

Two-way subgroup structure learning for high dimensional clustered survival data

Ye He¹, Dongsheng Tu², and Liu Liu³, and Ling Zhou ⁴

¹ *School of Mathematics Science, Sichuan Normal University, China*

² *Canadian Cancer Trials Group, Queen's University, Kingston, Canada*

³ *School of Mathematics Science, Chengdu University of Technology, China*

⁴ *Center of Statistical Research and School of Statistics and Data Science,
Southwestern University of Finance and Economics, China*

Supplementary Material

This web supplement contains seven sections. Section S1 presents the definition of the membership matrix, the regularity conditions of theorems and their proofs. Section S2 presents the procedures on calculating standard errors and the p -value. Section S3 presents the BIC for selecting tuning parameters. Section S4 presents the additional numerical studies, including the estimation result of Simulation Study 1,2,3,4,5, the Simulation Studies of thousands of gene/biomarkers, the Simulation Studies of how a gene signature derived from gene similarities can potentially lead to misleading conclusions, and the definition of the RI. Section S5 lists the completed results of PA.3. Section S6 lists the Case studies of CO.26. Section S7 lists the baseline characteristics of PA.3 and CO.26.

S1 Theoretical Properties

S1.1 The definition of membership matrix

Denote $\tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\psi} \tilde{\boldsymbol{\alpha}}_0$, where $\boldsymbol{\psi} = \{\psi_{i,j}\}$ is a $(pn + 2n + q + p) \times (pK_1 + 2K_1 + q + K_2)$ dimensional 0/1 membership matrix, $\lfloor x \rfloor$ be the largest integer less than x , and $\text{mod}(a/b) = a - \lfloor a/b \rfloor b$. $\psi_{i,j} = 1$ for $\lfloor (i-1)/(p+2) \rfloor + 1 \in \mathcal{G}_{0,k}$, $\lfloor (j-1)/(p+2) \rfloor + 1 = k$ or $i \in (n(p+2) + 1, n(p+2) + q)$, $j = i - n(p+2) + K_1(p+2)$ or $i - n(p+2) - q \in \mathcal{D}_{0,l}$, $j - K_1(p+2) - q = l$ and 0 otherwise. That is, $\psi_{i,j} = 1$ represents the $(\lfloor (i-1)/(p+2) \rfloor + 1)$ th subject belongs to the $(\lfloor (j-1)/(p+2) \rfloor + 1)$ th prognostic group or $(i - n(p+2) - q)$ th predictor belongs to the $(j - K_1(p+2) - q)$ th predictive group.

S1.2 Conditions

The following regularity conditions are required for the proofs of the major results:

- (C1) The end time of study τ satisfies that $\int_0^\tau \lambda_0(t) dt < \infty$.
- (C2) The at-risk Process $Y_{ij}(t) = I(t \leq \tilde{T}_{ij})$ is left-continuous with right-hand limits, and $P(Y_{ij}(t) = 1, \forall t \in [0, \tau]) > 0$.
- (C3) As $N_k \rightarrow \infty$, we have $\sup_{t \in [0, \tau]} |S_{N_k}^{(k,l)}(\tilde{\boldsymbol{\alpha}}^{(k)}, \mathbf{D}, t) - s^{(k,l)}(\tilde{\boldsymbol{\alpha}}^{(k)}, t)| \rightarrow 0$, where $s^{(k,0)}(., t)$, $s^{(k,1)}(., t)$, and $s^{(k,2)}(., t)$ are continuous in $\tilde{\boldsymbol{\alpha}} \in \mathcal{B}$

(where $\mathcal{B}$ is the neighborhood of $\tilde{\alpha}^{(0)}$), uniformly in $t \in [0, \tau]$; $s^{(k,0)}(., t)$ $s^{(k,1)}(., t)$, and $s^{(k,2)}(., t)$ are bounded on $\mathcal{B} \times [0, \tau]$; $s^{(k,0)}(., t)$ is bounded from zero on $\mathcal{B} \otimes [0, \tau]$.

(C4) The covariates satisfy $\sup_{i,j} \|\tilde{\mathbf{D}}_{ij}\|_\infty \leq c_1$ for some positive constant c_1 , where $\|\mathbf{a}\|_\infty = \max_j |a_j|$.

(C5) The Fisher information $\Sigma(\tilde{\alpha}_0, \tau)$ and the matrix $C(\tilde{\alpha}_{0,1}, \tau)$ defined in equation (S1.8) are positive definite.

Remark: Condition (C1)-(C4) guarantee the asymptotic normality of the maximum partial likelihood estimates. See Andersen and Gill (1982); Fan and Li (2001). $\mathbf{D}_{2,ij} = (\mathbf{z}_{ij}^\top, \sum_{l \in \mathcal{D}_{0,1}} x_{ij,l} w_{ij}, \dots, \sum_{l \in \mathcal{D}_{0,K_2}} x_{ij,l} w_{ij})^\top$, and we have $\|\sum_{l \in \mathcal{D}_{0,k}} x_{ij,l} w_{ij}\|_\infty < \infty$, to ensure $\sup_{i,j} \|\tilde{\mathbf{D}}_{ij}\|_\infty \leq c_1$. For fixed p or $p = o(N_{\min})$, the following assumption is satisfied. In the case where $p = O(N^\alpha)$, we require that the number of nonzero elements $s = \|\mathbf{g}\|_0$ satisfied $s = o(N_{\min})$. In the ultra-high dimensional setting where $p = \exp(N^\alpha)$ for some $\alpha \in (0, 1)$, the condition becomes $s \log(p) = o(N)$. See (Fan and Lv, 2008; Bickel et al., 2009). Condition (C5) guarantee the the covariance of the oracle estimator $(\boldsymbol{\gamma}^\top, \mathbf{g}^\top)^\top$ exist.

S1.3 Proof of Proposition 1

Proof: It is equivalent to show that $M_{\text{DCAF}}(\Theta^{(m+1)}, \alpha^{(m+1)}) \leq M_{\text{DCAF}}(\Theta^{(m)}, \alpha^{(m)})$, where the penalized negative log-likelihood is defined as

$$M_{\text{DCAF}}(\Theta, \alpha) := -S_{\text{DCAF}}(\Theta, \alpha) = -L_n(\theta, g, \gamma) + p_{\lambda_1}(\theta, \alpha^{sa}) + p_{\lambda_2}(g, \alpha_4) + p_{\lambda_3}(\alpha_1),$$

with $p_{\lambda_1}(\theta, \alpha^{sa}) = p_{\lambda_1}(\theta, \alpha_1, \alpha_2, \alpha_3)$ and $p_{\lambda_3}(\alpha_1) = \sum_{k=1}^{K_1} \sum_{j=1}^p p_{\lambda_3}(\alpha_{1,kj})$.

Step 1: Surrogate functions. Because the KHM penalty is non-convex in the subgroup centres α^{sa} and α_4 , the MM (majorization-minimization) principle is used to construct convex upper bounds (surrogates) that are tight at the current iterate. Define:

$$\begin{aligned} \tilde{M}_{\text{DCAF},1}(\theta, g, \gamma, \alpha^{sa}, \alpha_4; \alpha^{sa(m)}) \\ := -L_n(\theta, g, \gamma) + \lambda_1 \sum_{i=1}^n \left\{ \text{KHM}(s; \theta_i, \alpha^{sa(m)}) + \sum_{k=1}^{K_1} \omega_{i,k}^{(m)} (\|\theta_i - \alpha_k^{sa}\|^2 - \|\theta_i - \alpha_k^{sa(m)}\|^2) \right\} \\ + p_{\lambda_2}(g, \alpha_4) + p_{\lambda_3}(\alpha_1), \end{aligned}$$

$$\begin{aligned} \tilde{M}_{\text{DCAF},2}(\theta, g, \gamma, \alpha^{sa}, \alpha_4; \alpha_4^{(m)}) \\ := -L_n(\theta, g, \gamma) + p_{\lambda_1}(\theta, \alpha^{sa}) + \lambda_2 \sum_{j=1}^p \left\{ \text{KHM}(s; g_j, \alpha_4^{(m)}) + \sum_{k=1}^{K_2} \phi_{j,k}^{(m)} ((g_j - \alpha_{4,k})^2 - (g_j - \alpha_{4,k}^{(m)})^2) \right\} \\ + p_{\lambda_3}(\alpha_1). \end{aligned}$$

By construction, each surrogate majorizes M_{DCAF} and equals it at the

expansion point:

$$M_{\text{DCAF}}(\boldsymbol{\Theta}, \boldsymbol{\alpha}) \leq \tilde{M}_{\text{DCAF},1}(\boldsymbol{\Theta}, \boldsymbol{\alpha}; \boldsymbol{\alpha}^{sa(m)}), \quad \text{with equality when } \boldsymbol{\alpha}^{sa} = \boldsymbol{\alpha}^{sa(m)}, \quad (\text{S1.1})$$

$$M_{\text{DCAF}}(\boldsymbol{\Theta}, \boldsymbol{\alpha}) \leq \tilde{M}_{\text{DCAF},2}(\boldsymbol{\Theta}, \boldsymbol{\alpha}; \boldsymbol{\alpha}_4^{(m)}), \quad \text{with equality when } \boldsymbol{\alpha}_4 = \boldsymbol{\alpha}_4^{(m)}. \quad (\text{S1.2})$$

Step 2: Algorithm updates. At iteration m , the five block-coordinate updates are:

$$\begin{aligned} \boldsymbol{\theta}^{(m+1)} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} M_{\text{DCAF}}(\boldsymbol{\theta}, \boldsymbol{g}^{(m)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{(m)}), \\ \boldsymbol{g}^{(m+1)} &= \underset{\boldsymbol{g}}{\operatorname{argmin}} M_{\text{DCAF}}(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{g}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{(m)}), \\ \boldsymbol{\gamma}^{(m+1)} &= \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} M_{\text{DCAF}}(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{g}^{(m+1)}, \boldsymbol{\gamma}, \boldsymbol{\alpha}^{(m)}), \\ \boldsymbol{\alpha}^{sa(m+1)} &= \underset{\boldsymbol{\alpha}^{sa}}{\operatorname{argmin}} \tilde{M}_{\text{DCAF},1}(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{g}^{(m+1)}, \boldsymbol{\gamma}^{(m+1)}, \boldsymbol{\alpha}^{sa}, \boldsymbol{\alpha}_4^{(m)}; \boldsymbol{\alpha}^{sa(m)}), \\ \boldsymbol{\alpha}_4^{(m+1)} &= \underset{\boldsymbol{\alpha}_4}{\operatorname{argmin}} \tilde{M}_{\text{DCAF},2}(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{g}^{(m+1)}, \boldsymbol{\gamma}^{(m+1)}, \boldsymbol{\alpha}^{sa(m+1)}, \boldsymbol{\alpha}_4; \boldsymbol{\alpha}_4^{(m)}). \end{aligned}$$

Because $\boldsymbol{\gamma}$ enters M_{DCAF} only through $-L_n$ (and not through any penalty term), the update of $\boldsymbol{\gamma}^{(m+1)}$ minimizes M_{DCAF} over $\boldsymbol{\gamma}$ regardless of the value of $\boldsymbol{\alpha}$. Similarly, $\boldsymbol{g}^{(m+1)}$ minimizes M_{DCAF} over $\boldsymbol{g}$ for any fixed value of $\boldsymbol{\alpha}^{sa}$ (since p_{λ_1} does not depend on $\boldsymbol{g}$). These observations are used in Steps (a) and (d) of the chain below.

Step 3: Monotone descent chain. Using the above updates and surrogate

properties, we establish the following chain of inequalities:

$$\begin{aligned}
& -S_{\text{DCAF}}(\boldsymbol{\Theta}^{(m+1)}, \boldsymbol{\alpha}^{(m+1)}) \\
&= M_{\text{DCAF}}(\boldsymbol{\theta}^{(m+1)}, \mathbf{g}^{(m+1)}, \boldsymbol{\gamma}^{(m+1)}, \boldsymbol{\alpha}^{(m+1)}) \\
&\stackrel{(a)}{\leq} M_{\text{DCAF}}(\boldsymbol{\theta}^{(m+1)}, \mathbf{g}^{(m+1)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{(m+1)}) \\
&\stackrel{(b)}{\leq} \tilde{M}_{\text{DCAF},2}(\boldsymbol{\theta}^{(m+1)}, \mathbf{g}^{(m+1)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{sa(m+1)}, \boldsymbol{\alpha}_4^{(m+1)}; \boldsymbol{\alpha}_4^{(m)}) \\
&\stackrel{(c)}{\leq} \tilde{M}_{\text{DCAF},2}(\boldsymbol{\theta}^{(m+1)}, \mathbf{g}^{(m+1)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{sa(m+1)}, \boldsymbol{\alpha}_4^{(m)}; \boldsymbol{\alpha}_4^{(m)}) \\
&\stackrel{(d)}{=} M_{\text{DCAF}}(\boldsymbol{\theta}^{(m+1)}, \mathbf{g}^{(m+1)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{sa(m+1)}, \boldsymbol{\alpha}_4^{(m)}) \\
&\stackrel{(e)}{\leq} M_{\text{DCAF}}(\boldsymbol{\theta}^{(m+1)}, \mathbf{g}^{(m)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{sa(m+1)}, \boldsymbol{\alpha}_4^{(m)}) \\
&\stackrel{(f)}{\leq} \tilde{M}_{\text{DCAF},1}(\boldsymbol{\theta}^{(m+1)}, \mathbf{g}^{(m)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{sa(m+1)}, \boldsymbol{\alpha}_4^{(m)}; \boldsymbol{\alpha}^{sa(m)}) \\
&\stackrel{(g)}{\leq} \tilde{M}_{\text{DCAF},1}(\boldsymbol{\theta}^{(m+1)}, \mathbf{g}^{(m)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{sa(m)}, \boldsymbol{\alpha}_4^{(m)}; \boldsymbol{\alpha}^{sa(m)}) \\
&\stackrel{(h)}{=} M_{\text{DCAF}}(\boldsymbol{\theta}^{(m+1)}, \mathbf{g}^{(m)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{(m)}) \\
&\stackrel{(i)}{\leq} M_{\text{DCAF}}(\boldsymbol{\theta}^{(m)}, \mathbf{g}^{(m)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{(m)}) \\
&= -S_{\text{DCAF}}(\boldsymbol{\Theta}^{(m)}, \boldsymbol{\alpha}^{(m)}),
\end{aligned}$$

where each step is justified as follows:

- (a) $\boldsymbol{\gamma}^{(m+1)}$ minimizes M_{DCAF} over $\boldsymbol{\gamma}$ for any $\boldsymbol{\alpha}$ (since $\boldsymbol{\gamma}$ does not appear in any penalty term), so replacing $\boldsymbol{\gamma}^{(m+1)}$ with $\boldsymbol{\gamma}^{(m)}$ cannot decrease M_{DCAF} .
- (b) The surrogate majorizes the original objective (equation (S1.2)), so

$$M_{\text{DCAF}} \leq \tilde{M}_{\text{DCAF},2} \text{ at } \boldsymbol{\alpha}_4 = \boldsymbol{\alpha}_4^{(m+1)}.$$

(c) $\boldsymbol{\alpha}_4^{(m+1)}$ is the minimizer of $\tilde{M}_{\text{DCAF},2}$ over $\boldsymbol{\alpha}_4$, so substituting back the previous iterate $\boldsymbol{\alpha}_4^{(m)}$ can only increase the surrogate.

(d) The surrogate $\tilde{M}_{\text{DCAF},2}$ equals M_{DCAF} at the expansion point $\boldsymbol{\alpha}_4 = \boldsymbol{\alpha}_4^{(m)}$ from equation (S1.2).

(e) $\boldsymbol{g}^{(m+1)}$ minimizes M_{DCAF} over $\boldsymbol{g}$ for any $\boldsymbol{\alpha}^{sa}$ (since p_{λ_1} does not depend on $\boldsymbol{g}$), so replacing $\boldsymbol{g}^{(m+1)}$ with $\boldsymbol{g}^{(m)}$ cannot decrease M_{DCAF} .

(f) The surrogate majorizes the original objective (equation (S1.1)), so

$$M_{\text{DCAF}} \leq \tilde{M}_{\text{DCAF},1} \text{ at } \boldsymbol{\alpha}^{sa} = \boldsymbol{\alpha}^{sa(m+1)}.$$

(g) $\boldsymbol{\alpha}^{sa(m+1)}$ is the minimizer of $\tilde{M}_{\text{DCAF},1}$ over $\boldsymbol{\alpha}^{sa}$, so substituting back the previous iterate $\boldsymbol{\alpha}^{sa(m)}$ can only increase the surrogate.

(h) The surrogate $\tilde{M}_{\text{DCAF},1}$ equals M_{DCAF} at the expansion point $\boldsymbol{\alpha}^{sa} = \boldsymbol{\alpha}^{sa(m)}$ from equation (S1.1).

(i) $\boldsymbol{\theta}^{(m+1)}$ is the minimizer of M_{DCAF} over $\boldsymbol{\theta}$ at the current $(\boldsymbol{g}^{(m)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\alpha}^{(m)})$, so replacing $\boldsymbol{\theta}^{(m+1)}$ with $\boldsymbol{\theta}^{(m)}$ cannot decrease M_{DCAF} .

Combining the chain yields $-S_{\text{DCAF}}(\boldsymbol{\Theta}^{(m+1)}, \boldsymbol{\alpha}^{(m+1)}) \leq -S_{\text{DCAF}}(\boldsymbol{\Theta}^{(m)}, \boldsymbol{\alpha}^{(m)})$,

which is the desired result. $\square$

Theorem 1 (Theorem 1). *Assume that $(\mathbf{x}_{ij}, w_{ij}, \mathbf{Z}_{ij})$ are independent random samples for $i = [n]$, $j = [m]$, and that T_{ij} and C_{ij} are conditionally independent given $(\mathbf{x}_{ij}, w_{ij}, \mathbf{Z}_{ij})$. Under conditions (C1)–(C5), if $N_{\min} := \min_k N_k \rightarrow \infty$ and $\sqrt{N_{\min}} \lambda_3 = O_p(1)$, then there exists a local maximizer $\hat{\boldsymbol{\alpha}}^{or}$ of the oracle objective such that*

$$\|\hat{\boldsymbol{\alpha}}^{or} - \tilde{\boldsymbol{\alpha}}_0\| = O_p\left(1/\sqrt{N_{\min}}\right).$$

S1.4 Proof of Theorem 1

Let the filtration $\mathcal{F}_t$ be the statistical information accruing during the time $[0, t]$, namely, $\mathcal{F}_t = \sigma\{N_{ij}(u), Y_{ij}(u), i = [n], j = [m], 0 \leq u \leq t\}$, where $N_{ij}(u) = I(\tilde{T}_{ij} \leq u, \delta_{ij} = 1)$ and $Y_{ij}(u) = I(\tilde{T}_{ij} \geq u)$. Then, under the independent censoring scheme,

$$M_{ij}(t) = N_{ij}(t) - \int_0^t \lambda_{ij}(u) Y_{ij}(u) du, \quad i = [n], j = [m] \quad (\text{S1.3})$$

is an martingale.

Let $\tilde{\boldsymbol{\alpha}}^{or} = (\boldsymbol{\alpha}_{1,1}^{or\top}, \alpha_{2,1}^{or}, \alpha_{3,1}^{or}, \dots, \boldsymbol{\alpha}_{1,K_1}^{or\top}, \alpha_{2,K_1}^{or}, \alpha_{3,K_1}^{or}, \boldsymbol{\gamma}^{or\top}, \alpha_{4,1}^{or}, \dots, \alpha_{4,K_2}^{or})^\top$, $\tilde{\boldsymbol{\alpha}}^{(k)or} = (\boldsymbol{\alpha}_{1,k}^{or\top}, \alpha_{2,k}^{or}, \alpha_{3,k}^{or}, \boldsymbol{\gamma}^{or\top}, \alpha_{4,1}^{or}, \dots, \alpha_{4,K_2}^{or})^\top$, $k = [K_1]$. Denote

$$\begin{aligned} L_N(\tilde{\boldsymbol{\alpha}}^{or}, \tau) &= \sum_{k=1}^{K_1} \sum_{i \in \mathcal{G}_{0,k}} \sum_{j=1}^m \int_0^\tau \tilde{\boldsymbol{D}}_{ij}^\top \tilde{\boldsymbol{\alpha}}^{(k)or} dN_{ij}(t) \\ &\quad - \int_0^\tau \log \left\{ \sum_{k=1}^{K_1} \sum_{i \in \mathcal{G}_{0,k}} \sum_{j=1}^m Y_{ij}(t) \exp(\tilde{\boldsymbol{D}}_{ij}^\top \tilde{\boldsymbol{\alpha}}^{(k)or}) \right\} d\bar{N}(t) \end{aligned}$$

where $\bar{N}(t) = \sum_{k=1}^{K_1} \sum_{i \in \mathcal{G}_{0,k}} \sum_{j=1}^m N_{ij}(t)$.

