Published in final edited form as:
Ann Appl Stat. 2019 ; 13(3): 1957–1988. doi:10.1214/19-aoas1238.

RADIO-IBAG: RADIOMICS-BASED INTEGRATIVE BAYESIAN ANALYSIS OF MULTIPLATFORM GENOMIC DATA

Youyi Zhang[1], **Jeffrey S. Morris**[1], **Shivali Narang Aerry**[2], **Arvind U.K. Rao**[3], **Veerabhadran Baladandayuthapani**[*,3]

[1]The University of Texas MD Anderson Cancer Center

[2]Johns Hopkins University

[3]University of Michigan Ann Arbor

Abstract

Technological innovations have produced large multi-modal datasets that include imaging and multi-platform genomics data. Integrative analyses of such data have the potential to reveal important biological and clinical insights into complex diseases like cancer. In this paper, we present Bayesian approaches for integrative analysis of radiological imaging and multi-platform genomic data, wherein our goals are to simultaneously identify genomic and radiomic, i.e., radiology-based imaging markers, along with the latent associations between these two modalities, and to detect the overall prognostic relevance of the combined markers. For this task, we propose *Radio-iBAG: Radiomics-based Integrative Bayesian Analysis of Multiplatform Genomic Data*, a multi-scale Bayesian hierarchical model that involves several innovative strategies: it incorporates integrative analysis of multi-platform genomic data sets to capture fundamental biological relationships; explores the associations between radiomic markers accompanying genomic information with clinical outcomes; and detects genomic and radiomic markers associated with clinical prognosis. We also introduce the use of sparse Principal Component Analysis (sPCA) to extract a sparse set of approximately orthogonal meta-features each containing information from a set of related individual radiomic features, reducing dimensionality and combining like features. Our methods are motivated by and applied to The Cancer Genome Atlas glioblastoma multiforme data set, where-in we integrate magnetic resonance imaging-based biomarkers along with genomic, epigenomic and transcriptomic data. Our model identifies important magnetic resonance imaging features and the associated genomic platforms that are related with patient survival times.

1. Introduction.

In oncology, it is of critical importance to investigate both inter- and intra-tumor heterogeneity through an in-depth understanding of the complex interplay between genotypes and phenotypes, towards developing rational anti-cancer therapeutic strategies

[*]Correspondence, veerab@umich.edu.

SUPPLEMENTARY MATERIAL
Supplementary Materials for: Radio-iBAG: Radiomics-based Integrative Bayesian Analysis of Multiplatform Genomic Data (doi: 10.1214/00-AOASXXXXSUPP). Description of radiomic feature preprocessing procedure, relative analyzing results of our case study of glioblastoma multiforme (GBM) and hyperparameter setting and initial value specification in MCMC algorithm.

[Felipe De Sousa et al. (2013)]. The increased availability of complementary and matched molecular and imaging data allows for a thorough examination of tumor heterogeneity at multiple levels [Nicolasjilwan et al. (2015), Hu et al. (2017), Gutman et al. (2013)]. Investigations at the molecular level have been tremendously improved by the development of many genomic profiling technologies, including microarrays, next-generation sequencing, methylation arrays, and proteomic analyses. The Cancer Genome Atlas (TCGA) project, aiming to provide more comprehensive information of human cancer genomes by creating an “atlas” of high-throughput multiple genomic profiles across multiple cancers, was launched in 2005 as a publicly funded project [Tomczak, Czerwińska and Wiznerowicz (2015)]. The growing availability of such data has motivated the development of integrative analytical models that incorporate various genomic platforms to detect complex patterns of tumor heterogeneity that have predictive and prognostic ability [Wang et al. (2013)].

While genomic data provide information on the molecular characterization of a disease, imaging modalities such as X-ray radiography, magnetic resonance imaging (MRI), computed tomography, and positron emission tomography provide visual and broad resources for the acquisition of high-quality images and provide complementary quantitative information about the structural aspects of a disease. In the context of cancer, these imaging modalities provide a quantitative basis for detailed assessment of various features of the tumor that are associated with the development and progression of cancer. *Radiomics* is an emerging field with a goal of providing predictive or prognostic information by revealing quantitative mechanistic associations between radiologic images and clinical outcomes [Coroller et al. (2015), Aerts et al. (2014), Ganeshan et al. (2010), Lee et al. (2016)]. Radiomics, in general, involves the extraction and mining of various types of quantitative imaging features that are processed from high-throughput images obtained via different imaging modalities. These imaging features describe different morphological characteristics of a tumor, e.g., tumor shape features such as round or spiculated, total volume or surface area, intensity histogram features that describe the contrast intensity level, and textural features such as energy and entropy that evaluate tumor spatial heterogeneity. In particular, “texture analysis, which applies different statistical models and mathematical transforming methods to further evaluate a tumors intra-lesional heterogeneity, has become an active ongoing area of research [Castellano et al. (2004)]. In the context of glioblastoma multiforme (GBM), several studies have shown that the textural features from perfusion parametric maps provide useful information for predicting patients’ survival times [Lee et al. (2016)] and the features extracted from a gray-level co-occurrence matrix (GLCM) [Haralick, Shanmugam and Dinstein (1973), Castellano et al. (2004)] are effective in discriminating tumor volumetric phenotypes [Chaddad and Tanougast (2016)].

Radiomic and genomic features capture complementary characteristics of the underlying tumor, with radiomics capturing visual phenotypic information in the tumor and genomics capturing its underlying molecular biology. Thus, it is of interest to assess the interrelationships of these two types of features, a task termed *radiogenomics*, and then collectively assess how these inter-related features correlate with clinically relevant endpoints (e.g., survival, progression). From an analytical standpoint, radiogenomic analysis faces several key challenges. First, incorporating complex biological interactive mechanisms, both within and between multiple genomic platforms at the genomic (DNA), transcriptomic (mRNA) and epigenomic (methylation) levels, is understudied in the radiogenomic framework. Second, the highdimensional nature of both the quantitative features of images and genomic markers necessitates proper dimension reduction techniques and feature selection methods. Third, the analysis becomes more complicated when we wish to link clinical outcomes with genomic and radiomic outcomes in addition to modeling associations between the radiomic and genomic measurements to provide potentially biologically and clinically translatable results.

Multiple studies have addressed these challenges to various degrees. Taking advantage of multi-platform genomic data resources, additive models have been developed that treat the features from different platforms in the same models, although not explicitly modeling their interrelationships [Daemen et al. (2009), Lanckriet et al. (2004)]. Wang et al. (2013) proposed an integrative Bayesian analysis framework to integrate multi-platform genomic data using hierarchical models that capture the natural mechanistic relationships among the various molecular resolution levels. Jennings et al. (2012) generalized the method to integrate various types of genomic platforms with a single clinical outcome. These methods effectively capture the biological interaction within different molecular processes, but do not consider high dimensionality in the outcomes. Olivares et al. (2013) extended the above model with multivariate correlated imaging outcomes. This approach models image markers in separate linear models after applying a de-correlating procedure, but does not consider patient-specific clinical outcomes. Stingo et al. (2013) developed an integrative Bayesian modeling approach for imaging-genetics that incorporates the binary disease status as a clinical response, and developed a hierarchical mixture model that can select discriminatory imaging regions of interest and their relevant single-nucleotide polymorphisms simultaneously. Similarly, Batmanghelich et al. (2013) developed a joint probabilistic model of imaging and genetic features associated with disease measures, to provide insights into how imaging biomarkers can serve as intermediate phenotypes when detecting genetic and diagnostic associations. However, these approaches consider only individual platform and thus do not consider the interrelationships among the various molecular resolution levels in their analytical frameworks.

In this paper, we introduce Radio-iBAG: Radiomics-based integrative Bayesian analysis of multiplatform genomic data, an integrative multi-scale Bayesian framework to perform radiogenomic analyses. Our goal is three-fold: first, to detect explicit associations among different genomic platforms at the different molecular levels; second, to treat the radiomic-based biomarkers as an intermediate phenotype (i.e., endo-phenotype), evaluate the molecular underpinnings regulating these biomarkers and finally, evaluate the eventual associations with relevant patient-level clinical outcomes (e.g., survival times). To accomplish these tasks, we construct a multi-level regression-based modeling strategy: a first stage "*genomic model*" detects the complex biological mechanistic relationships among different genomic platforms, a second stage "*radiogenomic model*" subsequently discovers the underlying associations between gene-platform combinations and radiomic biomarkers. To assess clinical relevance, a third level model "*radiogenomic clinical model*" uncovers the associations between clinical outcomes and genomically-driven radiomic markers.

To address the high dimensionality in both the genomic and radiomic datasets, we utilize Bayesian shrinkage-based priors to achieve sparsity and regularization in the high-dimensional covariate space at various hierarchical levels. Specifically, we employ scale-mixture of normal representations, that allow adaptive shrinkage and borrowing strength within and across the different hierarchical levels. Our methodology is motivated by and applied to a GBM case study, wherein we discover multiple radiomic feature groups significantly associated with patients survival times along with their mechanism of action through multi-platform genomics.

In Section 2, we introduce our modeling scheme, major components, modeling methods and biomarker detection for each modeling stage. In Section 3, we illustrate our proposed model on the GBM case study with detailed description of the radiomic features and genomic profile datasets, modeling results and biological interpretations. In Section 4, we draw some conclusions and discuss some future extensions and advancements.

2. Method: Radio-iBAG Model.

2.1. Modeling stages.

As mentioned above, our core construction of the Radio-iBAG model framework consists of a multi-stage Bayesian hierarchical model. In the *genomic model*, we model the complex biological mechanistic relationship among genomic data from different platforms capturing information at various molecular resolution levels (e.g., gene expression, copy number and methylation). Subsequently, we carry the information garnered from the *genomic model* into the second stage, the *radiogenomic model*, to parse out the imaging-genomic correlations, which are then included as predictors in the third stage, the *radiogenomic clinical model*. This procedure delineates the image features that directly affect clinical outcomes, as well as those that appear to be modulated by combinations of genomic factors. This construction allows us to discover strong relationships between imaging and genomics data, among the genomic platforms, and identify which appear to be associated with clinical outcome.

Fig 1 illustrates the general multi-stage modeling scheme. In the first stage, multiplatform genomic data sets are expressed as data matrices: X_{mRNA}, X_{CN}, X_{miRNA} or X_{Methy}, each with rows as samples and columns as genelevel summaries of the respective platforms. In stage II, we consider radiomic features (RFs) that are preprocessed and extracted from imaging data sets, forming a data frame $\mathscr{I}$ with columns as different features and rows as samples. In the final stage, we incorporate into the model the clinical outcome, denoted as $\boldsymbol{Y}$, which is a vector with the number of elements as the sample size. The construction of each modeling stage is explained in detail in the ensuing sections.

A. Genomic Model—Our genomic model involves the integrative modeling of multiplatform genomic data sets. Modern genomics data is comprised of multiple platforms that contain measurements at various molecular resolution levels, from DNA to mRNA to proteins, and including epigenetic levels including alterations like methylation and microRNA (miRNA) that affect mRNA expression. These platforms capture complementary information at the different molecular resolution levels, and together provide a more complete picture of the underlying biology than any one platform. In this paper, we consider three genomic platforms: mRNA, DNA copy number (CN) and miRNA, but the general models we introduce can incorporate any other platforms capturing upstream genetic and epigenetic information, as well. Also, for a specific gene, we take only the genomic platforms mapped with this gene into our model, we do not consider modeling coexpression or coregulation of the neighboring genes or potential transcriptional regulators. Suppose $N_{\mathscr{G}}$ =number of patients with genomic information, J=total number of genomic platforms, and $P_{\mathscr{G}}$=number of target genes. For our particular case, using copy number alteration and miRNA as our upstream platforms, the gene expression level can be modeled and expressed as

$$X_{mRNA_g} = \underbrace{f_1\left(X_{miRNA_g}\right) + f_2\left(X_{CN_g}\right)}_{\text{upstream platform driven}} + \underbrace{\boldsymbol{O_g}}_{\text{explained by other factors}} \qquad (1)$$

where each $f_j(\cdot)$ is a smooth nonparametric function of the corresponding predictor modeled by a penalized spline formulation that allows us to capture flexible non-linear relationships. We assessed the nonlinearity of genelevel fits and show that GAM provides better fit that GLM for most genes (see supplementary materials Section S3). Other types of splines or alternative nonparametric models could also be used. Our analysis in this stage matches the first stage of the iBAG model [Wang et al. (2013)], whereby the gene expression of a given gene is modeled as explained by upstream factors, with the effects of upstream factors modeled nonparametrically as in [Jennings et al. (2015)] via a generalized additive model (GAM) [Hastie and Tibshirani (1990)]. In principle, the model can include any number of upstream (to mRNA) platform types, including methylation, copy number, loss of heterozygosity, methylation, miRNA, and transcription factors, as long as matched data are available.

