The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1808–1830
https://doi.org/10.1214/21-AOAS1458

RADIOHEAD: RADIOGENOMIC ANALYSIS INCORPORATING TUMOR HETEROGENEITY IN IMAGING THROUGH DENSITIES

BY SHARIQ MOHAMMED[1,*], KARTHIK BHARATH[2], SEBASTIAN KURTEK[3], ARVIND RAO[1,†] AND VEERABHADRAN BALADANDAYUTHAPANI[1,‡]

[1]*Department of Biostatistics, Department of Computational Medicine and Bioinformatics, University of Michigan,* [*]*shariqm@umich.edu;* [†]*ukarvind@umich.edu;* [‡]*veerab@umich.edu*

[2]*School of Mathematical Sciences, University of Nottingham, karthik.bharath@nottingham.ac.uk*

[3]*Department of Statistics, The Ohio State University, kurtek.1@stat.osu.edu*

Recent technological advancements have enabled detailed investigation of associations between the molecular architecture and tumor heterogeneity through multisource integration of radiological imaging and genomic (radiogenomic) data. In this paper we integrate and harness radiogenomic data in patients with lower grade gliomas (LGG), a type of brain cancer, in order to develop a regression framework called RADIOHEAD (RADIOgenomic analysis incorporating tumor HEterogeneity in imAging through Densities) to identify radiogenomic associations. Imaging data is represented through voxel-intensity probability density functions of tumor subregions obtained from multimodal magnetic resonance imaging and genomic data through molecular signatures in the form of pathway enrichment scores corresponding to their gene expression profiles. Employing a Riemannian-geometric framework for principal component analysis on the set of probability density functions, we map each probability density to a vector of principal component scores which are then included as predictors in a Bayesian regression model with the pathway enrichment scores as the response. Variable selection compatible with the grouping structure amongst the predictors induced through the tumor subregions is carried out under a group spike-and-slab prior. A Bayesian false discovery rate mechanism is then used to infer significant associations based on the posterior distribution of the regression coefficients. Our analyses reveal several pathways relevant to LGG etiology (such as synaptic transmission, nerve impulse and neurotransmitter pathways) to have significant associations with the corresponding imaging-based predictors.

1. Introduction. Gliomas are a group of tumors occurring in the brain and spinal cord, further categorized into subgroups. Lower grade gliomas (LGG) are characterized as World Health Organization grade II and III tumors, and they come from two different types of brain cells known as astrocytes and oligodendrocytes. The causes of these types of tumors are not well understood, and recent studies have examined their molecular characterization from datasets generated by The Cancer Genome Atlas (TCGA) and have associated disease prognosis with their underlying molecular architecture (Verhaak et al. (2010)). In the context of gliomas, there has been growing interest in exploring the underlying comprehensive molecular characterization (Fishbein et al. (2017), Noushmehr et al. (2010), Venneti and Huse (2015), Verhaak et al. (2014)). For example, Ceccarelli et al. (2016) studied the complete set of genes associated with diffuse grade II-III-IV gliomas from TCGA to identify molecular correlations by comprehensively analyzing the sequencing and array-based molecular profiling data and to improve disease classification and provide insights into the progression of the tumor from low- to high-grade.

Received July 2020; revised February 2021.

Key words and phrases. Fisher–Rao metric, group spike-and-slab, principal component analysis, radiogenomic associations.

Gliomas usually contain various heterogeneous subregions: edema, nonenhancing and enhancing core which reflect differences in tumor biology, have variable histologic and genomic phenotypes and exhibit highly variable clinical prognosis (Bakas et al. (2017a)). This intrinsic heterogeneity in tumor biology is also reflected in their radiographic phenotypes through different intensity profiles of the subregions in imaging. Such phenotypes can be obtained from images based on computed tomography (CT), positron emission tomography (PET) and magnetic resonance imaging (MRI); each of which allows integration with other data sources (e.g., genomics). Moreover, imaging and genomic data provide complementary information in terms of tumor heterogeneity and molecular characterization, respectively. Molecular classification of LGGs can be facilitated, and sometimes even validated, through radiogenomic analyses based on noninvasive medical image-derived features. Imaging features have been known to capture physiological and morphological heterogeneity of tumors as they progress from a single cell (Marusyk, Almendro and Polyak (2012)). Such studies have an important bearing on the design of personalized therapeutic strategies in cancer and, potentially, guide monitoring of disease development or progression for early stage cancers. Thus, examination of inter- and intra-tumor heterogeneity through imaging features as well as their potential association with genomic markers can lead to a better understanding of molecular signatures of LGGs.

In this work we focus on the MRI modality, as it furnishes a wide range of image contrasts at a high resolution, which can be used to exhibit and evaluate the location, growth and progression of tumors. Moreover, improved resolution of MRIs has facilitated the understanding of different aspects of tumor characteristics (Just (2014)). The apparent utility of MRI in studying heterogeneity of subregions of gliomas can be seen in Figures 2 and 3, where different intensity profiles disseminated across the multimodal MRI scans appear to exhibit complementary information. Studying heterogeneity in the subregions is now feasible, due to the availability of their gold standard labeling (Bakas et al. (2015)), which facilitates further radiomic and radiogenomic analyses.

1.1. *Voxel-intensity densities as an imaging feature*. Using the raw MRI scans as predictors in the modelling is a challenge, as we do not have an underlying atlas structure to compare between subjects that is commonly available for other imaging modalities such as neuroimaging studies (Ombao et al. (2016)). Diagnostic image-based features using voxel-level data have been utilized for modelling purposes (e.g., to visualize the progression/regression of tumors). However, one of the main drawbacks of existing studies is that only a few chosen summary statistics/metrics represent entire regions of interest. Some of these summary statistics include percentiles, extreme percentiles (e.g., 5th and 95th), quartiles, skewness, kurtosis, histographic pattern, range and mode of MRI-based voxel intensity histograms (Baek et al. (2012), Just (2011), Song et al. (2013)). Although such metrics have clear utility in the assessment of tumor heterogeneity, they generally do not provide a comprehensive representation due to: (a) the subjectivity in the choice of the number and location of summary features and (b) the limitation of these features in terms of capturing the entire information in a voxel-intensity distribution. As a result, any statistical analysis based on such an approach is unable to detect potential small-scale and sensitive changes in the tumor due to treatment effects (Just (2014)).

As an alternative to summary statistics, associations between genomic variables and tumor heterogeneity can be examined on different scales of the voxel intensity probability density function (PDF): while significant genomic variation might manifest as markedly distinct aspects of a PDF (e.g., number of modes, large changes in location of mean/mode), genomic variation (relative to the measurement scale) might show up in subtle, small-scale changes in overall shape of the PDF (e.g. slopes between modes) and sometimes in the tails. Indications

of such a behavior were evident in an unsupervised clustering setting in earlier work that considered entire voxel-intensity PDFs as data objects (Saha et al. (2016)). Including such small-scale changes without summarizing the entire PDF through coarse summary statistics could result in better correlative and predictive power of models associating genomic variables to radiographic phenotypes (Yang et al. (2020)).

In this article we propose to examine variations in the genomic signature of a tumor through changes, both large and subtle, in overall shape[1] of the PDF of voxel intensities, using a Riemannian-geometric framework on the space of PDFs. This space is a nonlinear, infinite-dimensional manifold, and the lack of a global linear structure brings about nontrivial challenges in their analyses. Here, we develop a regression framework called RADIOHEAD (RADIOgenomic analysis incorporating tumor HEterogeneity in imAging through Densities) to model associations between genomic variables characterizing the molecular signature of tumors and voxel-intensity PDFs from multimodal MRI scans. In what follows we use PDFs and densities interchangeably.

1.2. *RADIOHEAD modelling outline.* We propose an integrated end-to-end method: from MR images to evaluation of voxel-level density-based radiomic features, gene expression to associated pathway-level enrichment scores and the subsequent statistical modelling framework. Figure 1 shows the schematic workflow diagram for our method. For each patient we generate PDFs corresponding to three heterogeneous tumor subregions: (i) necrosis and nonenhancing, (ii) edema and (iii) enhancing core. The expression/activation of the pathways is evaluated by computing pathway enrichment scores through gene-set variation analysis (GSVA); these scores are subsequently used as a univariate response variable. We apply the proposed RADIOHEAD approach to the TCGA dataset of LGGs.

Fitting a model by regressing enrichment scores against multiple PDFs (one from each combination of tumor subregion and MRI sequence) poses two main challenges:

1. Each PDF is a nonnegative function which integrates to one and hence cannot be treated as a standard functional predictor;

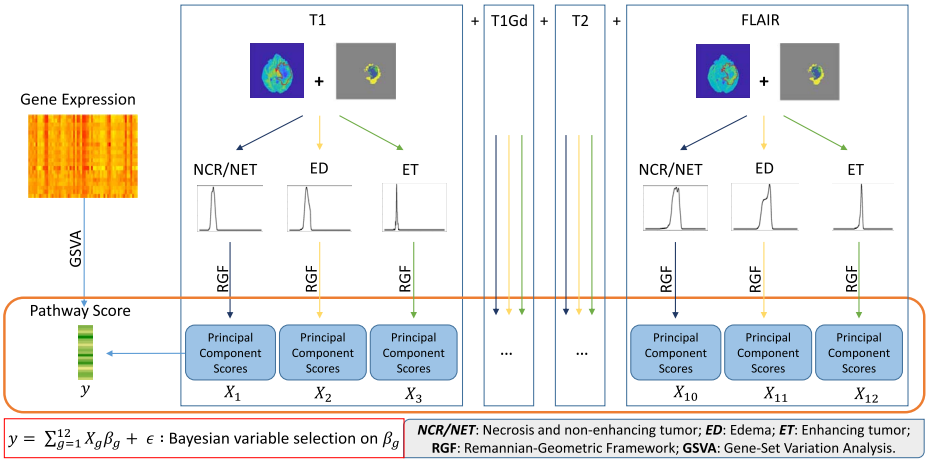


FIG. 1. *Schematic representation of the RADIOHEAD modelling approach. Pathway scores are constructed from gene expression using gene-set variation analysis (GSVA). From each MRI sequence we construct densities for each of the three tumor subregions and use them to construct principal component scores under a Reimannian-geometric framework. Pathway scores are used as a response and the principal component scores as predictors in the downstream analysis.*

[1]"(S)hape" is used in a nontechnical sense.

2. The grouping structure between tumor subregions needs to be incorporated while examining the functional relationship between a pathway score and its corresponding PDFs.

The first challenge is addressed by mapping each PDF to a finite-dimensional vector of principal component (PC) scores by carrying out Riemannian principal component analysis (PCA) on the sample of PDFs corresponding to each tumor subregion. These PC scores, corresponding to the multiple PDFs, act as imaging meta-features and are incorporated as individual predictors, which leads to a $p \gg n$ situation wherein the number of radiomic meta-features (p) is higher than the number of subjects (n). In the presence of uncertainty in the actual effects of small changes in the PC scores on the enrichment scores, it is natural to employ a Bayesian model for variable selection. To this end, we address the second challenge by using a group-structured continuous spike-and-slab prior (Andersen, Winther and Hansen (2014), Ishwaran and Rao (2005)) on the total set of PC scores in an effort to capture information on the biological structure in the data and to provide analyses that are more amenable to interpretation. The prior formulation also simplifies the computation by allowing for simple (conditional posteriors from standard distributions) and fast MCMC sampling (via Gibbs sampling). Other existing prior formulations incorporating group structure (Xu and Ghosh (2015), Yang and Narisetty (2020), Zhang et al. (2014)) could also be used. Furthermore, to address the issue of multiple comparisons, a Bayesian false discovery rate-based approach is used to build inference based on error rates.