For each $\tilde{\alpha}^{or}$ in a neighbourhood of $\tilde{\alpha}_0$,

$$\begin{aligned} 1/N \{L_N(\tilde{\alpha}^{or}, \tau) - L_N(\tilde{\alpha}_0, \tau)\} &= (\tilde{\alpha}^{or} - \tilde{\alpha}_0)^T A_N(\tilde{\alpha}_0, \tau) \\ &\quad - \frac{1}{2} (\tilde{\alpha}^{or} - \tilde{\alpha}_0)^T B_N(\tilde{\alpha}_0, \tau) (\tilde{\alpha}^{or} - \tilde{\alpha}_0) (1 + o_p(1)), \end{aligned}$$

where

$$A_N(\tilde{\alpha}_0, \tau) = 1/N \sum_{i=1}^n \sum_{j=1}^m \int_0^\tau \left\{ \mathbf{D}_{ij} - \frac{\sum_{i=1}^n \sum_{j=1}^m Y_{ij}(t) \mathbf{D}_{ij} \exp(\mathbf{D}_{ij}^\top \tilde{\alpha}_0)}{\sum_{i=1}^n \sum_{j=1}^m Y_{ij}(t) \exp(\mathbf{D}_{ij}^\top \tilde{\alpha}_0)} \right\} dN_{ij}(t),$$

and

$$\begin{aligned} B_N(\tilde{\alpha}_0, \tau) &= 1/N \int_0^\tau \left\{ \frac{\sum_{i=1}^n \sum_{j=1}^m Y_{ij}(t) \mathbf{D}_{ij}^{\otimes 2} \exp(\mathbf{D}_{ij}^\top \tilde{\alpha}_0)}{\sum_{i=1}^n \sum_{j=1}^m Y_{ij}(t) \exp(\mathbf{D}_{ij}^\top \tilde{\alpha}_0)} \right. \\ &\quad \left. - \left[\frac{\sum_{i=1}^n \sum_{j=1}^m Y_{ij}(t) \mathbf{D}_{ij} \exp(\mathbf{D}_{ij}^\top \tilde{\alpha}_0)}{\sum_{i=1}^n \sum_{j=1}^m Y_{ij}(t) \exp(\mathbf{D}_{ij}^\top \tilde{\alpha}_0)} \right]^{\otimes 2} \right\} d\bar{N}(t) \end{aligned}$$

where $\mathbf{D}_{ij} := (I(i \in \mathcal{G}_{0,1}) \mathbf{D}_{1,ij}^\top, \dots, I(i \in \mathcal{G}_{0,K_1}) \mathbf{D}_{1,ij}^\top, \mathbf{D}_{2,ij}^\top)^\top$. For $l =$

$0, 1, 2$, we denote $S_{N_k}^{(k,l)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t) = 1/N_k \sum_{i \in \mathcal{G}_{0,k}} \sum_{j=1}^{m_i} Y_{ij}(t) \tilde{\mathbf{D}}_{ij}^{\otimes l} \exp(\tilde{\mathbf{D}}_{ij}^\top \tilde{\alpha}_0^{(k)})$,

and $S_N^{(l)}(\tilde{\alpha}_0, \mathbf{D}, t) = 1/N \sum_{k=1}^{K_1} N_k S_{N_k}^{(k,l)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t)$. Denote the expectation

of $S_{N_k}^{(k,l)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t)$ and $S_N^{(l)}(\tilde{\alpha}_0, \mathbf{D}, t)$ by $s^{(k,l)}(\tilde{\alpha}_0^{(k)}, t)$ and $s^{(l)}(\tilde{\alpha}_0, t)$, respec-

tively. Under the assumption $N_k \rightarrow \infty$, we have $\sup_{t \in [0, \tau]} |S_{N_k}^{(k,l)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t) - s^{(k,l)}(\tilde{\alpha}_0^{(k)}, t)| \rightarrow 0$.

Redefine $S_{N_{k,1}}^{(k,1)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t) = 1/N_k \sum_{i \in \mathcal{G}_{0,k}} \sum_{j=1}^m Y_{ij}(t) \mathbf{D}_{1,ij} \exp(\mathbf{D}_{1,ij}^\top \tilde{\alpha}_0^{(k)})$,

and $S_{N_{k,2}}^{(k,1)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t) = 1/N_k \sum_{i \in \mathcal{G}_{0,k}} \sum_{j=1}^m Y_{ij}(t) \mathbf{D}_{2,ij} \exp(\mathbf{D}_{2,ij}^\top \tilde{\alpha}_0^{(k)})$, is the sub-

vector of $S_{N_k}^{(k,1)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t)$ correspond to the $\mathbf{D}_{1,ij}$ and $\mathbf{D}_{2,ij}$, respectively.

Similarity, Denote $S_{N_k, (1,1)}^{(k,2)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t)$, $S_{N_k, (2,2)}^{(k,2)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t)$ is the sub-matrix of $S_{N_k}^{(k,2)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t)$ correspond to $\mathbf{D}_{1,ij} \mathbf{D}_{1,ij}^\top$ and $\mathbf{D}_{2,ij} \mathbf{D}_{2,ij}^\top$, respectively. $S_{N_k, (1,2)}^{(k,2)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t)$ is the sub-matrix of $S_{N_k}^{(k,2)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t)$ correspond to $\mathbf{D}_{1,ij} \mathbf{D}_{2,ij}^\top$. $s^{(k,0)}(\tilde{\alpha}_0^{(k)}, t)$, $s_j^{(k,l)}(\tilde{\alpha}_0^{(k)}, t)$ and $s_{(j,m)}^{(k,l)}(\tilde{\alpha}_0^{(k)}, t)$ has the same definition, where $j = 1, 2$, $l, m = 1, 2$, $k = [K_1]$.

A substitution of (S1.3) into $A_N(\tilde{\alpha}_0, \tau)$ and $B_N(\tilde{\alpha}_0, \tau)$ gives

$$\begin{aligned} A_N(\tilde{\alpha}_0, \tau) &= 1/N \sum_{i=1}^n \sum_{j=1}^m \int_0^\tau \left\{ \mathbf{D}_{ij} - \frac{S_N^{(1)}(\tilde{\alpha}_0, \mathbf{D}, t)}{S_N^{(0)}(\tilde{\alpha}_0, \mathbf{D}, t)} \right\} dM_{ij}(t), \\ B_N(\tilde{\alpha}_0, \tau) &= 1/N \int_0^\tau \left\{ \frac{S_N^{(2)}(\tilde{\alpha}_0, \mathbf{D}, t)}{S_N^{(0)}(\tilde{\alpha}_0, \mathbf{D}, t)} - \left[\frac{S_N^{(1)}(\tilde{\alpha}_0, \mathbf{D}, t)}{S_N^{(0)}(\tilde{\alpha}_0, \mathbf{D}, t)} \right]^{\otimes 2} \right\} d\bar{M}(t) \\ &\quad + \int_0^\tau \left\{ S_N^{(2)}(\tilde{\alpha}_0, \mathbf{D}, t) - \frac{S_N^{(1)}(\tilde{\alpha}_0, \mathbf{D}, t)^{\otimes 2}}{S_N^{(0)}(\tilde{\alpha}_0, \mathbf{D}, t)} \right\} \lambda_0(t) dt \\ &= B_{N,1}(\tilde{\alpha}_0, \tau) + B_{N,2}(\tilde{\alpha}_0, \tau). \end{aligned}$$

We now derive the asymptotic normality of the term $A_N(\tilde{\alpha}_0, t)$,

$$\langle A_n, A_n \rangle(\tau) = 1/N \int_0^\tau \left\{ S_N^{(2)}(\tilde{\alpha}_0, \mathbf{D}, t) - \frac{S_N^{(1)}(\tilde{\alpha}_0, \mathbf{D}, t)^{\otimes 2}}{S_N^{(0)}(\tilde{\alpha}_0, \mathbf{D}, t)} \right\} \lambda_0(t) dt,$$

where

$$S_N^{(1)}(\tilde{\alpha}_0, \mathbf{D}, t) = \begin{pmatrix} N_1/N S_{N_1,1}^{(1,1)\top}(\tilde{\alpha}_0^{(1)}, \mathbf{D}, t) \\ \vdots \\ N_{K_1}/N S_{N_{K_1},1}^{(K_1,1)\top}(\tilde{\alpha}_0^{(K_1)}, \mathbf{D}, t) \\ \sum_{k=1}^{K_1} N_k/N S_{N_k,2}^{(k,1)}(\tilde{\alpha}_0^{(k)}, \mathbf{D}, t) \end{pmatrix},$$

and

$$S_N^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) = \begin{pmatrix} S_{N,(1,1)}^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) & S_{N,(1,2)}^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) S_{N,(1,2)}^{(2)\top}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) & S_{N,(2,2)}^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) \end{pmatrix},$$

where $S_{N,(1,1)}^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) = \text{diag}(N_1/N S_{N_1,(1,1)}^{(1,2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t), \dots, N_{K_1}/N S_{N_{K_1},(1,1)}^{(K_1,2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t))$,
 $S_{N,(1,2)}^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) = (S_{N_1,(1,2)}^{(1,2)\top}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t), \dots, S_{N_{K_1},(1,2)}^{(K_1,2)\top}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t))^\top$, and $S_{N,(2,2)}^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t)$
 $= \sum_{k=1}^{K_1} N_k/N S_{N_k,(2,2)}^{(k,2)}(\tilde{\boldsymbol{\alpha}}_0^{(k)}, \mathbf{D}, t)$.

By Large Number Theorem, and we can show that for any $0 \leq t \leq \tau$,

$$\begin{aligned} S_N^{(0)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) &= \sum_{k=1}^{K_1} \omega_k s^{(k,0)}(\tilde{\boldsymbol{\alpha}}_0^{(k)}, t) + O_p(1/\sqrt{N}) \\ &= s^{(0)}(\tilde{\boldsymbol{\alpha}}_0, t) + O_p(1/\sqrt{N}), \end{aligned}$$

$$\begin{aligned} S_N^{(1)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) &= \begin{pmatrix} \omega_1 s_1^{(1,1)}(\tilde{\boldsymbol{\alpha}}_0^{(1)}, t) \\ \vdots \\ \omega_{K_1} s_1^{(K_1,1)}(\tilde{\boldsymbol{\alpha}}_0^{(K_1)}, t) \\ \sum_{k=1}^{K_1} \omega_k s_2^{(k,1)}(\tilde{\boldsymbol{\alpha}}_0^{(k)}, t) \end{pmatrix} + \begin{pmatrix} O_p(\sqrt{N_1}/N) \\ \vdots \\ O_p(\sqrt{N_{K_1}}/N) \\ O_p(1/\sqrt{N}) \end{pmatrix} \\ &= s^{(1)}(\tilde{\boldsymbol{\alpha}}_0, t) + O_p(\sqrt{\max N_k}/N), \end{aligned}$$

and

$$S_{N,(1,1)}^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) = \begin{pmatrix} \omega_1 s_{(1,1)}^{(1,2)}(\tilde{\boldsymbol{\alpha}}_0^{(1)}, t) & 0 & \cdots & 0 \\ 0 & \omega_2 s_{(1,1)}^{(2,2)}(\tilde{\boldsymbol{\alpha}}_0^{(2)}, t) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_{K_1} s_{(1,1)}^{(K_1,2)}(\tilde{\boldsymbol{\alpha}}_0^{(K_1)}, t) \end{pmatrix} + \begin{pmatrix} O_p(\sqrt{N_1}/N) & 0 & \cdots & 0 \\ 0 & O_p(\sqrt{N_2}/N) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & O_p(\sqrt{N_{K_1}}/N) \end{pmatrix},$$

$$S_{N,(1,2)}^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) = \begin{pmatrix} \omega_1 s_{(1,2)}^{(1,2)}(\tilde{\boldsymbol{\alpha}}_0, t) \\ \vdots \\ \omega_{K_1} s_{(1,2)}^{(K_1,2)}(\tilde{\boldsymbol{\alpha}}_0, t) \end{pmatrix} + \begin{pmatrix} O_p(\sqrt{N_1}/N) \\ \vdots \\ O_p(\sqrt{N_{K_1}}/N) \end{pmatrix},$$

$$S_{N,(2,2)}^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) = \sum_{k=1}^{K_1} \omega_k s_{(2,2)}^{(k,2)}(\tilde{\boldsymbol{\alpha}}_0, t) + O_p(1/\sqrt{N}).$$

$$\begin{aligned}
 & S_N^{(2)}(\tilde{\boldsymbol{\alpha}}_0, \mathbf{D}, t) \\
 = & \begin{pmatrix} \omega_1 s_{(1,1)}^{(1,2)}(\tilde{\boldsymbol{\alpha}}_0^{(1)}, t) & 0 & \cdots & 0 & \omega_1 s_{(1,2)}^{(1,2)}(\tilde{\boldsymbol{\alpha}}_0, t) \\ 0 & \omega_2 s_{(1,1)}^{(2,2)}(\tilde{\boldsymbol{\alpha}}_0^{(2)}, t) & \cdots & 0 & \omega_2 s_{(1,2)}^{(2,2)}(\tilde{\boldsymbol{\alpha}}_0, t) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \omega_{K_1} s_{(1,1)}^{(K_1,2)}(\tilde{\boldsymbol{\alpha}}_0^{(K_1)}, t) & \omega_{K_1} s_{(1,2)}^{\top(K_1,2)}(\tilde{\boldsymbol{\alpha}}_0, t) \\ \omega_1 s_{(1,2)}^{(1,2)\top}(\tilde{\boldsymbol{\alpha}}_0, t) & \omega_2 s_{(1,2)}^{(2,2)\top}(\tilde{\boldsymbol{\alpha}}_0, t) & \cdots & \omega_{K_1} s_{(1,2)}^{(K_1,2)\top}(\tilde{\boldsymbol{\alpha}}_0, t) & \sum_{k=1}^{K_1} \omega_k s_{(2,2)}^{(k,2)}(\tilde{\boldsymbol{\alpha}}_0, t) \end{pmatrix} \\
 + & \begin{pmatrix} \sqrt{N_1}/N & 0 & \cdots & 0 & \sqrt{N_1}/N \\ 0 & \sqrt{N_2}/N & \cdots & 0 & \sqrt{N_2}/N \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sqrt{N_{K_1}}/N & \sqrt{N_{K_2}}/N \\ \sqrt{N_1}/N & \sqrt{N_1}/N & \cdots & \sqrt{N_{K_1}}/N & 1/\sqrt{N} \end{pmatrix} \\
 = & s^{(2)}(\tilde{\boldsymbol{\alpha}}_0, t) + O_p(\sqrt{\max N_k}/N).
 \end{aligned}$$

So we can get $E(A_N(\tilde{\boldsymbol{\alpha}}_0, t)^{\otimes 2}) = \Sigma(\tilde{\boldsymbol{\alpha}}_0, t)/N + O_p(\sqrt{\max N_k}/N^2)$, where

$$\Sigma(\tilde{\boldsymbol{\alpha}}_0, \tau) = \int_0^\tau \left\{ \frac{s^{(2)}(\tilde{\boldsymbol{\alpha}}_0, t)}{s^{(0)}(\tilde{\boldsymbol{\alpha}}_0, t)} - \left(\frac{s^{(1)}(\tilde{\boldsymbol{\alpha}}_0, t)}{s^{(0)}(\tilde{\boldsymbol{\alpha}}_0, t)} \right)^{\otimes 2} \right\} s^{(0)}(\tilde{\boldsymbol{\alpha}}_0, t) \lambda_0(t) dt.$$

Using Theorem 3.2 of Andersen and Gill (1982), it can be proved that

$$\sqrt{N} A_N(\tilde{\boldsymbol{\alpha}}_0, \tau) \rightarrow N(0, \Sigma(\tilde{\boldsymbol{\alpha}}_0, \tau)), \quad (\text{S1.4})$$

Next, we consider $B_N(\tilde{\boldsymbol{\alpha}}_0, \tau)$. It is obvious that $B_{N,1}(\tilde{\boldsymbol{\alpha}}_0, \tau) = o_p(1)$

since it is a stochastic integral against a martingale measure, and conse-

quently $B_N(\tilde{\alpha}_0, \tau) = \Sigma(\tilde{\alpha}_0, \tau) + o_p(1)$, since $B_{N,2}(\tilde{\alpha}_0, \tau)$ converges to the Fisher information $\Sigma(\tilde{\alpha}_0, \tau) > 0$.

To handle the case of potentially diverging K_1 , we employ an anisotropic scaling that reflects the different amounts of information available for group-specific parameters versus shared parameters. Partition $\mathbf{u}$ as $(\mathbf{u}_1^\top, \dots, \mathbf{u}_{K_1}^\top, \vartheta^\top)^\top$, where $\mathbf{u}_k \in \mathbb{R}^{p+2}$ corresponds to the group- k specific parameters $(\alpha_{1,k}, \alpha_{2,k}, \alpha_{3,k})$ and ϑ corresponds to the shared parameters (γ, α_4) . Define the anisotropic displacement

$$\delta(\mathbf{u}, \vartheta) = (\mathbf{u}_1^\top / \sqrt{N_1}, \dots, \mathbf{u}_{K_1}^\top / \sqrt{N_{K_1}}, \vartheta^\top / N)^\top.$$

We want to show that for any given $\varepsilon > 0$, there exists a large constant C_0 such that

$$P \left\{ \sup_{\sum_{k=1}^{K_1} \|\mathbf{u}_k\|^2 + \|\vartheta\|^2 = C_0^2 K_1} S_{DCAF}(\tilde{\alpha}_0 + \delta(\mathbf{u}, \vartheta)) < S_{DCAF}(\tilde{\alpha}_0) \right\} \geq 1 - \varepsilon. \quad (\text{S1.5})$$

By the compactness of the closed anisotropic ball $\mathcal{B}_n = \{\tilde{\alpha}_0 + \delta(\mathbf{u}, \vartheta) : \sum_{k=1}^{K_1} \|\mathbf{u}_k\|^2 + \|\vartheta\|^2 \leq C_0^2 K_1\}$ and the continuity of S_{DCAF} , (S1.5) implies that the global maximum of S_{DCAF} over $\mathcal{B}_n$ is attained in the interior. Hence there exists a local maximizer $\hat{\alpha}^{or}$ satisfying

$$\|\hat{\alpha}_{1,k}^{or} - \alpha_{0,1,k}\| = O_p(N_k^{-1/2}) = O_p(N_{\min}^{-1/2}), \quad \|(\hat{\gamma}^{or}, \hat{\alpha}_4^{or}) - (\gamma_0, \alpha_{4,0})\| = O_p(N^{-1/2}).$$

Consequently $\|\hat{\alpha}^{or} - \tilde{\alpha}_0\| = O_p(N_{\min}^{-1/2})$.

To establish (S1.5), define

$$\begin{aligned}
 D_N(\mathbf{u}, \vartheta) &\equiv 1/N[S_{DCAF}(\tilde{\boldsymbol{\alpha}}_0 + \delta(\mathbf{u}, \vartheta)) - S_{DCAF}(\tilde{\boldsymbol{\alpha}}_0)] \\
 &= 1/N\{L_N(\tilde{\boldsymbol{\alpha}}_0 + \delta(\mathbf{u}, \vartheta)) - L_N(\tilde{\boldsymbol{\alpha}}_0)\} \\
 &\quad - \lambda_3/N \sum_{k=1}^{K_1} \sum_{j=1}^p \left(\frac{|\alpha_{0,1,kj} + u_{1,kj}/\sqrt{N_k}|}{|\hat{\alpha}_{1,kj}^{\text{init}}|} - \frac{|\alpha_{0,1,kj}|}{|\hat{\alpha}_{1,kj}^{\text{init}}|} \right) \quad (\text{S1.6})
 \end{aligned}$$

where $\hat{\alpha}_{1,kj}^{\text{init}}$ denotes the ridge initial estimator used as the adaptive LASSO weight.

