# 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*

**Supplementary Material**

In this supplement, we provide derivation of the subgroup ATE estimators and variance expressions in Section S1, approximate null distributions of the test statistics in Section S2, additional numerical results in Section S3, and additional real data results in Section S4.

## S1 Derivation of the Subgroup ATE Estimators and Variance Expressions

Consider Model (2.3) in Section 2.1:

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

where $\epsilon_i^{(j)} \sim \mathcal{N}(0, {\sigma^{(j)}}^2)$ are independent errors with unknown variance ${\sigma^{(j)}}^2$, and $\hat{\delta}_i^{(j)} = I(\hat{g}(\boldsymbol{x}_i) \geq c_j)$. 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 $\boldsymbol{Z}^{(j)}$

denote the design matrices for prognostic covariates. Similarly, for the remaining subjects satisfying $\hat{\delta}_i^{(j)} = 0$, define $T^{(-j)}$, $y^{(-j)}$, $\epsilon^{(-j)}$ and $\boldsymbol{Z}^{(-j)}$. The model can be expressed in matrix form as

$$y^{(j)} = \begin{pmatrix} T^{(j)} & \boldsymbol{Z}^{(j)} \end{pmatrix} \begin{pmatrix} \alpha_{1,j} & \boldsymbol{\beta}^{(j)} \end{pmatrix}^\top + \epsilon^{(j)},$$

To estimate $\alpha_{1,j}$, we solve the following equation:

$$\begin{pmatrix} {T^{(j)}}^\top T^{(j)} & {T^{(j)}}^\top \boldsymbol{Z}^{(j)} \\ {\boldsymbol{Z}^{(j)}}^\top T^{(j)} & {\boldsymbol{Z}^{(j)}}^\top \boldsymbol{Z}^{(j)} \end{pmatrix} \begin{pmatrix} \alpha_{1,j} \\ \boldsymbol{\beta}^{(j)} \end{pmatrix} = \begin{pmatrix} {T^{(j)}}^\top y^{(j)} \\ {\boldsymbol{Z}^{(j)}}^\top y^{(j)} \end{pmatrix}.$$

For $m_1 \times m_2$ and $m_1 \times m_3$ matrices $U$ and $V$,

$$\begin{pmatrix} U^\top U & U^\top V \\ V^\top U & V^\top V \end{pmatrix}^{-1} = \begin{pmatrix} (U^\top U)^{-1} + U_V (V^\top N_U V)^{-1} U_V^\top & -U_V (V^\top N_U V)^{-1} \\ -(V^\top N_U V)^{-1} U_V^\top & (V^\top N_U V)^{-1} \end{pmatrix}, \tag{S1.1}$$

where $U_V = (U^\top U)^{-1} U^\top V$, $N_U = I - P_U$ and $P_U = U(U^\top U)^{-1} U^\top$, $I$ is an identity matrix of suitable dimension. To simplify the notations, we first treat $T^{(j)}$ and $\boldsymbol{Z}^{(j)}$ as $U$ and $V$, and based on Equation (S1.1), we can obtain

$$\begin{aligned}
\hat{\alpha}_{1,j} &= \left[(U^\top U)^{-1} + U_V(V^\top N_U V)^{-1}U_V^\top\right] U^\top y^{(j)} - U_V(V^\top N_U V)^{-1}V^\top y^{(j)} \\
&= \left[(U^\top U)^{-1}U^\top + (U^\top U)^{-1}U^\top V(V^\top N_U V)^{-1}V^\top U(U^\top U)^{-1}U^\top \right. \\
&\quad \left. -(U^\top U)^{-1}U^\top V(V^\top N_U V)^{-1}V^\top\right] y^{(j)} \\
&= \left[(U^\top U)^{-1}U^\top + (U^\top U)^{-1}U^\top V(V^\top N_U V)^{-1}V^\top \left(U(U^\top U)^{-1}U^\top - I\right)\right] y^{(j)} \\
&= \left[(U^\top U)^{-1}U^\top - (U^\top U)^{-1}U^\top V(V^\top N_U V)^{-1}V^\top N_U\right] y^{(j)} \\
&= (U^\top U)^{-1}U^\top \left[I - V(V^\top N_U V)^{-1}V^\top N_U\right] y^{(j)}.
\end{aligned} \tag{S1.2}$$

Then, based on Equation (S1.2), the variance of $\hat{\alpha}_{1,j}$, denoted as $\sigma^2_{\hat{\alpha}_{1,j}}$, is given by

$$\begin{aligned}
\sigma^2_\alpha &= \sigma^2_{1,j}(U^\top U)^{-1}U^\top \left[I - V(V^\top N_U V)^{-1}V^\top N_U\right] \\
&\quad \left[I - N_U V(V^\top N_U V)^{-1}V^\top\right] U(U^\top U)^{-1} \\
&= \sigma^2_{1,j}(U^\top U)^{-1}U^\top [I - N_U V(V^\top N_U V)^{-1}V^\top \\
&\quad - V(V^\top N_U V)^{-1}V^\top N_U + V(V^\top N_U V)^{-1}V^\top] U(U^\top U)^{-1}.
\end{aligned}$$