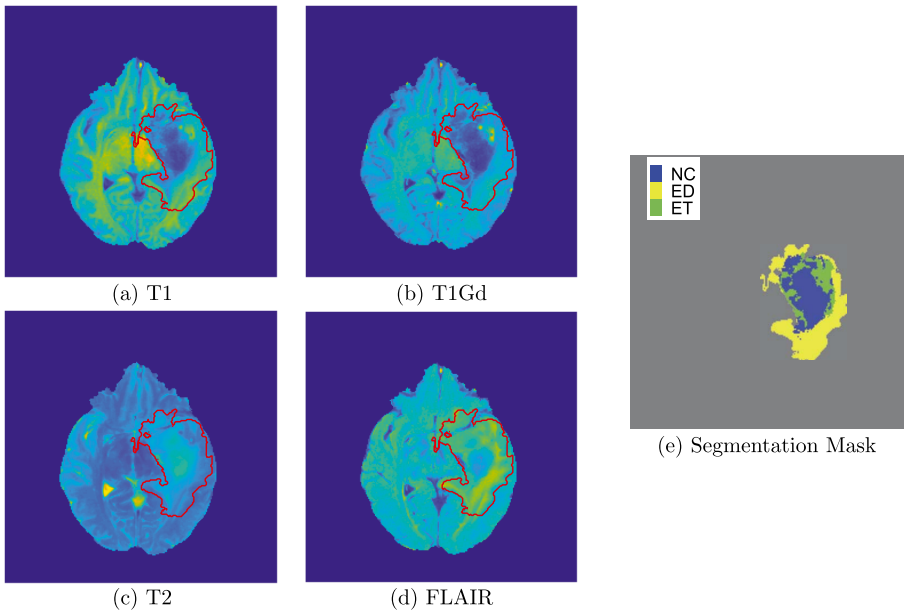
Section 2 describes the data along with the acquisition process and pre-processing steps. We describe the algorithm to compute the density-based PC scores in Section 3.1; the computation of GSVA-based enrichment scores is outlined in Appendix B. Section 3.2 describes the regression setup with densities as covariates. In Section 3.3 we describe the regression in terms of PC scores and the modelling approach based on Bayesian variable selection using the group spike-and-slab prior. The estimation and inference strategies follow in Sections 3.4 and 3.5. In Section 4 we present our results and describe the identified radiogenomic associations in LGG. We close with a brief Discussion and some directions for future work in Section 5.

2. Dataset description. We describe the data acquisition and pre-processing steps involved for the imaging and genomic data separately.

2.1. *Imaging data.* To conduct our analyses, we use MRI scans that include reliable tumor segmentations along with identified tumor subregions. We consider preoperative multi-institutional scans in the TCGA LGG collection, publicly available in The Cancer Imaging Archive (TCIA–Clark et al. (2013)). We obtain segmentation labels for these MRI scans using an automated method called GLISTRboost (Bakas et al. (2015), Bakas et al. (2017a)). Segmentation labels generate a mask for each subject's MRI scan which distinguishes between necrotic and nonenhancing tumor (NCR/NET or NC), peritumoral edema (ED) and enhancing tumor (ET).

MRI provides a wide range of imaging contrasts through multimodal images. The primary MRI sequences include: (a) native (T1), (b) post-contrast T1-weighted (T1Gd), (c) T2-weighted (T2) and (d) T2 fluid attenuated inversion recovery (FLAIR). Each of these sequences identifies different types of tissue and displays them using varying contrasts based on the tissue characteristics. We use LGG data for 65 subjects, obtained from Bakas et al. (2017b), which contain: (a) MRI scans based on all four sequences (T1, T1Gd, T2 and FLAIR) and (b) corresponding segmentation masks generated by GLISTRboost.

The structure of the data under study is as follows: each MRI scan is a three-dimensional array with the third axis representing different axial slices. For each subject we have four sequences, as described above, corresponding to four different 3D arrays accompanied by a



(a) T1 (b) T1Gd

(c) T2 (d) FLAIR

(e) Segmentation Mask

FIG. 2. *Figures* (*a*)–(*d*): *Axial slice of a skull-stripped brain MRI for a subject with LGG, shown for the four sequences T*1, *T*1*Gd*, *T*2 *and FLAIR, respectively. The segmented tumor region is displayed using a red boundary overlaid on the images. Figure* (*e*): *The corresponding subregion segmentation mask with the NC, ED and ET regions marked in different colors.*

unique segmentation mask that has a one-to-one correspondence with the voxels in the MRI scans. That is, there is a voxel-to-voxel correspondence across all four MRI sequences and the segmentation mask. An example of a single axial slice from a brain MRI for a subject with LGG, for the four aforementioned sequences, is shown in the left panel in Figure 2. The segmented tumor region is indicated by a red boundary overlaid on the images and is further classified into the tumor subregions NC, ED and ET, as shown in the right panel in Figure 2. The voxel intensity values of MRI scans are difficult to interpret and compare, as they are sensitive to the configuration of the MRI scanner. These values are not comparable either between study visits within a single subject or across different subjects which necessitates pre-processing of the images in terms of intensity value normalization. We address this issue through a biologically motivated normalization technique using the `R` package *WhiteStripe* (Shinohara et al. (2014)).

2.2. *Genomic data*. The genomic data was obtained from LinkedOmics[2] (Vasaikar et al. (2017)) which is a publicly available portal that includes multiomics data for LGG among many other cancer types. We consider the normalized gene-level RNA sequencing data from the primary solid tumor tissue using the Illumina HiSeq system (high-throughput sequencing) with expression values in $\log_2$ scale. The entire dataset contains gene expression data for 516 samples and 20,086 genes; we consider a subset of 65 matched samples corresponding to the imaging data described in Section 2.1. We consider the enriched pathways in LGG, as identified by Ceccarelli et al. (2016), hereafter referred to as *C-Pathways*.

We obtain the mapping from genes to pathways and use them along with the gene expression data to obtain *pathway scores*. These scores are numerical estimates of the relative enrichment of a pathway of interest across a sample population, and are computed using a

[2]www.linkedomics.org

nonparametric, unsupervised method called GSVA. It estimates a value per sample and pathway for the variation in the activity of a pathway within an entire gene expression set. In other words, it assesses the relative variability of gene expression in the pathway, as compared to expression of genes not in the pathway. The computation details of the pathway scores can be found in Appendix B. For the C-Pathways (such as ion transport and synaptic transmission) considered in this paper, the genes to pathway mappings are obtained from the molecular signature database (Liberzon et al. (2011)). For each gene-set within the collection, we construct the pathway score using gene-set variation analysis (Hänzelmann, Castelo and Guinney (2013)). Of the 22 C-Pathways we only include 21 of them, as the gene membership for one of the pathways was not available. The pathway scores are computed using the *GSVA* package in `R` obtained from Bioconductor (Gentleman et al. (2004)). Summary statistics for the pathway scores are shown in Table S1 of the Supplementary Material (Mohammed et al. (2021a)).

3. Statistical framework. Our main goal is to identify associations between imaging meta-features and gene expression-based pathway scores. In this section we first describe the Riemannian-geometric approach to construct the voxel PDF-based PC scores for each subject corresponding to a certain tumor subregion. We also define a formal regression model based on the group spike-and-slab prior as well as associated estimation and variable selection procedures.

3.1. *Density-based principal component scores.* We use R to index tumor subregions and M for the different MRI sequences. Consider MRI scans for n subjects from four sequences with the tumor masks containing the segmented tumor region and indicating the subregions. For a given sequence M, we construct the kernel density estimate $f_i^M(R)$, $i = 1, \ldots, n$ for the tumor subregion R in subject i, based on the voxel intensity values in the MRI scan at the array locations of region R obtained from the segmentation. Hence, for each subject i and each sequence M, we have PDF estimates denoted by $f_i^M(\text{NC})$, $f_i^M(\text{ET})$ and $f_i^M(\text{ED})$ corresponding to the necrotic and nonenhancing tumor core (NC), the peritumoral edema (ED) and the enhancing tumor (ET) subregions, respectively. Thus, we consider univariate kernel-density estimates for all tumor subregions and all subjects across the four imaging sequences. The density plots are displayed in Figure 3, where each row corresponds to a specific imaging sequence while each column corresponds to a tumor subregion. We compute the PC scores for each sequence M separately. For brevity, we shall drop the sequence indicator M from the densities and use f_{iR} instead of $f_i^M(R)$ for the remainder of this section.

The kernel density estimates (f_{iR} for all $i = 1, \ldots, n$ and $R \in \mathcal{T} = \{NC, ET, ED\}$) are proper PDFs and belong to the Banach manifold of all PDFs. The following description focuses on PDFs with domain $[0, 1]$; however, the methods apply to more general domains with small adjustments. PDFs are elements of the space $\mathcal{F} = \{f : [0, 1] \to \mathbb{R}_{>0} | \int_0^1 f(x)\,dx = 1\}$. To make $\mathcal{F}$ a Riemannian manifold and to facilitate computation on this space, we endow it with the Fisher–Rao (F-R) Riemannian metric (Kass and Vos (2011), Rao (1992), Srivastava, Jermyn and Joshi (2007)). For brevity, we omit the specific formula for this metric and simply mention that it is closely related to the Fisher information matrix and has useful statistical properties, for example, invariance to bijective and smooth transformations of the PDF domain (Čencov (1982)). Unfortunately, the F–R metric is difficult to use in practice, as the computation of geodesic paths and distances between PDFs is cumbersome and requires numerical methods for approximation. Thus, for simplification we further transform the kernel density estimates using a square-root transformation (Bhattacharyya (1943), Kurtek and Bharath (2015)). As a result, the space of PDFs becomes the positive orthant of the unit

