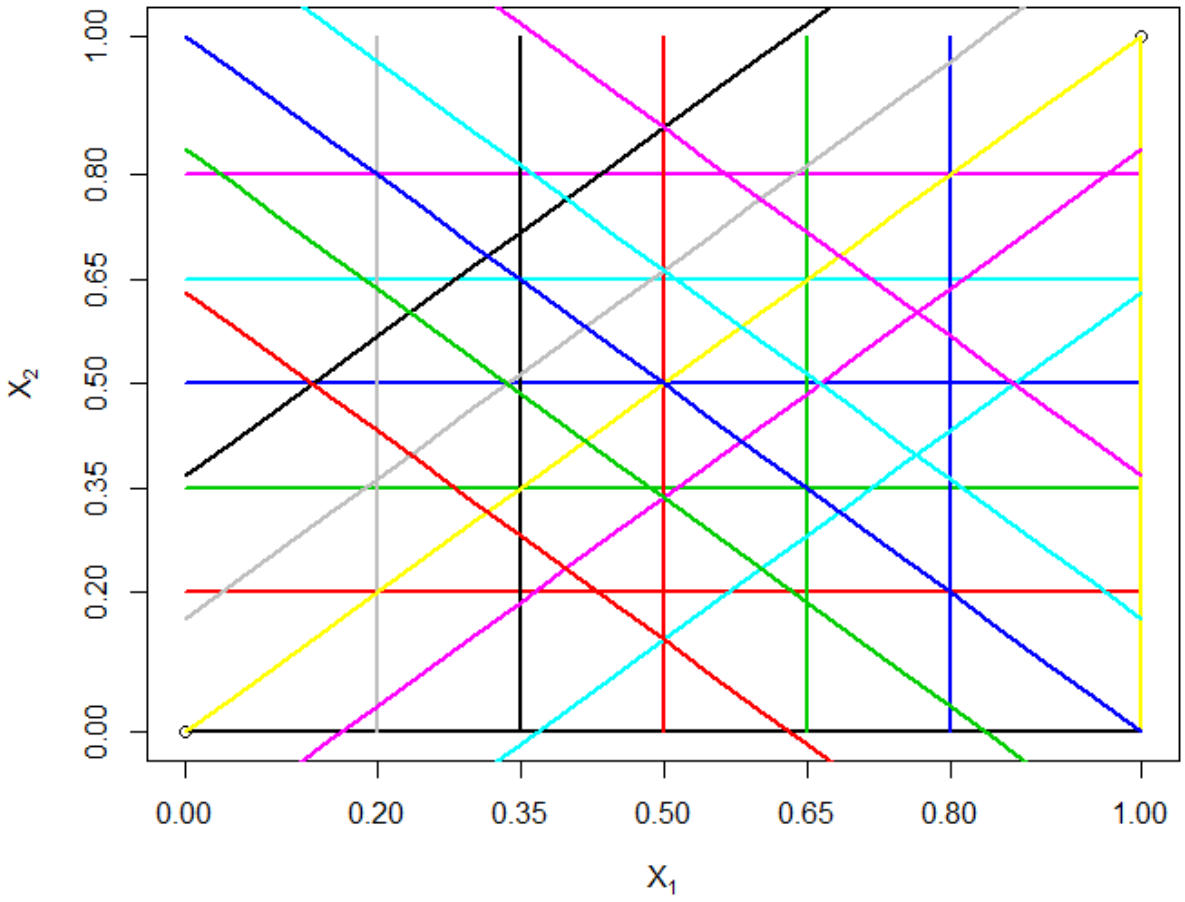


Table 1: Effect size in all (ES_{all}), effect size in the best subgroup (ES_S) and the prevalence of the best subgroup, π^*, corresponding to U_1 and U_2 for Model 3, $E[Y] = X^aT$.

		U_1		U_2	
a	ES_{all}	ES_S	π^*	ES_S	π^*
0.5	0.66	0.73	0.85	0.69	0.96
1	0.49	0.66	0.65	0.58	0.84
1.5	0.39	0.63	0.52	0.52	0.74
2	0.33	0.62	0.43	0.47	0.68

Table 2: Type I error rate where the best subgroup is estimated by maximizing utilities U_1 and U_2 with estimation by the non-parametric (NP) method and parametric (P) method based on linear model with interaction. The type I error rate is evaluated for tests Z_{All}, $Z_{All,S}$, $\tilde{Z}_S$, Z_S, and for the Hochberg (HC) procedure applied to Z_{All} and Z_S. $\tilde{Z}_S$ is a naïve test for the treatment effect in the subgroup that is not expected to preserve the type I error rate and Z_S is a permutation test in subgroup. The total sample size in the trial is 500 and the number of simulation runs is 10000.

