Statistica Sinica Preprint No: SS-2025-0498	
Title	Beyond a Single Biomarker: Adaptive Enrichment Design with Data-driven Subgroup Rules
Manuscript ID	SS-2025-0498
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0498
Complete List of Authors	Mingde Wu, Juan Shen, Ruqian Zhang and Emma Jingfei Zhang
Corresponding Authors	Juan Shen
E-mails	shenjuan@fudan.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

BEYOND A SINGLE BIOMARKER: ADAPTIVE ENRICHMENT DESIGN WITH DATA-DRIVEN SUBGROUP RULES

Mingde Wu[1], Juan Shen[1], Ruqian Zhang[1] and Jingfei Zhang[2]

[1]*Fudan University and* [2]*Emory University*

Abstract: Adaptive enrichment design plays a crucial role in clinical trials where treatment effects vary across subgroups. However, previous works often focus on subgroups determined by a single biomarker. In this paper, we extend the framework to allow subgroups to be defined by multiple biomarkers. We introduce a flexible subgroup modeling framework and employ variable selection methods to identify predictive and prognostic biomarkers simultaneously. Simulation studies demonstrate that the proposed adaptive enrichment design yields accurate subgroup-specific estimates, valid p-values, and achieves the desired statistical power via appropriate Stage 2 sample size determination. Finally, we apply the adaptive enrichment design to the National Supported Work dataset for empirical illustration.

Key words and phrases: Adaptive enrichment design, variable selection, subgroup selection and analysis, sample size determination.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

1. Introduction

Treatment effect heterogeneity across patient subpopulations, such as those defined by genomic or phenotypic biomarkers, is common in clinical trials and has important implications for treatment evaluation (Loh et al., 2019). Such heterogeneity has motivated the use of adaptive enrichment designs for identifying and evaluating patient subgroups with enhanced treatment effects, particularly when the relevant subgroups are unknown at the start of the trial (Bretz et al., 2009; Lai et al., 2019).

Adaptive enrichment designs often consist of two stages (Rosenblum et al., 2020; Lin et al., 2021), with the primary objective of identifying a patient subgroup that is more likely to benefit from the experimental treatment and subsequently confirming the treatment efficacy within the subgroup. In Stage 1, patients are enrolled from the overall population and randomly assigned to either the treatment arm or the control arm. At the end of Stage 1, accumulated clinical outcomes and biomarker information are used to identify potentially responsive subgroup defined by "predictive" biomarkers, where "predictive" biomarkers are those relevant to subgroup membership and "prognostic" biomarkers are those relevant to the response within subgroups (Italiano, 2011). In Stage 2, patients are enriched meaning that enrollment is restricted to patients belonging to this subgroup, to

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

further evaluate treatment effects, thereby improving statistical power while reducing the required sample size. Depending on the interim evidence from Stage 1, additional design adaptations may be implemented. First, if no promising subgroup is identified at the interim analysis, enrollment in Stage 2 continues in the full population. Second, if the observed treatment effect is insufficient to justify continuation, the trial may be terminated early for futility.

Since the subgroup structure is often unknown, identifying subgroups remains a central challenge in adaptive enrichment designs. Most existing approaches restrict subgroups to a prespecified structure. Stallard (2023) used a prespecified continuous biomarker to identify the subgroup, controlling family-wise Type I error rate in hypothesis testing. Frieri et al. (2023) considered continuous outcomes and a continuous biomarker, with a focus on sample size determination in Stages 1 and 2. Zhan et al. (2025) proposed a Bayesian procedure for assigning optimal treatment regimens to predefined patient subgroups. While these approaches make useful advances, restricting subgroups to known or single biomarkers may limit their applicability in many clinical settings, where treatment effect heterogeneity is commonly driven by multiple biomarkers. Moreover, when treatment effect heterogeneity is driven by multiple biomarkers, enrichment based on

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

a single predictive biomarker may inadequately capture the benefiting subpopulation, leading to inaccurate treatment effect estimation and inference.

In this work, we propose an adaptive enrichment design that allows subgroups to be defined by multiple predictive biomarkers. As the predictive biomarkers are not prespecified, we consider a data-driven variable selection procedure to identify the threshold-based construction of subgroups. We propose a two-stage hypothesis testing procedure for assessing treatment efficacy within a subgroup selected through an adaptive enrichment design. Additionally, we develop a Stage 2 sample size determination and a futility stopping boundary to achieve the desired statistical power while reducing experimental costs.

Our work is related to the extensive literature on subgroup identification. For example, Loh et al. (2019) provided a comparative review of several methods which can select predictive variables to identify subgroup defined by an indicator function, such as MOB (Zeileis et al., 2008; Seibold et al., 2016), GUIDE (Loh, 2002). For model-based subgroup identification, Takeishi (2025) adopted an L_1-norm penalty to select classification covariates that define subgroups, whereas Zhang et al. (2025) proposed a Bayesian framework for the joint selection of predictive and prognostic covariates. Beyond subgroup identification, inference issues have also attracted grow-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

ing interest, with particular emphasis on correcting bias in treatment effect estimation. Simon and Simon (2017) proposed bootstrap-based correction methods, and Wang and He (2023) provided a comprehensive review of four classes of approaches. These work focus on subgroup identification and inference and do not directly address how identified subgroups can be incorporated in adaptive enrichment designs with valid inference and design procedures.

Our work is also related to several other lines of research. Zhao et al. (2013) propose a data-driven, multi-biomarker approach to identify a promising target population for future studies, sharing our motivation of moving beyond single-biomarker subgroup definitions. Stallard and Kimani (2018) study post-selection estimation in multi-arm multi-stage trials with interim selection among treatment arms. Their setting differs from ours in that selection is among treatment arms rather than biomarker-defined subgroups. Guo et al. (2025) consider a two-stage framework for post-selection inference after data-driven subgroup selection with prespecified subgroups, and their work does not address Stage 2 sample size determination or futility stopping.

The main contribution of our work lies in the integration of four components within a unified adaptive enrichment design framework: (i) data-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

driven subgroup construction from multiple biomarkers via variable selection; (ii) a two-stage hypothesis testing procedure with valid family-wise Type I error rate (FWER) control; (iii) a Stage 2 sample size determination procedure; and (iv) a futility stopping boundary. This integration is not a simple combination of existing methods, which typically address one component at a time and do not account for how the other components interact. By incorporating all four components into a single framework, our approach enables adaptive enrichment trials with subgroup identification, confirmatory inference, and sample size planning, all carried out within a unified framework.

The rest of this paper is organized as follows. Section 2 describes the subgroup selection rules and the adaptive enrichment design procedure. Section 3 develops estimators of treatment effects and their variances, and presents the hypothesis testing procedure and sample size determination for Stage 2. Section 4 reports simulation experiments that evaluate the performance of the proposed adaptive enrichment procedure, including Type I error control and statistical power under the Stage 2 sample size determination. Section 5 illustrates the proposed methods using data from the National Supported Work study. Section 6 concludes the paper.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

2. Methodology

2.1 The Model

Suppose the data consist of n independent observations $\{y_i, t_i, \boldsymbol{z}_i, \boldsymbol{x}_i\}_{i=1}^n$, where $y_i \in \mathbb{R}$ is a continuous response, $t_i \in \{0, 1\}$ is a treatment variable ($t_i = 1$ if patient i receives treatment and $t_i = 0$ if patient i receives control), $\boldsymbol{z}_i \in \mathbb{R}^{p_1}$ denotes candidate prognostic covariates that affect the outcome regardless of treatment, and $\boldsymbol{x}_i \in \mathbb{R}^{p_2}$ denotes candidate predictive covariates that determine subgroup membership. The covariates are termed "candidate" because some may be inactive, meaning that they have no effect on the outcome or subgroup membership. Both p_1 and p_2 can be high-dimensional, where the number of candidate variables is large relative to the sample size.