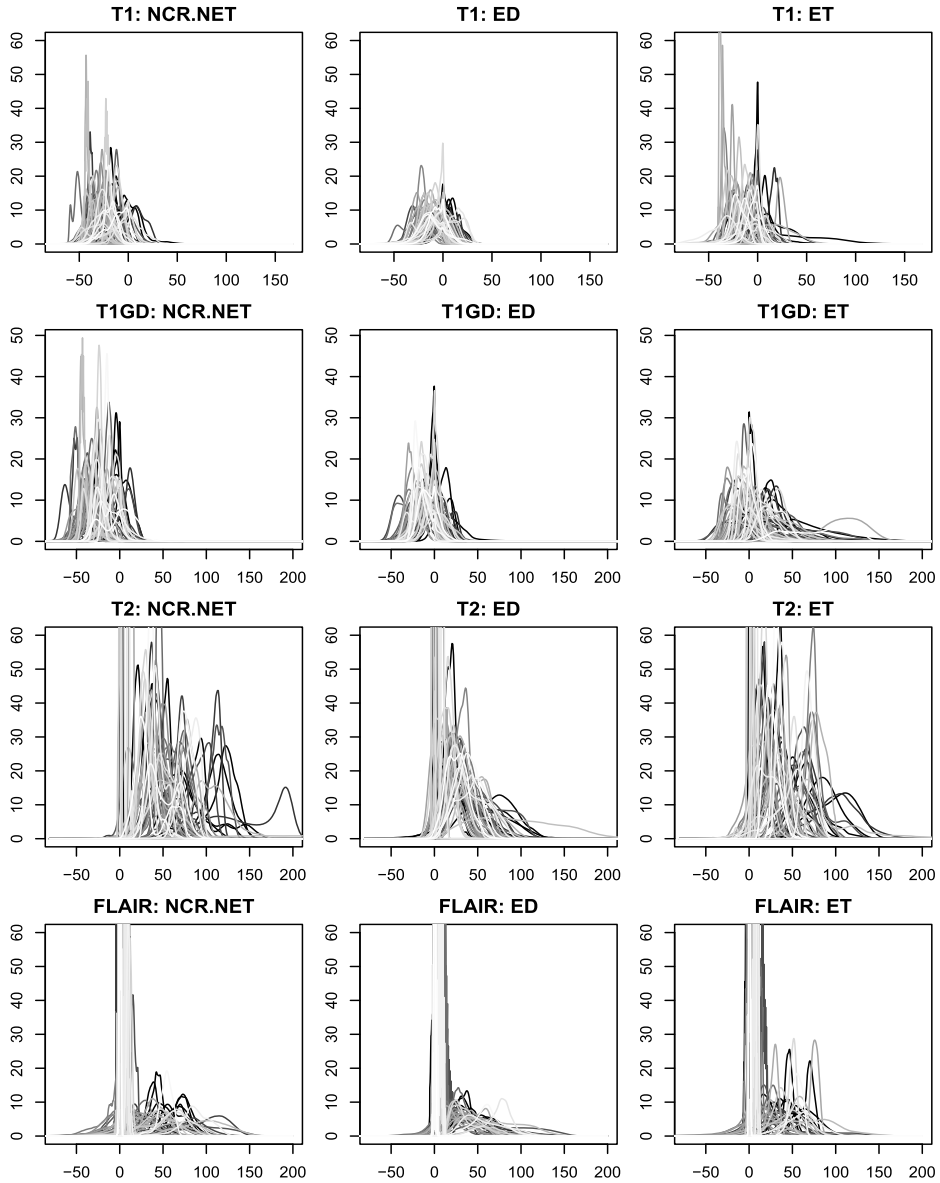


FIG. 3. *Kernel densities* $f_i^M(R)$ *for all subjects across all four MRI sequences and three tumor subregions. For visual convenience the* y*-axes are truncated for each of the subplots. The* x*-axis shows the voxel-intensity values*; *however, we transform them to* $[0, 1]$ *for each imaging sequence to compute the KDEs. Supplementary Material Figure S*1 *shows similar plots in color and Supplementary Material Figure S*2 *shows similar plots without truncation of the* y*-axis.*

sphere in $\mathbb{L}^2 := \mathbb{L}^2([0, 1])$, the geometry of which is well known and the F-R metric flattens to the standard $\mathbb{L}^2$ metric, enabling the computation of geodesic paths and distances in analytical form (Kurtek and Bharath (2015)). Briefly, this result provides simple tools for the statistical tasks of interest including: (a) definition of a distance between two densities, (b) computation of a Karcher mean of a sample of densities and (c) PCA of a sample of densities. We elaborate on these procedures next.

Distances between PDFs and their Karcher mean. Let $h_{iR} = +\sqrt{f_{iR}}$ denote the (positive) square-root densities (SRDs) corresponding to the kernel density estimates f_{iR} for all

$i = 1, \ldots, n$ and $R \in \mathcal{T}$. Each h_{iR} is an element of $\mathcal{H} = \{h : [0,1] \to \mathbb{R}_{>0} | \int_0^1 h^2(x)\,dx = 1\}$, the positive orthant of a unit sphere in $\mathbb{L}^2$, that is, $\mathcal{H}$ is the collection of SRDs corresponding to all PDFs in $\mathcal{F}$. Equipped with the standard $\mathbb{L}^2$ metric, $\mathcal{H}$ becomes a Riemannian manifold (recall that the $\mathbb{L}^2$ metric on $\mathcal{H}$ corresponds to the F–R metric on $\mathcal{F}$). Under this setup the geodesic distance between two densities $f_1, f_2 \in \mathcal{F}$, represented by their SRDs $h_1, h_2 \in \mathcal{H}$, is defined as the shortest great circle arc connecting them on $\mathcal{H}$: $d(f_1, f_2) = d(h_1, h_2)_{L^2} := \cos^{-1}(\langle h_1, h_2 \rangle) = \cos^{-1}(\int_0^1 h_1(x) h_2(x)\,dx) := \theta$. We can now compute the mean of a sample of SRDs using a generalized version of a mean on a metric space, called the Karcher mean (Dryden and Mardia (1998), Karcher (1977)). The sample Karcher mean $\bar{h}$ on $\mathcal{H}$ is defined as the minimizer of the variance functional $\mathcal{H} \ni h \mapsto \sum_{i=1}^n d(h, h_i)^2_{L^2}$. An algorithm for computing the Karcher mean is given in Section S1 of the Supplementary Material; Figure S3 shows the Karcher mean of the densities across all of the subjects for all tumor subregions and imaging sequences, overlaid within each subplot. The computations require two tools from differential geometry called the exponential and inverse-exponential maps. Let $T_h(\mathcal{H}) = \{\delta h | \langle \delta h, h \rangle = 0\}$ denote the tangent space at h. For $h \in \mathcal{H}$ and $\delta h \in T_h(\mathcal{H})$, the exponential map at h, $\exp : T_h(\mathcal{H}) \to \mathcal{H}$ is defined as $\exp_h(\delta h) = \cos(\|\delta h\|)h + \sin(\|\delta h\|)\delta h / \|\delta h\|$, where $\|\delta h\| = \sqrt{\int_0^1 \delta h^2(x)\,dx}$ is the $\mathbb{L}^2$ norm (Biliotti and Mercuri (2017)). The inverse-exponential map is denoted by $\exp_h^{-1} : \mathcal{H} \to T_h(\mathcal{H})$ and, for any $h_1, h_2 \in \mathcal{H}$, it is defined as $\exp_{h_1}^{-1}(h_2) = \theta[h_2 - \cos(\theta)h_1]/\sin(\theta)$, where $\theta = d(h_1, h_2)_{L^2}$ as before.[3]

Principal component analyses on a sample of PDFs. To perform PCA of a sample of SRDs (equivalently PDFs), we utilize the linear tangent space at the sample Karcher mean SRD. That is, we first project all SRDs onto this tangent space using the inverse-exponential map. The sample covariance matrix is then computed in the tangent space at the mean SRD, and PCA is applied through singular value decomposition (SVD) of this matrix. In practice, the densities and their corresponding SRDs are approximated using m-dimensional vectors, which specify the functional values at a set of m discrete points on the domain $[0, 1]$ resulting in $m \times m$-dimensional covariance matrices, where $m \gg n$. We describe the above step-by-step process in Algorithm 1.

The first L columns of U_R, denoted as $\tilde{U}_R \in \mathbb{R}^{m \times L}$, span the L-dimensional principal subspace of the given sample of densities. We can compute the principal coefficients as $X_R = V_R \tilde{U}_R$, where $V_R^\top = [\mathbf{v}_{1R} \mathbf{v}_{2R} \ldots \mathbf{v}_{nR}] \in \mathbb{R}^{m \times n}$ for each $R \in \mathcal{T}$. These principal coefficients X_R^M, referred to as PC scores, act as Euclidean coordinates corresponding to the kernel density estimates $f_i^M(R)$ generated from each MRI sequence M and will be used as predictors in our model. This procedure accomplishes two major goals: (1) it estimates orthogonal directions of variability in a sample of PDFs along with the amount of variability

Algorithm 1 PCA on $T_{\bar{h}}(\mathcal{H})$

1: Compute h_{iR} from f_{iR} (at m discrete points).
2: Compute the Karcher mean of h_{iR} for each tumor subregion $R \in \mathcal{T}$ as $\bar{h}_R$ (see Section S1 in the Supplementary Material).
3: Use the inverse-exponential map to compute $\mathbf{v}_{iR} = \exp_{\bar{h}_R}^{-1}(h_{iR}) \in T_{\bar{h}_R}(\mathcal{H})$.
4: Evaluate the sample covariance matrix $K_R = \frac{1}{n-1} \sum_{i=1}^n \mathbf{v}_{iR} \mathbf{v}_{iR}^\top \in \mathbb{R}^{m \times m}$ for each $R \in \mathcal{T}$.
5: Compute the SVD of $K_R = U_R \Sigma_R U_R^\top$.

[3]For the unit sphere in $\mathbb{L}^2$, strictly speaking, although the exponential map is well defined on the entire tangent space (Biliotti and Mercuri (2017)), the inverse-exponential map may not be. We eschew handling of this technical detail since this is not an issue when computing using the map in practice.

explained by each direction via the covariance decomposition, and (2) it performs dimension reduction by effectively exploring variability in the sample of PDFs through the primary modes of variation in the data.

3.2. *Regression with densities*. The PDFs $f_i^M(R)$ are representations of the heterogeneity in the tumor voxels from the imaging sequence M and the tumore subregion R for subject i. To identify radiogenomic associations, we build regression models with PDFs $f_i^M(R) \forall M, R$ as covariates. For ease of exposition, we drop the indices M and R and explicate the model for one density $f_i(t)$ for $t \in [0,1]$ for subject i as the covariate. Let $h_i(t)$ denote the corresponding SRD. If y_i corresponds to the pathway score for subject i, h_i can be related to y_i using the data-driven model

$$y_i = \beta_0 + \int_0^1 \exp_{\bar{h}}^{-1}(h_i(t))\beta(t)\,dt + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\bar{h} \in \mathcal{H}$ is the Karcher mean of SRDs $h_1, \ldots, h_n \in \mathcal{H}$. Here, $t \mapsto \beta(t)$ is the real-valued coefficient function, β_0 is a real-valued intercept and ϵ_i are i.i.d.$N(0, \sigma^2)$. We specify the model on the tangent space at the data-dependent Karcher mean $\bar{h}$. That is, $\bar{h}$ is the reference SRD for the inverse-exponential map. While this choice influences the model specification (as $\bar{h}$ changes with changing sample composition), it removes the arbitrariness associated with choosing the reference SRD. Effectively, $\exp_{\bar{h}}^{-1}(h_i(t))$ is the Riemannian-geometric equivalent of "centering" the functional covariate h_i. The amount of dependence of the model on $\bar{h}$ is directly dependent on the variability of the sample $h_1, \ldots, h_n$ which can be quantified using the geodesic distances between h_i and $\bar{h}$.

When it exists, the range of $\mathcal{H} \ni h \mapsto \exp_{\bar{h}}^{-1}(h)$ is contained within a linear subspace of $\mathbb{L}^2$, and we can thus express $\exp_{\bar{h}}^{-1}(h_i) = \sum_k \alpha_{ik}\phi_k$ for some sequence $(\alpha_{ik}, k \geq 1)$ with $\sum_k |\alpha_{ik}|^2 < \infty$, where $\{\phi_k, k = 1, 2, \ldots\}$ is an orthonormal set of basis functions for $\mathbb{L}^2$. Similarly, we can write $\beta = \sum_k \beta_k \phi_k$ for some sequence $(\beta_k, k \geq 1)$ with $\sum_k |\beta_k|^2 < \infty$. Hence, the model in equation (1) reduces to

$$y_i = \beta_0 + \sum_{k=1}^{\infty} \alpha_{ik}\beta_k + \epsilon_i, \tag{2}$$

since $\langle \phi_i, \phi_j \rangle = 1$ if $i = j$, and 0 otherwise. For a given gene-set we denote the pathway scores as $\mathbf{y} = (y_1, \ldots, y_n)^\top$, where y_i corresponds to the score for subject i. Having chosen $\bar{h}$, we truncate the number of basis functions at some positive integer $r_n < \infty$. The model in equation (2) is then further simplified as

$$\mathbf{y}_{n\times 1} = \beta_0 \mathbf{1}_n + A\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{1}_n \in \mathbb{R}^n$ is the vector with all entries as 1, row i of $A \in \mathbb{R}^{n\times r_n}$ is given as $(\alpha_{i1}, \ldots, \alpha_{ir_n})^\top \in \mathbb{R}^{r_n}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{r_n})^\top \in \mathbb{R}^{r_n}$. Let $A^\top A = PDP^\top$, where $P \in \mathbb{R}^{r_n \times r_n}$ is an orthogonal matrix of eigenvectors of $A^\top A$ and D is diagonal with $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_{s_n} > 0 = \lambda_{s_n+1} = \cdots \lambda_{r_n}$. If every $\lambda_k > 0 \forall k$, then $\boldsymbol{y}$ is regressed on the principal components of A, which is AP.

