Received: 11 January 2021 | Revised: 1 August 2021 | Accepted: 6 August 2021

DOI: 10.1111/biom.13544

Biometrics
A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY
WILEY

BIOMETRIC METHODOLOGY

Hierarchical cancer heterogeneity analysis based on histopathological imaging features

Mingyang Ren[1] | **Qingzhao Zhang**[2] | **Sanguo Zhang**[1] | **Tingyan Zhong**[3] | **Jian Huang**[4] | **Shuangge Ma**[5]

[1] School of Mathematics Sciences, University of Chinese Academy of Sciences, Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China

[2] MOE Key Laboratory of Economics, Department of Statistics, School of Economics, The Wang Yanan Institute for Studies in Economics and Fujian Key Lab of Statistics, Xiamen University, Xiamen, China

[3] SJTU-Yale Joint Center for Biostatistics, Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

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

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

Correspondence
Shuangge Ma, Department of Biostatistics, Yale School of Public Health, New Haven, CT06510, USA.
Email: shuangge.ma@yale.edu

Funding information
Yale Cancer Center Pilot Award; National Science Foundation, Grant/Award Number: 1916251; Natural Science Foundation of Beijing Municipality, Grant/Award Number: Z190004; National Institutes of Health, Grant/Award Numbers: CA241699, CA196530, CA204120; National Natural Science Foundation of China, Grant/Award Numbers: 12171454, 11971404, 12026604, Basic Scientific Project 71988101; Higher Education Discipline Innovation Project, Grant/Award Number: B13028; Pazhou Lab, Grant/Award Number: 510320

RR Reproducible Research

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues.

Abstract

In cancer research, supervised heterogeneity analysis has important implications. Such analysis has been traditionally based on clinical/demographic/molecular variables. Recently, histopathological imaging features, which are generated as a byproduct of biopsy, have been shown as effective for modeling cancer outcomes, and a handful of supervised heterogeneity analysis has been conducted based on such features. There are two types of histopathological imaging features, which are extracted based on specific biological knowledge and using automated imaging processing software, respectively. Using *both* types of histopathological imaging features, our goal is to conduct the first supervised cancer heterogeneity analysis *that satisfies a hierarchical structure*. That is, the first type of imaging features defines a rough structure, and the second type defines a nested and more refined structure. A penalization approach is developed, which has been motivated by but differs significantly from penalized fusion and sparse group penalization. It has satisfactory statistical and numerical properties. In the analysis of lung adenocarcinoma data, it identifies a heterogeneity structure significantly different from the alternatives and has satisfactory prediction and stability performance.

KEYWORDS
cancer, hierarchy, histopathological imaging, penalization, supervised heterogeneity analysis

1 | INTRODUCTION

Cancer is intrinsically heterogeneous. Heterogeneity analysis has played a pivotal role in cancer research, and can be roughly classified as unsupervised and supervised. The two types of analysis have different implications and utilization (Wiwie et al., 2015). In this study, we conduct supervised analysis, which may be clinically more relevant by taking into account cancer outcomes/phenotypes. In "traditional" supervised cancer heterogeneity analysis, clinical, demographic, and molecular variables have been used. For examples, Xie et al. (2020) conduct supervised heterogeneity analysis of lung carcinoma with gene expression data. Liu and Sunil (2020) perform supervised therapeutic heterogeneity analysis of blood cancer with drug treatment, copy number variation, and genome-wide transcriptional data. An alternative source of data comes from histopathological imaging—a "byproduct" of biopsy, which is routinely ordered for cancer and suspicious patients. Compared to some other types of data, especially molecular data, histopathological imaging data are much more cost-effective and broadly available (Cooper et al., 2015). Histopathological images have been long used for definitive diagnosis and staging. Relatively recently, studies have shown that histopathological imaging features can also be effective for modeling cancer biomarkers, prognosis, and other outcomes (Yu et al., 2016; Luo et al., 2017; Choi and Na, 2018). Among them, a handful have conducted supervised and unsupervised heterogeneity analysis. For example, Luo et al. (2017) examine lung cancer prognosis and show that imaging features associated with overall survival differ significantly across subgroups. This study may be limited by the subgroup information derived from pathological category as opposed to data, and by the simple Cox regression. As another example, Choi and Na (2018) conduct conventional hierarchical clustering analysis of histopathological imaging features, and suggest that immune cell composition in tumor microenvironment is heterogeneous across lung adenocarcinomas and closely associated with tumor metabolism and patient prognosis. Such studies have established the feasibility and effectiveness of heterogeneity analysis for cancer risk, prognosis, and biomarkers based on histopathological imaging features.

Two types of histopathological imaging features have been used in cancer modeling. The first type has been inspired by visual attributes of images traditionally identified by pathologists as important for cancer grading and diagnosis. Such features are at the clinical level and require prior knowledge of human cancer (Gurcan et al., 2009). This type of imaging features has been shown as informative for modeling cancer outcomes, especially including those of lung cancer (Wang et al., 2018). In contrast, features of the second type are captured by computer algorithms in an unsupervised and automated manner. They contain pixel-, object-, and semantic-level information that can go beyond what pathologists can generate via manual examination. These higher resolution features are logical expansions and may improve on the first type of features (Kothari et al., 2013; Aeffner et al., 2019). They have also been analyzed in recent cancer modeling (Sun et al., 2018; He et al., 2020).

In this study, for extracting the first type of imaging features, we adopt the pipeline sketched in the upper panel of Figure 1 (this figure appears in color in the electronic version of this article, and any mention of color refers to that version). Briefly, this pipeline applies ConvPath (Wang et al., 2019), a deep learning cell classification tool, and first identifies tumor, stroma, and lymphocyte cells in images. Based on the cell type information, it extracts six features that describe the perimeter and area of each type of cell. Such features have direct biological interpretations. On the negative side, this pipeline demands first manually labeling images by pathologists and can be time- and labor-intensive. In addition, the number of extracted features is small, limiting the amount of available information. The pipeline for extracting the second type of imaging features is sketched in the lower panel of Figure 1, and involves slide chopping, random selection of subimages, feature extraction using digital imaging processing software (CellProfiler in this study, Carpenter et al. (2006)), and feature averaging. Imaging features generated by this pipeline rely less on input from pathologists and can have a higher dimensionality. They can potentially capture more and higher-resolution information, such as on local anatomical structure and more global patterns of tumor cell and tumor nuclei (Yu et al., 2016). On the negative side, the extracted features do not have direct biological interpretations and may contain noise.

The goal of this study is to conduct supervised cancer heterogeneity analysis based on histopathological imaging features. Significantly different from the existing literature, we aim at utilizing both types of imaging features, which describe information contained in the same images using different techniques and from different angles. The scheme of our analysis is sketched in Figure 2A. With a group of subjects (upper panel), "ordinary" heterogeneity analysis using the first type of imaging features can separate them into subgroups (middle panel, where different colors represent different subgroups—this figure appears in color in the electronic version of this article, and any mention of color refers to that version). Here this type of imaging features is first considered as it has more direct biological interpretations. Then the question we seek to address is: *can a finer sub-subgrouping structure as shown*

