Received: 8 July 2023 | Revised: 11 January 2024 | Accepted: 19 March 2024

DOI: 10.1002/sim.10071

RESEARCH ARTICLE

Statistics in Medicine WILEY

Information-incorporated sparse hierarchical cancer heterogeneity analysis

Wei Han[1,2] | **Sanguo Zhang**[1,2] | **Shuangge Ma**[3] | **Mingyang Ren**[4]

[1]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

[2]Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China

[3]Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut,

[4]School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China

Correspondence
Mingyang Ren, School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China.
Email: renmingyang17@mails.ucas.ac.cn

Funding information
National Science Foundation, Grant/Award Number: 2209685; National Institutes of Health, Grant/Award Number: CA204120; National Natural Science Foundation of China, Grant/Award Number: 12171454

Cancer heterogeneity analysis is essential for precision medicine. Most of the existing heterogeneity analyses only consider a single type of data and ignore the possible sparsity of important features. In cancer clinical practice, it has been suggested that two types of data, pathological imaging and omics data, are commonly collected and can produce hierarchical heterogeneous structures, in which the refined sub-subgroup structure determined by omics features can be nested in the rough subgroup structure determined by the imaging features. Moreover, sparsity pursuit has extraordinary significance and is more challenging for heterogeneity analysis, because the important features may not be the same in different subgroups, which is ignored by the existing heterogeneity analyses. Fortunately, rich information from previous literature (for example, those deposited in PubMed) can be used to assist feature selection in the present study. Advancing from the existing analyses, in this study, we propose a novel sparse hierarchical heterogeneity analysis framework, which can integrate two types of features and incorporate prior knowledge to improve feature selection. The proposed approach has satisfactory statistical properties and competitive numerical performance. A TCGA real data analysis demonstrates the practical value of our approach in analyzing data heterogeneity and sparsity.

KEYWORDS
data integration, feature selection, heterogeneity analysis, hierarchy, prior information

1 | INTRODUCTION

Cancer is naturally heterogeneous,[1] and its heterogeneity analysis is essential for a deeper biological understanding of etiology and personalized medical decisions but challenging in statistical and biomedical research.[2,3] There are multiple levels of defining cancer heterogeneity, such as differences in cells of the same tumor and differences in tumors of the same subject. We focus on the differences across subjects in this study. Depending on the different presence of meaningful phenotypes, heterogeneity analysis can be roughly classified as supervised, semi-supervised, and unsupervised. We conduct supervised heterogeneity analysis in this article and note that different analysis schemes serve different purposes, and no one can dominate others.

In the early study of supervised heterogeneous data analysis, the finite mixture of regression (FMR) is widely considered.[4,5] The downside is that, like many classical approaches, FMR requires a pre-setting number of subgroups. Recent methods via penalized fusion have achieved data-dependently identification of the heterogeneous structure and in principle accommodation of subgroups as small as size one.[6-8] Due to its intuitive and explainable penalized

function, satisfactory statistical properties, and competitive numerical performance, it has shown great advantages in longitudinal heterogeneous data analysis,[9] robust heterogeneity analysis,[10] survival heterogeneity analysis,[11] functional heterogeneous data analysis,[12] and so on.

However, in most heterogeneity analyses, only a single type of data is considered. In cancer clinical practice, heterogeneity is commonly determined based on pathological images obtained from biopsy and visually examined by pathologists, which has been highly successful, and pathological images have been effective for staging and diagnosis.[13] Nonetheless, it has also been recognized that only a rough subgrouping can be led by pathological imaging data.[14] On the other hand, with the development of biotechnology and the storage of large-scale data, a large amount of high-throughput omics data is collected, which provides great convenience for the study of complex diseases, and also encourages researchers to obtain potential biological explanations by modeling in molecular biology.[15] In the literature, heterogeneity analysis on gene expression data has been successfully studied, and omics data-based analysis does not void analysis based on pathological imaging data and can obtain a refined heterogeneous structure nested in the known rough structure.[16]

In addition, most heterogeneity analyses mainly focus on the heterogeneity of important features without considering sparsity. The selection of important features has extraordinary significance and is more challenging for heterogeneity analysis, which can be divided into two levels, that is, only a few features are important for heterogeneity analysis, and the important variables may not be the same in different subgroups. For the selection of important gene features, a powerful and cost-effective type of auxiliary way is to mine large publication databases such as PubMed and extract useful information about important gene features associated with the cancer of interest from published studies. Several text mining tools, such as PubMatrix and VxInsight, are available, in addition, some direct text mining procedures are also feasible for this purpose. Specifically, in our analysis, we search a pair of keywords of the disease name and the candidate gene name in PubMed Advanced Search Builder, then a frequency matrix of term co-occurrence is obtained, indicating whether an association exists and its amount of evidence. It is noted that the search procedure in PubMed is easy and convenient to conduct in several minutes since only the co-occurrence frequency is demanded. For example, when we conduct a pair of keywords search in PubMed, EGFR and lung adenocarcinoma get more than 7000 hits, in comparison, AKT1 and lung adenocarcinoma get 160 hits, RXRB and lung adenocarcinoma get zero hit. Although not all publications are included in PubMed and information by such a crude and simple search may not be fully convincing, it is indisputable that EGFR is of far greater relevance than AKT1/RXRB in terms of genes associated with lung adenocarcinoma. We construct prior information set S^{p}, aiming to include important genes according to the occurrence frequency. We preset a threshold for example 300, then EGFR is included in S^{p}, while AKT1 and RXRB are excluded. In lung adenocarcinoma data analysis, we set four thresholds and obtain S^{p} under four scenarios, summarized in Table S7 (Supporting Information). By mining the previous publication databases, a large amount of such valuable and available information can be collected in a cost-effective way. To balance efficiency and safety, we only incorporate qualitative information instead of quantitative results by only adopting the suggested associations in penalized terms rather than their estimated effect sizes, which shares similar spirits with Prior Lasso in Jiang et al.[17] In existing studies, such a technique is popular and applicable in omics data analysis. Wang et al[18] proposed incorporating prior information for gene-environment interactions identification and selection. Yi et al[19] performed an information-incorporated GGM-based model to assist in estimating the network structure in gene expression data. To the best of our knowledge, incorporating prior information in a sparse hierarchical heterogeneity analysis framework has not been studied in the literature.

In this article, we consider integrating two types of features, pathological imaging and omics data, to conduct cancer hierarchical heterogeneity analysis. To improve feature selection, we would like to borrow additional information extracted from existing publications to reinforce sparse penalization. This study can contribute beyond the existing literature in the following ways. First, an innovative information-incorporated sparse hierarchical heterogeneity analysis framework is developed, which can combine the two different types of data to obtain a sensible hierarchical heterogeneous structure and utilize the prior information to better perform model selection. Second, the much-desired statistical consistency properties of the proposed approach are rigorously established, including accurate estimation of subgroup numbers, recovery of the unknown heterogeneous structures, and feature selection consistency. Theoretically and numerically, we show that the incorporation of trivial but useful information can significantly improve the accuracy of model selection as well as subject subgrouping and has robust performance even if incorrect information is involved. Last but not least, the application on lung adenocarcinoma serves as an example of the proposed approach and may provide a more effective way to extract useful information from The Cancer Genome Atlas (TCGA) and other cancer databases.

2 | METHODS

2.1 | Data and model setting

Consider n independently subjects with measurements $\{\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i\}_{i=1}^n$, where $\boldsymbol{X}_i = \left(X_{i1}, \ldots, X_{iq}\right)^{\mathrm{T}}$ are q-dimensional features, $\boldsymbol{Z}_i = \left(Z_{i1}, \ldots, Z_{ip}\right)^{\mathrm{T}}$ are p-dimensional features, and Y_i is the continuous response for the ith subject. As illustrated in Section 1, the first type of features $\boldsymbol{X}$ can be clinical data and pathological image data, which have low dimension and are all important for rough subgrouping, while only a small amount of components of gene features $\{\boldsymbol{Z}_i\}_{i=1}^n$ are important, leading to refined subgrouping. In our model, superior to previous studies, the hierarchical structure and data sparsity are taken into consideration. The rough heterogeneous structure with K_1 subgroups is denoted by $\left\{\mathcal{G}_1, \ldots, \mathcal{G}_{K_1}\right\}$ which is a mutually exclusive partition of $\{1, \ldots, n\}$, while a refined heterogeneous structure with K_2 sub-subgroups is denoted by $\left\{\mathcal{T}_1, \ldots, \mathcal{T}_{K_2}\right\}$. Moreover, hierarchy means that refined sub-subgroups are nested in rough subgroups, that is, subject indices in the same $\mathcal{T}$ must be in the same $\mathcal{G}$. Hence, there exists a partition of $\{1, \ldots, K_2\}$ denoted by $\left\{\mathcal{H}_1, \ldots, \mathcal{H}_{K_1}\right\}$ satisfying that $\mathcal{G}_{k_1} = \cup_{k_2 \in \mathcal{H}_{k_1}} \mathcal{T}_{k_2}$ for $k_1 = 1, \ldots, K_1$. We consider the hierarchical heterogeneity regression model,

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\xi}_{k_1} + \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\alpha}_{k_2} + \epsilon_i, \quad i \in \mathcal{T}_{k_2} \subset \mathcal{G}_{k_1}, k_2 \in \mathcal{H}_{k_1}, k_1 = 1, \ldots, K_1, \tag{1}$$

where $\boldsymbol{\xi}_{k_1} = \left(\xi_{k_1 1}, \ldots, \xi_{k_1 q}\right)^{\mathrm{T}}$ is q-dimensional coefficient in $\mathcal{G}_{k_1}$, $\boldsymbol{\alpha}_{k_2} = \left(\alpha_{k_2 1}, \ldots, \alpha_{k_2 p}\right)^{\mathrm{T}}$ is p-dimensional coefficient in $\mathcal{T}_{k_2}$, and ϵ_i is the random error with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.