Step 1 (Quadratic lower bound). By the second-order Taylor expansion of $1/N L_N$ at $\tilde{\boldsymbol{\alpha}}_0$ and $B_N(\tilde{\boldsymbol{\alpha}}_0, \tau) = \Sigma(\tilde{\boldsymbol{\alpha}}_0, \tau) + o_p(1)$,

$$1/N\{L_n(\tilde{\boldsymbol{\alpha}}_0 + \delta) - L_n(\tilde{\boldsymbol{\alpha}}_0)\} = \delta^\top A_N(\tilde{\boldsymbol{\alpha}}_0, \tau) - \frac{1}{2} \delta^\top B_N(\tilde{\boldsymbol{\alpha}}_0, \tau) \delta (1 + o_p(1)).$$

Under the anisotropic scaling, the group- k diagonal block of $B_N = B_N(\tilde{\boldsymbol{\alpha}}_0, \tau)$ satisfies $[B_N]_{kk} = (N_k/N)\Sigma_{kk} + o_p(1)$, so

$$\frac{\mathbf{u}_k^\top [B_N]_{kk} \mathbf{u}_k}{N_k} = \frac{\mathbf{u}_k^\top \Sigma_{kk} \mathbf{u}_k}{N} + o_p\left(\frac{\|\mathbf{u}_k\|^2}{N_k}\right).$$

The cross block satisfies $\|[B_N]_{ks}\|_2 = O(N_k/N)$, hence by Young's inequality,

$$\left| \frac{2\mathbf{u}_k^\top [B_N]_{ks} \vartheta}{\sqrt{N_k N}} \right| \leq \frac{2c(N_k/N)\|\mathbf{u}_k\|\|\vartheta\|}{\sqrt{N_k N}} = \frac{2c\sqrt{N_k}\|\mathbf{u}_k\|\|\vartheta\|}{N^{3/2}} \leq \frac{c(\|\mathbf{u}_k\|^2 + \|\vartheta\|^2)}{N}.$$

Since $\Sigma(\tilde{\boldsymbol{\alpha}}_0, \tau)$ is positive definite by Condition (C5), combining diagonal and cross-block contributions yields, for $C^2 = \sum_{k=1}^{K_1} \|\mathbf{u}_k\|^2 + \|\vartheta\|^2$,

$$\delta^\top B_N \delta \geq \frac{\lambda_{\min}(\Sigma)}{N} C^2 + o_p(1),$$

where $\lambda_{\min}(\Sigma) > 0$ is independent of K_1 and $\omega_k = N_k/N$.

Step 2 (First-order term). The group- k component of A_N is a sum of N_k martingale increments each scaled by $1/N$, so $A_N^{(k,1)} = O_p(\sqrt{N_k}/N)$.

Under the anisotropic displacement,

$$\frac{\mathbf{u}_k^\top}{\sqrt{N_k}} A_N^{(k,1)} = O_p\left(\frac{\|\mathbf{u}_k\|}{\sqrt{N_k}} \cdot \frac{\sqrt{N_k}}{N}\right) = O_p\left(\frac{\|\mathbf{u}_k\|}{N}\right).$$

Summing over all groups and shared parameters and applying Cauchy-Schwarz,

$$|\delta^\top A_N(\tilde{\alpha}_0, \tau)| = O_p\left(\frac{\sum_k \|\mathbf{u}_k\| + \|\vartheta\|}{N}\right) = O_p\left(\frac{C\sqrt{K_1+1}}{N}\right).$$

Step 3 (Choice of C_0). Set $C^2 = C_0^2 K_1$, i.e. $C = C_0\sqrt{K_1}$. Combining Steps 1 and 2,

$$\begin{aligned} 1/N\{L_n(\tilde{\alpha}_0 + \delta) - L_n(\tilde{\alpha}_0)\} &\leq -\frac{\lambda_{\min}(\Sigma)C_0^2 K_1}{2N}(1 + o_p(1)) + O_p\left(\frac{C_0 K_1}{N}\right) \\ &= \frac{C_0 K_1}{N} \left[-\frac{\lambda_{\min}(\Sigma)C_0}{2}(1 + o_p(1)) + O_p(1) \right]. \end{aligned}$$

For $C_0 > 4/\lambda_{\min}(\Sigma)$, the bracket is negative with probability at least $1 - \varepsilon/2$ for all large n .

Step 4 (Penalty term). Since the initial ridge estimator satisfies $\hat{\alpha}_{1,kj}^{\text{init}} - \alpha_{0,1,kj} = O_p(N_{\min}^{-1/2})$, a Taylor expansion gives, for each active coordinate $j \leq p_{s,k}$,

$$\frac{1}{|\hat{\alpha}_{1,kj}^{\text{init}}|} = \frac{1}{|\alpha_{0,1,kj}|} + \frac{O_p(1)}{\sqrt{N_{\min}}}.$$

Applying the triangle inequality to the penalty difference and summing over all active coordinates,

$$\begin{aligned} \frac{\lambda_3}{N} \sum_{k=1}^{K_1} \sum_{j=1}^{p_{s,k}} \frac{|u_{1,kj}|/\sqrt{N_k}}{|\hat{\alpha}_{1,kj}^{\text{init}}|} &\leq \frac{\lambda_3}{N\sqrt{N_{\min}}} \sum_{k=1}^{K_1} \sum_{j=1}^{p_{s,k}} \frac{|u_{1,kj}|}{|\alpha_{0,1,kj}|} \left(1 + \frac{O_p(1)}{\sqrt{N_{\min}}}\right) \\ &\leq \frac{\lambda_3 \sqrt{K_1} p_s C}{N\sqrt{N_{\min}} |\alpha_{0,\min}|} (1 + o_p(1)) = \frac{\lambda_3 C_0 K_1 \sqrt{p_s}}{N\sqrt{N_{\min}} |\alpha_{0,\min}|} (1 + o_p(1)), \end{aligned}$$

where $p_s = \max_k p_{s,k}$ and $|\alpha_{0,\min}| = \min_{k,j \leq p_{s,k}} |\alpha_{0,1,kj}| > 0$. Since $\sqrt{N_{\min}} \lambda_3 = O_p(1)$, this penalty contribution is $O_p(C_0 K_1 / (N N_{\min}))$, which is negligible relative to the quadratic $O(C_0^2 K_1 / N)$ since $N_{\min} \rightarrow \infty$.

Combining Steps 3 and 4, for $C_0 > 4/\lambda_{\min}(\Sigma)$ and all sufficiently large n , $D_N(\mathbf{u}, \vartheta) < 0$ holds on the anisotropic sphere $\sum_{k=1}^{K_1} \|\mathbf{u}_k\|^2 + \|\vartheta\|^2 = C_0^2 K_1$ with probability at least $1 - \varepsilon$. Therefore (S1.5) holds and it completes the proof.

S1.5 Proof of Theorem 2.

(i) Here we show that $\hat{\tilde{\alpha}}_{1,k,2}^{or} = 0$ for $k = [K_1]$. Following similar proofs as those in (Zhang and Lu, 2007), we want to show

$$S_{DCAF}(\tilde{\alpha}_1^{or}, \mathbf{0}) = \max_{\|\alpha_{1,k,2}^{or}\| \leq CN_{\min}^{-1/2}, k=[K_1]} S_{DCAF}(\tilde{\alpha}_1^{or}, \tilde{\alpha}_2^{or})$$

In other words, we will show that $\frac{\partial S_{DCAF}(\tilde{\alpha}^{or})}{\partial \alpha_{1,k,2,j}^{or}}$ and $\alpha_{1,k,2,j}^{or}$ have different signs for $\alpha_{1,k,2,j}^{or} \in (-CN_{\min}^{-1/2}, CN_{\min}^{-1/2})$, and $\alpha_{1,k,2,j}^{or}$ is the j -th element of $\alpha_{1,k}^{or}$ for

$j = p_{s,k} + 1, \dots, p$. For $j = p_{s,k} + 1, \dots, p$, we have

$$\frac{\partial S_{DCAF}(\tilde{\alpha}^{or})}{\partial \alpha_{1,k,2,j}} = \frac{\partial L_N(\tilde{\alpha}^{or})}{\partial \alpha_{1,k,2,j}^{or}} - N\lambda_3 \frac{\text{sign}(\alpha_{1,k,2,j}^{or})}{|\alpha_{1,k,2,j}^{or}|} = O_p(N^{1/2}) - (N\lambda_3)N_{\min}^{1/2} \frac{\text{sign}(\alpha_{1,k,2,j}^{or})}{|N_{\min}^{1/2} \alpha_{1,k,2,j}^{or}|}$$

Note that $N_{\min}^{1/2}(\alpha_{1,k,2,j}^{or} - 0) = O_p(1)$, so that we have

$$\frac{\partial S_{DCAF}(\tilde{\alpha}^{or})}{\partial \alpha_{1,k,2,j}^{or}} = N^{1/2} \left\{ O_p(1) - \sqrt{NN_{\min}} \lambda_3 \frac{\text{sign}(\alpha_{1,k,2,j}^{or})}{O_p(1)} \right\} \quad (\text{S1.7})$$

Since $N_{\min} \lambda_3 \rightarrow \infty$, the sign of $\frac{\partial S_{DCAF}(\tilde{\alpha}^{or})}{\partial \alpha_{1,k,2,j}^{or}}$ in (S1.7) is completely determined by the sign of $\alpha_{1,k,2,j}^{or}$, when $N_{\min}$ is large, and they always have different signs.

(ii) Here we have show the asymptotic normality of $\hat{\tilde{\alpha}}_1^{or}$. Denote $\check{\alpha}_k = (\alpha_{1,k,1}^{or\top}, \alpha_{2,k}^{or}, \alpha_{3,k}^{or})^\top$ for $k = 1, \dots, K_1$, $\check{\theta} = (\gamma^{or\top}, \alpha_4^{or\top})^\top$. $\check{\alpha}_{k,0} = (\alpha_{0,1,k,1}^\top, \alpha_{0,2,k}, \alpha_{0,3,k})^\top$ and $\check{\theta}_0 = (\gamma_0^\top, \alpha_{0,4}^\top)$ is the true value of $\check{\alpha}_k$ and $\check{\theta}$.

$$\begin{aligned} \mathbf{0} &= 1/N \frac{\partial S_{DCAF}(\tilde{\alpha}^{or})}{\partial \tilde{\alpha}^{or}} \Big|_{\{\tilde{\alpha}_1^{or}, \tilde{\alpha}_2^{or}\} = \{\hat{\alpha}_1, \mathbf{0}\}} \\ &= 1/N \frac{\partial L_N(\tilde{\alpha})}{\partial \tilde{\alpha}} \Big|_{\{\tilde{\alpha}_1^{or}, \tilde{\alpha}_2^{or}\} = \{\hat{\alpha}_1, \mathbf{0}\}} - \lambda_3 \left(\frac{\text{sign}(\hat{\alpha}_{1,1,1}^\top)}{\tilde{\alpha}_{1,1,1}^{0\top}}, 0, 0, \dots, \frac{\text{sign}(\hat{\alpha}_{1,K_1,1}^\top)}{\tilde{\alpha}_{1,K_1,1}^{0\top}}, 0, 0, \mathbf{0}_{q+K_2} \right) \\ &= \tilde{A}_n(\tilde{\alpha}_{0,1}, \tau) - \tilde{B}_n(\tilde{\alpha}_{0,1}, \tau)(\hat{\alpha}_1^{or} - \tilde{\alpha}_{0,1}) \\ &\quad + \Delta - \lambda_3 \left(\frac{\text{sign}(\alpha_{0,1,1,1}^\top)}{\tilde{\alpha}_{1,1,1}^{0\top}}, 0, 0, \dots, \frac{\text{sign}(\alpha_{0,1,K_1,1}^\top)}{\tilde{\alpha}_{1,K_1,1}^{0\top}}, 0, 0, \mathbf{0}_{q+K_2} \right), \end{aligned}$$

where $\tilde{A}_n(\tilde{\alpha}_{0,1}, \tau)$ be the first $\sum_{k=1}^{K_1} (p_{k,s} + 2) + K_2 + q$ elements of $A_n(\tilde{\alpha}_0, \tau)$,

and $\tilde{B}_n(\tilde{\alpha}_{0,1}, \tau)$ is the first $\sum_{k=1}^{K_1} (p_{k,s} + 2) + K_2 + q \times \sum_{k=1}^{K_1} (p_{k,s} + 2) + K_2 + q$

$$\text{submatrix of } B_n(\tilde{\boldsymbol{\alpha}}_0, \tau), \Delta = O_p \begin{pmatrix} \|\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_0\|_2^2 + \omega_1 \|\hat{\boldsymbol{\alpha}}_1^{or} - \check{\boldsymbol{\alpha}}_{1,0}\|_2^2 \\ \|\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_0\|_2^2 + \omega_2 \|\hat{\boldsymbol{\alpha}}_2^{or} - \check{\boldsymbol{\alpha}}_{2,0}\|_2^2 \\ \vdots \\ \|\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_0\|_2^2 + \omega_{K_1} \|\hat{\boldsymbol{\alpha}}_{K_1}^{or} - \check{\boldsymbol{\alpha}}_{K_1,0}\|_2^2 \\ \|\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_0\|_2^2 + \sum_{k=1}^{K_1} \omega_k \|\hat{\boldsymbol{\alpha}}_k^{or} - \check{\boldsymbol{\alpha}}_{k,0}\|_2^2 \end{pmatrix}.$$

The last equation is implied by $\text{sign}(\hat{\boldsymbol{\alpha}}_{1,k,1}) = \text{sign}(\boldsymbol{\alpha}_{0,1,k,1})$ for $k = [K_1]$

when $N_{\min}$ is large, since $\hat{\boldsymbol{\alpha}}_{1,k,1}^\top$ is a root- $N_{\min}$ consistent estimator of $\boldsymbol{\alpha}_{0,1,k,1}$, for $k = [K_1]$.

According to Theorem 3.2 of Andersen and Gill (1982), we have

$$\sqrt{N} \tilde{A}_n(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau) \rightarrow N(\mathbf{0}, \tilde{\Sigma}(\tilde{\boldsymbol{\alpha}}_{0,1}, t)), \text{ and } \tilde{B}_n(\tilde{\boldsymbol{\alpha}}_{10}, \tau) \rightarrow \tilde{\Sigma}(\tilde{\boldsymbol{\alpha}}_{0,1}, t) + o_p(1),$$

where $\tilde{\Sigma}(\tilde{\boldsymbol{\alpha}}_{0,1}, t)$ is first $\sum_{k=1}^{K_1} (p_{k,s} + 2) + K_2 + q \times \sum_{k=1}^{K_1} (p_{k,s} + 2) + K_2 + q$ sub-

matrix of $\Sigma(\boldsymbol{\alpha}_0, t)$. Denote $\mathbf{b} = \left(\frac{\text{sign}(\boldsymbol{\alpha}_{0,1,1,1}^\top)}{\hat{\boldsymbol{\alpha}}_{1,1,1}^{0^\top}}, 0, 0, \dots, \frac{\text{sign}(\boldsymbol{\alpha}_{0,1,K_1,1}^\top)}{\hat{\boldsymbol{\alpha}}_{1,K_1,1}^{0^\top}}, 0, 0, \mathbf{0}_{q+K_2} \right)$,

then

$$(\hat{\boldsymbol{\alpha}}_1^{or} - \tilde{\boldsymbol{\alpha}}_{0,1}) = \tilde{\Sigma}^{-1}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau) \left\{ \tilde{A}_n(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau) + \Delta - \lambda_3 \mathbf{b} \right\}.$$

Denote

$$\begin{aligned}
 \tilde{\Sigma}(\tilde{\alpha}_{0,1}, \tau) &= \begin{pmatrix} h(\tilde{\alpha}_{0,1}, \tau) & h_{12}(\tilde{\alpha}_{0,1}, \tau) \\ h_{12}(\tilde{\alpha}_{0,1}, \tau)^\top & h_{22}(\tilde{\alpha}_{0,1}, \tau) \end{pmatrix}, \quad h(\tilde{\alpha}_{0,1}, \tau) := \{h_{ij}(\tilde{\alpha}_{0,1}, \tau)\}_{1 \leq i, j \leq K_1}, \\
 h_{22}(\tilde{\alpha}_{0,1}, \tau) &= \int_0^\tau \left[\sum_{k=1}^{K_1} \omega_k s_{(2,2)}^{(k,2)}(\tilde{\alpha}_{0,1}, t) - \frac{\left\{ \sum_{k=1}^{K_1} \omega_k s_1^{(k,1)}(\tilde{\alpha}_{0,1}, t) \right\}^{\otimes 2}}{\sum_{k=1}^{K_1} \omega_k s^{(k,0)}(\tilde{\alpha}_{0,1}, t)} \right] \lambda_0(t) dt, \\
 h_{ii}(\tilde{\alpha}_{0,1}, \tau) &= \int_0^\tau \left[\omega_i s_{(1,1)}^{(i,2)}(\tilde{\alpha}_{0,1}, t) - \frac{\left\{ \omega_i s_1^{(i,1)}(\tilde{\alpha}_{0,1}, t) \right\}^{\otimes 2}}{\sum_{k=1}^{K_1} \omega_k s^{(k,0)}(\tilde{\alpha}_{0,1}, t)} \right] \lambda_0(t) dt, \quad 1 \leq i \leq K_1, \\
 h_{ij}(\tilde{\alpha}_{0,1}, \tau) &= - \int_0^\tau \frac{w_i w_j s_1^{(i,1)}(\tilde{\alpha}_{0,1}, t) s_1^{(j,1)}(\tilde{\alpha}_{0,1}, t)^\top}{\sum_{k=1}^{K_1} \omega_k s^{(k,0)}(\tilde{\alpha}_{0,1}, t)} \lambda_0(t) dt, \quad 1 \leq i \neq j \leq K_1, \\
 h_{12}(\tilde{\alpha}_{0,1}, \tau) &= \{h_{12,j}(\tilde{\alpha}_{0,1}, \tau)\}_{1 \leq j \leq K_1}, \quad \text{for } 1 \leq j \leq K_1, \\
 h_{12,j}(\tilde{\alpha}_{0,1}, \tau) &= \int_0^\tau \left[w_j s_{(1,2)}^{(j,2)}(\tilde{\alpha}_{0,1}, t) - \left\{ \frac{w_j s_1^{(j,1)}(\tilde{\alpha}_{0,1}, t) \sum_{k=1}^{K_1} \omega_k s_2^{(k,1)}(\tilde{\alpha}_{0,1}, t)^\top}{\sum_{k=1}^{K_1} \omega_k s^{(k,0)}(\tilde{\alpha}_{0,1}, t)} \right\} \right] \lambda_0(t) dt.
 \end{aligned}$$

We know that

$$\tilde{\Sigma}^{-1}(\tilde{\alpha}_{0,1}, \tau) = \begin{pmatrix} \Sigma_{11}(\tilde{\alpha}_{0,1}, \tau) & -h^{-1}(\tilde{\alpha}_{0,1}, \tau) h_{12}(\tilde{\alpha}_{0,1}, \tau) C^{-1}(\tilde{\alpha}_{0,1}, \tau) \\ -C^{-1}(\tilde{\alpha}_{0,1}, \tau) h_{12}^\top(\tilde{\alpha}_{0,1}, \tau) h^{-1}(\tilde{\alpha}_{0,1}, \tau) & C^{-1}(\tilde{\alpha}_{0,1}, \tau) \end{pmatrix} \quad (\text{S1.8})$$

where $\Sigma_{11}(\tilde{\alpha}_{0,1}, \tau) = h^{-1}(\tilde{\alpha}_{0,1}, \tau) + h^{-1}(\tilde{\alpha}_{0,1}, \tau) h_{12}(\tilde{\alpha}_{0,1}, \tau) C^{-1}(\tilde{\alpha}_{0,1}, \tau) h_{12}^\top(\tilde{\alpha}_{0,1}, \tau) h^{-1}(\tilde{\alpha}_{0,1}, \tau)$,

$C^{-1}(\tilde{\alpha}_{0,1}, \tau) = h_{22}(\tilde{\alpha}_{0,1}, \tau) h^{-1}(\tilde{\alpha}_{0,1}, \tau) h_{12}(\tilde{\alpha}_{0,1}, \tau)$.