We introduce the latent subgroup membership indicator $\delta_i \in \{0, 1\}$, where $\delta_i = 1$ indicates patient i belongs to the subgroup with enhanced treatment effect. Conditional on δ_i, treatment assignment, and prognostic covariates, we model the outcome as

$$\begin{aligned} y_i \mid (\delta_i, \boldsymbol{z}_i, \boldsymbol{x}_i, t_i) &= f(\delta_i, \boldsymbol{z}_i, t_i) + \epsilon_i, \\ \delta_i \mid (\boldsymbol{z}_i, \boldsymbol{x}_i) &\sim g(\boldsymbol{x}_i), \end{aligned} \tag{2.1}$$

where $f(\delta_i, \boldsymbol{z}_i, t_i)$ characterizes the mean outcome as a function of sub-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

group membership, treatment, and prognostic covariates, $g(\boldsymbol{x}_i)$ determines the subgroup membership based on the predictive variables $\boldsymbol{x}_i$, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are independent errors with unknown variance σ. Model (2.1) is general and includes several subgroup models as special cases. For example, when $g(\boldsymbol{x}_i) = I\left\{\boldsymbol{x}_i^\top \boldsymbol{\gamma} > a\right\}$ with parameters $\boldsymbol{\gamma}$ and a, the model corresponds to tree-based methods (Loh, 2002; Zeileis et al., 2008). When $g(\boldsymbol{x}_i)$ follows a logistic regression form, the model corresponds to a logistic–normal mixture model (Shen and He, 2015; Zhang et al., 2025).

In the following, we adopt a specific form of Model (2.1) based on the structured logistic–normal mixture model of Shen and He (2015). While our development focuses on this specification, the proposed adaptive enrichment design does not rely on this particular functional form and can in principle accommodate alternative choices of $f(\cdot)$ and $g(\cdot)$. We consider a two-component mixture model given by

$$\begin{aligned} f(\delta_i, \boldsymbol{z}_i, t_i) &= \boldsymbol{z}_i^\top \boldsymbol{\beta} + \delta_i t_i \alpha_1 + (1-\delta_i) t_i \alpha_2, \\ g(\boldsymbol{x}_i) &= \text{Bernoulli}\left[\frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\gamma})}\right], \end{aligned} \tag{2.2}$$

where α_1 and α_2 denote the subgroup-specific average treatment effects (ATEs) for the subgroup with enhanced treatment effect and its complement, respectively, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are coefficients for $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$, respectively.

As not all variables have effect on the outcome or subgroup membership,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

variable selection is performed on the prognostic covariates $\boldsymbol{z}_i$ and predictive covariates $\boldsymbol{x}_i$. To perform variable selection, we directly implement the Gibbs sampling algorithm proposed by Zhang et al. (2025), and obtain the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, denoted as $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, as well as the posterior inclusion probabilities for each covariate. These posterior inclusion probabilities are used to identify active prognostic and predictive variables.

Since $\delta_i = 1$ denotes the subgroup with a greater treatment effect, it follows that $\alpha_1 > \alpha_2$. We assume that patients with larger values of $\hat{g}(\boldsymbol{x}_i) = 1/[1 + \exp(-\boldsymbol{x}_i^\top \hat{\boldsymbol{\gamma}})]$ tend to experience more favorable treatment effects. We further assume that larger values of y_i correspond to better patient outcomes, as assumed in Frieri et al. (2023). Under the above assumptions, we define $\hat{\delta}_i = I\{\hat{g}(\boldsymbol{x}_i) \geq c\}$ where c is chosen from a threshold set $\mathbb{C}$ defined in Section 2.2, resulting in a sequence of nested subgroups, where a larger threshold c corresponds to a subgroup contained within that defined by a smaller c.

It is worth noting that the downstream inference procedure depends on the working model only through the estimated subgroup membership scores $\hat{g}(\boldsymbol{x}_i)$ used for threshold-based subgroup construction. Once consistent estimates of $\hat{g}(\boldsymbol{x}_i)$ are available, the subsequent subgroup selection and testing procedure remain valid regardless of the specific working model. This is

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

illustrated empirically in Settings M4–M6, where the true subgroup structure does not conform to the logistic-normal mixture model assumed in the working model, and further supported by the sensitivity analyses reported in Section S3.4 of the Supplementary Material.

2.2 Subgroup Threshold Determination

Subgroup membership can be determined by thresholding the estimated subgroup membership score $\hat{g}(\boldsymbol{x}_i)$ over a prespecified set of candidate thresholds $\mathbb{C} = \{c_1, \ldots, c_k\}$, ordered such that $c_1 > \cdots > c_k$. For example, $c_1, \ldots, c_k$ may be chosen as quantiles of $\hat{g}(\boldsymbol{x}_i)$. We assume that the thresholds are chosen so that, for $j = 1, \ldots, k-1$, there exists at least one subject i such that $c_j > \hat{g}(\boldsymbol{x}_i) \geq c_{j+1}$, and that $c_k \leq \hat{g}(\boldsymbol{x}_i)$ for all i. These conditions ensure that each candidate subgroup is nonempty, and that subgroups defined by larger thresholds are contained within those defined by smaller thresholds (Stallard, 2023). Correspondingly, the jth subgroup is defined as $\{i : \hat{g}(\boldsymbol{x}_i) \geq c_j\}$ and the corresponding subgroup size is denoted by $n_j = \mid \{i : \hat{g}(\boldsymbol{x}_i) \geq c_j\} \mid$, where $\mid \cdot \mid$ denotes cardinality. Under Model (2.2), each threshold $c_j \in \mathbb{C}$ defines a candidate subgroup. Let $\alpha_{1,j}$ denote the average treatment effect within this candidate subgroup, and let $\alpha_{2,j}$ denote the average treatment effect in its complement. These quantities are

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

threshold-specific and represent the treatment effects associated with the subgroup induced by c_j. Specifically, conditional on the subgroup defined by $\{\hat{g}(\boldsymbol{x}_i) \geq c_j\}$, the outcome model is given by

$$y_i = \boldsymbol{z}_i^\top \boldsymbol{\beta}^{(j)} + t_i \alpha_{1,j} + \epsilon_i^{(j)}, \quad i : \hat{\delta}_i^{(j)} = 1, \tag{2.3}$$

where $\epsilon_i^{(j)} \sim \mathcal{N}(0, {\sigma^{(j)}}^2)$ are independent errors with unknown variance ${\sigma^{(j)}}^2$, and $\hat{\delta}_i^{(j)} = I\{\hat{g}(\boldsymbol{x}_i) \geq c_j\}$. An analogous model can be defined for the complement subgroup with treatment effect $\alpha_{2,j}$.

A subgroup selection rule specifies how one candidate subgroup is chosen at the interim analysis for subsequent enrichment. Let $\hat{\alpha}_{1,j}$ and $\hat{\alpha}_{2,j}$ be the estimates of $\alpha_{1,j}$ and $\alpha_{2,j}$, respectively. We consider two subgroup selection rules. Let $J^{(1)}, J^{(2)} \in \mathbb{C}$ be the indexes of the best subgroups selected by Rules 1 and 2, respectively. Selection Rule 1 maximizes ATE in the subgroup, i.e., $J^{(1)} = \arg\max_{j=1,\ldots,k}\{\hat{\alpha}_{1,j}\}$. Selection Rule 2 maximizes the difference of ATEs in subgroup and its complement. Then the best subgroup is defined by $\hat{\delta}_i^* = I\{\hat{g}(\boldsymbol{x}_i) \geq c^*\}$, where c^* is the optimal threshold $c_{J^{(1)}}$ or $c_{J^{(2)}}$. The choice between the two selection rules should be driven by the scientific objective of the study. If the primary goal is to identify patients with the strongest individual-level response, Selection Rule 1 is more appropriate. If the goal is to identify a subgroup for which treatment provides a clear advantage relative to its complement, Selection

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

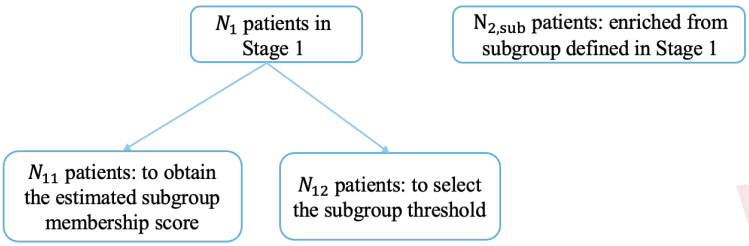

Figure 1: Sample partitioning at each stage.