In regression-based heterogeneity analysis, each subject has its own regression coefficients, and two subjects belong to the same subgroup if and only if they have the same regression coefficients.[6,7] For each ith subject, we model its own regression coefficients $\boldsymbol{\beta}_i = \left(\beta_{i1}, \ldots, \beta_{iq}\right)^{\mathrm{T}}$ and $\boldsymbol{\gamma}_i = \left(\gamma_{i1}, \ldots, \gamma_{ip}\right)^{\mathrm{T}}$. Hence, in (1), $\boldsymbol{\xi}_{k_1}$ equals the common value of $\boldsymbol{\beta}_i$ for $i \in \mathcal{G}_{k_1}$, and $\boldsymbol{\alpha}_{k_2}$ equals the common value of $\boldsymbol{\gamma}_i$ for $i \in \mathcal{T}_{k_2}$. We remark that the rough heterogeneous structure $\left\{\mathcal{G}_1, \ldots, \mathcal{G}_{K_1}\right\}$ is a known prior. By the known rough heterogeneous structure, define $\mathcal{M}_{\mathcal{G}} = \left\{\boldsymbol{\beta} \in \mathbb{R}^{nq} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_l, i, l \in \mathcal{G}_{k_1}, k_1 = 1, \ldots, K_1\right\}$. Our analysis goal is to determine K_2 and $\left\{\mathcal{T}_1, \ldots, \mathcal{T}_{K_2}\right\}$, and estimate coefficients $\boldsymbol{\xi}_{k_1}$ and $\boldsymbol{\alpha}_{k_2}$, ensuring data hierarchy and sparsity at the same time.

2.2 | Information-incorporated penalized estimation

We assume that the prior information S^{p}, the index set of the second type of important candidate features, is available. A typical and cost-effective way to obtain S^{p} is extracted from previous studies, explicitly illustrated in Section 1. To select important gene features, we additionally employ sparse penalties on coefficients $\boldsymbol{\gamma}$ via prior information S^{p}, and not impose them on $\boldsymbol{\beta}$ since the first type of features have low dimension and intuitively biological meanings. For simultaneous coefficients estimation, feature selection, and detection of the refined heterogeneous structure, we propose a multi-step **I**nformation-incorporated **S**parse **H**ierarchical **H**eterogeneous approach (ISHH).

Step 1: Information-guided step. Consider the following penalized objective function,

$$\begin{aligned} Q_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) = & \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}_i - \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\gamma}_i\right)^2 \\ & + \sum_{i=1}^{n} \sum_{j \notin S^{\mathrm{p}}} p\left(|\gamma_{ij}|; \lambda_1\right) + \sum_{k_1=1}^{K_1} \sum_{i<l, i, l \in \mathcal{G}_{k_1}} p\left(\left\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_l\right\|_2; \lambda_2\right), \end{aligned} \tag{2}$$

where $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_n^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathcal{M}_{\mathcal{G}}$, $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_n^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^{np}$, and $p(\cdot; \lambda)$ is the concave penalty with tuning parameter λ. In our implementation, we adopt the minimax concave penalty (MCP).[20] It is noted that the smoothly clipped absolute deviation penalty (SCAD)[21] and some alternatives are equally applicable. In Step 1, the prior information is fully trusted, and sparse penalties are only imposed on the coefficients which are not recommended. We obtain prior estimators $\left(\widehat{\boldsymbol{\beta}}^{\mathrm{p}}, \widehat{\boldsymbol{\gamma}}^{\mathrm{p}}\right)$ by minimizing $Q_1(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Then we compute $\widehat{Y}_i^{\mathrm{p}} = \boldsymbol{X}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_i^{\mathrm{p}} + \boldsymbol{Z}_i^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_i^{\mathrm{p}}$ for $i = 1, \ldots, n$, which are artificial responses totally guided by prior information.

Step 2: Information-incorporated step. Define $\widetilde{Y}_i = (1 - \tau) Y_i + \tau \widehat{Y}_i^{\mathrm{p}}$ as the weighted response, where $\tau \in [0, 1)$ is a tuning parameter to balance the observed data and the artificial data. Consider the following penalized objective function,

$$Q_2(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{n}\left(\widetilde{Y}_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_i\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{p} p\left(|\gamma_{ij}|;\lambda_1\right) + \sum_{k_1=1}^{K_1}\sum_{i<l,i,l\in\mathcal{G}_{k_1}} p\left(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_l\|_2;\lambda_2\right), \tag{3}$$

where notations have the same implications as above. We obtain final estimators $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})$ by minimizing $Q_2(\boldsymbol{\beta},\boldsymbol{\gamma})$. Denote $\widehat{\boldsymbol{\xi}} = \left(\widehat{\boldsymbol{\xi}}_1^{\mathrm{T}},\ldots,\widehat{\boldsymbol{\xi}}_{K_1}^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}},\ldots,\widehat{\boldsymbol{\alpha}}_{\widehat{K}_2}^{\mathrm{T}}\right)^{\mathrm{T}}$ as the distinct values of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, respectively. Then the number of sub-subgroups $\widehat{K}_2$ and the refined heterogeneous structure $\left\{\widehat{\mathcal{T}}_1,\ldots,\widehat{\mathcal{T}}_{\widehat{K}_2}\right\}$ are determined by checking the distinct values of $\widehat{\boldsymbol{\gamma}}$, where $\left\{\widehat{\mathcal{T}}_1,\ldots,\widehat{\mathcal{T}}_{\widehat{K}_2}\right\}$ constitutes mutually exclusive partitions of $\{1,\ldots,n\}$ with $\widehat{\mathcal{T}}_{k_2} = \left\{i : \widehat{\boldsymbol{\gamma}}_i = \widehat{\boldsymbol{\alpha}}_{k_2}, i = 1,\ldots,n\right\}$ for $k_2 = 1,\ldots,\widehat{K}_2$.

Interestingly, if introducing a parameter $\kappa = \frac{\tau}{1-\tau} \in [0,\infty)$, by adding the terms independent of $(\boldsymbol{\beta},\boldsymbol{\gamma})$ and scaling the objective function, minimizing (3) is equivalent to minimizing

$$\widetilde{Q}_2(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_i\right)^2 + \frac{\kappa}{2}\sum_{i=1}^{n}\left(\widehat{Y}_i^{\mathrm{p}} - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_i\right)^2 + (1+\kappa)\sum_{i=1}^{n}\sum_{j=1}^{p} p\left(|\gamma_{ij}|;\lambda_1\right) + (1+\kappa)\sum_{k_1=1}^{K_1}\sum_{i<l,i,l\in\mathcal{G}_{k_1}} p\left(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_l\|_2;\lambda_2\right).$$

Clearly, an extra loss term is caused by the information-guided step in this form of representation. In Step 2, the quadratic loss has two components, one is from the observed data and the other is from the artificial data by information-guided step. The balance parameter τ can be selected via some appropriate criteria, such as BIC, in practice. With trustworthy prior information of high quality, it tends to choose a larger τ to obtain more accurate estimators, which means that the credible prior information is effectively incorporated. When the quality of prior information is poor, a smaller τ is adaptively chosen based on the observed data, and (3) tends to a standard sparse and fusion-penalized objective function without prior. To some extent, it can compensate for the bias caused by incorrect prior information. In conclusion, τ is introduced to data-dependently balance two types of quadratic losses in a sense, making our proposed approach flexible and stable.

Remark 1. In our multi-step approach, we formulate pairwise fusion penalty with respect to $\boldsymbol{\gamma}$ to estimate refined heterogeneous structure nested in known prior rough heterogeneous structure. It makes sense since rough heterogeneity is usually suggested based on clinical features and visually examined by experts. When the interested outcome is completely new, and there is no prior subgroups knowledges, the objective function requires double pairwise fusion penalties to simultaneously estimate rough and refined heterogeneous structure. The double pairwise fusion penalties includes two terms of $\boldsymbol{\beta}$ and $\left(\boldsymbol{\beta}^{\mathrm{T}},\boldsymbol{\gamma}^{\mathrm{T}}\right)^{\mathrm{T}}$, respectively, where data hierarchy is guaranteed by the properties of group penalties. Theoretical investigation and efficient algorithm development are expected to be thoroughly studied in our future research.

Remark 2. Although the first step is totally guided by prior information, the combination of pairwise fusion penalty is necessary. If pairwise fusion penalty is removed, it is equivalent to conduct sparse regression for each sample point individually in information-guided step, resulting in unidentifiable coefficient estimation and unreliable artificial responses $\widehat{Y}_1^{\mathrm{p}},\ldots,\widehat{Y}_n^{\mathrm{p}}$. A comparative numerical experiment is conducted. Compared to Table 2 with original formulation, simulation results in Table S5 of Supporting Information shows much lower TPR and higher MSE, which validates our discussion.