Let $\mathbf{w}^D = \text{diag}(w_1, \dots, w_K)$, and $\mathbf{w} = (w_1, \dots, w_K)^\top$. Using the

Shermann-Morrison-Woodbury formula,

$$(A + uv^\top)^{-1} = A^{-1} - \sigma^{-1} A^{-1} uv^\top A^{-1}, \quad \sigma = 1 + v^\top A^{-1} u,$$

we have that

$$\begin{aligned} h^{-1}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau) &= O((\mathbf{w}^D - \mathbf{w}\mathbf{w}^\top)^{-1}) = O((\mathbf{w}^D)^{-1} + (1 + \mathbf{w}^\top(\mathbf{w}^D)^{-1}\mathbf{w})^{-1}(\mathbf{w}^D)^{-1}\mathbf{w}\mathbf{w}^\top(\mathbf{w}^D)^{-1}) \\ &= O(\mathbf{w}^D)^{-1} + 0.5\mathbf{1}\mathbf{1}^\top, \end{aligned}$$

$$h_{12}^\top(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau)h^{-1}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau)h_{12}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau) = O(1),$$

$$C^{-1}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau) = O(1).$$

So we have

$$\begin{aligned} (\hat{\boldsymbol{\alpha}}_1^{or} - \tilde{\boldsymbol{\alpha}}_{0,1}) &= \tilde{\Sigma}^{-1}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau)(\tilde{A}_n(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau) - \lambda_3 \mathbf{b}) + r_N. \end{aligned}$$

where $r_N = O_p \left(\begin{array}{c} (K_1 + \omega_1^{-1})\|\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_0\|_2^2 + \|\hat{\boldsymbol{\alpha}}_1^{or} - \check{\boldsymbol{\alpha}}_{1,0}\|_2^2 + \sum_{k=1}^{K_1} \omega_k \|\hat{\boldsymbol{\alpha}}_k^{or} - \check{\boldsymbol{\alpha}}_{k,0}\|_2^2 \\ (K_1 + \omega_2^{-1})\|\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_0\|_2^2 + \|\hat{\boldsymbol{\alpha}}_2^{or} - \check{\boldsymbol{\alpha}}_{2,0}\|_2^2 + \sum_{k=1}^{K_1} \omega_k \|\hat{\boldsymbol{\alpha}}_k^{or} - \check{\boldsymbol{\alpha}}_{k,0}\|_2^2 \\ \vdots \\ (K_1 + \omega_{K_1}^{-1})\|\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_0\|_2^2 + \|\hat{\boldsymbol{\alpha}}_{K_1}^{or} - \check{\boldsymbol{\alpha}}_{K_1,0}\|_2^2 + \sum_{k=1}^{K_1} \omega_k \|\hat{\boldsymbol{\alpha}}_k^{or} - \check{\boldsymbol{\alpha}}_{k,0}\|_2^2 \\ K_1\|\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_0\|_2^2 + \sum_{k=1}^{K_1} \omega_k \|\hat{\boldsymbol{\alpha}}_k^{or} - \check{\boldsymbol{\alpha}}_{k,0}\|_2^2 \end{array} \right)$

Under the condition $N < N_{\min}^2$ and $K_1 = O(N^{1/2-\delta})$ for some $0 < \delta \leq 1/2$,

if $\sqrt{N_{\min}}\lambda_3 \rightarrow 0$, we have for $j = 1, 2, \dots, K_1$,

$$\sqrt{N_k} \begin{pmatrix} \hat{\boldsymbol{\alpha}}_{1,k,1}^{or} - \boldsymbol{\alpha}_{0,1,k,1} \\ \hat{\alpha}_{2,k}^{or} - \alpha_{0,2} \\ \hat{\alpha}_{3,k}^{or} - \alpha_{0,3} \end{pmatrix} \rightarrow N(\mathbf{0}, \tilde{\Sigma}_{11,k}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau)),$$

and

$$\sqrt{N} \begin{pmatrix} \hat{\boldsymbol{\gamma}}^{or} - \boldsymbol{\gamma}_0 \\ \hat{\boldsymbol{\alpha}}_4^{or} - \boldsymbol{\alpha}_{0,4} \end{pmatrix} \rightarrow N(\mathbf{0}, C^{-1}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau)),$$

where $\tilde{\Sigma}_{11,k}(\tilde{\alpha}_{0,1}, \tau) = \omega_k \Sigma_{11,k}(\tilde{\alpha}_{0,1}, \tau)$, and $\Sigma_{11,k}(\tilde{\alpha}_{0,1}, \tau)$ is the $(p_{k_s} + 2) \times (p_{k_s} + 2)$ submatrix of $\Sigma_{11}(\tilde{\alpha}_{0,1}, \tau)$.

S1.6 Proof of Theorem 3

Let

$$\begin{aligned} b_1 &= \min_{k \neq k'} \|(\alpha_{1,k,1}^\top, \alpha_{2,k}, \alpha_{3,k}) - (\alpha_{1,k',1}^\top, \alpha_{2,k'}, \alpha_{3,k'})\|_2, \\ b_2 &= \min_{k \neq k'} |\alpha_{4,k} - \alpha_{4,k'}|. \end{aligned}$$

Denote $\theta_i = (\zeta_i^\top, \eta_i, v_i)^\top$, $\theta = (\theta_1^\top, \dots, \theta_n^\top)^\top$. Define the oracle parameter space as

$$\begin{aligned} \mathcal{M}_{\mathcal{G}_{0,1}} &= \{\theta \in R^{n(p+2)} : \theta_i = \theta_j = (\alpha_{1,k}^\top, \alpha_{2,k}, \alpha_{3,k})^\top, \text{ for any } i, j \in \mathcal{G}_{0,k}, 1 \leq k \leq K_1\}, \\ \mathcal{M}_{\mathcal{D}_{0,2}} &= \{g \in R^p : g_j = g_{j'} = \alpha_{4,k}, \text{ for any } j, j' \in \mathcal{D}_{0,k}, 1 \leq k \leq K_2\}. \end{aligned}$$

Define the mapping $T_1^* : R^{n(p+2)} \rightarrow R^{K_1(p+2)}$ and $T_2^* : R^p \rightarrow R^{K_2}$ as

$$\begin{aligned} T_1^*(\theta) &= \{|\mathcal{G}_{0,k}|^{-1} \sum_{i \in \mathcal{G}_{0,k}} \theta_i, k = [K_1]\}, \\ T_2^*(g) &= \{|\mathcal{D}_{0,k}|^{-1} \sum_{j \in \mathcal{D}_{0,k}} g_j, k = [K_2]\}, \end{aligned}$$

respectively. Let the one-to-one mapping $T_1 : \mathcal{M}_{\mathcal{G}_{0,1}} \rightarrow R^{K_1(p+2)}$, such that $T_1(\theta)$ is the $K_1(p+2)$ vector whose k -th sub-vector equal to the common values of θ_i for $i \in \mathcal{G}_{0,k}$ and satisfying $T_1(\theta) = T_1^*(\theta)$. $T_2 : \mathcal{M}_{\mathcal{D}_{0,2}} \rightarrow R^{K_2}$, such that $T_2(g)$ is the $K_2 \times 1$ vector whose k -th sub-vector equal to the

common values of g_i for $i \in \mathcal{D}_{0,k}$ satisfying $T_2(\mathbf{g}) = T_2^*(\mathbf{g})$. For any vector $\boldsymbol{\theta} \in R^{n(p+2)}$ and $\mathbf{g} \in R^p$, set $\boldsymbol{\alpha}^{sa} = T_1^*(\boldsymbol{\theta})$, $\boldsymbol{\alpha}_4 = T_2^*(\mathbf{g})$ and $\boldsymbol{\theta}^* = T_1^{-1}(T_1^*(\boldsymbol{\theta}))$ and $\mathbf{g}^* = T_2^{-1}(T_2^*(\mathbf{g}))$. Clearly, for any vector $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}^* \in \mathcal{M}_{\mathcal{G}_0,1}$ and $\mathbf{g}^* \in \mathcal{M}_{\mathcal{D}_0,2}$. Define

$$S_N^{\mathcal{M}}(\boldsymbol{\theta}^*, \mathbf{g}^*, \boldsymbol{\gamma}) = L_{a,n}(\boldsymbol{\theta}^*, \mathbf{g}^*, \boldsymbol{\gamma}) - p_{\lambda_1}(\boldsymbol{\theta}^*, T_1^*(\boldsymbol{\theta})) - p_{\lambda_2}(\mathbf{g}^*, T_2^*(\mathbf{g})) - p_{\lambda_3}(\boldsymbol{\alpha}_1),$$

$$S_N(\boldsymbol{\theta}, \mathbf{g}, \boldsymbol{\gamma}) = S_{DCAF}(\boldsymbol{\Theta}, \boldsymbol{\alpha}) = L_n(\boldsymbol{\Theta}) - p_{\lambda_1}(\boldsymbol{\theta}, \boldsymbol{\alpha}^{sa}) - p_{\lambda_2}(\mathbf{g}, \boldsymbol{\alpha}_4) - p_{\lambda_3}(\boldsymbol{\alpha}_1),$$

where $L_{a,n}(\boldsymbol{\theta}^*, \mathbf{g}^*, \boldsymbol{\gamma}) = \tilde{L}_N(\boldsymbol{\alpha}^{sa}, \boldsymbol{\alpha}_4, \boldsymbol{\gamma}, \tau)$, $p_{\lambda_1}(\boldsymbol{\theta}^*, T_1^*(\boldsymbol{\theta})) = \lambda_1 \sum_{i=1}^n (1/K_1 \sum_{k=1}^{K_1} \|\boldsymbol{\theta}_i^* - \boldsymbol{\alpha}_k^{sa}\|^{2s})^{1/s}$, $p_{\lambda_2}(\mathbf{g}^*, T_2^*(\mathbf{g})) = \lambda_2 \sum_{j=1}^p (1/K_2 \sum_{k=1}^{K_2} \|g_j^* - \alpha_{4,k}\|^{2s})^{1/s}$, and

$$\begin{aligned} & \tilde{L}_N(\boldsymbol{\alpha}^{sa}, \boldsymbol{\alpha}_4, \tilde{\boldsymbol{\gamma}}, \tau) \\ &= \sum_{k=1}^{K_1} \sum_{i \in \mathcal{G}_{0,k}} \sum_{j=1}^{m_i} \int_0^\tau \left[\boldsymbol{\alpha}_k^{sa\top} (\mathbf{x}_{ij}^\top, 1, w_{ij})^\top + \boldsymbol{\gamma}^\top \mathbf{z}_{ij} + \sum_{l=1}^{K_2} \left(\sum_{s \in \mathcal{D}_{0,l}} x_{ij,s} w_{ij} \right) \alpha_{4,l} \right] dN_{ij}(t) \\ & - \int_0^t \log \left\{ \sum_{k=1}^{K_1} \sum_{i \in \mathcal{G}_{0,k}} \sum_{j=1}^{m_i} Y_{ij}(s) \exp \left[\boldsymbol{\alpha}_k^{sa\top} (\mathbf{x}_{ij}^\top, 1, w_{ij})^\top + \boldsymbol{\gamma}^\top \mathbf{z}_{ij} + \sum_{l=1}^{K_2} \left(\sum_{s \in \mathcal{D}_{0,l}} x_{ij,s} w_{ij} \right) \alpha_{4,l} \right] \right\} d\bar{N}(t), \end{aligned}$$

Consider the neighborhood of $(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, \mathbf{g}_0)$, that is, $\boldsymbol{\Theta}_{\phi_n} = \{\boldsymbol{\theta} \in R^{n(p+2)}, \boldsymbol{\gamma} \in$

$R^q, \mathbf{g} \in R^p : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|_2 \leq \phi_n, \max_i \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0,i}\|_2 \leq \phi_n, \max_j |g_j - g_{0,j}| \leq \phi_n\}$.

where $\phi_n \rightarrow 0$ as $N_{\min}$ goes to infinity. By the result in Theorem 1, there

is an event E_1 such that on the event E_1 , $(\hat{\boldsymbol{\theta}}^{or}, \hat{\boldsymbol{\gamma}}^{or}, \hat{\mathbf{g}}^{or}) \in \boldsymbol{\Theta}_{\phi_n}$ on the

event E_1 . We show that $(\hat{\boldsymbol{\theta}}^{or}, \hat{\boldsymbol{\gamma}}^{or}, \hat{\mathbf{g}}^{or})$ is the strictly local maximizer of the

objective function with the probability approaching 1 through the following

two steps.

- (i) On the Event E_1 , for any $(\boldsymbol{\theta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{g}^\top)^\top \in \boldsymbol{\Theta}_{\phi_n}$, if $((\boldsymbol{\theta}^*)^\top, \boldsymbol{\gamma}^\top, (\boldsymbol{g}^*)^\top)^\top \neq ((\hat{\boldsymbol{\theta}}^{or})^\top, (\hat{\boldsymbol{\gamma}}^{or})^\top, (\hat{\boldsymbol{g}}^{or})^\top)^\top$, then $S_N(\boldsymbol{\theta}^*, \boldsymbol{\gamma}, \boldsymbol{g}^*) < S_N(\hat{\boldsymbol{\theta}}^{or}, \hat{\boldsymbol{\gamma}}^{or}, \hat{\boldsymbol{g}}^{or})$.
- (ii) For any $(\boldsymbol{\theta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{g}^\top)^\top \in \boldsymbol{\Theta}_{\phi_n}$, there is a neighborhood of $((\hat{\boldsymbol{\theta}}^{or})^\top, (\hat{\boldsymbol{\gamma}}^{or})^\top, (\hat{\boldsymbol{g}}^{or})^\top)^\top$ denoted by $\boldsymbol{\Theta}_{\phi_n}^{or}$, such that $S_N(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{g}) \leq S_N(\boldsymbol{\theta}^*, \boldsymbol{\gamma}, \boldsymbol{g}^*)$ for any $(\boldsymbol{\theta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{g}^\top)^\top \in \boldsymbol{\Theta}_{\phi_n} \cap \boldsymbol{\Theta}_{\phi_n}^{or}$.

Therefore, by the results of (i) and (ii), we have $S_N(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{g}; T_1^*(\boldsymbol{\theta}), T_2^*(\boldsymbol{g})) \leq S_N(\hat{\boldsymbol{\theta}}^{or}, \hat{\boldsymbol{\gamma}}^{or}, \hat{\boldsymbol{g}}^{or}; T_1(\hat{\boldsymbol{\theta}}^{or}), T_2(\hat{\boldsymbol{g}}^{or}))$ for any $(\boldsymbol{\theta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{g}^\top)^\top \in \boldsymbol{\Theta}_{\phi_n} \cap \boldsymbol{\Theta}_{\phi_n}^{or}$ and $(\boldsymbol{\theta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{g}^\top)^\top \neq ((\hat{\boldsymbol{\theta}}^{or})^\top, (\hat{\boldsymbol{\gamma}}^{or})^\top, (\hat{\boldsymbol{g}}^{or})^\top)^\top$, so that $((\hat{\boldsymbol{\theta}}^{or})^\top, (\hat{\boldsymbol{\gamma}}^{or})^\top, (\hat{\boldsymbol{g}}^{or})^\top)^\top$ is a strict local maximizer of $S_N(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{g})$ given in (2.5).

In the following we prove the result in (i), note that when $s < 0$,

$$\begin{aligned} & p_{\lambda_1}(\boldsymbol{\theta}^*, T_1(\boldsymbol{\theta}^*)) + p_{\lambda_2}(\boldsymbol{g}^*, T_2(\boldsymbol{g}^*)) \\ &= \lambda_1 \sum_{k=1}^{K_1} |\mathcal{G}_{0,k}| \left(1/K_1 \sum_{k'=1}^{K_1} \|\boldsymbol{\alpha}_k^{sa} - \boldsymbol{\alpha}_{k'}^{sa}\|_2^{2s} \right)^{1/s} + \lambda_2 \sum_{k=1}^{K_2} |\mathcal{D}_{0,k}| \left(1/K_2 \sum_{k'=1}^{K_2} |\alpha_{4,k} - \alpha_{4,k'}|^{2s} \right)^{1/s} \\ &= 0, \end{aligned}$$

where the last equation follows from the fact that $1/K \sum_{k=1}^K a_k^{-2} = 0$ when

some $a_k = 0$. Similarly, we have

$$\begin{aligned} & p_{\lambda_1}(\hat{\boldsymbol{\theta}}^{or}, T_1(\hat{\boldsymbol{\theta}}^{or})) + p_{\lambda_2}(\hat{\boldsymbol{g}}^{or}, T_2(\hat{\boldsymbol{g}}^{or})) \\ &= \lambda_1 \sum_{k=1}^{K_1} |\mathcal{G}_{0,k}| \left(1/K_1 \sum_{k'=1}^{K_1} \|\hat{\boldsymbol{\alpha}}_k^{sa,or} - \hat{\boldsymbol{\alpha}}_{k'}^{sa,or}\|_2^{2s} \right)^{1/s} + \lambda_2 \sum_{k=1}^{K_2} |\mathcal{D}_{0,k}| \left(1/K_2 \sum_{k'=1}^{K_2} |\hat{\alpha}_{4,k}^{or} - \hat{\alpha}_{4,k'}^{or}|^{2s} \right)^{1/s} \\ &= 0. \end{aligned}$$

Note that $\left((\hat{\boldsymbol{\theta}}^{or})^\top, (\hat{\boldsymbol{\gamma}}^{or})^\top, (\hat{\boldsymbol{g}}^{or})^\top\right)^\top$ is the local maximizer of $L_{a,n}(\boldsymbol{\theta}, \boldsymbol{g}, \boldsymbol{\gamma}) - p_{\lambda_3}(\boldsymbol{\alpha}_1)$, then $S_N^{\mathcal{M}}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}, \boldsymbol{g}^*; T_1(\boldsymbol{\theta}^*), T_2(\boldsymbol{g}^*)) \leq S_N^{\mathcal{M}}(\hat{\boldsymbol{\theta}}^{or}, \hat{\boldsymbol{\gamma}}^{or}, \hat{\boldsymbol{g}}^{or}; T_1(\hat{\boldsymbol{\theta}}^{or}), T_2(\hat{\boldsymbol{g}}^{or}))$, we have $S_N^{\mathcal{M}}(\hat{\boldsymbol{\theta}}^{or}, \hat{\boldsymbol{\gamma}}^{or}, \hat{\boldsymbol{g}}^{or}; T_1(\hat{\boldsymbol{\theta}}^{or}), T_2(\hat{\boldsymbol{g}}^{or})) = S_N(\hat{\boldsymbol{\theta}}^{or}, \hat{\boldsymbol{\gamma}}^{or}, \hat{\boldsymbol{g}}^{or}; T_1(\hat{\boldsymbol{\theta}}^{or}), T_2(\hat{\boldsymbol{g}}^{or}))$, and $S_N^{\mathcal{M}}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}, \boldsymbol{g}^*; T_1(\boldsymbol{\theta}^*), T_2(\boldsymbol{g}^*)) = S_N(\boldsymbol{\theta}^*, \boldsymbol{\gamma}, \boldsymbol{g}^*; T_1(\boldsymbol{\theta}^*), T_2(\boldsymbol{g}^*))$, and the result in (i) is proved.

Next we prove the result in (ii). For $(\boldsymbol{\theta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{g}^\top)^\top \in \boldsymbol{\Theta}_{\phi_n} \cap \mathcal{D}_n$, by Taylor's expansion, we have

$$\begin{aligned} & S_N(\boldsymbol{\theta}, \boldsymbol{g}, \boldsymbol{\gamma}; \boldsymbol{\alpha}^{sa}, \boldsymbol{\alpha}_4) - S_N(\boldsymbol{\theta}^*, \boldsymbol{g}^*, \boldsymbol{\gamma}; T_1^*(\boldsymbol{\theta}), T_2^*(\boldsymbol{g})) \\ &= \frac{\partial L_n(\boldsymbol{\Theta})}{\partial(\boldsymbol{\theta}^\top, \boldsymbol{g}^\top)} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^m, \boldsymbol{g}=\boldsymbol{g}^m} \begin{pmatrix} \boldsymbol{\theta} - \boldsymbol{\theta}^* \\ \boldsymbol{g} - \boldsymbol{g}^* \end{pmatrix} - p_{\lambda_1}(\boldsymbol{\theta}, \boldsymbol{\alpha}^{sa}) - p_{\lambda_2}(\boldsymbol{g}, \boldsymbol{\alpha}_4) \\ &= \Gamma_1 - p_{\lambda_1}(\boldsymbol{\theta}, \boldsymbol{\alpha}^{sa}) - p_{\lambda_2}(\boldsymbol{g}, \boldsymbol{\alpha}_4), \end{aligned}$$

where $\boldsymbol{\theta}^m$ is a vector between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$, and $\boldsymbol{g}^m$ is a vector between $\boldsymbol{g}$ and $\boldsymbol{g}^*$.