The model in equation (3) depends on the choice of the orthonormal basis $\{\phi_k\}$ of $\mathbb{L}^2$ or, in other words, the matrix A and its eigenvectors in P. We use a PC basis for two reasons: (i) it is the optimal empirical orthogonal basis (see, e.g., Chapter 6 of Ramsay and Silverman (2005)) for data on $\mathbb{L}^2$, of which $\mathcal{T}_{\bar{h}}(\mathcal{H})$ is a linear subspace, and (ii) the map $\exp_{\bar{h}}^{-1}(h_i) \mapsto (\alpha_{i1}, \alpha_{i2}, \ldots)$ is an isometry, and, for a fixed positive integer r_n, the corresponding full isometry group is $O(r_n)$ (the set of square orthogonal matrices in dimension r_n). From the perspective of (ii), choosing another orthornormal basis and truncating at r_n

amounts to an orthogonal transform of the corresponding coefficients. Thus, we are effectively regressing the score y_i of the ith subject on the "optimal" r_n-dimensional linear representation of the SRD h_i in the tangent space $\mathcal{T}_{\bar{h}}(\mathcal{H})$ of the sample Karcher mean $\bar{h}$.

The model in equation (3) corresponds to one PDF as a covariate for each subject i. However, from our imaging data, we have 12 PDFs, from four imaging sequences and three tumor subregions, as covariates for each subject. Hence, the model in equation (1) can be extended as

$$y_i = \beta_0 + \sum_M \sum_R \int_0^1 \exp_{\bar{h}_R^M}^{-1}\big(h_{iR}^M(t)\big)\beta_M^R(t)\,dt + \epsilon_i, \tag{4}$$

where $h_{iR}^M(t)$ is the SRD for the PDF $f_{iR}^M(t)$ and $\beta_M^R(t)$ is the coefficient function corresponding to the tumor subregion R in imaging sequence M. Here, $\bar{h}_R^M$ is the sample Karcher mean of $h_{1R}^M(t), \ldots, h_{nR}^M(t)$. Each of the integrals in equation (4) can be reduced to the PC regression form in equation (3). In Section 3.3 we directly work with the PC regression form with the 12 groups of PCs as covariates.

3.3. *Regression with PC scores.* The PDFs belong to a function space, and they carry rich information of the voxel intensities of different tumor subregions at different scales. As a consequence, they also result in a large number (greater than the number of subjects) of principal components across sequences and tumor subregions. This $p \gg n$ situation necessitates the use of variable selection approaches that can induce sparsity as well as regularization. As the PC scores are surrogates for the entire density, it is natural to model the aspects of the density not captured through the scores (such as information on the tumor subregions) using a Bayesian approach by appropriately placing a prior on the high-dimensional feature space. Consequently, this allows us to construct and to assess posterior distributions of coefficients for inference.

Our goal is to identify the density-based principal components across tumor subregions that are significantly associated with the expression levels in the gene set considered. We address this problem from a Bayesian perspective and use the continuous spike-and-slab prior (George and McCulloch (1997), Ishwaran and Rao (2005)) which has inherent variable selection properties. We model the pathway scores $\mathbf{y}$ using principal component scores obtained from all of the tumor subregions and MRI sequences as the predictors. In other words, we assume

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{5}$$

where

$$\mathbf{X} = [X_{NC}^{T1} X_{ED}^{T1} X_{ET}^{T1} X_{NC}^{T1Gd} X_{ED}^{T1Gd} X_{ET}^{T1Gd} \\ X_{NC}^{T2} X_{ED}^{T2} X_{ET}^{T2} X_{NC}^{\mathrm{FLAIR}} X_{ED}^{\mathrm{FLAIR}} X_{ET}^{\mathrm{FLAIR}}]$$

corresponds to the $n \times L$ matrix of predictors containing the principal component scores. The normality assumption is reasonable here since the pathway scores are unimodal and approximately normal by construction (Hänzelmann, Castelo and Guinney (2013)). The model can also be adapted to categorical or survival response types by incorporating latent variable approaches. Here, L is defined as the total number of principal components considered across all sequences and tumor subregions: $L = \sum_M \sum_R L_R^M$, where L_R^M corresponds to the number of columns in X_R^M for $R \in \mathcal{T}$ and M belongs to the four different sequences. We choose L_R^M based on a threshold for the total variation explained by the chosen number of principal components. In the coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^L$, each component is the coefficient corresponding to the principal component from each tumor subregion R and each MRI sequence M; σ^2 is the variance parameter.

Group spike-and-slab prior. Our aim is to identify the tumor subregions in a specific sequence (through the principal components) influencing the pathway scores. This translates to identifying the nonzero coefficients of the model in equation (5). However, the PC scores within each X_R^M contain rich information about the small-scale variability in the densities for region R in sequence type M. The number of principal components to include is dictated by the cumulative amount of variability explained by them. As these densities belong to a function space, capturing variability requires including a large number of PC scores. Moreover, each of these principal components captures different aspects of the variability for the same group, that is, (M, R) pair, and, hence, they will need to be evaluated as a group. Incorporating this grouping structure into the modelling framework, we rewrite the model in equation (5) as

$$\mathbf{y} \sim N\left(\sum_{g=1}^{G} X_g \boldsymbol{\beta}_g, \sigma^2 \mathbf{I}_n\right), \tag{6}$$

where $G = 4 \times 3$, as we have 12 groups arising from four MRI sequences and three tumor subregions. Here, $\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gL_g})^\top$, where L_g is the number of principal components included for the gth group of covariates X_g. Note that our covariates have a clear grouping structure, where each group corresponds to the principal components of a tumor subregion within an imaging sequence. We now introduce a group spike-and-slab prior onto the coefficients $\boldsymbol{\beta}_g$ to identify the groups X_g influencing the pathway scores. Consider the following prior structure:

$$\begin{aligned}
\beta_{gk} &\overset{\text{ind}}{\sim} N\big(0, \sigma^2 \zeta_g \nu_{gk}^2\big), \\
\zeta_g &\overset{\text{iid}}{\sim} (1-w)\delta_{v_0}(\zeta_g) + w\delta_1(\zeta_g), \\
w &\sim U(0,1), \\
\nu_{gk}^{-2} &\overset{\text{iid}}{\sim} \text{Gamma}(a_1, a_2), \\
\sigma^{-2} &\sim \text{Gamma}(b_1, b_2),
\end{aligned} \tag{7}$$

where $\zeta_g \nu_{gk}^2$ is the hypervariance of β_{gk} with ζ_g acting as the group-level indicator variable taking values 1 or v_0 (a small number > 0) with probability w or $1 - w$, respectively. If $\zeta_g = 1$, the hypervariance is dictated by the Inverse-Gamma prior on ν_{gk}^2; if $\zeta_g = v_0$, the prior on β_{gk} is concentrated at 0 allowing for shrinkage of the coefficient parameter β_{gk}. The choice of hyperparameters a_1 and a_2 should be such that we have a continuous bimodal prior on β_{gk}. Further, w acts as the complexity parameter, indicating the proportion of groups with nonzero coefficients, and has a continuous uniform prior on $(0, 1)$. We consider an Inverse-Gamma prior on the variance parameter σ^2. Note that the group structure is incorporated into the variable selection through the indicator ζ_g, which impacts the variance of the parameter β_{gk}. That is, if a specific group is not selected, the hypervariance for the coefficients corresponding to all columns in X_g is small, leading to the prior on β_{gk} being concentrated at zero and vice-versa.

3.4. *Estimation.* For the model in equation (6) and the group spike-and-slab prior structure in equation (7), the full posterior distribution is provided in Section S2 of the Supplementary Material. Let us define $\boldsymbol{\Gamma}_g = \text{diag}(\gamma_{g1}, \ldots, \gamma_{gL_g})$ and $\boldsymbol{\Gamma} = \text{block-diag}(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_G)$, where $\gamma_{gk} = \zeta_g \nu_{gk}^2$. The conditional posteriors for all of the parameters arise from standard distributions, and, hence, we can use Markov chain Monte Carlo (MCMC) sampling procedures, such as Gibbs sampling. Details of the Gibbs sampling approach along with the

Algorithm 2 Gibbs Sampling for Estimation

1: **for** T iterations **do**

2: Sample $\boldsymbol{\beta}_g$ from $\boldsymbol{\beta}_g|\zeta_g, \nu_{gk}^{-2}, \sigma^{-2} \sim N(\boldsymbol{\Sigma}\mathbf{X}^\top\mathbf{y}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Gamma}^{-1})^{-1}$.

3: Sample ζ_g from $\zeta_g|\beta_{gk}, \nu_{gk}^2, w, \sigma^{-2} \sim \frac{w_{1g}}{w_{1g}+w_{2g}}\delta_{v_0}(\cdot) + \frac{w_{2g}}{w_{1g}+w_{2g}}\delta_1(\cdot)$, where

$$w_{1g} = (1-w)v_0^{-\frac{L_g}{2}} \exp\left(-\sum_{k=1}^{L_g} \frac{\beta_{gk}^2}{2\sigma^2 v_0 \nu_{gk}^2}\right) \quad \text{and} \quad w_{2g} = w \exp\left(-\sum_{k=1}^{L_g} \frac{\beta_{gk}^2}{2\sigma^2 \nu_{gk}^2}\right).$$

4: Sample ν_{gk}^{-2} from $\nu_{gk}^{-2}|\beta_{gk}, \zeta_g, \sigma^{-2} \overset{\text{ind}}{\sim} \text{Gamma}(a_1 + \frac{1}{2}, a_2 + \frac{\beta_{gk}^2}{2\sigma^2\zeta_g})$.

5: Sample w from $w|\zeta_g \overset{\text{ind}}{\sim} \text{Beta}(1 + \#\{\zeta_g = 1\}, 1 + \#\{\zeta_g = v_0\})$.

6: Sample σ^{-2} from

$$\sigma^{-2}|\boldsymbol{\beta}_g, \zeta_g, \nu_{gk}^{-2} \overset{\text{ind}}{\sim} \text{Gamma}\left(b_1 + \frac{n + \sum_{g=1}^G L_g}{2}, b_2 + \frac{1}{2}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\beta}]\right).$$

conditional posteriors for the parameters $\boldsymbol{\beta}_g$, ζ_g, ν_{gk}^{-2}, w and σ^{-2} are given in Algorithm 2. Since we are modelling data from each gene set separately, the estimation can be run in parallel across all pathways making the analysis computationally feasible.