2.3 | Computation

We derive an ADMM algorithm for optimizing objective functions in our proposed approach. Define a $n \times K_1$ matrix $\widetilde{\boldsymbol{L}}_1$ with $\widetilde{l}_{ik_1} = 1$ for $i \in \mathcal{G}_{k_1}$ and $\widetilde{l}_{ik_1} = 0$ otherwise. Define $\boldsymbol{L}_1 = \widetilde{\boldsymbol{L}}_1 \otimes \boldsymbol{I}_q$, where $\boldsymbol{I}_q$ is a $q \times q$ identity matrix and $\otimes$ is the

Kronecker product. For $\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}$ and $\boldsymbol{\xi} = \left(\boldsymbol{\xi}_1^{\mathrm{T}}, \ldots, \boldsymbol{\xi}_{K_1}^{\mathrm{T}}\right)^{\mathrm{T}}$, $\boldsymbol{\beta} = \boldsymbol{L}_1\boldsymbol{\xi}$. In a general manner, we derive an ADMM algorithm for minimizing the following type of function, which is the unified form of our objective functions in two steps.

$$\frac{1}{2}\|\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\xi} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \sum_{i=1}^{n}\sum_{j=1}^{p} p\left(|\gamma_{ij}|; \lambda_1\right) + \sum_{k_1=1}^{K_1}\sum_{i<l, i,l\in\mathcal{G}_{k_1}} p\left(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_l\|_2; \lambda_2\right), \tag{4}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\mathbf{X} = \mathrm{diag}\left(\boldsymbol{X}_1^{\mathrm{T}}, \ldots, \boldsymbol{X}_n^{\mathrm{T}}\right)$, $\mathbf{Z} = \mathrm{diag}\left(\boldsymbol{Z}_1^{\mathrm{T}}, \ldots, \boldsymbol{Z}_n^{\mathrm{T}}\right)$, and $\widetilde{\mathbf{X}} = \mathbf{X}\boldsymbol{L}_1$. Define the counts of difference within subgroups by $N \triangleq \frac{1}{2}\sum_{k_1=1}^{K_1}\left|\mathcal{G}_{k_1}\right|\left(\left|\mathcal{G}_{k_1}\right| - 1\right)$, where $|\cdot|$ is the cardinality of the set. By introducing two new parameters $\boldsymbol{\eta} = \left\{\boldsymbol{\eta}_{il}^{\mathrm{T}}, i < l, i, l \in \mathcal{G}_{k_1}, k_1 = 1, \ldots, K_1\right\}^{\mathrm{T}} \in \mathbb{R}^{Np}$ and $\boldsymbol{\delta} = \left(\boldsymbol{\delta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\delta}_n^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^{np}$ to replace $\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_l$ and γ_{ij} in penalty terms, respectively, minimizing (4) is equivalent to the following constrained minimization problem,

$$\begin{aligned}
\mathcal{L}_0(\boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}) = \frac{1}{2}\|\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\xi} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \sum_{i=1}^{n}\sum_{j=1}^{p} p\left(|\delta_{ij}|; \lambda_1\right) + \sum_{k_1=1}^{K_1}\sum_{i<l, i,l\in\mathcal{G}_{k_1}} p\left(\|\boldsymbol{\eta}_{il}\|_2; \lambda_2\right),\\
\text{s.t.} \quad \gamma_{ij} - \delta_{ij} = 0, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p.\\
\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_l - \boldsymbol{\eta}_{il} = 0, \quad i < l, \quad i, l \in \mathcal{G}_{k_1}, \quad k_1 = 1, \ldots, K_1.
\end{aligned}$$

Then the augmented Lagrangian function is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{u}) = \mathcal{L}_0(\boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}) + \sum_{i=1}^{n}\boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\gamma}_i - \boldsymbol{\delta}_i\right) + \frac{\vartheta}{2}\sum_{i=1}^{n}\|\boldsymbol{\gamma}_i - \boldsymbol{\delta}_i\|_2^2\\
+ \sum_{k_1=1}^{K_1}\sum_{i<l, i,l\in\mathcal{G}_{k_1}} \boldsymbol{u}_{il}^{\mathrm{T}}\left(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_l - \boldsymbol{\eta}_{il}\right) + \frac{\vartheta}{2}\sum_{k_1=1}^{K_1}\sum_{i<l, i,l\in\mathcal{G}_{k_1}} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_l - \boldsymbol{\eta}_{il}\|_2^2,
\end{aligned}$$

where $\boldsymbol{v} = \left(\boldsymbol{v}_1^{\mathrm{T}}, \ldots, \boldsymbol{v}_n^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^{np}$ and $\boldsymbol{u} = \left\{\boldsymbol{u}_{il}^{\mathrm{T}}, i < l, i, l \in \mathcal{G}_{k_1}, k_1 = 1, \ldots, K_1\right\}^{\mathrm{T}} \in \mathbb{R}^{Np}$ are Lagrange multipliers, and ϑ is a fixed ADMM algorithm penalty parameter. Then the standard ADMM optimization procedures[22] can be applied to find the local minimizer of $\mathcal{L}(\boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{u})$. The computational complexity of the algorithm is $O\left(\sum_{k_1=1}^{K_1}\left|\mathcal{G}_{k_1}\right|^2 p\right)$, which is similar as the computational complexity $O\left(n^2 p\right)$ in the existing literature.[6,7,11,23] The details are available in Supporting Information.

Comparing with the representation of (4), our two-step approach includes sum of sparse penalties with respect to $j \in \{1, \ldots, p\} \setminus S^{\mathrm{p}}$ rather than $j \in \{1, \ldots, p\}$ in Step 1, and includes the weighted response $\widetilde{\boldsymbol{Y}}$ rather than $\boldsymbol{Y}$ in Step 2. Similar to the optimization procedure of (4), the objective functions in our approach can be effectively optimized with a few modifications. Details of iteration formulas in Steps 1 and 2 are available in Supporting Information. These steps are repeated until convergence. Convergence of the algorithm follows from Ma and Huang[6] and is achieved in all of our numerical analyses.

For initial values, we first capture an initial refined heterogeneous structure following the initial value generation method in Ma et al,[7] and then apply Lasso regression to obtain sparse initial values of coefficients in each refined subgroup. Let $C_n = \log(n(q+p))$, we choose the tuning parameters by minimizing a modified BIC criterion following Ma and Huang,[6]

$$\begin{aligned}
\mathrm{BIC}(\tau, \lambda_1, \lambda_2) = \log\left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_i(\tau, \lambda_1, \lambda_2) - \boldsymbol{Z}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_i(\tau, \lambda_1, \lambda_2)\right)^2\right\}\\
+ C_n\frac{\log n}{n}\{K_1 q + s\},
\end{aligned}$$

where s is the number of nonzero coefficients in $\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\alpha}}_{\widehat{K}_2}^{\mathrm{T}}\right)^{\mathrm{T}}$. We suggest grid search of λ_1 and λ_2 while holding $\tau = 0$, and then conduct line search of τ while holding the optimal λ_1 and λ_2.

3 | STATISTICAL PROPERTIES

By the true refined heterogeneous structure $\{\mathcal{T}_1, \ldots, \mathcal{T}_{K_2}\}$, define $T_{\max} = \max_{1\leqslant k_2\leqslant K_2} |\mathcal{T}_{k_2}|$ and $T_{\min} = \min_{1\leqslant k_2\leqslant K_2} |\mathcal{T}_{k_2}|$, where $|\cdot|$ is the cardinality of the set. The true values of coefficients are denoted by $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}_1^{*\mathrm{T}}, \ldots, \boldsymbol{\beta}_n^{*\mathrm{T}}\right)^{\mathrm{T}}$ and $\boldsymbol{\gamma}^* = \left(\boldsymbol{\gamma}_1^{*\mathrm{T}}, \ldots, \boldsymbol{\gamma}_n^{*\mathrm{T}}\right)^{\mathrm{T}}$, by which, we define $d_n = \min_{i\in\mathcal{T}_{k_2}, l\in\mathcal{T}_{k_2'}, 1\leqslant k_2\neq k_2'\leqslant K_2} \left\|\boldsymbol{\gamma}_i^* - \boldsymbol{\gamma}_l^*\right\|_2$ as the minimal differences of the common values between two refined subgroups. Define $\mathcal{A}_i = \left\{j : \gamma_{ij}^* \neq 0\right\}$ for $i = 1, \ldots, n$, and $b_n = \min_{j\in\mathcal{A}_i, 1\leqslant i\leqslant n} \left|\gamma_{ij}^*\right|$. We assume several mild and sensible conditions described as follows, and then establish theoretical results when the number of parameters tends to infinity as the sample size increases.

Condition 1. For some constant $M_0 > 0$ and any $i = 1, \ldots, n$, the elements of $\boldsymbol{W}_i = \left(\boldsymbol{X}_i^{\mathrm{T}}, \boldsymbol{Z}_i^{\mathrm{T}}\right)^{\mathrm{T}}$ are bounded by $[-M_0, M_0]$. For some constants $C_1, C_2 > 0$ and any $k_2 = 1, \ldots, K_2$, the smallest eigenvalue of positive definite matrix $\frac{1}{|\mathcal{T}_{k_2}|}\sum_{i\in\mathcal{T}_{k_2}} \boldsymbol{W}_i\boldsymbol{W}_i^{\mathrm{T}}$ is larger than C_1, and $\frac{1}{|\mathcal{T}_{k_2}|}\sum_{i\in\mathcal{T}_{k_2}} \left\|\boldsymbol{W}_i\boldsymbol{W}_i^{\mathrm{T}}\right\|_2 \leqslant C_2$.

Condition 2. Random error $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$ follows sub-Gaussian distribution, that is, for any $\boldsymbol{a} \in \mathbb{R}^n$ and $x > 0$, $\Pr\left(\left|\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\epsilon}\right| > \|\boldsymbol{a}\|_2 x\right) \leqslant 2\exp(-cx^2)$, where c is a positive constant.

Condition 3. The penalty $p(t;\lambda)$ is non-decreasing and concave on $[0,\infty)$. There exists a constant $a > 0$ such that $p(t;\lambda)$ is a constant for all $t \geqslant a\lambda$, and $p(0;\lambda) = 0$. The derivative $p'(t;\lambda)$ is continuous, bounded by λ and satisfies $\lim_{t\to 0^+} p'(t;\lambda) = \lambda$.

Condition 4. (i) $T_{\max} = O(T_{\min})$. (ii) $T_{\min} \gg \max\left\{K_2, (q+p)\log(T_{\min})\right\}$.

Condition 5. The prior estimators satisfy $\sup_{1\leqslant i\leqslant n}\left\|\widehat{\boldsymbol{\beta}}_i^{\mathrm{p}} - \boldsymbol{\beta}_i^*\right\|_2 + \sup_{1\leqslant i\leqslant n}\left\|\widehat{\boldsymbol{\gamma}}_i^{\mathrm{p}} - \boldsymbol{\gamma}_i^*\right\|_2 \leqslant C_0\tau^{-1}(1-\tau)(q+p)^{\frac{1}{2}}T_{\min}^{-\frac{1}{2}}\left(\log(T_{\min})\right)^{\frac{1}{2}}$ with probability tending to 1 as $n\to\infty$, where C_0 is a positive constant.

