Received: 26 December 2020 | Accepted: 2 August 2022

DOI: 10.1111/biom.13740

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Integrative Bayesian models using Post-selective inference: A case study in radiogenomics

Snigdha Panigrahi[1] | **Shariq Mohammed**[2,3] | **Arvind Rao**[2,3,4,5] | **Veerabhadran Baladandayuthapani**[2,3]

[1]Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA

[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

[3]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA

[4]Department of Biomedical Engineering, University of Michigan, Ann Arbor, Michigan, USA

[5]Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Snigdha Panigrahi, Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA.
Email: psnigdha@umich.edu

Funding information
Division of Cancer Prevention, National Cancer Institute, Grant/Award Number: R37CA214955-01A1; NIH, Grant/Award Numbers: R01-CA160736, R21-CA220299; NSF, Grant/Award Numbers: 1463233, 1951980, 2113342; CCSG, Grant/Award Number: P30 CA046592

Abstract

Integrative analyses based on statistically relevant associations between genomics and a wealth of intermediary phenotypes (such as imaging) provide vital insights into their clinical relevance in terms of the disease mechanisms. Estimates for uncertainty in the resulting integrative models are however unreliable unless inference accounts for the selection of these associations with accuracy. In this paper, we develop selection-aware Bayesian methods, which (1) counteract the impact of model selection bias through a "selection-aware posterior" in a flexible class of integrative Bayesian models post a selection of promising variables via ℓ_1-regularized algorithms; (2) strike an inevitable trade-off between the quality of model selection and inferential power when the same data set is used for both selection and uncertainty estimation. Central to our methodological development, a carefully constructed conditional likelihood function deployed with a reparameterization mapping provides tractable updates when gradient-based Markov chain Monte Carlo (MCMC) sampling is used for estimating uncertainties from the selection-aware posterior. Applying our methods to a radiogenomic analysis, we successfully recover several important gene pathways and estimate uncertainties for their associations with patient survival times.

KEYWORDS

Bayesian methods, conditional inference, genomic data, integrative models, postselection inference, radiogenomics, sparse regression

1 | INTRODUCTION

Our methodology in the present paper is motivated by a radiogenomic analysis in low-grade gliomas (LGG), a type of brain cancers. Briefly, a radiogenomic analysis finds associations between imaging outcomes obtained from radiological imaging modalities, for example, magnetic resonance imaging (MRI), with molecular and genomic markers. We pursue post-selective inference, alternately called "selection-aware" inference, in a two-stage integrative modeling framework built on a sequential flow of information: genomics to imaging to clinical outcomes. In the first stage, we glean important genomic variables, which help us identify the variables associated with (multiple) imaging outcomes, called radiogenomic variables (Zhang et al., 2019). The second stage then assesses the clinical relevance of these radiogenomic variables on clinical outcomes. In some sense, this follows the natural

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Biometrics* published by Wiley Periodicals LLC on behalf of International Biometric Society.

progression of cancer, where genomic changes initiate tumor formation and development that are subsequently assessed using imaging, and finally manifest clinical outcomes (e.g., survival) are assessed as the eventual clinical phenotypes.

There exist fundamental gaps between the use of integrative models based on the selected associations between different modalities of information and reliable estimation of uncertainties for the matched parameters. To bridge the void, we recognize the call for "selection-aware inference" in order to systematically counteract the bias incurred in the selection of integrative models (Benjamini & Yekutieli, 2005; Berk et al., 2013; Lee et al., 2016). Intertwined with this goal, we notice an inevitable trade-off of information between the quality of model selection and inferential power for uncertainty estimates from these models. In all realistic scenarios, the extent to which this balance is accomplished has severe implications on the number of discoveries and the power of making these discoveries. Impaired by the high-dimensional nature of genomic variables, known to share complex correlation structures and sparse in terms of their associations with the outcomes, and by the availability of (relatively) small sample sizes, the repercussions of unreliable inference and low power of discoveries can be quite profound in a radiogenomic case study. Our Bayesian methods in the paper demonstrate the potential of reusing samples toward two indispensable goals of inference: (1) an effective integrative modeling of clinical outcomes with interpretable parameters in terms of their mechanisms and (2) a significant reduction in the variance of the matched estimates for the parameters within the integrative models while overcoming the hazardous effects of selection bias at the same time. Before proceeding further, we provide an overview of our methodological development through a schematic snapshot of the integrative pipeline for inference and draw connections with related literature.