In summary, we obtain

$$\begin{aligned}
\hat{\alpha}_{1,j} &= \left(T^{(j)\top}T^{(j)}\right)^{-1} T^{(j)\top}\left[I - \boldsymbol{Z}^{(j)}\left(\boldsymbol{Z}^{(j)\top}N_{T^{(j)}}\boldsymbol{Z}^{(j)}\right)^{-1}\boldsymbol{Z}^{(j)\top}N_{T^{(j)}}\right] y^{(j)}, \\
\sigma^2_{\hat{\alpha}_{1,j}} &= \sigma^{(j)2}(T^{(j)\top}T^{(j)})^{-1}T^{(j)\top}\left[I + \boldsymbol{Z}^{(j)}\left(\boldsymbol{Z}^{(j)\top}N_{T^{(j)}}\boldsymbol{Z}^{(j)}\right)^{-1}\boldsymbol{Z}^{(j)\top}\right] \\
&\quad T^{(j)}\left(T^{(j)\top}T^{(j)}\right)^{-1},
\end{aligned} \tag{S1.3}$$

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

Similarly, we obtain

$$\begin{aligned}\hat{\alpha}_{2,j} =& \left(T^{(-j)\top}T^{(-j)}\right)^{-1} T^{(-j)\top}\left[I - \boldsymbol{Z}^{(-j)}\left(\boldsymbol{Z}^{(-j)\top}N_{T^{(-j)}}\boldsymbol{Z}^{(-j)}\right)^{-1}\right.\\ &\left.\boldsymbol{Z}^{(-j)\top}N_{T^{(-j)}}\right] y^{(-j)},\\ \sigma^2_{\hat{\alpha}_{2,j}} =& \sigma'^{(j)^2}(T^{(-j)\top}T^{(-j)})^{-1}T^{(-j)\top}\left[I + \boldsymbol{Z}^{(-j)}\left(\boldsymbol{Z}^{(-j)\top}N_{T^{(-j)}}\boldsymbol{Z}^{(-j)}\right)^{-1}\right.\\ &\left.\boldsymbol{Z}^{(-j)\top}\right] T^{(-j)}\left(T^{(-j)\top}T^{(-j)}\right)^{-1},\end{aligned} \tag{S1.4}$$

where $N_{T^{(-j)}} = I - P_{T^{(-j)}}$ and $P_{T^{(-j)}} = T^{(-j)}(T^{(-j)\top}T^{(-j)})^{-1}T^{(-j)\top}$, $\sigma'^{(j)^2}$ denotes the unknown variance of independent errors in the Model for the complement subgroup.

To simplify the notations, we define $\boldsymbol{D}^{(j)} = (y^{(j)}, T^{(j)}, \boldsymbol{Z}^{(j)}, \boldsymbol{X}^{(j)}, P_{T^{(j)}}, N_{T^{(j)}})$ and $\boldsymbol{D}^{(-j)} = (y^{(-j)}, T^{(-j)}, \boldsymbol{Z}^{(-j)}, \boldsymbol{X}^{(-j)}, P_{T^{(-j)}}, N_{T^{(-j)}})$. 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. In (S1.3), substituting $D^{(j)}$ with $D^{(s1)}$ for Stage 1 or $D^{(s2)}$ for Stage 2, and substituting $\sigma^{(j)^2}$ with $\sigma'^2$ yields the subgroup ATE estimators $\hat{\alpha}_1^{'(s1)}$ and $\hat{\alpha}_1^{'(s2)}$ for Stages 1 and 2, respectively, together with their estimated variances.

# S2 Approximate Null Distributions of the Test Statistics

Following the work of Stallard (2023), we first introduce the notations used in the computation of approximate null distributions of the test statistics. For $j = 1, \ldots, k$, let $\boldsymbol{A}_{1:k}^{(j)}$ be a $k \times k$ matrix whose diagonal elements are 1, whose $j$th column has entries -1 except for the $(j, j)$ element, and whose remaining entries are 0. For $1 \leq i < l \leq k$, define $\boldsymbol{A}_{i:l}^{(j)}$ as the submatrix of $\boldsymbol{A}_{1:k}^{(j)}$ obtained by extracting rows i to l and columns i to l. For a vector $\boldsymbol{v} = (v_1, \ldots, v_k)^\top$, let $\boldsymbol{v}_{i:l} = (v_i, \ldots, v_l)^\top$ for $1 \leq i < l \leq k$, and let $\boldsymbol{v}^{-1} = (v_1^{-1}, \ldots, v_k^{-1})^\top$. Define $\boldsymbol{\Sigma}(\boldsymbol{v})$ as the $k \times k$ matrix with entries $\boldsymbol{\Sigma}(\boldsymbol{v})_{i,j} = v_{\max\{i,j\}}$. Finally, define $\boldsymbol{C}(\boldsymbol{v})$ as a $(k-1) \times k$ matrix with $\boldsymbol{C}(\boldsymbol{v})_{i,i} = v_k(v_k - v_i)^{-1}$, $\boldsymbol{C}(\boldsymbol{v})_{i,k} = -\boldsymbol{C}(\boldsymbol{v})_{i,i}$, and all other entries equal to 0 for $i = 1, \ldots, k-1$.