Condition 1 is widely used in statistical theoretical analysis. The sub-Gaussian assumption in Condition 2 is common in regression-based heterogeneity analysis.[7] SCAD, MCP, and several other concave penalties satisfy Condition 3.[20,21] Condition 4 makes some assumptions on the true refined heterogeneous structure. Specifically, Condition 4(i) imposes the balance condition on the samples sizes between sub-subgroups, and Condition 4(ii) requires that the subgroup size is far greater than the dimension of features, which still allows p to tend to infinity. Condition 5 concerns with prior estimators obtained in the information-guided step and a similar condition has been assumed in Jiang et al.[17] It suggests that information with higher quality leads to more reliable prior estimators corresponding to a larger τ, and Condition 5 is satisfied when $\tau = 0$. Theoretical results are established as follows and the rigorous proof is available in Supporting Information.

Theorem 1. *Define* $\phi_n = (1-\tau)(q+p)^{\frac{1}{2}}T_{\min}^{-\frac{1}{2}}\left(\log(T_{\min})\right)^{\frac{1}{2}}$. *Suppose that* $\lambda_1 < a^{-1}b_n$, $\lambda_2 < \left(a+\frac{1}{2}\right)^{-1}d_n$, $\lambda_2 > T_{\min}^{-\frac{1}{2}}\lambda_1$, *and* $\min\{\lambda_1,\lambda_2\} \gg \phi_n$. *Under Condition 1-5, as* $n\to\infty$, *there exists a local minimizer* $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ *of* $Q_2(\boldsymbol{\beta},\boldsymbol{\gamma})$ *such that*

(1) *(Estimation consistency)* $\sup_{1\leqslant i\leqslant n}\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^*\right\|_2 + \sup_{1\leqslant i\leqslant n}\left\|\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\right\|_2 = O_p(\phi_n)$.
(2) *(Variable selection consistency)* $\Pr\left(\widehat{\gamma}_{ij} = 0, j\notin\mathcal{A}_i, i = 1, \ldots, n\right) \to 1$.
(3) *(Subgroup number consistency)* $\Pr\left(\widehat{K}_2 = K_2\right) \to 1$.
(4) *(Heterogeneous structure consistency)* $\Pr\left(\widehat{\mathcal{T}}_{k_2} = \mathcal{T}_{k_2}, k_2 = 1, \ldots, K_2\right) \to 1$.

In the literature, quite a lot of published studies about heterogeneity analysis via fusion penalties consider a simpler scenario without sparsity or hierarchical heterogeneous structures,[6,11,12,23,24] they only establish theoretical results on coefficients estimation consistency and heterogeneous structure consistency. To the best of our knowledge, the variable selection consistency and sparsity properties on fusion-based heterogeneity models have not been investigated in the literature. More importantly, we further theoretically find out how prior information assists feature selection and coefficients estimation. Comparing with scenarios without prior information (also included in our analysis with $\tau = 0$), our theoretical advancement concludes that the estimation convergence rate is shrunk with a factor $1-\tau$, which leads to a huge reduction when accurate information is incorporated and τ is close to 1. In a word, our theoretical development is

considerably more challenging since it simultaneously demands new investigation on prior information, heterogeneous structures, and sparsity.

4 | SIMULATION

Consider n independent subjects with $(q+p)$-dimensional covariates. In our simulation, let $n = 120$, $q = 3$, and $p = 12$. We generate $\boldsymbol{X}_i$ independently and identically distributed from q-dimensional multivariate normal distribution $N(\mathbf{0}_q, \boldsymbol{\Sigma}_1)$, where $\boldsymbol{\Sigma}_1 = (\sigma_{1jm})_{1\leqslant j,m\leqslant q}$ and $\sigma_{1jm} = \mathbb{I}_{\{j=m\}} + 0.3\mathbb{I}_{\{j\neq m\}}$. We generate $\boldsymbol{Z}_i$ independently and identically distributed from $N(\mathbf{0}_p, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\Sigma}_2 = (\sigma_{2jm})_{1\leqslant j,m\leqslant p}$ has auto-regressive structure with parameter ρ or banded structure, that is, $\sigma_{2jm} = \mathbb{I}_{\{j=m\}} + \rho^{|j-m|}\mathbb{I}_{\{j\neq m\}}$ ($\rho_1 = 0.3$ denoted as AR1 and $\rho_2 = 0.7$ denoted as AR2) or $\sigma_{2jm} = \mathbb{I}_{\{j=m\}} + 0.5\mathbb{I}_{\{|j-m|=1\}}$ (denoted as Banded). Let $\boldsymbol{\mu}_x = (\mu, \mu, \mu)$ and $\boldsymbol{\mu}_z = (\mu, \mu, \mu, \mu, \mathbf{0}_{p-4})$ and consider two signal levels with $\mu_1 = 1$ and $\mu_2 = 2$. There are $K_1 = 2$ rough subgroups with coefficients $\boldsymbol{\beta}_i$ equal to $\boldsymbol{\mu}_x$ and $-\boldsymbol{\mu}_x$, and $K_2 = 4$ refined sub-subgroups with coefficients $\boldsymbol{\gamma}_i$ equal to $\frac{1}{2}\boldsymbol{\mu}_z$, $2\boldsymbol{\mu}_z$, $-\frac{1}{2}\boldsymbol{\mu}_z$ and $-2\boldsymbol{\mu}_z$. The random errors are independently and identically generated from $N(0, 0.5^2)$. Consider balanced and imbalanced settings to accommodate more general real-world situations, where two proportions of group sizes in refined structure are $P_1 = (0.25, 0.25, 0.25, 0.25)$ and $P_2 = (0.2, 0.2, 0.3, 0.3)$, and accordingly, two proportions of group sizes in rough structure are $\widetilde{P}_1 = (0.5, 0.5)$ and $\widetilde{P}_2 = (0.4, 0.6)$.

In each simulated data setting, we consider information of high/medium/low/worst-quality, respectively, that is, S^{p} can be $S_1^{\mathrm{p}} = \{1, 2, 3, 4\}$ (exactly correct information), $S_2^{\mathrm{p}} = \{1, 2, 3, 5\}$ (75% correct information and 25% wrong information), $S_3^{\mathrm{p}} = \{1, 2, 5, 6\}$ (50% correct information and 50% wrong information), and $S_4^{\mathrm{p}} = \{1, 5, 6, 7\}$ (25% correct information and 75% wrong information). In reality, we usually encounter a large amount of correct information and a small amount of wrong information, which coincides with the medium/low-quality information in our simulation settings. Although scenarios with high and worst-quality information are not common in reality, our proposed approach is still applicable by selecting the appropriate parameter τ. Our approaches are denoted as ISHH-S_1^{p}, ISHH-S_2^{p}, ISHH-S_3^{p}, ISHH-S_4^{p}. To show effectiveness of information incorporation, by setting $\tau = 0$ in Step 2, we also perform sparse hierarchical heterogeneity analysis without any prior information, denoted as "NOT incorporated."

For each setting, we generate 100 replicates. We adopt the following measures to assess performance: (a1) Mean, median, and standard deviation of $\widehat{K}_2$; (a2) Percentage of $\widehat{K}_2$ equal to K_2; (a3) Rand Index (RI) of refined heterogeneous structure consistency; (b) The true positive rate (TPR) and false positive rate (FPR) of the feature selection; (c) Mean squared errors (MSEs) of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. In practice, (a1), (a2), and (a3) are used for assessing refined subgrouping, (b) is used for assessing feature selection, (c) is used for assessing coefficients estimation. In addition, we observe the selection of τ with respect to our approaches with different quality information incorporated.

To demonstrate the competitive performance of our proposed approach, we consider some alternatives to conduct sparse hierarchical heterogeneity analyses. (1) The oracle approach (denoted as oracle) with known true refined heterogeneous structure but unknown important features. The oracle approach is regarded as a baseline that only adopts a standard sparsity penalty with enjoying true hierarchical heterogeneous structures. Then, under homogeneity assumption for each k_1th subgroup in known rough heterogeneous structure, a Lasso regression is applied to $\{Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\}_{i\in\mathcal{G}_{k_1}}$. In this manner, for fixed $k_1 = 1, \ldots, K_1$, homogeneous coefficients $\left(\widehat{\boldsymbol{\xi}}_{k_1}, \widetilde{\boldsymbol{\alpha}}_{k_1}\right)$ are obtained in each rough subgroup $\mathcal{G}_{k_1}$. Furthermore, we consider the following four alternatives: (2) The sparse finite mixture regression (denoted as sparseFMR), in which FMR is applied to $\left\{Y_i - \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\xi}}_{k_1}, \boldsymbol{Z}_i\right\}_{i\in\mathcal{G}_{k_1}}$ and Lasso is adopted for selecting relevant features. It is noted that the sparse FMR approach is widely used in heterogeneous data analysis with feature selection. (3) The K-means clustering method is applied to obtain a refined structure, and then a Lasso regression is applied to $\left\{Y_i - \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\xi}}_{k_1}, \boldsymbol{Z}_i\right\}_{i\in\widehat{\mathcal{T}}_{k_2}}$ for each refined subgroup $\widehat{\mathcal{T}}_{k_2}$ contained in $\mathcal{G}_{k_1}$. For K-means clustering methods, we consider clustering based on $\{Y_i\}_{i\in\mathcal{G}_{k_1}}$ (denoted as K-means1+Lasso), $\left\{Y_i - \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\xi}}_{k_1}\right\}_{i\in\mathcal{G}_{k_1}}$ (denoted as K-means2+Lasso), and $\left\{Y_i - \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\xi}}_{k_1} - \boldsymbol{Z}_i^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}}_{k_1}\right\}_{i\in\mathcal{G}_{k_1}}$ (denoted as K-means3+Lasso). Sparse coefficient estimation and hierarchical heterogeneous structures can be realized in all the above approaches. In comparison with our proposed approach, sparseFMR, K-means1+Lasso, K-means2+Lasso, and K-means3+Lasso all need a pre-setting number of refined subgroups, which is set as the true value. Our simulation results are summarized in Tables 1 and S1-S4 (Supporting Information), and also visualized in Figures 1 and S1-S3 (Supporting Information). Throughout the whole simulation, our proposed approach shows highly competitive performance.