2 | SCHEMATIC OVERVIEW AND RELATED LITERATURE

2.1 | Overview

Introducing some basic notations, we denote the outcome variable, the matrix of p explanatory variables and the matrix of L intermediary outcomes, all measured across the same set of n samples, by $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{G} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{\mathcal{I}} \in \mathbb{R}^{n \times L}$, respectively. These measurements represent in the radiogenomic case study the clinical outcome, the genomic variables, and the imaging outcomes, respectively. We let $\boldsymbol{\mathcal{I}}_l$ stand for the imaging outcome l, which is the l-th column of $\boldsymbol{\mathcal{I}}$, and let $\mathbf{G}_F$ represent the submatrix of $\mathbf{G}$ containing the subset of columns indexed by $F \subset \{1, 2, \ldots, p\}$. Through the paper, we use the notation $\rho(\mathbf{d}; \boldsymbol{\beta}_E, \boldsymbol{\Sigma}_E)$ for a normal density function with mean and covariance $\boldsymbol{\beta}_E$ and $\boldsymbol{\Sigma}_E$, respectively, evaluated at $\mathbf{d}$, $\text{diag}(\mathbf{V})$ for a diagonal matrix with the vector $\mathbf{V}$ along the diagonal, $\text{vec}(V_1, \ldots, V_k)$ to denote a vector with the entries V_j, $j \in \{1, \ldots, k\}$, and use $[\mathbf{A}]_j$ to denote the j-th column vector of the matrix $\mathbf{A}$ wherever needed.

Figure 1A depicts how the selection steps inform our integrative models. Divided into two stages, we deploy a multiple regression framework in the first step of this pipeline in order to select genomic variables associated with at least one of the intermediary imaging outcomes. We index the selected set of genomic variables by $\bar{F} \subset \{1, 2, \ldots, p\}$. In the second step, we select from $\bar{F}$ the variables that are further associated with the clinical outcome. Calling this set E with cardinality $|E|$, the output at this stage results in an integrative radiogenomic clinical model [**RgCM**], using $\mathcal{E}(E) = \bar{E} \subset \{1, 2, \ldots, p\}$, where $\mathcal{E}$ is a deterministic mapping that takes E as the input and returns $\bar{E}$.

Figure 1B outlines the two core inferential results we establish in the paper to validate the use of integrative models through a "selection-aware posterior." In a nutshell, this selection-aware posterior uses a conditional likelihood, jointly with a prior post selection. The conditional likelihood is obtained by conditioning on the event that we select genomic variables in the set E. Theorem 1 identifies a simplified expression for the conditional likelihood function, in the sense that the truncation region associated with the conditioning event can be very simply expressed in terms of some sign constraints. Theorem 2 enables us tractable optimization-based updates [**OP**] to sample from a working version of the selection-aware posterior [**SaP**] through a reparameterization mapping [**RP**]. Incurring no additional cost, we construct from the reparameterization mapping samples for the parameters within the integrative radiogenomic clinical model [**RgCM**].

2.2 | Related work

Our approach of using a "selection-aware posterior" is anchored within a conditional proposal for Bayesian models post selection in Yekutieli (2012) and Panigrahi et al. (2021). Such an approach advocates the use of a conditional likelihood to discard the information from data consumed in model selection. Adopting the Bayesian perspective as opposed to a frequentist solution to the post-selective problem (Lee et al., 2016; Panigrahi et al., 2019; Tian & Taylor, 2018, among others) admits several flexibilities for estimating uncertainties. The latter line of work outlines inference for real-valued parameters in models where the outcome mean is simply modeled as $\mu \in \mathbb{R}^n$, without specifying a relation with the explanatory variables. Our Bayesian prescription on the other hand permits a joint estimation of