The terms in the model are described and interpreted as follows:

- X_{mRNA_g} is the expression of gene g with dimension $N_{\mathscr{G}} \times 1$, $g = 1, 2, \ldots, P_{\mathscr{G}}$
- X_{mRNA_g} is an aggregated miRNA expression value that integrates information across miRNAs that have been documented to regulate the expression of gene g. For a given gene, there exist multiple miRNAs that interact with this gene, and here we construct gene-level summaries of these miRNAs that condense their activity into a lower dimension using principal components, as described in detail in Section 3. The gene-level summaries X_{miRNAg} have dimension $N_{\mathscr{G}} \times M_{miRNA_g}$, where M_{miRNA_g} denotes the number of gene-level summary vectors for the g^{th} gene.
- X_{CN_g} are gene-level summaries of the CN alteration for the g^{th} gene with dimension $N_{\mathscr{G}} \times M_{CN_g}$. Similarly, as there are multiple CN alteration values from different markers within the same gene, M_{CN_g} denotes the number of gene-level summary vectors.

- O_g represents the “other” part of gene expression that is not captured by the modeled upstream factors, but instead attributed to other upstream factors not in the model, and is of dimension $N_{\mathscr{G}} \times 1$.

This model is fit separately for each gene, and effectively partitions the information contained in the mRNA measurements into an additive set of components, with each component capturing the part of mRNA expression explained by a particular upstream platform. We call these parts different genomic platform components. For gene *g*, the components can be estimated based on the following formula: $G_{miR_g} = \hat{f}_1\left(X_{miRNA_q}\right), G_{CN_g} = \hat{f}_2\left(X_{CN_g}\right)$ *and* $(G_{O_g} = X_{mRNA_g} - \hat{f}_1\left(X_{miRNA_g}\right) - \hat{f}_2\left(X_{CN_g}\right)$. Repeating the same procedure for all the genes, we combine the components grouped by platform, forming different genomic platform combinations: $\mathbf{G}_{\mathrm{mi}\mathbf{R}} = \left\{G_{miR_1}, G_{miR_2}, \ldots, G_{miR_{P_{\mathscr{G}}}}\right\}, \mathbf{G}_{\mathbf{CN}} = \left\{G_{CN_1}, G_{CN_2}, \ldots, G_{CN_{P_{\mathscr{G}}}}\right\}$ *and* $(\mathbf{G}_{\mathbf{O}} = \left\{G_{O_1}, G_{O_2}, \ldots, G_{O_{P_{\mathscr{G}}}}\right\}$. These combinations represent the gene expression level attributed to miRNA, CN and other factors, respectively, for all $P_{\mathscr{G}}$ target genes of interest.

At times, not all samples with genomic data have radiomic data, as in our GBM example. In that case, we denote $N_{\mathscr{G}\mathscr{I}}(N_{\mathscr{G}\mathscr{I}} \subseteq N_{\mathscr{G}})$ as the sample size of their intersection. We carry forward the corresponding subset of the estimated gene platform combinations $\mathbf{G}_{\mathbf{miR}}$, $\mathbf{G}_{\mathbf{CN}}$, $\mathbf{G}_{\mathbf{O}}$, each with dimension $N_{\mathscr{G}\mathscr{I}} \times P_{\mathscr{G}}$, as predictors into the second-stage *radiogenomic model.*

B. Radiogenomic Model—

The goal of the second stage *radiogenomic model* is to find gene-platform combinations that appear to be associated with radiomic markers, and to partition the radiomic markers into the parts modulated by different gene effects carried from the *genomic model* and those that are not modulated by the modeled genomic factors. The model can be written as

$$\begin{aligned}\mathscr{I} &= \mathscr{I}_g + \mathscr{I}_{\bar{g}} \\ &= \underbrace{\boldsymbol{G_{miR}B_{miR}} + \boldsymbol{G_{CN}B_{CN}} + \boldsymbol{G_{O}B_{O}}}_{\text{Genomically driven}} + \underbrace{\mathscr{I}_{\bar{g}}}_{\text{Non-genomically driven}}\end{aligned} \tag{2}$$

The terms in the model can be expressed and interpreted as follows:

- $\mathscr{I}$ denotes a $N_{\mathscr{G}\mathscr{I}} \times K$ matrix in which K is the number of general RFs (individual radiomic features or Radiomic-meta-Features (RmFs) that we constructed from high dimensional RFs that are highly correlated, described in detail in section 3.2).
- $\boldsymbol{B_{miR}}$ is of dimension $P_{\mathscr{G}} \times K$, with columns as the vectors of the expression effects for corresponding radiomic markers through miRNA;
- $\boldsymbol{B_{CN}}$ is of dimension $P_{\mathscr{G}} \times K$, with columns as the vectors of the expression effects for corresponding radiomic markers through CN;

- $\boldsymbol{B_O}$ is of dimension $P_{\mathscr{G}} \times K$, with columns as the vectors of the expression effects for the corresponding radiomic markers through “other” genomic mechanistic factors;
- $\mathbf{G_{miR}}$, $\mathbf{G_{CN}}$, $\mathbf{G_O}$ are the estimated gene expression components described in part A.

Associations are detected by examining the coefficients’ posterior probabilities based on Markov chain Monte Carlo (MCMC) samples, and estimates given by posterior means (detailed information in Section 2.3). To achieve the segmentation of the radiomic features, we can estimate each component by $\widehat{\mathscr{I}}_{\boldsymbol{CN}} = \mathbf{G}_{\text{CN}}\widehat{\boldsymbol{B}}_{\boldsymbol{CN}}, \widehat{\mathscr{I}}_{\boldsymbol{miR}} = \mathbf{G}_{\boldsymbol{miR}}\widehat{\boldsymbol{B}}_{\boldsymbol{miR}}, \widehat{\mathscr{I}}_{\boldsymbol{O}} = \mathbf{G}_{\boldsymbol{O}}\widehat{\boldsymbol{B}}_{\boldsymbol{O}}$. The final nongene-driven part can be estimated by $\widehat{\mathscr{I}}_{\bar{\boldsymbol{g}}} = \mathscr{I} - \mathbf{G}_{\boldsymbol{CN}}\widehat{\boldsymbol{B}}_{\boldsymbol{CN}} - \mathbf{G}_{\mathbf{miR}}\widehat{\boldsymbol{B}}_{\boldsymbol{miR}} - \boldsymbol{G_O}\widehat{\boldsymbol{B}}_{\boldsymbol{O}}$. We then further carry the above four components into the final stage, the *radiogenomic clinical model*.

C. Radiogenomic Clinical Model

The third-stage model relates the various radiogenomic marker combinations from the second stage model to a clinical outcome (e.g., survival time in our context). The model can be expressed as

$$\boldsymbol{Y} = \mathscr{I}_{\boldsymbol{CN}}\boldsymbol{\alpha_1} + \mathscr{I}_{\boldsymbol{miR}}\boldsymbol{\alpha_2} + \mathscr{I}_{\boldsymbol{O}}\boldsymbol{\alpha_3} + \mathscr{I}_{\bar{\boldsymbol{g}}}\boldsymbol{\alpha_4} + \boldsymbol{\epsilon} \qquad (3)$$

where Y is the clinical outcome with dimension $N_{\mathscr{G}\mathscr{I}\mathscr{C}} \times 1$ and $N_{\mathscr{G}\mathscr{I}\mathscr{C}}(N_{\mathscr{G}\mathscr{I}\mathscr{C}} \subseteq N_{\mathscr{G}\mathscr{I}} \subseteq N_{\mathscr{G}})$ is the sample size of the intersection of the genetic, image and clinical data sets. $\mathscr{I}_{\boldsymbol{CN}}$ is the CN modulated radiomic marker component matrix. Similarly, $\mathscr{I}_{\boldsymbol{miR}}$ denotes the microRNA modulated part; $\mathscr{I}_{\boldsymbol{O}}$ is the part of radiomic features explained by a genomic factor but modulated by something other than CN or miRNA; and $\mathscr{I}_{\bar{\boldsymbol{g}}}$ denotes the part of the radiomic feature not regulated by genes in the model. All four radiomic marker components have the dimension $N_{\mathscr{G}\mathscr{I}\mathscr{C}} \times K$. $\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$, $\boldsymbol{\alpha_3}$, $\boldsymbol{\alpha_4}$ denote the corresponding image marker combination effects. $\boldsymbol{\epsilon}$ is the error term for modeling the clinical outcome. In our GBM application, where the clinical outcome is survival time, we use an accelerated failure time (AFT) model, with Y as the log-transformed survival time [Wei (1992)]. However, for the general analytical process, our outcome $\boldsymbol{Y}$ can involve any clinical measurements with suitable regression model determined by the type of outcome (e.g., logistic models for binary outcomes or Cox proportional hazards models in the presence of censored outcomes.)

Our goal in this final stage is to identify radiomic markers associated with clinical outcome, either modulated by genomic factors or not. We identify these factors by estimation and Bayesian posterior inference of $\boldsymbol{\alpha} = \{\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \boldsymbol{\alpha_3}, \boldsymbol{\alpha_4}\}$, and then can characterize these effects in more detail by tracing information back through the earlier stage models. For example, if a particular radiomic feature is related to clinical outcome through a genomic effect, we can examine the corresponding second stage model to identify which genes are driving such effects, and then the first stage model for those genes to find which upstream platforms most strongly modulate the expression of those genes. In this way, the radio-iBAG model can not only detect clinically relevant radiomic features, but provide a thorough summary of the radiomic-genomic and multi-platform genomic interrelationships that appear to modulate these factors.

2.2. Radio-iBAG Model Estimation.

Our second- and third-stage models involve multiple genes and/or RFs, so it is necessary to introduce sparsity into the regression models to regularize the fitting and to obtain a relatively smaller and more interpretable set of radiogenomic factors that appear to be related to the clinical outcome. This can be done using penalized likelihood or other regularization techniques, but here we use a Bayesian approach and induce sparsity through the prior distributions on the regression parameters.

Some commonly used sparsity priors involve a discrete-mixture prior consisting of a point mass at zero for noise and a continuous density distribution for signals, for example a normal distribution and a point mass at zero [Mitchell and Beauchamp (1988)]. Other types of sparsity priors do not have a zero component, but instead are absolutely continuous distributions that accomplish sparsity via nonlinear shrinkage, which can often be accomplished using a normal scale mixture prior distribution. Examples include a normal-exponential (Bayesian lasso) [Park and Casella (2008)], Horseshoe [Carvalho, Polson and Scott (2010)], generalized double pareto [Armagan, Dunson and Lee (2013)], Dirichlet Laplace [Bhattacharya et al. (2015)], and Normal Gamma [Griffin et al. (2010)]. While the Bayesian lasso, which is a Bayesian analog to the commonly-used lasso [Tibshirani (1996)], is commonly used, it has limited flexibility given it is determined by a single hyperparameter that regulates both sparsity and the tails. We instead use the normal-gamma (NG) prior [Griffin et al. (2010)], which contains a second hyperparameter, and thus can better handle sparsity as well as flexibility to manage the tails and yield to more accurate coefficient estimates, as described and illustrated via multiple simulation settings in [Griffin et al. (2010)]. We apply this prior in both stage II *radiogenomic* and stage III *radiogenomic clinical* models. Further, we allow the sparsity hyperparameters to be indexed by platform, which enables borrowing of strength across genes in determining the desired sparsity and tail levels on a platform specific basis.

To estimate the coefficient vector, for the k^{th} RF, we assign the NG prior distribution to $\boldsymbol{\beta_k} = \left\{\boldsymbol{\beta^k_{miR}}, \boldsymbol{\beta^k_{CN}}, \boldsymbol{\beta^k_{O}}\right\}$, each part of the coefficient vector being assigned with a particular set of the hyperparameters. In this way, we allow priors settings that incoporate multi-scale datasets. More specifically, suppose our genomic platform combination predictors can be expressed as $X = \{\mathbf{G_{miR}}, \mathbf{G_{CN}}, \mathbf{G_O}\}$, then the linear regression model and its hierarchical prior setting can be expressed as

$$\mathscr{I}_{\boldsymbol{k}} = X\boldsymbol{\beta_k} + \mathscr{I}_{\boldsymbol{k\bar{g}}}$$

$$\mathscr{I}_{\boldsymbol{k}} \sim Normal\left(X\boldsymbol{\beta_k}, \sigma_k^2 \boldsymbol{I}_{\boldsymbol{N}_{\mathscr{G}\mathscr{I}}}\right)$$

$$\boldsymbol{\beta_k} \sim Normal\left(\mathbf{0}_{\widetilde{\boldsymbol{P}}}, D_\psi\right)$$

$D_\psi = diag\left(\psi_{1,1}, \psi_{1,2}, \ldots, \psi_{1,P_1}, \psi_{2,1}, \psi_{2,2}, \ldots, \psi_{2,P_2}, \ldots, \psi_{J,1}, \psi_{J,2}, \ldots, \psi_{J,P_J}\right)$, where $\widetilde{P} = P_1 + P_2 + \ldots + P_J$ is the total number of predictors (dimension of X), J denotes the total number of platform types ($j = 1,2,3, \ldots, J$, here our $J = 3$), and P_j denotes the total number of genomic features (each subindexed as g) for the j^{th} genomic platform type. Our estimation of the scale parameters and the main coefficients ($\boldsymbol{\beta_k}$) is processed by applying the NG prior $\psi_{j,g} \sim Gamma\left(\lambda_j, 1/\left(2\gamma_j^2\right)\right)$ for the j^{th} platform. Also, the hyper-prior $\lambda_j \sim exp(c)\ and\ \gamma_j^{-2} \sim Gamma\left(\tilde{a}, \tilde{b}/\left(2\lambda_j\right)\right)$ are assigned to induce greater flexibility and completeness in shrinkage estimation. To complete our prior specification, we assume a conjugate *InverseGamma*(a,b) prior on σ_k^2. Here, we let each genomic platform combination (platform type) share the same set of hyperparameters (λ_j, γ_j^2), thus maintaining the grouped structure at the shrinkage level. For implementation, we utilize Markov Chain Monte Carlo (MCMC) based Bayesian sampling techniques such as Gibbs sampling and Metropolis-Hastings. The posterior means calculated from MCMC samples are used to obtain the parameter estimations, and the corresponding posterior probabilities are used to conduct signal detection. The details for the posterior distribution and MCMC sampling are shown in Appendix A.