Rule 2 is preferred. Importantly, both rules are accommodated within our framework while maintaining valid inference.

2.3 Adaptive Enrichment Design

We propose a two-stage adaptive enrichment design to assess ATE in the subgroup with heterogeneous treatment effects. Figure 1 illustrates the sample partitioning at each stage, and Figure 2 provides a flowchart of the design.

In Stage 1, we enroll N_1 patients and randomly assign them to treatment or control. To separate subgroup score estimation from threshold selection, we partition the Stage 1 data into two independent subsamples of sizes N_{11} and $N_{12} = N_1 - N_{11}$. Using the first subsample of size N_{11}, we estimate the subgroup membership score $\hat{g}(\boldsymbol{x}_i)$ via a data-driven subgroup identification procedure; for example, under Model (2.2), the Bayesian approach of Zhang

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

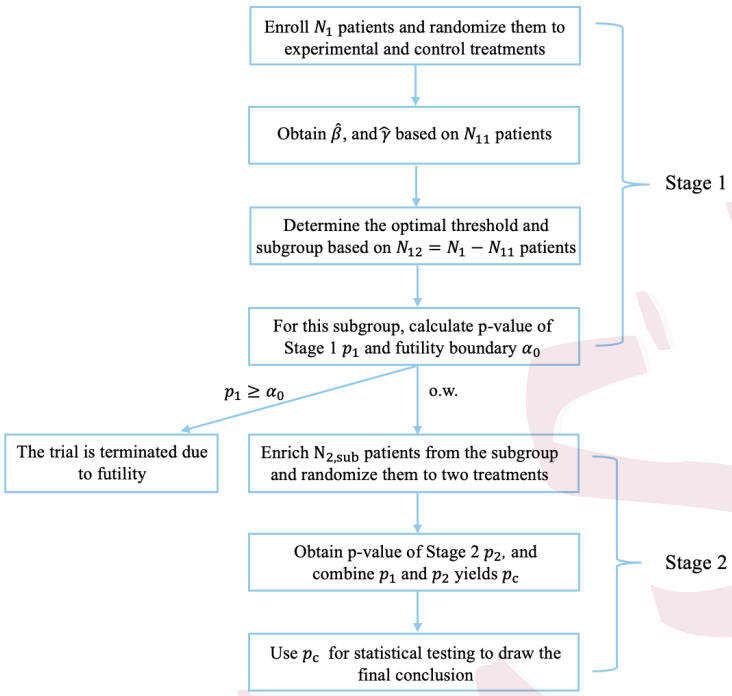

Figure 2: Flowchart of the proposed two-stage adaptive enrichment design, where N_1 and $N_{2,\text{sub}}$ are sample sizes in Stage 1 and Stage 2, respectively, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are the estimates of coefficients, and p_c is the combined p-value for testing.

et al. (2025) may be employed. Using the remaining N_{12} patients, we apply the subgroup selection rule described in Section 2.2 to determine the optimal threshold c^*. The selected subgroup at the end of Stage 1 is thus defined by $\hat{g}(\boldsymbol{x}_i) \geq c^*$. Based on the selected subgroup, we estimate the subgroup

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

average treatment effect from

$$y_i = \boldsymbol{z}_i^\top \boldsymbol{\beta}' + t_i \alpha_1' + \epsilon_i', \qquad i : \hat{\delta}_i^* = 1, \tag{2.4}$$

where $\epsilon_i' \sim \mathcal{N}(0, \sigma'^2)$ are independent errors with unknown variance σ'^2. For the null hypothesis $H_{0,J^{(r)}} : \alpha_{1,J^{(r)}} \leq 0$, as defined in Section 3.2 and based on Selection Rule r for $r = 1, 2$, we then compute the corresponding Stage 1 p-value, denoted by p_1. We also provide a futility stopping boundary for the p-value in Stage 1, denoted by α_0, as detailed in Section 3.4. If $p_1 > \alpha_0$, the trial is terminated for futility; otherwise, it proceeds to Stage 2.

Remark 1. The average treatment effect α_1' in Model (2.4) differs from the subgroup effect α_1 in the working model (2.2), as $\hat{\delta}_i^*$ is an estimate of δ_i. Specifically, when $g(\boldsymbol{x}_i)$ follows a logistic regression form as in Model (2.2), we have

$$\alpha_1' = Pr(\delta_i = 1 | \delta_i^* = 1) \cdot \alpha_1 + Pr(\delta_i = 0 | \delta_i^* = 1) \cdot \alpha_2,$$

where $\delta_i \sim$ Bernoulli $\left[1/\{1 + \exp(-\boldsymbol{x}_i^\top \boldsymbol{\gamma})\}\right]$ and $\delta_i^* = I\{1/[1+\exp(-\boldsymbol{x}_i^\top \hat{\boldsymbol{\gamma}})] > c^*\}$. When $g(\boldsymbol{x}_i) = I\{\boldsymbol{x}_i^\top \boldsymbol{\gamma} > a\}$ takes an indicator-function form, and $\boldsymbol{\gamma}$ and a are accurately estimated, then $\hat{\delta}_i^* = \delta_i$ and hence $\alpha_1' = \alpha_1$.

In Stage 2, we enroll $N_{2,\text{sub}}$ patients from the selected subgroup defined by $\hat{g}(\boldsymbol{x}_i) \geq c^*$, and randomly assign them to either the treatment or control.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

The value of $N_{2,\text{sub}}$ is either prespecified or determined in a data-driven way to achieve the target statistical power while controlling the Type I error rate.

Based on these $N_{2,\text{sub}}$ patients, we estimate the subgroup specific ATE α_1' under Model (2.4) via Equation (3.1), denoted by $\hat{\alpha}_1^{'(s2)}$, and obtain the p-value p_2 in Stage 2 for the null hypothesis $H_0' : \alpha_1' \leq 0$ defined in Section 3.2. The p-values in Stage 1 and Stage 2 are then combined to yield a combined p-value p_c according to Equation (3.5), which is used for the final hypothesis test to confirm whether the treatment is superior to the control treatment within the selected subgroup.

3. Estimation and Testing

3.1 Estimation

In this subsection, we present the estimation of ATE under Model (2.3). For subjects satisfying $\hat{\delta}_i^{(j)} = 1$, let $T^{(j)} = (t_1, \ldots, t_{n_j})^\top$, $y^{(j)} = (y_1, \ldots, y_{n_j})^\top$, and $\epsilon^{(j)} = (\epsilon_1^{(j)}, \ldots, \epsilon_{n_j}^{(j)})^\top$, and let $\boldsymbol{Z}^{(j)}$ and $\boldsymbol{X}^{(j)}$ denote the design matrices for prognostic and predictive covariates, respectively. Define the projection matrices

$$P_{T^{(j)}} = T^{(j)}({T^{(j)}}^\top T^{(j)})^{-1}{T^{(j)}}^\top, \quad N_{T^{(j)}} = I - P_{T^{(j)}}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