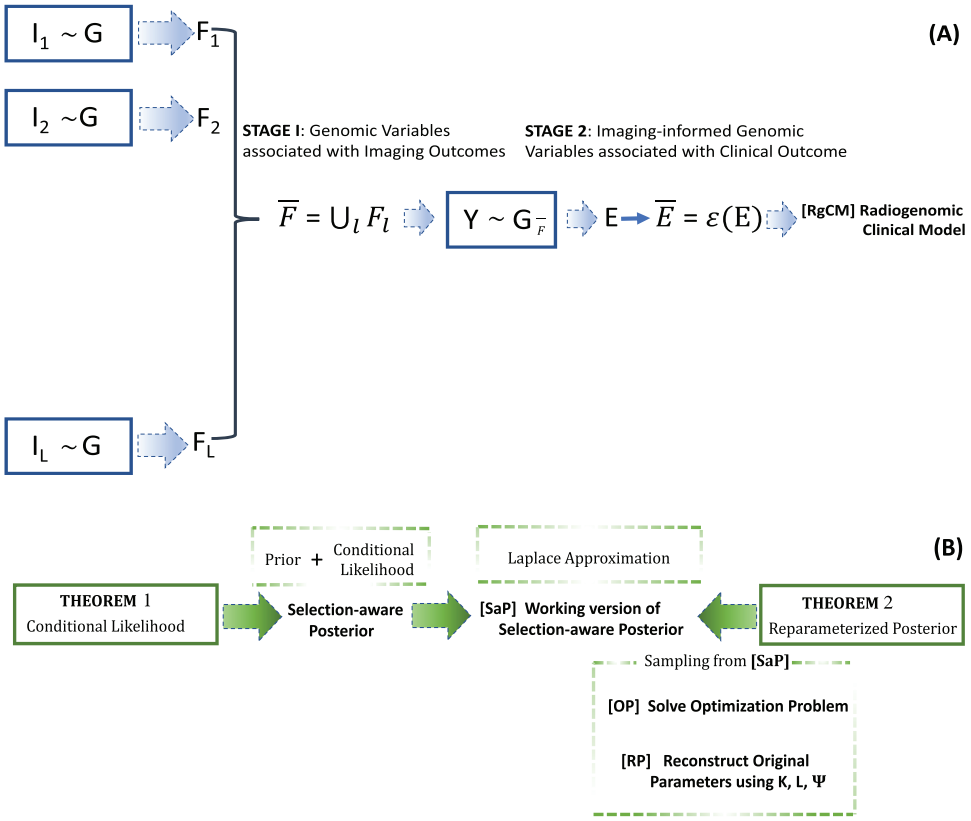


FIGURE 1 Panel (A). Schematic representation of the inputs and outputs of the selection pipeline divided into two stages: The first stage gives a candidate set of genomic variables associated with the intermediary imaging phenotypes; the second stage selects promising variables from an imaging-informed set of genomic variables (from the first stage) that are statistically associated with the clinical outcome. The final output of the algorithm, E, imaging-informed genomic variables associated with the clinical outcome, (potentially) interacts with preexisting knowledge for the genomic variables to determine our radiogenomic clinical models. Panel (B): Schematic representation of the methodological development: Theorem 1 identifies a simplified expression for the conditional likelihood function that conditions out the observed event wherein the pipeline selects the set of variables indexed by E. In conjunction with the prior post selection, the conditional likelihood leads us to a selection-aware posterior. Theorem 2 provides us tractable optimization-based updates [**OU**] to sample from a working version of the selection-aware posterior [**SaP**] through a reparameterization mapping [**RP**]. Incurring no additional cost, we construct from the reparameterization mapping samples for our original target, the parameters within the integrative radiogenomic clinical model [**RgCM**].

vector-valued parameters and functions thereof in a flexible class of models, including models based on an interplay between the selection output and prior knowledge.

As a prelude to the technical development in the paper, let $\boldsymbol{\beta}_{\bar{E}}$ be the parameters in the radiogenomic clinical [**RgCM**]. Our starting point for inference is the selection-aware posterior for $\boldsymbol{\beta}_{\bar{E}}$, proportional to:

$$\{\mathbb{P}[\,\mathcal{E}_0 \mid \boldsymbol{\beta}_{\bar{E}}]\}^{-1} \cdot \pi(\mathbf{D}; \boldsymbol{\beta}_{\bar{E}}),$$

where $\mathbb{P}[\,\mathcal{E}_0 \mid \boldsymbol{\beta}_{\bar{E}}]$ is the probability for the event of selection $\mathcal{E}_0$, the set of all realizations of data leading to the selected set E, and $\pi(\mathbf{D}; \boldsymbol{\beta}_{\bar{E}})$ is the usual (ignoring the effects for selection) posterior based on observed data $\mathbf{D}$. Previous proposals in Panigrahi et al. (2021) and Panigrahi and Taylor (2018, 2022) establish a statistically consistent approximation for the probability of selection and enable a working version for the otherwise intractable selection-aware posterior. Sampling from the working posterior is, however, an arduous task in the high-dimensional regime, due to impediments from both geometric and analytic angles. Geometrically, the conditional likelihood is not easily amenable for inference, since it is truncated to a rather complicated event of selection. From the analytic perspective, several variables must be integrated out to calculate the probability of selection, which sets apart our selection-aware posterior from the usual posterior without adjustments for selection.