For Γ_1 , let

$$\begin{aligned} U_{1i} &= \frac{\partial L_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\theta}_i} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^m, \boldsymbol{g}=\boldsymbol{g}^m} = \sum_{j=1}^{m_i} \int_0^\tau (\mathbf{x}_{ij}^\top, 1, w_{ij})^\top dN_{ij}(t) \\ &\quad - \int_0^\tau \frac{\sum_{j=1}^{m_i} Y_{ij}(t) (\mathbf{x}_{ij}^\top, 1, w_{ij})^\top \exp((\boldsymbol{\theta}_i^m)^\top (\mathbf{x}_{ij}^\top, 1, w_{ij})^\top + \boldsymbol{\gamma}^\top \mathbf{z}_{ij} + (\boldsymbol{g}^m)^\top \mathbf{x}_{ij} w_{ij})}{\sum_{i=1}^n \sum_{j=1}^{m_i} Y_{ij}(t) \exp((\boldsymbol{\theta}_i^m)^\top (\mathbf{x}_{ij}^\top, 1, w_{ij})^\top + \boldsymbol{\gamma}^\top \mathbf{z}_{ij} + (\boldsymbol{g}^m)^\top \mathbf{x}_{ij} w_{ij})} d\bar{N}(t), \end{aligned}$$

and

$$\begin{aligned} U_{2l} &= \frac{\partial L_{a,n}(\boldsymbol{\Theta})}{\partial g_l} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^m, \boldsymbol{g}=\boldsymbol{g}^m} = \sum_{i=1}^n \sum_{j=1}^{m_i} \int_0^\tau x_{ij,l} w_{ij} dN_{ij}(t) \\ &\quad - \int_0^\tau \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} Y_{ij}(t) x_{ij,l} w_{ij} \exp((\boldsymbol{\theta}_i^m)^\top (\mathbf{x}_{ij}^\top, 1, w_{ij})^\top + \boldsymbol{\gamma}^\top \mathbf{z}_{ij} + (\boldsymbol{g}^m)^\top \mathbf{x}_{ij} w_{ij})}{\sum_{i=1}^n \sum_{j=1}^{m_i} Y_{ij}(t) \exp((\boldsymbol{\theta}_i^m)^\top (\mathbf{x}_{ij}^\top, 1, w_{ij})^\top + \boldsymbol{\gamma}^\top \mathbf{z}_{ij} + (\boldsymbol{g}^m)^\top \mathbf{x}_{ij} w_{ij})} d\bar{N}(t), \end{aligned}$$

where $\bar{N}(t) = \sum_{i=1}^n \sum_{j=1}^{m_i} N_{ij}(t)$. Then after some calculation, we have

$$\begin{aligned}
 \Gamma_1 &= \sum_{i=1}^n U_{1i}^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) + \sum_{l=1}^p U_{2l} (g_l - g_l^*) \\
 &= \sum_{k=1}^{K_1} \sum_{i \in \mathcal{G}_{0,k}} U_{1i}^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) + \sum_{l=1}^{K_2} \sum_{j \in \mathcal{D}_{0,l}} U_{2j} (g_j - g_j^*) \\
 &\stackrel{(a)}{=} \sum_{k=1}^{K_1} \sum_{i,j \in \mathcal{G}_{0,k}} \frac{U_{1i}^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{|\mathcal{G}_{0,k}|} + \sum_{l=1}^{K_2} \sum_{j,i \in \mathcal{D}_{0,l}} \frac{U_{2j} (g_j - g_i)}{|\mathcal{D}_{0,l}|} \\
 &= \sum_{k=1}^{K_1} \sum_{i,j \in \mathcal{G}_{0,k}} \frac{(U_{1i} - U_{1j})^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{2|\mathcal{G}_{0,k}|} + \sum_{l=1}^{K_2} \sum_{i,j \in \mathcal{D}_{0,l}} \frac{(U_{2i} - U_{2j})(g_i - g_j)}{2|\mathcal{D}_{0,l}|} \\
 &\stackrel{(b)}{\leq} \sum_{k=1}^{K_1} \sum_{i,j \in \mathcal{G}_{0,k}, i < j} \frac{2 \max_l \|U_{1l}\| \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|}{|\mathcal{G}_{0,k}|} + \sum_{l=1}^{K_2} \sum_{i,j \in \mathcal{D}_{0,l}, i < j} \frac{2 \max_k |U_{2l}| |g_i - g_j|}{|\mathcal{D}_{0,l}|}.
 \end{aligned}$$

(a) is due to the definition of $\boldsymbol{\theta}_i^* = \frac{1}{|\mathcal{G}_{0,k}|} \sum_{j \in \mathcal{G}_{0,k}} \boldsymbol{\theta}_j$, $\sum_{i \in \mathcal{G}_{0,k}} U_{1i}^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) = \sum_{i \in \mathcal{G}_{0,k}} U_{1i}^\top (\boldsymbol{\theta}_i - \frac{1}{|\mathcal{G}_{0,k}|} \sum_{j \in \mathcal{G}_{0,k}} \boldsymbol{\theta}_j) = \sum_{i,j \in \mathcal{G}_{0,k}} \frac{U_{1i}^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{|\mathcal{G}_{0,k}|}$, similarly, $\sum_{j \in \mathcal{D}_{0,l}} U_{2j} (g_j - g_j^*) = \sum_{j,i \in \mathcal{D}_{0,l}} \frac{U_{2j} (g_j - g_i)}{|\mathcal{D}_{0,l}|}$. (b) is due to the Cauchy-Swarch inequality, $(U_{1i} - U_{1j})^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) \leq \|U_{1i} - U_{1j}\| \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| \leq 2 \max_l \|U_{1l}\| \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|$, similarly, $(U_{2i} - U_{2j})(g_i - g_j) \leq 2 \max_l |U_{2l}| |g_i - g_j|$.

When $s < 0$, using $\max_i \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0,i}\|_2 \leq \phi_n$ and $b_1, b_2 \gg \phi_n$, we have

$$\begin{aligned}
 &p_{\lambda_1}(\boldsymbol{\theta}, \boldsymbol{\alpha}^{sa}) + p_{\lambda_2}(\mathbf{g}, \boldsymbol{\alpha}_4) \\
 &= \lambda_1 \sum_{k=1}^{K_1} \sum_{i \in \mathcal{G}_{0,k}} \left(1/K_1 \sum_{k'=1}^{K_1} \|\boldsymbol{\theta}_i - \boldsymbol{\alpha}_{k'}^{sa}\|_2^{2s} \right)^{1/s} + \lambda_2 \sum_{k=1}^{K_2} \sum_{j \in \mathcal{D}_{0,k}} \left(1/K_2 \sum_{k'=1}^{K_2} |g_j - \alpha_{4,k'}|^{2s} \right)^{1/s} \\
 &= \lambda_1 \sum_{k=1}^{K_1} |\mathcal{G}_{0,k}| K_1^{-1/s} \phi_n^2 + \lambda_2 \sum_{k=1}^{K_2} |\mathcal{D}_{0,k}| \phi_n^2 K_2^{-1/s}.
 \end{aligned}$$

Combining the above two inequalities, we have

$$\begin{aligned}
& \Gamma_1 - p_{\lambda_1}(\boldsymbol{\theta}, \boldsymbol{\alpha}^{sa}) - p_{\lambda_2}(\mathbf{g}, \boldsymbol{\alpha}_4) \\
& \leq \sum_{k=1}^{K_1} \sum_{i,j \in \mathcal{G}_{0,k}, i < j} \frac{2 \max_l \|U_{1l}\| \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|}{|\mathcal{G}_{0,k}|} - \lambda_1 \sum_{k=1}^{K_1} |\mathcal{G}_{0,k}| K_1^{-1/s} \phi_n^2 \\
& \quad + \sum_{l=1}^{K_2} \sum_{i,j \in \mathcal{D}_{0,l}, i < j} \frac{2 \max_k |U_{2k}| |g_i - g_j|}{|\mathcal{D}_{0,k}|} - \lambda_2 \sum_{k=1}^{K_2} |\mathcal{D}_{0,k}| \phi_n^2 K_2^{-1/s} \\
& = \sum_{k=1}^{K_1} |\mathcal{G}_{0,k}| \phi_n \left(2 \max_l \|U_{1l}\| - \lambda_1 K_1^{-1/s} \phi_n \right) + \sum_{k=1}^{K_2} |\mathcal{D}_{0,k}| \phi_n \left(2 \max_l \|U_{2l}\| - \lambda_2 K_2^{-1/s} \phi_n \right) \\
& \leq 0,
\end{aligned}$$

where the last inequality holds under the conditions that $\lambda_1 K_1^{-1/s} \phi_n > 2 \max_l \|U_{1l}\|$ and $\lambda_2 K_2^{-1/s} \phi_n > 2 \max_l \|U_{2l}\|$, and $\max_l \|U_{1l}\| < c_1$ and $\max_j \|U_{2j}\| < c_1$ from condition (C2) for some positive constant c , so that the result in (ii) is proved.

S2 Procedures on calculating standard errors and the p -values

S2.1 Splitting procedure

- With the full data, determine the number of the subgroup K_1 , K_2 and model S ;
- Let B be a large positive integer, for each $b = 1, \dots, B$,

- (i) randomly split the data into equal halves (split in country or hospital), D_1^b and D_2^b ;
 - (ii) on D_1^b , apply S to select a subset of predictors, denote $\hat{S}_1^b$;
 - (iii) on D_2^b , for each $j = 1, \dots, p * K_1 - 1 + K_2 + q$, append j to $\hat{S}_1^b$ such that $\hat{S}_{+j1}^b = j \cup \hat{S}_1^b$, fit a Cox proportional hazard model on $\hat{S}_{+j1}^b$, denote as $\tilde{\beta}_{j1}^b$.
- Fusing: the estimator for $j = 1, \dots, p * K_1 - 1 + K_2 + q$ is $\hat{\beta}_{j1} = \frac{1}{B} \sum_{b=1}^B \tilde{\beta}_{j1}^b$;
 - covert (ii) and (iii), $\hat{\beta}_{j2} = \frac{1}{B} \sum_{b=1}^B \tilde{\beta}_{j2}^b$;
 - Average: the final estimator for $j = 1, \dots, p * K_1 - 1 + K_2 + q$ is $\hat{\beta}_j = \frac{\hat{\beta}_{j1} + \hat{\beta}_{j2}}{2}$.

S2.2 A variance estimator via the functional Delta method

Let $J_{bi} \in \{0, 1\}$ indicate whether the i -th observation is in the b^{th} subsample of the second data part, and $J_i = \frac{\sum_{b=1}^B J_{bi}}{B}$. For each $i = 1, \dots, n$, we define the re-sampling covariance between J_{bi} and $\tilde{\beta}_{j1}^b$ as

$$s_{ij} = \frac{\sum_{b=1}^B (J_{bi} - J_i)(\tilde{\beta}_{j1}^b - \hat{\beta}_{j1})}{B}.$$

Define $S_j = (s_{1j}, s_{2j}, \dots, s_{nj})$, and let $n_1 = |D_2^b|$. Then, a variance estimator for $\hat{\beta}_{j1}$ is

$$\hat{V}_{j1} = \frac{n(n-1)}{(n-n_1)^2} S_j^T S_j. \quad (\text{S2.9})$$

Furthermore, for a finite B, we propose a bias corrected version of (S2.9):

$$\hat{V}_{j1}^B = \hat{V}_{j1} - \frac{nn_1}{B(n-n_1)} \left\{ B^{-1} \sum_{b=1}^B (\tilde{\beta}_{j1}^b - \hat{\beta}_{j1})^2 \right\}.$$

Finally, a variances estimator for $\hat{\beta}_j$ as follows,

$$\hat{V}_j = \frac{\hat{V}_{j1} + \hat{V}_{j2}}{4} \quad \text{or} \quad \hat{V}_j^B = \frac{\hat{V}_{j1}^B + \hat{V}_{j2}^B}{4}. \quad (\text{S2.10})$$

When the sample size is large, we only choose $\hat{\beta}_{j1}$ or $\hat{\beta}_{j2}$ as the estimator, and $\hat{V}_{j1}^B$ or $\hat{V}_{j2}^B$ as the corresponding variances estimator. The p -value of testing $H_0 : \beta_j = 0$ is $1.96 \times (1 - \Phi(|\hat{\beta}_j| / \sqrt{\hat{V}_j}))$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

S3 Tuning Parameter Selection

In Section 2, five parameters need to be selected: the number of two-way subgroups K_1, K_2 , the thresholding parameters λ_1, λ_2 and the shrinkage parameter λ_3 . We propose to use the BIC to simultaneously select $K_1, K_2, \lambda_1, \lambda_2$ and λ_3 by minimizing

$$\text{BIC} = -L_n(\boldsymbol{\theta}, \mathbf{g}, \boldsymbol{\gamma})/N + C_n \log N/N(|\mathcal{S}| + K_1 + K_2 + q), \quad (\text{S3.11})$$

where $N = \sum_{i=1}^n m_i$, $C_n = C \log(n(p+2)+p+q)$, $|\mathcal{S}| = \sum_{k=1}^{K_1} \sum_{j=1}^p I(\alpha_{1,kj} \neq 0)$, and C is a constant. Different values of λ_r , $r = 1, 2, 3$ correspond to distinct estimates of $\boldsymbol{\theta}$, $\mathbf{g}$ and $\boldsymbol{\gamma}$. It is important to note that λ_1 and λ_2 impact the variability of $\boldsymbol{\theta}_i$ and g_j within each group, respectively. A higher $\lambda_1(\lambda_2)$ results in reduced variability of $\boldsymbol{\theta}_i(g_j)$ within each group. For instance, when $\lambda_1 = \infty$, $\boldsymbol{\theta}_i \equiv \boldsymbol{\alpha}_k$ for $i \in \mathcal{G}_k$; whereas when $\lambda_i = 0$, $\boldsymbol{\theta}_i$ is the solution that minimizes the partial likelihood. λ_3 influences the sparsity level of $\boldsymbol{\alpha}_{1,k}$. A larger λ_3 leads to increased sparsity of $\boldsymbol{\alpha}_{1,k}$. Consequently, λ_1 and λ_2 influence the first term of equation (S3.11) by altering $\boldsymbol{\theta}$, $\mathbf{g}$ values and subsequently $\boldsymbol{\gamma}$. λ_3 affects the first term of equation (S3.11) by altering $\boldsymbol{\alpha}_{1,k}$ values, which in turn impacts $\boldsymbol{\theta}_i$, $\mathbf{g}$, and $\boldsymbol{\gamma}$. Furthermore, λ_3 impacts the second term of equation (S3.11) through $|\mathcal{S}| := \sum_{k=1}^{K_1} \sum_{j=1}^p I(\alpha_{1,kj} \neq 0)$. In practice, we adopt a hierarchical search strategy: for each candidate pair (K_1, K_2) , we first narrow the range of λ_1 and λ_2 using a coarse grid, then refine jointly with λ_3 on a finer grid, and select the combination that minimizes BIC. Warm starts from the nearest previous configuration are used to reduce computation. We found that the BIC proposed in (S3.11) worked well, as shown by the results of Simulation Studies.

S4 Additional numerical studies

S4.1 The estimation results of Simulation Study 1

The estimation results of the and Setting 1 and Setting 2 are presented in Table S1 and Table S2, respectively.

S4. ADDITIONAL NUMERICAL STUDIES

Table S1: The simulation results of Setting 1 of Simulation Study 1

(n_1, n_2, n_3) = (10, 10, 10)		Rig-ref			Pairwise			FMCox			DCAF			Oracle		
30%	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
α_1	0.060	0.287	0.294	0.049	0.238	0.243	0.039	0.242	0.245	0.047	0.174	0.180	0.050	0.170	0.178	
α_2	0.086	0.290	0.302	0.089	0.265	0.280	0.070	0.261	0.271	0.034	0.215	0.218	0.032	0.197	0.199	
α_3	0.018	0.229	0.229	0.026	0.224	0.226	0.016	0.223	0.224	0.049	0.217	0.223	0.054	0.219	0.225	
Accuracy	1*			0.740			0.960			1			1			
RI	0.941			0.937			0.947			0.982			1			
15%	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
α_1	0.012	0.186	0.186	0.035	0.216	0.219	0.015	0.177	0.177	0.041	0.155	0.160	0.044	0.154	0.161	
α_2	0.060	0.211	0.219	0.051	0.226	0.231	0.052	0.211	0.217	0.034	0.188	0.191	0.030	0.177	0.180	
α_3	0.016	0.202	0.202	0.011	0.202	0.202	0.026	0.204	0.205	0.043	0.194	0.199	0.048	0.193	0.199	
Accuracy	1*			0.790			0.950			1			1			
RI	0.966			0.967			0.969			0.990			1			
(n_1, n_2, n_3) = (15, 7, 8)		Rig-ref			Pairwise			FMCox			DCAF			Oracle		
30%	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
α_1	0.057	0.294	0.299	0.053	0.339	0.344	0.056	0.294	0.299	0.045	0.214	0.219	0.056	0.205	0.212	
α_2	0.020	0.191	0.192	0.032	0.412	0.413	0.010	0.190	0.191	0.030	0.173	0.176	0.034	0.168	0.171	
α_3	0.014	0.199	0.200	0.015	0.270	0.270	0.014	0.199	0.199	0.040	0.190	0.194	0.045	0.191	0.196	
Accuracy	1*			0.710			0.995			1			1			
RI	0.966			0.951			0.967			0.991			1			
15%	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
α_1	0.021	0.241	0.242	0.043	0.244	0.247	0.016	0.227	0.228	0.045	0.190	0.195	0.051	0.185	0.192	
α_2	0.014	0.181	0.181	0.037	0.211	0.214	0.016	0.175	0.175	0.028	0.161	0.164	0.032	0.154	0.158	
α_3	0.017	0.188	0.189	0.016	0.190	0.191	0.022	0.180	0.182	0.038	0.176	0.181	0.042	0.177	0.182	
Accuracy	1*			0.810			0.990			1			1			
RI	0.967			0.968			0.981			0.994			1			

S4.2 The estimation results of Simulation Study 2

Substudy 2.1(Continuous biomarkers) In this substudy, the survival time for the j -th subject in the i -th cluster was generated from $T_{ij} = \varepsilon_{ij} / \exp(\eta_i + \boldsymbol{\zeta}_i^\top \mathbf{x}_{ij} + v_i w_{ij} + \mathbf{g}^\top \mathbf{x}_{ij} w_{ij} + \boldsymbol{\gamma}^\top \mathbf{z}_{ij})$, $i = [n], j = [m_i]$, where the treatment variable $w_{ij} \stackrel{i.i.d.}{\sim} \text{Bernoulli}(0.5)$, and $\mathbf{z}_{ij} = (z_{ij,1}, \dots, z_{ij,q})^\top$ were confounder variables with $q = 2$, $z_{ij,1} \stackrel{i.i.d.}{\sim} \text{Bernoulli}(0.5)$, and $z_{ij,2} \stackrel{i.i.d.}{\sim} N(0,1)$. $\boldsymbol{\nu} = (v_1, \dots, v_n)^\top = (\underbrace{\alpha_{3,1}, \dots, \alpha_{3,1}}_{n_1}, \underbrace{\alpha_{3,2}, \dots, \alpha_{3,2}}_{n_2}, \underbrace{\alpha_{3,3}, \dots, \alpha_{3,3}}_{n_3})^\top$, where $\alpha_{3,1} = -0.5$, $\alpha_{3,2} = 0$, and $\alpha_{3,3} = 0.5$. The predictive effect $\mathbf{g} = (g_1, \dots, g_p)^\top = (\underbrace{\alpha_{4,1}, \dots, \alpha_{4,1}}_{p/2}, \underbrace{\alpha_{4,2}, \dots, \alpha_{4,2}}_{p/2})^\top$