Let $\boldsymbol{D}^{(j)} = (y^{(j)}, T^{(j)}, \boldsymbol{Z}^{(j)}, \boldsymbol{X}^{(j)}, P_{T^{(j)}}, N_{T^{(j)}})$. Under Model (2.3), we can estimate $\alpha_{1,j}$ as

$$\hat{\alpha}_{1,j} = \left(T^{(j)\top} T^{(j)}\right)^{-1} T^{(j)\top} \left[I - \boldsymbol{Z}^{(j)} \left(\boldsymbol{Z}^{(j)\top} N_{T^{(j)}} \boldsymbol{Z}^{(j)}\right)^{-1} \boldsymbol{Z}^{(j)\top} N_{T^{(j)}}\right] y^{(j)}, \tag{3.1}$$

where I is an identity matrix of suitable dimension. In addition, the variance of $\hat{\alpha}_{1,j}$ are given by

$$\begin{aligned}\sigma^2_{\hat{\alpha}_{1,j}} =& {\sigma^{(j)}}^2 (T^{(j)\top} T^{(j)})^{-1} T^{(j)\top} \left[I + \boldsymbol{Z}^{(j)} \left(\boldsymbol{Z}^{(j)\top} N_{T^{(j)}} \boldsymbol{Z}^{(j)}\right)^{-1} \boldsymbol{Z}^{(j)\top}\right] \\ & T^{(j)} \left(T^{(j)\top} T^{(j)}\right)^{-1}.\end{aligned} \tag{3.2}$$

In Stage 1, after selecting the optimal subgroup rule, we obtain $\hat{\delta}_i^* = I\{\hat{g}(\boldsymbol{x}_i) \geq c^*\}$. For subjects satisfying $\hat{\delta}_i^* = 1$, we define $\boldsymbol{D}^{(s1)} = (y^{(s1)}, T^{(s1)}, \boldsymbol{Z}^{(s1)}, \boldsymbol{X}^{(s1)}, P_{T^{(s1)}}, N_{T^{(s1)}})$ for Stage 1, and $\boldsymbol{D}^{(s2)} = (y^{(s2)}, T^{(s2)}, \boldsymbol{Z}^{(s2)}, \boldsymbol{X}^{(s2)}, P_{T^{(s2)}}, N_{T^{(s2)}})$ for Stage 2, analogously. Substituting $D^{(j)}$ in (3.1)–(3.2) with $D^{(s1)}$ for Stage 1 or $D^{(s2)}$ for Stage 2, and substituting ${\sigma^{(j)}}^2$ with σ'^2 in (3.2) yields the subgroup ATE estimators $\hat{\alpha}_1^{\prime(s1)}$ and $\hat{\alpha}_1^{\prime(s2)}$ for Stages 1 and Stage 2, respectively, together with their estimated variances.

3.2 Testing

To evaluate whether the experimental treatment is superior to the control treatment within a subgroup of interest, we first test the null hypothesis $H_{0,J^{(r)}} : \alpha_{1,J^{(r)}} \leq 0$ versus $H_{1,J^{(r)}} : \alpha_{1,J^{(r)}} > 0$ in Stage 1, where $\alpha_{1,J^{(r)}}$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

denotes the average treatment effect within $J^{(r)}$th subgroup based on Selection Rule r for $r = 1, 2$, and then test $H_0' : \alpha_1' \leq 0 \quad \text{versus} \quad H_1' : \alpha_1' > 0$ in Stage 2, where α_1' denotes the average treatment effect in the selected subgroup, at significance level α.

In what follows, we provide a combined p-value approach that integrates the p-values in Stage 1 and Stage 2 to test whether the experimental treatment yields a better average treatment effect in the selected subgroup, while controlling the family-wise Type I error rate (FWER) over $\mathcal{H} = \{H_{0,1}, ..., H_{0,k}, H_0'\}$.

First, we obtain the p-value from Stage 1. Let $J^{(r)}$ be the subgroup index selected from $\{1, \ldots, k\}$ under Selection Rule r defined in Section 2.2. We compute the estimate $\hat{\alpha}_1^{'(s1)}$ and its variance $\hat{\sigma}^2_{\hat{\alpha}_1^{'(s1)}}$ according to Equations (3.1) and (3.2), respectively. The resulting test statistic is $z = \hat{\alpha}_1^{'(s1)} / \hat{\sigma}^2_{\hat{\alpha}_1^{'(s1)}}$ for Selection Rule 1, and for Selection Rule 2 the expression of the test statistic is in the Supplementary Material. We calculate the p-value from Stage 1 as:

$$p_1 = \max_{i=1,\ldots,J^{(r)}} \{1 - F_i^{(r)}(z)\} \quad \text{for } r = 1, 2, \tag{3.3}$$

where $F_i^{(r)}(z)$ denotes the corresponding approximate null distribution of z. The test statistic and its approximate null distribution are both proposed by Stallard (2023). The expressions of $F_i^{(r)}(z)$ under Selection Rule r for

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

$r = 1, 2$, are shown in the Supplementary Material.

For the p-value of Stage 2, we compute the estimate $\hat{\alpha}_1^{\prime(s2)}$ and its variance $\hat{\sigma}^2_{\hat{\alpha}_1^{\prime(s2)}}$ according to Equations (3.1) and (3.2), respectively. The resulting test statistic is $\hat{\alpha}_1^{\prime(s2)}/\hat{\sigma}^2_{\hat{\alpha}_1^{\prime(s2)}}$, and the corresponding p-value is then calculated as follows:

$$p_2 = 1 - \Phi(\hat{\alpha}_1^{\prime(s2)}/\hat{\sigma}^2_{\hat{\alpha}_1^{\prime(s2)}}), \tag{3.4}$$

where Φ is the standard normal distribution function.

Following Bauer and Kohne (1994), we use the inverse normal method to combine p_1 and p_2 based on Equation (3.3) and (3.4):

$$p_c = 1 - \Phi\big[w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2)\big], \tag{3.5}$$

where $w_1 = \sqrt{N_{12,\text{sub}}/(N_{12,\text{sub}} + N_{2,\text{sub}})}$ and $w_2 = \sqrt{N_{2,\text{sub}}/(N_{12,\text{sub}} + N_{2,\text{sub}})}$ satisfying $w_1^2 + w_2^2 = 1$, and $N_{12,\text{sub}}$ is the subgroup sample size in N_{12} patients.

Remark 2. While this global testing step provides strong error-rate guarantees, it may introduce some conservativeness, since it accounts for all candidate subgroups simultaneously rather than focusing on the selected one. An alternative approach is the method proposed by Guo et al. (2025), which combines data from the subgroup selection and validation phases and can serve as a less conservative testing procedure. However, the method of

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

Guo et al. (2025) is designed primarily for post-selection inference with prespecified candidate subgroups, and does not directly accommodate the Stage 2 sample size determination procedure of Section 3.3 or the futility stopping boundary of Section 3.4, both of which are essential components of a deployable adaptive enrichment trial design. A numerical comparison between the proposed combined p-value approach, the procedure relying solely on the Stage 2 p-value, and the method of Guo et al. (2025) is provided in Section S3.3 of the Supplementary Material.