3.5. *False discovery rate-based variable selection.* The MCMC samples explore the distribution of the coefficients corresponding to the principal components of each of the subgroups, as guided by the data. There are different ways of summarizing the information from these MCMC samples. We could use the posterior mode (maximum a posteriori or MAP estimate) of the coefficients β_{gk} and conduct conditional inference based on these point estimates. While this approach provides interpretable point estimates, it does not yield exact zero values as estimates for the coefficients corresponding to principal components not associated with the response; it also does not make use of the complete posterior samples. We use Bayesian model averaging (Hoeting et al. (1999)) which builds inference based on various configurations visited by the MCMC sampler. This approach adequately accounts for the uncertainty in the data and allows for variable selection through downstream inference based on error rates. In this paper we use a multiplicity-adjusted inference for regression on each pathway separately, since in each of these regressions we are trying to infer from the estimates of β_{gk} if they are zero or not. The variable selection also contributes to the multiplicity correction by inducing sparsity. In our Discussion we present results of the false discovery rate (FDR)-based variable selection approach using Bayesian model averaging combined with the MAP estimates.

From the model in equation (6), for each β_{gk}, we obtain S samples $\beta_{gk}^{(1)}, \ldots, \beta_{gk}^{(S)}$ from the posterior distribution. For any given threshold $c > 0$, we can empirically compute $p_{gk} = \frac{1}{S}\sum_{s=1}^S I(|\beta_{gk}^{(s)}| \le c)$ which can be interpreted as the local FDR (Morris et al. (2008)); then, $(1 - p_{gk})$ is the probability that the principal component k from group g significantly impacts the pathway score. Owing to the inherent variable selection property of the group spike-and-slab prior in equation (7), for some g and k it is almost certain that the corresponding β_{gk} is close to zero. The value of p_{gk} for such a β_{gk} is large, and it is almost certain that including such a nonzero coefficient is an inferential error. We also expect some of the coefficients to have moderate values for p_{gk}. Furthermore, we expect to have some coefficients

β_{gk} such that the corresponding p_{gk} are close to zero and they almost certainly influence the pathway score.

Based on this discussion, we assume that the principal component k from group g will be included in the estimation as a significant coefficient if $p_{gk} < \phi$. Note that p_{gk} is a Bayesian q-value or an estimate of the local FDR (Storey (2003)). This threshold ϕ can be determined based on different criteria, such as Bayesian utility considerations (Müller et al. (2004)), or by controlling false-positive/false-negative errors, or the average Bayesian FDR. We determine a threshold ϕ_α, which controls the overall average FDR at some level α, so that we expect only $100\alpha\%$ of the elements of the set $\{(g,k)|p_{gk} < \phi_\alpha\}$ to actually be false-positive inclusions in terms of associations with the pathway scores. To compute the threshold ϕ_α, we sort the posterior inclusion probabilities p_{gk} across all principal components $k = 1, \ldots, L_g$ and groups $g = 1, \ldots, G$, and denote the sorted probabilities as $p_{(l)}$ for $l = 1, \ldots, L = \sum_{g=1}^{G} L_g$. We then compute $\phi_\alpha = p_{(u)}$, where $u = \max\{l^*|\frac{1}{l^*}\sum_{l=1}^{l^*} p_{(l)} \leq \alpha\}$. The set of principal components k from group g with $p_{gk} < \phi_\alpha$, that is, $\{(g,k)|p_{gk} < \phi_\alpha\}$, can then be claimed to be significantly associated with the pathway score based on an average Bayesian FDR of α.

In summary, we start with the MRI scans for each patient and identify the three tumor subregions. Based on these subregions, we construct imaging-based meta-features through PCA on the space of voxel-intensity PDFs using a Riemannian-geometric framework. The resulting PC scores are used as predictors in a regression model with the pathway score as a response. The pathway scores act as genomic markers capturing the enrichment activity in a gene-set. We then use a group structured spike-and-slab prior which captures the natural grouping of the principal components arising from various tumor subregions to identify radiogenomic associations. We use Gibbs sampling for estimation and an FDR-based criterion for variable selection. The complete approach of RADIOHEAD is outlined in Algorithm 3 of Appendix A.

4. Radiogenomic analyses of lower grade gliomas. We consider the imaging and matched genomic data described in Sections 2.1 and 2.2, respectively, which comprises 65 samples. However, four of the 65 samples do not posses segmentation labels for all three tumor subregions and hence are dropped from the analysis, resulting in a final sample size of 61. For each patient, we have $G = 12$ groups arising from four MRI sequences (T1, T1Gd, T2 and FLAIR) and three tumor subregions (NC, ED and ET). First, the density estimates are obtained using the *ksdensity* function in `MATLAB` software which uses an optimal value for estimating normal densities using Silverman's rule as the default bandwidth (Silverman (1986)). We present a sensitivity analysis to assess the differences in the density estimates based on the choice of bandwidth in Section S9 of the Supplementary Material. The results indicate reasonable consistency in the computed density estimates. We then compute the PC scores for these 61 subjects for each of the 12 groups. The number of principal components included within each group is decided such that the included principal components cumulatively explain 99.99% of the total variance. For each of the four imaging sequences, we display the cumulative percentage of variance explained by the principal components in Supplementary Material Figure S4. Note that this cut-off of 99.99% results in choosing a different number of principal components across each group g. Although the choice of this cut-off could include a large number of PCs, any overfitting concerns are addressed by regularization via the spike-and-slab prior that incorporates explicit shrinkage on the regression coefficients (Morris and Carroll (2006), Scheipl, Fahrmeir and Kneib (2012)). We include a total of 143 covariates across all of the 12 groups for the LGG data. We discuss results of a sensitivity analysis to assess the effect of sample composition on the computation of PC bases in Section S8 of the Supplementary Material. We consider only the C-Pathways (Ceccarelli et al. (2016)), and the corresponding pathway scores are computed for the 61 subjects as

described in Appendix B. We provide an R package, *RADIOHEAD*,[4] which includes all relevant code, including the data under consideration, that is, the pathway scores corresponding to C-Pathways and PC scores along with their grouping labels for the 61 LGG subjects.

Prior elicitation and MCMC settings. In our model we have shape (a_1, b_1) and rate (a_2, b_2) hyperparameters corresponding to σ^2 and ν_{gk}^2 in equation (7). We choose these hyperparameters so as to have noninformative/vague priors with $a_1 = a_2 = 0.001$ and $b_1 = b_2 = 0.001$: the mean is 1 with a large variance. The other hyperparameter is v_0, one of the two possible values of the indicator ζ_g. We choose $v_0 = 0.005$ to be close to zero which generates continuous bimodal priors for β_{gk}. We perform a sensitivity analysis based on different values for v_0. These results are included in Section S7 of the Supplementary Material. We run the MCMC chain for 10^5 iterations and discard the first 20,000 samples as burn-in. The final estimates are based on MCMC samples with a thinning of 125 iterations to reduce autocorrelation. In Supplementary Material Figures S9 and S10 we show the posterior densities and trace plots corresponding to randomly chosen β_{gk}s for the transmission of the nerve impulse pathway showing good convergence of the parameters. In Supplementary Material Figure S11 we present boxplots for the potential scale reduction factors (Gelman and Rubin (1992)) computed based on the MCMC samples of β_{gk} from seven different chains. This plot indicates convergence of the MCMC samples across multiple chains.

The results from the regression of these pathway scores on the imaging predictors through the corresponding PC scores are shown in Figure 4. We display only the gene sets that have at least one significantly associated covariate among all of the gene sets in the C-Pathways. Hence, any pathway not shown indicates no significant association between that pathway and the imaging predictors. Similarly, any principal components for any of the 12 groups not listed in this figure are not significantly associated with any of the C-Pathways. Each cell in Figure 4 represents the magnitude of the estimated (MAP) coefficients $\hat{\beta}_{gk}$, and the overlaid symbol denotes its sign; the significantly associated PCs are determined using FDR-based variable selection on the MCMC samples, as described in Section 3.5. For example, we see that the scores of the first principal component of enhancing tumor (`T1_ET.1`) subregion have a significant association with the transmission of nerve impulse gene-set. The average Bayesian FDR is controlled at the level $\alpha = 0.05$; we use a threshold $c = 0.001$ to compute the values for p_{gk} across all of the pathway score regressions. The value of c is chosen such that it is comparable to the bandwidth used to compute the kernel density estimates, which in turn is essential in computing the MAP estimate from the MCMC chain. Diagnostics for the linear model in equation (6) reveal no obvious violations of modelling assumptions (Figures S5–S8 in the Supplementary Material).

Effect of sample composition. As the computation of the pathway scores can be sensitive to the samples in the patient cohort, the associations identified by our model are dependent on the sample composition. A visual illustration of the distribution of the pathway scores (using violin plots) is provided in the Supplementary Material Figures S12–S18. To address this issue, we calibrate the results from our model by computing the pathway scores corresponding to the 61 subjects in three different scenarios. For the calibration we include genomic data from TCGA for additional glioma patients (including glioblastoma multiforme (GBM)). The three scenarios include computing the pathway scores with: (a) the 61 LGG subjects, (b) 516 LGG subjects and (c) 516 LGG and 153 GBM subjects. We build the model in equation (6) for

[4]The R package can be found in a file in the Supplementary Material that accompanies this article (Mohammed et al. (2021b)). For the most recent version of the code, see www.github.com/bayesrx/RADIOHEAD.

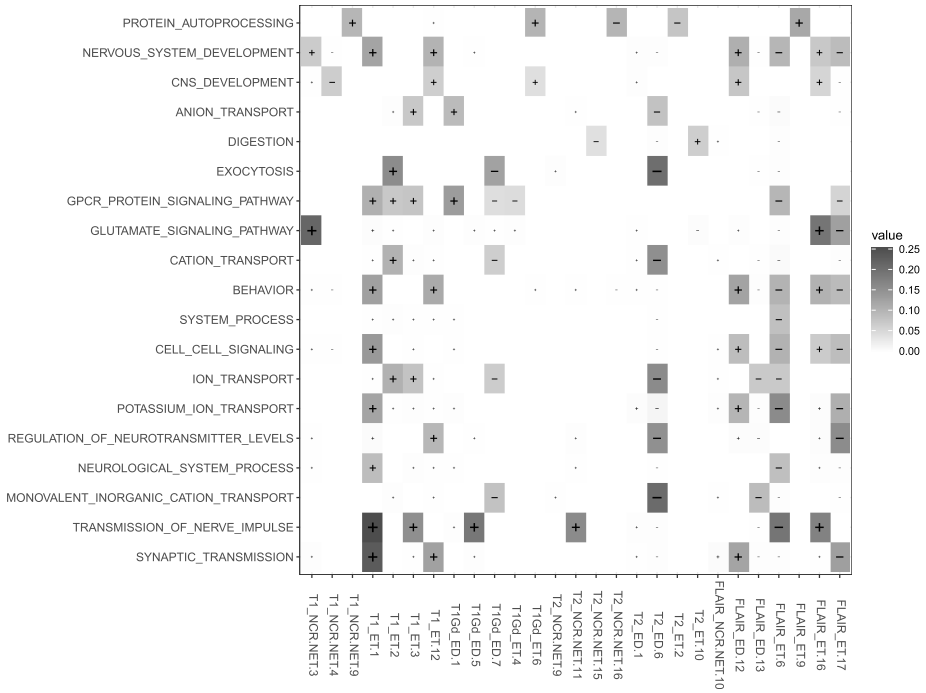


FIG. 4. *Posterior estimates of β_{gk}, after FDR-based variable selection, corresponding to different PC scores across MRI sequences and tumor subregions. Each row corresponds to a pathway from the C-Pathways. The average Bayesian FDR is controlled at the level $\alpha = 0.05$. Values on the gray-scale indicate the magnitude of $\hat{\beta}_{gk}$ and the overlaid symbol (+/−) indicates its sign. The size of the symbol is proportional to the magnitude of $\hat{\beta}_{gk}$. Lack of a symbol denotes a zero estimate indicating no significant association.*

all three cases and carry out the estimation and inference, as described in Section 3. The results presented earlier in Figure 4 correspond to the first case, where the pathway scores were computed with the $n = 61$ LGG subjects only. However, in Supplementary Material Figures S19–S21 we present plots for the estimated coefficients (rows and columns are matched in these plots) when the pathway scores are computed, as described in cases (a)–(c), respectively. These plots are summarized in Figure 5 with pairwise scatterplots of the estimated coefficients from the three different cases; for example, the top-right plot in Figure 5 corresponds to the scatterplot of estimated coefficients when the pathway scores were computed with 61 LGG subjects vs. all 669 glioma subjects (LGG+GBM). The triangles indicate coefficients that are selected as significantly associated in both cases, whereas the circles indicate coefficients that were not selected in one of the two cases. From Figure 5 we see that, across all three pairwise comparisons, we estimate many coefficients to be similar (as indicated by the solid line $y = x$).