For testing $H_{0,J^{(r)}} : \alpha_{1,J^{(r)}} \leq 0$, where $\alpha_{1,J^{(r)}}$ denotes the average treatment effect within $J^{(r)}$th subgroup based on Selection Rule $r$ for $r = 1, 2$, let $z^{(1)} = \hat{\alpha}_1^{\prime(s1)} / \hat{\sigma}^2_{\hat{\alpha}_1^{\prime(s1)}}$ be the test statistic for Selection Rule 1, and $z^{(2)} = (\hat{\alpha}_1^{\prime(s1)} - \hat{\alpha}_2^{\prime(s1)})(\mathrm{var}(\hat{\alpha}_1^{\prime(s1)} - \hat{\alpha}_2^{\prime(s1)}))^{-1/2}$ be the test statistic for Selection Rule 2. The approximate null distributions of the test statistics, for $i = 1, \ldots, J^{(r)}$ under Selection Rule $r$ for $r = 1, 2$, are proposed by Stallard

(2023):

$$
\begin{aligned}
F_i^{(1)}(z) &= \sum_{j=i}^{k} \Phi^{(k-i+1)} \left( z n_j^{-1/2} \boldsymbol{A}_{i:k}^{(j)} \mathbf{1}, \mathbf{0}, \boldsymbol{A}_{i:k}^{(j)} \boldsymbol{\Sigma}(\boldsymbol{n}_{i:k}^{-1}) \boldsymbol{A}_{i:k}^{(j)\top} \right), \\
F_i^{(2)}(z) &= \sum_{j=i}^{k-1} \Phi^{(k-i)} \Big( z n_k (n_k - n_j)^{-1} n_j^{-1/2} \boldsymbol{A}_{i:k-1}^{(j)} \mathbf{1}, \mathbf{0}, \boldsymbol{C}^{(j)}(\boldsymbol{n}_{i:k}) \\
&\qquad \boldsymbol{\Sigma}(\boldsymbol{n}_{i:k}^{-1}) \boldsymbol{C}^{(j)}(\boldsymbol{n}_{i:k})^\top + n_k (n_k - n_j)^{-2} \boldsymbol{A}_{i:k-1}^{(j)} \mathbf{1}\mathbf{1}^\top \boldsymbol{A}_{i:k-1}^{(j)\top} \Big),
\end{aligned}
\tag{S2.5}
$$

where $F_i^{(r)}(z)$ corresponds to test statistic under Selection Rule $r$, $z = z^{(1)}$ or $z = z^{(2)}$, $\boldsymbol{C}^{(j)}(\boldsymbol{n}_{i:k}) = \boldsymbol{A}_{i:k-1}^{(j)} \boldsymbol{C}(\boldsymbol{n}_{i:k})$, and $\boldsymbol{n} = (n_1, \ldots, n_k)^\top$, $n_j$ is the sample size within the $j$th subgroup.

# S3 Additional Numerical Results

## S3.1 Comparison of Subgroup Identification Methods

In this subsection, we assess more subgroup selection methods, including MOB (Zeileis et al., 2008) and GUIDE (Loh, 2002) in the comparison of subgroup ATE estimates, to evaluate different subgroup identification methods in adaptive enrichment designs when subgroups are defined by multiple biomarkers.

Following Section 4.2, we focus on M5 and M6, and report the results in Figure 1 based on 200 simulations. 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

