Published in final edited form as:

J Biopharm Stat. 2019 ; 29(4): 685–695. doi:10.1080/10543406.2019.1633655.

ESTIMATING THE SUBGROUP AND TESTING FOR TREATMENT EFFECT IN A POST-HOC ANALYSIS OF A CLINICAL TRIAL WITH A BIOMARKER

Neha Joshi^a, Jason Fine^a, Rong Chu^b, Anastasia Ivanova^{a,*}

^aDepartment of Biostatistics, CB #7420, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420, USA

^bAgensys, Inc., Santa Monica, California, USA

Abstract

We 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.

Keywords

Subgroup; biomarker; cross-validation

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; Alosch 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.

* aivanova@bios.unc.edu.

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 k th step, $k = 1, \dots, K$, 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 k th 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 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. 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.

In this paper, we consider the case of a single biomarker. In Section 2, for several models for biomarker and treatment response we derive the subgroup that maximizes the utility function equal to the non-centrality parameter and another utility function that give more weight to larger subgroups. We describe several procedures to test for the treatment effect in Section 3. Simulations in Section 4 compare a non-parametric method of subgroup estimation and the method based on fitting a linear model with treatment by biomarker interaction. We illustrate the methods with an example in Section 5 and present conclusions in Section 6.

2. Defining and estimating the subgroup

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 $\mu_T(c) = E[Y|X > c, T = 1]$ be the treatment response in subjects with $X > c$ receiving treatment and $\mu_C(c) = E[Y|X > c, T = 0]$ 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 $\pi(c) = P[X > c]$. 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

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

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

$c^* = \arg \max \{ \pi(c)^\gamma [\mu_T(c) - \mu_C(c)] \}$. Lai, Lavori and Liao (2014) considered

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

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 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 $X \sim \text{Uniform}(0, 1)$ in Model 3.

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

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

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

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.

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

$$\begin{pmatrix} Y_T \\ X \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_T \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_T^2 & \rho_{TX}\sigma_T \\ \rho_{TX}\sigma_T & 1 \end{pmatrix} \right), \begin{pmatrix} Y_C \\ X \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_C \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_C^2 & \rho_{CX}\sigma_C \\ \rho_{CX}\sigma_C & 1 \end{pmatrix} \right).$$

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.

Model 3. The model with treatment by biomarker interaction is

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

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

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.

3. Testing for the treatment effect

In Section 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 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 $\alpha/2$.

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. 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 $0.5P[X > \hat{c}_1] + 0.5P[X > \hat{c}_2]$.

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 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 with $\sigma_C^2 = 1$. 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 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 $Z_{All, S}$.

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.

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 $U(c, \gamma) = \pi(c)^\gamma [\mu_T(c) - \mu_C(c)]$ 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.

Acknowledgements

Dr. Ivanova's work was supported in part by the NIH grant P01 CA142538. We thank Agensys for providing the phase 2 trial data. We thank the associate editor and reviewers for helpful comments.

Appendix.: DERIVATIONS OF THE OPTIMAL BIOMARKER CUT-OFF c^* FOR MODELS 1–3

Model 1. For 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 is $c^* = -\infty$ because

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

Model 2. In the bivariate normal model, using standard formulas (Arnold et al., 1993)

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

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

where $\phi(c)$ is a normal density and Φ is normal cumulative distribution function. We

maximize $\frac{E(Y_T - Y_C | X > c) \pi^\gamma}{\sqrt{\text{Var}(Y_T - Y_C | X > c)}}$ with $\gamma = 1/2$ if U_1 is used or $\gamma = 3/4$ if U_2 is used. There is no closed form formulae for the optimal cut-off.

Model 3. The mean and the variance in the overall population with size N are

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

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

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

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

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

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

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

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

Then we maximize

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

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

REFERENCES

- Alosh M, Huque MF (2009). A flexible strategy for testing subgroups and overall population. *Statistics in Medicine* 28: 3–23. [PubMed: 18985704]
- Arnold BC, Beaver RJ, Groeneveld RA, Meeker WQ (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika* 58: 471.
- Diao G, Dong J, Zeng D, Ke C, Rong A, Ibrahim JG (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. [PubMed: 21815180]
- 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. [PubMed: 16278411]
- Freidlin B, Jiang W, Simon R (2010). The cross-validated adaptive signature design. *Clinical Cancer Research* 16(2): 691–698. [PubMed: 20068112]
- 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. [PubMed: 22328327]
- 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. [PubMed: 17596577]
- Kehl V, Ulm K (2006). Responder identification in clinical trials with censored data. *Computational Statistics & Data Analysis* 50: 1338–1355.
- Lai T, Lavori P, Liao O (2014). Adaptive choice of patient subgroup for comparing two treatments. *Contemporary Clinical Trials* 39(2): 191–200. [PubMed: 25205644]
- 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. [PubMed: 21786278]
- Renfro LA, Coughlin CM, Grothey AM, Sargent DJ (2014). Adaptive randomized phase II design for biomarker threshold selection and independent evaluation. *Chinese Clinical Oncology* 3(1): 3489. [PubMed: 25485277]
- Song Y, Chi GY (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* 26: 3535–3549. [PubMed: 17266164]
- Su XG, Tsai CL, Wang HS, Nickerson DM, Li BG (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. [PubMed: 23525452]

- Wolpin BM, O'Reilly EM, Ko YJ, Blaszkowsky LS, Rarick M, Rocha-Lima CM, 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. [PubMed: 23448807]
- 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.

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, $E[Y] = X^a T$.

a	U_1			U_2	
	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	$\frac{0.72(U_1)}{0.66(U_2)}$	$\frac{0.77(U_1)}{0.66(U_2)}$	-
			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 $\delta, \rho_T, \rho_C, \sigma_T^2$. 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}, Z_S , and HC in each scenario is in bold.

δ	ρ_T	ρ_C	σ_T^2	Method	π^*	Z_{All}	$Z_{All,S}$	Z_S	HC
0.25	0.25	0.10	2	True	$\pi_{U_1}^* = 0.55$ $\pi_{U_2}^* = 0.75$	0.63	0.74(U_1) 0.72(U_2)	0.84(U_1) 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	$\pi_{U_1}^* = 1.00$ $\pi_{U_2}^* = 1.00$	0.72	0.72	0.72	-
				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	$\pi_{U_1}^* = 1.00$ $\pi_{U_2}^* = 1.00$	0.80	0.80	0.80	-
				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
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
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
			$\pi_{U_1}^* = 0.44$	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	-