For the *radiogenomic clinical model*, we utilize similar NG prior distributions, the only difference being that our group structure is determined by the RF combinations. We assign the same hyperparameters for the partitioned RFs that belong to the same combination/group. Suppose our predictor set estimated from stage II can be expressed as $\mathscr{I} = \left\{\mathscr{I}_{\boldsymbol{CN}}, \mathscr{I}_{\boldsymbol{miR}}, \mathscr{I}_{\boldsymbol{O}}, \mathscr{I}_{\bar{\boldsymbol{g}}}\right\}$, and the effect parameter $\boldsymbol{\alpha} = \{\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \boldsymbol{\alpha_3}, \boldsymbol{\alpha_4}\}$, then the model and prior construction can be expressed as

$$\begin{aligned} &\boldsymbol{Y} = \mathscr{I}\boldsymbol{\alpha} + \epsilon \\ &\boldsymbol{Y} \sim Normal\left(\mathscr{I}\boldsymbol{\alpha}, \sigma^2 \boldsymbol{I}_{\boldsymbol{N}_{\mathscr{G},\mathscr{I}c}}\right) \\ &\boldsymbol{\alpha} \sim Normal\left(\mathbf{0}, D_\psi\right) \end{aligned}$$

$D_\psi = diag(\psi_{1,1}, \psi_{1,2}, \ldots, \psi_{1,K}, \psi_{2,1}, \psi_{2,2}, \ldots, \psi_{2,K}, \ldots, \psi_{J,1}, \psi_{J,2}, \ldots, \psi_{J,K})$, where J denotes the total number of different RF combination types ($j = 1,2,3, \ldots, J$, our $J = 4$), k denotes the RF index ($k = 1,2,3, \ldots, K$). Further, we assign our prior and hyper-prior distributions as $\psi_{j,k} \sim Gamma\left(\lambda_j, 1/\left(2\gamma_j^2\right)\right)$, $\sigma^2 \sim\ InverseGamma\ (u_1, u_2)$, $\lambda_j \sim exp(d)$, and $1/\left(2\gamma_j^2\right) \sim Gamma\left(\tilde{e}, \tilde{f}/\left(2\lambda_j\right)\right)$. Note that for censored sample i, we sample $\boldsymbol{Y_i}$ from complete conditional distribution which is normal distribution with left truncation at t_i that represents the follow-up time. Finally, RF combination selection is based on the posterior probability of the MCMC samples. Details about the posterior distribution and sampling methods are provided in Appendix A.

2.3. Radiomic and Genomic Marker Selection.

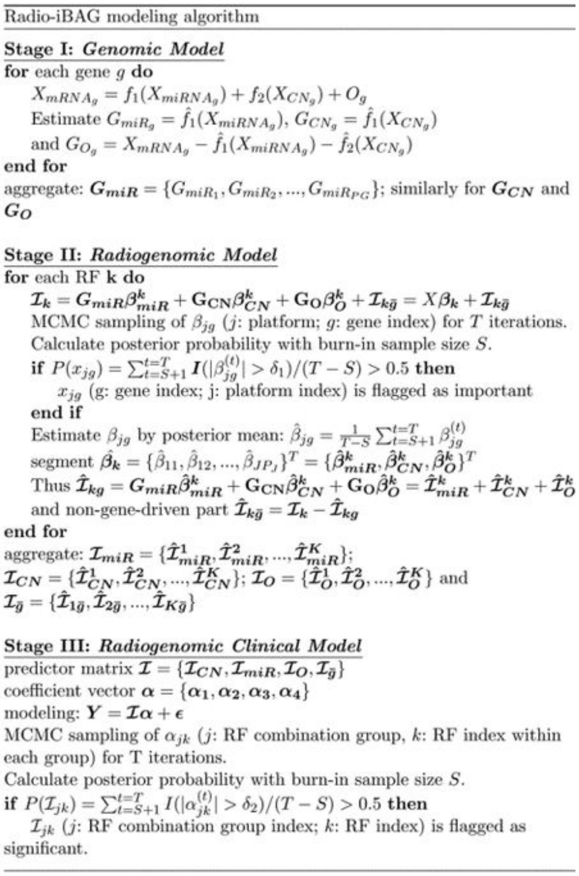
For marker/feature selection we propose a thresholding procedure for the various regression models. Specifically for the *radiogenomic clinical model*, we choose a thresholding criteria considering both the effective size and clinical interpretability. For example, in the GBM case study, we apply the AFT model with the log-transformed survival time as the clinical outcome. In our analysis, considering that the survival times are measured in months, which is comparatively small, we choose to apply $\log_2$-based transformation, which leads to better interpretability and a simpler calculation. Based on this setting, the region for detecting the coefficients of the image markers becomes $\alpha_{jk} \in (-\infty, \delta_{\pm}^{*}) \cup (\delta_{+}^{*}, \infty)$, where we denote $\delta_{\pm}^{*}\ as\ log_2(1 - \delta_2)$ and δ_{+}^{*} as $log_2(1 + \delta_2)$, particularly, α_{jk} is the coefficient of the k^{th} radiomic marker modulated by the j^{th} genomic platform ($j = 1,2, \ldots, J$, $k = 1,2, \ldots, K$). Moreover, δ_2 is determined to achieve the proper effect size and is interpreted as the percentage change in survival time, e.g., for the GBM data analysis, we choose $\delta_2 = 0.05$, which corresponds to 5% change in survival time. More specifically, we denote $P_{+}\left(\mathcal{I}_{jk}\right) = \Sigma_{t = S + 1}^{t = T}\boldsymbol{I}\left(\alpha_{jk}^{(t)} > \delta_{+}^{*}\right)/(T - S)$ and $P_{-}\left(\mathcal{I}_{jk}\right) = \Sigma_{t = S + 1}^{t = T}\boldsymbol{I}\left(\alpha_{jk}^{(t)} < \delta_{\pm}^{*}\right)/(T - S)$ where t denotes the t^{th} MCMC iteration, S denotes the burn-in sample size and T represents the total number of MCMC iterations. We flag $\mathcal{I}_{jk}$ to be positively significant if $P_{+}\left(\mathcal{I}_{jk}\right) > 0.5$ or negatively significant if $P_{-}\left(\mathcal{I}_{jk}\right) > 0.5$ [Barbieri and Berger (2004)].

Analogously, for the *radiogenomic model*, considering δ-fold or larger variation in the response for a unit change in a particular predictor is defined as a standard in the significance detection, which corresponds to $\beta_{jg} \in (-\infty, -\delta) \cup (\delta, \infty)$ and β_{jg} is the coefficient of the j^{th} platform of the g^{th} gene in the analysis. Once a proper threshold δ_1 is determined, the posterior probability is defined as $P\left(x_{jg}\right) = \Sigma_{t = S + 1}^{t = T}\ \boldsymbol{I}\left(\left|\beta_{jg}^{(t)}\right| > \delta_1\right)/(T - S)$, where S is the burn-in sample size and T is the total number of MCMC iterations. Feature x_{jg} in the gene-platform combinations is highlighted to be ‘significant’ if $P(x_{jg}) > 0.5$.

Radio-iBAG modeling algorithm provides a concise summary of Radio-iBAG model implementation and genomic/radiomic marker selection.

3. Radiogenomic Mapping of Glioblastoma Multiforme.

Glioblastoma Multiforme (GBM) is an aggressive and malignant form of primary brain cancer. It is the highest grade glial tumor, with a median survival time of 14.6 months following standard treatment options and typically 3 months without treatment [Stupp et al. (2009)]. Although different treatment approaches that include radiation, surgery and chemotherapy have been developed and applied in clinical practice, the overall mortality rate still remains high, mainly due to the tumors resistance to treatment [Bleeker, Molenaar and Leenstra (2012)] and the complexity of its primary biological mechanism.



```
Radio-iBAG modeling algorithm

Stage I: Genomic Model
for each gene $g$ do
    $X_{mRNA_g} = f_1(X_{miRNA_g}) + f_2(X_{CN_g}) + O_g$
    Estimate $G_{miR_g} = \hat{f}_1(X_{miRNA_g})$, $G_{CN_g} = \hat{f}_1(X_{CN_g})$
    and $G_{O_g} = X_{mRNA_g} - \hat{f}_1(X_{miRNA_g}) - \hat{f}_2(X_{CN_g})$
end for
aggregate: $\boldsymbol{G_{miR}} = \{G_{miR_1}, G_{miR_2}, ..., G_{miR_{PG}}\}$; similarly for $\boldsymbol{G_{CN}}$ and
$\boldsymbol{G_O}$

Stage II: Radiogenomic Model
for each RF k do
    $\boldsymbol{\mathcal{I}_k} = \boldsymbol{G_{miR}}\beta^{\boldsymbol{k}}_{\boldsymbol{miR}} + \mathbf{G_{CN}}\beta^{\boldsymbol{k}}_{\boldsymbol{CN}} + \mathbf{G_O}\beta^{\boldsymbol{k}}_{\boldsymbol{O}} + \boldsymbol{\mathcal{I}_{k\bar{g}}} = X\boldsymbol{\beta_k} + \boldsymbol{\mathcal{I}_{k\bar{g}}}$
    MCMC sampling of $\beta_{jg}$ ($j$: platform; $g$: gene index) for $T$ iterations.
    Calculate posterior probability with burn-in sample size $S$.
    if $P(x_{jg}) = \sum_{t=S+1}^{t=T} \boldsymbol{I}(|\beta_{jg}^{(t)}| > \delta_1)/(T-S) > 0.5$ then
        $x_{jg}$ (g: gene index; j: platform index) is flagged as important
    end if
    Estimate $\beta_{jg}$ by posterior mean: $\hat{\beta}_{jg} = \frac{1}{T-S}\sum_{t=S+1}^{t=T} \beta_{jg}^{(t)}$
    segment $\boldsymbol{\hat{\beta}_k} = \{\hat{\beta}_{11}, \hat{\beta}_{12}, ..., \hat{\beta}_{JP_J}\}^T = \{\boldsymbol{\hat{\beta}^k_{miR}}, \boldsymbol{\hat{\beta}^k_{CN}}, \boldsymbol{\hat{\beta}^k_O}\}^T$
    Thus $\boldsymbol{\hat{\mathcal{I}}_{kg}} = \boldsymbol{G_{miR}}\boldsymbol{\hat{\beta}^k_{miR}} + \mathbf{G_{CN}}\boldsymbol{\hat{\beta}^k_{CN}} + \mathbf{G_O}\boldsymbol{\hat{\beta}^k_O} = \boldsymbol{\hat{\mathcal{I}}^k_{miR}} + \boldsymbol{\hat{\mathcal{I}}^k_{CN}} + \boldsymbol{\hat{\mathcal{I}}^k_O}$
    and non-gene-driven part $\boldsymbol{\hat{\mathcal{I}}_{k\bar{g}}} = \boldsymbol{\mathcal{I}_k} - \boldsymbol{\hat{\mathcal{I}}_{kg}}$
end for
aggregate: $\boldsymbol{\mathcal{I}_{miR}} = \{\boldsymbol{\hat{\mathcal{I}}^1_{miR}}, \boldsymbol{\hat{\mathcal{I}}^2_{miR}}, ..., \boldsymbol{\hat{\mathcal{I}}^K_{miR}}\}$;
$\boldsymbol{\mathcal{I}_{CN}} = \{\boldsymbol{\hat{\mathcal{I}}^1_{CN}}, \boldsymbol{\hat{\mathcal{I}}^2_{CN}}, ..., \boldsymbol{\hat{\mathcal{I}}^K_{CN}}\}$; $\boldsymbol{\mathcal{I}_O} = \{\boldsymbol{\hat{\mathcal{I}}^1_O}, \boldsymbol{\hat{\mathcal{I}}^2_O}, ..., \boldsymbol{\hat{\mathcal{I}}^K_O}\}$ and
$\boldsymbol{\mathcal{I}_{\bar{g}}} = \{\boldsymbol{\hat{\mathcal{I}}_{1\bar{g}}}, \boldsymbol{\hat{\mathcal{I}}_{2\bar{g}}}, ..., \boldsymbol{\hat{\mathcal{I}}_{K\bar{g}}}\}$

Stage III: Radiogenomic Clinical Model
predictor matrix $\boldsymbol{\mathcal{I}} = \{\boldsymbol{\mathcal{I}_{CN}}, \boldsymbol{\mathcal{I}_{miR}}, \boldsymbol{\mathcal{I}_O}, \boldsymbol{\mathcal{I}_{\bar{g}}}\}$
coefficient vector $\boldsymbol{\alpha} = \{\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \boldsymbol{\alpha_3}, \boldsymbol{\alpha_4}\}$
modeling: $\boldsymbol{Y} = \boldsymbol{\mathcal{I}\alpha} + \boldsymbol{\epsilon}$
MCMC sampling of $\alpha_{jk}$ ($j$: RF combination group, $k$: RF index within
each group) for T iterations.
Calculate posterior probability with burn-in sample size $S$.
if $P(\mathcal{I}_{jk}) = \sum_{t=S+1}^{t=T} I(|\alpha_{jk}^{(t)}| > \delta_2)/(T-S) > 0.5$ then
    $\mathcal{I}_{jk}$ ($j$: RF combination group index; $k$: RF index) is flagged as
significant.
```

Currently, at the molecular level, TCGA provides data sets with multiple genomic platforms, including methylation, CN alteration, and gene expression. Studies based on TCGA platform have identified distinct molecular subclasses of GBM, resembling stages in neurogenesis that are relevant to prognosis [Verhaak et al. (2010)]. Also, with the availability of standardized medical image annotations from The Cancer Imaging Archive (TCIA), multiple studies currently focus on the detection of radiomic imaging variables associated with clinical outcomes [Chaddad and Tanougast (2016), Kickingereder et al. (2016)]. Relevant studies have shown that quantitative imaging features extracted from different modalities provide strong prognostic information [Nicolasjilwan et al. (2015), Lee et al. (2016)].