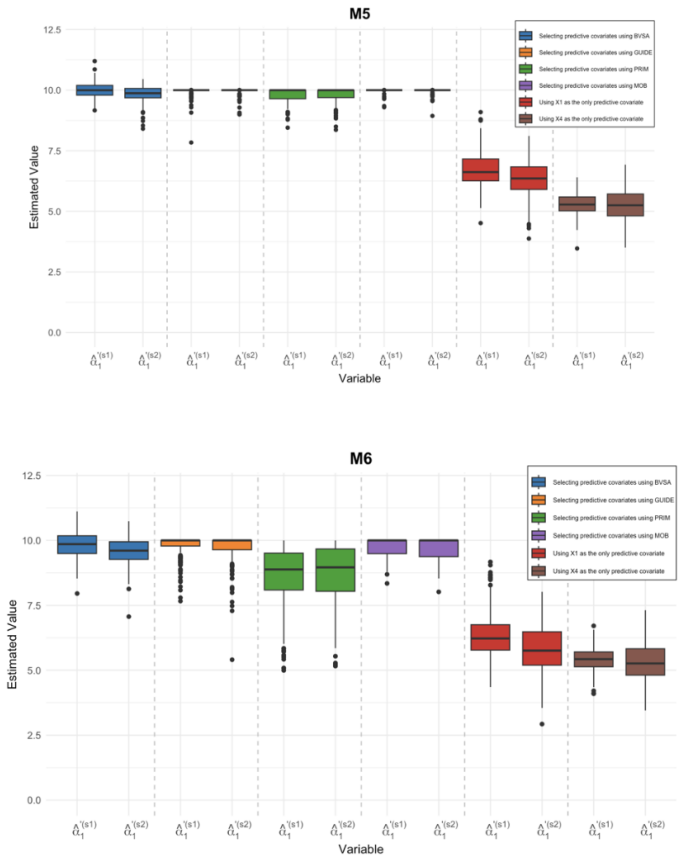


Figure 1: Comparison of subgroup ATE estimates obtained from six methods, where $\hat{\alpha}_1^{'(s1)}$ and $\hat{\alpha}_1^{'(s2)}$ denote the subgroup ATE estimates in Stage 1 and Stage 2, respectively.

are determined using BVSA (Zhang et al., 2025) and kept identical across all methods. Results from four predictive variable selection approaches show satisfactory performance in setting M5, with ATE estimates close to the true value. In contrast, the methods relying on a single predictive variable (Stallard, 2023) 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, GUIDE and MOB achieve the best estimation performance. PRIM (Chen et al., 2015) performs slightly worse, with lower estimates and larger variance, but remains satisfactory, while the single predictive variable methods again perform poorly, with median estimates not exceeding 65% of the true effect. Overall, across these two settings, variable selection approaches consistently outperform single predictive variable methods. Table 1 further provides numerical summaries of the subgroup ATE estimates, including empirical mean, bias, MSE, and standard error.

### S3.2 Oracle Comparison and Robustness to Data Splitting

To assess the efficiency loss due to data splitting and the robustness of the proposed procedure to different splits, we conduct additional numerical experiments under Settings M1-M3 with $\alpha_1 = 0.9$ and $\alpha_2 = -0.5$, based on 200 simulations at the 5% significance level.

We consider two oracle benchmarks: (i) an oracle assuming the true subgroup membership scores are known, and (ii) an oracle assuming the true predictive variables are known. Table 2 reports the statistical power and average Stage 2 sample size for the proposed method and the two

Table 1: Numerical summaries of subgroup ATE estimates in Stages 1 and 2 for Models M5 and M6, including mean, bias, MSE, and standard error, based on 200 simulations under $\alpha_1 = 10$ and $\alpha_2 = 0$.

| | M5 | | | | M6 | | | |
|---|---|---|---|---|---|---|---|---|
| | Bias in Stage 1 (MSE) | Bias in Stage 2 (MSE) | Mean in Stage 1 (SE) | Mean in Stage 2 (SE) | Bias in Stage 1 (MSE) | Bias in Stage 2 (MSE) | Mean in Stage 1 (SE) | Mean in Stage 2 (SE) |
| BVSA | -0.01 (0.098) | -0.16 (0.144) | 9.99 (0.022) | 9.84 (0.024) | -0.17 (0.290) | -0.44 (0.493) | 9.83 (0.036) | 9.56 (0.039) |
| GUIDE | -0.05 (0.041) | -0.04 (0.024) | 9.95 (0.014) | 9.96 (0.011) | -0.21 (0.230) | -0.28 (0.423) | 9.79 (0.031) | 9.72 (0.042) |
| PRIM | -0.22 (0.143) | -0.20 (0.142) | 9.78 (0.022) | 9.80 (0.022) | -1.40 (3.535) | -1.34 (3.378) | 8.60 (0.089) | 8.66 (0.089) |
| MOB | -0.01 (0.006) | -0.02 (0.010) | 9.99 (0.006) | 9.98 (0.007) | -0.25 (0.200) | -0.33 (0.291) | 9.75 (0.026) | 9.67 (0.030) |
| $x_1$ | -3.25 (11.178) | -3.67 (14.007) | 6.75 (0.054) | 6.33 (0.053) | -3.60 (13.778) | -4.22 (18.537) | 6.40 (0.065) | 5.78 (0.060) |
| $x_4$ | -4.70 (22.275) | -4.74 (22.950) | 5.30 (0.033) | 5.26 (0.051) | -4.59 (21.285) | -4.71 (22.722) | 5.41 (0.033) | 5.29 (0.054) |

oracle benchmarks. As expected, the oracle with known true subgroup membership scores achieves the highest power and requires the smallest Stage 2 sample size. The oracle based on known true predictive variables performs closer to the proposed method. The proposed method shows some efficiency loss relative to the oracle benchmarks, but maintains high power across all settings.