Table S2: The simulation results of Setting 2 of Simulation Study 1

(n_1, n_2, n_3)		Rig-ref			FMCox			DCAF			Oracle		
= (10, 10, 10)													
30%	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
α_1	0.782	0.579	0.973	0.278	0.464	0.541	0.053	0.196	0.204	0.052	0.187	0.195	
α_2	0.580	0.534	0.789	0.280	0.338	0.439	0.027	0.134	0.137	0.024	0.132	0.134	
α_3	0.320	0.163	0.359	0.110	0.181	0.211	0.025	0.144	0.147	0.023	0.145	0.146	
Sensitivity		0.715			0.953			1.000			1		
Specificity		0.950			0.974			0.998			1		
Accuracy		1*			1*			1			1		
RI		0.834			0.892			0.998			1		
15%	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
α_1	0.578	0.325	0.664	0.250	0.301	0.391	0.045	0.171	0.177	0.042	0.165	0.170	
α_2	0.510	0.259	0.572	0.332	0.251	0.416	0.016	0.119	0.120	0.015	0.117	0.118	
α_3	0.341	0.148	0.371	0.155	0.173	0.232	0.024	0.134	0.136	0.023	0.134	0.136	
Sensitivity		0.940			0.994			1.000			1		
Specificity		0.952			0.971			0.999			1		
Accuracy		1*			1*			1			1		
RI		0.862			0.902			0.999			1		
(n_1, n_2, n_3)		Rig-ref			FMCox			DCAF			Oracle		
= (15, 7, 8)													
30%	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
α_1	0.774	0.459	0.900	0.348	0.365	0.504	0.050	0.165	0.172	0.046	0.161	0.168	
α_2	0.712	0.341	0.790	0.517	0.250	0.574	0.034	0.161	0.164	0.032	0.158	0.161	
α_3	0.390	0.152	0.418	0.197	0.175	0.264	0.028	0.154	0.157	0.026	0.154	0.157	
Sensitivity		0.820			0.951			0.999			1		
Specificity		0.947			0.958			0.998			1		
Accuracy		1*			1*			1			1		
RI		0.760			0.802			0.999			1		
15%	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
α_1	0.658	0.245	0.703	0.440	0.227	0.495	0.043	0.145	0.151	0.039	0.142	0.147	
α_2	0.695	0.248	0.738	0.609	0.229	0.651	0.025	0.144	0.146	0.024	0.140	0.143	
α_3	0.409	0.134	0.430	0.271	0.147	0.308	0.026	0.137	0.140	0.025	0.137	0.139	
Sensitivity		0.915			0.947			1.000			1		
Specificity		0.949			0.968			0.998			1		
Accuracy		1*			1*			1			1		
RI		0.778			0.799			1.000			1		

with $\alpha_{4,1} = -0.25$, $\alpha_{4,2} = 0.25$, $\gamma = (0, -0.5)^\top$, 30% censoring rate considered. The other assumptions are similar with that in Setting 2 except $s_{ij} = 0.3^{|i-j|}$.

Substudy 2.2 (Binary biomarkers) We considered $\mathbf{x}_{ij}$ as biomarkers with values 0 or 1, which was also widespread in clinical trials. The settings are similar with Substudy 2.1 except that the predictive effect $\mathbf{g} = (g_1, \dots, g_p)^\top = (\underbrace{\alpha_{4,1}, \dots, \alpha_{4,1}}_{p^1}, \underbrace{\alpha_{4,2}, \dots, \alpha_{4,2}}_{p^2}, \underbrace{\alpha_{4,3}, \dots, \alpha_{4,3}}_{p^3})^\top$ with $\alpha_{4,1} = -0.75$, $\alpha_{4,2} = 0$, and $\alpha_{4,3} = 0.75$. Each element of $\mathbf{x}_{ij}$ was independently generated from the binomial distribution with mean 0.5.

The estimation results of Subtype 2.1 and Subtype 2.2 are presented in Table S3 and Table S4, respectively

S4.3 Simulation Study 3: Robustness of the proposed method when there are no clusters and validity of the BIC criterion

In this simulation study, we assume $K_1 = K_2 = 1$. The survival time for the j -th subject in the i -th cluster was generated from

$$T_{ij} = \varepsilon_{ij} / \exp(\eta_i + v_i w_{ij} + \boldsymbol{\zeta}_i^\top \mathbf{x}_{ij} + \mathbf{g}^\top \mathbf{x}_{ij} w_{ij} + \boldsymbol{\gamma}^\top \mathbf{z}_{ij}), i = [n], j = [m_i],$$

where $\eta_i \equiv 0$, $v_i \equiv -0.5$, $\boldsymbol{\zeta}_i \equiv \boldsymbol{\alpha} = (2, 1, 1.5, \mathbf{0}_{p-3}^\top)^\top$, $\mathbf{g} = (\underbrace{-0.25, \dots, -0.25}_p)^\top$, and $\boldsymbol{\gamma} = (0, 0.5)^\top$. The setting of C_{ij} , ε_{ij} , $\mathbf{x}_{ij}$, and $\mathbf{z}_{ij}$ are similar with Substudy 2.1 of the Simulation Study 2. We simulated 200 data sets consisting of $n = 30$, $p = 50$, and m_i were randomly sampled from 5 to 30.

We compared our proposed DCAF (K_1 and K_2 were selected by BIC criterion) with MisDCAF($K_1 = 3$, $K_2 = 2$) and the Oracle estimator. Table S5 summarizes the Bias, SD, RMSE of $\boldsymbol{\Theta}$. We can see from Table S5 that when homogeneous components are misclassified into different subgroups, estimates and prediction are still valid at the price of estimation efficiency and prediction power loss. However, it is worth noting that correctly identifying biomarkers with heterogeneous effects and choosing appropriate number of clusters are crucial in terms of estimation and prediction accuracy, as shown in He et al. (2023).

Table S3: The simulation results of Substudy 2.1 of Simulation Study 2

Case 1 $m_i = (5, 50)$		Rig-ref			DCAF			Oracle		
$(n_1, n_2, n_3) = (10, 10, 10)$	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
$\tilde{\alpha}$	0.961	0.303	1.008	0.016	0.140	0.141	0.023	0.132	0.134	
α_4	1.053	0.125	1.060	0.003	0.014	0.014	0.004	0.013	0.013	
γ	0.372	0.080	0.381	0.007	0.071	0.071	0.008	0.071	0.071	
Sensitivity		–			0.995			1		
Specificity		–			0.989			1		
RI_{ζ}		0.918			0.989			1		
RI_g		0.836			0.995			1		
$(n_1, n_2, n_3) = (15, 7, 8)$	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
$\tilde{\alpha}$	0.946	0.302	0.993	0.016	0.140	0.141	0.024	0.131	0.133	
α_4	1.039	0.143	1.048	0.003	0.014	0.014	0.004	0.012	0.013	
γ	0.372	0.076	0.379	0.008	0.071	0.072	0.009	0.071	0.071	
Sensitivity		–			0.979			1		
Specificity		–			0.989			1		
RI_{ζ}		0.887			0.987			1		
RI_g		0.836			0.996			1		
Case2 $m_i = (5, 30)$		Rig-ref			DCAF			Oracle		
$(n_1, n_2, n_3) = (10, 10, 10)$	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
$\tilde{\alpha}$	0.937	0.423	1.028	0.027	0.194	0.196	0.031	0.171	0.174	
α_4	0.896	0.123	0.904	0.002	0.021	0.021	0.005	0.016	0.016	
γ	0.359	0.099	0.373	0.004	0.094	0.094	0.011	0.091	0.092	
Sensitivity		–			0.985			1		
Specificity		–			0.982			1		
RI_{ζ}		0.829			0.980			1		
RI_g		0.726			0.964			1		
$(n_1, n_2, n_3) = (15, 7, 8)$	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
$\tilde{\alpha}$	0.932	0.426	1.025	0.030	0.192	0.194	0.032	0.165	0.169	
α_4	0.868	0.118	0.876	0.003	0.022	0.022	0.005	0.016	0.016	
γ	0.358	0.089	0.369	0.006	0.093	0.094	0.013	0.089	0.090	
Sensitivity		–			0.975			1		
Specificity		–			0.981			1		
RI_{ζ}		0.782			0.972			1		
RI_g		0.717			0.965			1		

During 200 replicates, the proportion of correctly choosing $(K_1, K_2) = (1, 1)$ based on the BIC criterion defined in expression (S3.11) of the main manuscript was 100%, representing no clusters. Furthermore, we consider the Case 2 of Substudy 2.1 to show the validity of the proposed BIC criterion when $(K_1, K_2) = (3, 2)$. In 200 replicates, 190 runs generate convergence results. The correctly selected proportions of $K_1 = 3$ and $K_2 = 2$ are 94% and 89%, respectively. Clearly, as the number of subgroups increases, the proportion of correctly identifying number

S4. ADDITIONAL NUMERICAL STUDIES

of subgroups decreases, leading to reduced estimation and prediction accuracy.

We also conducted a sensitivity analysis regarding the assumption of homogeneous membership within the centre. In particular, members associated with a single center may belong to two or three subgroups. The survival time for the j -th subject in the i -th cluster is given by

$$T_{ij} = \varepsilon_{ij} / \exp(\eta_i + \boldsymbol{\beta}_i^\top \mathbf{x}_{ij}), i = [30], j = [10],$$

where $\varepsilon_{ij} \sim \text{Exponential}(1)$, and $\mathbf{x}_{ij}$ is drawn from the multivariate normal distribution with mean $\mathbf{0}_p$ and covariance matrix $(s_{ij})_{p \times p}$, where $s_{ij} = 0.3^{|i-j|}$. We set $K = 3$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n)^\top =$

$(\underbrace{\boldsymbol{\alpha}_{1,1}, \dots, \boldsymbol{\alpha}_{1,1}}_{n_1}, \underbrace{\boldsymbol{\alpha}_{1,2}, \dots, \boldsymbol{\alpha}_{1,2}}_{n_2}, \underbrace{\boldsymbol{\alpha}_{1,3}, \dots, \boldsymbol{\alpha}_{1,3}}_{n_2})^\top$. Here, $\boldsymbol{\alpha}_{1,1} = (-2, -3, 0, \mathbf{0}_{p-3}^\top)^\top$, $\boldsymbol{\alpha}_{1,2} = (0, -1, -1.5, \mathbf{0}_{p-3}^\top)^\top$, and $\boldsymbol{\alpha}_{1,3} = (2, 1, 1.5, \mathbf{0}_{p-3}^\top)^\top$. $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^\top = (\underbrace{\alpha_{2,1}, \dots, \alpha_{2,1}}_{n_1}, \underbrace{\alpha_{2,2}, \dots, \alpha_{2,2}}_{n_2}, \underbrace{\alpha_{2,3}, \dots, \alpha_{2,3}}_{n_3})^\top$, where $\alpha_{2,1} = -1$, $\alpha_{2,2} = 0$, and $\alpha_{2,3} = 1$. We consider $p = 50$, with a censoring rate of 30%, and we explore five situations: (S1) One center contains heterogeneous individual

groups, while the other centers consist of homogeneous individuals. In this center, individuals are divided into two groups with ratios of 1 : 1, 7 : 3 and 8 : 2, leading to scenarios (S1.1), (S1.2), and (S1.3), respectively. (S2) Two centers from the same cluster contain heterogeneous individual groups, while the remaining centers are homogeneous. Each of the two centers has two different groups with proportions of 8 : 2 and 9 : 1, respectively, and the minority groups belong to two different clusters. (S3) Similar to (S2), but the minority groups in two centers belong to the same cluster. (S4) The same setting as (S2), with the two centers belonging to different clusters. (S5) Similar to (S3), but with the two centers belong to different clusters.

The estimation results are presented in Table S6, where ‘1*’ indicates the true number of subgroups and the significant variables are known in advance. For comparison, we also include Oracle methods and FMCox methods; the former assumes prior knowledge of the number of subgroups and significant variables, while the latter performs subgroup analysis among centres and individuals, denoted as FMCox and FMCox-I, respectively.

From Table S6, it is evident that the proposed DCAF is robust against misclassification in scenario (S1), as its estimation accuracy—measured by RMSE, sensitivity, specificity, and RI—is closest to that of the Oracle estimators. In contrast, the FMCox and FMCox-I methods show bias, with limited performance in RI, Sensitivity and Specificity across all scenarios, the FMCox-I perform worst. The proposed DCAF estimators exhibit bias when the two misclassified groups from the two centres differ, as seen in Scenarios (S2) and (S4). However, they are unbiased when the misclassified groups from both centres are the same, as illustrated in Scenarios (S3) and (S5).

S4.4 Simulation Study 4: Confirmatory semi-simulation using the covariates from the CO.26 study

The survival time was generated from $T_{ij} = \varepsilon_{ij} / \exp(\eta_i + \zeta_i^\top \mathbf{x}_{ij} + v_i w_{ij} + \mathbf{g}^\top \mathbf{x}_{ij} w_{ij} + \boldsymbol{\gamma}^\top \mathbf{z}_{ij})$, $i = [n], j = [m_i]$, where $\varepsilon_{ij} \sim \text{Uniform}(0, 1)$, and w_{ij} , $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ were from the real data CO.26 study. w_{ij} equals 1 if the patient received the new treatment and 0 otherwise. $\mathbf{x}_{ij} = (x_{ij,1}, \dots, x_{ij,14})^\top$ were the biomarkers with 0 or 1, where x_1 denotes the mutation status of the gene BRCA1, and x_2 denotes the mutation status of the gene SMAD4. $\mathbf{z}_{ij} = (z_{ij,1}, \dots, z_{ij,6})^\top$ were confounder variables, representing the age, sex, ECOG performance status (ECOG), presence of liver metastases (liver), time from initial cancer diagnosis (diag), and number of disease sites (num), respectively. In this simulation study, $n = 10$, and m_i ranges from 4 to 24. We set $K_1 = 2$, $K_2 = 3$. The centres of the $\boldsymbol{\zeta}$ are $\boldsymbol{\alpha}_{1,1} = (0, c_1, \mathbf{0}_{12}^\top)^\top$ and $\boldsymbol{\alpha}_{1,2} = (c_2, 0, \mathbf{0}_{12}^\top)^\top$. The centres of $\boldsymbol{\eta}$ are $\alpha_{2,1} = 0$, $\alpha_{2,2} = 1.34$, and the centres of the $\boldsymbol{\nu}$ are $\alpha_{3,1} = 0.87$, $\alpha_{3,2} = -0.98$. The centres of the predictive effect $\mathbf{g} = (g_1, \dots, g_{14})^\top$ are denoted as $\alpha_{4,1}$, $\alpha_{4,2}$, and $\alpha_{4,3}$. $\boldsymbol{\gamma} = (-0.02, 0, 0.95, 0.73, -0.01, -0.95)^\top$. To evaluate the performance of our proposed method in variable selection and subgroup identification, we consider four cases: (Case 5, weak signal

S4. ADDITIONAL NUMERICAL STUDIES

and moderate separation level) $c_1 = 0.73$, $c_2 = 1.23$, $\alpha_{4,1} = -1$, $\alpha_{4,2} = 0$, $\alpha_{4,3} = 1$; (Case 6, strong signal and low separation level) $c_1 = 1$, $c_2 = 1.5$, $\alpha_{4,1} = -0.75$, $\alpha_{4,2} = 0$, $\alpha_{4,3} = 0.75$; (Case 7, strong signal and moderate separation level) $c_1 = 1$, $c_2 = 1.5$, $\alpha_{4,1} = -1$, $\alpha_{4,2} = 0$, $\alpha_{4,3} = 1$; and (Case 8, strong signal and high separation level) $c_1 = 1$, $c_2 = 1.5$, $\alpha_{4,1} = -1.5$, $\alpha_{4,2} = 0$, $\alpha_{4,3} = 1.5$. The censoring time was generated from $C_{ij} \sim \text{Uniform}(0, c)$, where c was selected to obtain the desired censoring rate of 5% – 6%.

The estimation results for Case 5-8 are summarized in Table S7. In particular, from Table S7, we found that: (1) With regard to the prediction accuracy of the centres, the proposed DCAF performs better than Rig-ref in all Cases, but the proposed DCAF performs worse than Oracle method. This is not surprising because the Oracle assumes that the number and membership of subgroups and significant variables are known in advance. In addition, the gap between DCAF and Oracle decreases as the signal of the non-zero coefficients increases and separation level of the predictive subgroups increases. Particularly, the average amplification of the RMSE decreased from 9.2%(Case 5, small signal of non-zero coefficients and low separation level of the predictive subgroups) to 4.6%(Case 8, large signal of non-zero coefficients and high separation level of the predictive subgroups). (2) For variable selection, the sensitivity and the specificity increases as the signal of the non-zero coefficients increase, especially the sensitivity increased from 0.915(Case 5, small signal of non-zero coefficients) to 0.945(Case 7, large signal of non-zero coefficients), indicating that our DCAF exhibits the oracle property. (3) As the separation level between different subgroups increases, the proportion of correctly identifying the membership of subgroups increased from 0.865(Case 6, low separation level of the predictive subgroups) to 0.958(Case 8, high separation level of the predictive subgroups), leading to increased estimation and prediction accuracy.

S4.5 Simulation Study 5: Performance of proposed strategy for calculation of standard errors

To evaluate the performance of the estimated standard errors (SEs) via the data splitting method, we consider the Case 7 in Simulation Study 4. In Table S8, we report the average SEs, empirical standard deviations (SDs) and coverage probabilities of the 90%, 95%, 99% confidence intervals (Cov) for our method and Oracle estimators for some of the coefficients. From Table S8, we can see that the calculated SEs are consistent with SDs, which confirm that the proposed procedure can be used to calculate the standard errors and p -values. We may conclude that newly calculated standard errors are able to provide meaningful inference results.

S4.6 The Simulation Studies of thousands of gene/biomarkers

In this study, we consider three prognostic subgroups and four predictive subgroups, with $m = 100, n = 30$ and $p = 5000$. In particular, the survival time for the j th subject in the i th cluster was generated from $T_{ij} = \{-\log(\varepsilon_{ij})/\exp(\eta_i + \boldsymbol{\zeta}_i^\top \mathbf{x}_{ij} + v_i w_{ij} + \mathbf{g}^\top \mathbf{x}_{ij} w_{ij} + \boldsymbol{\gamma}^\top \mathbf{z}_{ij})\}^{1/4}$, $i = [n], j = [m]$, where the error $\varepsilon_{ij} \stackrel{i.i.d.}{\sim} \text{Uniform}(0, 1)$, and the treatment variable $w_{ij} \stackrel{i.i.d.}{\sim} \text{Bernoulli}(0.5)$. $\mathbf{x}_{ij}$ was generated from the multivariate normal distribution with mean $\mathbf{0}_p$ and covariance matrix $(s_{ij})_{p \times p}$, where $s_{ij} = 0.3^{|i-j|}$. $\mathbf{z}_{ij} = (z_{ij,1}, z_{ij,2})^\top$ were confounder variables with $z_{ij,1} \stackrel{i.i.d.}{\sim} \text{Bernoulli}(0.5)$ and $z_{ij,2} \stackrel{i.i.d.}{\sim} N(0, 1)$. $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)^\top =$

$$\begin{aligned} & (\underbrace{\alpha_{1,1}, \dots, \alpha_{1,1}}_{n_1}, \underbrace{\alpha_{1,2}, \dots, \alpha_{1,2}}_{n_2}, \underbrace{\alpha_{1,3}, \dots, \alpha_{1,3}}_{n_3})^\top, \boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^\top = (\underbrace{\alpha_{2,1}, \dots, \alpha_{2,1}}_{n_1}, \underbrace{\alpha_{2,2}, \dots, \alpha_{2,2}}_{n_2}, \\ & \underbrace{\alpha_{2,3}, \dots, \alpha_{2,3}}_{n_3})^\top, \boldsymbol{\nu} = (v_1, \dots, v_n)^\top = (\underbrace{\alpha_{3,1}, \dots, \alpha_{3,1}}_{n_1}, \underbrace{\alpha_{3,2}, \dots, \alpha_{3,2}}_{n_2}, \underbrace{\alpha_{3,3}, \dots, \alpha_{3,3}}_{n_3})^\top, \text{ where} \\ & \alpha_{1,1} = (-2, -3, 0, \mathbf{0}_{p-3}^\top)^\top, \alpha_{1,2} = (0, -1, -1.5, \mathbf{0}_{p-3}^\top)^\top, \alpha_{1,3} = (2, 1, 1.5, \mathbf{0}_{p-3}^\top)^\top, \alpha_{2,1} = -1, \\ & \alpha_{2,2} = 0, \text{ and } \alpha_{2,3} = 1, \alpha_{3,1} = -0.5, \alpha_{3,2} = 0, \text{ and } \alpha_{3,3} = 0.5. \text{ The predictive effect} \\ & \mathbf{g} = (g_1, \dots, g_p)^\top = (\underbrace{\alpha_{4,1}, \dots, \alpha_{4,1}}_{15}, \underbrace{\alpha_{4,2}, \dots, \alpha_{4,2}}_{15}, \underbrace{\alpha_{4,3}, \dots, \alpha_{4,3}}_{15}, \underbrace{\alpha_{4,4}, \dots, \alpha_{4,4}}_{15}, \mathbf{0}_{p-60})^\top \text{ with} \\ & (\alpha_{4,1}, \alpha_{4,2}, \alpha_{4,3}, \alpha_{4,4}) = (-2r, -r, r, 2r), r = 1, \text{ and the confounder coefficients } \boldsymbol{\gamma} = (0, -0.5)^\top. \end{aligned}$$

S4. ADDITIONAL NUMERICAL STUDIES

The censoring time was generated from $C_{ij} \sim \text{Uniform}(0, c)$, where c was selected to obtain the desired censoring rate of 20%. All the results are calculated based on 100 independent Monte Carlo repetitions.

