Received: 7 April 2021 | Revised: 20 September 2021 | Accepted: 9 December 2021

DOI: 10.1002/sim.9309

RESEARCH ARTICLE

Statistics in Medicine WILEY

Bayesian hierarchical finite mixture of regression for histopathological imaging-based cancer data analysis

Yunju Im[1] | **Yuan Huang**[1] | **Jian Huang**[2] | **Shuangge Ma**[1]

[1]Department of Biostatistics, Yale University, New Haven, Connecticut, USA

[2]Department of Statistics and Actuarial Science, University of Iowa, Iowa, Iowa, USA

Correspondence
Shuangge Ma, Department of Biostatistics, Yale University, 60 College ST, New Haven, CT 06520, USA.
Email: shuangge.ma@yale.edu

Funding information
National Institutes of Health, Grant/Award Number: CA241699, CA196530, CA204120; National Science Foundation, Grant/Award Number: 1916251; Yale Cancer Center Pilot Award

Cancer is heterogeneous, and for seemingly similar cancer patients, the associations between an outcome/phenotype and covariates can be different. To describe such differences, finite mixture of regression (FMR) and other modeling techniques have been developed. "Classic" FMR analysis has usually been based on clinical, demographic, and molecular variables. More recently, histopathological imaging data—which is a byproduct of biopsy and enjoys broader data availability and higher cost-effectiveness—has been increasingly used in cancer modeling, although it is noted that its application to cancer FMR analysis still remains limited. In this article, we further advance cancer FMR analysis based on histopathological imaging data. Significantly advancing from the existing analyses under heterogeneity and homogeneity, our goal is to *simultaneously* use two types of histopathological imaging features, which are extracted based on domain-specific biomedical knowledge and using automated signal processing software, respectively. A significant modeling/methodological advancement is that, to reflect the "increased resolution" of the second type of imaging features over the first type, we impose a hierarchy in the mixture structures. An effective and flexible Bayesian approach is proposed. Simulation shows its competitiveness over several alternatives. The TCGA lung cancer data is analyzed, and interesting heterogeneous structures different from using the alternatives are found. Overall, this study provides a new venue for FMR analysis for cancer and other complex diseases.

KEYWORDS
Bayesian estimation, cancer, finite mixture of regression, hierarchy, histopathological imaging data

1 | INTRODUCTION

Heterogeneity is a hallmark of cancer. Interestingly, the definitions of cancer heterogeneity are very "heterogeneous", ranging from very "micro" (eg, pertained to differences between cancer cells) to very "macro" (eg, pertained to clinical characteristics). In this article, we consider an analysis which has been popular in the statistical literature, in which there is a cancer outcome/phenotype of interest, and its associations with a set of covariates differ across seemingly similar patients.[1,2] For this analysis, the most popular technique is perhaps FMR (finite mixture of regression), under which subjects form subgroups, those in the same subgroup share the same regression model (for the outcome/phenotype of interest), and different subgroups have different regression models. FMR is a relatively mature technique, and there have been extensive methodological developments[3] and applications.[1]

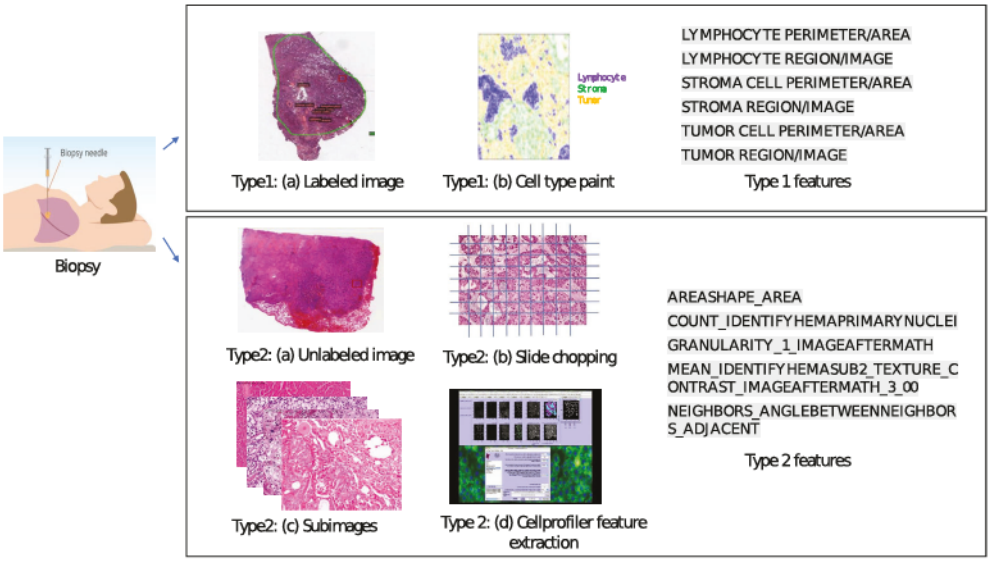

FIGURE 1 Pipelines for extracting Type 1 (upper) and Type 2 (lower) imaging features

In early cancer FMR analysis, clinical, and demographic variables have been extensively analyzed. With the development of profiling techniques, molecular data has brought additional insight beyond clinical/demographic variables, leading to finer model structures. Relatively recently, an alternative source data coming from histopathological images has drawn increasing attention. Such images are generated by biopsy, which is routinely ordered for cancer patients and suspected ones. As such, compared to molecular and some other types of data, histopathological imaging data can be highly advantageous with broader availability and higher cost-effectiveness. Histopathological images contain rich information on tumors' micro properties and microenvironment, and have been long used for the purposes of diagnosis and staging. In a series of recent studies, histopathological imaging features have been successfully used for modeling cancer outcomes and phenotypes.[4] Here it is noted that only a handful of studies have analyzed histopathological imaging data for quantifying differences across cancer patients (especially under the regression framework).[5,6]

This study aims to further advance histopathological imaging data-based mixture modeling analysis, which can complement the existing cancer mixture modeling analysis based on other types of data as well as that based on histopathological imaging data but with simple statistical techniques. More importantly, significantly advancing from the existing literature, we will *simultaneously use two types of features extracted from the same histopathological images*. As shown in the upper panel of Figure 1, the generation of the first type of imaging features involves labeling regions of interest, painting different types of cells using different colors, and then computing a small number of features based on the painting.[7] Such features have been motivated by long clinical practice and have a strong biological ground. On the negative side, this process is labor-intensive, and the extracted information is limited by the available biological knowledge. Applications of the first type of imaging features have been considered by Li et al,[8] Zhang et al,[9] and others. The pipeline for extracting the second type of imaging features is sketched in the lower panel of Figure 1 and based on automated signal processing. With unlabeled images, it involves chopping images into small subimages, randomly selecting subimages, extracting features using signal processing software such as *CellProfiler*,[10,11] and averaging over subimages to generate final imaging features. They demand very limited human labor and have been shown as able to extract features not visible to human eyes. On the negative side, the extracted features do not have direct biological interpretations. In addition, this type of imaging features usually has a higher dimensionality than the first type and may contain substantial noises. Applications of the second type of imaging features have been examined by Yu et al,[4] Luo et al,[5] and others. Here it is worth re-emphasizing that the two types of imaging features are generated on the same pathological slides. In the current clinical practice, the first type of features is "automatically" generated, and the second type of features is getting increasingly popular.

When including both types of imaging features in the same analysis, we note that they have a natural order. In particular, the second type of imaging features has a shorter history, has been developed to represent information not visible to human eyes (as reflected in the first type of features), and may provide finer information. Accordingly, this study targets