Method	Z_{All}	$Z_{All,S}$	$\tilde{Z}_S$	Z_S	HC
Null scenario with no biomarker (X) or treatment effect $X \sim N(0,1)$					
U_1, NP	0.048	0.051	0.063	0.050	0.048
U_2, NP	0.048	0.051	0.062	0.050	0.047
U_1, P	0.048	0.050	0.059	0.052	0.048
U_2, P	0.048	0.049	0.058	0.051	0.045
Null scenario with no biomarker (X) or treatment effect $X \sim U(0,1)$					
U_1, NP	0.045	0.050	0.063	0.049	0.045
U_2, NP	0.045	0.050	0.063	0.050	0.046
U_1, P	0.050	0.049	0.053	0.049	0.047
U_2, P	0.050	0.051	0.053	0.053	0.047

Table 3: Change-point model with parameters δ, θ and π_0. Best subgroup is estimated by maximizing utilities U_1 and U_2 with estimation by the non-parametric (NP) method and parametric (P) method based on linear model with interaction. Column π^* shows the median, 25% and 75% for the prevalence of the estimated subgroup. Power is for tests Z_{All}, $Z_{All,S}$, Z_S, which is a permutation-based test of the treatment effect in the subgroup, and for the Hochberg (HC) procedure applied to Z_{All} and Z_S. The best power for each test $Z_{All,S}$, Z_S, and HC in each scenario is in bold.

θ	δ	π_0	Method	$\hat{\pi}^*$	Z_{All}	$Z_{All,S}$	Z_S	HC
0.10	0.28	0.40	True	$\pi^*_{U_1}=0.40$, $\pi^*_{U_2}=1$	0.66	0.72(U_1) 0.66(U_2)	0.77(U_1) 0.66(U_2)	-
			U_1, NP	0.54 (0.40, 0.66)	0.66	0.66	0.63	**0.69**
			U_2, NP	0.67 (0.55, 0.80)	0.66	**0.67**	0.64	0.68
			U_1, P	0.63 (0.50, 0.81)	0.66	0.65	0.61	0.65
			U_2, P	0. 81 (0. 69, 0.91)	0.66	**0.67**	**0.65**	0.66
0.03	0.35	0.50	True	$\pi^*_{U_1}=0.50$, $\pi^*_{U_2}=0.50$	0.63	0.75	0.85	-
			U_1, NP	0.52 (0.43, 0.64)	0.63	**0.68**	0.68	**0.70**
			U_2, NP	0.64 (0.53, 0. 75)	0.63	**0.68**	**0.70**	**0.70**
			U_1, P	0.58 (0.42, 0.74)	0.63	0.63	0.62	0.65

			U_2, P	0.77 (0.65, 0.87)	0.63	**0.68**	0.68	0.67
0.18	0.18	0.50	True	$\pi^*_{U_1}=1,\ \pi^*_{U_2}=1$	0.85	0.85	0.85	-
			U_1, NP	0.65 (0.52,0.78)	0.85	0.81	0.74	**0.83**
			U_2, NP	0.79 (0.67, 0.91)	0.85	0.83	0.77	**0.83**
			U_1, P	0.79 (0.69, 0.72)	0.85	0.81	0.75	0.82
			U_2, P	0.90 (0.80, 0.97)	0.85	**0.84**	**0.81**	**0.83**

Table 4: Bivariate normal model with parameters δ, ρ_T, ρ_C, σ^2_T. Best subgroup is estimated by maximizing utilities U_1 and U_2 with estimation by the non-parametric (NP) method and parametric (P) method based on linear model with interaction. Column π^* shows the median, 25% and 75% for the prevalence of the estimated subgroup. Power is for tests Z_{All}, $Z_{All,S}$, Z_S, which is a permutation based test of the treatment effect in the subgroup, and for the Hochberg (HC) procedure applied to Z_{All} and Z_S. The best power for each test $Z_{All,S}$, Z_S, and HC in each scenario is in bold.