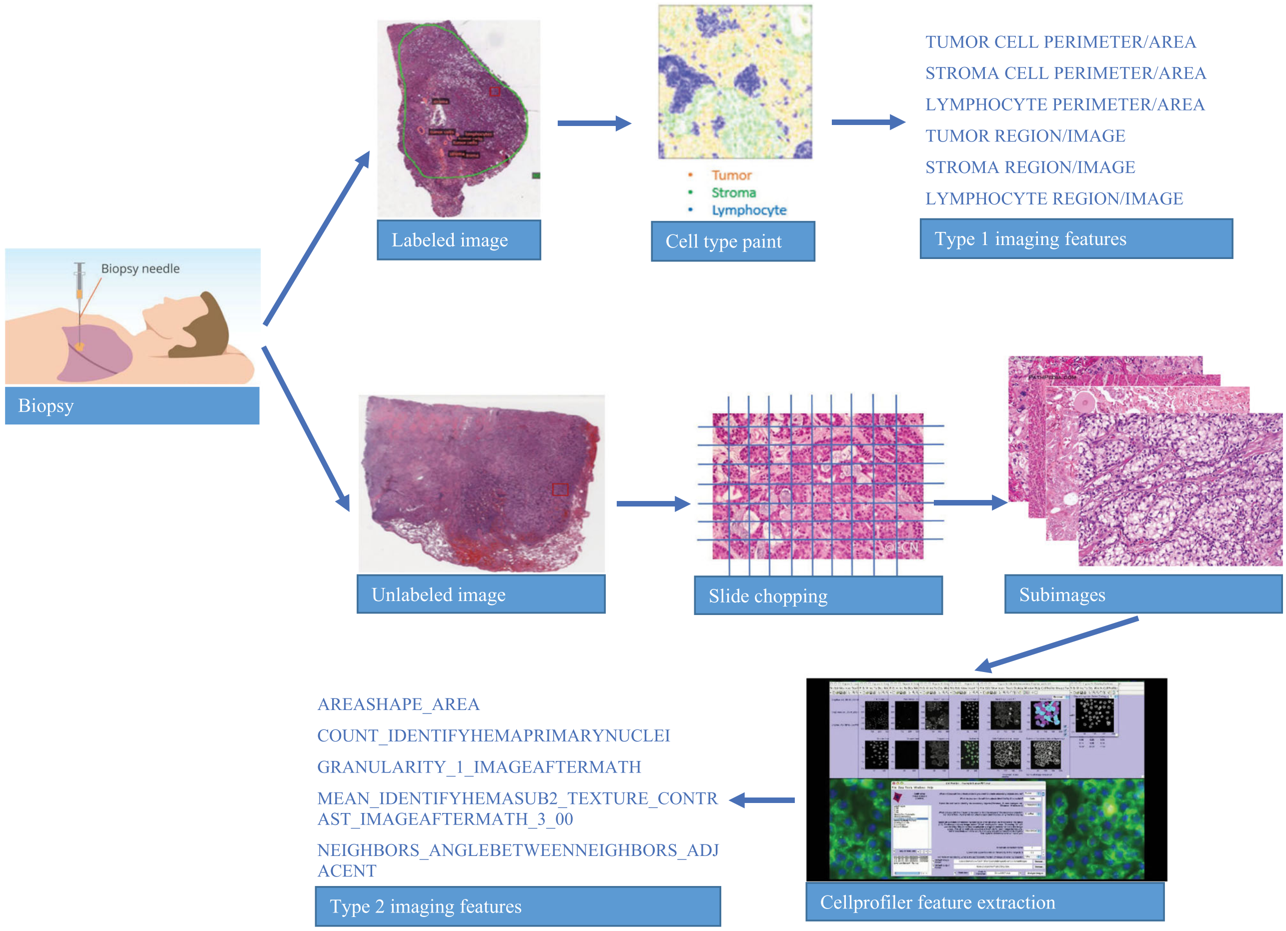
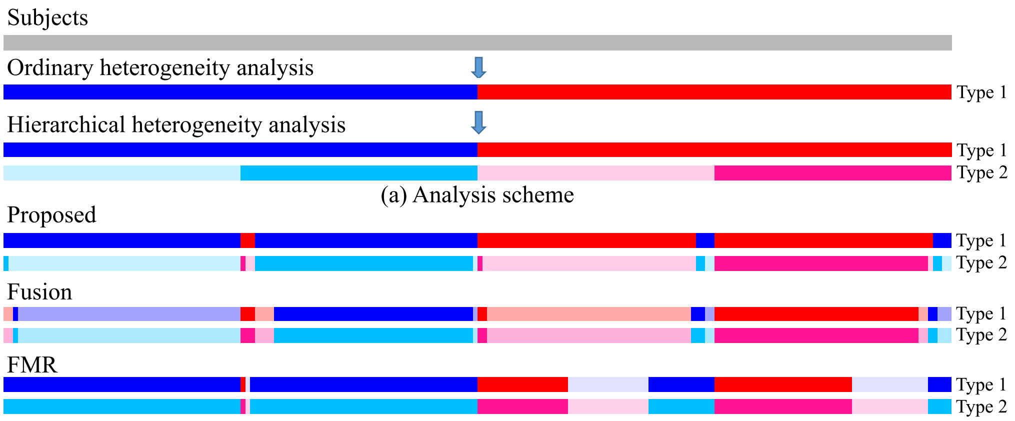

TUMOR CELL PERIMETER/AREA
STROMA CELL PERIMETER/AREA
LYMPHOCYTE PERIMETER/AREA
TUMOR REGION/IMAGE
STROMA REGION/IMAGE
LYMPHOCYTE REGION/IMAGE
Labeled image
Tumor
Stroma
Lymphocyte
Cell type paint
Type 1 imaging features
Biopsy needle
Biopsy
Unlabeled image
Slide chopping
Subimages
AREASHAPE_AREA
COUNT_IDENTIFYHEMAPRIMARYNUCLEI
GRANULARITY_1_IMAGEAFTERMATH
MEAN_IDENTIFYHEMASUB2_TEXTURE_CONTRAST_IMAGEAFTERMATH_3_00
NEIGHBORS_ANGLEBETWEENNEIGHBORS_ADJACENT
Type 2 imaging features
Cellprofiler feature extraction

FIGURE 1 Pipelines for extracting the two types of histopathological imaging features



(a) Analysis scheme

(b) Identified hierarchical structures

FIGURE 2 Analysis scheme and result from one simulated dataset: different colors represent different (sub-)subgroups. Type 1/2: using the two types of imaging features. (A) Upper: ungrouped subjects. Middle: with "ordinary" heterogeneity analysis and Type 1 imaging features, subjects are separated into two subgroups. Lower: with the proposed hierarchical heterogeneity analysis, subjects are separated into two subgroups based on Type 1 imaging features and then further separated into four sub-subgroups based on Type 2 imaging features. The lengths of the color bars are proportional to the sizes of (sub-)subgroups. (B) With one simulation replicate, subgrouping structures are identified by the proposed method and two alternatives (which generate the same subgrouping structures with the two types of imaging features). The same position on different bars represents the same sample

in the lower panel be further obtained? This is sensible as the second type of imaging features has a higher resolution and may contain finer information to identify more subtle sub-subgrouping structures. In the lower panel of Figure 2A, with the natural order of the two types of imaging features, the heterogeneity structure has a hierarchy, which, to the best of our knowledge, has not been studied in the literature.

The idea of obtaining more refined cancer heterogeneity structures with the increased resolution of information/techniques is not new. Take breast cancer as an example. In the past, with information/technique limitations, it was considered as a single disease. With the development of high-throughput profiling and information contained in gene expressions, it was separated into five subtypes: Luminal A, Luminal B, HER2-enriched, Triple-negative, and Claudin-low (Prat et al., 2015). Further advancements in sequencing have suggested that these subtypes may contain finer structures. For example, a recent study (Gong et al., 2021) suggests that the Triple-Negative subtype can be further separated into three sub-subtypes (lipogenic, glycolytic, and mixed). In a sense, this study pursues the "histopathological imaging analogy" of such analysis.

For supervised heterogeneity analysis, there are multiple techniques, with finite mixture of regression (FMR) perhaps being the most popular (Khalili and Chen, 2007; Städler et al., 2010). A relatively more recent technique is penalized fusion (Ma and Huang, 2017; Chen et al., 2020). Compared to FMR and some other techniques, it determines the number (and hence structure) of subgroups in a less ad hoc way. It can in principle accommodate subgroups as small as size one. Its competitive statistical and numerical properties have been well established (Wang et al., 2018; Yang et al., 2019). It is especially noted that penalized fusion has been applied to cancer studies and histopathological imaging data (He et al., 2020). To more intuitively demonstrate the different working characteristics of the proposed and alternative methods, we simulate one data set with the true heterogeneity structure shown in the lower panel of Figure 2A. In Figure 2B, we show the estimated heterogeneity structures using the proposed method, a direct application of penalized fusion, and FMR. The differences are obvious, and the proposed approach can more satisfactorily identify the hierarchical heterogeneity structure.

This study can complement and advance from the existing ones in the following important aspects. First, it advances from the existing heterogeneity analysis by introducing the novel and sensible hierarchical structure, which can provide information on the similarity of subjects at multiple levels. It is noted that this hierarchical heterogeneity structure and hence proposed analysis are also applicable to other settings, as long as an order of variables can be defined. Second, a novel analysis approach is developed. It is built on but significantly advances from penalized fusion and sparse group penalization and can generate the sophisticated heterogeneity structure as sketched in the lower panel of Figure 2A in a single estimation. Third, this study provides an alternative way of conducting cancer heterogeneity analysis, which includes some existing analyses as special cases. Last but not least, it may also provide additional insights into the heterogeneous associations between histopathological imaging features and an important lung cancer biomarker.