To evaluate robustness to the choice of data split in Stage 1, we repeat the splitting procedure 10 times under different random splits and report the

resulting combined p-values in Table 3. The results show that the combined p-values are significant for most random splits across most scenarios, with some variability present due to the randomness of data splitting, indicating that the proposed procedure is reasonably robust to the choice of split.

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 $\alpha_1 = 0.9$ and $\alpha_2 = -0.5$.

| | **Statistical Power** | | | **Average Stage 2 Sample Size** | | |
|---|---|---|---|---|---|---|
| | Proposed | Known True Subgroup Membership Scores | Known True Predictive Variables | Proposed | Known True Subgroup Membership Scores | Known True Predictive Variables |
| M1_Rule1 | 97.69% | 98.50% | 96.30% | 49 | 26 | 42 |
| M1_Rule2 | 94.97% | 98.98% | 96.65% | 60 | 35 | 66 |
| M2_Rule1 | 96.36% | 98.98% | 96.39% | 41 | 28 | 43 |
| M2_Rule2 | 93.59% | 100.00% | 94.25% | 56 | 29 | 59 |
| M3_Rule1 | 94.33% | 99.00% | 96.70% | 51 | 26 | 49 |
| M3_Rule2 | 96.92% | 98.96% | 95.65% | 60 | 33 | 57 |

## S3.3 Comparison with Alternative Testing Procedures

To investigate the potential conservativeness of the global testing step in Stage 1, we compare the proposed method with two alternative procedures:

Table 3: Combined p-value based on different random data splits in stage 1, with sample size determination for Stage 2 targeting 90% power and early futility stopping at the 5% significance level under $\alpha_1 = 0.9$ and $\alpha_2 = -0.5$. Values less than 0.001 are reported as "<0.001"; otherwise, p-values are reported as decimals rounded to three decimal places.

| | Combined P-value | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Split 1 | Split 2 | Split 3 | Split 4 | Split 5 | Split 6 | Split 7 | Split 8 | Split 9 | Split 10 |
| M1_Rule1 | 0.001 | <0.001 | 0.463 | <0.001 | <0.001 | <0.001 | <0.001 | 0.002 | 0.022 | <0.001 |
| M1_Rule2 | 0.055 | 0.106 | 0.001 | 0.002 | 0.003 | 0.017 | <0.001 | 0.003 | <0.001 | 0.002 |
| M2_Rule1 | 0.001 | <0.001 | 0.012 | <0.001 | 0.044 | <0.001 | <0.001 | 0.002 | <0.001 | <0.001 |
| M2_Rule2 | 0.025 | 0.002 | <0.001 | 0.040 | <0.001 | 0.002 | <0.001 | <0.001 | <0.001 | <0.001 |
| M3_Rule1 | <0.001 | 0.004 | <0.001 | 0.007 | 0.065 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| M3_Rule2 | 0.133 | <0.001 | 0.121 | 0.005 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| M4_Rule1 | <0.001 | <0.001 | 0.001 | <0.001 | <0.001 | 0.002 | <0.001 | <0.001 | <0.001 | <0.001 |
| M4_Rule2 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| M5_Rule1 | 0.003 | 0.005 | 0.005 | 0.001 | <0.001 | <0.001 | <0.001 | <0.001 | 0.001 | 0.004 |
| M5_Rule2 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| M6_Rule1 | <0.001 | 0.002 | <0.001 | <0.001 | 0.003 | 0.004 | 0.001 | <0.001 | <0.001 | 0.018 |
| M6_Rule2 | <0.001 | <0.001 | <0.001 | <0.001 | 0.001 | <0.001 | 0.002 | <0.001 | <0.001 | <0.001 |

(i) an approach relying solely on the Stage 2 p-value, and (ii) the method proposed by Guo et al. (2025). The results are summarized in Table 4, based on 200 simulations under $\alpha_1 = 0.9$ and $\alpha_2 = -0.5$. For the proposed method, both Stage 2 sample size determination and early futility stopping are incorporated. For the other two approaches, the Stage 2 sample size is fixed at 100, without early futility stopping.

The results show that relying solely on the Stage 2 p-value yields sub-

stantially lower power than the proposed method across all settings. For the method of Guo et al. (2025), in correctly specified settings M1-M3, the proposed method achieves higher power, whereas in misspecified settings M4-M6, the method of Guo et al. (2025) achieves slightly higher power. This suggests that the method of Guo et al. (2025) may serve as a useful alternative when applicable, though its direct application is limited to settings corresponding to Selection Rule 1. Overall, the power advantage of the proposed method is largely attributable to the adaptive Stage 2 sample size determination, which compensates for the conservativeness introduced by the global testing step.