δ	ρ_T	ρ_C	σ^2_T	Method	π^*	Z_{All}	$Z_{All,S}$	Z_S	HC
0.25	0.25	0.10	2	True	$\pi^*_{U_1} = 0.55$	0.63	0.74(U_1)	0.84(U_1)	-
					$\pi^*_{U_2} = 0.75$		0.72(U_2)	0.75(U_2)	
				U_1, NP	0.51 (0.39, 0.62)	0.63	0.71	0.75	**0.75**
				U_2, NP	0.64 (0.53, 0.76)	0.63	**0.72**	0.76	0.74
				U_1, P	0.45 (0.30, 0.60)	0.63	0.68	0.72	0.73
				U_2, P	0.68 (0.58, 0.78)	0.63	**0.72**	**0.77**	**0.75**
0.28	0	0	2	True	$\pi^*_{U_1} = 1.00$	0.72	0.72	0.72	-
					$\pi^*_{U_2} = 1.00$				
				U_1, NP	0.72 (0.59, 0.87)	0.72	0.63	0.52	0.65
				U_2, NP	0.85 (0.70, 0.94)	0.72	0.66	0.57	0.65

				U_1, P	0.88 (0.66, 0.99)	0.72	0.67	0.59	0.66
				U_2, P	0.95 (0.84, 0.99)	0.72	**0.69**	**0.64**	**0.67**
0.25	0.20	0.20	1	True	$\pi^*_{U_1} = 1.00$	0.80	0.80	0.80	-
					$\pi^*_{U_2} = 1.00$				
				U_1, NP	0.67 (0.54, 0.85)	0.80	0.72	0.60	0.75
				U_2, NP	0.84 (0.67, 0.94)	0.80	0.75	0.67	**0.76**
				U_1, P	0.89 (0.73, 0. 99)	0.80	0.74	0.68	0.74
				U_2, P	0.96 (0.87,0.99)	0.80	**0.77**	**0.73**	**0.76**

Table 5: A linear model with interaction $E[Y] = X^{a}T$, total sample size of N. Best subgroup is estimated by maximizing utilities U_1 and U_2 with estimation by the non-parametric (NP) method and parametric (P) method based on linear model with interaction. Column π^* shows the median, 25% and 75% for the prevalence of the estimated subgroup. Power is for tests Z_{All}, $Z_{All,S}$, Z_S, which is a permutation based test of the treatment effect in the subgroup, and for the Hochberg (HC) procedure applied to Z_{All} and Z_S. The best power for each test $Z_{All,S}$, Z_S, and HC in each scenario is in bold.

a	*N*	Method	π^*	Z_{All}	$Z_{All,S}$	Z_S	HC
1	100	True	$\pi^*_{U_1} = 0.66$	0.70	0.74(U_1)	0.77(U_1)	-
			$\pi^*_{U_2} = 0.84$		0.73(U_2)	0.72(U_2)	
		U_1, NP	0.67 (0.56, 0.76)	0.70	**0.74**	0.69	**0.73**
		U_2, NP	0.73 (0.63, 0.83)	0.70	**0.74**	0.69	0.72
		U_1, P	0.68 (0.56, 0.79)	0.70	0.73	0.70	**0.73**
		U_2, P	0.78 (0.69, 0.88)	0.70	**0.74**	**0.72**	**0.73**
1.5	156	True	$\pi^*_{U_1} = 0.53$	0.70	0.76(U_1)	0.83(U_1)	-
			$\pi^*_{U_2} = 0.74$		0.75(U_2)	0.74(U_2)	
		U_1, NP	0.54 (0.43, 0.66)	0.70	**0.73**	0.69	0.73
		U_2, NP	0.65 (0.54, 0.77)	0.70	**0.73**	0.70	0.72

		U_1, P	0.58 (0.44, 0.71)	0.70	0.72	0.70	**0.74**
		U_2, P	0.74 (0.64, 0.84)	0.70	**0.73**	**0.74**	0.73
2	228	True	$\pi^*_{U_1} = 0.44$	0.71	0.80(U_1)	0.88(U_1)	-
			$\pi^*_{U_2} = 0.68$		0.77(U_2)	0.76(U_2)	
		U_1, NP	0.47 (0.36, 0.60)	0.71	**0.76**	0.76	**0.78**
		U_2, NP	0.61 (0.48, 0.73)	0.71	**0.76**	0.76	0.76
		U_1, P	0.51 (0.37, 0.64)	0.71	0.74	0.75	0.77
		U_2, P	0.71 (0.61, 0.80)	0.71	**0.76**	**0.79**	**0.78**

Table 6: Data analysis of a phase 2 study of 1C4D4 in patients with metastatic pancreatic cancer. Best subgroup is selected based on utilities U_1 and U_2 with estimation by the non-parametric (NP) approach using the logrank test and parametric approach (P) by fitting a Cox model with interaction. The adjusted Hochberg p-value is to test the intersection hypothesis of no treatment effect in all and in the subgroup.