In the screen step, we generate a screening set $\hat{\mathcal{S}}$ for prognostic and predictive biomarkers, respectively. To evaluate the quality of the screened set $\hat{\mathcal{S}}$, we report the true positive proportion (TPP) of $\hat{\mathcal{S}}$. The TPP is defined as $\text{TPP}(\hat{\mathcal{S}}) = \frac{|\hat{\mathcal{S}} \cap \mathcal{S}^0|}{|\hat{\mathcal{S}}|}$, where $\mathcal{S}^0$ represents the true support set, and $|\mathcal{S}^0|$ indicates the cardinality of $\mathcal{S}^0$. RI, Sensitivity and Specificity were used to evaluate the performance of our DCAF.

Figure S1 shows the boxplot illustrating the TPP, RI, sensitivity, and specificity for prognostic and predictive biomarkers. The proposed DCAF demonstrates strong performance in terms of high TPP, RI, sensitivity and specificity when there are thousands of biomarkers under the assumption of sparsity.

S4.7 A gene signature derived from gene similarities can potentially lead to misleading conclusions

The survival time for the j th subject in the i th center was generated from $T_{ij} = \{-\log(\varepsilon_{ij})/\exp(\mathbf{g}^\top \mathbf{x}_{ij})\}^{1/6}$, $i = [200]$, $j = [1]$, where the error $\varepsilon_{ij} \stackrel{i.i.d.}{\sim} \text{Uniform}(0, 1)$, $\mathbf{x}_{ij}$ was generated from the multivariate normal distribution with mean $\mathbf{0}_p$ and covariance matrix $\mathbf{S}$. The censoring time was generated from $C_{ij} \sim \text{Uniform}(0, c)$, where c was selected to obtain the desired censoring rate of 25%. We set $p = 60$ and $K = 4$ coefficient groups of size 15, which take value $(-2r, -r, r, 2r)$ with $r = 1$.

Three covariance matrix structures are considered.

- The (k, l) element of $\mathbf{S}$ being $0.5^{(k \neq l)}$ (Scenario 1).
- $\mathbf{S}$ is a block diagonal matrix with diagonal blocks $\mathbf{1}_{15}\mathbf{1}_{15}^\top$, $\mathbf{1}_{15}\mathbf{1}_{15}^\top$, and $\mathbf{\Sigma}_{30 \times 30}$, where the

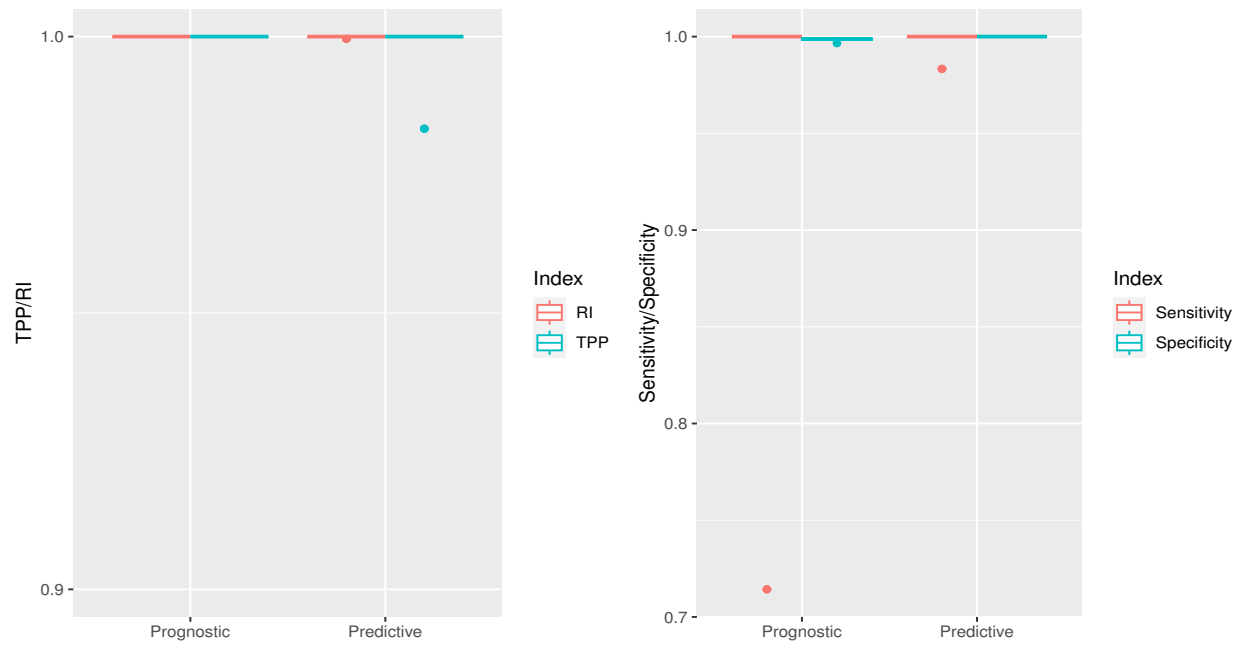


Figure S1: The simulation results including the TPP, RI, Sensitivity and Specificity for biomarkers of dimension $p = 5000$ under $n = 30, m = 100$.

(k, l) element of $\Sigma_{30 \times 30}$ being $0.5^{(k \neq l)}$ (Scenario 2).

- $\mathbf{S} = \Sigma + \mathbf{I}$, where Σ is a block diagonal matrix with diagonal blocks $\mathbf{1}_{15}\mathbf{1}_{15}^\top, \mathbf{1}_{15}\mathbf{1}_{15}^\top, \mathbf{1}_{15}\mathbf{1}_{15}^\top$, $\mathbf{1}_{15}\mathbf{1}_{15}^\top$, and $\mathbf{I}$ is a diagonal matrix with 1 diagonal (Scenario 3).

In Scenario 3, four equally groups having pairwise correlations being around 0.5 are generated. Clearly, the similarities among covariates align entirely with their impact on the survival outcome in Scenario 3. In Scenario 1, the similarities among covariates are unrelated to their effects, and Scenario 2 falls between these two extremes. Particularly, we compare four methods: KHM, K-means clustering on the solution provided by maximum partial likelihood estimator (MPL+), Unsupervised Hierarchical Clustering (Reynolds et al., 2006) based on Euclidean distance (HclustDist) or correlation distance (HclustCor).

The boxplots of RI across the three Scenarios are presented in Figure S2. Our DCAF demonstrate consistent performance across all three scenarios. In particular, DCAF shows notably higher RI compared to other methods in Scenarios 1 and 2. HclustCor and HclustDist achieve higher grouping accuracy only when the grouping structure perfectly aligns with covariate similarity. Otherwise, the grouping accuracy diminishes significantly.

S4.8 The definition of RI

Rand index (RI) (Rand, 1971) which is between 0 and 1 and takes the form $RI = (TP + TN) / (TP + FP + FN + TN)$, where TP is the number of true positive decisions which assign two observations from the same ground truth group to the same cluster, TN is the number of true negative decisions which assign two observations from different groups to different clusters, FP is the number of false positive decisions which assign two observations from different groups to the same cluster, and FN is the number of false negative decisions which assign two observations from the same group to different clusters. The Rand index is between 0 and 1.

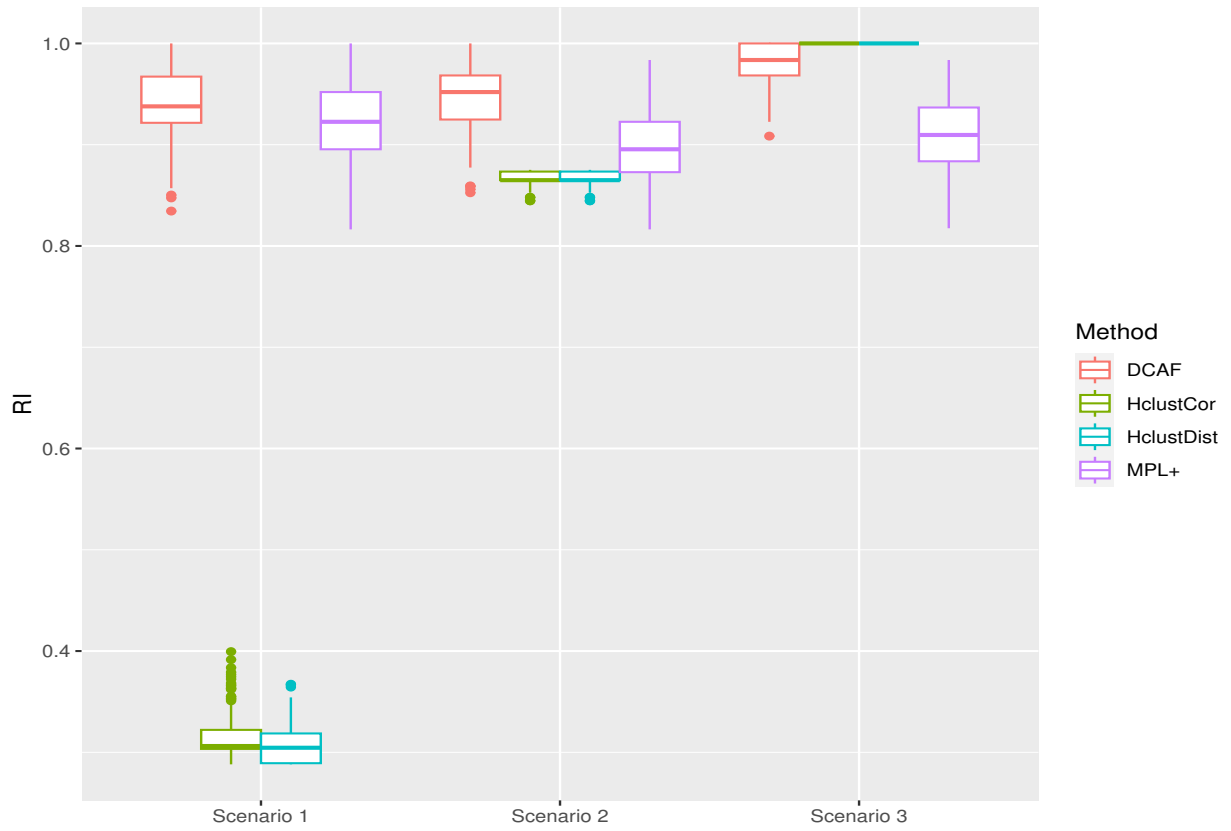


Figure S2: RI under three covariance structures for various methods including DCAF, HclustCor, HclustDist, and method MPL+.

S5 The completed results of PA.3

The prognostic subgroup identified by FMCox and DCAF are presented as follows,

- DCAF, $\mathcal{G}_1$: Argentina, Australia, Brazil, Canada, Romania, Singapore, the USA; $\mathcal{G}_2$: Belgium, Chile, Germany, the UK, Israel, New Zealand, Poland
- FMCox, $\mathcal{G}_1$: Argentina, Australia, Brazil, Canada, Romania; $\mathcal{G}_2$: Belgium, Chile, Germany, the UK, Israel, New Zealand, Poland, Singapore, the USA

The completed results of PA.3 related to the prognostic groups and confounder variables are presented in Table S9.

S6 CO.26

Combined Immune Checkpoint Inhibition vs Best Supportive Care Alone in Patients With Advanced Colorectal Cancer (CO.26) was a trial analyzed in 2018 by the Canadian Cancer Trials Group, which enrolled 180 patients from 27 cancer centres across Canada between August 2016 and June 2017. Each patient was randomly assigned to receive either 75 mg of tremelimumab every 28 days for the first 4 cycles plus 1500 mg durvalumab every 28 days, or best supportive care alone (BSC) in a 2:1 ratio. In an unpublished companion study, mutation statuses of 14 key genes were assessed from the pre-treatment blood samples of 156 patients using cell-free DNA (cfDNA) mutational analysis performed with the GuardantOMNITM assay (2.1Mb). We denote $\mathbf{x}_{ijk}$ as the mutation status of the k -th gene measured from the sample of the j -th patient enrolled from the i -th centre, which is a binary variable taking a value 1 if the gene is mutated and 0 if not. For the same patient, denote $\mathbf{z}_{ij}$ as a vector of confounding variables observed, which include age, sex(1 for male vs. 0 for female), ECOG performance status (ECOG), presence of liver metastases (liver), the duration (in months) from the initial

cancer diagnosis to the patient's enrollment in the clinical study (diag), and the number of disease sites (num). These confounding variables are balanced between two treatment groups as shown in Table S13.

We applied the DCAF procedure proposed in this paper to identify both prognostic and predictive subgroups based on the mutation statuses of the 14 genes of interest. Minimization of BIC defined in (S3.11) led to $K_1 = 2$, $K_2 = 3$ and $\lambda_1 = 1, \lambda_2 = 0.5, \lambda_3 = 0.008$, implying two major prognostic subgroups and three predictive subgroups.

The results for the two subgroups of centres identified are presented in Table S10. In the first subgroup, treatment with combined immune checkpoint inhibition had significantly worse survival in comparison with treatment with best supportive care alone (hr=2.631, p -value=0.019) and, after adjusting for the treatment effect, mutations of NRAS gene, a gene belongs to a class of genes known as oncogenes, and BRCA1 gene, a gene that produces proteins to repair damaged DNA, also yielded to significantly worse survival. Treatment with combined immune checkpoint inhibition had significantly better survival in comparison with treatment with best supportive care alone, however, in the second subgroup (hr=0.370, p -value=0.038). After adjusting for the effect of the treatment, mutation of SMAD4 gene, a gene that provides instructions for making a protein involved in transmitting chemical signals from the cell surface to the nucleus, predicts worse survival for patients in the second subgroup but patients with mutations of TP53 and PFD genes would have to significantly better survival. The prognostic effects for some of these genes were also found in other studies. For example, Oh et al. (2018) found that BRCA1 mutation was associated with increased risk of colorectal cancer; NRAS was found significantly associated with a low response rate by De Roock et al. (2010), and Wasserman et al. (2019) found that SMAD4 mutation correlated with worse clinical outcome. All confounding factors except sex were also found as significant prognostic factors. FMCox

method also identified two subgroups of centres with a only difference that the center CANW was moved from the first subgroup to the second. Treatment with combined immune checkpoint inhibition was significantly associated with the survival in, however, the second subgroup of centres ($hr=0.475$, $p\text{-value}=0.004$). In the first subgroup, combined immune checkpoint inhibition did not lead to significantly better survival compared to best supportive care alone ($hr=0.800$, $p\text{-value}=0.205$). The C-index of FMCox method is 0.656, which is also lower than that of DCAF method (0.762), suggesting that FMCox method may also have misclassified some cases in this study.

In contrast, only BRCA1 and SMAD4 were found as significant prognostic biomarkers by Cox-Lasso method but BRCA1 was no longer significant by the Homo-Lasso method. In addition, the cluster-specific fixed intercept in $\mathcal{G}_1$ is set to 1 for identification purposes, and the estimated cluster-specific fixed intercept in $\mathcal{G}_2$ is 4.617 ($p\text{-value}< 0.001$), indicating that the heterogeneity of the clinical outcomes among patients in $\mathcal{G}_2$ is significantly higher than that among patients in $\mathcal{G}_1$. To further understand the differences between two subgroups identified by DCAF, we compared the baseline characteristics of patients enrolled from these two subgroups and noted a variation in racial distribution: centres in $\mathcal{G}_1$ had a lower percentage of white patients (81% vs. 90%) and a higher percentage of Asian patients (14% vs. 6%) compared to centers in $\mathcal{G}_2$. This suggests that gene mutations may have different effects for patients across different racial and ethnic groups, warranting further investigation.

Among the three predictive subgroups identified by the proposed DCAF procedure, the results presented in Table S11 indicate a significant interaction between the treatment and $\mathcal{D}_1$ ($est=-1.067$, $p\text{-value}<0.001$) with a better treatment effect in this subgroup and also between the treatment and $\mathcal{D}_3$ ($est=0.647$, $p\text{-value}=0.001$) with, however, a worse treatment effect in this subgroup. Meanwhile, there was a weak interaction between the treatment and $\mathcal{D}_2$ ($est=-0.144$,

p -value=0.064). The interaction effect can be explained as, for example, patients with high values of KRAS/PFD (in $\mathcal{D}_1$) benefits more from the main treatment, whereas the opposite is true for biomarkers in $\mathcal{D}_3$. Alternatively, we can conclude that the treatment becomes more efficient as the expression level of KRAS/PFD increases, with the reverse effect observed for biomarkers in $\mathcal{D}_3$).

No subgroup with a significant interaction with the treatment was found, however, by the Homo-Lasso approach and Cox Lasso method. Results from the analyses by our proposed method would provide important information to clinicians for the selection of treatment to patients with advanced colorectal cancer.

S7 Baseline Characteristics of PA.3 and CO.26

The Baseline Characteristic of PA.3 and Co.26 are presented in Table S12 and Table S13.

Bibliography

- Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann Stat*, 10(4):1100–1120.
- Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.
- De Roock, W., Claes, B., Bernasconi, D., De Schutter, J., Biesmans, B., Fountzilas, G., et al. (2010). Effects of kras, braf, nras, and pik3ca mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol*, 11(8):753–762.

- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*, 96(456):1348–1360.
- Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J R Stat Soc Ser B (Stat Methodol)*, 70:849–911.
- He, Y., Zhou, L., Xia, Y., and Lin, H. (2023). Center-augmented ℓ_2 -type regularization for subgroup learning. *Biometrics*, 79(3):2157–2170.
- Oh, M., McBride, A., Yun, S., Bhattacharjee, S., Slack, M., Martin, J. R., et al. (2018). Brca1 and brca2 gene mutations and colorectal cancer risk: Systematic review and meta-analysis. *J Natl Cancer I*, 110(11):1178–1189.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc*, 66(336):846–850.
- Reynolds, A. P., Richards, G., Iglesia, B. D. L., and Rayward-Smith, V. J. (2006). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504.
- Wasserman, I., Lee, L. H., Ogino, S., Marco, M. R., Wu, C., and et al. (2019). Smad4 loss in colorectal cancer patients correlates with recurrence, loss of immune infiltrate, and chemoresistance. *Clin Cancer Res*, 25(6):1948–1956.
- Zhang, H. and Lu, W. (2007). Adaptive lasso for Cox’s proportional hazards model. *Biometrika*, 94(3):691–703.