TABLE 1 Mean of selected τ, mean, median, standard deviation (SD) of $\widehat{K}_2$ and percentage (per) of $\widehat{K}_2$ equal to the true number of subgroups under 100 simulated replicates with $\mu = 1$.

		Balanced design					Imbalanced design				
Cor structure	**Methods**	τ	**Mean**	**Median**	**SD**	**Per**	τ	**Mean**	**Median**	**SD**	**Per**
AR1	NOT incorporated	0.000	4.44	4.00	1.844	0.51	0.000	4.30	4.00	1.521	0.45
	ISHH-S_1^p	0.724	3.68	4.00	0.601	0.72	0.726	3.69	4.00	0.662	0.71
	ISHH-S_2^p	0.680	3.68	4.00	0.601	0.72	0.668	3.69	4.00	0.662	0.71
	ISHH-S_3^p	0.652	3.68	4.00	0.601	0.72	0.636	3.68	4.00	0.634	0.71
	ISHH-S_4^p	0.634	3.68	4.00	0.601	0.72	0.615	3.68	4.00	0.634	0.71
AR2	NOT incorporated	0.000	4.18	4.00	0.757	0.82	0.000	4.29	4.00	0.844	0.72
	ISHH-S_1^p	0.639	3.98	4.00	0.284	0.92	0.663	4.06	4.00	0.509	0.80
	ISHH-S_2^p	0.620	3.98	4.00	0.284	0.92	0.609	4.04	4.00	0.470	0.81
	ISHH-S_3^p	0.558	3.98	4.00	0.284	0.92	0.558	4.04	4.00	0.470	0.81
	ISHH-S_4^p	0.539	3.98	4.00	0.284	0.92	0.542	4.04	4.00	0.470	0.81
Banded	NOT incorporated	0.000	4.52	4.00	1.480	0.53	0.000	4.51	4.00	1.453	0.55
	ISHH-S_1^p	0.702	3.78	4.00	0.462	0.80	0.695	3.85	4.00	0.520	0.77
	ISHH-S_2^p	0.661	3.77	4.00	0.468	0.79	0.645	3.83	4.00	0.473	0.78
	ISHH-S_3^p	0.632	3.77	4.00	0.468	0.79	0.606	3.86	4.00	0.513	0.78
	ISHH-S_4^p	0.599	3.79	4.00	0.456	0.81	0.590	3.85	4.00	0.479	0.78

First, we compare the performance of proposed approaches with ($\tau \neq 0$) and without ($\tau = 0$) prior information incorporated. Take the setting of balanced design and $\mu = 1$ as an example (Tables 1,2 and Figure 1). Compared with the approach without information incorporated, our approaches significantly improve (at least 10% in the majority of cases) the percentage of identifying the number of refined subgroups whatever scenario of information is incorporated. It is observed that our approaches achieve 92% percentage of correct identification under AR2, suggesting the highest estimation performance. Numerical results suggest that information-incorporated approaches can contribute to estimating the number of refined subgroups. Although there is little difference in RIs of two paradigms (both of high accuracy around 0.8), it is observed that FPR is significantly reduced by the incorporated information with maintaining a high TPR. For examples, under AR1, FPRs/TPRs are 0.295/0.998(Not incorporated), 0.085/0.997(ISHH-S_1^p), 0.106/0.990(ISHH-S_2^p), 0.129/0.983(ISHH-S_3^p), and 0.153/0.977(ISHH-S_4^p). We analyze the reasons behind this in Remark 3. Moreover, the boxplots show that the adaptively selected weighted parameter τ decreases along with the worse information, which is as expected. Whatever the quality of the information, our approaches are virtually unaffected by identifying refined heterogeneous structures. More importantly, even if the worst case occurs, the proposed approach data-dependently selects a reasonable τ, leading the robust estimation, which is still highly comparable to the alternatives.

Moreover, compared with the alternatives, our proposed approaches show remarkable performance overall (Table 2 and Figure 1). Numerical results suggest that our proposed approaches have higher RIs than K-means+Lasso methods under any scenario of information. In the whole, our proposed approaches are pretty competitive on TPR and FPR performance compared with K-means+Lasso methods and may have superior performance when incorporating information of high quality. Although sparseFMR outperforms slightly on RIs, it leads to an undesirable FPR which fails on feature selection tasks. For coefficients estimation, our proposed approaches consistently outperform all alternatives. For examples, under Banded, MSEs of $\boldsymbol{\gamma}$ are 0.974 (sparseFMR), 0.624 (K-means1+Lasso), 0.669 (K-means2+Lasso), 0.579 (K-means3+Lasso), 0.504 (ISHH-S_1^p), 0.512 (ISHH-S_2^p), 0.515 (ISHH-S_3^p), and 0.522 (ISHH-S_4^p). It is particularly noted that our proposed approach data-drivenly obtains the number of refined subgroups while all alternatives are given a true value. Other simulation settings show similar results, which are provided in Supporting Information. All in all, even if some helpful additional information about the heterogeneous structure is unknown, our proposed approach still outperforms the competitors in the vast majority of cases.

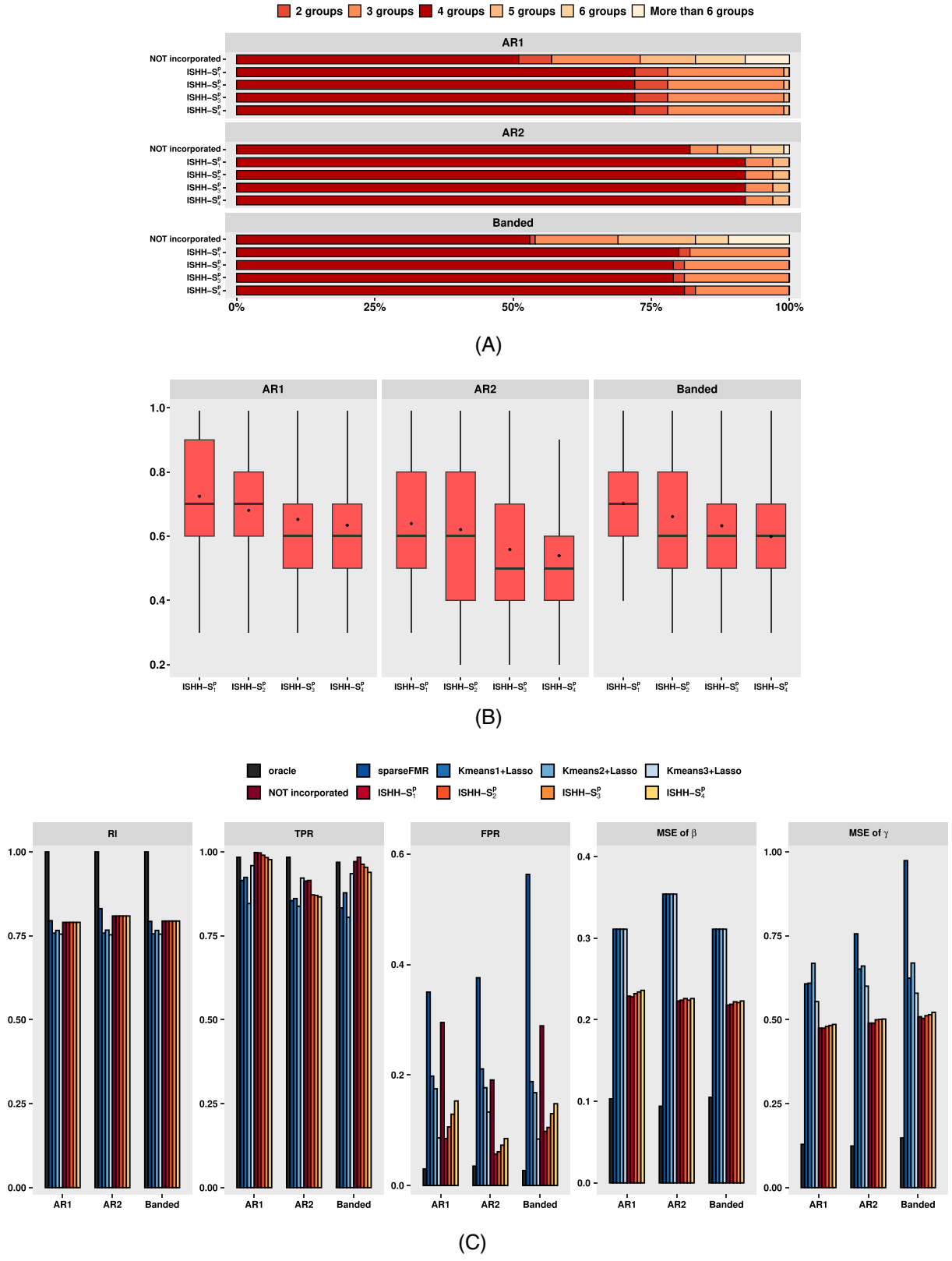

FIGURE 1 Simulation results under 100 simulated replicates with balanced design and $\mu = 1$. (A) Percentage of $\widehat{K}_2$ equal to candidate values. (B) Boxplots of selected τ. (C) Measures comparison with alternatives.

TABLE 2 Simulation results under 100 simulated replicates with balanced design and $\mu = 1$; in each cell, mean (SD).