Method	Cutoff	Prevalence of the estimated subgroup with H-SCORE > cutoff	Median survival in 1C4D4 arm	Median survival in control arm	P-value in the subgroup	Hochberg p-value
U_1, NP	0.5	0.78	8.08	5.03	0.50	0.38
U_2, NP	0.5	0.78	8.08	5.03	0.50	0.38
U_1, P	55.5	0.66	9.17	5.52	0.57	0.38
U_2, P	55.5	0.66	9.17	5.52	0.57	0.38
All patients	0	1	7.92	5.52	0.19	-

Table 7 : Proportion of times covariates were selected using the overlapping group LASSO method for Models 1-6 using 2-biomarkers. Predictive column is the proportion of trials where the right set of predictive biomarkers were selected. Subset is the proportion of trials when an exact subset of the right set of predictive biomarkers were selected. Noise is the proportion to trials when at least one noise biomarker was selected. No Biomarker is the proportion of trials where no biomarker (predictive or noise) was selected.

Method	Models	Predictive Set	Subset of the Predictive Set	Noise	No Biomarker
LM	**1**	0	-	1	0
	2	1	0	-	0
	3	1	0	-	0
	4	1	0	-	0
	5	1	0	-	0
	6	1	0	-	0
OGLASSO	1	0.16	-	0.82	0.02
	2	0.81	0.15	-	0.04
	3	0.76	0.19	-	0.05

4	0.80	0.18	-	0.03
5	0.69	0.23	-	0.07
6	0.78	0.17	-	0.04

Table 8 : Results for comparing design U_2U_1 using LM and OGLASSO for models D1-D5. Column π shows the true subgroup prevalence in the first row, and the median, 25% and 75% for the prevalence of the estimated subgroup at the second interim analysis for each of the designs. Column %*U* shows the median, 25% and 75% of the percentage of the true utility estimated in the design at the second interim analysis. Power is for tests Z_{All}, based on all subjects enrolled and Z'. The power incorporating the single futility look at the second interim analysis is given by Z'_f. The proportion of trials stopped for futility is given by p_f. Total sample size used is *n* = 360. The best power and %*U* for the best method for each model is in bold.

Model	π	%*U*	Z'	Z'_f	p_f
D1, $\pi_{U_1} = 0.60$, $\pi_{U_2} = 0.60$, Z_{All} = **0.78**					
LM	0.70 (0.55, 0.83)	**84 (77, 89)**	0.87	0.87	0.04
OGLASSO	0.80 (0.62, 0.99)	**84 (81, 88)**	**0.88**	0.87	0.04
D2, $\pi_{U_1} = 0.64$, $\pi_{U_2} = 0.64$, Z_{All} = **0.73**					
LM	0.69 (0.54, 0.82)	84 (74, 90)	**0.82**	0.82	0.05
OGLASSO	0.81 (0.60, 0.99)	**85 (80, 88)**	0.81	0.80	0.06
D3, $\pi_{U_1} = 0.21$, $\pi_{U_2} = 1.00$, Z_{All} = **0.53**					
LM	0.58 (0.41, 0.74)	**75 (71, 79)**	0.66	0.65	0.10
OGLASSO	0.67 (0.44, 0.94)	74 (71, 77)	**0.68**	0.67	0.11

D4, $\pi_{U_1} = 0.46$, $\pi_{U_2} = 0.46$, Z_{All} **= 0.66**

LM	0.64 (0.48, 0.79)	**75 (70, 80)**	**0.86**	0.85	0.05
OGLASSO	0.73 (0.54, 0.94)	73 (68, 79)	0.83	0.82	0.05

D5, $\pi_{U_1} = 1.00$, $\pi_{U_2} = 1.00$, Z_{All} **= 0.81**

LM	0.75 (0.57, 0.89)	87 (76, 94)	0.81	0.81	0.05
OGLASSO	0.91 (0.69, 1.00)	**95 (83, 100)**	**0.82**	0.81	0.05

Table 9 : Type I error rate for the enrichment design for tests Z' and Z'_f. The proportion of trials stopped for futility, p_f are also presented. Total sample size in the trial is n = 360.

Z'	Z'_f	p_f
0.051	0.022	0.74

Table 10 : Comparison for four methods: LM_{max}, LM_{avg}, hyperplane and hyperplane with penalty in terms of $\%U$, and estimated prevalence $\hat{\pi}^*$ in a parallel trial. Median values are reported. $\%X$ denotes the proportion of trials for which the right set of biomarkers were used to draw the hyperplane. The $\%X$ values for scenarios where the best subgroup is based both biomarkers is in ().