The availability of such large-scale data resources (TCIA and TCGA) makes it feasible to perform radiogenomic mapping in GBM to explore the complex associations between molecular features and imaging features for this particular cancer type. In this section, we apply our integrative multi-stage Bayesian hierarchical model with the data from patients with GBM and matched with TCGA and TCIA platforms, to discover radiogenomic associations characterizing these data and identify RmFs and genomic markers associated with GBM prognosis. More details of the genomic and imaging data sets are provided hereafter.

3.1. Data Description.

3.1.1. Radiomic and clinical data description.—Among 304 GBM patients with available genomic records, 78 matched patients ($N_{\mathscr{G}\mathscr{J}}$=78) also have MRI T1-weighted post contrast images and T2-weighted fluid attenuated inversion recovery (T2-weighted FLAIR) images available from TCIA for texture analysis. Image preprocessing procedures, including steps such as non-uniformity normalization (N3) correction, registration, segmentation, isotropic voxel-reslicing and image filtering, were performed prior to texture feature extraction. For this analysis, we derived textural features from the axial 2D slice that has the largest tumor area [Zhou et al. (2014)]. Our textural features were obtained from a two-step process: 1) Image filtering, 2) Haralick features[1] derivation [Haralick, Shanmugam and Dinstein (1973)] [Haralick (1979)] and summary measures calculation. These image pre-processing steps as well as the texture feature calculations are described in detail in the Supplementary Section S2.1.

For the radiomic data set, we had 972 RFs that could be categorized into 20 groups based on how they were calculated. The group names and corresponding descriptions are provided in the Supplementary S2.2. They cover the features of both T2-weighted FLAIR and T1-weighted post-contrast MRI modalities with different type of features: texture features, histogram features and regional features and with two types of ratio based normalization methods.

For clinical outcomes, we utilized overall survival times (in months) as the response in our integrative analysis. For the clinical model, we used data from $N_{\mathscr{G}\mathscr{J}\mathscr{C}} = N_{\mathscr{G}\mathscr{J}} = 78$ GBM patients with matching multi-platform genomic, radiomic, and clinical data, and with 9 patients having censored clinical outcomes. We applied the AFT model using the $\log_2$ transformed survival time $\log_2(T_i)$ as the response, where T_i is the survival time in months after diagnosis for patient i, and imputed the survival time for censored samples simultaneously.

3.1.2. Genomic data description.—Our gene expression data set is level 3 (summarized per gene), and was downloaded and processed by TCGA Assembler [Zhu, Qiu and Ji (2014)] with open-source software and related instructions available in public. The CN data set is level 2 (probe-level) data obtained from TCGA Portal from the HG CGH 244A platform with normalized records of CN alteration for each probe. The miRNA data set was also acquired from TCGA Portal with 534 miRNA records and 575 samples in total.

In our analysis, we focus on 49 genes that are members of signaling pathways that have previously been detected associated with GBM (RTK/PI3K, P53, and RB pathways [Network et al. (2008)] and 304 patients ($N_{\mathscr{G}} = 304$) with records available for mRNA, CN and miRNA. The sample sizes and the specific types of the raw datasets for all genomic platform, radiomic data and clinical data are illustrated via diagram in Supplementary Fig S3 with description in Section S2.3. The genomic datasets used in the first stage are all

[1] Features generated using variòus metrices of the co-occurence matrices are called "Haralick features" after the publication of [Haralick, Shanmugam and Dinstein (1973)].

continuous and the descriptions of the raw data structure (for 304 samples) of different genomic platforms are given in below:

- mRNA (304 × 49) contains gene expression levels for each gene and each patient.
- Copy number (304 × 491) contains the CN alteration data (columns) for each sample (rows). There exist multiple copy number markers per gene, and the columns of the data set are sorted by gene. Also, one gene, HRAS, does not have CN alteration information, thus, any variance of gene expression contributed by CN changes will be captured by the factor "others" in this analysis; in other words, for gene HRAS, the corresponding column in matrix $\boldsymbol{G_{CN}}$ is set as zero.
- miRNA (304 × 522) contains miRNA values for each gene (column) and patient (row) based on the miRNA-mRNA interaction membership matrix, with records coming from targetHub [Manyam et al. (2013)], which collected miRNA-mRNA interaction records based on 5 external databases, and multiMiR [Ru et al. (2014)] is based on 14 external databases, including validation databases, prediction databases and drug-associated databases. There exist multiple miRNA records corresponding to one gene, and the columns of the miRNA data set are ordered by gene.

We wish to obtain gene-level summaries for each platform based on these raw data sets. Considering that a given gene can contain multiple values from different markers for both miRNA and CN alteration records, and including all these records into the *genomic model* is computationally expensive and inefficient, the gene-level summaries that can be carried into the modeling stage need to be generated. There are different ways to obtain gene-level summaries, e.g., taking the average, selecting the top most correlated records, or extracting the top principal components via PCA. For the analysis of GBM data, CN alteration and miRNA, in each case, we perform PCA on the genomic platform data set mapped to a gene and keep the top principal components with cumulative variance that explain up to 90% of the total variance. In this way, we regard the remaining records as capturing most of the information of the genomic platform data. Specifically, for gene *g*, the gene-level summaries for each platform can be expressed as X_{miRNAg} and X_{CN_g}, which have been denoted in Methods. Our genomic model is conducted based on these three data sets, X_{mRNA_g}, X_{miRNA_g} and X_{CN_g}.

As described in Section 2.1, our *genomic model* uses the GAM to fit the model and estimate the partitioned mRNA that is modulated by different genomic platforms. To implement the GAM algorithm, we utilized Woods R package "mgcv" and exploited its option for the automatic smoothness selection for the penalty parameter based on generalized cross-validation [Wood (2001)]. Subsequently, for each gene, we calculated the proportion of the mRNA variance explained by each platform. We assume that if a genomic platform does not explain much variation in mRNA expression, it will not have a significant impact on image features. Thus, for $\mathbf{G_{miR}}$, $\mathbf{G_{CN}}$ and $\mathbf{G_O}$, we filtered out the genomic platform features that explain less than 10% of the total variance of gene expression, leaving the remaining features to be carried forth into the *radiogenomic model*.

3.2. Estimation of Radiomic-meta-Features.

One of the critical challenges in fitting the radiogenomic model is the high dimensionality and redundancy of the set of radiomic features (RFs). In our GBM case study, the preprocessed RF data set has 972 features, and contains many features within the same type of radiomic class but with different settings, e.g. filtering scales. Thus, there are extensive correlation among many RFs with high magnitudes up to 0.99, as can be seen in the correlation heatmap shown in Supplementary Fig S1. Facing these challenges, we utilize a new radiomic strategy of empirically constructing radiomic meta features (RmFs) comprised by a linear combination a sparse subset of highly correlated RFs. Each RmF defines a factor capturing one aspect of the fundamental structure in the radiomic features, and together the relatively small number of RmFs retain a vast majority of information contained in the set of 972 RFs. To our knowledge, this strategy has not been applied in the radiomics literature to date, and may be useful in other contexts. We construct the RmFs by applying sparse principal component analysis (sPCA) [Zou, Hastie and Tibshirani (2006)] which incorporates a regularization technique such as the lasso or elastic net to induce sparsity in the principal component loadings. This has the advantage of interpretability over general principal components that do not in general yield sparse loadings, in the case of our GBM application yielding RmFs that are reasonably intuitive and interpretable (see Section 3).

This algorithm offers a parsimonious way to obtain more comprehensive representation of radiomic features, which contain the maximum information of the original radiomic data. While not strictly orthogonal like PCs, the SPCs are approximately orthogonal so it is reasonable to model these RmFs as independent imaging features in the second stage radiogenomic model. The sparse loadings for the RmFs for our GBM application are shown in Fig 2, and by contrast, the non-sparse loadings for ordinary PCA are shown in the Supplementary Fig 2S.

Let $\mathscr{M}$ be an $N_{\mathscr{G}\mathscr{I}} \times P$ matrix (typically with $P > \ > N_{\mathscr{G}\mathscr{I}}$) with the rows being the subjects and the columns the P (=972) RFs. The sparse PCA is applied as follows:

- Apply ordinary PCA to $\mathscr{M}$ and record the number of top principal components with the cumulative variance explaining up to $100(1-\alpha)\%$ (eg. 90%) of the total variance. Each PC is regarded as a linear combination of the original features with its loadings can be estimated by regressing the PC on these features. Sparsity in loadings results from adding regularization terms in the regressions.
- The general sPCA algorithm and its numerical computation procedure are described by [Zou, Hastie and Tibshirani (2006)]. In most cases, the number of features is typically much bigger than the sample size; hence, the simplified version of the general sPCA described in the paper should be applied here. The mathematical formulation of sPCA is illustrated in the Supplementary file Section S2.2. To implement the algorithm, we utilized the R package “elasticnet” [Zou and Hastie (2005)], with K (the number of principal components based on the ordinary PCA) principal components and vectors of λ_j (L1 norm regularization parameter for each loading vector). The parameter λ_j can be

chosen by cross validation, or various values can be tried to find one that results in the desired level of sparsity.

Suppose V is our final matrix of loadings with dimensionality $P \times K$, the projected imaging features matrix (PC score matrix) is then $\mathscr{I}_{(N_{\mathscr{G},\mathscr{I}} \times K)} = \mathscr{M}_{(N_{\mathscr{G},\mathscr{I}} \times P)} V_{(P \times K)}$. We define the vectors of this matrix as RmFs, which contain the majority of the information of the original radiomic data. These features are further regarded as predictors in the analysis of the *radiogenomic clinical model*.

3.3. Results Using the Radio-iBAG Model.

3.3.1. Radiomic-meta-Feature Estimation.—We conducted sPCA with the regularization parameter $\lambda = 2.5$ for each principal component, leading to 22 top principal scores that explain 80.7% of the total variance. We also explored large range of λ with the corresponding loadings and the cumulative variance that are showed in Supplement Section S6. We chose $\lambda = 2.5$ given its balance in the sparsity of the loadings which leads to good interpretation and the cumulative variance that could be attained. We call these 22 principal scores Radiomic-meta-features (RmFs) as discussed in section 3.2. To summarize the RmFs, Fig 2 plots a heatmap of the squared loading proportions within the 20 broad categories of RFs to show which feature types dominate each RmF.

This figure reveals that many of the RMFs appear to be interpretable in the sense of summarizing certain aspects of the images, including morphological imaging features that can be directly visualized, eg. unformity, tumor area, mean intensity, etc. To further illustrate their interpretability, we pick out three example RmFs and in Fig 3 plot T1-Post Contrast images for the four tumors with highest and lowest values of the corresponding RmF scores, rescaled to [0,1]. RmF 21 has the largest loading values for feature categories indicating tumor area (T1_Region, F_Region). The first column of Fig 3 shows that samples with higher values of RmF 21 tend to have larger tumor area. RmF 14 has non-zero loadings inversely proportional to pixel intensity variance measures, and thus can be construed as representative of local pixel heterogeneity. From the second column, it is evident that larger RmF 14 (smaller variance) leads to lower local pixel heterogeneity. The third column of Fig 3 shows the sorted RmF 17, whose loadings are dominated by the imaging intensity histogram feature "uniformity", which represents how non-uniform of the overall gray-level pixel intensities. The gray level of the magnified tumor region shows that when RmF value gets larger, the tumor surface gets more non-uniform. These RmFs quantitatively capture these fundamental features of the images.