Cor structure	Methods	RI	TPR	FPR	MSE of β	MSE of γ
AR1	oracle	1.000 (0.000)	0.984 (0.031)	0.030 (0.034)	0.103 (0.036)	0.129 (0.028)
	sparseFMR	0.795 (0.035)	0.915 (0.057)	0.350 (0.166)	0.311 (0.097)	0.607 (0.125)
	K-means1+Lasso	0.758 (0.007)	0.924 (0.069)	0.198 (0.077)	0.311 (0.097)	0.609 (0.041)
	K-means2+Lasso	0.766 (0.010)	0.846 (0.106)	0.175 (0.086)	0.311 (0.097)	0.668 (0.050)
	K-means3+Lasso	0.755 (0.006)	0.959 (0.063)	0.086 (0.053)	0.311 (0.097)	0.554 (0.039)
	NOT incorporated	0.790 (0.017)	0.998 (0.013)	0.295 (0.130)	0.229 (0.093)	0.475 (0.035)
	ISHH-S_1^{p}	0.790 (0.017)	0.997 (0.016)	0.085 (0.084)	0.228 (0.091)	0.475 (0.037)
	ISHH-S_2^{p}	0.790 (0.017)	0.990 (0.029)	0.106 (0.093)	0.232 (0.092)	0.480 (0.038)
	ISHH-S_3^{p}	0.790 (0.017)	0.983 (0.038)	0.129 (0.097)	0.234 (0.092)	0.483 (0.039)
	ISHH-S_4^{p}	0.790 (0.017)	0.977 (0.044)	0.153 (0.104)	0.236 (0.092)	0.486 (0.039)
AR2	oracle	1.000 (0.000)	0.984 (0.034)	0.035 (0.033)	0.094 (0.032)	0.124 (0.025)
	sparseFMR	0.831 (0.048)	0.855 (0.074)	0.376 (0.194)	0.354 (0.119)	0.756 (0.317)
	K-means1+Lasso	0.758 (0.008)	0.861 (0.080)	0.211 (0.061)	0.354 (0.119)	0.651 (0.056)
	K-means2+Lasso	0.767 (0.010)	0.838 (0.083)	0.177 (0.067)	0.354 (0.119)	0.660 (0.053)
	K-means3+Lasso	0.753 (0.005)	0.922 (0.069)	0.133 (0.063)	0.354 (0.119)	0.600 (0.051)
	NOT incorporated	0.809 (0.010)	0.913 (0.077)	0.191 (0.097)	0.223 (0.096)	0.490 (0.042)
	ISHH-S_1^{p}	0.809 (0.009)	0.915 (0.076)	0.057 (0.059)	0.224 (0.095)	0.490 (0.043)
	ISHH-S_2^{p}	0.809 (0.009)	0.872 (0.077)	0.061 (0.056)	0.226 (0.095)	0.500 (0.042)
	ISHH-S_3^{p}	0.809 (0.009)	0.870 (0.085)	0.073 (0.063)	0.224 (0.095)	0.501 (0.043)
	ISHH-S_4^{p}	0.809 (0.009)	0.866 (0.084)	0.085 (0.063)	0.226 (0.096)	0.502 (0.043)
Banded	oracle	1.000 (0.000)	0.969 (0.047)	0.027 (0.031)	0.105 (0.034)	0.148 (0.031)
	sparseFMR	0.793 (0.031)	0.833 (0.081)	0.563 (0.163)	0.311 (0.092)	0.974 (0.285)
	K-means1+Lasso	0.756 (0.006)	0.878 (0.076)	0.188 (0.074)	0.311 (0.092)	0.624 (0.039)
	K-means2+Lasso	0.766 (0.010)	0.805 (0.100)	0.168 (0.076)	0.311 (0.092)	0.669 (0.043)
	K-means3+Lasso	0.755 (0.006)	0.935 (0.063)	0.084 (0.050)	0.311 (0.092)	0.579 (0.041)
	NOT incorporated	0.794 (0.014)	0.971 (0.058)	0.289 (0.116)	0.218 (0.082)	0.509 (0.044)
	ISHH-S_1^{p}	0.794 (0.014)	0.984 (0.042)	0.098 (0.072)	0.219 (0.086)	0.504 (0.044)
	ISHH-S_2^{p}	0.794 (0.014)	0.963 (0.062)	0.105 (0.079)	0.222 (0.087)	0.512 (0.044)
	ISHH-S_3^{p}	0.794 (0.014)	0.954 (0.068)	0.130 (0.073)	0.221 (0.088)	0.515 (0.045)
	ISHH-S_4^{p}	0.794 (0.014)	0.939 (0.074)	0.148 (0.082)	0.223 (0.087)	0.522 (0.046)

Remark 3. The incorporation of prior information can significantly reduce FPR. Denote S as the set of ground-true important features and S^{C} as the set of ground-true unimportant features. Note that the false positive means the times of features predicted as important in S^{C} of n subjects. In information-guided step, $j \notin S^{\mathrm{P}}$ is equivalent to $j \in \left(S \setminus S^{\mathrm{P}}\right) \cup \left(S^{\mathrm{C}} \setminus S^{\mathrm{P}}\right)$. Thus our two-step approach imposes double sparsity penalization on features in $S^{\mathrm{C}} \setminus S^{\mathrm{P}}$, leading to fewer chances to be predicted as important. Even if the prior information is worst, there are still numerous features in $S^{\mathrm{C}} \setminus S^{\mathrm{P}}$ of n subjects imposed on double sparsity penalization, resulting in much lower FPR than no incorporation.

5 | REAL DATA ANALYSIS

The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) is a recent collective effort jointly organized by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), which has published

comprehensive clinical, omics, and imaging data over various cancer types. Note that some recent studies have explored the heterogeneity of lung cancer based on TCGA imaging data and provided a brand new understanding of disease biology and prognosis.[13] Other heterogeneity analysis on omics data has been also successfully conducted and led to meaningful biomedical findings.[16] In this section, we analyze the pathological imaging data and the gene expression data on lung adenocarcinoma (LUAD), which is a heterogeneous subtype of non-small cell lung cancer. The response variable of interest is the forced expiratory volume (FEV1), which is the percentage compared to a normal value reference range of the volume of air that a patient can forcibly exhale from the lungs in one second before using the bronchodilator. FEV1 is regarded as a crucial biomarker for lung cancer capacity and prognosis. One can refer to Wang et al[25,26] for more detailed information on extracting pathological imaging features. We focus on genes in the non-small cell lung cancer pathway. Considering the limitation of sample size and the burden of estimation efficiency, we use the prescreening technique to remove meaningless genes, and then select candidate gene features with frequencies larger than 50. Overall, the final analyzed data contains 6 pathological imaging features, 24 gene features, and a response variable for 116 subjects. Specific biological interpretations of the 6 extracted imaging features are provided in Table S6 and names of 24 candidate gene features are provided in Table S7 (Supporting Information). The histogram of scaled FEV1 in Figure S4 (Supporting Information) shows a mixed distribution of biomarkers from 116 subjects and indicates the sample heterogeneity.

The first type of features is clinical, which are informative for modeling a rough heterogeneous structure for subjects based on expert experience or some existing mature statistical methods. Here, we refer to standard fusion-based regression[7] to pre-subgroup all subjects into two rough subgroups with sizes 43 and 73. The second type of features is molecular, which provides the possibility for identifying a latent refined heterogeneous structure. For the prior information on feature importance, we conduct a search of co-occurrence frequencies of gene names and lung adenocarcinoma in PubMed publication database. The searching results with respect to 24 candidate gene features are significantly different as shown in Figure 2. To construct the prior information set S^p, we select gene features whose co-occurrence frequencies with lung adenocarcinoma are larger than a threshold. To obtain a more credible and convincing result, we consider four scenarios with the threshold set as 300, 400, 500, and 800, then the corresponding numbers of suggested gene features are 12, 9, 7, and 5. The detailed prior information is provided in Table S7 (Supporting Information).

The analysis results using our proposed approaches are summarized in Table 3. For sensitivity analysis with respect to different thresholds, we present the similarity measures of subgrouping structure and feature selection under four scenarios in Tables S9 and S10 (Supporting Information). The estimation of refined heterogeneous structure is identical and all similarity measures of feature selection are larger than 0.94, which validate that our approach is stable for different thresholds. In practice, we suggest to determine the optimal scenario with the smallest modified BIC. Under the above four scenarios, as shown in Table 3, four distinct refined sub-subgroups are identified, with sizes 22, 21, 48, and 25,

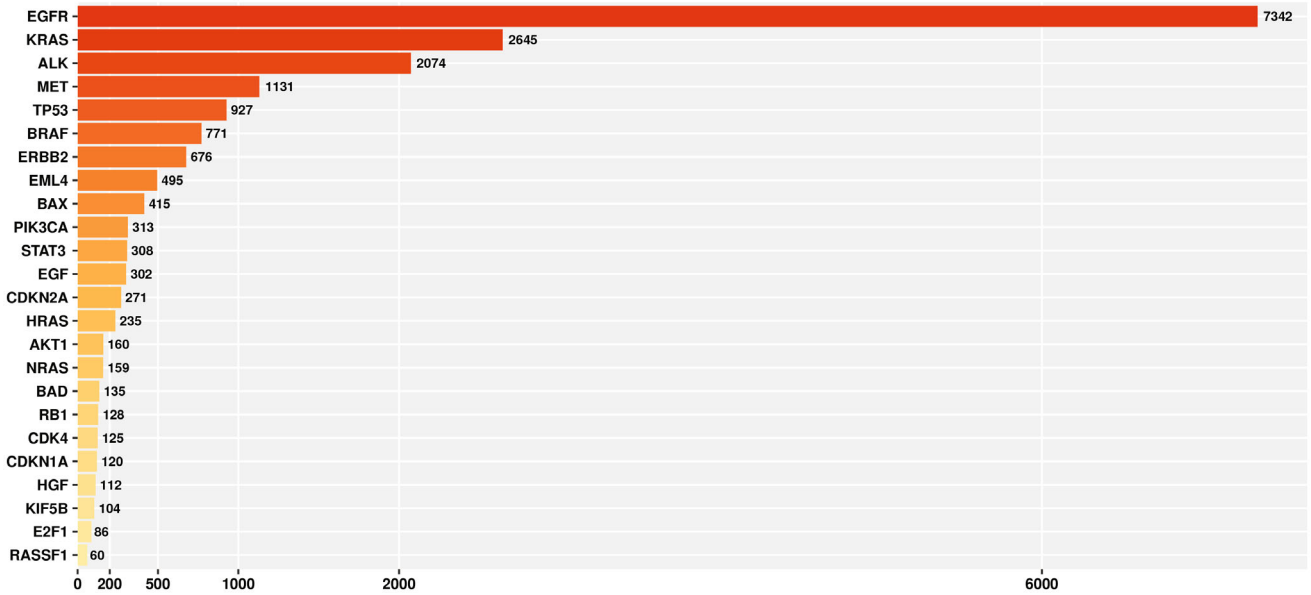

FIGURE 2 Analysis of LUAD data. Co-occurrence frequencies of publications with respect to Gene+LUAD in PubMed.