	LM_{avg}		Hyperplane			Hyperplane with penalty		
Model	$\%U$	$\hat{\pi}^*$	$\%U$	$\hat{\pi}^*$	$\%X$	$\%U$	$\hat{\pi}^*$	$\%X$
1	89	0.66	**91**	0.65	68	88	0.65	61
2	85	0.64	**90**	0.65	66	89	0.65	62
3	74	0.52	**89**	0.65	62	88	0.65	60
4	85	0.52	82	0.50	60	**87**	0.50	65
5	79	0.50	81	0.50	60	**83**	0.50	63
6	68	0.42	75	0.50	54	**75**	0.50	56
7	**78**	0.35	70	0.35	(48)	70	0.35	(46)
8	**71**	0.36	68	0.35	(45)	68	0.35	(44)
9	**58**	0.33	61	0.35	(41)	61	0.35	(40)
10	**94**	0.88	86	0.80	12	88	0.80	30

11	74	0.54	82	0.65	8	**83**	0.65	16
12	**89**	0.79	82	0.65	8	83	0.65	15

Table 11 : Comparison for four methods: LM_{max}, LM_{avg}, hyperplane and hyperplane with penalty in terms of $\%U$, and estimated prevalence $\hat{\pi}^*$ in a crossover trial. Median values are reported. $\%X$ denotes the proportion of trials for which the right set of biomarkers were used to draw the hyperplane. The $\%X$ values for scenarios where the best subgroup is based both biomarkers is in ( ).

	LM_{avg}		LM_{max}		Hyperplane			Hyperplane with penalty		
Model	**$\%U$**	$\hat{\pi}^*$	**$\%U$**	$\hat{\pi}^*$	**$\%U$**	$\hat{\pi}^*$	$\%X$	**$\%U$**	$\hat{\pi}^*$	$\%X$
1	89	0.66	82	0.98	**90**	0.65	66	89	0.65	62
2	85	0.63	81	0.98	**90**	0.65	64	89	0.65	61
3	73	0.53	81	0.99	**88**	0.65	60	**88**	0.65	60
4	**86**	0.51	67	0.96	80	0.50	59	82	0.50	62
5	79	0.50	67	0.98	76	0.50	56	**80**	0.50	59
6	67	0.42	66	0.98	**74**	0.50	52	**74**	0.50	53
7	**79**	0.34	51	0.97	69	0.35	(43)	69	0.35	(43)
8	**71**	0.36	51	0.98	64	0.35	(41)	64	0.35	(40)
9	58	0.33	51	0.98	**60**	0.35	(37)	**60**	0.35	(37)
10	**94**	0.88	99	0.99	88	0.80	13	89	0.80	29

11	74	0.54	99	0.99	82	0.65	9	**83**	0.65	15
12	**88**	0.78	88	0.78	82	0.65	9	82	0.65	15

Supplemental Table 1: Models used to generate data to compare the subgroup estimations methods. #Outcome represents the number of outcomes associated with the treatment and $\%U_{all}$ is the $\%U$ value including all subjects and. S^* is the true subgroup, δ^* is the true treatment difference and π^* is the underlying prevalence of the subgroup. U is the utility value in the model.

Model	#Outcome	$\%U_{all}$	S^*	δ^*	π^*	U
1	3					
2	2	87	$X_1 > 0.35$	0.50	0.65	0.40
3	1					
4	3					
5	2	48	$X_1 > 0.57$	0.50	0.43	0.33
6	1					
7	3					
8	2	64	$X_1 > 0.50 \,\&\, X_2 > 0.50$	0.50	0.25	0.25
9	1					
10	3					
11	1	100	$X_1 > 0 \,\&\, X_2 > 0$	0.30	1.00	0.30
12	1					

Supplemental Table 2: Comparison for three methods LM_{avg} , hyperplane and hyperplane with penalty in terms of $\%U$, estimated prevalence $\hat{\pi}^*$ and mean treatment difference $\hat{\delta}^*$ in a parallel group trial. Median, 25th and 75th percentile reported.