Table S4: The simulation results of Substudy 2.2 of Simulation Study 2

		Rig-ref			DCAF			Oracle		
Case 3 $n = 30$		Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
	$\tilde{\alpha}$	0.250	0.477	0.538	0.056	0.389	0.394	0.140	0.363	0.389
	α_4	0.072	0.108	0.130	0.016	0.069	0.071	0.020	0.051	0.055
	γ	0.041	0.104	0.112	0.011	0.098	0.099	0.004	0.089	0.089
	Sensitivity		–			0.989			1	
	Specificity		–			0.985			1	
	RI_{ζ}		0.953			0.974			1	
	RI_g		0.817			0.940			1	
Case 4 $n = 60$		Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
	$\tilde{\alpha}$	0.126	0.282	0.309	0.036	0.251	0.254	0.140	0.247	0.283
	α_4	0.040	0.056	0.069	0.003	0.035	0.035	0.006	0.033	0.033
	γ	0.017	0.065	0.067	0.001	0.061	0.061	0.004	0.060	0.060
	Sensitivity		–			0.999			1	
	Specificity		–			0.995			1	
	RI_{ζ}		0.972			0.979			1	
	RI_g		0.941			0.997			1	

Table S5: The simulation results of Simulation Study 3

		Mis-DCAF			DCAF			Oracle		
		Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
	θ	0.098	0.358	0.371	0.015	0.094	0.095	0.015	0.094	0.095
	g	0.024	0.094	0.097	0.002	0.012	0.013	0.002	0.012	0.013
	γ	0.038	0.120	0.126	0.008	0.093	0.094	0.008	0.093	0.094
	Sensitivity		–			1			1	
	Specificity		–			1			1	

Table S6: The simulation results of (S1)-(S5) of Simulation Study 3

$(n_1, n_2, n_3) = (10, 10, 10)$				FMCox-I			FMCox			DCAF			Oracle	
		Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
(S1.1)	α_1	0.519	1.685	1.763	0.481	0.681	0.834	0.090	0.275	0.289	0.058	0.234	0.241	
	α_2	1.070	1.605	1.929	0.270	0.484	0.554	0.027	0.183	0.185	0.027	0.152	0.154	
	α_3	0.241	0.962	0.992	0.070	0.259	0.268	0.005	0.165	0.165	0.040	0.166	0.170	
	Sensitivity		0.878			0.968			0.986			1*		
	Specificity		0.390			0.949			0.993			1*		
	RI		0.614			0.873			0.968			1*		
	(S1.2)	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
	α_1	0.506	1.337	1.429	0.414	0.569	0.704	0.086	0.281	0.294	0.057	0.230	0.237	
(S1.3)	α_2	0.992	1.853	2.102	0.277	0.526	0.595	0.014	0.181	0.181	0.028	0.153	0.156	
	α_3	0.163	0.941	0.954	0.076	0.272	0.282	0.006	0.169	0.170	0.039	0.166	0.171	
	Sensitivity		0.882			0.968			0.988			1*		
	Specificity		0.370			0.952			0.993			1*		
	RI		0.615			0.878			0.974			1*		
	(S2)1%	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
	α_1	0.399	1.092	1.162	0.586	0.681	0.899	0.183	0.374	0.416	0.057	0.229	0.236	
	α_2	1.052	2.034	2.290	0.376	0.680	0.777	0.047	0.200	0.206	0.027	0.153	0.155	
(S3)	α_3	0.122	1.090	1.097	0.090	0.216	0.234	0.034	0.178	0.181	0.039	0.166	0.171	
	Sensitivity		0.869			0.963			0.986			1*		
	Specificity		0.365			0.940			0.993			1*		
	RI		0.616			0.865			0.970			1*		
	(S4)	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
	α_1	0.398	1.438	1.492	0.428	0.579	0.720	0.105	0.297	0.315	0.058	0.229	0.236	
	α_2	0.818	1.825	2.000	0.289	0.511	0.587	0.012	0.178	0.178	0.027	0.153	0.155	
	α_3	0.206	1.152	1.170	0.086	0.250	0.264	0.004	0.168	0.168	0.039	0.166	0.171	
(S5)	Sensitivity		0.872			0.968			0.987			1*		
	Specificity		0.366			0.950			0.993			1*		
	RI		0.616			0.872			0.973			1*		
	(S5)	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
	α_1	0.541	1.142	1.263	0.432	0.636	0.769	0.096	0.275	0.291	0.056	0.227	0.234	
	α_2	1.015	1.805	2.071	0.320	0.569	0.653	0.045	0.198	0.203	0.027	0.153	0.156	
	α_3	0.348	1.415	1.458	0.063	0.311	0.317	0.013	0.170	0.171	0.037	0.165	0.169	
	Sensitivity		0.884			0.965			0.982			1*		
(S6)	Specificity		0.382			0.931			0.994			1*		
	RI		0.615			0.870			0.975			1*		
	(S6)	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	
	α_1	0.378	1.399	1.450	0.462	0.705	0.843	0.086	0.275	0.289	0.056	0.227	0.234	
	α_2	0.882	1.885	2.082	0.337	0.472	0.580	0.034	0.197	0.200	0.028	0.154	0.156	
	α_3	0.308	1.237	1.274	0.100	0.242	0.261	0.009	0.169	0.169	0.038	0.163	0.167	
	Sensitivity		0.872			0.964			0.993			1*		
	Specificity		0.367			0.944			0.993			1*		
RI		0.616			0.862			0.968			1*			

Table S7: The results of Simulation Study 4

Rig-ref				DCAF			Oracle		
Case 5	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
$\tilde{\alpha}$	0.103	0.460	0.471	0.125	0.417	0.436	0.092	0.384	0.395
α_4	0.114	0.272	0.295	0.025	0.207	0.208	0.064	0.183	0.193
γ	0.044	0.197	0.202	0.072	0.198	0.211	0.046	0.191	0.196
Sensitivity	–			0.915			1		
Specificity	–			0.984			1		
RI_{ζ}	0.934			0.957			1		
RI_g	0.854			0.906			1		
Case 6	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
$\tilde{\alpha}$	0.073	0.456	0.462	0.109	0.416	0.430	0.102	0.388	0.401
α_4	0.081	0.235	0.248	0.020	0.199	0.200	0.053	0.179	0.186
γ	0.050	0.197	0.203	0.067	0.201	0.212	0.048	0.193	0.199
Sensitivity	–			0.950			1		
Specificity	–			0.988			1		
RI_{ζ}	0.932			0.968			1		
RI_g	0.854			0.865			1		
Case 7	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
$\tilde{\alpha}$	0.098	0.469	0.479	0.108	0.417	0.431	0.099	0.387	0.399
α_4	0.116	0.272	0.296	0.019	0.204	0.205	0.063	0.183	0.193
γ	0.045	0.196	0.201	0.066	0.198	0.209	0.046	0.191	0.197
Sensitivity	–			0.945			1		
Specificity	–			0.987			1		
RI_{ζ}	0.933			0.965			1		
RI_g	0.856			0.908			1		
Case 8	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
$\tilde{\alpha}$	0.150	0.506	0.527	0.109	0.392	0.407	0.102	0.381	0.394
α_4	0.210	0.353	0.410	0.014	0.225	0.225	0.085	0.191	0.209
γ	0.045	0.197	0.202	0.057	0.196	0.204	0.043	0.190	0.195
Sensitivity	–			0.940			1		
Specificity	–			0.987			1		
RI_{ζ}	0.933			0.966			1		
RI_g	0.856			0.958			1		

Table S8: Results on calculating standard errors for Simulation Study 5

Case 7	DCAF						Oracle	
	Bias	SD	SE	90%Cov	95%Cov	99%Cov	Bias	SD
$\alpha_{2,2}$	0.168	0.385	0.372	0.849	0.935	0.978	0.131	0.376
$\alpha_{3,1}$	0.063	0.429	0.423	0.908	0.935	0.989	0.070	0.396
$\alpha_{3,2}$	-0.102	0.447	0.422	0.849	0.919	0.989	-0.065	0.415
c_1	0.168	0.385	0.372	0.849	0.935	0.978	0.096	0.382
c_2	-0.008	0.532	0.361	0.854	0.914	1.000	0.117	0.362
$\alpha_{4,1}$	-0.022	0.273	0.276	0.919	0.962	0.995	-0.088	0.259
$\alpha_{4,2}$	-0.001	0.116	0.102	0.870	0.903	0.957	0.003	0.075
$\alpha_{4,3}$	0.026	0.191	0.175	0.876	0.935	0.978	0.064	0.167
γ_1	-0.001	0.010	0.056	1.000	1.000	1.000	-0.001	0.010
γ_2	-0.025	0.220	0.207	0.914	0.962	0.989	0.003	0.210
γ_3	0.123	0.247	0.228	0.838	0.903	0.962	0.083	0.238
γ_4	0.062	0.253	0.237	0.827	0.930	0.984	0.049	0.239
γ_5	-0.000	0.003	0.055	1.000	1.000	1.000	-0.000	0.003
γ_6	0.083	0.249	0.231	0.843	0.919	0.968	-0.058	0.247

Table S9: The estimated coefficients(est), hazard ratio(hr), standard errors(SE) and p -values for Prognostic biomarkers and Confounder Variables identified by the proposed DCAF, the FMCox, the Homo-Lasso, and the Cox-Lasso based on PA.3

Method	Prognostic Variables													
(C-index)														
DCAF (0.695)	$\mathcal{G}_1$	Intercept	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	PDK1	EGFR	AXL	PF4	MUC1		
	est	0.275	-0.332	0.215	0.162	0.176	-0.284	-0.238	-0.209	-0.177	-0.311	-0.023		
	hr	1.317	0.717	1.239	1.176	1.192	0.753	0.788	0.811	0.838	0.733	0.977		
	se	0.216	0.142	0.082	0.077	0.077	0.105	0.091	0.084	0.103	0.122	0.096		
	p-value	0.199	0.019	0.008	0.034	0.022	0.007	0.009	0.013	0.084	0.011	0.795		
	$\mathcal{G}_2$	-	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	PDK1	EGFR	AXL	PF4	MUC1		
	est	-	0.391	-0.172	-0.007	0.295	-0.216	-0.249	-0.235	-0.152	-0.146	0.052		
	hr	-	1.479	0.842	0.993	1.343	0.805	0.780	0.791	0.859	0.864	1.054		
	se	-	0.285	0.176	0.193	0.165	0.189	0.239	0.210	0.157	0.218	0.211		
	p-value	-	0.167	0.321	0.950	0.073	0.248	0.292	0.259	0.329	0.493	0.788		
FMCox (0.651)	$\mathcal{G}_1$	-	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	IL8.4B	EGFR	AXL	PF4			
	est	-	-0.198	0.151	0.186	0.230	-0.198	0.143	-0.218	-0.231	-0.079			
	hr	-	0.821	1.163	1.204	1.259	0.820	1.155	0.804	0.794	0.924			
	se	-	0.127	0.067	0.080	0.081	0.110	0.084	0.094	0.086	0.070			
	p-value	-	0.118	0.023	0.020	0.005	0.070	0.086	0.020	0.007	0.256			
	$\mathcal{G}_2$	-	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	IL8.4B	EGFR	AXL	PF4			
	est	-	-0.197	0.449	0.610	0.265	-0.192	0.484	-0.066	-0.224	0.003			
	hr	-	0.821	1.567	1.841	1.303	0.826	1.622	0.936	0.799	1.003			
	se	-	0.147	0.120	0.118	0.093	0.122	0.131	0.109	0.097	0.189			
	p-value	-	0.174	< 0.001	< 0.001	0.004	0.114	< 0.001	0.533	0.020	0.969			
Homo-Lasso (0.694)		CXCL6	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	PDK1	EGFR	AXL	IL8.4B	CPA1.6B	MUC1	MMMP.8
	est	-0.100	-0.189	0.167	0.269	0.261	-0.243	-0.180	-0.129	-0.123	0.165	-0.093	0.131	0.074
	hr	0.905	0.828	1.181	1.308	1.298	0.785	0.835	0.879	0.885	1.179	0.912	1.141	1.076
	se	0.124	0.167	0.091	0.088	0.088	0.126	0.140	0.141	0.107	0.148	0.114	0.132	0.106
	p-value	0.413	0.252	0.066	0.002	0.003	0.053	0.196	0.354	0.248	0.261	0.408	0.312	0.480
Cox-Lasso (0.702)		CXCL6	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	PDK1	EGFR	AXL	IL8.4B	CPA1.6B	MUC1	IGFB.2
	est	-0.116	-0.194	0.165	0.261	0.257	-0.198	-0.187	-0.158	-0.164	0.171	-0.094	0.114	-0.093
	hr	0.891	0.824	1.179	1.298	1.293	0.820	0.830	0.854	0.849	1.187	0.910	1.120	0.911
	se	0.138	0.175	0.092	0.088	0.088	0.148	0.141	0.128	0.110	0.161	0.112	0.124	0.140
	p-value	0.395	0.264	0.072	0.003	0.003	0.178	0.180	0.214	0.134	0.283	0.390	0.351	0.496
Confounder Variables														
DCAF					Homo-Lasso					Cox-Lasso				
variable	est	hr	se	p-value	est	hr	se	p-value		est	hr	se	p-value	
age	0.003	1.003	0.006	0.612		0.005	1.005	0.008	0.573		0.004	1.004	0.009	0.650
sex	0.219	1.244	0.125	0.078		0.142	1.152	0.160	0.370		0.154	1.166	0.162	0.336
PF	-0.613	0.542	0.168	< 0.001		-0.483	0.617	0.211	0.021		-0.506	0.603	0.210	0.016
DIS	0.487	1.627	0.153	0.001		0.475	1.607	0.192	0.013		0.516	1.676	0.198	0.009

Table S10: The estimated coefficients(est), hazard ratio(hr), standard errors(SE) and p -values for prognostic biomarkers identified by the proposed DCAF, the FMCox, the Homo-Lasso, and the Cox-Lasso based on CO.26

Method		Prognostic Variables								
(C-index)										
DCAF (0.762)	$\mathcal{G}_1$		Trt	BRCA1	SMAD4	TP53	PFD	NRAS	ATM	PIK3CA
	est		0.968	0.836	−0.115	0.195	−0.047	0.867	0.397	0.024
	hr		2.631	2.306	0.891	1.216	0.954	2.679	1.488	1.025
	SE		0.415	0.325	0.355	0.273	0.338	0.415	0.297	0.265
	p-value		0.019	0.010	0.731	0.465	0.872	0.019	0.178	0.909
	$\mathcal{G}_2$	Intercept	Trt	BRCA1	SMAD4	TP53	PFD	NRAS	ATM	PIK3CA
	est	1.530	−0.994	0.347	1.292	−0.846	−1.269	0.176	−0.084	−0.539
	hr	4.617	0.370	1.415	3.641	0.429	0.281	1.192	0.919	0.583
	SE	0.425	0.480	0.497	0.403	0.366	0.584	0.499	0.400	0.331
	p-value	< 0.001	0.038	0.475	0.001	0.020	0.029	0.710	0.817	0.101
FMCox (0.656)	$\mathcal{G}_1$		Trt	BRCA1	SMAD4	TP53	PFD	KRAS	BRAFV	
	est		−0.223	0.712	0.346	0.282	−0.205	−0.733	1.304	
	hr		0.800	2.038	1.413	1.326	0.814	0.481	3.683	
	SE		0.177	0.307	0.217	0.227	0.510	0.199	0.304	
	p-value		0.205	0.020	0.109	0.209	0.673	< 0.001	< 0.001	
	$\mathcal{G}_2$		Trt	BRCA1	SMAD4	TP53	PFD	KRAS	BRAFV	
	est		−0.745	0.368	1.716	0.074	−0.902	0.343	0.174	
	hr		0.475	1.445	5.564	1.076	0.406	1.409	1.190	
	SE		0.257	0.440	0.329	0.280	0.329	0.282	0.386	
	p-value		0.004	0.395	< 0.001	0.777	0.006	0.220	0.638	
Homo-Lasso (0.719)			Trt	BRCA1	SMAD4	TP53	PIK3CA	NRAS	ATM	
	est		−0.356	0.669	0.834	0.058	0.188	0.536	0.071	
	hr		0.700	1.953	2.302	1.060	1.207	1.710	1.074	
	SE		0.468	0.376	0.326	0.315	0.293	0.409	0.343	
	p-value		0.438	0.074	0.010	0.836	0.511	0.186	0.818	
Cox-Lasso (0.719)			Trt	BRCA1	SMAD4	TP53	PIK3CA	NRAS	ATM	
	est		−0.459	0.573	0.815	0.020	0.120	0.347	0.014	
	hr		0.632	1.773	2.259	1.021	1.127	1.414	1.015	
	SE		0.304	0.288	0.286	0.293	0.295	0.455	0.338	
	p-value		0.128	0.046	0.004	0.926	0.672	0.437	0.947	

Table S11: The estimated coefficients(est), hazard ratios(hr), standard errors(SE) and p -values for Predictive biomarkers and Confounder Variables for the benefit of combined immune checkpoint inhibition identified by the proposed DCAF, the FMCox, the Homo-Lasso, and the Cox-Lasso based on CO.26

Predictive Biomarkers												
DCAF		variable			est	hr	SE	p-value				
$\mathcal{D}_1$	KRAS,	PFD			−1.067	0.344	0.277	< 0.001				
	NRAS,	BRAFV,	ATM,									
$\mathcal{D}_2$	BRCA1,	BRCA2,	MRG,									
	TP53,	APC,	TERT		−0.144	0.866	0.078	0.064				
$\mathcal{D}_3$	Polymerase,	PIK3CA,	SMAD4		0.647	1.910	0.193	0.001				
Homo-Lasso					est	hr	SE	p-value				
$\mathcal{D}$					−0.023	0.977	0.094	0.789				
Cox-Lasso		variable			est	hr	SE	p-value				
−								−				
Confounder Variables												
DCAF					Homo-Lasso				Cox-Lasso			
variable	est	hr	SE	p-value	est	hr	SE	p-value	est	hr	SE	p-value
age	−0.023	0.977	0.010	0.028	−0.013	0.987	0.013	0.327	−0.012	0.988	0.013	0.324
sex	0.355	1.427	0.210	0.089	0.153	1.166	0.276	0.567	0.132	1.142	0.266	0.606
ECOG	1.017	2.764	0.269	< 0.001	1.010	2.745	0.342	0.003	1.027	2.793	0.342	0.003
liver	0.806	2.238	0.258	0.002	0.551	1.736	0.310	0.074	0.529	1.697	0.288	0.065
diag	−0.011	0.989	0.003	< 0.001	−0.008	0.992	0.005	0.082	−0.008	0.992	0.005	0.077
num	−0.995	0.370	0.243	< 0.001	−0.664	0.515	0.305	0.029	−0.647	0.524	0.295	0.028

MRG: Mismatch.Repair.Gene

Table S12: Baseline Characteristics of the erlotinib plus gemcitabine(Treatment) vs placebo plus gemcitabine(Control) Groups based on PA.3

Variables	Control (n=245)	Treatment (n=233)	P-Value
age, Mean(SD)	63.431(10.186)	63.824(10.334)	0.666
sex			
Male	141(57.6%)	114(48.9%)	0.072
Female	105(42.4%)	120(51.1%)	
ECOG performance status(PF)			
0,1	200(81.6%)	188(80.7%)	0.883
2	45(18.4%)	45(19.3%)	
Extent of disease(DIS)			
1,Distant metastatic	179(73.1%)	176(75.5%)	0.607
0,Locally advanced	66(26.9%)	57(24.5%)	

Table S13: Baseline Characteristics of the Combined Immune Checkpoint Inhibition(Treatment) vs Best Supportive Care Alone(Control) Groups based on CO.26

Variables	Control (n=46)	Treatment (n=110)	P-Value
age, Mean(SD)	65.326(9.252)	64.300(9.923)	0.598
Sex			
Male	34(73.9%)	70(63.6%)	0.291
Female	12(26.1%)	40(35.4%)	
ECOG performance status(ECOG)			
1	13(28.9%)	31(28.2%)	1.000
2	33(71.1%)	79(71.8%)	
Presence of liver metastases(liver)			
Yes	36(78.3%)	75(68.2%)	0.283
No	10(22.7%)	35(32.8%)	
Time (months) from initial cancer diagnosis(diag), Mean (SD)	75.200(43.097)	76.882(44.440)	0.816
Number of disease sites(num)			
>2	20(65.2%)	83(75.5%)	0.268
≤ 2	16(34.8%)	27(24.5%)	