2 | METHODS

Consider n independent subjects with measurements $\{y_i, \boldsymbol{x}_i, \boldsymbol{z}_i\}_{i=1}^n$, where $\boldsymbol{x}_i = (x_{i1}, \dots, x_{iq})^\top$ and $\boldsymbol{z}_i = (z_{i1}, \dots, z_{ip})^\top$. Here for subject i, y_i is the response variable, and $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are the first and second type of imaging features, respectively. Consider the heterogeneity model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_i + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_i + \epsilon_i,\ i = 1, \dots, n,$$

where $\boldsymbol{\beta}_i$ and $\boldsymbol{\gamma}_i$ are the q- and p-dimensional vectors of unknown regression coefficients, respectively, and ϵ_i is the random error with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. Intercept is omitted for the simplicity of notation. We consider linear regression for a continuous response, which matches the data analysis in Section 4. Note that the proposed approach is potentially applicable to other types of response/model. Each subject is flexibly modeled to have its own regression coefficients, and two subjects belong to the same (sub)subgroup if and only if they have the same regression model/coefficients.

Significantly advancing from the existing literature, we consider a more sophisticated heterogeneity structure as sketched in the lower panel of Figure 2A, where $\boldsymbol{\beta}_i$'s define a "rough" heterogeneity structure with K_1 subgroups, and $\boldsymbol{\gamma}_i$'s define a more refined heterogeneity structure with K_2 sub-subgroups. Denote $\{\mathcal{G}_1^*, \dots, \mathcal{G}_{K_1}^*\}$ as the collection of subject index sets of the K_1 subgroups, and $\{\mathcal{T}_1^*, \dots, \mathcal{T}_{K_2}^*\}$ as the collection of subject index sets of the K_2 sub-subgroups. The hierarchy of heterogeneity amounts to a nested structure. That is, there exists a mutually exclusive partition of $\{1, \dots, K_2\}$: $\{\mathcal{H}_1, \dots, \mathcal{H}_{K_1}\}$ satisfying $\mathcal{G}_{k_1}^* = \bigcup_{k_2 \in \mathcal{H}_{k_1}} \mathcal{T}_{k_2}^*$, $1 \leqslant k_1 \leqslant K_1, 1 \leqslant k_2 \leqslant K_2$.

2.1 | Penalized estimation

For simultaneous estimation and determination of the heterogeneity structure, we propose the penalized objective function:

$$Q(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}_i - \boldsymbol{z}_i^\top\boldsymbol{\gamma}_i)^2 + \sum_{1\leqslant j<m\leqslant n} p\left(\sqrt{\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_m\|_2^2+\|\boldsymbol{\gamma}_j-\boldsymbol{\gamma}_m\|_2^2},\lambda_1\right) + \sum_{1\leqslant j<m\leqslant n} p\left(\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_m\|_2,\lambda_2\right), \tag{2.1}$$

where $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^\top,\dots,\boldsymbol{\beta}_n^\top)^\top$, $\boldsymbol{\gamma}=(\boldsymbol{\gamma}_1^\top,\dots,\boldsymbol{\gamma}_n^\top)^\top$, and $p(\cdot,\lambda)$ is a concave penalty function with tuning parameter $\lambda > 0$. In our numerical study, we adopt MCP (Minimax Concave Penalty; Zhang (2010)), and note that SCAD (Smoothly Clipped Absolute Deviation Penalty) and some other penalties are also applicable. Consider $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\operatorname{argmin}}\, Q(\boldsymbol{\beta},\boldsymbol{\gamma})$. Denote $\{\widehat{\boldsymbol{\alpha}}_1,\dots,\widehat{\boldsymbol{\alpha}}_{\widehat{K}_1}\}$ and $\{\widehat{\boldsymbol{\delta}}_1,\dots,\widehat{\boldsymbol{\delta}}_{\widehat{K}_2}\}$ as the distinct values of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, respectively. Then $\{\widehat{\mathcal{G}}_1,\dots,\widehat{\mathcal{G}}_{\widehat{K}_1}\}$ and $\{\widehat{\mathcal{T}}_1,\dots,\widehat{\mathcal{T}}_{\widehat{K}_2}\}$ constitute mutually exclusive partitions of $\{1,\dots,n\}$, where $\widehat{\mathcal{G}}_{k_1}=\{i:\widehat{\boldsymbol{\beta}}_i=\widehat{\boldsymbol{\alpha}}_{k_1}, i=1,\dots,n\}$, $k_1=1,\dots,\widehat{K}_1$ and $\widehat{\mathcal{T}}_{k_2}=\{i:\widehat{\boldsymbol{\gamma}}_i=\widehat{\boldsymbol{\delta}}_{k_2}, i=1,\dots,n\}, k_2=1,\dots,\widehat{K}_2$. Collectively, they fully determine the heterogeneity structure.

Rationale In the objective function, the first term measures lack of fit (and can be replaced with the negative log likelihood for another type of data/model). The proposed penalty is built on the penalized fusion technique. For the proposed regression models, under the standard penalized fusion (Ma and Huang, 2017), the penalty is $\sum_{1\leqslant j<m\leqslant n} p(\|(\boldsymbol{\beta}_j^\top,\boldsymbol{\gamma}_j^\top)^\top-(\boldsymbol{\beta}_m^\top,\boldsymbol{\gamma}_m^\top)^\top\|_2)$. It can shrink some differences to exactly zero, and subjects with exactly the same regression coefficients are assigned to the same subgroup. However, using the standard penalized fusion, different types of variables can only produce the same subgrouping structure. Beyond penalized fusion, the proposed penalty also shares some similar spirit with sparse group penalization. That is, a hierarchical penalty is imposed on the two types of variables, achieving a nested subgrouping structure. The resulted estimates have the following possibilities: (a) $\widehat{\boldsymbol{\beta}}_j=\widehat{\boldsymbol{\beta}}_m$ and $\widehat{\boldsymbol{\gamma}}_j=\widehat{\boldsymbol{\gamma}}_m$. Then subjects j and m are assigned to the same subgroup (of the rough heterogeneity structure) and the same sub-subgroup (of the finer heterogeneity structure); (b) $\widehat{\boldsymbol{\beta}}_j\neq\widehat{\boldsymbol{\beta}}_m$ and $\widehat{\boldsymbol{\gamma}}_j\neq\widehat{\boldsymbol{\gamma}}_m$. Then subjects j and m are assigned to different subgroups and different sub-subgroups; and (c) $\widehat{\boldsymbol{\beta}}_j=\widehat{\boldsymbol{\beta}}_m$ but $\widehat{\boldsymbol{\gamma}}_j\neq\widehat{\boldsymbol{\gamma}}_m$. Then subjects j and m are assigned to the same subgroup but different sub-subgroups. Note that the term related to $\boldsymbol{\gamma}_j-\boldsymbol{\gamma}_m$ only appears in the first group penalty. Due to the "all in or all out" property of group penalization, the case with $\widehat{\boldsymbol{\beta}}_j\neq\widehat{\boldsymbol{\beta}}_m$ and $\widehat{\boldsymbol{\gamma}}_j=\widehat{\boldsymbol{\gamma}}_m$ cannot happen. This approach has a mechanism similar to the hierarchical structured sparsity penalty (Bien et al., 2013).

Combining the above scenarios, we see that, if subjects j and m are assigned to the same sub-subgroup, then they must be assigned to the same subgroup. On the other hand, when subjects j and m are assigned to the same subgroup, they can be assigned to the same sub-subgroup or different sub-subgroups. As such, the proposed approach can achieve the nested structure sketched in Figure 2A.

We note that, in some studies, the second type of imaging features is high dimensional and noisy. Directly analyzing such data demands imposing additional sparsity or low-rank (Jing et al., 2018) constraints. We have very briefly explored such possibilities and found that they make computation intractable and also introduce tremendous theoretical challenges. In this study, we focus on the scenario where the second type of imaging features is low dimensional and noise is not of a serious concern. If needed, this can be achieved with prescreening, which is a popular technique and adopted in our data analysis.