TABLE 3 Analysis of LUAD data using the proposed approaches under four scenarios.

ISHH-Scenario1 ($\tau = 0.6$)				
	Rough structure			
Imaging features	**1**		**2**	
LymphocytesPN	0.6889		0.0698	
StromaPN	−1.2379		−0.4867	
TumorPN	−0.1543		0.4140	
LymphocytesSN	−0.9344		0.0418	
StromaSN	1.0045		0.1126	
TumorSN	−0.9561		−0.1160	
	Refined structure (group size)			
Gene features	**1-1 (22)**	**1-2 (21)**	**2-1 (48)**	**2-2 (25)**
CDKN2A	0.0000	0.0592	0.0000	0.0000
CDK4	0.0000	0.0000	0.0000	0.0804
RB1	0.0571	−0.0158	0.0000	0.1236
E2F1	0.0000	0.0000	0.0000	−0.2439
KRAS	0.0000	0.0000	−0.0613	0.0281
RASSF1	0.1254	0.0000	0.0000	−0.0497
PIK3CA	−0.2403	0.0000	0.0000	−0.0505
AKT1	0.0000	0.0000	0.0000	0.3699
BAD	0.3757	0.0000	0.0000	0.0609
EGF	0.0416	−0.7518	0.0000	−0.1005
EGFR	0.2644	−0.0441	−0.0758	0.0000
ERBB2	−0.2738	0.0000	0.0000	−0.2370
HGF	0.0000	0.0000	0.0000	0.0777
MET	0.0000	0.0000	−0.0563	1.1186
HRAS	0.0000	−0.3945	−0.0047	0.0000
NRAS	0.0000	0.0000	−0.0011	0.2795
BRAF	0.0000	0.4644	0.0000	0.0000
TP53	0.0000	0.0124	0.0000	0.0568
CDKN1A	0.1755	0.0000	0.0000	0.0284
BAX	0.0927	0.0000	0.0000	0.3173
KIF5B	0.0000	0.0000	0.0000	0.0288
EML4	0.0000	0.0000	0.0000	0.0000
ALK	−1.3267	0.0000	0.0000	0.0000
STAT3	0.0000	0.0000	0.0000	0.0000
ISHH-Scenario2 ($\tau = 0.6$)				
	Rough structure			
Imaging features	**1**		**2**	
LymphocytesPN	0.4912		0.0664	
StromaPN	−1.1957		−0.5046	

(Continues)

TABLE 3 (Continued)

ISHH-Scenario1 ($\tau = 0.6$)				
	Rough structure			
Imaging features	**1**		**2**	
TumorPN	−0.2061		0.4308	
LymphocytesSN	−0.7517		0.0426	
StromaSN	0.9944		0.1167	
TumorSN	−0.9568		−0.1201	
	Refined structure (group size)			
Gene features	**1-1 (22)**	**1-2 (21)**	**2-1 (48)**	**2-2 (25)**
CDKN2A	0.0000	0.0763	0.0000	0.0000
CDK4	0.0000	0.0000	0.0000	0.0817
RB1	0.0409	−0.0308	0.0000	0.1245
E2F1	0.0000	0.0000	0.0000	−0.2415
KRAS	0.0000	0.0000	−0.0591	0.0329
RASSF1	0.1079	0.0000	0.0000	−0.0596
PIK3CA	−0.2105	0.0000	0.0000	−0.0502
AKT1	0.0000	0.0000	0.0000	0.3682
BAD	0.3711	0.0000	0.0000	0.0595
EGF	0.0000	−0.8115	0.0000	−0.0808
EGFR	0.2710	−0.0327	−0.0701	0.0000
ERBB2	−0.2799	0.0000	0.0000	−0.2395
HGF	0.0000	0.0000	0.0000	0.0747
MET	0.0000	0.0000	−0.0567	1.1182
HRAS	0.0000	−0.3901	−0.0022	0.0000
NRAS	0.0000	0.0000	0.0000	0.2883
BRAF	0.0000	0.4612	0.0000	0.0000
TP53	0.0000	0.0142	0.0000	0.0621
CDKN1A	0.1723	0.0000	0.0000	0.0161
BAX	0.0925	0.0000	0.0000	0.3159
KIF5B	0.0000	0.0000	0.0000	0.0228
EML4	0.0000	0.0000	0.0000	0.0000
ALK	−1.3250	0.0000	0.0000	0.0000
STAT3	0.0000	0.0000	0.0000	0.0000
ISHH-Scenario2 ($\tau = 0.8$)				
	Rough structure			
Imaging features	**1**		**2**	
LymphocytesPN	0.4362		0.0705	
StromaPN	−1.1827		−0.4924	
TumorPN	−0.2136		0.4188	

(Continues)

TABLE 3 (Continued)

ISHH-Scenario1 ($\tau = 0.8$)				
	Rough structure			
Imaging features	**1**		**2**	
LymphocytesSN	−0.6942		0.0383	
StromaSN	0.9853		0.1138	
TumorSN	−0.9511		−0.1170	
	Refined structure (group size)			
Gene features	**1-1 (22)**	**1-2 (21)**	**2-1 (48)**	**2-2 (25)**
CDKN2A	0.0000	0.0700	0.0000	0.0000
CDK4	0.0000	0.0000	0.0000	0.0807
RB1	0.0577	−0.0220	0.0000	0.1273
E2F1	0.0000	0.0000	0.0000	−0.2437
KRAS	0.0000	0.0000	−0.0725	0.0186
RASSF1	0.1032	0.0000	0.0000	−0.0618
PIK3CA	−0.2070	0.0000	0.0000	−0.0511
AKT1	0.0000	0.0000	0.0000	0.3682
BAD	0.3734	0.0000	0.0000	0.0574
EGF	0.0000	−0.8080	0.0000	−0.0804
EGFR	0.2647	−0.0410	−0.0673	0.0000
ERBB2	−0.2746	0.0000	0.0000	−0.2378
HGF	0.0000	0.0000	0.0000	0.0730
MET	0.0000	0.0000	−0.0575	1.1181
HRAS	0.0000	−0.3792	0.0000	0.0000
NRAS	0.0000	0.0000	0.0000	0.2886
BRAF	0.0000	0.4559	0.0000	0.0000
TP53	0.0000	0.0188	0.0000	0.0621
CDKN1A	0.1698	0.0000	0.0000	0.0144
BAX	0.0733	0.0000	0.0000	0.3160
KIF5B	0.0000	0.0000	0.0000	0.0241
EML4	0.0000	0.0000	0.0000	0.0000
ALK	−1.3282	0.0000	0.0000	0.0000
STAT3	0.0000	0.0000	0.0000	0.0000
ISHH-Scenario2 ($\tau = 0.8$)				
	Rough structure			
Imaging features	**1**		**2**	
LymphocytesPN	0.4526		0.0737	
StromaPN	−1.1619		−0.4689	
TumorPN	−0.2170		0.4000	

(Continues)

TABLE 3 (Continued)

ISHH-Scenario1 ($\tau = 0.8$)				
	Rough structure			
Imaging features	**1**		**2**	
LymphocytesSN	−0.7067		0.0333	
StromaSN	0.9758		0.1078	
TumorSN	−0.9408		−0.1106	
	Refined structure (group size)			
Gene features	**1-1 (22)**	**1-2 (21)**	**2-1 (48)**	**2-2 (25)**
CDKN2A	0.0000	0.0563	0.0000	0.0000
CDK4	0.0000	0.0000	0.0000	0.0810
RB1	0.0598	−0.0212	0.0000	0.1270
E2F1	0.0000	0.0000	0.0000	−0.2411
KRAS	0.0000	0.0000	−0.0917	0.0000
RASSF1	0.1011	0.0000	0.0000	−0.0632
PIK3CA	−0.2113	0.0000	0.0000	−0.0460
AKT1	0.0000	0.0000	0.0000	0.3673
BAD	0.3817	0.0000	0.0000	0.0574
EGF	0.0000	−0.8033	0.0000	−0.0792
EGFR	0.2568	−0.0577	−0.0694	0.0000
ERBB2	−0.2318	0.0000	0.0000	−0.2285
HGF	0.0000	0.0000	0.0000	0.0692
MET	0.0000	0.0000	−0.0578	1.1175
HRAS	0.0000	−0.3769	0.0000	0.0000
NRAS	0.0000	0.0000	0.0000	0.2938
BRAF	0.0000	0.4522	0.0000	0.0000
TP53	0.0000	0.0190	0.0000	0.0596
CDKN1A	0.1678	0.0000	0.0000	0.0121
BAX	0.0681	0.0000	0.0000	0.3155
KIF5B	0.0000	0.0000	0.0000	0.0264
EML4	0.0000	0.0000	0.0000	0.0000
ALK	−1.3294	0.0000	0.0000	0.0000
STAT3	0.0000	0.0000	0.0000	0.0000

Note: The estimated regression coefficients.

respectively, indicating that the two rough subgroups are both split into two refined sub-subgroups. As shown in Table 3, it is observed that the four sub-subgroups have significantly different models with different estimated coefficients, which also validates the necessity to perform refined subgroup analysis for individuals. It is also noted that a large number of estimated coefficients with respect to gene features are zero, which is caused by the sparsity penalties. Some gene features (such as EGFR, ERBB2, MET, BRAF, BAX, and ALK) have implications on some sub-subgroups while they are omitted in others. Such gene features have significantly different levels of effects between different sub-subgroups, which indicates that they are favorable for subgrouping and highly compatible with our prior information. In addition, as shown in Table 3, the optimal τ under the four scenarios are all larger than 0.5, which means that prior information is not fully trusted, but

tends to be adopted to a greater extent. It is more conducive for the model to make use of the additional information and is also consistent with our intuition.