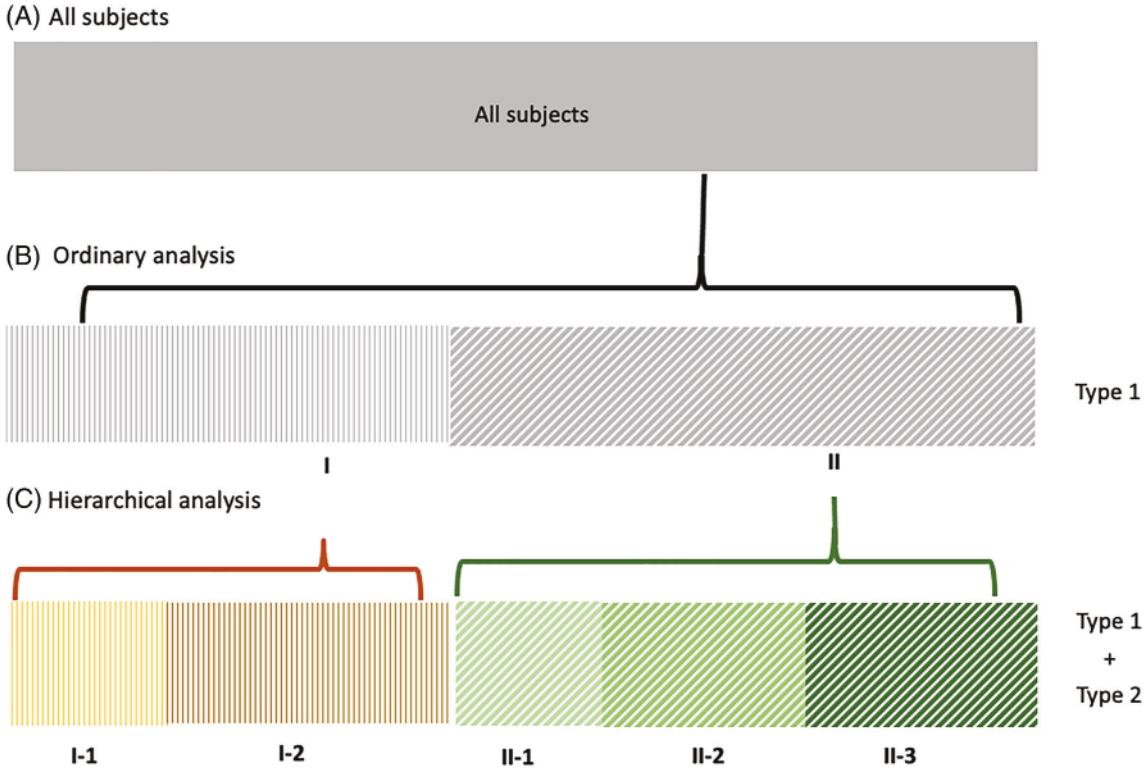

FIGURE 2 Analysis scheme: different colors represent different (sub-)subgroups

at addressing the question: *once a "rough" subgrouping structure has been identified using the first type of imaging features, can we further refine the structure by adding the second type of imaging features?* This analysis scheme is sketched in Figure 2. With a set of subjects (upper panel), "ordinary" mixture modeling analysis separates them into subgroups (middle panel). In our analysis, as shown in the lower panel, the goal is to separate subjects into "rough" subgroups based on the first type of imaging features and, *at the same time*, further separate subjects in the same subgroups into sub-subgroups by additionally using the second type of imaging features. Hierarchical grouping structures are very common in unsupervised clustering analysis. In clustering analysis, distances are calculated based on observed covariate values. In our analysis, "distance" between two subjects is defined based on their unobservable regression models. Model estimation and grouping need to be simultaneously conducted, making the analysis significantly more challenging and different from hierarchical clustering.

Most of the existing mixture modeling studies conduct "one-level" analysis. In such studies, both frequentist and Bayesian techniques have been adopted, and it has been well recognized that both types of techniques have pros and cons, with no one dominating the other. As such, it is of interest to develop new analysis methods in both domains. We refer to the works by McLachlan et al[3] and Schlattmann[1] for representative examples of the existing methods. In a recent study,[12] mixture modeling with a hierarchical structure is conducted using the penalization technique. It involves pairwise fusion, which is computationally highly expensive. In this article, we focus on the scenario with low-dimensional covariates—which the logical first step, but note that in the literature there have also been developments tailored to high-dimensional data.[2,13,14]

This study may have high significance in the following aspects. First, it can further advance cancer mixture modeling analysis. In particular, it is based on histopathological imaging data, which has great potential but has been limitedly studied for cancer modeling. Second, the analysis scheme, as sketched in Figure 2, significantly differs from the existing one-level analysis. It is noted that, when the second type of imaging features brings no new information, it simplifies to the "ordinary" mixture modeling analysis and hence includes it as a special case. It is also noted that this analysis scheme is not limited to histopathological imaging data—in fact, it can be studied as long as there are multiple types of covariates collected from the same subjects and with a natural order. Third, a new Bayesian approach is developed, which may further enrich the family of Bayesian FMR techniques. It is fundamentally different from the penalization technique.[12] Its computational and empirical properties are carefully investigated. Last but not least, this study may provide a new

look at the TCGA data and lung cancer heterogeneity and also serve as a prototype for future analysis using the proposed approach.

2 | METHODS

2.1 | Modeling

Consider a set of n independent subjects $\{(y_i, \boldsymbol{x}_i, \boldsymbol{w}_i),\ i = 1, \ldots, n\}$. For subject i, y_i is the response variable. $\boldsymbol{x}_i$ is the p-dimensional vector for conducting the "rough" subgrouping. In data analysis, beyond the first type of imaging features, $\boldsymbol{x}_i$ may also include demographic and clinical variables—such variables have been traditionally considered along with manually examining histopathological images to make cancer diagnosis, staging, etc. $\boldsymbol{w}_i$ is the q-dimensional vector for conducting the refined subgrouping. In our analysis, it contains the second type of imaging features. We first consider a continuous response with a Gaussian distribution, which matches the data analyzed in Section 4:

$$y_i|\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\beta}_i, \boldsymbol{\theta}_i, \sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_i + \boldsymbol{w}_i^T\boldsymbol{\theta}_i,\ \sigma^2\right),\ i = 1, \ldots, n. \tag{1}$$

The n subjects are assumed to be independent, as in the published mixture modeling studies. For the TCGA study to be analyzed in Section 4 and many others, this independence assumption is sensible. Exploring more general correlation assumption is beyond our scope. Here $\boldsymbol{\beta}_i$ and $\boldsymbol{\theta}_i$ are the vectors of regression coefficients for $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$, respectively, and the superscript "T" represents transpose. As in many published studies, we conclude two subjects as in the same subgroup (sub-subgroup) if and only if they have the same $\boldsymbol{\beta}_i$ ($\boldsymbol{\theta}_i$). As such, determining the mixture modeling structure amounts to examining the values of regression coefficients. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^T$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)^T$ denote the vectors of the latent subgroup and sub-subgroup memberships for the n subjects. Advancing from the existing one-level analyses, to accommodate the more complex subgrouping structure, two latent membership vectors are needed. Denote K as the number of subgroups defined based on $\boldsymbol{x}$ (ie, there are K distinct values of $\boldsymbol{\beta}_i$'s). In general, it is challenging to determine the value of K. In some of the existing Bayesian studies, K is assumed to be known. In our analysis, we take a more flexible strategy, do not assume a known value, and also impose a prior on K. Such a strategy has been referred to as Mixture of Finite Mixture (MFM) modeling.[15] Given K, let $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}_1^{*T}, \ldots, \boldsymbol{\beta}_K^{*T}\right)^T$ denote the vector composed of all subgroup-specific regression coefficients for $\boldsymbol{x}$. Here we note that $\boldsymbol{\beta}_d^* = \left(\beta_{d,0}^*, \beta_{d,1}^*, \ldots, \beta_{d,p}^*\right)^T$ is $(p+1)$-dimensional, and the first component corresponds to intercept. Without mixture modeling, intercept can be easily omitted with normalization. However, this is not the case in our analysis. Let $\boldsymbol{p}_1 = (p_{1,1}, \ldots, p_{1,K})^T$ be the vector of the relative sizes of the K subgroups, where the subscript "1" stresses the first level of mixture modeling. Now consider the dth subgroup. Conditioning on K, denote K_d as the number of sub-subgroups for $d = 1, \ldots, K$. Here it is also flexibly assumed that K_d is unknown, and a prior will be imposed. Let $\boldsymbol{\theta}_d^* = \left(\boldsymbol{\theta}_{1|d}^{*T}, \ldots, \boldsymbol{\theta}_{K_d|d}^{*T}\right)^T$ denote the sub-subgroup-specific regression coefficients for $\boldsymbol{w}$, where $\boldsymbol{\theta}_{j|d}^* = \left(\theta_{j|d,1}^*, \ldots, \theta_{j|d,q}^*\right)^T$ is q-dimensional and corresponds to the jth sub-subgroup (within the dth subgroup). Let $\boldsymbol{p}_{2,d} = (p_{2,1}, \ldots, p_{2,K_d})^T$ be the vector of the relative sizes of the K_d sub-subgroups. Collectively, denote $\boldsymbol{\theta}^* = \left(\boldsymbol{\theta}_1^{*T}, \ldots, \boldsymbol{\theta}_K^{*T}\right)^T$ and $\boldsymbol{p}_2 = \left(\boldsymbol{p}_{2,1}^T, \ldots, \boldsymbol{p}_{2,K}^T\right)^T$.