### S3.4 Sensitivity Analysis under Model Misspecification

To further assess the robustness of the proposed method to model misspecification, we conduct sensitivity analyses under three additional scenarios for Settings M1–M3:

(i) replacing the normal error $\epsilon_i \sim N(0,1)$ with heavier-tailed $t(3)$ noise;

(ii) adding an interaction term $X_1X_2$ to the outcome model;

(iii) adding a quadratic term $X_2^2$ to the outcome model.

These scenarios introduce deviations from the logistic-normal mixture model assumed in Stage 1.

Table 4: Comparison of statistical power based on 200 simulations under $\alpha_1 = 0.9$ and $\alpha_2 = -0.5$. For the proposed method, both Stage 2 sample size determination and early futility stopping are incorporated. For the other two approaches, the Stage 2 sample size is fixed at 100, without early futility stopping.

| | Statistical Power | | |
|---|---|---|---|
| | Proposed | Relying Solely on the Stage 2 P-value | The Method Proposed by Guo et al. (2025) |
| M1_Rule1 | 97.69% | 74.50% | 91.50% |
| M1_Rule2 | 94.97% | 66.00% | – |
| M2_Rule1 | 96.36% | 68.50% | 90.50% |
| M2_Rule2 | 93.59% | 62.00% | – |
| M3_Rule1 | 94.33% | 57.00% | 86.00% |
| M3_Rule2 | 96.92% | 53.00% | – |
| M4_Rule1 | 97.95% | 93.50% | 99.00% |
| M4_Rule2 | 98.40% | 89.50% | – |
| M5_Rule1 | 95.31% | 86.50% | 97.00% |
| M5_Rule2 | 98.34% | 77.50% | – |
| M6_Rule1 | 99.49% | 93.00% | 99.50% |
| M6_Rule2 | 99.47% | 87.00% | – |

The results are reported in Table 5, based on 200 simulations under $\alpha_1 = 0.9$ and $\alpha_2 = -0.5$. Across all settings and selection rules, the proposed method maintains statistical power at or above 80% under model misspecification, though some power loss is observed relative to the correctly specified case, particularly when interaction or quadratic terms are present. The required Stage 2 sample sizes increase accordingly, reflect-

ing the greater difficulty of subgroup identification under misspecification. These results confirm that the testing procedure, which depends on the working model only through the estimated subgroup membership scores used for thresholding, is reasonably robust to moderate deviations from the working model.

Table 5: Sensitivity analysis of statistical power and average Stage 2 sample size, based on 200 simulations, with Stage 2 sample size determination targeting 90% power and early futility stopping at the 5% significance level under $\alpha_1 = 0.9$ and $\alpha_2 = -0.5$.

| | **Statistical Power** | | | | **Average Stage 2 Sample Size** | | | |
|---|---|---|---|---|---|---|---|---|
| | Baseline | $t(3)$ Noise | Interaction Term $X_1X_2$ Added | Quadratic Term $X_2^2$ Added | Baseline | $t(3)$ Noise | Interaction Term $X_1X_2$ Added | Quadratic Term $X_2^2$ Added |
| M1_Rule1 | 97.69% | 89.39% | 93.27% | 87.39% | 49 | 86 | 73 | 64 |
| M1_Rule2 | 94.97% | 90.63% | 87.37% | 83.15% | 60 | 94 | 90 | 79 |
| M2_Rule1 | 96.36% | 95.45% | 85.86% | 84.17% | 41 | 89 | 69 | 67 |
| M2_Rule2 | 93.59% | 92.06% | 91.21% | 85.00% | 56 | 104 | 87 | 81 |
| M3_Rule1 | 94.33% | 94.83% | 93.33% | 90.91% | 51 | 97 | 80 | 81 |
| M3_Rule2 | 96.92% | 96.30% | 96.51% | 89.47% | 60 | 96 | 83 | 88 |

## S3.5 FWER Control Beyond the Global Null

We note that under the nested subgroup structure, constructing a beyond-global-null configuration necessarily violates the assumption in Stallard (2023) that the outcome is positively associated with a biomarker. Con-

sequently, in our correctly specified model, it is not possible to establish a configuration that extends beyond the global null hypothesis. To assess FWER control beyond the global null, we consider the following setting:

$$Y = 1 + X_1 + X_2 + X_4 + I(X_6 = 2) + X_7 + \alpha_1 t \times I(-1.5 < X_1 < 0.5) + \epsilon,$$

where $\alpha_1 = 1.5$. The data are generated following the same procedure as described in Section 4.1, except for the outcome model specified above. This setting violates the positive association assumption in Stallard (2023) as the treatment effect is nonzero only within the interior region $-1.5 < X_1 < 0.5$.