The p-value of Stage 1, p_1, strongly controls the FWER for the family of hypotheses $\{H_{0,1}, ..., H_{0,k}\}$ via the closed testing procedure (Marcus et al., 1976; Stallard, 2023); and the p-value of Stage 2, p_2, is uniformly distributed under the null hypothesis H_0'. We note that what is required here is not exact recovery of the true latent subgroup membership at the individual level, which is generally not achievable when $g(\boldsymbol{x}_i)$ takes a logistic form, but rather a sufficiently accurate estimation of the subgroup membership score $\hat{g}(\boldsymbol{x}_i)$ for threshold-based subgroup construction. The Bayesian variable selection procedure of Zhang et al. (2025) provides consistent estimation of the subgroup membership score with high probability, which supports the accuracy of the selected subgroup. Under these conditions, the inverse normal combination test preserves strong FWER control over $\mathcal{H}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

3.3 Sample Size Determination for Stage 2

To reduce experimental costs, we propose a sample size determination approach for Stage 2 that achieves the target statistical power $1-\beta$ at significance level α. Based on the combined p-value (3.5), the combined test statistic is given as:

$$z_c = w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2) = w_1\Phi^{-1}(1-p_1) + w_2\hat{\alpha}_1^{\prime(s2)}/\hat{\sigma}^2_{\hat{\alpha}_1^{\prime(s2)}},$$

where $\hat{\alpha}_1^{\prime(s2)}/\hat{\sigma}^2_{\hat{\alpha}_1^{\prime(s2)}}$ is the test statistic from Stage 2. Denote the target mean, a practically meaningful threshold, in Stage 2 by d and let $\mu = w_1\Phi^{-1}(1-p_1) + w_2 d/\hat{\sigma}^2_{\hat{\alpha}_1^{\prime(s2)}}$. Assume that $z_c \sim N(\mu, 1)$ asymptotically under H_1, then

$$\begin{aligned} 1-\beta &= \Pr(Z_c > Z_{1-\alpha} \mid H_1) = \Pr(Z_c - \mu > Z_{1-\alpha} - \mu \mid H_1) \\ &= 1 - \Phi(Z_{1-\alpha} - \mu) = \Phi(\mu - Z_{1-\alpha}), \end{aligned} \tag{3.6}$$

where $Z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. Letting $Z_{1-\beta} = \Phi^{-1}(1-\beta)$, Equation (3.6) becomes

$$w_1\Phi^{-1}(1-p_1) + w_2 d/\hat{\sigma}^2_{\hat{\alpha}_1^{\prime(s2)}} = Z_{1-\alpha} + Z_{1-\beta}. \tag{3.7}$$

Since $1/\hat{\sigma}^2_{\hat{\alpha}_1^{\prime(s2)}}$ is asymptotically proportional to the sample size in different samples, we can use the sample of $N_{12,\text{sub}}$ in Stage 1 to estimate this proportion s, and then let $1/\hat{\sigma}^2_{\hat{\alpha}_1^{\prime(s2)}} \approx sN_{2,\text{sub}}$. Under the choice of

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

$w_1 = \sqrt{N_{12,\text{sub}}/(N_{12,\text{sub}} + N_{2,\text{sub}})}$ and $w_2 = \sqrt{N_{2,\text{sub}}/(N_{12,\text{sub}} + N_{2,\text{sub}})}$, to obtain the desired sample size $N_{2,\text{sub}}$, we solve

$$sd^2 N_{2,\text{sub}}^2 + [2\sqrt{sN_{12,\text{sub}}}d\Phi^{-1}(1-p_1) - (Z_{1-\alpha} + Z_{1-\beta})^2]N_{2,\text{sub}} + \left[\{\Phi^{-1}(1-p_1)\}^2 - (Z_{1-\alpha} + Z_{1-\beta})^2\right] N_{12,\text{sub}} \approx 0,$$

with the discriminant

$$D = (Z_{1-\alpha}+Z_{1-\beta})^2[(Z_{1-\alpha}+Z_{1-\beta})^2+4sd^2N_{12,\text{sub}}-4\sqrt{sN_{12,\text{sub}}}d\Phi^{-1}(1-p_1)].$$

Here, if $D < 0$, it indicates that p_1 is very small. In this case, we set N_2 equal to $N_{2,\min}$, where $N_{2,\min}$ denotes the minimal required sample size in Stage 2. For $D \geq 0$,

$$N_{2,\text{sub}} = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 - 2\sqrt{sN_{12,\text{sub}}}d\Phi^{-1}(1-p_1)}{2sd^2} + (Z_{1-\alpha} + Z_{1-\beta}) \cdot \frac{\left[(Z_{1-\alpha} + Z_{1-\beta})^2 + 4sd^2N_{12,\text{sub}} - 4\sqrt{sN_{12,\text{sub}}}d\Phi^{-1}(1-p_1)\right]^{1/2}}{2sd^2}. \tag{3.8}$$

Equation (3.8) provides the sample size determination in Stage 2 to achieve the target statistical power while controlling Type I error, thereby improving design efficiency and reducing experimental costs.

3.4 Futility Stopping Boundary

We further construct a futility stopping boundary. If achieving the desired statistical power requires an excessively large sample size in Stage 2, due

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

to a relatively large p-value from Stage 1, then the trial can be terminated early for reasons of utility and cost-efficiency.

Let α_0 and $N_{2,\max}$ denote the futility boundary and the maximum affordable sample size in Stage 2, respectively. If $p_1 > \alpha_0$, the trial is terminated early. Based on Equation (3.7), for $w_1 = \sqrt{N_{12,\text{sub}}/(N_{12,\text{sub}} + N_{2,\max})}$ and $w_2 = \sqrt{N_{2,\max}/(N_{12,\text{sub}} + N_{2,\max})}$, we have

$$\sqrt{N_{12,\text{sub}}}\Phi^{-1}(1-\alpha_0) + N_{2,\max}\sqrt{s}d \approx (Z_{1-\alpha} + Z_{1-\beta})\sqrt{N_{12,\text{sub}} + N_{2,\max}},$$

then the boundary given $N_{2,\max}$ is given by

$$\alpha_0 \approx 1 - \Phi\left[(Z_{1-\alpha} + Z_{1-\beta})\sqrt{1 + N_{2,\max}/N_{12,\text{sub}}} - N_{2,\max}\sqrt{s}d/\sqrt{N_{12,\text{sub}}}\right]. \tag{3.9}$$

4. Simulation Studies

In this section, we use simulation studies to evaluate the validity of the combined p-value, the performance of ATE estimates, and statistical power under the proposed adaptive enrichment design.

4.1 Setting

We set the sample sizes in Stage 1 to be $N_{11} = N_{12} = 250$. The Stage 2 sample size $N_{2,\text{sub}}$ is either prespecified or determined using Equation (3.8).

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

For sample size determination for Stage 2 and futility stopping boundary, we set the target ATE $d = \max\{\hat{\alpha}_1^{'(s1)}, 0.01\}$ since a very small subgroup ATE is considered negligible, where $\hat{\alpha}_1^{'(s1)}$ is the ATE in Model (2.4) based on the $N_{12,\text{sub}}$ patients. We further set $N_{2,\min} = 20$ and $N_{2,max} = 250$.

For settings M1-M3, we generate data from Model (2.2). Each row of $\boldsymbol{X}$ is generated independently from a multivariate normal distribution with unit marginal variances and pairwise covariance 0.25, and the dimension of $\boldsymbol{X}$ is set to p. An intercept column is added to $\boldsymbol{X}$, and we set $\boldsymbol{Z} = \boldsymbol{X}$. We consider different $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as follows:

$M1 : \boldsymbol{\beta} = \boldsymbol{\gamma} = (1, -1.5, 2, -2.5, 3)^T \quad \text{with } p = 4,$

$M2 : \boldsymbol{\beta} = \boldsymbol{\gamma} = (1, -1.5, 2, -2.5, 3, 0, \ldots, 0)^T \quad \text{with } p = 10,$

$M3 : \boldsymbol{\beta} = \boldsymbol{\gamma} = (1, -1.5, 2, -2.5, 3, 0, \ldots, 0)^T \quad \text{with } p = 50.$

Similar to the settings of Zhang et al. (2025) and Loh et al. (2019), we also consider subgroups defined through splitting rules. We generate ten predictor variables as follows: X_1 and X_2 follow standard normal distribution with a covariance of 0.5; X_3 is standard normal; X_4 follows an exponential distribution with mean 1; X_5 follows a Bernoulli distribution with success probability of 0.5; X_6 takes values from 0, 1, or 2 with equal probability; X_7 to X_{10} are jointly standard normal with pairwise covariance 0.5. The treatment variable t_i is independently generated from a Bernoulli

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

distribution with success probability 0.5. Then we consider the following settings:

$M4 : Y = 1 + X_1 + X_2 + X_4 + I(X_6 = 2) + X_7 + \alpha_1 t \times I(X_1 > 0) + \alpha_2 t \times [1 - I(X_1 > 0)] + \epsilon,$

$M5 : Y = 1 + X_1 + X_2 + \alpha_1 t \times I(X_1 > -1, X_4 < 1) + \alpha_2 t \times [1 - I(X_1 > -1, X_4 < 1)] + \epsilon,$

$M6 : Y = 1 + X_1 + \alpha_1 t \times I(X_1 > -1.5, X_4 < 2, X_6 \neq 0) + \alpha_2 t \times [1 - I(X_1 > -1.5, X_4 < 2, X_6 \neq 0)] + \epsilon,$

where ϵ is standard normal noise.