We make the following distributional specifications:

$$y_i|\delta_i = d, \rho_i = c, \boldsymbol{\delta}_{-i}, \boldsymbol{\rho}_{-i}\boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \boldsymbol{p}_1, \boldsymbol{p}_2, \sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_d^* + \boldsymbol{w}_i^T\boldsymbol{\theta}_{c|d}^*,\ \sigma^2\right),\ i = 1, \ldots, n;\ c = 1, \ldots, K_d,\ d = 1, \ldots, K,$$

$$\rho_i|\delta_i = d, \boldsymbol{p}_{2,d}, K_d, K \overset{\text{iid}}{\sim} \text{Multinomial}(1; p_{2,1}, \ldots, p_{2,K_d}),\ i = 1, \ldots, n, \tag{2a}$$

$$\boldsymbol{\theta}_{c|d}^*|K_d, K \overset{\text{iid}}{\sim} \mathrm{N}_q(m_\theta, \Sigma_\theta),\ c = 1, \ldots, K_d,\ d = 1, \ldots, K, \tag{2b}$$

$$p_{2,1}, \ldots, p_{2,K_d}|K_d, K \sim \text{Dirichlet}(\gamma_2, \ldots, \gamma_2),\ d = 1, \ldots, K, \tag{2c}$$

$$K_d|K \overset{\text{iid}}{\sim} p_{K_d}(\cdot),\ d = 1, 2, \ldots, K, \tag{2d}$$

$$\delta_i|\boldsymbol{p}_1, K \overset{\text{iid}}{\sim} \text{Multinomial}(1; p_{1,1}, \cdots, p_{1,K}),\ i = 1, \ldots, n, \tag{2e}$$

$$\boldsymbol{\beta}_d^*|K \overset{\text{iid}}{\sim} \text{N}_p(m_\beta, \Sigma_\beta),\ d = 1, \ldots, K, \tag{2f}$$

$$p_{1,1}, \ldots, p_{1,K}|K \sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_1), \tag{2g}$$

$$K \sim p_K(\cdot), \tag{2h}$$

$$\sigma^2 \sim \text{Inverse Gamma}(a_0, b_0). \tag{2i}$$

Here, loosely speaking, the MFM technique is applied at two levels. In particular, (2e) to (2h) define the first level subgrouping structure. Within each subgroup, (2a) to (2d) define the second level sub-subgrouping structure. By applying the MFM technique within each subgroup, the hierarchical structure is guaranteed. For each level, the distributional specifications have high similarity with those in the existing studies. For example, the Gaussian distribution for y_i, multinomial distributions for ρ_i and δ_i, Gaussian distributions for $\boldsymbol{\theta}$'s and $\boldsymbol{\beta}$'s, and inverse Gamma distribution for σ^2 have all been extensively adopted. For K and K_d's, multiple proper priors with support on $\{1, 2, \ldots\}$, for example Geometric and Poisson, can be taken. In our numerical study, we take the Geometric prior. For the sizes of the subgroups and sub-subgroups (within subgroups), Dirichlet distributions are assumed. It is noted that we assume symmetric Dirichlet distributions with single parameters, which may be less flexible than those with possibly different parameters. This simplification has been taken in published studies[15] and can significantly simplify computation: it allows the subgroup and sub-subgroup sizes and numbers to be marginalized by summation or integration, removing the trans-dimensional problem that arises from different dimensional parameters at different values of the numbers of (sub-)subgroups. Under such a specification, the distribution of the partition of subjects takes a form that can be readily computed.[15]

Different from the "standard" MFM where partition is defined only by $\boldsymbol{\delta}$, the above model induces a hierarchical partition of subjects, C, defined by the (sub-)subgroup memberships $(\boldsymbol{\delta}, \boldsymbol{\rho})$. By extending published results,[15] we obtain the distribution over C: $p(C) = p(\boldsymbol{\rho}|\boldsymbol{\delta})p(\boldsymbol{\delta}) = \left[\prod_{d\in\boldsymbol{\delta}} p(\boldsymbol{\rho}_d)\right] p(\boldsymbol{\delta})$, where $\boldsymbol{\rho}_d = \{\rho_i : i \in R_d\}$ and $R_d = \{i : \delta_i = d\}$. Following the literature,[15] we have $p(\boldsymbol{\delta}) = V_n(t) \prod_{d\in\boldsymbol{\delta}} \gamma_1^{(n_d)}$ and $p(\boldsymbol{\rho}_d) = V_{n_d}(t_d) \prod_{c\in\boldsymbol{\rho}_d} \gamma_2^{(n_{c|d})}$, where $n_d = |\{i : \delta_i = d\}|$ is the size of the dth subgroup, t is the number of distinct values of $\boldsymbol{\delta}$, $n_{c|d} = |\{i : \delta_i = d, \rho_i = c\}|$ is the size of the cth sub-subgroup of the dth subgroup, $t_d = |\boldsymbol{\rho}_d|$ is the number of unique values in $\boldsymbol{\rho}_d$, and $V_n(t) = \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma_1 k)^{(n)}} p_K(k)$, $a^{(b)} = a(a+1)\ldots(a+b-1)$, $a^{(b)} = a(a-1)\ldots(a-b+1)$ with $a^{(0)} = a_{(0)} = 1$.

2.2 | Computation

We develop a Gibbs sampler to draw samples from the posterior distribution of interest, which can be written as:

$$p(\boldsymbol{\delta}, \boldsymbol{\rho}, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \sigma^2|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{\rho}, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \sigma^2) p(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*|\boldsymbol{\delta}, \boldsymbol{\rho}) p(\boldsymbol{\delta}, \boldsymbol{\rho}) p(\sigma^2). \tag{3}$$

In (3), the dependence on $\boldsymbol{x}$ and $\boldsymbol{w}$ is suppressed for convenience, and we continue this convention throughout this article unless otherwise specified. We adopt the auxiliary variable method,[15,16] which consecutively updates latent group memberships, parameters for heterogeneous effects, and parameters for homogeneous effects. As described above, $(\boldsymbol{\delta}, \boldsymbol{\rho})$ induces partitions of $[n] := \{1, \ldots, n\}$ that have a hierarchical structure. In the algorithm presented below, as in published studies, the values of $(\boldsymbol{\delta}, \boldsymbol{\rho})$ do not carry any numerical meaning and are solely used to indicate which subjects belong to which (sub-)subgroups.

We first describe updating the latent (sub-)subgroup memberships. Notation-wise, for a vector $v = (v_1, \ldots, v_n)^T \in R^n$, where v can be $\boldsymbol{y}$, $\boldsymbol{\delta}$, or $\boldsymbol{\rho}$, let v_{-i} denote v with its ith component removed. For any d (which takes value in the components of $\boldsymbol{\delta}$), let $v_d = \{v_j : \delta_j = d\}$, n_d denote its size, $v_{d,-i}$ denote v_d with its ith component removed, and $n_{d,-i}$ denote its size. For c (which takes value in the components of $\boldsymbol{\rho}_d$), let $R_{c|d} = \{j : \delta_j = d, \rho_j = c\}$. $\boldsymbol{\rho}_{c|d} = \{\rho_j : j \in R_{c|d}\}$, and denote its size as $n_{c|d}$. Let $\boldsymbol{\rho}_{c|d,-i}$ denote the vector $\boldsymbol{\rho}_{c|d}$ with components corresponding to the ith subject removed, and denote its size as $n_{c|d,-i}$. Let t and t_d denote the numbers of unique values in $\boldsymbol{\delta}_{-i}$ and $\boldsymbol{\rho}_{d,-i}$, respectively. Let $\boldsymbol{\beta}_a^{*T} = \left(\boldsymbol{\beta}_{t+1}^{*T}, \ldots, \boldsymbol{\beta}_{t+m}^{*T}\right)$ and $\boldsymbol{\theta}_a^{*T} = \left(\boldsymbol{\theta}_{a,1}^{*T}, \ldots, \boldsymbol{\theta}_{a,t}^{*T}, \boldsymbol{\theta}_{a,t+1}^{*T}, \ldots, \boldsymbol{\theta}_{a,t+m}^{*T}\right)$, where $\boldsymbol{\theta}_{a,d}^{*T} = \left(\boldsymbol{\theta}_{t_d+1}^{*T}, \ldots, \boldsymbol{\theta}_{t_d+m}^{*T}\right)$ for $d = 1, \ldots, t$, denote a set of m and $(t+m)m$ auxiliary variables that are identically and independently distributed from their priors specified in the previous subsection. Then

the full conditional distribution of (δ_i, ρ_i) is:

$$p(\delta_i = d, \rho_i = c|\boldsymbol{\delta}_{-i}, \boldsymbol{\rho}_{-i}) \propto \begin{cases} (n_{d,-i} + \gamma_1)\frac{n_{s|d,-i}+\gamma_2}{\frac{V_{n_{d,-i}+1}(t_d+1)}{V_{n_{d,-i}+1}(t_d)}\gamma_2+n_{d,-i}+\gamma_2 t_d} & \text{if } d = \delta_k \text{ and } c = \rho_k \text{ for some } k \neq i, \\ (n_{d,-i} + \gamma_1)\frac{\frac{V_{n_{d,-i}+1}(t_d+1)}{V_{n_{d,-i}+1}(t_d)}\alpha_2}{\frac{V_{n_{d,-i}+1}(t_d+1)}{V_{n_{d,-i}+1}(t_d)}\gamma_2+n_{d,-i}+\gamma_2 t_d} & \text{if } d = \delta_k \text{ and } c \neq \rho_k \text{ for some } k \neq i, \\ \frac{V_n(t+1)}{V_n(t)}\gamma_1 & \text{if } d \neq \delta_k \text{ and } c \neq \rho_k \text{ for all } k \neq i. \end{cases}$$

$$p(\delta_i = d, \rho_i = c|\boldsymbol{\delta}_{-i}, \boldsymbol{\rho}_{-i}, \boldsymbol{\beta}^*, \boldsymbol{\beta}^*_a, \boldsymbol{\theta}^*, \boldsymbol{\theta}^*_a, \sigma^2, \boldsymbol{y}) \propto \begin{cases} (n_{d,-i} + \gamma_1)\frac{n_{c|d,-i}+\gamma_2}{\frac{V_{n_{d,-i}+1}(t_d+1)}{V_{n_{d,-i}+1}(t_d)}\frac{\gamma_2}{m}+n_{d,-i}+\gamma_2 t_d} f(y_i|\boldsymbol{\beta}^*_d, \boldsymbol{\theta}^*_{c|d}, \sigma^2) & \text{if } d = \delta_k \text{ and } c = \rho_k \text{ for some } k \neq i, \\ (n_{d,-i} + \gamma_1)\frac{\frac{V_{n_{d,-i}+1}(t_d+1)}{V_{n_{d,-i}+1}(t_d)}\frac{\gamma_2}{m}}{\frac{V_{n_{d,-i}+1}(t_d+1)}{V_{n_{d,-i}+1}(t_d)}\frac{\gamma_2}{m}+n_{d,-i}+\gamma_2 t_d} f(y_i|\boldsymbol{\beta}^*_d, \boldsymbol{\theta}^*_{c|d}, \sigma^2) & \text{if } d = \delta_k \text{ and } c \neq \rho_k \text{ for some } k \neq i, \\ \frac{V_n(t+1)}{V_n(t)}\frac{\gamma_1}{m^2} f(y_i|\boldsymbol{\beta}^*_d, \boldsymbol{\theta}^*_{c|d}, \sigma^2) & \text{if } d \neq \delta_k \text{ and } c \neq \rho_k \text{ for all } k \neq i. \end{cases}$$

Updating the rest of the parameters is relatively straightforward. Specifically, for $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}^*$, we sample from:

$$p(\boldsymbol{\beta}^*|\boldsymbol{\delta}, \boldsymbol{\theta}^*, \sigma^2, \boldsymbol{y}) = \prod_{d\in\boldsymbol{\delta}} p(\boldsymbol{\beta}^*_d|\boldsymbol{\delta}, \boldsymbol{\theta}^*, \sigma^2, \boldsymbol{y}) \propto \prod_{d\in\boldsymbol{\delta}} f(\boldsymbol{y}_d|\boldsymbol{\beta}^*_d, \boldsymbol{\theta}^*_d, \sigma^2) p(\boldsymbol{\beta}^*_d),$$
$$p(\boldsymbol{\theta}^*|\boldsymbol{\delta}, \boldsymbol{\beta}^*, \sigma^2, \boldsymbol{y}) = \prod_{d\in\boldsymbol{\delta}}\prod_{c\in\boldsymbol{\rho}_d} p(\boldsymbol{\theta}^*_{c|d}|\boldsymbol{\delta}, \boldsymbol{\beta}^*, \sigma^2, \boldsymbol{y}) \propto \prod_{d\in\boldsymbol{\delta}}\prod_{c\in\boldsymbol{\rho}_d} f(\boldsymbol{y}_{c|d}|\boldsymbol{\beta}^*_d, \boldsymbol{\theta}^*_{c|d}, \sigma^2) p(\boldsymbol{\theta}^*_{c|d}).$$

Under the priors specified above, the full conditional distribution of $\boldsymbol{\beta}^*_d$ is:

$$\boldsymbol{\beta}^*_d|\boldsymbol{\theta}^*_d, \sigma^2, \boldsymbol{y}_d \sim \mathrm{N}(\mu_{\beta_d}, \sigma^2\Omega_{\beta_d}),$$

where $\Omega_{\beta_d} = \left[\sum_{c\in\boldsymbol{\rho}_d}\sum_{i\in R_{c|d}} \boldsymbol{x}_i\boldsymbol{x}_i^T + \sigma^2\Sigma_\beta^{-1}\right]^{-1}$, $\mu_{\beta_d} = \Omega_{\beta_d}\left[\sum_{c\in\boldsymbol{\rho}_d}\sum_{i\in R_{c|d}} \boldsymbol{x}_i(y_i - \boldsymbol{w}_i^T\boldsymbol{\theta}^*_{c|d}) + \sigma^2\Sigma_\beta^{-1}m_\beta\right]$. The full conditional distribution of $\boldsymbol{\theta}^*_{c|d}$ is:

$$\boldsymbol{\theta}^*_{c|d}|\boldsymbol{\beta}^*_d, \sigma^2, \boldsymbol{y}_{c|d} \sim \mathrm{N}(\mu_{\theta_{c|d}}, \sigma^2\Omega_{\theta_{c|d}}),$$

where $\Omega_{\theta_{c|d}} = \left[\sum_{i\in R_{c|d}} \boldsymbol{w}_i\boldsymbol{w}_i^T + \sigma^2\Sigma_\theta^{-1}\right]^{-1}$ and $\mu_{\theta_{c|d}} = \Omega_{\theta_{c|d}}\left[\sum_{i\in R_{c|d}} \boldsymbol{w}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*_d) + \sigma^2\Sigma_\theta^{-1}m_\theta\right]$. Lastly, the full conditional distribution of σ^2 is:

$$\sigma^2|\boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \boldsymbol{y} \sim \text{Inverse Gamma}(a_1, b_1),$$

where $a_1 = \frac{n}{2} + a_0$ and $b_1 = \frac{1}{2}\sum_{d\in\boldsymbol{\delta}}\sum_{c\in\boldsymbol{\rho}_d}\sum_{i\in R_{c|d}} (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*_d - \boldsymbol{w}_i^T\boldsymbol{\theta}^*_{c|d})^2 + b_0$.

To facilitate data analysis within and beyond this study, we develop Julia code and make it publicly available at www.github.com/shuanggema.

3 | SIMULATION

Simulation is conducted to assess the proposed approach and gauge against relevant alternatives. In what follows, we consider a sample with 200 independent subjects generated from model (1). For subject i, $\boldsymbol{x}_i$ follows a multivariate normal distribution with marginal means one and covariance matrix $\Sigma_x = (\rho_{ij})_{p\times p}$. $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iq})^T$ follows a multivariate

normal distribution with marginal means one and covariance matrix $\Sigma_w = (\rho_{ij})_{q\times q}$. We consider four setups, which have different ways of generating $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$. In setup 1, both $\boldsymbol{x}_i$'s and $\boldsymbol{w}_i$'s have an auto-regressive (AR) correlation structure with $\rho_{ij} = \rho^{|i-j|}$ and $\rho = 0.3$. In setup 2, we randomly select $0.3p$ components of $\boldsymbol{x}_i$ and $0.4q$ components of $\boldsymbol{w}_i$ generated in setup 1 and dichotomize at 1. Here we note that histopathological imaging features have continuous distributions. We also simulate binary variables with the consideration that certain clinical/demographic variables (eg, sex) may be categorical, and that the proposed approach can be potentially applied to other data settings that contain categorical components. In setup 3, both $\boldsymbol{x}_i$'s and $\boldsymbol{w}_i$ have a banded correlation (BC) structure, where $\rho_{ij} = \mathbb{1}_{(i=j)} + 0.33\mathbb{1}_{(|i-j|=1)}$. In setup 4, we randomly select $0.3p$ components of $\boldsymbol{x}_i$ and $0.4q$ components of $\boldsymbol{w}_i$ generated in setup 3 and dichotomize at 1. The random errors are independently generated from a normal distribution with mean 0. We consider two variance values: $\sigma^2 = 0.25$ and 0.5.