We use these RmF as quantifications of the radiomic data in our modeling, with the radiomic model fit to Rmf matrix $\mathscr{I}$, which is of dimension 78 × 22, with RmFs as columns and subjects as rows.

3.3.2. Radio-iBAG Modeling Results.—Our model shows proper convergence and it is not sensitive to the choice of the hyperparameters based on the model checking results respectively described and shown in Supplementary file Section S4.2 and Section S7. After model fitting, the information about the prognostic radiogenomic features can be explored in the following sequence: RmFs that significantly influence the survival time, either positively or negatively, are selected using our criteria outlined in Section 2.3. For each selected RmF, the important RF groups comprising this RmF can be identified by evaluating the sPCA loading information as shown in Fig 2. To obtain significant genes and genomic platforms for the selected RmFs, we then trace back to the radiogenomic model and the genomics model, to identify which genes, if any, are associated with that RmF, and then which upstream platforms appear to be modulating the genomic effect. The specific results for each stage are described here.

Radiomic Results.: We use posterior probabilities to detect significant radiomic signals as well as genomic platform factors in both stage II and stage III based on the median probability criteria described in Section 2.3. Fig 4 shows the posterior probabilities used to select the positively and negatively significant clinical RmF combinations. The results show that more RmF are positively significant for the prognostic outcome, with 1 unit change leading to at least 5% increase in survival time ($\delta_2 = 0.05$, the results using alternative thresholding $\delta_2 = 0.02$ and $\delta_2 = 0.08$ are shown in Supplementary File Section S4.3). Also negatively selected significant RmF combinations have the interpretation as 1 unit change leading to at least 5% decrease in survival time. From Fig 4(a) and Fig 4(b), we see that RmF 7 and RmF 8 have a positively significant influence on the survival time, with the parts that are modulated by genes through their copy number effects (G_{CN}). RmF 1 and RmF 3 are negatively associated with survival with the parts that are modulated by genes through their copy number effects (G_{CN}). RmF 1, RmF 4, RmF 8, RmF 18 and RmF 21 are positively related with survival via genemic effects not modulated by CN and miRNA. RmF 10 and RmF 19 are negatively associated with survival through genomic effects not modulated by miRNA nor CN. RmF 13, RmF 14 and RmF 21 are positively associated with survival apart from genetic modulated factors.

To interpret the flagged RmFs, we turn to Fig 2 (as well as the Supplementary Material.xls file), which illustrates how much variance each RF group contributes to the corresponding RmF combinations. RmF 8 is found to be positively associated with survival through CN effects, and Fig 2 shows that RmF 8 is dominated by the the RF groups “T1_LoG_Tex_R1” and “T1_LoG_Tex_Fine”. RF names and their brief interpretations are shown in the table of Section S2.2 of the Supplementary Materials. In general, we see that texture features derived from T1-weighted post contrast images processed with R1 normalizing approach tend to be more significant, and based on the actual loading values (Supplementary Material Excel file), we found Haralick features to be important, including sum average and inverse difference moment. As another example, RmF 19 modulated by gene expression not explained by miRNA or CN changes ($\boldsymbol{G_O}$) is detected to be negatively associated with survival, and for this RmF the dominant RF group is the Haralick features extracted from T2-weighted FLAIR images, especially with exact features named cluster shade, cluster prominence, energy and contrast. Additionally, RmF 21, which is found to be positively associated with survival both through genomic factors explained by“other” and the non-gene driven part. Further checking found that RmF 21 is associated with T1-weighted post contrast and T2- weighted FLAIR tumor areas. This indicates tumor area, as one of the major regional features, associated with the survival time and seemingly moderated by gene expression of signaling pathway genes, in part regulated by some genomic transcriptional factors other than CN or miRNA.

Radiomic Biological Significance.: In general, more radiomic features extracted from T1-weighted post contrast MRI images are selected to be clinically significant and most of them appear to be associated with genomic effects in signaling pathways. This is not unexpected given the fact that recent studies in literature showed that genomics are expected to be most related to T1-weighted post contrast images rather than T2-weighted FLAIR preprocessed ones. More specifically, RmF 14, which mostly captures the contrast margin of the enhancing MRI image, the magnitude and the loading information (included in Supplementary excel file), indicate that higher texture feature *sum of average* or lower texture feature *sum of variance* derived from the contrast of the edges leads to longer survival times. The detection of RmF 10 shows that histogram features, derived from T2-weighted FLAIR image pixel intensity and representing the global summary of the enhancement, are selected to be primarily affecting patients' survival. It has been shown that the overall intensity is correlated with blood flow vasoconstriction. Moreover, we see associations with several key genes PDGFRA and TP53, with genomic transcriptional factors that affect the uniformity of the overall pixel intensity. In addition, RmF 21, with both genomic transcriptional factor driven part and non-gene driven part, are also selected to be significant in influencing patients' survival time. Since the region feature, more specifically, tumor area, that captures most of the variation of RmF 21, our conclusion indicates that tumor area calculated from both T1-weighted post contrast and T2-weighted FLAIR images, are clinically important, larger area results in shorter survival times.

Genomic Results.: For the selected RmFs, we trace back to stage II and obtain the regulating genes that significantly affect the RmFs through specific genomic platforms (CN, miRNA or others), as shown in Fig 5, Fig 6 and Fig 7. To flag genes as associated wtih the RmFs, we compute the posterior probabilities of the magnitude exceeding a pre-specified threshold. For our analysis, we present the results with the setting $\delta_1 = 0.075$ in this section since it gave us the best balance between the signal and sparsity (the results when setting the threshold $\delta_1 = 0.05, 0.075, 0.1$ are shown in supplementary file Section S4.3). For the flagged genes, we traced back through the stage I model to acquire the percentage values (marked in blue) that represent the proportion of the mRNA variance that is explained by the corresponding genomic platform. For example, RmF 8 modulated by the CN combination is selected to be important, referring to the top left graph in Fig 5, genes GRB2, PIK3CB, SPRY2 and TP53 are selected as important, affecting RmF 8 through CN alteration. For gene SPRY2, 20.3% of its mRNA is explained by CN alteration. Also, genes PDGFRA and TP53 are selected as significantly influencing RmF 10 via other transcriptional factors. For TP53 in particular, the genomic factors (other than CN and miRNA) explains 86.2% of its mRNA variance.

For the results of stage I, after performing genomic modeling and filtering out the genomic platforms that did not explain much of the variance of gene expression (discussed earlier), there are 92 markers in the remaining gene-platform combinations (miRNA: 12; CN: 31; Others: 49) being carried into stage II, the radiogenomics model, as predictors. Fig 8 presents the overall genomic and radiomics results: RmF 7 and RmF 8, modulated by CN, are selected to be positively associated with survival time. Furthermore, 4 genes (GRB2, PIK3CB, SPRY2 and TP53), with their part of gene expression (mRNA) explained by CN alteration, are detected as being significantly associated with these RmFs. For the transcription modulated part, RmF 10 and RmF 19 are detected as being negatively important and associated with gene ERBB2, TP53 and PDGFRA; while RmF 21 is positively significant and associated with genes CDKN2A, ERBB2, MDM2, PDGFRA, PIK3C2G and PIK3CG. For the non-gene-driven factors, RmF 13, RmF 14 and RmF 21 are positively significant.

Genomic Biological Significance.: Result table shows that gene EGFR is selected to be significant for multiple flagged RmFs. It agrees with the literature that the aberrations and gene expression of EGFR, with its full name as “epidermal growth factor receptor”, have been associated with the classical subtype of GBM among 4 major subtypes (proneural, classical, mesenchymal and neural), defined based on transcription data analysis [Verhaak et al. (2010)]. This particular subtype accounts for ~ 25%−30% of GBM cases. The amplification of the EGFR gene is the most common genomic change that leads to overexpression of the receptor variant III (EGFRvIII), and 20% or less EGFRvIII in GBM is significantly related to longer overall patient survival [Montano et al. (2011)]. Moreover, PDGFRA is another gene which has been flagged as important for multiple RmFs. It was found that for the proneural subtype, platelet-derived growth factor (PDGF) receptors (PDGFRAs) have been found to represent gene [Verhaak et al. (2010)]. Also, PDGFR has been positively correlated with patient survival time and its critical role in oncology has been well described in the context of gliomas [Nazarenko et al. (2012)]. Gene TP53 is selected to significantly influence RmF 8 via its mRNA explained by CN, and specifically, TP53 has been found to be the main hub gene that acts as tumor suppressor through comparative analyses of CN and mRNA expression in GBM tumor and xenografts Hodgson et al. (2009). The study illustrated that loss of TP53 function in GBM leads to transcriptional upregulation in gene expression network.

MDM2 is a commonly known oncogene that inhibits the tumor suppressor TP53; its overexpression and amplification have been studied through the analysis of CN alterations and gene expression profiles in previous studies [Yin et al. (2009)]. The gene CDKN2A, with other transcription factors accounting for its expression, has been found to be significantly associated with tumor area for both T1-weighted post contrast and T2-weighted FLAIR. CDKN2A belongs to the RB1 pathway, serves as a cyclin-dependent kinase inhibitor, and has been detected to be important [Solomon et al. (2008)]. It has been reported that loss of RB1 expression occurs in up to 25% of glioblastomas. Changes in RB1 expression have been associated with alterations in tumor cell proliferation and survival [Kim et al. (2011), Nakamura et al. (1997)]. Also, the assessment of RB1 promoter hypermethylation showed a clear correlation between the loss of RB1 expression and promoter hypermethylation [Nakamura et al. (2001)]. Analysis of GBM on the molecular level (TCGA data), using fluorescence in situ hybridization and immunohistochemistry, showed that alterations in RB1 occur more commonly in the proneural subtype of GBM.

4. Discussion and Conclusion.

This article presents the radio-iBAG model, a general framework for multi-scale integrative Bayesian analysis of radiogenomics data. Our hierarchical models incorporate biological mechanistic relationships among multiple genomic platforms, radiomic feature analysis and radiogenomic analysis with relevant clinical outcomes. There are three key features of this modeling strategy: (1) Multiple genomic platform profiles are incorporated in the radiogenomics framework; (2) For model fitting, high dimensionality with a pre-defined group structure in the covariates can be addressed through Bayesian shrinkage priors. In particular, we choose the normal gamma prior due to its flexibility in both shrinkage and parameter estimation; and finally (3) Investigating the relationship between clinical outcomes and radiomic features containing genomic information allows us to identify clinically significant genes, radiomic features and more importantly, the hidden associations between these two data modalities. We note that although our modeling strategy is motivated by an imaging genomics study in GBM, our methodology is general and can be applied to any other disease domain which generates quantitative imaging data with matched genomic data. This includes neurological diseases where the imaging features could be computed from structural or functional neuroimaging assays [Azadeh et al. (2016)].

We applied our methodology to the analysis of radiomic and genomic data sets of GBM. Our model analyzed the relationship between the survival times of patients and the RmFs modulated by various gene-platform combinations. Our analysis identified several RmFs that significantly impact survival times as well as identified the key radiomic features driving these factors. These results revealed that some of the most prognostically important radiomic features include tumor area, intensity histogram uniformity, and Haralick features derived from the GLCM, including energy contrast, inverse difference moment, and entropy for both T1-weighted post-contrast and T2-weighted flair images. Based on the results of modeling the relationship between RmFs and multi-platform genomic measurements, for each detected RmF, we subsequently identified which gene-platform combinations modulated that RmF. This allows us to detect prognostic RmFs modulated by upstream molecular platforms such as copy number, microRNA or other factors. Furthermore, we were able to identify which genes and platforms were driving these associations.

In summary, the advantages of applying integrative analysis of multiplatform genomic profiles in this framework are illustrated through the hierarchical back-tracking, which allows us to discover strong associations and interrelationships among the clinical, image, and genomic factors that may help elucidate the underlying biology. Most of the significant genes identified in our analysis have been shown to be biologically and clinically relevant to GBM molecular subclassifications, cancer development, or therapeutic strategies.

Several possible future extensions and generalizations could be explored based on our Radio-iBAG framework. For example, in our methodology, we applied a multi-stage modeling strategy in doing integrative analysis. A possible advancement may be using a joint model to capture all the relationships among different platforms simultaneously and maintain the detective power with interpretable results. One other possible direction may be incorporating pathway information as another hierarchy into the model structure or considering more complicated biological mechanisms at molecular level, e.g, hidden associations between a gene and other platforms of the neighboring genes, into the modeling framework, e.g. as considered in McGuffey et al. (2018). Another possible future extension may be involving histological images of different tumor tissue regions as another imaging modality into the study, which will provide more pathological based interpretable radiogenomic relationships along with relevant clinical outcomes. We leave these tasks for future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

VB's work was supported by NIH grant R01-CA160736, NIH grant R0116073 and NSF grant 1463233, Cancer Center Support Grant (CCSG) from NIH/NCI (P30-CA016672). JSM's work was supported by NCI grants CA-178744, CA160736, and CA-016672, and NSF grant 1550088 and the M.D. Anderson Cancer Center Colorectal Cancer Moonshot, and the Del and Dennis McCarthy Distinguished Professorship funds. We also acknowledge Dr. Ganesh Rao, M.D, for GBM MRI segmentation, imaging preprocessing and radiomic feature extraction. We also would like to acknowledge the editor, associate editor and the reviewers who provided valuable comments to our manuscript, which led to more sufficient research descriptions and results.