Biological associations. We now focus on those pathways and coefficients whose estimates are consistent across all three cases (within a deviation of ± 0.1); these coefficients are the triangles lying within the dotted lines parallel to $y = x$ in Figure 5. We plot these estimated coefficients across the three cases in Figure 6. These plots include pathways related to synaptic transmission, ion transport, glutamate signaling, G protein receptor signaling, exocytosis, nervous system development and protein autoprocessing. Here, we focus on two major findings in terms of the magnitudes of the different associations:

1. The transmission of nerve impulse pathway is associated with the enhancing tumor region from the T1 and FLAIR imaging sequences. The region enhanced in both of the sequences could potentially indicate demyelination due to glioma invasion which could, in turn,

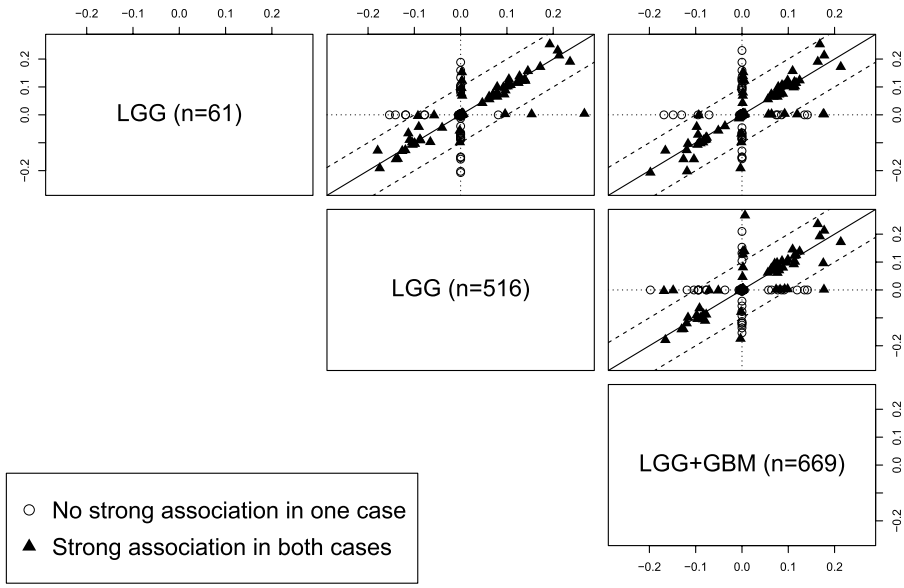


FIG. 5. *Scatterplots of the estimated coefficients when the pathway scores are constructed using*: (*a*) 61 *LGG subjects for which imaging data was available*, (*b*) 516 *LGG subjects from TCGA and* (*c*) 516 *LGG and* 153 *GBM subjects from TCGA*.

lead to disruption in transmission of nerve impulses. This association between the metabolic activity and the infiltrating tumor region is identified by our model. It is also known that neuronal activity promotes glioma growth (Venkatesh et al. (2015)) which is supported by the associations of the transmission of nerve impulse pathway with these imaging predictors.

2. The association of glutamate signaling pathway with the enhancing tumor region from the FLAIR sequence and the necrotic and nonenhancing region from the T1 sequence highlights metabolic activity related to the infiltration of the tumor. In the mammalian central nervous system (CNS), glutamate is a major excitatory neurotransmitter, and experimental evidence suggests that glutamate receptor antagonists may limit tumor growth (Brocke et al. (2010)).

The aforementioned associations indicate that a deeper validation of these phenotypes is essential to better understand tumor etiology which may illuminate more specific nuances. Accordingly, we list some of our other findings:

1. Ion channels are important regulators in cell proliferation, migration and apoptosis, and play an important role in the pathology of glioma. Biological processes can be disrupted, or cancer progression can be influenced, by malfunction and/or aberrant expression of ion channels (Wang et al. (2015)). Our model identifies these connections via associations of the imaging predictors with ion transport pathways, such as potassium ion transport, cell signaling, behavior and anion transport.

2. G protein-coupled receptor (GPCR) signaling affects tumor growth, metastasis and angiogenesis (Cherry and Stella (2014)). Our model identifies this association with the pathway score for GPCR protein signaling.

3. The inhibition of lysosome exocytosis from glioma cells is known to play an important modulatory role in their migration and invasion (Liu, Zhou and Zhu (2012)). Such influences are identified through the radiogenomic association with the exocytosis pathway.

5. Discussion. In this paper we propose a statistical framework for integrating multimodal data from both radiological images and genomic profiles. This model aims to identify

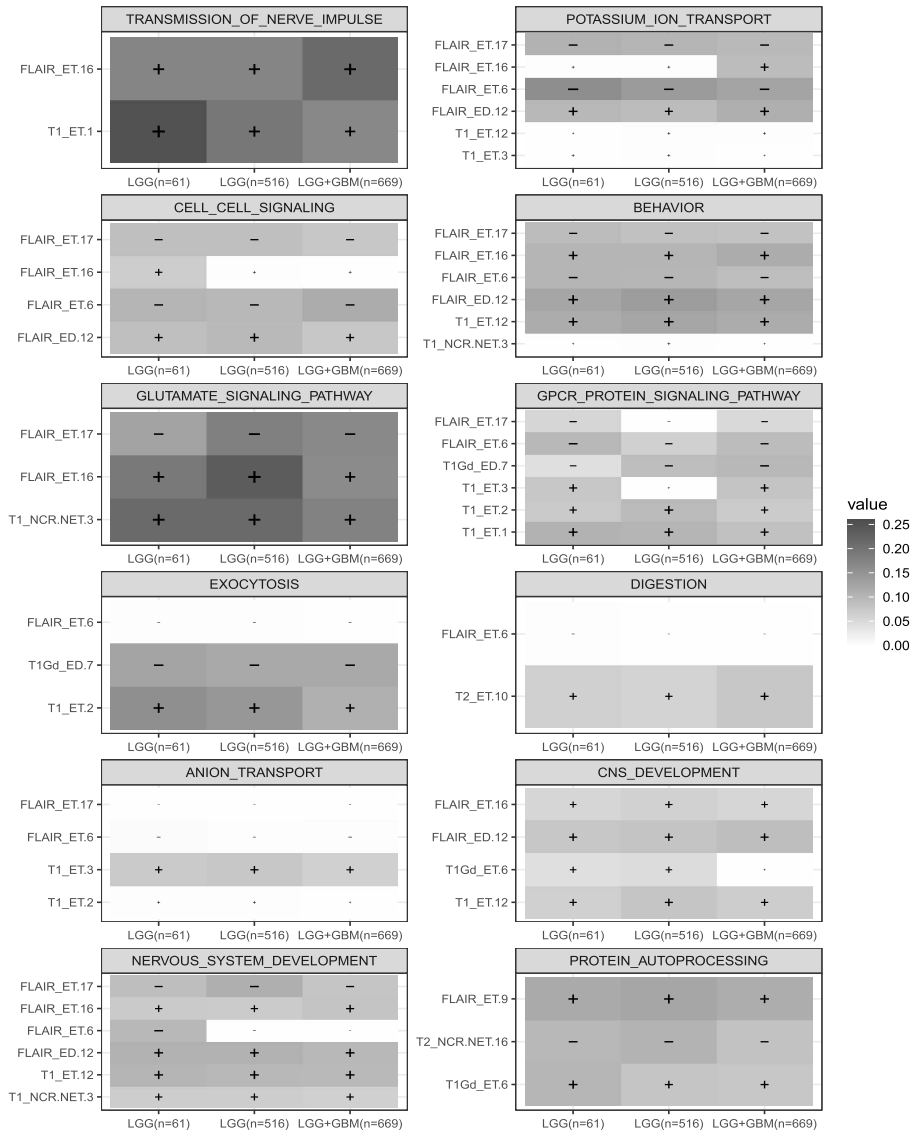


FIG. 6. *Posterior estimates of β_{gk}, after FDR-based variable selection, corresponding to different PC scores across MRI sequences and tumor subregions. Each panel corresponds to a pathway from the C-Pathways. Each column within the panel corresponds to the sample composition used for calibration to compute the pathway scores. The color on the gray-scale indicates the magnitude of $\hat{\beta}_{gk}$ and the overlaid symbol (+/-) indicates its sign. The size of the symbol is proportional to the magnitude of $\hat{\beta}_{gk}$.*

underlying radiogenomic associations, that is, associations between the radiological characteristics extracted from MRI images and molecular underpinnings encoded in gene expression data. Toward this end, from the transcriptomic profiling data we have constructed pathway scores corresponding to those pathways that are known to have influence specifically in LGGs; from the radiological imaging data we have constructed meta-features based on voxel intensities of tumor subregions through PDF-based approaches which effectively capture tumor heterogeneity. These meta-features, constructed from multiple MR sequences, are then used as covariates in a model with pathway scores as responses. We use a Bayesian variable selection strategy by employing a continuous spike-and-slab prior with a grouping structure which accounts for the inherent grouping in the imaging meta-features. This approach identi-

fies many underlying associations between gene pathway activations and image-based tumor characteristics.

We note that, although we incorporate the grouping structure in the RADIOHEAD framework, we are not (explicitly) interested in the associations with the entire PDF. That is, our inference is not based only on groups where $\hat{\beta}_{gk} \neq 0$ for all k in a given group g. Instead, our focus is to identify associations with *any* aspect of the PDFs. The imaging meta-features (PC scores) facilitate evaluation of any underlying associations of the genomic markers with various aspects of the PDFs. Furthermore, inference on the group-level indicator is not feasible in our model setup, as ζ_g is not identifiable. Such an inference is inferior in performance under cases with high within-group sparsity, even under a model which has identifiability of the group-level indicator (Yang and Narisetty (2020)). We demonstrate this using a simulation study described in Section S6 of the Supplementary Material.