There are two subgroups, and within each subgroup, there are two sub-subgroups. As such, there are two subgroup specific regression parameters: $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$; and there are four sub-subgroup specific parameters: $\boldsymbol{\theta}_{1|1}^*$, $\boldsymbol{\theta}_{2|1}^*$, $\boldsymbol{\theta}_{1|2}^*$, and $\boldsymbol{\theta}_{2|2}^*$, with the first two coupled with $\boldsymbol{\beta}_1^*$, and the latter two coupled with $\boldsymbol{\beta}_2^*$. We set $(p, q) = (4, 4)$ and $(p, q) = (9, 5)$. We have also examined other p, q values of the same order and made similar observations. Each subgroup has a unique intercept. In addition, the first two and four components of $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ have nonzero coefficients, respectively. The nonzero components of $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are generated from Unif$(-1.2, -0.8)$ and Unif$(0.8, 1.2)$, respectively. And the nonzero components of $\boldsymbol{\theta}_{j|1}^*$ and $\boldsymbol{\theta}_{j|2}^*$, for $j = 1, 2$, are generated from Unif$(-1.2, -0.8)$ and Unif$(0.8, 1.2)$, respectively. For $i = 1, \ldots, n$, $\boldsymbol{\beta}_i$ equals $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ with probabilities p_1 and p_2, respectively, and denote $pr = (p_1, p_2)$. For subjects belonging to the subgroup that corresponds to $\boldsymbol{\beta}_1^*$, $\boldsymbol{\theta}_i$ equals $\boldsymbol{\theta}_{1|1}^*$ and $\boldsymbol{\theta}_{2|1}^*$ with probabilities p_1 and p_2, respectively. Similar settings and notations hold for the other subgroup/sub-subgroups. We set $pr = (0.5, 0.5)$ and $(0.4, 0.6)$. We have also examined highly imbalanced cases and found inferior results, which is as expected (details omitted). For each specific setting, 100 replicates are simulated.

When applying the proposed method, we set Geometric(s) priors with $s = 0.1$ for K and K_d, and $(\gamma_1, \gamma_2, m_\beta, \Sigma_\beta, m_\theta, \Sigma_\theta, a_0, b_0) = (1, 1, \mathbf{0}_p, 10\mathrm{I}_{p+1}, \mathbf{0}_q, 10\mathrm{I}_q, 1, 1)$. Here we set the hyperparameter values as $s = 1$ and $\gamma_1 = \gamma_2 = 1$, reflecting our vague prior knowledge on the number of subgroups and their sizes. For the hyperparameters like Σ_β and Σ_θ that do not have straightforward interpretations and default values, we conduct sensitivity analysis. As shown in Table S4 (Supporting Information), the analysis results are not sensitive to their values. We run the Gibbs sampler described in Section 2.2 for 20,000 iterations, with the first half discarded as burn-in. Convergence of the chain is assessed by inspecting trace plots of individual parameters. For all of our simulation settings/datasets, satisfactory convergence is observed. Computation is affordable. For example, for one simulated dataset with $p = 9$ and $q = 5$, it takes about one minute on a laptop with standard configurations.

To gain more insight into the working characteristics of the proposed approach, we examine the uncertainty in estimation. With one simulated dataset, in Figure S3 (Supporting Information), we show the posterior distributions of the numbers of subgroups and sub-subgroups. It is observed that the values are highly "concentrated" on the true values (of two and four). The uncertainty of grouping can be assessed by the posterior similarity matrix, whose (i, j)th entry represents the posterior probability of subjects i and j belonging to the same (sub-)subgroup. This is graphically presented in Figure S4 (Supporting Information). Note that subjects have been rearranged to improve visualization. The darker (red) color corresponds to a higher posterior probability. Clear (sub-)subgrouping structures are observed. We have also examined a few other plots and made similar observations.

For comparison, we consider the following relevant alternatives. (a) BFMRp: this is a Bayesian MFM approach, has a prior on K, and applies FMR to $y_i \sim \boldsymbol{x}_i + \boldsymbol{w}_i$. It determines the number of subgroups and subgrouping structure in a way similar to the proposed approach. However, it is limited to one-level subgrouping. (b) BFMRf: this is also a Bayesian FMR approach. The difference from the above approach is that the value of K is prespecified and fixed (see below for its candidate values). (c) The one-step FMR approach, denoted as FMR1. This is a frequentist approach and realized using R package `flexmix`. (d) The response-based clustering approach, denoted as Respclust1, first clusters subjects into K subgroups using K-means based on the response variable, and then conducts linear regression with MCP penalization[17] within each subgroup. Here we note that penalization is not necessary with the low dimensionality. We apply it to be consistent with the literature and find that it changes estimation minimally. (e) The residual-based clustering approach, denoted as Resiclust1, first conducts linear regression with MCP penalization under the homogeneity assumption, and then separates subjects into K subgroups based on the residuals using K-means. Linear regression with MCP penalization is then conducted within each subgroup. It is noted that the above five approaches can only generate one-level subgrouping. To better "match" the proposed approach, for the three frequentist approaches, we also consider their two-step versions, which can achieve two levels of grouping. It may also be possible to repeatedly apply the two Bayesian approaches to generate two-level groupings. However, this would involve fixing random variables at for example modes of their

distributions—-this is not very natural and hence not pursued. We additionally consider the following approaches. (f) The two-step FMR approach, denoted as FMR2, conducts FMR with $y_i \sim \boldsymbol{x}_i + \boldsymbol{w}_i$, followed by FMR with $y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_i \sim \boldsymbol{w}_i$ for each subgroup. (g) The two-step response-based clustering approach, denoted as Respclust2, first separates subjects into K subgroups using K-means based on the response variable, and then repeat the same procedure for each subgroup. For estimation with each (sub-)subgroup, linear regression with MCP is conducted. (h) The two-step residual-based clustering approach, denoted as Resiclust2, first conducts linear regression assuming homogeneity and applies MCP penalization. Then subjects are clustered based on the residuals. The same procedure is then repeated for each subgroup. For approaches (b) to (e), the number of subgroups is chosen from $\{2, 3, 4, 5, 6\}$. For approaches (f) to (h), the numbers of subgroups and sub-subgroups are chosen from $\{2, 3\}$. We note that limiting the numbers of (sub-)subgroups close to the true may favor the alternative methods. We note that there exist other alternatives. The above may be "sufficiently relevant" and can be readily realized.

To evaluate the accuracy of identifying subgrouping structures, we report the mean and standard deviation (SD) of $\hat{K}$ (the estimated number of subgroups) and $\sum \hat{K}_d$ (the estimated total number of sub-subgroups). For the proposed approach, such estimates are generated using the marginal posterior modes. For the alternatives, the numbers of subgroups and sub-subgroups are determined data-dependently following published practice. For further examination, we also compute the mean and SD of grouping accuracy for both the subgroups and sub-subgroups. In particular, we adopt the adjusted Rand index (ARI)[18] as the accuracy measure. It quantifies the similarity between two grouping configurations, with a higher value indicating a higher accuracy. When there is a perfect match, ARI takes the value of 1. It can be negative when the similarity between two configurations is less than expected under a random assignment. With the proposed and Bayesian alternative methods, accuracy is computed as the median of the ARIs between the truth and (sub-)subgrouping configuration in each MCMC iteration. With the frequentist alternatives, the calculation is more straightforward. We also evaluate the accuracy of estimating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ using mean squared error (MSE). For the proposed and alternative Bayesian methods, the MSEs for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are calculated as $\sqrt{\frac{\sum_{i=1}^{n}\sum_{m=1}^{M}||\hat{\boldsymbol{\beta}}_i^{(m)}-\boldsymbol{\beta}_i||_2^2}{nMp}}$ and $\sqrt{\frac{\sum_{i=1}^{n}\sum_{m=1}^{M}||\hat{\boldsymbol{\theta}}_i^{(m)}-\boldsymbol{\theta}_i||_2^2}{nMq}}$, where $\hat{\boldsymbol{\beta}}_i^{(m)}$ and $\hat{\boldsymbol{\theta}}_i^{(m)}$ are the mth posterior samples generated by the Gibbs sampler for the ith individual after burn-in, and M is the number of MCMC iterations after burn-in. For the frequentist alternatives, the MSEs of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are calculated as $\sqrt{\frac{\sum_{i=1}^{n}||\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i||_2^2}{np}}$ and $\sqrt{\frac{\sum_{i=1}^{n}||\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i||_2^2}{nq}}$.