In order to explore the rationality and biological meanings of refined subgrouping, we compare some clinical features across the estimated refined sub-subgroups. Specifically, the carbon monoxide diffusion and the percentage values to represent the result of forced expiratory volume in one second divided by the forced vital capacity (FVC) pre/post-bronchodilator among refined sub-subgroups are analyzed using ANOVA. Such clinical features indicate pulmonary ventilatory function and are informative for lung cancer staging and prognosis. The suggested P-values are 0.0586, 0.0666, and 0.0139, respectively, all of which suggest significant differences between the estimated refined heterogeneous structure. Note that the analyzed clinical features are not included in the aforementioned heterogeneity analyses, as a result, there is no concern about over-fitting. In a sense, this analysis contributes to the validity of the estimated refined heterogeneous structure.

We also use alternative approaches to analyze LUAD data for comparison. In fairness, we fix the number of refined sub-subgroups as four in sparseFMR and K-means+Lasso methods. All alternatives can obtain non-trivial heterogeneous refined structures and sparse coefficients estimation. The estimation results are summarized in Tables S8 and S11 (Supporting Information). It is observed that different approaches lead to different refined heterogeneous structures and coefficients estimation. We compute similarity measures for pairs of heterogeneous structures with respect to different approaches, and the similarity values are all larger than 0.7 as shown in Table S10 (Supporting Information). In particular, our proposed approach has moderate subgrouping similarity with the alternatives, which are 0.760 (sparseFMR), 0.777 (K-means1+Lasso), 0.763 (K-means2+Lasso), and 0.778 (K-means3+Lasso). This finding is reasonable since the data heterogeneity is intrinsic and different analysis schemes yield moderately similar heterogeneous structures. Moreover, compared with the alternatives, our proposed approaches also have similar results on gene feature selection to some extent. However, it is especially noted that all alternatives are pre-given a fixed number of components, which is not realistic and not flexible as our proposed approaches.

For the evaluation of the stability and the prediction performance, we consider the following procedure. For each subject, we remove itself and apply our approaches to the remaining 115 subjects, then compare the heterogeneous refined structure obtained on the whole data with the one subject removed. The mean (SD) similarity measures through 116 subjects are 0.8970(0.0728) (ISHH-Scenario1), 0.8970(0.0728) (ISHH-Scenario2), 0.9000(0.0717) (ISHH-Scenario3), and 0.8969(0.0726) (ISHH-Scenario4), which provides persuasive support for the stability of our proposed approaches. In addition, we make predictions on the removed subject by our approaches and the alternatives. Through all 116 subjects, the reported prediction MSEs are 1.1797 (ISHH-Scenario1), 1.1778 (ISHH-Scenario2), 1.1804 (ISHH-Scenario3), and 1.1801 (ISHH-Scenario4). As for the alternatives, the best prediction MSE is 1.1929 and the worst is 1.4852, both worse than our approaches under four scenarios. Overall, our proposed approach can produce a stable refined heterogeneous structure, select important gene features within sub-subgroups reliably, and enjoy favorable prediction performance.

6 | DISCUSSION

In this article, we have conducted a supervised regression-based hierarchical heterogeneity analysis integrating two types of data. Penalized fusion has been adopted for identifying a latent refined heterogeneous structure and additional penalization has been adopted for the feature sparsity. A significant advancement of this study is the incorporation of prior information on feature identification. We have made great use of additional information extracted from existing publications, which alleviated the dilemma of lack of information to some extent. In the two-step approach, a proper balance between the observed data and prior information is data-drivenly achieved. The theoretical achievements have provided a solid ground for our proposed approaches. Simulation has shown competitive and stable performance even if the prior information is partially correct. The LUAD data analysis has demonstrated practical biomedical findings.

This study can be potentially extended in multiple ways. Our proposed approach is not limited to imaging data and gene expression data, moreover, it can be conducted with any two types of data with hierarchy. And the least square loss can be replaced by other lack-of-fit measures corresponding to data distributions. Furthermore, it is of interest to develop a more reliable framework to incorporate various types of prior information, such as involving an index to measure the authenticity of information and generating qualitative rather than quantitative information.

FUNDING INFORMATION

This work was partially supported by the National Natural Science Foundation of China (12171454), Fundamental Research Funds for the Central Universities, NIH (CA204120), and NSF (2209685).

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings in this article are openly available in TCGA (The Cancer Genome Atlas) at https://portal.gdc.cancer.gov/projects/TCGA-LUAD. In summary of lung adenocarcinoma data, the pathological imaging data is available at https://wiki.cancerimagingarchive.net/display/Public/TCGA-LUAD, and the omics data and clinical data of patients are available at https://www.cbioportal.org/study/summary?id=luad_tcga. More detailed access to analyzed lung adenocarcinoma data is provided in Supporting Information. R programs including codes for the proposed method and reproduction of numerical simulations and real data analysis are available at GitHub (https://github.com/HHanWei/ISHH).

ORCID

Wei Han https://orcid.org/0009-0004-1344-9365
Shuangge Ma https://orcid.org/0000-0001-9001-4999
Mingyang Ren https://orcid.org/0000-0002-8061-9940

REFERENCES

1. Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat Rev Cancer*. 2014;14(8):535-546.
2. Bianchini G, Balko JM, Mayer IA, Sanders ME, Gianni L. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nat Rev Clin Oncol*. 2016;13(11):674-690.
3. Ma H, Zeng D, Liu Y. Learning individualized treatment rules with many treatments: a supervised clustering approach using adaptive fusion. *Adv Neural Inf Process Syst*. 2022;35:15956-15969.
4. Khalili A, Chen J. Variable selection in finite mixture of regression models. *J Am Stat Assoc*. 2007;102(479):1025-1038.
5. Städler N, Bühlmann P, Van De Geer S. l_1-penalization for mixture regression models. *TEST*. 2010;19:209-256.
6. Ma S, Huang J. A concave pairwise fusion approach to subgroup analysis. *J Am Stat Assoc*. 2017;112(517):410-423.
7. Ma S, Huang J, Zhang Z, Liu M. Exploration of heterogeneous treatment effects via concave fusion. *Int J Biostat*. 2019;16(1):20180026.
8. Chen J, Tran-Dinh Q, Kosorok MR, Liu Y. Identifying heterogeneous effect using latent supervised clustering with adaptive fusion. *J Comput Graph Stat*. 2021;30(1):43-54.
9. Zhu X, Qu A. Cluster analysis of longitudinal profiles with subgroups. *Electron J Stat*. 2018;12(1):171-193.
10. Zhang Y, Wang HJ, Zhu Z. Robust subgroup identification. *Stat Sin*. 2019;29(4):1873-1889.
11. Hu X, Huang J, Liu L, Sun D, Zhao X. Subgroup analysis in the heterogeneous cox model. *Stat Med*. 2021;40(3):739-757.
12. Zhang X, Zhang Q, Ma S, Fang K. Subgroup analysis for high-dimensional functional regression. *J Multivar Anal*. 2022;192:105100.
13. Luo X, Zang X, Yang L, et al. Comprehensive computational pathological image analysis predicts lung cancer prognosis. *J Thorac Oncol*. 2017;12(3):501-509.
14. Choi H, Na KJ. Integrative analysis of imaging and transcriptomic data of the immune landscape associated with tumor metabolism in lung adenocarcinoma: clinical and prognostic implications. *Theranostics*. 2018;8(7):1956-1965.
15. Connor AA, Gallinger S. Pancreatic cancer evolution and heterogeneity: integrating omics and clinical data. *Nat Rev Cancer*. 2022;22(3):131-142.
16. Yu KH, Berry GJ, Rubin DL, Re C, Altman RB, Snyder M. Association of omics features with histopathology patterns in lung adenocarcinoma. *Cell Syst*. 2017;5(6):620-627.
17. Jiang Y, He Y, Zhang H. Variable selection with prior information for generalized linear models via the prior LASSO method. *J Am Stat Assoc*. 2016;111(513):355-376.
18. Wang X, Xu Y, Ma S. Identifying gene-environment interactions incorporating prior information. *Stat Med*. 2019;38(9):1620-1633.
19. Yi H, Zhang Q, Lin C, Ma S. Information-incorporated Gaussian graphical model for gene expression data. *Biometrics*. 2022;78(2):512-523.
20. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894-942.
21. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348-1360.
22. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Vol 3. Boston, MA: Now Foundations and Trends; 2011:1-122.
23. Liu L, Lin L. Subgroup analysis for heterogeneous additive partially linear models and its application to car sales data. *Comput Stat Data Anal*. 2019;138:239-259.

24. Chen K, Huang R, Chan NH, Yau CY. Subgroup analysis of zero-inflated Poisson regression model with applications to insurance data. *Insur Math Econ*. 2019;86:8-18.
25. Wang S, Wang T, Yang L, et al. ConvPath: a software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. *EBioMedicine*. 2019;50:103-110.
26. Wang S, Rong R, Yang DM, et al. Computational staining of pathology images to study the tumor microenvironment in lung cancer. *Cancer Res*. 2020;80(10):2056-2066.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Han W, Zhang S, Ma S, Ren M. Information-incorporated sparse hierarchical cancer heterogeneity analysis. *Statistics in Medicine*. 2024;43(11):2280-2297. doi: 10.1002/sim.10071