We evaluate the Type I error as the proportion of simulations in which the optimal threshold $c^*$ exceeds 0.5, among those in which only $X_1$ is selected as the predictive variable. In this case, the threshold set $\mathbb{C}$ is simply constructed as the deciles of $X_1$ rather than the estimated subgroup membership score $\hat{g}(\boldsymbol{x}_i)$. A threshold exceeding 0.5 would correspond to incorrectly selecting a subgroup that excludes part of the truly beneficial region defined by $-1.5 < X_1 < 0.5$.

The results are summarized in Table 6. When $\alpha_1 = 1.5$, among the 200 simulations, 138 simulations select only $X_1$ as the predictive variable, among which 2 simulations select a threshold exceeding 0.5 under Selection Rule 1, yielding an empirical Type I error of 1.45%, while no simulation exceeds the threshold under Selection Rule 2. The results are well below

the nominal level of 5%, providing empirical evidence that the proposed procedure controls the Type I error beyond the global null.

Table 6: Empirical Type I error rates beyond the global null among simulations in which only $X_1$ is selected as the predictive variable, based on 200 simulations at the 5% significance level, where the true subgroup is defined by $I(-1.5 < X_1 < 0.5)$ with $\alpha_1 = 1.5$ and $\alpha_2 = 0$.

| | **Type I Error** |
|---|---|
| | $\alpha_1 = 1.5$ |
| Rule1 | 1.45% |
| Rule2 | 0% |

## S4 Additional Real Data Results

In this section, we apply our proposed adaptive enrichment design approach to the AIDS Clinical Trials Group (ACTG) 320 study data. These 852 patients are randomly allocated to one of the two treatment regimens. In the experimental group, patients are administered the new drug in combination with the placebo, i.e., $t = 1$, whereas the control group indicates the patients only receive the placebo, i.e., $t = 0$. Following Shen and He (2015) and Zhang et al. (2025), CD4 count change at 24th week is used as the response, and we include sex, age, weight, race (categories: White =

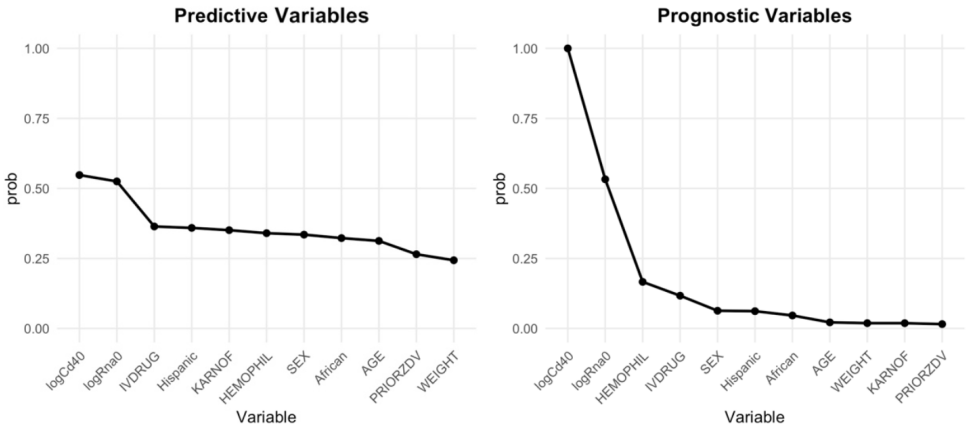


Figure 2: Posterior inclusion probabilities for ACTG 320 study data. The left panel displays the probabilities of predictive variables, and the right panel displays those of prognostic variables.

1, African = 2, Hispanic = 3), injection-drug use (IVDRUG), hemophilia (HEMOPHIL), Karnofsky score (KARNOF), months of prior zidovudine therapy (PRIORZDV), baseline CD4 counts and baseline RNA concentration on the logarithm scale (denoted as logCd40 and logRna0, respectively) as the possible predictive and prognostic variables. Standardization is applied to all continuous covariates. We aim to investigate whether the new drug has a beneficial effect on the progression of AIDS.

Given a total of $N = 852$ patients, we assume that we have $N_1 = 450$ patients in Stage 1, with $N_{11} = 150$ and $N_{12} = 300$, and $N_{2,\text{sub}}$ is determined

based on results in Stage 1. For illustration, we randomly draw $N_{2,\text{sub}}$ patients from the remaining 402 patients and evaluate the performance. Firstly, we apply BVSA (Zhang et al., 2025) to $N_{11} = 150$ patients and obtain the estimated subgroup membership score. Figure 2 presents the posterior inclusion probabilities of predictive and prognostic variables. For predictive variables, we select logCd40 and logRna0 since their probabilities exceed 0.5, while the probabilities of other variables are less than 0.4. These two predictive variables are also selected by Cai et al. (2010) and Zhang et al. (2025). Similarly, for prognostic variables, we select logCd40 and logRna0 since their probabilities are larger than 0.5, while others are no more than 0.25.