Utility in using densities. Data integration from multiple modalities comes with computational and modelling challenges. For the imaging data, MRIs facilitate the characterization of tumor subregions and are obtained from four different sequences. The tumor subregions are represented as voxel intensity values, and standard analyses utilize summaries from the histograms of these voxel intensity values. As an improved alternative, we have used the complete information from the voxel intensities through smoothed histograms (kernel density estimates). Next, we show the benefits of this more comprehensive representation by considering seven different cases as potential predictors: (a) mean, (b) mean, first and third quartiles (Q_1 and Q_3), (c) five-number summary, (d) mean, standard deviation, skewness and kurtosis, (e) deciles, (f) 15 equally spaced percentiles and (g) 20 equally spaced percentiles. The summary statistics are computed across all of the 12 groups separately. In each of these seven cases, we employ the RADIOHEAD pipeline which uses the group spike-and-slab prior and FDR-based variable selection to identify associations. The issue of multicollinearity within the predictors is handled by the shrinkage properties of the spike-and-slab prior. These seven cases also include scenarios where the number of predictors is higher/lower compared to the 143 predictors across groups from the PC scores. The results based on these seven cases are presented in Supplementary Material Figures S22–S25. We see that having just the mean or just the mean, Q_1 and Q_3, does not identify any associations with the pathway scores. However, adding more summary statistics describing the histogram aids in identifying associations. But, as we will see next, the PC scores offer more relevant information about the densities rather than including a larger number of summary statistics as covariates (cases (f) and (g)). Hence, using the PDF-derived PC scores has a higher utility in terms of understanding the pathway scores. In Figure 7 we show the boxplots of the Spearman correlations between computed (observed) and fitted (using estimated coefficients of density-based meta-features/summary statistics from RADIOHEAD) pathway scores, that is, Spearman correlation between y and $X\hat{\boldsymbol{\beta}}$, respectively. These correlations are computed separately by considering cases (a)–(g) and density-based PC scores as predictors.

Additionally, since the computation of the pathway score was dependent on the sample composition, for the case with density-based PC scores as predictors we also include boxplots of Spearman correlations for three different computations of pathway scores, as described in Section 4. The width of these boxplots is proportional to the number of pathways exhibiting significant associations with at least one of the imaging meta-features. In Supplementary Material Figure S25 we also show the Spearman correlations between the computed pathway scores and the fitted pathway scores. This figure demonstrates that we are able to better understand the underlying radiogenomic associations through our modelling approach when the density-based meta-features are considered as covariates. Furthermore,

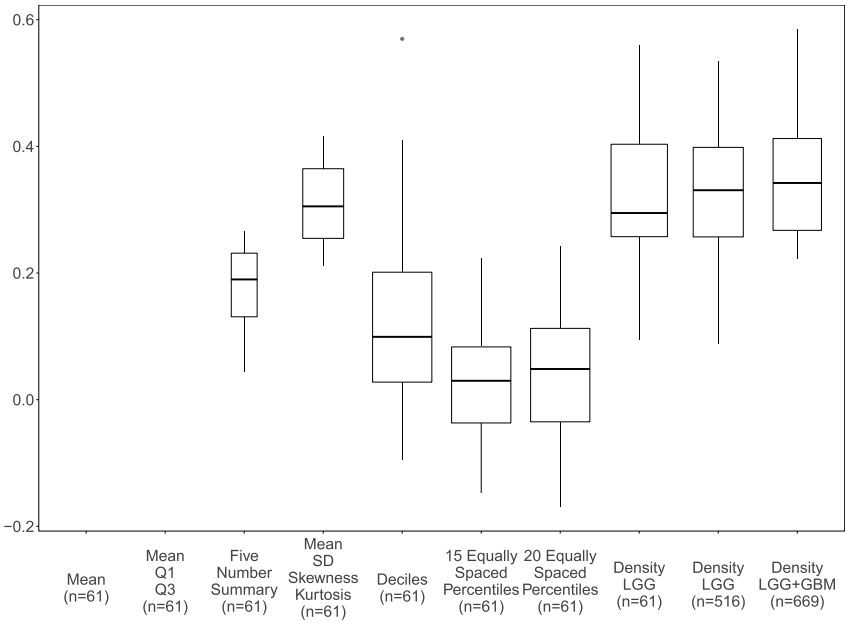


FIG. 7. *Boxplot of Spearman correlations between computed and fitted pathway scores* ($X\hat{\boldsymbol{\beta}}$) *using RADIOHEAD, while different sets of covariates are considered. The width of the boxplots is proportional to the number of pathways exhibiting significant associations with at least one of the imaging meta-features.*

our model can be used in other applications (including other cancers and disease systems) involving imaging and genomic data, as the methodology is readily generalizable to different application domains.

Future work. Although we see promise in the proposed modelling framework to identify radiogenomic associations in LGG, there are certain directions which can be further explored. While using density-based features extracted from multimodal MRI scans does facilitate modelling and provide improved performance, these densities do not explicitly utilize potentially important spatial information in their construction. Incorporating voxel-based spatial information, in addition to intensity values, is nontrivial and will be explored in our future studies. The current model explores linear relationships between the PC scores and pathway scores which could be further extended to investigate nonlinear associations as well. Such analyses will better inform the understanding of the inter- and intra-tumor heterogeneity in LGG. Other directions could be to: (a) extend the framework to incorporate dependencies between pathways (data derived or based on canonical topology) or (b) use gene-level data instead of pathways while incorporating cross-correlations between the genes. Our framework could also be explored further with other forms of pan-omic data, such as epigenomic and proteomic data. Furthermore, our findings could be used to build predictive models for clinical phenotypes (such as survival or progression) that include biologically relevant information based on radiogenomic associations. This provides a statistically-informed strategy to incorporate relevant information for the prediction of clinical phenotypes from complex data.

APPENDIX A: OVERALL OUTLINE OF RADIOHEAD

Here, we describe the algorithm with an outline of the overall approach of this paper to identify the radiogenomic associations by modelling the genomic-based pathway scores using the radiomic-based PC scores.

Algorithm 3 Outline of RADIOHEAD

1: **for** each MRI sequence $M = \text{T1, T1Gd, T2, FLAIR}$ **do**
2: **for** each tumor subregion $R = \text{NC, ET, ED}$ **do**
3: **for** each subject $i = 1, \ldots, n$ **do** Compute the kernel densities $f_i^M(R)$.
4: Compute the principal component scores X_R^M using PCA in Algorithm 1.
5: Consider a pathway of interest and compute pathway scores $\mathbf{y} = (y_1, \ldots, y_n)^\top$ (as described in Appendix B) with the sample $i = 1, \ldots, n$ in the cohort.
6: Bayesian Modelling
a: Model:

$$\mathbf{y} \sim N\Bigg(\sum_{g=1}^{(4\times 3)} X_g \boldsymbol{\beta}_g, \sigma^2 \mathbf{I}_n\Bigg); \qquad \beta_{gk} \overset{\text{ind}}{\sim} N\big(0, \sigma^2 \zeta_g \nu_{gk}^2\big);$$

$$\zeta_g \overset{\text{iid}}{\sim} (1-w)\delta_{v_0}(\zeta_g) + w\delta_1(\zeta_g); \qquad w \sim U(0,1);$$

$$\nu_{gk}^{-2} \overset{\text{iid}}{\sim} \text{Gamma}(a_1, a_2); \qquad \sigma^{-2} \sim \text{Gamma}(b_1, b_2).$$

b: Gibbs sampling for the parameters $\boldsymbol{\beta}_g, \zeta_g, \nu_{gk}^{-2}, w, \sigma^{-2}$ as described in Algorithm 2.
c: FDR-based variable selection as described in Section 3.5 to identify nonzero β_{gk}.

APPENDIX B: COMPUTATION OF PATHWAY SCORES

Instead of directly including the gene expression profiles in the model, we use the corresponding pathway scores. Pathway-based methods offer a significant benefit in terms of interpretability as gene function is exerted collectively and may vary based on several factors, such as disease state, genetic modification or environmental stimuli. As mentioned in Hänzelmann, Castelo and Guinney (2013), using gene sets obtained by organizing genes provides an intuitive and stable context for assessing biological activity. We compute these gene-set scores using gene-set variation analysis (GSVA) (Hänzelmann, Castelo and Guinney (2013)) which is a gene-set enrichment method that estimates variation of pathway activity over a sample population in an unsupervised manner. We provide a brief overview of the GSVA procedure next.

Let Z denote the $p \times n$ matrix of normalized gene-expression values of p genes for n samples ($p \gg n$) and a collection of gene sets $G = \{g_1, \ldots, g_m\}$. The expression profile for gene i is defined as $z_i = (z_{i1}, \ldots, z_{in})$, and each gene set is a subset of genes with its cardinality being denoted by $|g_k|$. First, GSVA evaluates whether a gene i is highly or lowly expressed in sample j in the context of the sample population distribution. An expression-level statistic is computed so that distinct expression profiles can be compared on the same scale. For each z_i, a nonparametric kernel estimation of its cumulative density function is performed using a Gaussian kernel to compute $\hat{F}_{s_i}(z_{ij}) = \frac{1}{n}\sum_{r=1}^{n} \Phi(\frac{z_{ij}-z_{ir}}{s_i})$, where s_i is the gene-specific bandwidth parameter controlling the resolution of the kernel estimation. These statistics $\hat{F}_{s_i}(z_{ij})$ are converted to ranks $r_{(i)j}$ for each sample j and further normalized using $t_{ij} = |\frac{p}{2} - r_{(i)j}|$. We use these t_{ij} to compute a Kolmogorov–Smirnov (KS)-type random walk statistic for $l = 1, \ldots, p$ as

$$\eta_{jk}(l) = \frac{\sum_{i=1}^{l} |t_{ij}|^\tau I(u_{(i)} \in g_k)}{\sum_{i=1}^{p} |t_{ij}|^\tau I(u_{(i)} \in g_k)} - \frac{\sum_{i=1}^{l} I(u_{(i)} \in g_k)}{p - |g_k|},$$

where τ is a parameter describing the weight of the tail and $I(u_{(i)} \in g_k)$ is an indicator taking the value 1 if the gene corresponding to the rank i expression-level statistic belongs to the gene-set g_k. The statistic $\eta_{jk}(l)$ produces a distribution over the genes by identifying

whether the genes in a gene set are more likely to belong to either tail of the rank distribution. This KS-like statistic is now converted into an enrichment score of the pathway using $S_{jk} = \max_l(0, \eta_{jk}(l)) - \min_l(0, \eta_{jk}(l))$. Hänzelmann, Castelo and Guinney (2013) note that S_{jk} has a clear biological interpretation, as it emphasizes genes in pathways that are concordantly activated in one direction only, that is, ones that are either over-expressed or under-expressed relative to the overall population. Low enrichment is shown for pathways containing genes strongly acting in both directions.

Acknowledgments. We would like to extend our gratitude to Kirsten Herold from the Writing Lab at the U-M School of Public Health. We acknowledge the efforts of the anonymous Associate Editor and referees, whose comments have strengthened this paper.

Funding. All of the authors acknowledge support by the NCI grant R37-CA214955. SM was partially supported by Precision Health at The University of Michigan (U-M). SM and AR were partially supported by U-M institutional research funds. SK and KB were partially supported by the NSF grants DMS 1613054 and DMS 2015374. SK was also partially supported by the NSF grant CCF 1740761. VB was supported by NIH grants R01-CA160736, R21-CA220299, P30 CA 46592 and NSF grant 1463233 and start-up funds from the U-M Rogel Cancer Center and School of Public Health.