	LM_{avg}			Hyperplane		
	%*U*	$\hat{\pi}^*$	$\hat{\delta}^*$	**%*U***	$\hat{\pi}^*$	$\hat{\delta}^*$
1	89 (85, 93)	0.66 (0.56, 0.77)	0.45 (0.40, 0.49)	91 (82, 100)	0.65 (0.65, 0.65)	0.49 (0.41, 0.50)
2	85 (79, 89)	0.64 (0.51, 0.77)	0.43 (0.39, 0.49)	90 (82,100)	0.65 (0.65, 0.65)	0.49 (0.41, 0.50)
3	74 (60, 82)	0.52 (0.35, 0.70)	0.40 (0.34, 0.49)	89 (80, 100)	0.65 (0.50, 0.65)	0.49 (0.40, 0.50)
4	85 (79, 91)	0.52 (0.45, 0.67)	0.40 (0.34, 0.44)	82 (71, 91)	0.50 (0.35, 0.50)	0.42 (0.33, 0.48)
5	79 (71, 87)	0.50 (0.39, 0.64)	0.38 (0.30, 0.45)	81 (70, 91)	0.50 (0.35, 0.50)	0.42 (0.32, 0.48)
6	68 (54, 79)	0.42 (0.27, 0.60)	0.34 (0.24, 0.44)	75 (68, 91)	0.50 (0.35, 0.50)	0.42 (0.30, 0.47)
7	78 (67, 85)	0.35 (0.26, 0.50)	0.33 (0.24, 0.44)	70 (59, 80)	0.35 (0.20, 0.50)	0.25 (0.19, 0.34)
8	71 (59, 82)	0.36 (0.25, 0.54)	0.28 (0.19, 0.39)	68 (50, 79)	0.35 (0.20, 0.50)	0.25 (0.16, 0.34)
9	58 (30, 73)	0.33 (0.22, 0.52)	0.21 (0.12, 0.33)	61 (35, 78)	0.35 (0.20, 0.50)	0.24 (0.12, 0.33)
10	94 (88, 98)	0.88 (0.78, 0.97)	0.30 (0.30, 0.30)	86 (76, 90)	0.80 (0.65, 0.80)	0.30 (0.29, 0.30)
11	74 (59, 85)	0.54 (0.35, 0.71)	0.30 (0.29, 0.31)	82 (70, 90)	0.65 (0.50, 0.80)	0.30 (0.29, 0.30)
12	89 (81, 95)	0.79 (0.67, 0.90)	0.30 (0.30, 0.30)	82 (71, 90)	0.65 (0.50, 0.80)	0.30 (0.29, 0.30)

Hyperplane with penalty

Model	**%*U***	$\hat{\pi}^*$	$\hat{\delta}^*$
1	88 (80, 100)	0.65 (0.65, 1.00)	0.49 (0.33, 0.50)
2	89 (80, 100)	0.65 (0.65, 0.80)	0.50 (0.34, 0.50)
3	88 (80, 100)	0.65 (0.65, 0.80)	0.49 (0.38, 0.50)
4	87 (70, 92)	0.50 (0.35, 0.50)	0.43 (0.33, 0.49)
5	83 (70, 91)	0.50 (0.35, 0.50)	0.42 (0.33, 0.49)
6	75 (67, 91)	0.50 (0.35, 0.50)	0.42 (0.30, 0.48)
7	70 (55, 79)	0.35 (0.20, 0.50)	0.25 (0.19, 0.34)
8	68 (50, 79)	0.35 (0.20, 0.50)	0.25 (0.16, 0.34)
9	61 (35, 77)	0.35 (0.20, 0.50)	0.24 (0.12, 0.33)
10	88 (78, 98)	0.80 (0.65, 1.00)	0.30 (0.29, 0.30)
11	83 (70, 91)	0.65 (0.50, 0.80)	0.30 (0.29, 0.30)
12	83 (71, 91)	0.65 (0.50, 0.80)	0.30 (0.29, 0.30)

Supplemental Table 3: Comparison for four methods: LM_{max}, LM_{avg}, the hyperplane method and the hyperplane with penalty in terms of $\%U$, estimated prevalence $\hat{\pi}^*$ and mean treatment difference $\hat{\delta}^*$ in a crossover trial. Median, 25th and 75th percentile reported.