APPENDIX A: APPENDIX SECTION

A. Full conditi006Fnal posterior distribution.

The general posterior distribution of the coefficient parameter as well as other hyperparameters for the regression model either for the *Radiogenomic Model* or the *Radiogenomic Clinical Model* are shown below.

Consider the linear regression formula: $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

In the *radiogenomic model*, $\boldsymbol{Y}$ denotes the specific RF, X is the matrix of the genomic platform combinations. In the *radiogenomic clinical model*, $\boldsymbol{Y}$ denotes the clinical outcome, X represents the RF combinations modulated by different gene expression parts explained by different genomic platforms. The full posterior distributions are

$$\begin{aligned}
&\boldsymbol{\beta}\Big|rest{\sim}Normal\left(\left(X^TX+\sigma^2D_\tau^{-1}\right)^{-1}X^TY,\left(X^TX+\sigma^2D_\tau^{-1}\right)^{-1}\sigma^2\right)\\
&\sigma^2\Big|result{\sim}IG\left(a+n/2,b+(Y-X\beta)^T(Y-X\beta)/2\right)\\
&\psi_{ji}\Big|rest{\sim}GIG\left(a=\gamma_j^{-2},b=\beta_{ji}^2,p=\lambda_j-\frac{1}{2}\right)\\
&\lambda_j\Big|rest{\sim}\left(1/\lambda_j\right)^{\tilde{a}}exp\left\{-\tilde{b}\gamma_j^{-2}/\left(2\lambda_j\right)-c\lambda_j\right\}\times\prod_{i=1}^{p_j}\psi_{ji}^{\lambda_j}/\left\{\left(\Gamma\left(\lambda_j\right)\right)^{p_j}\left(2\gamma_j^2\right)^{p_j\lambda_j}\right\}\\
&\gamma_j^{-2}\Big|rest{\sim}Gamma\left(\tilde{a}+p_j\lambda_j,\left(\tilde{b}/\lambda_j+\sum_{i=1}^{pj}\psi_{ji}\right)/2\right)
\end{aligned}$$

If applying to the *radiogenomic clinical model*, j denotes the RF combination that are modulated by the gene expression that is explained by the j^{th} platform, and k represents the

k^{th} RF; if applying to the *radiogenomic model*, j is the genomic platform type index, i is the gene index.

Specifically, λ_j is sampled through the Metropolis-Hastings method, the proposed family is $exp\left(\sigma_\lambda^2 z\right)\lambda_j$, and z comes from the standard normal distribution. The acceptance rate is controlled between 20% and 30%.

REFERENCES

Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications 5.

Armagan A, Dunson DB and Lee J (2013). Generalized double Pareto shrinkage. Statistica Sinica 23 119. [PubMed: 24478567]

Azadeh S, Hobbs BP, Ma L, Nielsen DA, Moeller FG and Baladandayuthapani V (2016). Integrative Bayesian analysis of neuroimaging-genetic data with application to cocaine dependence. NeuroImage 125 813–824. [PubMed: 26484829]

Barbieri MM and Berger JO (2004). Optimal predictive model selection. Annals of Statistics 870–897.

Batmanghelich NK, Dalca AV, Sabuncu MR and Golland P (2013). Joint modeling of imaging and genetics. In Information Processing in Medical Imaging 766–777. Springer. [PubMed: 24684016]

Bhattacharya A, Pati D, Pillai NS and Dunson DB (2015). Dirichlet-Laplace priors for optimal shrinkage. Journal of the American Statistical Association 110 1479–1490. [PubMed: 27019543]

Bleeker FE, Molenaar RJ and Leenstra S (2012). Recent advances in the molecular understanding of glioblastoma. Journal of Neuro-oncology 108 11–27. [PubMed: 22270850]

Carvalho CM, Polson NG and Scott JG (2010). The horseshoe estimator for sparse signals. Biometrika asq017.

Castellano G, Bonilha L, Li L and Cendes F (2004). Texture analysis of medical images. Clinical Radiology 59 1061–1069. [PubMed: 15556588]

Chaddad A and Tanougast C (2016). Extracted magnetic resonance texture features discriminate between phenotypes and are associated with overall survival in glioblastoma multiforme patients. Medical & Biological Engineering & Computing 1–12.

Coroller TP, Grossmann P, Hou Y, Velazquez ER, Leijenaar RT, Hermann G, Lambin P, Haibe-Kains B, Mak RH and Aerts HJ (2015). CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiotherapy and Oncology 114 345–350. [PubMed: 25746350]

Daemen A, Gevaert O, Ojeda F, Debucquoy A, Suykens JA, Sempoux C, Machiels J-P, Haustermans K and De Moor B (2009). A kernel-based integration of genome-wide data for clinical decision support. Genome Medicine 1 39. [PubMed: 19356222]

Felipe De Sousa EM, Vermeulen L, Fessler E and Medema JP (2013). Cancer heterogeneitya multifaceted view. EMBO Reports 14 686–695. [PubMed: 23846313]

Ganeshan B, Abaleke S, Young R, Chatwin CR and Miles KA (2010). Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. Cancer Imaging 10 137–143. [PubMed: 20605762]

Griffin JE, Brown PJ et al. (2010). Inference with normal-gamma prior distributions in regression problems. Bayesian Analysis 5 171–188.

Gutman DA, Cooper LA, Hwang SN, Holder CA, Gao J, Aurora TD, Dunn WD Jr, Scarpace L, Mikkelsen T, Jain R et al. (2013). MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. Radiology 267 560–569. [PubMed: 23392431]

Haralick RM (1979). Statistical and structural approaches to texture. Proceedings of the IEEE 67 786–804.

Haralick RM, Shanmugam K and Dinstein IH (1973). Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on 6 610–621.

Hastie TJ and Tibshirani RJ (1990). Generalized additive models 43 CRC Press.

Hodgson JG, Yeh R-F, Ray A, Wang NJ, Smirnov I, Yu M, Hariono S, Silber J, Feiler HS, Gray JW et al. (2009). Comparative analyses of gene copy number and mRNA expression in glioblastoma multiforme tumors and xenografts. Neuro-oncology 11 477–487. [PubMed: 19139420]

Hu LS, Ning S, Eschbacher JM, Baxter LC, Gaw N, Ranjbar S, Plasencia J, Dueck AC, Peng S, Smith KA et al. (2017). Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. Neuro-Oncology 19 128–137. [PubMed: 27502248]

Jennings EM, Morris JS, Carroll RJ, Manyam GC and Baladandayuthapani V (2012). Hierarchical Bayesian methods for integration of various types of genomics data. In Genomic Signal Processing and Statistics,(GENSIPS), 2012 IEEE International Workshop on 5–8. IEEE.

Jennings EM, Morris JS, Manyam GC, Carroll RJ and Baladandayuthapani V (2015). Bayesian models for flexible integrative analysis of multiplatform genomics data Book: Integrating omics data: statistical and computational methods. Cambridge University Press;1 edition.

Kickingereder P, Burth S, Wick A, Götz M, Eidel O, Schlemmer H-P, Maier-Hein KH, Wick W, Bendszus M, Radbruch A et al. (2016). Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. Radiology 280 880–889. [PubMed: 27326665]

Kim Y-H, Lachuer J, Mittelbronn M, Paulus W, Brokinkel B, Keyvani K, Sure U, Wrede K, Nobusawa S, Nakazato Y et al. (2011). Alterations in the RB1 Pathway in Low-grade Diffuse Gliomas Lacking Common Genetic Alterations. Brain Pathology 21 645–651. [PubMed: 21470325]

Lanckriet GR, De Bie T, Cristianini N, Jordan MI and Noble WS (2004). A statistical framework for genomic data fusion. Bioinformatics 20 2626–2635. [PubMed: 15130933]

Lee J, Jain R, Khalil K, Griffith B, Bosca R, Rao G and Rao A (2016). Texture feature ratios from relative CBV maps of perfusion MRI are associated with patient survival in glioblastoma. American Journal of Neuroradiology 37 37–43. [PubMed: 26471746]

Manyam G, Ivan C, Calin GA and Coombes KR (2013). targetHub: a programmable interface for miRNA–gene interactions. Bioinformatics 29 2657–2658. [PubMed: 24013925]

McGuffey et al, C. R. M. G. B. V. Morris JS (2018). piBAG: Pathway-based integrative Bayesian modeling of multiplatform genomics data. Under Review.

Mitchell TJ and Beauchamp JJ (1988). Bayesian variable selection in linear regression. Journal of the American Statistical Association 83 1023–1032.

Montano N, Cenci T, Martini M, DAlessandris QG, Pelacchi F, RicciVitiani L, Maira G, De Maria R, Larocca LM and Pallini R (2011). Expression of EGFRvIII in glioblastoma: prognostic significance revisited. Neoplasia 13 1113–IN6. [PubMed: 22241957]

Nakamura M, Konishi N, Tsunoda S, Hiasa Y, Tsuzuki T, Inui T and Sakaki T (1997). Retinoblastoma protein expression and MIB-1 correlate with survival of patients with malignant astrocytoma. Cancer 80 242–249. [PubMed: 9217037]

Nakamura M, Yonekawa Y, Kleihues P and Ohgaki H (2001). Promoter hypermethylation of the RB1 gene in glioblastomas. Laboratory Investigation 81 77–82. [PubMed: 11204276]

Nazarenko I, Hede S-M, He X, Hedrén A, Thompson J, Lindström MS and Nistér M (2012). PDGF and PDGF receptors in glioma. Upsala journal of Medical Sciences 117 99–112. [PubMed: 22509804]

Network CGATR et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455 1061. [PubMed: 18772890]

Nicolasjilwan M, Hu Y, Yan C, Meerzaman D, Holder CA, Gutman D, Jain R, Colen R, Rubin DL, Zinn PO et al. (2015). Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. Journal of Neuroradiology 42 212–221. [PubMed: 24997477]

Olivares RJ, Rao A, Rao G, Morris JS and Baladandayuthapani V (2013). Integrative analysis of multi-modal correlated imaging-genomics data in glioblastoma. In Genomic Signal Processing and Statistics (GENSIPS), 2013 IEEE International Workshop on 5–8. IEEE.

Park T and Casella G (2008). The bayesian lasso. Journal of the American Statistical Association 103 681–686.

Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, Bowler R, Mahaffey S, Rossi S, Calin GA, Bemis L et al. (2014). The multiMiR R package and database: integration of microRNA–target interactions along with their disease and drug associations. Nucleic Acids Research 42 e133–e133. [PubMed: 25063298]

Solomon DA, Kim J-S, Jenkins S, Ressom H, Huang M, Coppa N, Mabanta L, Bigner D, Yan H, Jean W et al. (2008). Identification of p18INK4c as a tumor suppressor gene in glioblastoma multiforme. Cancer Research 68 2564–2569. [PubMed: 18381405]

Stingo FC, Guindani M, Vannucci M and Calhoun VD (2013). An integrative Bayesian modeling approach to imaging genetics. Journal of the American Statistical Association 108 876–891.

Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJ, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K et al. (2009). Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. The Lancet Oncology 10 459–466. [PubMed: 19269895]

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 267–288.

Tomczak K, Czerwińska P and Wiznerowicz M (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology 19 A68. [PubMed: 25691825]

Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP 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. [PubMed: 20129251]

Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G and Do K-A (2013). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics 29 149–159. [PubMed: 23142963]

Wei L (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in Medicine 11 1871–1879. [PubMed: 1480879]

Wood SN (2001). mgcv: GAMs and generalized ridge regression for R. R news 1 20–25.

Yin D, Ogawa S, Kawamata N, Tunici P, Finocchiaro G, Eoli M, Ruckert C, Huynh T, Liu G, Kato M et al. (2009). High-resolution genomic copy number profiling of glioblastoma multiforme by single nucleotide polymorphism DNA microarray. Molecular Cancer Research 7 665–677. [PubMed: 19435819]

Zhou M, Hall L, Goldgof D, Russo R, Balagurunathan Y, Gillies R and Gatenby R (2014). Radiologically defined ecological dynamics and clinical outcomes in glioblastoma multiforme: preliminary results. Translational Oncology 7 5–13. [PubMed: 24772202]

Zhu Y, Qiu P and Ji Y (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. Nature Methods 11 599–600. [PubMed: 24874569]

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 301–320.

Zou H, Hastie T and Tibshirani R (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics 15 265–286.