SUPPLEMENTARY MATERIAL

Supplement to "RADIOHEAD: Radiogenomic analysis incorporating tumor heterogeneity in imaging through densities" (DOI: 10.1214/21-AOAS1458SUPPA; .pdf). All of the details in the text which were referenced as Supplementary Material are provided in this file. This includes details of: (a) computation of the Karcher mean, (b) full posterior distribution, (c) calibration of pathway scores, (d) utility of densities as predictors, (e) inference on group-level indicator; and results of the analysis to assess sensitivity in (f) the parameter estimates based on the choice of hyperparameters, (g) the estimated principal component bases based on the sample composition, and (h) the density estimates based on the choice of bandwidth.

R **package RADIOHEAD** (DOI: 10.1214/21-AOAS1458SUPPB; .zip). The R package RADIOHEAD can be found in this file. For the most recent version of the code, see www.github.com/bayesrx/RADIOHEAD.

REFERENCES

ANDERSEN, M. R., WINTHER, O. and HANSEN, L. K. (2014). Bayesian inference for structured spike and slab priors. In *Advances in Neural Information Processing Systems* 1745–1753.

BAEK, H. J., KIM, H. S., KIM, N., CHOI, Y. J. and KIM, Y. J. (2012). Percent change of perfusion skewness and kurtosis: A potential imaging biomarker for early treatment response in patients with newly diagnosed glioblastomas. *Radiology* **264** 834–843.

BAKAS, S., ZENG, K., SOTIRAS, A., RATHORE, S., AKBARI, H., GAONKAR, B., ROZYCKI, M., PATI, S. and DAVATZIKOS, C. (2015). GLISTRboost: Combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In *International Workshop on Brainlesion*: *Glioma*, *Multiple Sclerosis*, *Stroke and Traumatic Brain Injuries* 144–155. Springer, Berlin.

BAKAS, S., AKBARI, H., SOTIRAS, A., BILELLO, M., ROZYCKI, M., KIRBY, J. S., FREYMANN, J. B., FARAHANI, K. and DAVATZIKOS, C. (2017a). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **4** 170117.

BAKAS, S., AKBARI, H., SOTIRAS, A., BILELLO, M., ROZYCKI, M., KIRBY, J., FREYMANN, J., FARAHANI, K. and DAVATZIKOS, C. (2017b). Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection. Available at https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF.

BHATTACHARYYA, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **35** 99–109. MR0010358

BILIOTTI, L. and MERCURI, F. (2017). Riemannian Hilbert manifolds. In *Hermitian–Grassmannian Submanifolds. Springer Proc. Math. Stat.* **203** 261–271. Springer, Singapore. MR3710849 https://doi.org/10.1007/978-981-10-5556-0_2

BROCKE, K. S., STAUFNER, C., LUKSCH, H., GEIGER, K. D., STEPULAK, A., MARZAHN, J., SCHACKERT, G., TEMME, A. and IKONOMIDOU, C. (2010). Glutamate receptors in pediatric tumors of the central nervous system. *Cancer Biol. Ther.* **9** 455–468.

CECCARELLI, M., BARTHEL, F. P., MALTA, T. M., SABEDOT, T. S., SALAMA, S. R., MURRAY, B. A., MOROZOVA, O., NEWTON, Y., RADENBAUGH, A. et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164** 550–563.

ČENCOV, N. N. (1982). *Statistical Decision Rules and Optimal Inference. Translations of Mathematical Monographs* **53**. Amer. Math. Soc., Providence, RI. MR0645898

CHERRY, A. E. and STELLA, N. (2014). G protein-coupled receptors as oncogenic signals in glioma: Emerging therapeutic avenues. *Neuroscience* **278** 222–236.

CLARK, K., VENDT, B., SMITH, K., FREYMANN, J., KIRBY, J., KOPPEL, P., MOORE, S., PHILLIPS, S., MAFFITT, D. et al. (2013). The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **26** 1045–1057.

DRYDEN, I. L. and MARDIA, K. V. (1998). *Statistical Shape Analysis. Wiley Series in Probability and Statistics: Probability and Statistics*. Wiley, Chichester. MR1646114

FISHBEIN, L., LESHCHINER, I., WALTER, V., DANILOVA, L., ROBERTSON, A. G., JOHNSON, A. R., LICHTENBERG, T. M., MURRAY, B. A., GHAYEE, H. K. et al. (2017). Comprehensive molecular characterization of pheochromocytoma and paraganglioma. *Cancer Cell* **31** 181–193. https://doi.org/10.1016/j.ccell.2017.01.001

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y. et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5** R80.

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* 339–373.

HÄNZELMANN, S., CASTELO, R. and GUINNEY, J. (2013). GSVA: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14** 7. https://doi.org/10.1186/1471-2105-14-7

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* 382–401.

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158 https://doi.org/10.1214/009053604000001147

JUST, N. (2011). Histogram analysis of the microvasculature of intracerebral human and murine glioma xenografts. *Magn. Reson. Med.* **65** 778–789.

JUST, N. (2014). Improving tumour heterogeneity MRI assessment with histograms. *Br. J. Cancer* **111** 2205–2213. https://doi.org/10.1038/bjc.2014.512

KARCHER, H. (1977). Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.* **30** 509–541. MR0442975 https://doi.org/10.1002/cpa.3160300502

KASS, R. E. and VOS, P. W. (2011). *Geometrical Foundations of Asymptotic Inference* **908**. Wiley, New York.

KURTEK, S. and BHARATH, K. (2015). Bayesian sensitivity analysis with the Fisher–Rao metric. *Biometrika* **102** 601–616. MR3394278 https://doi.org/10.1093/biomet/asv026

LIBERZON, A., SUBRAMANIAN, A., PINCHBACK, R., THORVALDSDÓTTIR, H., TAMAYO, P. and MESIROV, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27** 1739–1740.

LIU, Y., ZHOU, Y. and ZHU, K. (2012). Inhibition of glioma cell lysosome exocytosis inhibits glioma invasion. *PLoS ONE* **7** e45910.

MARUSYK, A., ALMENDRO, V. and POLYAK, K. (2012). Intra-tumour heterogeneity: A looking glass for cancer? *Nat. Rev. Cancer* **12** 323–334. https://doi.org/10.1038/nrc3261

MOHAMMED, S., BHARATH, K., KURTEK, S., RAO, A. and BALADANDAYUTHAPANI, V. (2021a). Supplement to "RADIOHEAD: Radiogenomic analysis incorporating tumor heterogeneity in imagine through densities." https://doi.org/10.1214/21-AOAS1458SUPPA

MOHAMMED, S., BHARATH, K., KURTEK, S., RAO, A. and BALADANDAYUTHAPANI, V. (2021b). Code for "RADIOHEAD: Radiogenomic analysis incorporating tumor heterogeneity in imagine through densities." https://doi.org/10.1214/21-AOAS1458SUPPB

MORRIS, J. S. and CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 179–199. MR2188981 https://doi.org/10.1111/j.1467-9868.2006.00539.x

MORRIS, J. S., BROWN, P. J., HERRICK, R. C., BAGGERLY, K. A. and COOMBES, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics* **64** 479–489.

MÜLLER, P., PARMIGIANI, G., ROBERT, C. and ROUSSEAU, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *J. Amer. Statist. Assoc.* **99** 990–1001. MR2109489 https://doi.org/10.1198/016214504000001646

NOUSHMEHR, H., WEISENBERGER, D. J., DIEFES, K., PHILLIPS, H. S., PUJARA, K., BERMAN, B. P., PAN, F., PELLOSKI, C. E., SULMAN, E. P. et al. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17** 510–522.

OMBAO, H., LINDQUIST, M., THOMPSON, W. and ASTON, J. (2016). *Handbook of Neuroimaging Data Analysis*. CRC Press, Boca Raton.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2168993

RAO, C. R. (1992). Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in Statistics* 235–247. Springer, Berlin.

SAHA, A., BANERJEE, S., KURTEK, S., NARANG, S., LEE, J., RAO, G., MARTINEZ, J., BHARATH, K., RAO, A. U. et al. (2016). DEMARCATE: Density-based magnetic resonance image clustering for assessing tumor heterogeneity in cancer. *NeuroImage Clin.* **12** 132–143.

SCHEIPL, F., FAHRMEIR, L. and KNEIB, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *J. Amer. Statist. Assoc.* **107** 1518–1532. MR3036413 https://doi.org/10.1080/01621459.2012.737742

SHINOHARA, R. T., SWEENEY, E. M., GOLDSMITH, J., SHIEE, N., MATEEN, F. J., CALABRESI, P. A., JARSO, S., PHAM, D. L., REICH, D. S. et al. (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage Clin.* **6** 9–19.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0848134 https://doi.org/10.1007/978-1-4899-3324-9

SONG, Y. S., CHOI, S. H., PARK, C.-K., YI, K. S., LEE, W. J., YUN, T. J., KIM, T. M., LEE, S.-H., KIM, J.-H. et al. (2013). True progression versus pseudoprogression in the treatment of glioblastomas: A comparison study of normalized cerebral blood volume and apparent diffusion coefficient by histogram analysis. *Korean J. Radiol.* **14** 662–672.

SRIVASTAVA, A., JERMYN, I. H. and JOSHI, S. H. (2007). Riemannian analysis of probability density functions with applications in vision. In *IEEE Conference on Computer Vision and Pattern Recognition* 1–8.

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398 https://doi.org/10.1214/aos/1074290335

VASAIKAR, S. V., STRAUB, P., WANG, J. and ZHANG, B. (2017). LinkedOmics: Analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* **46** D956–D963.

VENKATESH, H. S., JOHUNG, T. B., CARETTI, V., NOLL, A., TANG, Y., NAGARAJA, S., GIBSON, E. M., MOUNT, C. W., POLEPALLI, J. et al. (2015). Neuronal activity promotes glioma growth through neuroligin-3 secretion. *Cell* **161** 803–816.

VENNETI, S. and HUSE, J. T. (2015). The evolving molecular genetics of low-grade glioma. *Adv. Anat. Pathol.* **22** 94–101. https://doi.org/10.1097/PAP.0000000000000049

VERHAAK, R. G., HOADLEY, K. A., PURDOM, E., WANG, V., QI, Y., WILKERSON, M. D., MILLER, C. R., DING, L., GOLUB, T. et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17** 98–110.

VERHAAK, R. G. W., COOPER, L. A. D., SALAMA, S. S., ALDAPE, K., YUNG, W. K. A. and BRAT, D. J. (2014). Abstract 936: Comprehensive and integrative genomic characterization of diffuse lower grade gliomas. *Cancer Res.* **74** 936–936.

WANG, R., GURGUIS, C. I., GU, W., KO, E. A., LIM, I., BANG, H., ZHOU, T. and KO, J.-H. (2015). Ion channel gene expression predicts survival in glioma patients. *Sci. Rep.* **5** 11593.

XU, X. and GHOSH, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Anal.* **10** 909–936. MR3432244 https://doi.org/10.1214/14-BA929

YANG, X. and NARISETTY, N. N. (2020). Consistent group selection with Bayesian high dimensional modeling. *Bayesian Anal.* **15** 909–935. MR4132654 https://doi.org/10.1214/19-BA1178

YANG, H., BALADANDAYUTHAPANI, V., RAO, A. U. K. and MORRIS, J. S. (2020). Quantile function on scalar regression analysis for distributional data. *J. Amer. Statist. Assoc.* **115** 90–106. MR4078447 https://doi.org/10.1080/01621459.2019.1609969

ZHANG, L., BALADANDAYUTHAPANI, V., MALLICK, B. K., MANYAM, G. C., THOMPSON, P. A., BONDY, M. L. and DO, K.-A. (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 595–620. MR3258055 https://doi.org/10.1111/rssc.12053