We first apply the method in Section 2.1 to $N_{11} = 250$ sub-patients in Stage 1 to obtain the estimated subgroup membership score $\hat{g}(\boldsymbol{x}_i)$. The remaining $N_{12} = 250$ patients in Stage 1 are then used to determine the threshold c^* from set $\mathbb{C}$, where $\mathbb{C} = \{c_1, \ldots, c_k\}$ is chosen as the deciles of the estimated subgroup membership score $\hat{g}(\boldsymbol{x}_i)$, yielding $k = 10$ candidate thresholds. The detailed procedure is described in Section 2.3.

To assess the validity of the p-values and evaluate the control of Type I error, we set $\alpha_1 = \alpha_2 = 0$, and adopt the calculated $N_{2,\text{sub}}$. To evaluate the performance of ATE estimation, we consider the setting $\alpha_1 = 10$, $\alpha_2 = 0$ and $N_{2,\text{sub}} = 100$ and compare the Bayesian variable selection approach (BVSA) (Zhang et al., 2025) with PRIM (Chen et al., 2015) and single pre-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

Table 1: Proportions of identifying a subgroup under the no-subgroup scenario ($\alpha_1 = \alpha_2 = 0$) across six settings based on 1,000 simulations.

Setting	M1	M2	M3	M4	M5	M6
Proportion	1.30%	1.90%	2.20%	4.80%	3.70%	3.30%

dictive variable approaches. Comparisons with other tree-based subgroup identification methods, such as MOB (Zeileis et al., 2008) and GUIDE (Loh, 2002), are presented in the Supplementary Material S3.1. Finally, we consider the scenario $\alpha_1 = 0.8$ and $\alpha_2 = -0.4$ to evaluate statistical power and the effectiveness of Stage 2 sample size determination in achieving the target power level $1 - \beta = 0.9$. Additional numerical results are provided in the Supplementary Material.

4.2 Numerical Results

In this subsection, we provide the simulation results for the assessment of p-value validity, comparison of subgroup ATE estimates and statistical power.

Table 1 reports the proportions of subgroup identification across M1 to M6 based on 1,000 simulations under the no-subgroup scenario with $\alpha_1 = \alpha_2 = 0$. Figure 3 presents the corresponding p-value distributions for

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

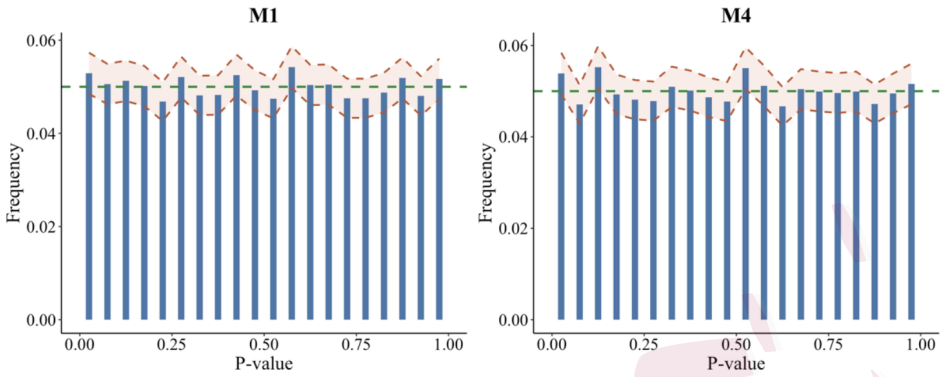

Figure 3: Empirical p-value distributions for two settings under the no-subgroup scenario ($\alpha_1 = \alpha_2 = 0$), based on 10,000 replications. The dashed lines represent the 95% Monte Carlo confidence bands around the nominal significance level of 0.05.

M1 and M4 based on 10,000 simulations. Results for M1 and M4 are shown as representative examples, as the results for other settings are similar. When no true subgroup exists, the probability of incorrectly identifying a subgroup does not exceed 5%, and the p-values are approximately uniformly distributed, with the nominal level of 0.05 falling within the 95% Monte Carlo confidence bands across all settings, confirming that the observed deviations are within the range of expected simulation variability. This provides empirical evidence for the validity of the p-values and the effective control of the Type I error.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

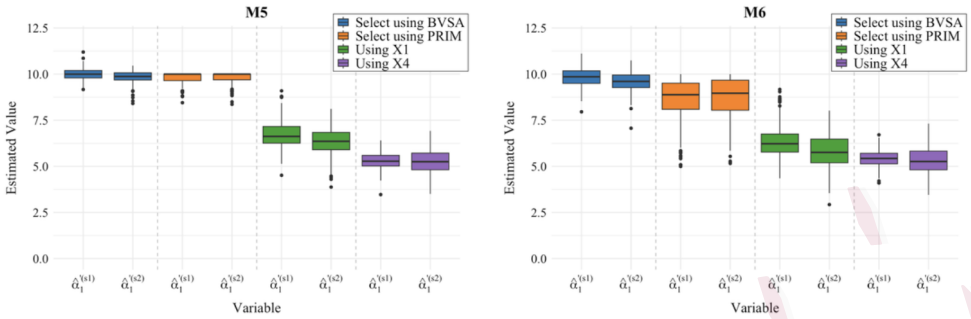

Figure 4: Comparison of subgroup ATE estimates obtained from four methods, where $\hat{\alpha}_1^{'(s1)}$ and $\hat{\alpha}_1^{'(s2)}$ denotes the subgroup ATE estimates in Stage 1 and Stage 2, respectively.

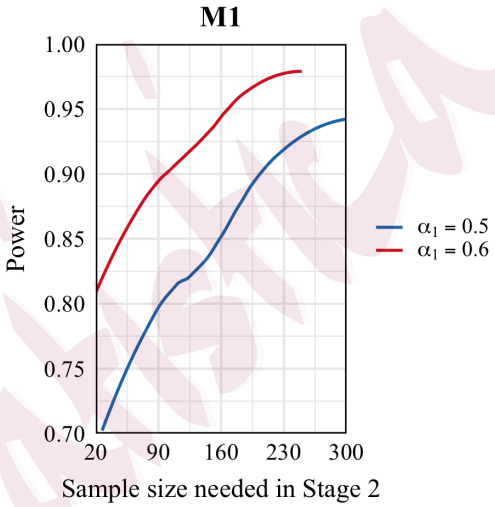

Figure 5: Sample size needed in Stage 2 to achieve desired power under $\alpha_1 \in \{0.5, 0.6\}$ and $\alpha_2 = 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

We focus on M5 and M6 to assess the performance of subgroup ATE estimation, as these settings involve complex subgroup structures defined by two or more biomarkers and facilitate comparison between variable selection approaches and the single predictive variable method. We consider the comparison of ATE estimates obtained from four different approaches including two predictive variable selection approaches: BVSA (Zhang et al., 2025), PRIM, and two single biomarker approaches using either X_1 or X_4 (Stallard, 2023). We set $\alpha_1 = 10$, $\alpha_2 = 0$, $N_{11} = N_{12} = 250$, and $N_{2,\text{sub}} = 100$. For a fair comparison, the prognostic variables are kept identical across all methods and determined using BVSA.