	LM_{avg}			LM_{max}		
	%*U*	$\hat{\pi}^*$	$\hat{\delta}^*$	**%*U***	$\hat{\pi}^*$	$\hat{\delta}^*$
1	89 (85, 92)	0.66 (0.56, 0.76)	0.45 (0.40, 0.49)	82 (81, 83)	0.98 (0.95, 0.99)	0.33 (0.33, 0.34)
2	85 (79, 89)	0.63 (0.51, 0.76)	0.43 (0.38, 0.49)	81 (80, 83)	0.98 (0.96, 0.99)	0.33 (0.33, 0.34)
3	73 (60, 82)	0.53 (0.36, 0.69)	0.40 (0.34, 0.49)	81 (80, 82)	0.99 (0.97, 0.99)	0.33 (0.32, 0.33)
4	86 (79, 91)	0.51 (0.44, 0.61)	0.40 (0.34, 0.45)	67 (66, 69)	0.96 (0.91, 0.99)	0.22 (0.22, 0.33)
5	79 (71, 88)	0.50 (0.39, 0.63)	0.38 (0.31, 0.45)	67 (65, 68)	0.98 (0.94, 0.99)	0.22 (0.21, 0.33)
6	67 (54, 78)	0.42 (0.28, 0.49)	0.33 (0.25, 0.43)	66 (65, 67)	0.98 (0.96, 0.99)	0.22 (0.21, 0.22)
7	79 (67, 85)	0.34 (0.26, 0.49)	0.34 (0.24, 0.41)	51 (50, 53)	0.97 (0.93, 0.99)	0.13 (0.12, 0.13)
8	71 (58, 82)	0.36 (0.26, 0.55)	0.28 (0.19, 0.39)	51 (50, 52)	0.98 (0.95, 0.99)	0.13 (0.12, 0.13)
9	58 (30, 73)	0.33 (0.24, 0.53)	0.20 (0.12, 0.33)	51 (49, 52)	0.98 (0.96, 0.99)	0.13 (0.12, 0.13)
10	94 (88, 98)	0.88 (0.78, 0.96)	0.30 (0.30, 0.30)	99 (98, 100)	0.99 (0.98, 1.00)	0.30 (0.20, 0.30)
11	74 (59, 84)	0.54 (0.35, 0.71)	0.30 (0.30, 0.30)	99 (98, 100)	0.99 (0.97, 0.99)	0.30 (0.30, 0.30)
12	88 (81, 95)	0.78 (0.66, 0.89)	0.30 (0.30, 0.30)	88 (81, 95)	0.78 (0.66, 0.89)	0.30 (0.30, 0.30)

	Hyperplane			Hyperplane with penalty		
Model	**%*U***	$\hat{\pi}^*$	$\hat{\delta}^*$	**%*U***	$\hat{\pi}^*$	$\hat{\delta}^*$
1	90 (82, 100)	0.65 (0.65, 0.80)	0.50 (0.40, 0.50)	89 (81, 100)	0.65 (0.65, 0.80)	0.49 (0.33, 0.50)
2	90 (82, 100)	0.65 (0.65, 0.80)	0.50 (0.40, 0.50)	89 (81, 100)	0.65 (0.65, 0.80)	0.49 (0.38, 0.50)
3	88 (79, 100)	0.65 (0.50, 0.65)	0.49 (0.40, 0.50)	88 (79, 100)	0.65 (0.65, 0.80)	0.49 (0.38, 0.50)
4	80 (71, 91)	0.50 (0.35, 0.65)	0.42 (0.32, 0.46)	82 (71, 91)	0.50 (0.35, 0.65)	0.42 (0.32, 0.48)
5	76 (70, 91)	0.50 (0.35, 0.65)	0.42 (0.31, 0.46)	80 (69, 91)	0.50 (0.35, 0.65)	0.42 (0.31, 0.46)
6	74 (67, 91)	0.50 (0.35, 0.65)	0.39 (0.29, 0.44)	74 (67, 91)	0.50 (0.35, 0.65)	0.39 (0.29, 0.45)
7	69 (55, 79)	0.35 (0.20, 0.50)	0.25 (0.16, 0.34)	69 (66, 79)	0.35 (0.20, 0.50)	0.25 (0.17, 0.34)
8	64 (48, 79)	0.35 (0.20, 0.50)	0.25 (0.15, 0.33)	64 (48, 79)	0.35 (0.20, 0.50)	0.25 (0.15, 0.33)
9	60 (35, 74)	0.35 (0.20, 0.50)	0.24 (0.12, 0.27)	60 (34, 78)	0.35 (0.20, 0.50)	0.24 (0.12, 0.26)
10	88 (78, 90)	0.80 (0.65, 0.80)	0.30 (0.30, 0.30)	89 (79, 98)	0.80 (0.65, 1.00)	0.30 (0.30, 0.30)
11	82 (71, 90)	0.65 (0.50, 0.80)	0.30 (0.30, 0.30)	83 (71, 90)	0.65 (0.50, 0.80)	0.30 (0.30, 0.30)
12	82 (71, 90)	0.65 (0.50, 0.80)	0.30 (0.30, 0.30)	82 (71, 90)	0.65 (0.50, 0.80)	0.30 (0.30, 0.30)

REFERENCES

Alosh, M., Huque, M. F. (2009). A flexible strategy for testing subgroups and overall population. *Statistics in Medicine* 28: 3-23.

Arnold, B. C., Beaver, R. J., Groeneveld, R. A., Meeker, W. Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika* 58: 471.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). Classification and Regression Trees. *Chapman & Hall/CRC*.

Breiman, L. (2001). Random forests. *Machine Learning* 45:5–32.