Results for the balanced design with $(p, q) = (4, 4)$, $(p, q) = (9, 5)$ and $\sigma^2 = 0.5$ are summarized in Tables 1 and 2, respectively. Results for the rest of the simulation settings are provided in Supporting Information. Across all simulation settings, the proposed approach is observed to have competitive performance. Specifically, it can very accurately estimate the numbers of subgroups and sub-subgroups, while the alternatives tend to over-estimate. Consider for example setup 1 in Table 1. For the (number of subgroups, total number of sub-subgroups) dual, the proposed approach has on average (2.010, 4.040). The five one-step approaches have the estimated numbers of subgroups as 4.160 (BFMRp), 2.560 (BFMRf), 4.830 (FMR1), 2.860 (Resiclust1), and 2.890 (Respclust1). The three two-step alternatives have (2.970, 6.720) for FMR2, (2.440, 5.570) for Resiclust2, and (2.440, 5.630) for Respclust2. The proposed approach is also observed to have competitive ARI values. For the same specific setting, it has a significantly higher ARI value for subgrouping. For sub-subgrouping, its ARI value is slightly higher than that of BFMRp and FMR1 but much higher than the other alternatives. The proposed approach has more accurate estimation. For example, for this specific setting and for estimating $\boldsymbol{\beta}$, the average MSE values are 0.511 (proposed), 0.754 (BFMRp), 0.960 (BFMRf), 0.633 (FMR1), 0.784 (FMR2), 1.292 (Resiclust1), 1.226 (Resiclust2), 1.279 (Respclust1), and 1.216 (Respclust2).

4 | DATA ANALYSIS

TCGA (The Cancer Genome Atlas) provides one of the most comprehensive and extensively analyzed data sources. High-quality data has been published for multiple cancers. TCGA clinical and molecular data has been analyzed in a large number of publications, and there has been a growing interest in its histopathological imaging data. Imaging-based studies have been conducted by Azuaje et al,[19] Jain and Massoud,[20] Xu et al,[21] and others, which have convincingly demonstrated the effectiveness of imaging features for modeling cancer outcomes/phenotypes. In this article, we analyze data on lung adenocarcinoma (LUAD), which is the most common histological subtype of lung cancer. The differences across lung cancer patients have been studied in many publications including a few that analyze histopathological imaging data.[4] On the

TABLE 1 Simulation results with $(p, q, \sigma^2) = (4, 4, 0.5)$: mean(SD) based on 100 replicates

		β			θ		
	$pr = (0.5, 0.5)$	$\hat{K}$	ARI	MSE	$\sum \hat{K}_d$	ARI	MSE
Setup 1	Proposed	2.010(0.10)	0.654(0.05)	0.511(0.04)	4.040(0.20)	0.689(0.04)	0.644(0.06)
	BFMRp	4.160(0.37)	0.324(0.03)	0.754(0.10)		0.680(0.04)	0.756(0.09)
	BFMRf	2.560(0.77)	0.161(0.06)	0.960(0.15)		0.307(0.14)	0.976(0.12)
	FMR1	4.830(0.62)	0.315(0.05)	0.633(0.17)		0.669(0.07)	0.695(0.18)
	FMR2	2.970(0.17)	0.211(0.05)	0.784(0.12)	6.720(0.98)	0.403(0.08)	0.939(0.17)
	Resiclust1	2.860(1.10)	0.122(0.05)	1.292(0.29)		0.313(0.07)	0.884(0.08)
	Resiclust2	2.440(0.50)	0.128(0.04)	1.226(0.28)	5.570(1.57)	0.361(0.09)	0.846(0.07)
	Respclust1	2.890(1.12)	0.121(0.04)	1.279(0.28)		0.313(0.07)	0.876(0.06)
	Respclust2	2.440(0.50)	0.127(0.04)	1.216(0.28)	5.630(1.57)	0.276(0.05)	0.905(0.08)
Setup 2	Proposed	2.010(0.10)	0.622(0.06)	0.548(0.05)	4.010(0.10)	0.647(0.05)	0.713(0.05)
	BFMRp	4.280(0.45)	0.308(0.04)	0.874(0.12)		0.631(0.05)	0.890(0.11)
	BFMRf	2.660(0.93)	0.155(0.07)	0.999(0.15)		0.287(0.14)	1.048(0.13)
	FMR1	4.910(0.68)	0.292(0.06)	0.769(0.25)		0.616(0.10)	0.833(0.24)
	FMR2	2.970(0.17)	0.202(0.06)	0.812(0.16)	6.780(0.97)	0.376(0.09)	1.024(0.21)
	Resiclust1	2.780(1.05)	0.134(0.06)	1.182(0.27)		0.309(0.07)	0.898(0.08)
	Resiclust2	2.400(0.49)	0.140(0.05)	1.131(0.22)	5.290(1.55)	0.367(0.09)	0.864(0.10)
	Respclust1	2.830(1.07)	0.133(0.05)	1.182(0.27)		0.310(0.07)	0.893(0.06)
	Respclust2	2.410(0.49)	0.141(0.05)	1.131(0.24)	5.510(1.43)	0.277(0.05)	0.903(0.06)
Setup 3	Proposed	2.030(0.17)	0.671(0.06)	0.506(0.04)	4.050(0.22)	0.695(0.04)	0.641(0.06)
	BFMRp	4.250(0.48)	0.332(0.03)	0.783(0.13)		0.686(0.05)	0.775(0.12)
	BFMRf	2.570(0.88)	0.166(0.07)	0.984(0.17)		0.308(0.16)	0.982(0.15)
	FMR1	4.830(0.60)	0.323(0.04)	0.645(0.18)		0.681(0.08)	0.686(0.14)
	FMR2	3.000(0.00)	0.214(0.04)	0.796(0.14)	7.000(0.83)	0.404(0.07)	0.923(0.15)
	Resiclust1	2.760(1.20)	0.128(0.05)	1.301(0.27)		0.295(0.06)	0.894(0.07)
	Resiclust2	2.360(0.48)	0.137(0.05)	1.219(0.23)	5.290(1.54)	0.371(0.09)	0.853(0.07)
	Respclust1	2.730(1.15)	0.128(0.05)	1.296(0.26)		0.296(0.06)	0.890(0.07)
	Respclust2	2.360(0.48)	0.137(0.05)	1.228(0.25)	5.320(1.52)	0.287(0.05)	0.894(0.09)
Setup 4	Proposed	2.000(0.00)	0.638(0.05)	0.541(0.05)	4.040(0.20)	0.660(0.04)	0.704(0.05)
	BFMRp	4.240(0.50)	0.316(0.04)	0.858(0.13)		0.648(0.05)	0.875(0.11)
	BFMRf	2.510(0.76)	0.146(0.07)	1.037(0.17)		0.278(0.14)	1.064(0.13)
	FMR1	4.800(0.70)	0.306(0.06)	0.737(0.26)		0.638(0.09)	0.797(0.24)
	FMR2	2.980(0.14)	0.214(0.05)	0.827(0.26)	6.760(0.84)	0.395(0.08)	1.027(0.18)
	Resiclust1	2.840(1.13)	0.136(0.06)	1.208(0.24)		0.309(0.07)	0.904(0.07)
	Resiclust2	2.380(0.49)	0.146(0.06)	1.156(0.22)	5.420(1.57)	0.364(0.08)	0.871(0.07)
	Respclust1	2.840(1.13)	0.135(0.06)	1.217(0.27)		0.311(0.07)	0.906(0.08)
	Respclust2	2.380(0.49)	0.146(0.06)	1.147(0.23)	5.450(1.40)	0.285(0.05)	0.904(0.06)

Note: $\hat{K}\left(\sum \hat{K}_d\right)$: estimated number of subgroups (sub-subgroups).
Abbreviations: ARI, adjusted Rand index; MSE, mean squared error.