2.2 | Statistical properties

Denote the true values of parameters as $\boldsymbol{\beta}^*=(\boldsymbol{\beta}_1^{*\top},\dots,\boldsymbol{\beta}_n^{*\top})^\top$ and $\boldsymbol{\gamma}^*=(\boldsymbol{\gamma}_1^{*\top},\dots,\boldsymbol{\gamma}_n^{*\top})^\top$. Define the minimal differences of the common values between two (sub-)subgroups as

$$b_1 = \min_{j\in\mathcal{T}_k^*, m\in\mathcal{T}_{k'}^*, 1\leqslant k\neq k'\leqslant K_2} \|\boldsymbol{\gamma}_j^*-\boldsymbol{\gamma}_m^*\|_2,\ b_2 = \min_{j\in\mathcal{G}_k^*, m\in\mathcal{G}_{k'}^*, 1\leqslant k\neq k'\leqslant K_1} \|\boldsymbol{\beta}_j^*-\boldsymbol{\beta}_m^*\|_2.$$

Denote $|\mathcal{T}_{\min}| = \min_{1\leqslant k_2\leqslant K_2} |\mathcal{T}_{k_2}^*|$, where $|\cdot|$ is the cardinality of the set. Define the 2-norm of the q-dimensional vector $\boldsymbol{d}=(d_1,\dots,d_q)^\top$ as $\|\boldsymbol{d}\|_2=\sqrt{\sum_{j=1}^q d_j^2}$. Let $\psi_{\min}(\boldsymbol{A})$ denote the smallest eigenvalues of matrix $\boldsymbol{A}$. Mild assumptions are made and described in the Supporting Information, with which we can establish the following consistency properties.

Theorem 1. *Suppose that Conditions A.1–A.3 (Supplementary Information) hold. Assume that* $|\mathcal{T}_{\min}| \gg \sqrt{(K_1 q + K_2 p)n\log n}$, $b_1 > (a+0.5)\lambda_1$, $b_2 > (a+0.5)\lambda_2$, *where a is the regularization parameter of the concave penalty defined in Condition A.1, and* $\min\{\lambda_1,\lambda_2\} \gg |\mathcal{T}_{\min}|^{-1}\sqrt{(K_1 q + K_2 p)n\log n}$. *Then we have:*

(1) (Consistency of identifying the number of (sub-)subgroups) $\mathrm{pr}(\widehat{K}_1=K_1,\widehat{K}_2=K_2)\rightarrow 1$.

(2) *(Consistency of subgrouping)* $\mathrm{pr}(\widehat{\mathcal{G}}_{k_1} = \mathcal{G}^*_{k_1},$ *for* $k_1 = 1, \ldots, K_1) \to 1$ *and* $\mathrm{pr}(\widehat{\mathcal{T}}_{k_2} = \mathcal{T}^*_{k_2},$ *for* $k_2 = 1, \ldots, K_2) \to 1.$

(3) *(Rate of convergence) There exists a local minimum of* (2.1) *that satisfies:*

$$\sup_i \|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}^*_i\|_2 + \sup_i \|\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}^*_i\|_2$$
$$= O_p\left(|\mathcal{T}_{\min}|^{-1}\sqrt{(K_1 q + K_2 p) n \log n}\right).$$

This theorem shows that the proposed approach has much-desired consistency properties. In particular, it can accurately estimate the number of (sub-)subgroups and recover the subgrouping structures with a high probability. Such results, particularly the consistency of the number of (sub-)subgroups, can be extremely difficult to establish with some other heterogeneity analysis techniques. This approach also has the "ordinary" estimation consistency. With the special design of the penalty, it easily follows that the heterogeneity hierarchy holds. Proof is provided in the Supporting Information. We note that although certain steps have roots in Ma and Huang (2017), Yang et al. (2019) and other published studies, as the penalty is significantly different, the theoretical development is nontrivial. For example, more complex consistency conditions on subgrouping recovery theory need to be established under the hierarchical nested penalty.

2.3 | Computation

We derive an ADMM (Alternating Direction Method of Multipliers) algorithm for optimizing (2.1). Some developments are specific to MCP, and optimizing with other penalties demands minor modifications. The objective function can be reformulated as

$$\mathcal{L}_0(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{v}, \boldsymbol{\omega}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_i - \boldsymbol{z}_i^\top \boldsymbol{\gamma}_i)^2$$
$$+ \sum_{1 \leqslant j < m \leqslant n} p\left(\sqrt{\|\boldsymbol{v}_{jm}\|_2^2 + \|\boldsymbol{\omega}_{jm}\|_2^2}, \lambda_1\right)$$
$$+ \sum_{1 \leqslant j < m \leqslant n} p\left(\|\boldsymbol{v}_{jm}\|_2, \lambda_2\right),$$

subject to $\boldsymbol{v}_{jm} = \boldsymbol{\beta}_j - \boldsymbol{\beta}_m$, $\boldsymbol{\omega}_{jm} = \boldsymbol{\gamma}_j - \boldsymbol{\gamma}_m$, where $\boldsymbol{v} = \{\boldsymbol{v}_{jm}^\top, j < m\}^\top$ and $\boldsymbol{\omega} = \{\boldsymbol{\omega}_{jm}^\top, j < m\}^\top$. By the augmented Lagrangian method, the estimates can be obtained by minimizing:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{v}, \boldsymbol{\omega}, \boldsymbol{\xi}, \boldsymbol{\eta}) = \mathcal{L}_0(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{v}, \boldsymbol{\omega})$$
$$+ \sum_{j<m} \boldsymbol{\xi}_{jm}^\top(\boldsymbol{\beta}_j - \boldsymbol{\beta}_m - \boldsymbol{v}_{jm})$$
$$+ \frac{\kappa}{2}\sum_{j<m}\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m - \boldsymbol{v}_{jm}\|_2^2$$
$$+ \sum_{j<m} \boldsymbol{\eta}_{jm}^\top(\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_m - \boldsymbol{\omega}_{jm})$$
$$+ \frac{\kappa}{2}\sum_{j<m}\|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_m - \boldsymbol{\omega}_{jm}\|_2^2,$$

where the dual variables $\boldsymbol{\xi} = \{\boldsymbol{\xi}_{jm}^\top, j < m\}^\top$ and $\boldsymbol{\eta} = \{\boldsymbol{\eta}_{jm}^\top, j < m\}^\top$ are the Lagrange multipliers, $\boldsymbol{\xi}_{jm}$ and $\boldsymbol{\eta}_{jm}$ are q- and p-dimensional vectors, and κ is a penalty parameter.

Overall, we propose an iterative algorithm, with the FMR estimates as the initial for $(\boldsymbol{\beta}^{(0)\top}, \boldsymbol{\gamma}^{(0)\top})^\top$, $\boldsymbol{v}_{jm}^{(0)} = \boldsymbol{\beta}_j^{(0)} - \boldsymbol{\beta}_m^{(0)}$, $\boldsymbol{\omega}_{jm}^{(0)} = \boldsymbol{\gamma}_j^{(0)} - \boldsymbol{\gamma}_m^{(0)}$, $\boldsymbol{\xi}^{(0)} = \mathbf{0}$, and $\boldsymbol{\eta}^{(0)} = \mathbf{0}$. As iteration $t+1$, the updates are

$$(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}) = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\mathrm{argmin}}\,\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{v}^{(t)}, \boldsymbol{\omega}^{(t)}, \boldsymbol{\xi}^{(t)}, \boldsymbol{\eta}^{(t)}), \quad (2.2)$$

$$(\boldsymbol{v}^{(t+1)}, \boldsymbol{\omega}^{(t+1)}) = \underset{\boldsymbol{v},\boldsymbol{\omega}}{\mathrm{argmin}}\,\mathcal{L}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}, \boldsymbol{v}, \boldsymbol{\omega}, \boldsymbol{\xi}^{(t)}, \boldsymbol{\eta}^{(t)}), \quad (2.3)$$

$$\boldsymbol{\xi}_{jm}^{(t+1)} = \boldsymbol{\xi}_{jm}^{(t)} + \kappa(\boldsymbol{\beta}_j^{(t+1)} - \boldsymbol{\beta}_m^{(t+1)} - \boldsymbol{v}_{jm}^{(t+1)}), \quad (2.4)$$

$$\boldsymbol{\eta}_{jm}^{(t+1)} = \boldsymbol{\eta}_{jm}^{(t)} + \kappa(\boldsymbol{\gamma}_j^{(t+1)} - \boldsymbol{\gamma}_m^{(t+1)} - \boldsymbol{\omega}_{jm}^{(t+1)}). \quad (2.5)$$

Details of (2.2) and (2.3) are provided in the Supporting Information. These updates are repeated until convergence, which is concluded when the difference between two consecutive estimates is smaller than a predefined cutoff. Convergence of this algorithm follows from Ma and Huang (2017) and is achieved in all of our numerical analyses. This algorithm adopts ADMM, which is similar to Ma and Huang (2017). However, it significantly differs in solving (2.3) by involving the complex hierarchical groupwise thresholding operator. Denote $C_n = \log[n(p+q)]$. To select (λ_1, λ_2), following Ma and Huang (2017), we conduct a grid search and minimize the modified BIC-type criterion:

$$\mathrm{BIC}(\lambda_1, \lambda_2) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\{y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_i(\lambda_1, \lambda_2) - \boldsymbol{z}_i^\top \widehat{\boldsymbol{\gamma}}_i(\lambda_1, \lambda_2)\}^2\right]$$

$$+C_n \frac{\log n}{n}\left\{\widehat{K}_1(\lambda_1,\lambda_2)q + \widehat{K}_2(\lambda_1,\lambda_2)p\right\}.$$

Here it is noted that, although this criterion has been shown as numerically satisfactory, its theoretical properties, for example, uniqueness of optimizer, remains unestablished.