Chen, Y. C., Lee, U. J., Tsai, C. A., Chen, J. J. (2018). Development of predictive signatures for treatment selection in precision medicine with survival outcomes. *Pharmaceutical statistics*, *17*(2), 105-116.

Cortes, C., Vapnik, V. (1995). Support-vector network. *Machine Learning* 20:1–25.

Diao, G., Dong, J., Zeng, D., Ke, C., Rong, A., Ibrahim, J.G. (2019). Biomarker threshold adaptive designs for survival endpoints. *Journal of Biopharmaceutical Statistics* in press.

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

Freidlin, B., Simon, R. (2005). Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 11: 7872–7878.

Freidlin, B., Jiang, W., Simon, R. (2010). The cross-validated adaptive signature design. *Clinical Cancer Research* 16(2): 691–698.

Guo, W., Ji, Y., & Catenacci, D. (2016). A subgroup cluster-based Bayesian adaptive design for precision medicine. *Biometrics*, 73(2), 367-377.

Jenkins, M., Stone, A., Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* 10: 347–56.

Jiang, W., Freidlin, B., Simon, R. (2007). Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute* 99: 1036–1043.

Kehl, V., Ulm, K. (2006). Responder identification in clinical trials with censored data. *Computational Statistics & Data Analysis* 50: 1338–1355.

Kulev I., Pu P., and Faltings B. (2018). A Bayesian Approach to Intervention-Based Clustering. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 9, 4, Article 44.

Lai, T., Lavori, P., Liao, O. (2014). Adaptive choice of patient subgroup for comparing two treatments. *Contemporary Clinical Trials* 39(2): 191-200.

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.

Liu, A., Li, Q., Liu, C., Yu, K. F., Yuan, V. W. (2010). A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. *Clinical Trials (London, England)* 7(5) 537–545.

Loh, W.Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 1:14-23.

Loh, W.Y., He, X., Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine* 34:1818–1833.

Loh, W.-Y., Fu, H., Man, M., Champion, V. and Yu, M. (2016), Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables, *Statistics in Medicine*, vol. 35, 4837-4855.

Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., Posch, M. (2015). Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics* 26(1): 99-119.

Pei H., Lai, T., Liao, O. (2012). Futility stopping in clinical trials. *Statistics and its Interface* Volume 5 415-423.

Renfro, L. A., Coughlin, C. M., Grothey, A. M., Sargent, D. J. (2014). Adaptive randomized phase II design for biomarker threshold selection and independent evaluation. *Chinese Clinical Oncology* 3(1): 3489.

Song, Y., Chi, G. Y. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* 26: 3535–3549.

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

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

Simon R., Simon N. (2017). Inference for multimarker adaptive enrichment trials. *Statistics in Medicine* 36:4083-4093.

Tian, L., Alizadeh, A. A., Gentles, A., J., Tibshirani, R. (2012). A Simple Method for Detecting Interactions between a Treatment and a Large Number of Covariates. Journal of the American Statistical Association, 109:508, 1517-1532.Wang, S., Hung, Hung, J. H. M., O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal*, 51: 358-374.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B.* Vol. 58, No. 1, pages 267-288.

Wassmer, G., Dragalin, V. (2015). Designing issues in confirmatory adaptive population enrichment trials. *Journal of Biopharmaceutical Statistics*. 25:4, 651-669.

Wolpin, B. M, O'Reilly, E. M., Ko, Y. J., Blaszkowsky, L. S., Rarick, M., Rocha-Lima, C. M., Ritch, P., Chan, E., Spratlin, J., Macarulla, T., Mcwhirter, E., Pezet, T., Lichinister, M., Roman, L., Hartford, A., Morrison, K., Jackson, L., Vincent, M., Reyno, L., Hidalgo, M. (2013). Global, Multicenter, Randomized, Phase II Trial of Gemcitabine and Gemcitabine Plus AGS-1C4D4 in Patients with Previously Untreated, Metastatic Pancreatic Cancer. *Annals of Oncology* 24: 1792-1801.

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. (2006). *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

Zeng, Y., Breheny, P. (2016). Overlapping Group Logistic Regression with Applications to Genetic Pathway Selection. *Cancer informatics*, *15*, 179-87.

Zhang, Z., Li, M., Lin, M., Soon, G., Greene, T., Shen, C. (2017), Subgroup selection in adaptive signature designs of confirmatory clinical trials. *Journal of the Royal Statistical Society C*, 66: 345–361.

Zhang Z, Chen R, Soon G, Zhang H. (2018). Treatment evaluation for a data-driven subgroup in adaptive enrichment designs of clinical trials. *Statistics in Medicine* 37:1–11.