Figure 4 summarizes the comparison results based on 200 simulations. For M5, where the true subgroup is defined by X_1 and X_4, both BVSA and PRIM show satisfactory performance, with ATE estimates close to the true value. In contrast, the methods relying on a single predictive variable yield substantially biased estimates, with median values falling below 70% of the true effect. For M6, where the subgroup structure becomes more complex with three biomarkers, BVSA remains the best-performing approach. PRIM performs well, with its median estimate reaching approximately 90% of the true value, but it exhibits larger variance. The single predictive variable methods again perform poorly, with median estimates not exceeding

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

65% of the true effect. Overall, these results demonstrate that the proposed adaptive enrichment design framework provides robust and accurate ATE estimation across diverse subgroup structures. Furthermore, variable selection approaches outperform single predictive variable methods, highlighting the importance of data-driven subgroup rules in adaptive enrichment design. Additional numerical summaries are provided in S3.1 of the Supplementary Material.

Taking M1 as an example, we illustrate the relationship between statistical power and sample size in Stage 2 based on 200 simulations at the 5% significance level under $\alpha_1 \in \{0.5, 0.6\}$ and $\alpha_2 = 0$. The results are shown in Figure 5, and those for the other settings are similar. These results demonstrate that the sample size in Stage 2 can be appropriately determined to achieve a target statistical power.

Table 2 reports the statistical power and average sample size in Stage 2 based on 200 simulations at the 5% significance level under $\alpha_1 = 0.8$ and $\alpha_2 = -0.4$, where the sample size in Stage 2 is calculated using Equation (3.8) to target 90% power, with early futility stopping implemented via Equation (3.9). Here, "M1_Rule1" denotes M1 under Selection Rule 1, and the remaining notations are defined analogously. Across different settings and selection rules, all powers reach the target level of 90%. In addition,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

Table 2: Statistical power and average sample size in Stage 2, based on 200 simulations, with sample size determination for Stage 2 targeting 90% power and early futility stopping at the 5% significance level under $\alpha_1 = 0.8$ and $\alpha_2 = -0.4$.

	Statistical Power	**Average Stage 2 Sample Size**
M1_Rule1	91.41%	76
M1_Rule2	92.42%	76
M2_Rule1	91.96%	79
M2_Rule2	91.96%	82
M3_Rule1	91.28%	86
M3_Rule2	93.33%	89
M4_Rule1	95.88%	40
M4_Rule2	95.88%	41
M5_Rule1	96.94%	42
M5_Rule2	95.41%	49
M6_Rule1	95.45%	41
M6_Rule2	93.43%	47

for M4 to M6, the required Stage 2 sample sizes do not exceed 50, while for the M1 to M3 with more complex subgroup structures, the required sample sizes remain below 90. These findings indicate that proposed sample size determination for Stage 2 is effective, allowing early stopping for futility or determining an adequate sample size to achieve the desired power. Consequently, this approach can substantially reduce trial costs, addressing a key

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

concern in practical clinical trials.

5. Real Data Example

In this section, we apply our proposed adaptive enrichment design approach to the dataset from the National Supported Work (NSW) study (LaLonde, 1986).

This dataset concerns the provision of work experience to disadvantaged workers from 1975 to 1978 and includes 722 workers. These workers are allocated to the training group and control group, denoted as $t = 1$ and $t = 0$ respectively. Following Zhang et al. (2025), the earning increases, defined as the difference of earnings between 1975 and 1978, is considered as the response, and we include age, years of education (education), marriage status (married), race (White, Black and Hispanic), unemployment status prior to receiving work experience (unemploy) and the earnings in 1975 (RE75) as possible predictive and prognostic variables. Standardization is applied to all continuous covariates. We seek to examine whether the provision of work experience contributes to higher earning increases.

Given a total of $N = 722$ workers, we assume that we have $N_1 = 400$ workers in Stage 1, with $N_{11} = 100$ and $N_{12} = 300$. We then calculate the required sample size $N_{2,\text{sub}}$ based on results in Stage 1. For illustration,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

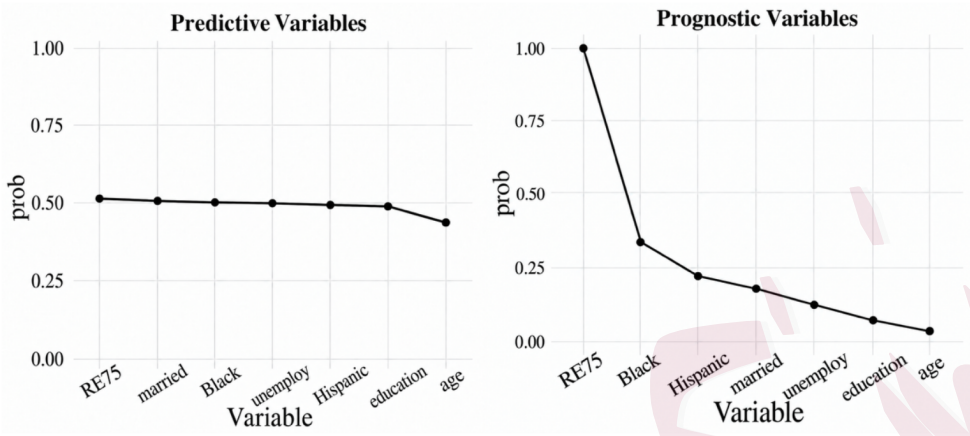

Figure 6: Posterior inclusion probabilities for NSW dataset. The left panel displays the probabilities of predictive variables, and the right panel displays those of prognostic variables.

we randomly draw $N_{2,\text{sub}}$ subjects from the remaining 322 subjects and evaluate the performance. Firstly, we apply BVSA (Zhang et al., 2025) to $N_{11} = 100$ workers and obtain the estimated subgroup membership score. Figure 6 displays the posterior inclusion probabilities of predictive and prognostic variables. For predictive variables, we select all variables since their posterior inclusion probabilities are similar, ranging between 0.4 and 0.55. For prognostic variables, only RE75 is selected, as its posterior inclusion probability equals 1, whereas those of the remaining variables are all below 0.3.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

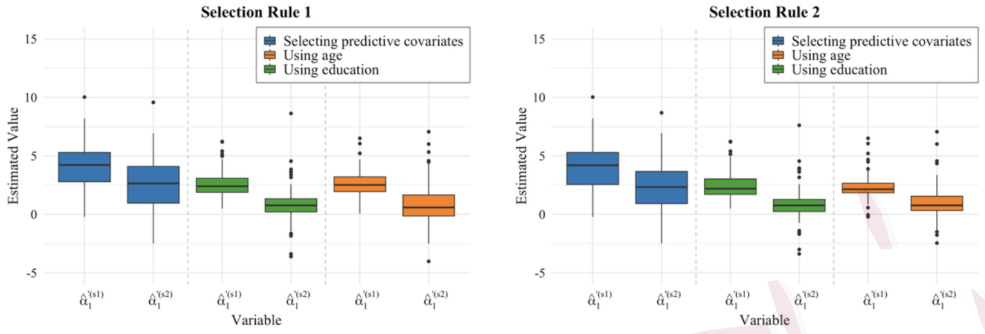

Figure 7: Comparison of ATE estimates from three methods under two selection rules for the NSW dataset, based on 100 simulations, where $\hat{\alpha}_1^{\prime(s1)}$ and $\hat{\alpha}_1^{\prime(s2)}$ denote the Stage 1 and Stage 2 ATE estimates, respectively.

Given the active predictive and prognostic variables, and the subgroup membership score, the threshold c^* is determined using $N_{12} = 300$ workers. The results of the proposed adaptive enrichment design are shown in Figure 7 and Table 3, based on 100 simulations, incorporating sample size determination for Stage 2 and futility stopping. We compare BVSA with single predictive variable approach (Stallard, 2023), using the same set of prognostic variables in both approaches. For the single predictive variable, we use education or age, since education is selected by Zhang et al. (2025) and age is selected by PRIM in Loh et al. (2019).