3 | SIMULATION

For $n = 120$ independent subjects, we generate $\boldsymbol{x}_i$'s from a q-dimensional normal distribution $N(0, \Sigma_1)$, where $\Sigma_1 = (\sigma_{1jm})_{1\leqslant j,m\leqslant p}$ with $\sigma_{1jm} = 1$ if $j = m$, and =0.2 otherwise. Then we randomly select $0.4q$ components and dichotomize at zero, leading to $0.6q$ continuous and $0.4q$ binary components. We simulate $\boldsymbol{z}_i$'s from a p-dimensional normal distribution $N(0, \Sigma_2)$. Two structures of the covariance matrix $\Sigma = (\sigma_{2jm})_{1\leqslant j,m\leqslant p}$ are considered. The first structure is auto-regressive (AR) correlation given by $\sigma_{2jm} = \rho^{|j-m|}$ with $\rho = 0.25$ and 0.75 (denoted as AR1 and AR2, respectively). The second structure is banded correlation. Here two scenarios are considered and denoted as B1 and B2 respectively. Under B1, $\sigma_{2jm} = I(j = m) + 0.3I(|j - m| = 1)$. Under B2, $\sigma_{2jm} = I(j = m) + 0.6I(|j - m| = 1) + 0.3I(|j - m| = 2)$. Then we randomly select $0.4p$ components and dichotomize at zero. For the correlation between $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, we consider the uncorrelated and weakly correlated structures, under which the correlation coefficients between the jth component of $\boldsymbol{x}_i$ and the mth component of $\boldsymbol{z}_i$ are 0 and $0.05\min\{\min_{1\leqslant j,m\leqslant p}(\sigma_{1jm}), \min_{1\leqslant j,m\leqslant p}(\sigma_{2jm})\}$, respectively. We note that the imaging features analyzed in the next section are all continuous. Here we generate covariates that also have binary components to be more general, as the proposed approach is also applicable to other data scenarios that may have discrete components. We have experimented with all covariates being continuous and observed similar patterns as described below.

There are $K_1 = 2$ subgroups with $\boldsymbol{\beta}_i$ equal to $(\mu, \mu, \mathbf{0}_{q-2})$ and $(-\mu, -\mu, \mathbf{0}_{q-2})$, respectively. There are $K_2 = 4$ sub-subgroups with $\boldsymbol{\gamma}_i$ equal to $(1.5\mu, 1.5\mu, 1.5\mu, 1.5\mu, \mathbf{0}_{p-4})$, $(0.5\mu, 0.5\mu, 0.5\mu, 0.5\mu, \mathbf{0}_{p-4})$, $(-0.5\mu, -0.5\mu, -0.5\mu, -0.5\mu, \mathbf{0}_{p-4})$, and $(-1.5\mu, -1.5\mu, -1.5\mu, -1.5\mu, \mathbf{0}_{p-4})$, respectively. We set $q = 5$ and $p = 10$ and consider two signal levels with $\mu = 2$ and 0.6. For the proportions of subjects in the sub-subgroups, we consider (1/4,1/4,1/4,1/4) and (1/6,1/6,1/3,1/3), which represent balanced and imbalanced designs. The random errors are generated from $N(0, 0.5)$. For each simulation setting, we generate 100 replicates.

To gauge performance of the proposed approach, we consider the following competitors: (a) the "ordinary" penalized fusion approach (denoted as fusion1) as developed in Ma and Huang (2017); (b) the one-step FMR approach (denoted as FMR1), which applies the FMR technique to $y_i \sim \boldsymbol{x}_i + \boldsymbol{z}_i, i = 1, \ldots, n$; (c) the one-step response-based clustering approach (respclust1), which first clusters the subjects based on the response variable, and then applies "linear regression + MCP" to each subgroup. Here we note that penalization is not necessary. It is applied to be more coherent with the proposed penalized estimation; (d) the one-step residual-based clustering (resiclust1), which applies "linear regression + MCP" under the homogeneity assumption $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\gamma} + \epsilon_i$, groups subjects based on the residuals $y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{z}_i^\top \widehat{\boldsymbol{\gamma}}$, and then applies "linear regression + MCP" again to each subgroup. The above approaches can only generate one-level heterogeneity structures; (e) the two-step FMR approach (FMR2), which first applies FMR to $y_i \sim \boldsymbol{x}_i + \boldsymbol{z}_i, i = 1, \ldots, n$, and then applies FMR again to $y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_i \sim \boldsymbol{z}_i$ within each subgroup. By repeatedly applying the FMR technique, this approach can generate two-level heterogeneity structures, which can be more comparable to the proposed approach; (f) the two-step response-based clustering (respclust2), which first clusters all subjects based on the response variable, and then clusters the subjects within each subgroup using the same technique; and (g) the two-step residual-based clustering (resiclust2), which is similar to (f) but is based on residuals as in (d). For fusion1, the number of subgroups is determined in a similar way as for the proposed approach. For approaches (b), (c), and (d), the number of subgroups is selected from $\{2, 3, 4, 5, 6\}$. For approaches (e), (f), and (g), the number of (sub)-subgroups is selected from $\{2, 3\}$. It is noted that, comparatively, the proposed approach can much more easily identify the number of (sub)-subgroups. To the best of our knowledge, there is no existing approach that directly generates heterogeneity structures similar to the proposed one.

The following measures are adopted to evaluate performance: (a) percentage of $\widehat{K}_{1,2}$ equal to $K_{1,2}$ (denoted by "per"), (b) mean and sd of $\widehat{K}_{1,2}$, (c) subgrouping consistency defined by

$$\mathrm{SC}(\widehat{\varphi}, \varphi) : = \binom{n}{2}^{-1} \left| \left\{(i,j) : I\left(\widehat{\varphi}(\boldsymbol{x}_i, \boldsymbol{z}_i) = \widehat{\varphi}\left(\boldsymbol{x}_j, \boldsymbol{z}_j\right)\right) \right.\right.$$
$$\left.\left. = I\left(\varphi(\boldsymbol{x}_i, \boldsymbol{z}_i) = \varphi\left(\boldsymbol{x}_j, \boldsymbol{z}_j\right)\right); i < j\right\} \right|,$$

where $\widehat{\varphi}$ and φ are the estimated and true (sub)subgrouping memberships, respectively. This is essentially the Rand index (Rand, 1971), which has been commonly adopted to measure clustering accuracy. In particular, it measures the percentage of correct decisions (if pairs of subjects in different subgroups are assigned to different subgroups, and pairs from the same subgroup are assigned to the same subgroup) out of a total of $\binom{n}{2}$ possible decisions. It is calculated for the two heterogeneity levels separately. We also consider: (d) mean squared error (MSE) for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$: $\mathrm{MSE}(\boldsymbol{\beta}) = \sqrt{\frac{\sum_{i=1}^{n} \|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|_2^2}{nq}}$ and $\mathrm{MSE}(\boldsymbol{\gamma}) = \sqrt{\frac{\sum_{i=1}^{n} \|\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\|_2^2}{np}}$.

Penalized fusion is usually computationally expensive. However, simulation suggests that that proposed analysis is still feasible. For analyzing one replicate, it takes about 15 minutes on a laptop with regular configurations. To gain more insight into the operating characteristics of the proposed approach, in the Supporting Information, we show the modified BIC criterion (Figure C1), estimated K_1 and K_2 (Figure C2), and estimates for one component of $\boldsymbol{\beta}$ and one component of $\boldsymbol{\gamma}$ (Figure C3), as a function of λ_1 and λ_2, for one simulated data set. In Figure C1, there is one unique optimizer. The paths are "well-behaved" and similar to those of other penalized estimations.