TABLE 2 Simulation results with $(p, q, \sigma^2) = (9, 5, 0.5)$: mean(SD) based on 100 replicates

		β			θ		
	$pr = (0.5, 0.5)$	$\hat{K}$	ARI	MSE	$\sum \hat{K}_d$	ARI	MSE
Setup 1	Proposed	1.940(0.24)	0.638(0.16)	0.431(0.12)	3.960(0.49)	0.666(0.13)	0.654(0.19)
	BFMRp	5.440(0.72)	0.240(0.05)	1.416(0.24)		0.532(0.08)	1.330(0.20)
	BFMRf	2.550(1.00)	0.114(0.07)	1.006(0.19)		0.185(0.10)	1.056(0.15)
	FMR1	5.500(0.82)	0.097(0.08)	1.509(0.45)		0.230(0.16)	1.373(0.35)
	FMR2	2.700(0.46)	0.155(0.09)	0.923(0.33)	6.560(1.60)	0.258(0.12)	1.127(0.32)
	Resiclust1	3.230(1.35)	0.123(0.06)	0.949(0.21)		0.300(0.06)	0.807(0.07)
	Resiclust2	2.390(0.49)	0.140(0.06)	0.902(0.20)	5.450(1.59)	0.347(0.08)	0.795(0.11)
	Respclust1	3.270(1.36)	0.121(0.06)	0.956(0.22)		0.305(0.06)	0.799(0.06)
	Respclust2	2.420(0.50)	0.140(0.06)	0.883(0.20)	5.460(1.53)	0.276(0.04)	0.825(0.13)
Setup 2	Proposed	1.930(0.26)	0.591(0.16)	0.461(0.09)	3.970(0.540)	0.620(0.13)	0.776(0.19)
	BFMRp	5.590(0.61)	0.208(0.05)	1.583(0.19)		0.450(0.09)	1.639(0.18)
	BFMRf	2.420(0.93)	0.101(0.08)	1.026(0.19)		0.146(0.09)	1.190(0.16)
	FMR1	5.460(1.02)	0.074(0.06)	1.600(0.43)		0.163(0.12)	1.718(0.39)
	FMR2	2.550(0.50)	0.157(0.09)	0.974(0.35)	6.220(1.52)	0.230(0.11)	1.339(0.38)
	Resiclust1	3.170(1.33)	0.157(0.08)	0.864(0.20)		0.307(0.06)	0.835(0.08)
	Resiclust2	2.420(0.50)	0.174(0.08)	0.804(0.17)	5.530(1.49)	0.347(0.09)	0.807(0.07)
	Respclust1	3.160(1.35)	0.159(0.08)	0.861(0.21)		0.308(0.06)	0.834(0.10)
	Respclust2	2.420(0.50)	0.176(0.07)	0.801(0.17)	5.480(1.50)	0.283(0.05)	0.837(0.11)
Setup 3	Proposed	1.970(0.17)	0.638(0.12)	0.412(0.08)	3.990(0.33)	0.661(0.10)	0.624(0.20)
	BFMRp	5.410(0.65)	0.248(0.04)	1.422(0.21)		0.543(1.32)	0.075(0.18)
	BFMRf	2.290(0.78)	0.110(0.07)	1.011(0.19)		0.173(0.09)	1.071(0.15)
	FMR1	5.450(0.97)	0.101(0.08)	1.506(0.54)		0.236(0.15)	1.385(0.37)
	FMR2	2.680(0.47)	0.149(0.08)	0.953(0.40)	6.400(1.49)	0.250(0.11)	1.141(0.28)
	Resiclust1	2.760(1.12)	0.124(0.05)	0.933(0.19)		0.298(0.07)	0.807(0.06)
	Resiclust2	2.370(0.49)	0.134(0.05)	0.894(0.19)	5.460(1.55)	0.350(0.08)	0.787(0.08)
	Respclust1	2.740(1.11)	0.126(0.04)	0.939(0.30)		0.302(0.07)	0.814(0.15)
	Respclust2	2.370(0.49)	0.134(0.04)	0.890(0.19)	5.360(1.47)	0.274(0.04)	0.810(0.05)
Setup 4	Proposed	1.910(0.29)	0.587(0.18)	0.476(0.12)	4.020(0.64)	0.609(0.14)	0.796(0.27)
	BFMRp	5.640(0.63)	0.207(0.05)	1.634(0.20)		0.448(0.08)	1.665(0.19)
	BFMRf	2.370(0.69)	0.103(0.08)	1.027(0.20)		0.151(0.10)	1.199(0.24)
	FMR1	5.480(0.99)	0.072(0.07)	1.577(0.44)		0.160(0.13)	1.690(0.43)
	FMR2	2.550(0.50)	0.147(0.09)	1.015(0.43)	6.100(1.57)	0.224(0.11)	1.364(0.39)
	Resiclust1	3.050(1.26)	0.154(0.06)	0.829(0.18)		0.316(0.07)	0.824(0.07)
	Resiclust2	2.470(0.50)	0.172(0.06)	0.795(0.17)	5.650(1.56)	0.335(0.08)	0.845(0.29)
	Respclust1	3.040(1.29)	0.154(0.06)	0.820(0.18)		0.311(0.07)	0.819(0.07)
	Respclust2	2.450(0.50)	0.171(0.06)	0.791(0.17)	5.620(1.48)	0.269(0.05)	0.838(0.09)

Note: $\hat{K}(\sum \hat{K}_d)$: estimated number of subgroups (sub-subgroups).
Abbreviations: ARI, adjusted Rand index; MSE, mean squared error.

TABLE 3 Data analysis: estimated coefficients (90% credible intervals)

	Subgroup		
Demographic/clinical variables + Type 1 imaging features	**1**		**2**
Sex	−1.043 (−1.31, −0.76)		0.880 (0.35, 1.39)
Age	0.626 (−0.02, 1.33)		−1.025 (−2.39, 0.21)
tumor_size	1.453 (0.47, 2.46)		−1.917 (−3.87, 0.90)
LymphocytesPN	1.676 (0.40, 2.73)		−0.105 (−1.66, 1.34)
StromaPN	0.661 (−2.25, 3.35)		−2.265 (−4.76, 0.70)
TumorPN	1.965 (0.21, 3.77)		−0.045 (−2.14, 1.77)
StromaSN	−0.846 (−2.38, 0.81)		1.844 (0.07, 3.40)
	Sub-subgroup		
Type 2 imaging features	**1-1**	**1-2**	**2-1**
AreaShape_Zernike_8_2	0.344 (−0.14, 0.84)	−0.131 (−0.66, 0.42)	2.054 (1.27, 2.78)
Granularity_12_ImageAfterMath	0.144 (−0.53, 0.81)	0.453 (−1.43, 2.27)	−1.693 (−2.74, -0.42)
Texture_Contrast_maskosingray_3_03	2.949 (1.34, 4.45)	−1.093 (−3.94, 2.07)	−0.036 (-2.44, 2.19)
Texture_SumVariance_maskosingray_3_01	−2.557 (−4.15, −0.84)	−0.621 (−3.68, 2.46)	2.389 (-0.03, 4.81)
Threshold_FinalThreshold_Identifyhemasub2	−2.098 (−3.13, −1.04)	−0.605 (−2.00, 0.78)	−1.809 (−3.65, −0.08)

other hand, it is also commonly agreed that additional analysis is needed. Data is directly downloaded from TCGA. The response variable is FEV1, which measures the percentage comparison to a normal value reference range of the volume air that a patient can forcibly exhale from the lungs in one second prebronchodilator. FEV1 has been studied in many publications and shown as an important marker for lung capacity, cancer prognosis, and other outcomes. Exploratory analysis suggests a square root transformation (which is a special case of the Box-Cox transformation). The histogram in Figure S5 (Supporting Information) shows two peaks, suggesting that it may be of interest to conduct mixture modeling—this has been ignored in many published studies.

TCGA whole-slide histopathological images have been captured at 20× or 40× magnification by the Aperio medical scanner. Raw data is in the svs format. The pipelines for extracting the two types of imaging features have been briefly described in Section 1. In Table 3, the four Type 1 imaging features are as follows: LymphocytesPN—Perimeter of lymphocyte cell region/square root of image size, StromaPN—perimeter of stromal cell region/square root of image size, TumorPN—perimeter of tumor cell region/square root of image size, and StromaSN—size of stromal cell region/image size. It is noted that two other features (namely, LymphocytesSN—size of lymphocyte cell region/image size, and TumorSN—size of tumor cell region/image size) have also been extracted. They are removed from analysis because of very high correlations with the above features. As shown in Table 3, for subgrouping, we also consider two demographic variables (sex and age) and one clinical variable (tumor size), all of which have been suggested as having critical implications for lung cancer modeling. Using the automated pipeline described in the lower panel of Figure 1, a total of 229 features can be extracted. However, as acknowledged in the literature,[22] most of these features may not be relevant for the response variable. As such, we conduct a supervised screening with linear regression and select the top five features with the smallest marginal P-values, whose names are also provided in Table 3.

For the proposed approach, we assign a Geometric prior—denoted as Geometric(s)—on K and K_d with $s = 0.5$. This prior assigns a 99% probability to K(or K_d) ≤ 7 – a limited number of (sub-)subgroups is sensible given the limited sample size. An informative inverse gamma prior with $(a_0, b_0) = (20, 2)$ is assigned on σ^2. For the rest of the hyperparmeters, we use the same values as in Section 3: $(\gamma_1, \gamma_2, m_\beta, \Sigma_\beta, m_\theta, \Sigma_\theta) = (1, 1, \mathbf{0}_p, 10\mathrm{I}_{p+1}, \mathbf{0}_q, 10\mathrm{I}_q)$. In Tables S11 and S12 (Supporting Information), we report sensitivity analysis results, which show that the estimation results are fairly robust to the hyperparameter values.