Simplifying the complex geometry of the truncation region and bypassing intensive integrations, our new methods in the paper quite remarkably facilitate tractable updates for gradient-based sampling from the selection-aware posterior. We solve only an $|E|$-dimensional convex optimization for each update. A reparameterization mapping applied to a carefully constructed conditional likelihood function reduces the effective dimension of our

inferential updates. The trade-off that our selection-aware Bayesian methods strike between selection and inference stands in stark contrast with sample splitting (Hurvich & Tsai, 1990, for example), a simple tool of choice for practitioners to counter the effect of selection bias. With sample sizes as small as 60, ignoring a fraction of the samples for either of the two goals is highly suboptimal for integrative inference. Our numerical experiments illustrate this trade-off between selection and inference, and highlight the advantages our methods enjoy over splitting at different resolutions in terms of the support recovery of models and the power of their uncertainty estimates.

We structure the remaining paper as follows. Section 3 outlines our integrative modeling framework. Section 4 discusses a selection-aware posterior and develops a tractable Bayesian framework amenable for gradient-based sampling from this posterior. Section 5 explores the potential of our proposal by simulating integrative models using the actual genomic and imaging measurements from a radiogenomic case study. Section 6 constructs uncertainty estimates for the effects of biologically relevant gene pathways in our integrative models using imaging and clinical outcomes for LGG.

3 | MODELING FRAMEWORK

Adopting a two-stage selection for integrative modeling, we solve L LASSO regression problems (Tibshirani, 1996) to select the promising explanatory variables associated with $\boldsymbol{\mathcal{I}}_l$, an intermediary (imaging) outcome:

$$(\hat{\boldsymbol{\alpha}}_{F_l}, \mathbf{0}) = \arg\min_{\alpha} \frac{1}{2}\|\boldsymbol{\mathcal{I}}_l - \mathbf{G}\boldsymbol{\alpha}\|_2^2 + \lambda_l\|\boldsymbol{\alpha}\|_1 \text{ for } l \in \{1, 2, \dots, L\}; \quad (1)$$

λ_l is the tuning parameter for the ℓ_1 penalty. Note that $\bar{F} = \cup_{l=1}^{L} F_l$ with cardinality $|\bar{F}|$ gives us a candidate set of explanatory genomic variables such that each variable is associated with at least one of the intermediary phenotypic outcomes.

We follow (1) by solving a randomized version of the LASSO regression (Panigrahi et al., 2019; Panigrahi & Taylor, 2018; Tian & Taylor, 2018; Tian et al., 2016) which introduces a trade-off between the quality of model selection and inferential power. This strategy perturbs the canonical algorithm with a randomization variable $\mathbf{r} \sim N(0, \eta^2 \cdot \mathbf{I})$ independent of $\mathbf{y}$ and $\boldsymbol{\mathcal{I}}_l$, $l \in \{1, 2 \dots, L\}$ to solve:

$$(\hat{\boldsymbol{\beta}}_E^{\text{LASSO}}, \mathbf{0}) = \arg\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{G}_{\bar{F}}\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\Lambda}\boldsymbol{\beta}\|_1 + \frac{\epsilon}{2} \cdot \|\boldsymbol{\beta}\|^2 - \mathbf{r}^T\boldsymbol{\beta}. \quad (2)$$

In (2), $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_{\bar{F}})$, the diagonal entries are ℓ_1-penalty weights for the $|\bar{F}|$ variables that are selected as significant associations from the previous step. Setting these weights to be inversely proportional to the number of times an explanatory variable is selected across the L regularized queries in the previous step is one such concrete way to incorporate the relative importance of each variable in (2). We identify E as the set of nonzero LASSO estimates, giving us the set of explanatory variables that are statistically associated with both the intermediary and clinical outcomes. Noticeably, the optimization objective (2) differs from a canonical version of LASSO through an additional term, which is linear in the randomization instance $\mathbf{r}$ and an ℓ_2 penalty with a small positive coefficient $\epsilon > 0$. We use a small value for the ℓ_2 penalty in the objective to ensure the existence of a solution for the randomized problem.