Summary results for the scenario with correlated $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, balanced design, and $\mu = 2$ are provided in Figure 3. The rest of the results are provided in Figures C4– C10 (Supporting Information). The corresponding numerical values are provided in Tables C1– C8 (Supporting Information). Across the whole spectrum of simulation, the proposed approach is observed to have highly competitive performance. Consider, for example, Figure 3 and the scenario with correlated $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, balanced design, $\mu = 2$, and AR1 correlation. For the heterogeneity defined associated with $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, respectively, the proposed approach has per values 0.88 and 0.80, suggesting high accuracy. In contrast, all the alternatives fail in this aspect, in particular with respect to $\boldsymbol{z}_i$. Overall, the proposed approach also has high SC values, although it is noted that for some scenarios, it not necessarily has the highest values (but very close). It is also observed to have high estimation accuracy. For example, for estimating $\boldsymbol{\beta}$, the MSE values are 0.970 (fusion1), 1.430 (FMR1), 1.617 (repclust1), 1.860 (resiclust1), 1.423 (FMR2), 1.157 (respclust2), 1.310 (resiclust2), and 0.942 (proposed).

3.1 | Additional exploration

Although the heterogeneity hierarchy is biologically and statistically sensible, it may still be of interest to explore what may happen when it is violated. In the middle panel of Figure C11 (Supporting Information), we sketch the setting where the heterogeneity structures defined by $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ violate the hierarchy. Examining the proposed penalty and theoretical development suggests that our approach will generate a heterogeneity structure as represented in the lower panel of Figure C11. However, the estimation consistency may still hold. To explore this numerically, we simulate data with $K_1 = 2$ subgroups based on $\boldsymbol{x}$ and $K_2 = 3$ sub-subgroups based on $\boldsymbol{z}$. Overall, four subject groups can be obtained with the combination of these two. The regression coefficients are $(2, 2, \mathbf{0}_{q-2}, 3, 3, 3, 3, \mathbf{0}_{p-4})$, $(2, 2, \mathbf{0}_{p+q-2})$, $(-2, -2, \mathbf{0}_{p+q-2})$, and $(-2, -2, \mathbf{0}_{q-2}, -3, -3, -3, -3, \mathbf{0}_{p-4})$. Other settings are the same as those described above. The results are summarized in Figure C12 and Table C9 (Supporting Information). As expected, the subgrouping accuracy is unsatisfactory. However, the proposed approach still has competitive estimation performance. In practical data analysis, the similarity of regression coefficients of different sub-subgroups may hint a violation of the hierarchy assumption.

4 | DATA ANALYSIS

The Cancer Genome Atlas (TCGA) is a collective effort organized by the NIH and has published high-quality clinical, molecular, and imaging data on multiple cancer types. Compared to clinical and molecular data, the TCGA histopathological imaging data has been less analyzed. Studies based on the TCGA imaging data include Yu et al. (2017), Sun et al. (2018), Wang et al. (2018), and others, which adopt somewhat simple analysis methods. Here we analyze data on lung adenocarcinoma (LUAD), which are publicly available at the TCGA data portal (TCGA, 2021). The heterogeneity of lung cancer and possibility of studying such heterogeneity based on imaging features have been established in Luo et al. (2017) and other published studies. The response variable is FEV1, which is the percentage comparison to a normal value reference range of the volume of air that a patient can forcibly exhale from the lungs in one second pre-bronchodilator. It is an important indicator of lung function and an established biomarker for lung cancer prognosis and other outcomes. It is correlated with its post-bronchodilator counterpart, which is also an important biomarker for lung cancer capacity and prognosis. It is noted that the post-bronchodilator measure is often not measured in population studies (Mannino et al., 2011; Balte et al., 2020). The histogram in Figure C13 (Supporting Information) suggests a mixture distribution (heterogeneity).

In the TCGA study, whole-slide histopathological images and tissue images are captured at 20* or 40* magnification by the Aperio medical scanner and saved in the svs format. The pipeline for extracting the first type of imaging features has been briefly described in Section 1 and Figure 1. We refer to Wang et al. (2019) for more

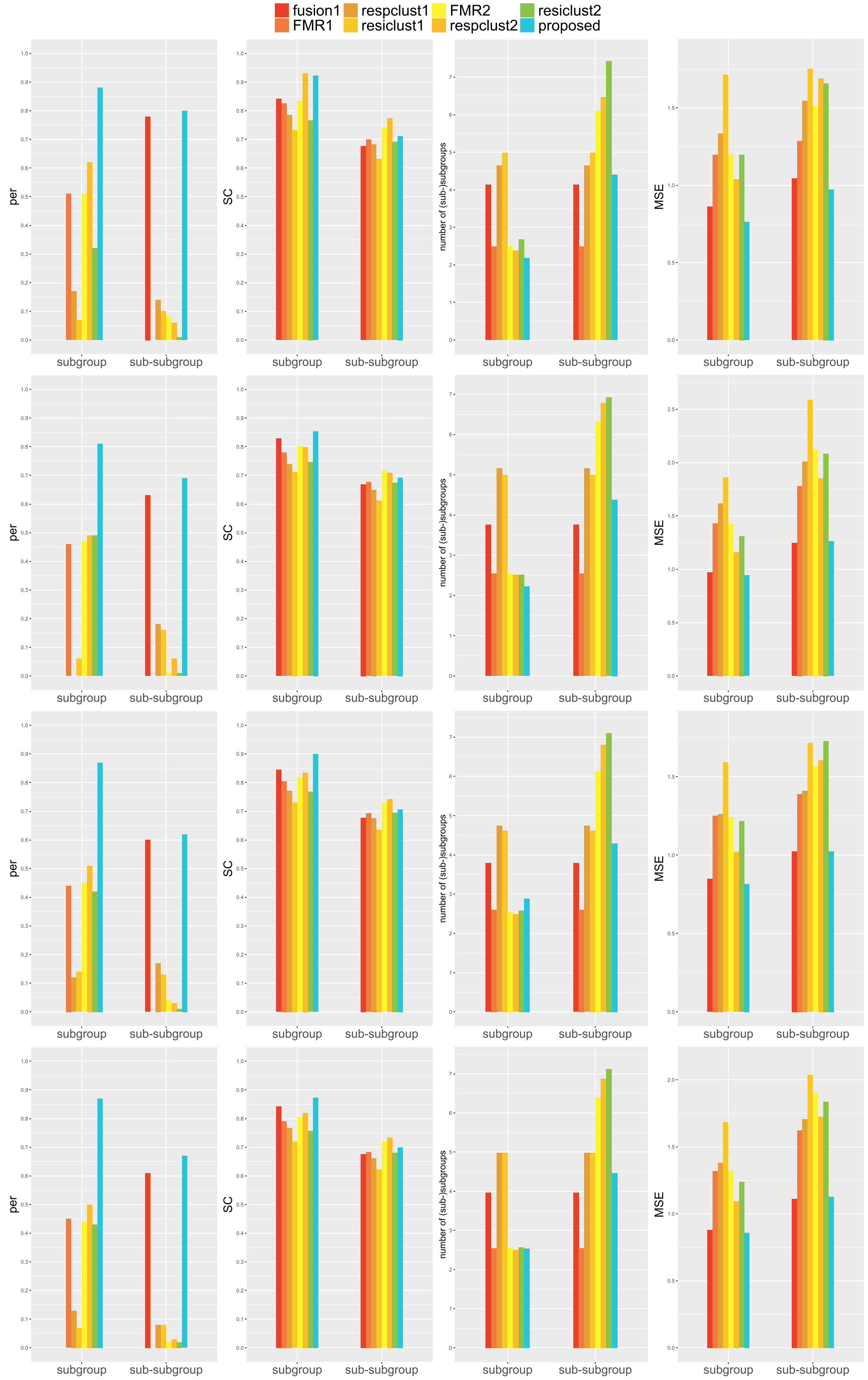


FIGURE 3 Simulation results for the scenario with correlated $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, balanced design, and $\mu = 2$. Correlation structures from top to bottom: AR1, AR2, B1, and B2. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