Three independent MCMC runs are performed for 100 000 iterations, with the first half discarded as burn-in. We carefully compare the three chains and observe high agreement, which suggests satisfactory MCMC convergence. As an example, in Figure S7 (Supporting Information), we show the pairwise comparison of the estimates for the

(sub-)subgroup-specific regression coefficients. The final results are obtained by pooling the outputs from the three chains.

The posterior distribution of the number of subgroups suggests that there are most likely two subgroups, with very small probabilities for the other values. On average, these subgroups have sizes 74 and 42, respectively. Furthermore, Figure S8 (Supporting Information) suggests that there are most likely two sub-subgroups within subgroup 1, and they have average sizes 59.9 and 14.1, respectively. In contrast, the second subgroup is highly unlikely to be further split. In Table 3, we provide the posterior means and 90% credible intervals for the regression coefficients. It is observed that different (sub-)subgroups have quite different regression models. Some relatively wide credible intervals are at least partly caused by the limited sample size. Here we note that all inferences are based on the postprocessed MCMC draws to address the label switching problems,[23,24] conditioning on the number of most likely (sub-)subgroups.

Data is also analyzed using the alternative approaches considered in simulation. Their estimation results are presented in Tables S13 to S20 (Supporting Information). In Table S21 (Supporting Information), for both subgrouping and sub-subgrouping, we compute the concordance values between the heterogeneity structures obtained using different methods. It is noted that the Bayesian alternative BFMRp does not identify any subgrouping structure. With BFMRf, the estimates for two subgroups are returned. However, a closer examination suggests that in the MCMC runs, the two subgroups are highly imbalanced, the small subgroup is unstable, and all the subjects are concluded as in one group based on their inclusion probabilities. Overall, it is observed that the proposed approach leads to heterogeneity structures and estimates significantly different from those of the alternatives. We note that our heterogeneity analysis, similar to many published mixture modeling studies, is defined based on a specific biomarker. Our literature search does not suggest a good way of evaluating the validity of such analysis—this is "worsened" by the limited research and understanding of histopathological imaging features. For the (sub-)subgroups obtained using the proposed approach, we briefly compare clinical outcomes and find significant differences—this is "as expected" as FEV1 is a strong biomarker for many clinical outcomes. This along with the simulation results can provide some support to our analysis results.

5 | DISCUSSION

In this article, we have conducted cancer finite mixture modeling using histopathological imaging data, which can provide a viable alternative to molecular and some other types of expensive data. The most significant advancement is that, recognizing that there are two types of imaging processing pipelines and extracted features with a natural order, we have develop a novel approach that can generate sub-subgrouping (based on the second type of imaging features) within subgrouping (based on the first type of imaging features). We have noted that hierarchical grouping structures are common in some other types of analysis such as unsupervised clustering. The proposed approach is built on "familiar" Bayesian techniques—thus having a strong statistical ground—and significantly differs from the existing ones in terms of the analysis objective and procedures. It has been effectively realized using a Gibbs sampling approach, and the accompanying Julia code has been made available to facilitate additional applications. In simulation, the proposed approach has been observed to have satisfactory performance and significantly outperforms several highly relevant competitors. We recognize that there are many other approaches that can realize finite mixture modeling. The adopted alternatives are popular and, equally importantly, can be readily realized (many alternatives do not have public software and are difficult to code). We postpone more extensive comparison to future research. With the TCGA LUAD data, for the association between FEV1 and imaging features+clinical/demographic variables, two subgroups and two sub-subgroups within the first subgroup have been identified. This heterogeneity structure is more refined than those in the literature and may provide additional insight into this important biomarker and lung cancer. The discrepancies in the grouping structures identified by the proposed and alternative approaches may further suggest the necessity of this analysis.

This study can be potentially extended in multiple directions. The proposed analysis pertains to estimation only, and there is a lack of consideration on model complexity. In some data analysis with limited sample sizes, it is not desirable to have too many (sub-)subgroups/"too refined" structures. It may be of interest to further take into account the numbers of subgroups and sub-subgroups. An obvious potential extension is to accommodate high-dimensional variables by imposing sparsity. Conceptually, this can be achieved by imposing spike-and-slab priors to the regression coefficients. Significant computational developments may be needed and are postponed to the future. Objectively evaluating analysis results,

especially with practical data, is desirable but highly challenging. Bayesian analysis may be further challenged as the final results are based on averaging over a large number of iterations (and grouping structures and estimates change across iterations). In addition, the subgrouping is defined based on regression models/coefficients. As such, tightness measures commonly adopted in unsupervised clustering analysis may not be applicable. Biological studies that explore implications of the identified mixture structures and model estimations will be of significant interest to ultimately prove the validity of our analysis. It is also of interest to explore more applications of the hierarchical heterogeneity structure and proposed estimation approach. Examining the proposed analysis suggests that it can be relatively "independent" of histopathological images. As long as there are multiple types of variables with a natural order, the proposed analysis may be conducted.

ACKNOWLEDGEMENTS

We thank the associate editor and two reviewers for their careful review and insightful comments, which have led to a significant improvement of this article. This study was supported by NSF (1916251), NIH (CA241699, CA196530, CA204120), and a Yale Cancer Center Pilot Award.

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/.

ORCID

Shuangge Ma https://orcid.org/0000-0001-9001-4999

REFERENCES

1. Schlattmann P. *Medical Applications of Finite Mixture Models*. New York, NY: Springer; 2009.
2. Khalili A, Chen J. Variable selection in finite mixture of regression models. *J Am Stat Assoc*. 2007;102:1025-1038.
3. McLachlan GJ, Lee SX, Rathnayake SI. Finite mixture models. *Annu Rev Stat Appl*. 2019;6:355-378.
4. Yu K-H, Zhang C, Berry GJ, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun*. 2016;7(1):1-10.
5. 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.
6. 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.
7. 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.
8. Li Q, Wang X, Liang F, Xiao G. A Bayesian mark interaction model for analysis of tumor pathology images. *Ann Appl Stat*. 2019;13(3):1708-1732.
9. Zhang S, Fan Y, Zhong T, Ma S. Histopathological imaging features-versus molecular measurements-based cancer prognosis modeling. *Sci Rep*. 2020;10(1):1-9.
10. Carpenter AE, Jones TR, Lamprecht MR, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol*. 2006;7(10):1-11.
11. Wiesmann V, Franz D, Held C, Münzenmayer C, Palmisano R, Wittenberg T. Review of free software tools for image analysis of fluorescence cell micrographs. *J Microsc*. 2015;257(1):39-53.
12. Ren M, Zhang Q, Zhang S, Zhong T, Huang J, Ma S. Hierarchical cancer heterogeneity analysis based on histopathological imaging features. *Biometrics*. 2021.
13. Städler N, Bühlmann P, Geer S. L_1 penalization for mixture regression models. *TEST*. 2010;19:209-256.
14. Lee K-J, Chen R-B, Wu YN. Bayesian variable selection for finite mixture model of linear regressions. *Comput Stat Data Anal*. 2016;95:1-16.
15. Miller JW, Harrison MT. Mixture models with a prior on the number of components. *J Am Stat Assoc*. 2018;113:340-356.
16. Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat*. 2000;9:249-265.
17. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894-942.
18. Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985;2:193-218.
19. Azuaje F, Kim S-Y, Perez HD, Dittmar G. Connecting histopathology imaging and proteomics in kidney cancer through machine learning. *J Clin Med*. 2019;8(10):1535.
20. Jain MS, Massoud TF. Predicting tumour mutational burden from histopathological images using multiscale deep learning. *Nature Mach Intell*. 2020;2(6):356-362.
21. Xu S, Lu Z, Shao W, et al. Integrative analysis of histopathological images and chromatin accessibility data for estrogen receptor-positive breast cancer. *BMC Med Genet*. 2020;13(Suppl 11):1-12.

22. He B, Zhong T, Huang J, Liu Y, Zhang Q, Ma S. Histopathological imaging-based cancer heterogeneity analysis via penalized fusion with model averaging. *Biometrics*. 2020;8(7):1956-1965.
23. Jasra A, Holmes CC, Stephens DA. Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Stat Sci*. 2005;20(1):50-67.
24. Stephens M. Dealing with label switching in mixture models. *J Royal Stat Soc Ser B (Stat Methodol)*. 2000;62:795-809.

SUPPORTING INFORMATION

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

How to cite this article: Im Y, Huang Y, Huang J, Ma S. Bayesian hierarchical finite mixture of regression for histopathological imaging-based cancer data analysis. *Statistics in Medicine*. 2022;41(6):1009-1022. doi: 10.1002/sim.9309