Figure 7 shows that, under both selection rules, the BVSA-based approach yields larger median subgroup ATE estimates than the single vari-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

Table 3: Comparison of statistical power, average sample size in Stage 2, mean subgroup ATE estimates in Stage 1 and 2, and their standard errors (SEs) for NSW dataset between the predictive variables selection method and the single-variable selection method at a significant level of 5%, based on 100 simulations with sample size determination for Stage 2 and futility stopping.

	Predictive Variable Selection		**Using "education"**		**Using "age"**	
Selection Rule	Rule 1	Rule 2	Rule 1	Rule 2	Rule 1	Rule 2
Statistical Power	**95.00%**	**96.00%**	53.01%	60.87%	41.76%	46.39%
Average Sample Size in Stage 2 (SE)	**41**(3.65)	**60**(7.34)	132(8.81)	155(11.04)	128(8.35)	176(9.38)
Mean of ATE Estimates in Stage 1 (SE)	**4.07**(0.19)	**4.00**(0.20)	2.65(0.13)	2.51(0.15)	2.69(0.11)	2.37(0.11)
Mean of ATE Estimates in Stage 2 (SE)	**2.65**(0.22)	**2.39**(0.21)	0.82(0.18)	0.84(0.19)	0.84(0.18)	0.91(0.14)

able approach Stallard (2023) in both Stage 1 and Stage 2. Table 3 further demonstrates BVSA attains statistical power of at least 90% based on the combined p-value, while those of single variable method do not exceed 61%. In addition, the average sample size in Stage 2 required by BVSA is considerably smaller than that required by their methods. The mean ATE estimates in Stage 1 and 2 are larger when Bayesian variable selection is

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

used. Overall, these results demonstrate that using variable selection in adaptive enrichment design is both cost-efficient and effective in detecting heterogeneous effects.

6. Discussion

In this paper, we propose a two-stage adaptive enrichment design framework based on latent subgroup identification from multiple biomarkers. Under this framework, we are able to select active predictive variables, rather than relying on a continuous biomarker as in Stallard (2023), to accommodate complex subgroup structures. Furthermore, we develop a Stage 2 sample size determination procedure and a futility stopping boundary to reduce experimental costs. Simulation studies and a real-data example demonstrate that the proposed framework achieves favorable performance.

Our proposed method has several limitations. First, its performance depends on the choice of subgroup identification method in Stage 1. The logistic-normal mixture model offers advantages in high-dimensional settings through variable selection, but the resulting subgroup definition is less interpretable and exact recovery of the true subgroup membership at the individual level is generally not feasible. Tree-based methods yield simple and interpretable subgroup definitions, but may suffer from computa-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

tional burden and reduced variable selection accuracy in high-dimensional settings. In addition, the current framework is restricted to a two-stage design with a continuous outcome, and extensions to survival endpoints or designs with multiple interim decision points are left for future work.

Future work can consider more general scenarios. First, we can consider survival time as the response variable, which serves as a primary interest in many clinical trials. Second, we can allow for multiple interim decision points in Stage 1, enabling earlier futility stopping and potentially further reducing trial costs. Third, the subgroup selection criterion could be extended to incorporate more general utility functions, for example by accounting for subgroup size, cost-effectiveness, or benefit-risk tradeoffs, as discussed in Luo and Guo (2023). Such an extension would require further theoretical development, including the derivation of appropriate null distributions for the test statistics under utility-based selection rules.

Supplementary Material

The online Supplementary Material contains the derivation of the subgroup ATE estimators and variance expressions, approximate null distributions of the test statistics, additional numerical results and real data results.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

Acknowledgments

This work was partially supported by the National Nature and Science Foundation of China (12331009).

References

Bauer, P. and K. Kohne (1994). Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics 50*(4), 1029–1041.

Bretz, F., F. Koenig, W. Brannath, E. Glimm, and M. Posch (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine 28*(8), 1181–1217.

Chen, G., H. Zhong, A. Belousov, and V. Devanarayan (2015). A PRIM approach to predictive-signature development for patient stratification. *Statistics in Medicine 34*(2), 317–342.

Frieri, R., W. F. Rosenberger, N. Flournoy, and Z. Lin (2023). Design Considerations for Two-Stage Enrichment Clinical Trials. *Biometrics 79*(3), 2565–2576.

Guo, X., J. Zhou, and X. He (2025). Inference with combined data from subgroup selection and validation phases in clinical trials. *The Annals of Applied Statistics 19*(3), 2088–2104.

Italiano, A. (2011). Prognostic or predictive? It's time to get back to definitions! *Journal of Clinical Oncology 29*(35), 4718–4718.

Lai, T. L., P. W. Lavori, and K. W. Tsang (2019). Adaptive enrichment designs for confirmatory trials. *Statistics in Medicine 38*(4), 613–624.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

REFERENCES

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review 76*(4), 604–620.

Lin, Z., N. Flournoy, and W. F. Rosenberger (2021). Inference for a two-stage enrichment design. *The Annals of Statistics 49*(5), 2697–2720.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica 12*, 361–386.

Loh, W.-Y., L. Cao, and P. Zhou (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *WIREs Data Mining and Knowledge Discovery 9*(5), e1326.

Luo, Y. and X. Guo (2023). Inference on tree-structured subgroups with subgroup size and subgroup effect relationship in clinical trials. *Statistics in Medicine 42*(27), 5039–5053.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*(3), 655–660.

Rosenblum, M., E. X. Fang, and H. Liu (2020). Optimal, two-stage, adaptive enrichment designs for randomized trials, using sparse linear programming. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*(3), 749–772.

Seibold, H., A. Zeileis, and T. Hothorn (2016). Model-based recursive partitioning for subgroup analyses. *The International Journal of Biostatistics 12*(1), 45–63.

Shen, J. and X. He (2015). Inference for Subgroup Analysis With a Structured Logistic-Normal

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

Mixture Model. *Journal of the American Statistical Association 110*(509), 303–312.

Simon, R. and N. Simon (2017). Inference for multimarker adaptive enrichment trials. *Statistics in Medicine 36*(26), 4083–4093.

Stallard, N. (2023). Adaptive enrichment designs with a continuous biomarker. *Biometrics 79*(1), 9–19.

Stallard, N. and P. K. Kimani (2018). Uniformly minimum variance conditionally unbiased estimation in multi-arm multi-stage clinical trials. *Biometrika 105*(2), 495–501.

Takeishi, S. (2025). A shrinkage likelihood ratio test for high-dimensional subgroup analysis with a logistic-normal mixture model. *Electronic Journal of Statistics 19*(1), 1942–1980.

Wang, J. and X. He (2023). Subgroup analysis and adaptive experiments crave for debiasing. *WIREs Computational Statistics 15*(6), e1614.

Zeileis, A., T. Hothorn, and K. Hornik (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics 17*(2), 492–514.

Zhan, D., Y. Ouyang, F. Vila-Rodriguez, M. E. Karim, and H. Wong (2025). Bayesian adaptive enrichment design in multi-arm clinical trials: The bayesaet package for r users. *Computer Methods and Programs in Biomedicine 268*, 108833.

Zhang, R., N. N. Narisetty, X. He, and J. Shen (2025). Bayesian variable selection on structured logistic-normal mixture models for subgroup analysis. *Electronic Journal of Statistics 19*(1), 2876–2922.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0498

REFERENCES

Zhao, L., L. Tian, T. Cai, B. Claggett, and L.-J. Wei (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association 108*(502), 527–539.

Mingde Wu

Department of Statistics and Data Science, Fudan University, Shanghai, China.

E-mail: 25110690005@m.fudan.edu.cn

Juan Shen

Department of Statistics and Data Science, Fudan University, Shanghai, China.

E-mail: shenjuan@fudan.edu.cn

Ruqian Zhang

Department of Statistics and Data Science, Fudan University, Shanghai, China.

E-mail: rqzhang20@fudan.edu.cn

Jingfei Zhang

Goizueta Business School, Emory University, Atlanta, GA 30322, USA.

E-mail: emma.zhang@emory.edu