TABLE 1 Data analysis using the proposed approach: estimated coefficients

	Subgroup			
Type 1 imaging feature[a]	**1**	**2**	**3**	
LymphocytesPN	−3.9241	−1.3239	0.9290	
StromaPN	0.6681	−2.1004	0.4450	
TumorPN	−0.3196	1.4985	−0.3917	
LymphocytesSN	2.3989	1.9573	−0.9580	
StromaSN	−0.5659	0.9379	0.0290	
TumorSN	−0.8804	−0.6502	−0.2264	
	Sub-subgroup			
Selected Type 2 imaging feature	1-1	2-1	3-1	3-2
AreaShape-Center-Y	−0.2939	0.8353	0.0059	0.0052
AreaShape-Zernike-8-2	−0.8493	1.1858	−0.3053	0.1607
Granularity-12-ImageAfterMath	−0.8566	0.6822	−0.0904	−0.0731
Texture-Contrast-maskosingray-3-03	−1.1070	−0.1602	0.3224	1.7454
Texture-Correlation-maskosingray-3-01	−0.7805	−0.1187	0.0598	0.9032
Texture-DifferenceVariance-ImageAfterMath-3-03	−1.0674	0.0939	−0.1574	2.0068
Texture-SumVariance-ImageAfterMath-3-00	−0.1428	0.5558	−0.3443	0.4095
Texture-SumVariance-maskosingray-3-01	1.0721	−0.1465	−0.1281	−3.7819
Threshold-FinalThreshold-Identifyhemasub2	−0.0651	−1.0249	−0.3460	−2.3063
Threshold-SumOfEntropies-Identifyeosinprimarycytoplasm	−0.2158	−0.4746	−0.0795	−2.5133

[a] 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. LymphocytesSN: Size of lymphocyte cell region/image size. StromaSN: Size of stromal cell region/image size. TumorSN: Size of tumor cell region/image size.

detailed information. Information on the six extracted features is available in Table 1. In the extraction of the second type of imaging features, in slide chopping, the subimages have 500×500 pixels, and 20 subimages are randomly selected for downstream analysis. Note that this is the common practice in published studies (Yu et al., 2017) and balances computational efficiency and effectiveness, and that these parameters are consistent with the literature (Zhu et al., 2016; Luo et al., 2017). There are 229 features available for analysis, most of which are not expected to be informative (Yu et al., 2017; Zhong et al., 2019). As such, we conduct a supervised prescreening, with FEV1 as the response variable, using the ridge fusion-based regression. This technique modifies the pairwise fusion approach by revising the penalty to ridge (Ma and Huang, 2017) and can be computationally more efficient. Then the top 10 features (Table 1), ranked based on residuals, are selected for downstream analysis, which can significantly reduce computational cost. Sensitivity analysis on the number of Type 2 features is performed (Table C10, Supporting Information), and it is observed that the (sub-)subgrouping structure is not sensitive to this choice. Overall, the analyzed data contains the six Type 1 imaging features, 10 Type 2 imaging features, and response variable for 118 subjects.

The analysis results are summarized in Table 1. Three subgroups are identified, with sizes 23, 23, and 72, respectively. More detailed information on which subjects belong to these subgroups is available from the authors. The three subgroups have significantly different estimated regression coefficients. With the second type of imaging features, the third subgroup is further split into two sub-subgroups, with sizes 49 and 23, respectively. This shows that the proposed approach does not force further splitting, as for the first two subgroups. The coefficients for sub-subgroups 3-1 and 3-2 are significantly different, suggesting no obvious violation of the hierarchy.

It has been recognized in the literature that the biological implications of histopathological imaging features, especially Type 2, are not clear. As such, we are not able to further interpret the heterogeneous regression models. To gain "indirect support," we compare across the identified (sub-)subgroups samples' clinical features, including pre- and post-bronchodilator FEV1, carbon monoxide diffusion, tumor size, disease free survival, and overall survival in Figure C14, as well as smoking status and tumor site in Table C11, in Supporting Information. Significant differences are observed, suggesting that the identified (sub-)subgroups may have clinically meaningful differences. It is observed that most of the relative differences in clinical outcomes are larger than 50% or even 100%, which indicates that they have significant differences across (sub-)subgroups. Moreover, the three subgroups differ significantly in smoking history. The difference in tumor site is significant between sub-subgroups 3-1 and 3-2. It is noted that, except for pre-bronchodilator FEV1, these clinical features are not included in model fitting. As such, there is no concern on over-fitting. This analysis can support the validity of the identified heterogeneity structure to a great extent.

4.1 | Analysis with alternative methods

Data are also analyzed using the alternative methods. The estimation results are provided in Tables C12– C18 (Supporting Information). It is observed that different approaches lead to significantly different heterogeneity structures and estimates. In Table C19, the concordance values between the different heterogeneity structures are shown. The proposed approach has a higher concordance with fusion1, which is as expected since they have similar analysis schemes. It has moderate concordance with the other approaches.

In evaluation, we first remove one subject and then conduct estimation. Prediction is made for the one removed subject, we use the model that most of the removed subject's co-members (in the analysis using the whole data) have for prediction. More specifically, denote the subject index sets of the $K^{(0)}$ sub-subgroups in the analysis using the whole data as $\{\mathcal{T}_1^{(0)}, \dots, \mathcal{T}_{K^{(0)}}^{(0)}\}$, and the subject index sets of the $K^{(l)}$ sub-subgroups in the analysis after removing the lth subject as $\{\mathcal{T}_1^{(l)}, \dots, \mathcal{T}_{K^{(l)}}^{(l)}\}$. For any l, there exists $k^{(0)}$ satisfying $l \in \mathcal{T}_{k^{(0)}}^{(0)}$. Denote $k^{(l)} = \max_{k=1,\dots,K^{(0)}} |\mathcal{T}_{k^{(0)}}^{(0)} \bigcap \mathcal{T}_k^{(l)}|$, and we use the model corresponding to $\mathcal{T}_{k^{(l)}}^{(l)}$ for prediction. This can be slightly ad hoc, however, literature search does not suggest a more appropriate approach. This procedure is cycled through all subjects, and the prediction MSE values are 0.907 (fusion1), 0.991 (FMR1), 1.130 (respclust1), 1.124 (resiclust1), 1.024 (FMR2), 1.012 (respclust2), 1.081 (resiclust2), and 0.869 (proposed).

As a byproduct of the above procedure, we are also able to evaluate subgrouping stability of the proposed approach. In particular, we compare the heterogeneity structure obtained using all subjects (denoted as φ_0) against that with the lth subject removed (denoted as φ_l), and compute the similarity measure as

$$\binom{n-1}{2}^{-1} |\{(i,j) : I[\varphi_l(\boldsymbol{x}_i,\boldsymbol{z}_i) = \varphi_l(\boldsymbol{x}_i,\boldsymbol{z}_i)] = I[\varphi_0(\boldsymbol{x}_i,\boldsymbol{z}_i) = \varphi_0(\boldsymbol{x}_i,\boldsymbol{z}_i)]; i < j, i, j \neq l\}|,$$

where $l = 1, \dots, n$. Note that this needs to be computed for the subgrouping and sub-subgrouping separately. The mean similarity measure values are 0.843 and 0.913, respectively. The high values provide additional confidence to the proposed analysis. Here we do not pursue the comparison of stability as it may not be sensible to do so. In particular, the "most stable" approach is the one grouping all subjects together and having stability measure 1.

4.2 | Data-based simulation

Realizing that the parametric-model-based simulation reported in Section 3 may be somewhat simplified, we conduct additional simulation based on the LUAD data. In particular, for each simulation replicate, we randomly select 96 (out of the 118) subjects and adopt the observed imaging feature values. The rest of the simulation settings are similar to those in Section 3. Results are summarized in Figure C15 and Table C20 (Supporting Information). Similar to in Section 3, the proposed approach is observed to have superior performance.

4.3 | Additional PCA-based analysis

To accommodate the high dimensionality of the second type of imaging features, we have conducted a prescreening. To "complete the picture," we further consider a dimension reduction-based alternative. In particular, we conduct PCA (principal component analysis) with the second type of imaging features, select the top 10 PCs, and then apply the proposed approach. The estimation results are summarized in Table C21 (Supporting Information). Two subgroups are identified, with sizes 23 and 95, respectively. The second subgroup is further split into four sub-subgroups, with sizes 23, 23, 23, and 26. With the same procedure as described above, the prediction MSE is computed as 0.920. The significant differences between the prescreening and dimension reduction-based analyses are not surprising. Our literature review suggests that there may be more variable selection-based analysis. We leave the systematic comparison between the two types of techniques to future research.

5 | DISCUSSION

For supervised heterogeneity analysis, which is of critical importance in cancer and other biomedical studies, this study has first proposed an innovative hierarchy. Although limited to the analysis of histopathological imaging features in this article, the hierarchy can have far broader applications. In particular, as long as there is a natural order in the "resolution" of two types of variables (for example, clinical biomarkers + molecular measurements, and main effects + higher order interactions), the hierarchy strategy may be applicable. We have then developed a novel approach to achieve estimation and, more importantly, the hierarchical heterogeneity structure in a single step. The proposed approach has been partly motivated by penalized fusion and sparse group penalization but significantly advances from them. Statistical properties and computation have been carefully studied. It is noted that although having some roots in the existing penalized fusion literature, with the significantly different analysis contexts and penalty form, the developments have been highly nontrivial. Both simulation and data analysis have shown satisfactory performance of the proposed approach. In the analysis of TCGA LUAD data, it has identified a finer heterogeneity structure significantly different from the alternatives.