Given the active predictive and prognostic variables, and the subgroup membership score, the threshold $c^*$ is determined using $N_{12} = 300$ patients. The results of the proposed adaptive enrichment design are shown in Figure 3 and Table 7, based on 100 simulations, incorporating sample size calculation for Stage 2 and futility stopping. We set the minimal sample size in Stage 2 to be 20. We conduct a comparison between BVSA (Zhang et al., 2025) and the single predictive variable approach (Stallard, 2023), using the same set of prognostic variables in both approaches. For the single predictive variable, we use logCd40 or logRna0.

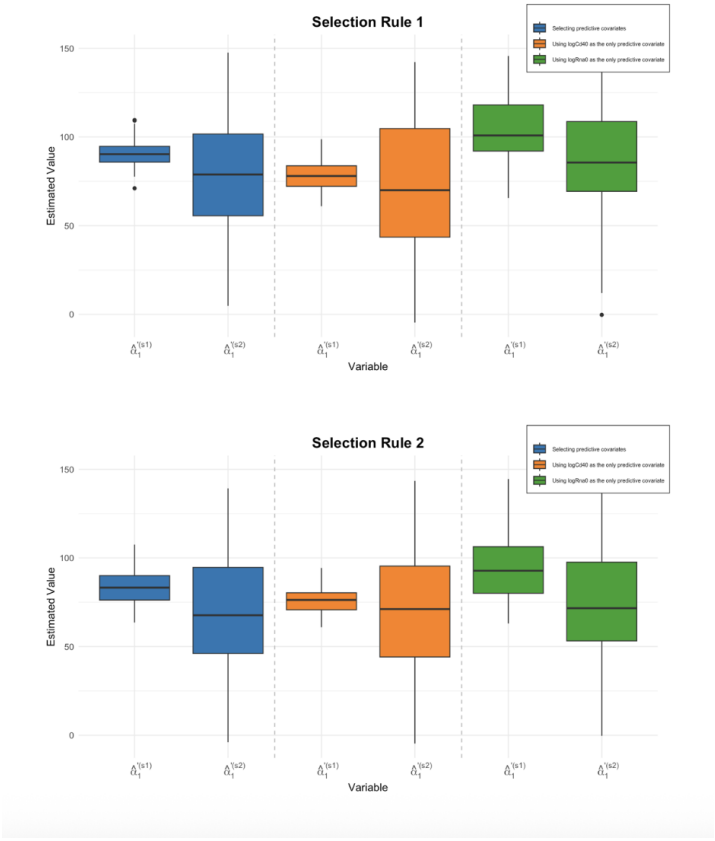


Figure 3: Comparison of ATE estimates from three methods under two selection rules for the ACTG dataset, based on 100 simulations, where $\hat{\alpha}_1^{'(s1)}$ and $\hat{\alpha}_1^{'(s2)}$ denote the Stage 1 and Stage 2 ATE estimates, respectively.

Figure 3 shows that, under both selection rules, the BVSA-based approach produces ATE estimates similar to those from the single predictive variable approach, with both methods demonstrating good performance. Table 7 indicates that these two methods attain 100% statistical power at a significant level of 5%, and the average sample sizes in Stage 2 required

Table 7: 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 ACTG 320 study data 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 calculation for Stage 2 and futility stopping.

| | **Predictive Variable Selection** | | **Using "logCd40"** | | **Using "logRna0"** | |
|---|---|---|---|---|---|---|
| **Selection Rule** | Rule 1 | Rule 2 | Rule 1 | Rule 2 | Rule 1 | Rule 2 |
| **Statistical Power** | 100% | 100% | 100% | 100% | 100% | 100% |
| **Average Sample Size in Stage 2** | 20(0) | 20(0) | 20(0) | 20(0) | 23(0.41) | 21 |
| **Mean of ATE Estimates in Stage 1 (SE)** | 90.62(0.73) | 82.73(0.88) | 78.08(0.79) | 76.24(0.72) | 111.43(2.52) | 101.24(2.76) |
| **Mean of ATE Estimates in Stage 2 (SE)** | 82.44(4.06) | 74.78(4.32) | 76.08(4.11) | 73.04(3.89) | 95.61(4.10) | 85.46(4.53) |

by selecting variables and using "logCd40" are both equal to 20, which are slightly smaller than that required by using "logRna0". BVSA yields mean estimate higher than that from using "logCd40" but with greater variability, and lower than that from "logRna0" while being more precise. Overall, these results suggest that, for the ACTG 320 data, using variable selection in adaptive enrichment design is comparable to, and not inferior to, the method of Stallard (2023).

## References

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

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.

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.

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

Shen, J. and X. He (2015). Inference for Subgroup Analysis With a Structured Logistic-Normal Mixture Model. *Journal of the American Statistical Association 110*(509), 303–312.

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

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

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.