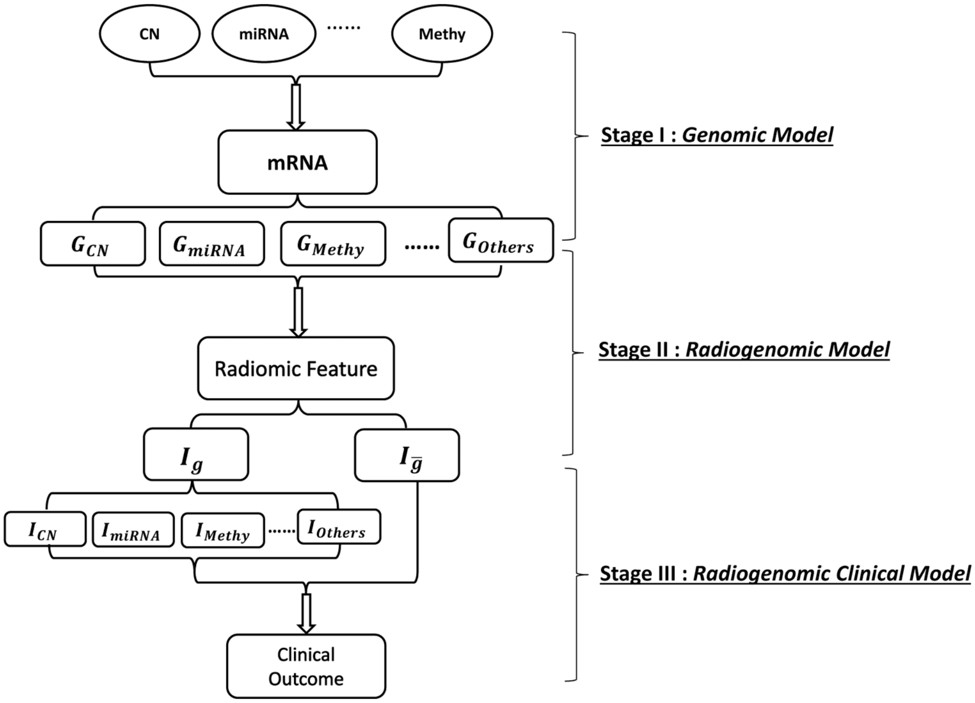


Fig 1: Schematic representation of the multi-stage modeling process.

In stage I, for each gene, model the relationship between mRNA and different upstream genomic platforms and partition mRNA expression into multiple parts explained by different genomic platforms, CN: copy number alteration, miRNA: microRNA, Methy: methylation, Others: gene expression that is explained by other factors; In stage II, for each radiomic marker, apply Bayesian hierarchical model and partition the radiomic marker into different parts modulated by multiple mRNA factors that are explained by various gene-platform combinations and regard the residual as a non-gene-driven part denoted as $I_{\bar{g}}$; In stage III, apply Bayesian hierarchical model to investigate the relationship between segmented radiomic factors with clinical outcome.

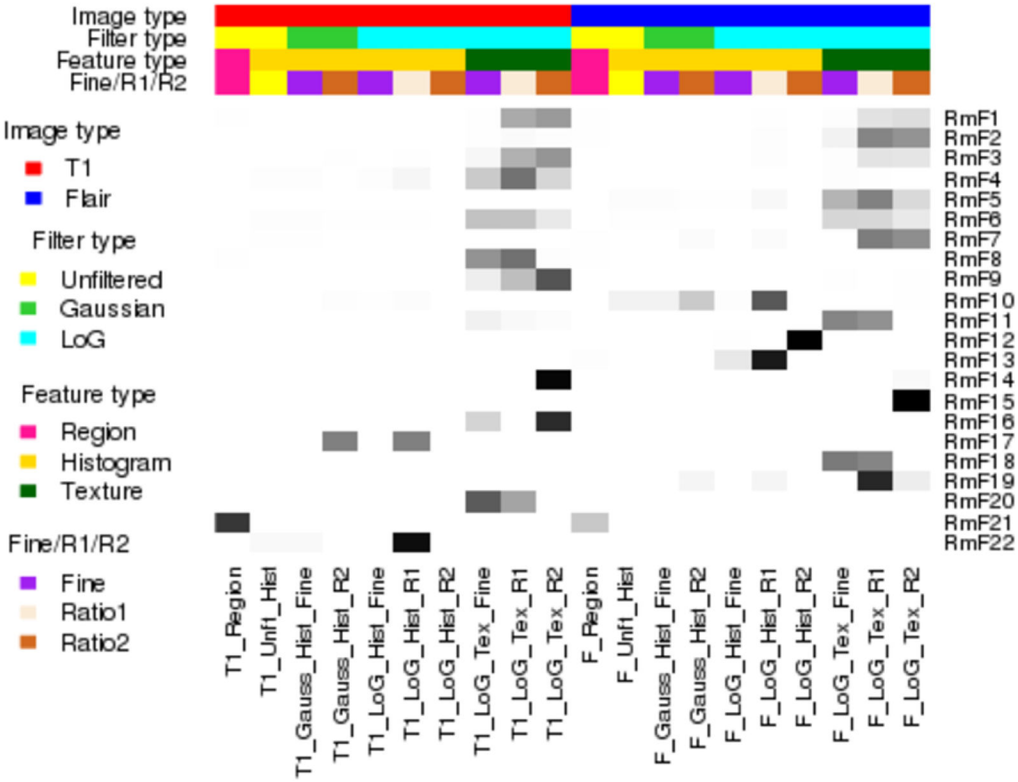


Fig 2: Squared loading proportion for each RF group.

For each of the 22 radiomic-meta-features (RmFs), the sum of the squared loadings of each group is calculated, divided by the total sum of the squared loadings, which equals exactly 1. The heatmap shows this values in grey level, interpreted the RF group importance for each RmF. The grey level ranging from white to black matches the proportional values ranging from 0 to 1.

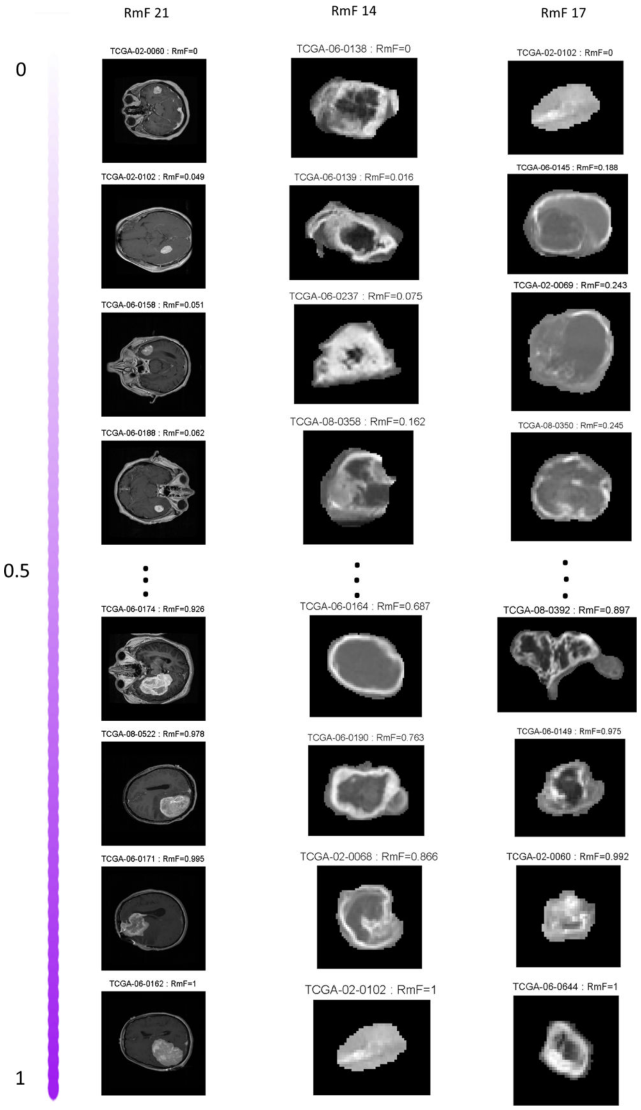
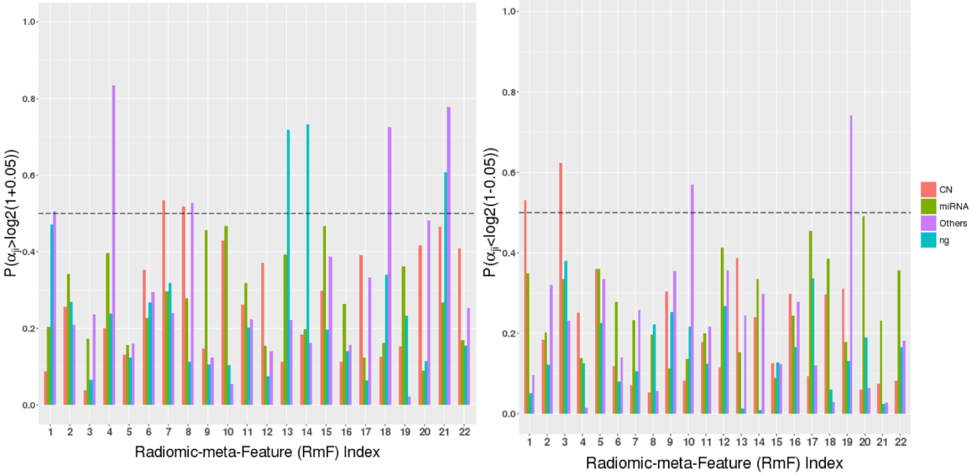


Fig 3:

T1-Post Contrast images are shown based on the sorted results of 3 representitive RmFs: RmF 21 mainly accounts for tumor area; RmF 14 mainly represents tumor pixel heterogeneity; RmF 17 represents tumor uniformity. The RmF values are all scaled from 0 to 1.



(a) positively significant RmF combinations

(b) negatively significant RmF combinations

Fig 4:

Results of stage III (*radiogenomic clinical model*): Detecting postively and negatively significant RmF combinations. Each RmF is segmented into 4 parts, of which 3 parts are modulated by different genomic platform combinations denoted as $\mathcal{I}_{\boldsymbol{CN}}, \mathcal{I}_{\boldsymbol{miR}}$, and $\mathcal{I}_{\boldsymbol{O}}$. The 4th part is modulated by unknown/unmeasured factors represented as $\mathcal{I}_{\bar{g}}$ (“ng” in the legend). The barplot shows the posterior probabilities that the coefficient for each part $\alpha_{jk} > \delta_{+}^{*}$, where α_{jk} denotes the k^{th} RmF modulated by the j^{th} genomic platform. For each RmF, the probabilities of these 4 components, CN, miRNA, others, and ng, are respectively shown in red, green, purple and blue. Each probability in Fig (a) shows that 1 unit increment in the RmF component leads to at least 5% increase in survival time. Each probability in Fig (b) shows that 1 unit increment in the RmF component leads to at least 5% decrease in survival time. We consider the markers to be significant if this posterior probability is larger than 0.5.

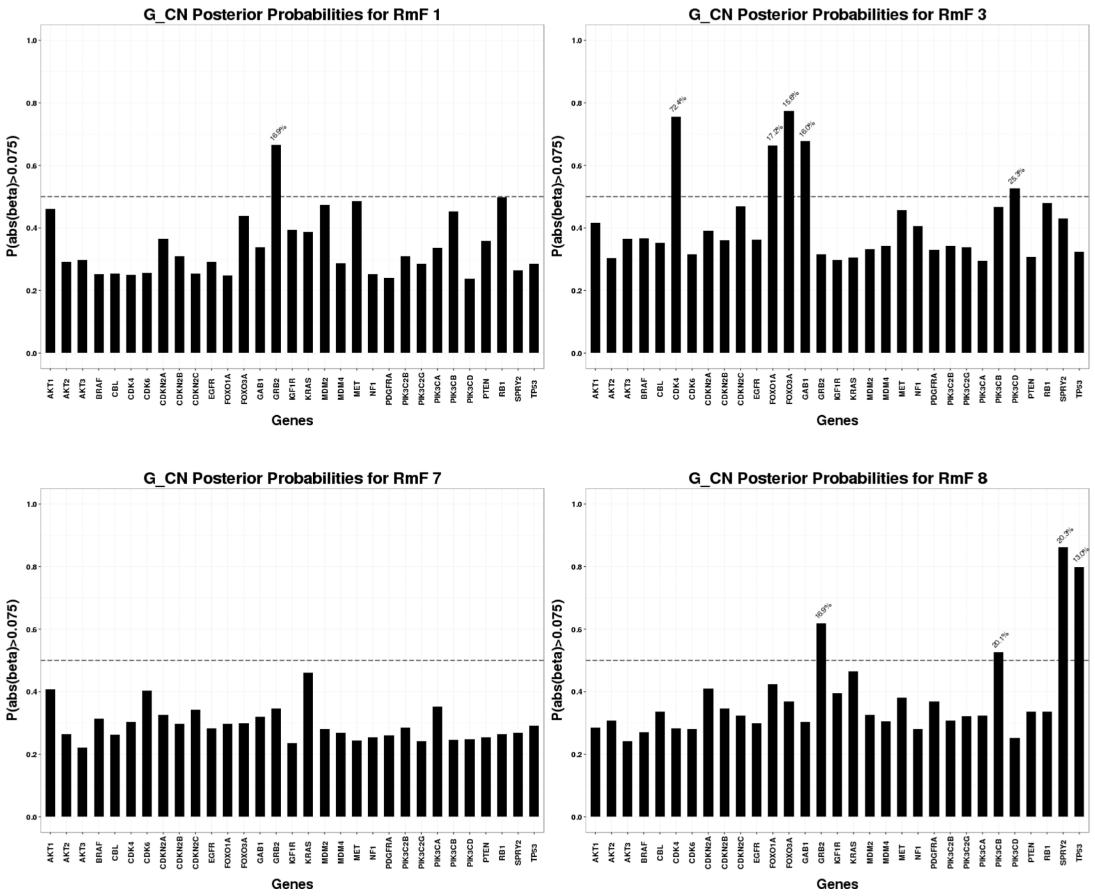


Fig 5:

Significant genomic CN combinations

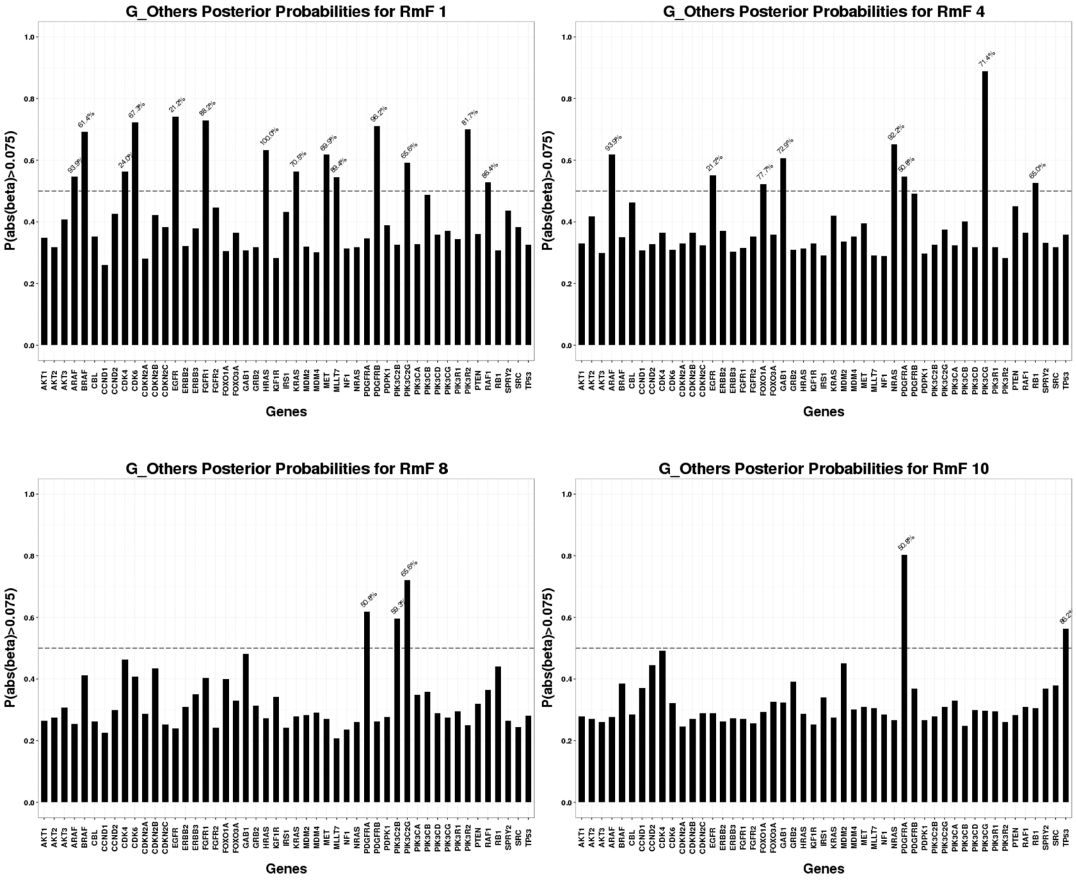


Fig 6:

Significant genomic mRNA “Other” combinations

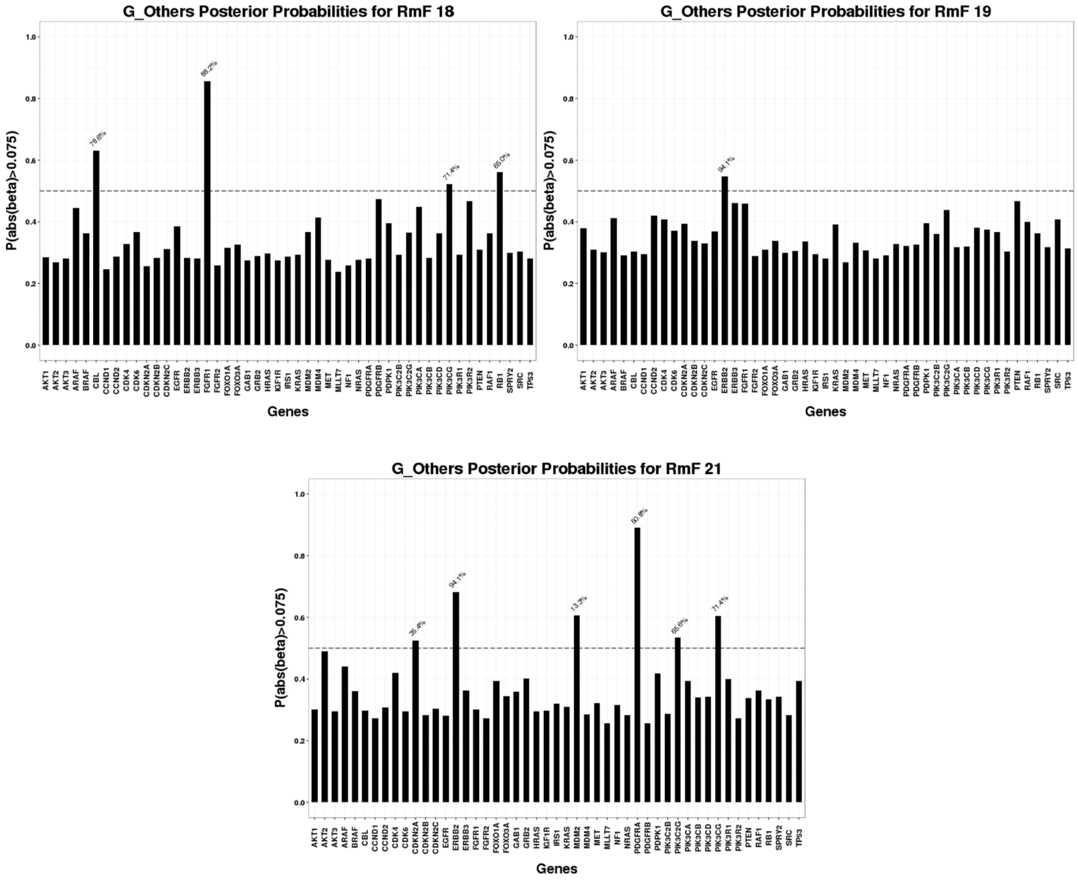
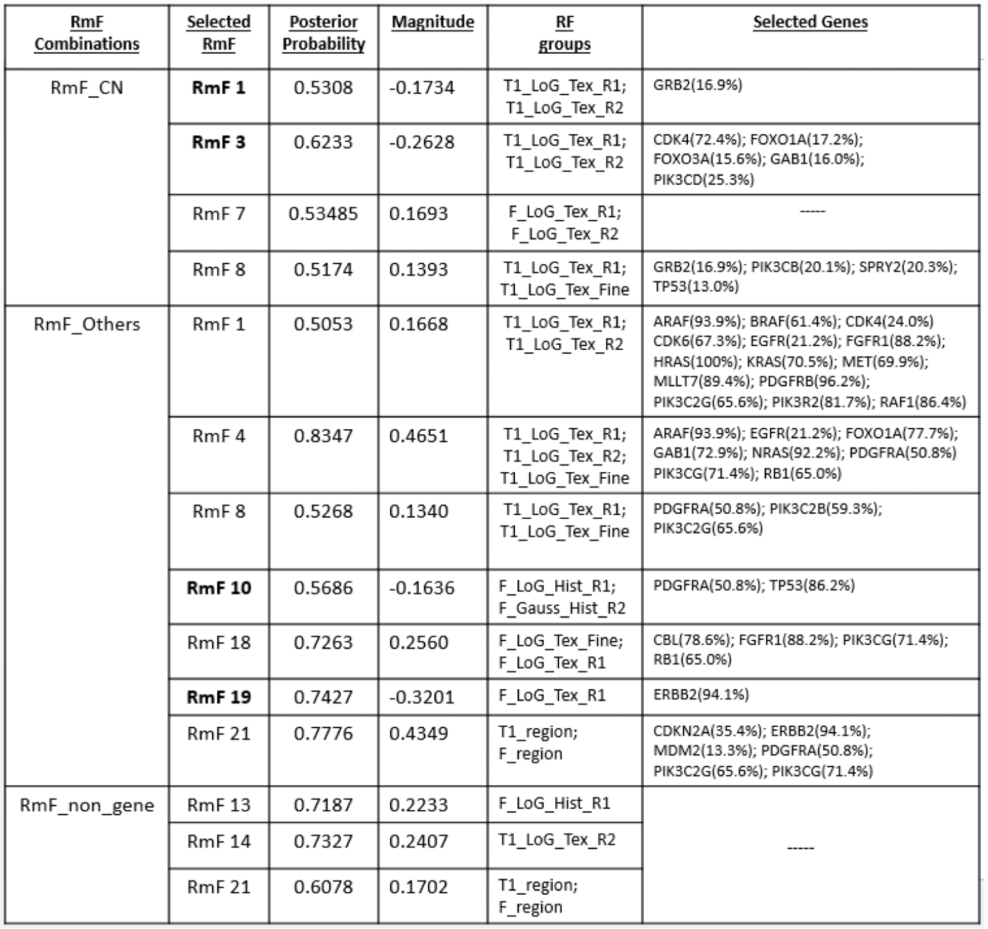


Fig 7:

Significant genomic mRNA “Other” combinations



RmF Combinations	Selected RmF	Posterior Probability	Magnitude	RF groups	Selected Genes
RmF_CN	**RmF 1**	0.5308	-0.1734	T1_LoG_Tex_R1; T1_LoG_Tex_R2	GRB2(16.9%)
	RmF 3	0.6233	-0.2628	T1_LoG_Tex_R1; T1_LoG_Tex_R2	CDK4(72.4%); FOXO1A(17.2%); FOXO3A(15.6%); GAB1(16.0%); PIK3CD(25.3%)
	RmF 7	0.53485	0.1693	F_LoG_Tex_R1; F_LoG_Tex_R2	-----
	RmF 8	0.5174	0.1393	T1_LoG_Tex_R1; T1_LoG_Tex_Fine	GRB2(16.9%); PIK3CB(20.1%); SPRY2(20.3%); TP53(13.0%)
RmF_Others	RmF 1	0.5053	0.1668	T1_LoG_Tex_R1; T1_LoG_Tex_R2	ARAF(93.9%); BRAF(61.4%); CDK4(24.0%) CDK6(67.3%); EGFR(21.2%); FGFR1(88.2%); HRAS(100%); KRAS(70.5%); MET(69.9%); MLLT7(89.4%); PDGFRB(96.2%); PIK3C2G(65.6%); PIK3R2(81.7%); RAF1(86.4%)

	RmF 4	0.8347	0.4651	T1_LoG_Tex_R1;
				T1_LoG_Tex_R2;
				T1_LoG_Tex_Fine
					ARAF(93.9%); EGFR(21.2%); FOXO1A(77.7%); GAB1(72.9%); NRAS(92.2%); PDGFRA(50.8%) PIK3CG(71.4%); RB1(65.0%)

	RmF 8	0.5268	0.1340	T1_LoG_Tex_R1; T1_LoG_Tex_Fine	PDGFRA(50.8%); PIK3C2B(59.3%); PIK3C2G(65.6%)
	RmF 10	0.5686	-0.1636	F_LoG_Hist_R1; F_Gauss_Hist_R2	PDGFRA(50.8%); TP53(86.2%)
	RmF 18	0.7263	0.2560	F_LoG_Tex_Fine; F_LoG_Tex_R1	CBL(78.6%); FGFR1(88.2%); PIK3CG(71.4%); RB1(65.0%)
	RmF 19	0.7427	-0.3201	F_LoG_Tex_R1	ERBB2(94.1%)
	RmF 21	0.7776	0.4349	T1_region; F_region	CDKN2A(35.4%); ERBB2(94.1%); MDM2(13.3%); PDGFRA(50.8%); PIK3C2G(65.6%); PIK3CG(71.4%)
RmF_non_gene	RmF 13	0.7187	0.2233	F_LoG_Hist_R1	-----
	RmF 14	0.7327	0.2407	T1_LoG_Tex_R2	
	RmF 21	0.6078	0.1702	T1_region; F_region	

Fig 8: Results: Significant RmFs, genes and genomic platforms.

Four categories of RmF combinations are listed in the first column, where "non_gene" denotes "$\bar{g}$," which is the non-gene-driven part of the RmF. For each category, several significant RmFs detected from the clinical model are listed in the second column, with unbolded indicating positive ones; bolded indicating negative ones. Posterior probability of the important radiomic markers is shown in column 3. For each selected RmF, several RF groups are selected based on RmF description heatmap (Fig 2). For each significant RmF combination, significant genes are listed with the percentage of how much the variance of mRNA is explained by the specific genomic platform.