There are several possible future directions. For example, it is of interest to conduct inference and test whether the proposed hierarchical structure can be simplified or is violated for practical data. This demands a new direction of investigation into inference under penalization. The proposed approach has been designed for low-dimensional variables. Its computational complexity is of the order $O(n^2(q+p))$. The high computational cost problem has been well noted even with the standard penalized fusion (Ma and Huang, 2017; Liu and Lin, 2019; Zhang et al., 2019). It is of interest to extend it to a high-dimensional approach involving additional variable selection or dimension reduction components. Although seemingly "straightforward," our brief exploration has suggested that tremendous computational and theoretical developments will be needed. It is also of interest to apply the proposed approach to more extensive data analysis. Last but not least, with the lack of direct interpretations of imaging features and lack of independent validation data, we have not been able to fully explore clinical implications of our findings. With the high significance of lung cancer heterogeneity, more careful biological/clinical examination may be warranted.

ACKNOWLEDGMENTS

We thank the editors and reviewers for their insightful comments. This work was supported by National Natural Science Foundation of China (12171454, 12026604, 11971404, Basic Scientific Project 71988101), Pazhou Lab (Guangzhou 510320, China), Beijing Natural Science Foundation (Z190004), 111 Project (B13028), NSF (1916251), NIH (CA241699, CA204120, CA196530), and a Yale Cancer Center Pilot Award.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are openly available in TCGA (The Cancer Genome Atlas) at https://portal.gdc.cancer.gov/projects/TCGA-LUAD.

ORCID

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

REFERENCES

Aeffner, F., Zarella, M.D., Buchbinder, N., Bui, M.M., Goodman, M. R., Hartman, D. J. et al. (2019) Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. *Journal of Pathology Informatics*, 10, 9.

Balte, P., Chaves, P., Couper, D., Enright, P., Jacobs, D., Kalhan, R. et al. (2020) Association of nonobstructive chronic bronchitis with respiratory health outcomes in adults. *JAMA Internal Medicine*, 180, 676–686.

Bien, J., Taylor, J.& Tibshirani, R. (2013) A lasso for hierarchical interactions. *Annals of Statistics*, 41, 1111.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O. et al. (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7, R100.

Chen, J., Tran-Dinh, Q., Kosorok, M.& Liu, Y. (2020) Identifying heterogeneous effect using latent supervised clustering with adaptive fusion. *Journal of Computational and Graphical Statistics*, 30, 43–54.

Choi, H.& Na, K. (2018) Integrative analysis of imaging and transcriptomic data of the immune landscape associated with tumor metabolism in lung adenocarcinoma: clinical and prognostic implications. *Theranostics*, 8, 1956–1965.

Cooper, L. A., Kong, J., Gutman, D. A., Dunn, W. D., Nalisnik, M.& Brat, D. J. (2015) Novel genotype-phenotype associations in human cancers enabled by advanced molecular platforms and computational analysis of whole slide images. *Laboratory Investigation*, 95, 366–376.

Gong, Y., Ji, P., Yang, Y., Xie, S., Yu, T., Xiao, Y. et al. (2021) Metabolic-pathway-based subtyping of triple-negative breast cancer reveals potential therapeutic targets. *Cell Metabolism*, 33, 51–64.

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M.& Yener, B. (2009) Histopathological image analysis: a review. *IEEE Reviews in Biomedical Engineering*, 2, 147–171.

He, B., Zhong, T., Huang, J., Liu, Y., Zhang, Q. & Ma, S. (2020) Histopathological imaging-based cancer heterogeneity analysis via penalized fusion with model averaging. *Biometrics*. https://doi.org/10.1111/biom.13357.

Jing, P., Su, Y., Nie, L., Gu, H., Liu, J.& Wang, M. (2018) A framework of joint low-rank and sparse regression for image memorability prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 29, 1296–1309.

Khalili, A.& Chen, J. (2007) Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102, 1025–1038.

Kothari, S., Phan, J.H., Stokes, T.H.& Wang, M.D. (2013) Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association*, 20, 1099–1108.

Liu, L.& Lin, L. (2019) Subgroup analysis for heterogeneous additive partially linear models and its application to car sales data. *Computational Statistics & Data Analysis*, 138, 239–259.

Liu, H.& Sunil R., (2020) Generalized finite mixture of multivariate regressions with applications to therapeutic biomarker identification. *Statistics in Medicine*, 39, 4301–4324.

Luo, X., Zang, X., Yang, L., Huang, J., Liang, F., Rodriguez-Canales, J. et al. (2017) Comprehensive computational pathological image analysis predicts lung cancer prognosis. *Journal of Thoracic Oncology*, 12, 501–509.

Mannino, D., Diaz-Guzman, E.& Buist, S. (2011) Pre-and post-bronchodilator lung function as predictors of mortality in the Lung health study. *Respiratory Research*, 12, 1–7.

Ma, S.& Huang, J. (2017) A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112, 410–423.

Prat, A., Pineda, E., Adamo, B., Galván, P., Fernández, A., Gaba, L. et al. (2015) Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast*, 24, S26–S35.

Rand, W. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846–850.

Städler, N., Bühlmann, P.& Van De Geer, S. (2010) ℓ_1-penalization for mixture regression model. *Test*, 19, 209–256.

Sun, D., Li, A., Tang, B.& Wang, M. (2018) Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Computer Methods and Programs in Biomedicine*, 161, 45–53.

TCGA. (The Cancer Genome Atlas) Available at: https://portal.gdc.cancer.gov/projects/TCGA-LUAD. [Accessed 8th January 2021].

Wang, B., Zhang, Y., Sun, W.& Fang, Y. (2018) Sparse convex clustering. *Journal of Computational and Graphical Statistics*, 27, 393–403.

Wang, S., Chen, A., Yang, L., Cai, L., Xie, Y., Fujimoto, J. et al. (2018) Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Scientific Reports*, 8, 1–9.

Wang, S., Wang, T.& Yang, L. (2019) ConvPath: a software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. *Ebiomedicine*, 50, 103–110.

Wiwie, C., Baumbach, J.& Röttger, R. (2015) Comparing the performance of biomedical clustering methods. *Nature Methods*, 12, 1033.

Xie, J., Lin, Y., Yan, X.& Tang, N. (2020) Category-adaptive variable screening for ultra-high dimensional heterogeneous categorical data. *Journal of the American Statistical Association*, 115, 747–760.

Yang, X., Yan, X.& Huang, J. (2019) High-dimensional integrative analysis with homogeneity and sparsity recovery. *Journal of Multivariate Analysis*, 174, 104529.

Yu, K. H., Berry, G. J., Rubin, D. L., Re, C., Altman, R. B.& Snyder, M. (2017) Association of omics features with histopathology patterns in lung adenocarcinoma. *Cell Systems*, 5, 620–627.

Yu, K. H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L. et al. (2016) Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communication*, 7, 12474.

Zhang, C. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942.

Zhang, Y., Wang, H.& Zhu, Z. (2019) Robust subgroup identification. *Statistica Sinica*, 29, 1873–1889.

Zhong, T., Wu, M.& Ma, S. (2019) Examination of independent prognostic power of gene expressions and histopathological imaging features in cancer. *Cancers*, 11, 361.

Zhu, X., Yao, J., Luo, X., Xiao, G., Xie, Y., Gazdar, A. et al. (2016) Lung cancer survival prediction from pathological images and genetic data–an integration study. In: *2016 IEEE 13th international symposium on biomedical imaging*. Piscataway, NJ: IEEE, pp. 1173–1176.

SUPPORTING INFORMATION

Additional theoretical results (referenced in Section 2.2), details of the computational algorithm (referenced in Section 2.3), and numerical results (referenced in Sections 3 and 4) are available at the Biometrics website on Wiley Online Library. R programs implementing the proposed method are available at https://github.com/shuanggema and the Biometrics website on Wiley Online Library.

How to cite this article: Ren M, Zhang Q, Zhang S, Zhong T, Huang J, Ma S. Hierarchical cancer heterogeneity analysis based on histopathological imaging features. *Biometrics*. 2021;1–13.
https://doi.org/10.1111/biom.13544