Turning to an integrative model post selection, we define:

$$\mathcal{E} : E \to \bar{E}, \quad (3)$$

a mapping applied to E that returns $\bar{E}$, a subset of the p explanatory variables, with cardinality $|\bar{E}| = q$. Specifically, (3) allows us the flexibility to incorporate interactions between preexisting knowledge about the explanatory variables with the output of the two-stage pipeline in (1) and (2). For instance, adding variables to the selected set E that might have been missed in the ℓ_1 regularized selection steps, based on pathway annotations or previously validated clinical analyses, are examples of some practical choices for this mapping. Then, we assume a linear dependence between $\mathbf{y}$ and $\mathbf{G}_{\bar{E}}$ under a fixed predictor matrix framework. Letting $\mathbf{G}_{i,\bar{E}}^T$ denote the i-th row of the matrix $\mathbf{G}_{\bar{E}}$, each sample i is identically and independently distributed as follows:

$$[\mathbf{RgCM}]: \quad \mathrm{y}_i = \mathbf{G}_{i,\bar{E}}^T\boldsymbol{\beta}_{\bar{E}} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2). \quad (4)$$

This is our primary outcome model of interest.

For modeling the associations between the intermediary outcomes and explanatory variables, we assume for now

$$\boldsymbol{\mathcal{I}}_{i,l} = \mathbf{G}_{i;F_l}^T\boldsymbol{\alpha}_{F_l} + \Psi_{i,l}, \; l \in \{1, 2, \dots, L\}, \; \boldsymbol{\Psi}_i = (\Psi_{i,1}, \dots, \Psi_{i,L}) \overset{i.i.d.}{\sim} N(0, \boldsymbol{\Sigma}_I); \quad (5)$$

$\boldsymbol{\mathcal{I}}_{i,l}$ is a sample for the intermediary outcome $\boldsymbol{\mathcal{I}}_l$ indexed by i, $\boldsymbol{\Psi}_i$ is independent of ϵ_i in the primary model (4). The validity of our inferential approach is not tied to the selected linear model between the explanatory and intermediary variables in (5). Our methods rely simply on the independence between the model errors ϵ_i and $\boldsymbol{\Psi}_i$ for each data sample; the proof for Proposition 1 in the next section justifies this observation. We specify a linear

model for the intermediary outcomes only for the sake of simple exposition in the remaining paper.

Notice, the intermediary outcome models inform our primary outcome model via F_l for $l \in \{1, 2, \dots, L\}$ that in turn determine the candidate set of explanatory variables in (2) for a downstream modeling of the clinical outcome through the selected set E.

Completing the model specification in a Bayesian framework, we impose a rich family of scale-mixture Gaussian priors $\pi(\cdot)$ on our selection-aware parameters $\boldsymbol{\beta}_{\bar{E}}$ (Park & Casella, 2008) :

$$\boldsymbol{\beta}_{\bar{E}} \mid \eta_1^2, \eta_2^2, \dots, \eta_q^2 \sim N(0, \sigma^2 \text{diag}(\eta_1^2, \eta_2^2, \dots, \eta_q^2)); \quad (6)$$

and $\eta_j^2 \sim \text{Exp}(2^{-1}\lambda^2)$ for $j \in \bar{E}$.

4 | SELECTION-AWARE POSTERIOR INFERENCE

4.1 | Selection-aware posterior

Recall, the model in (4) is dependent on the sets of selected variables F_l for $l \in \{1, 2, \dots, L\}$ and E. We call the respective random variables $\widehat{E}(\mathbf{Y}, \mathbf{R})$ and $\widehat{F}_l(\mathbf{I}_l)$, highlighting their dependence on the outcome variable $\mathbf{Y}$, the intermediary variable $\mathbf{I}_l$, and the randomization variable $\mathbf{R}$. That is, we observe the following realizations from the model selection pipeline in Section 3:

$$\widehat{E}(\mathbf{Y}, \mathbf{R}) = E \text{ and } \widehat{F}_l(\mathbf{I}_l) = F_l \text{ for } l = 1, 2, \dots, L. \quad (7)$$

Explicitly accounting for the selection-aware nature of our modeling framework, a likelihood conditioned upon observing (7) discards the information from our samples utilized for model selection. We provide in Proposition 1 the form of the conditional likelihood in terms of the parameters within the primary outcome model (4).

Proposition 1. *Let $\bar{E}$ be defined according to (3). Let $\widehat{\boldsymbol{\beta}}_{\bar{E}} = (\mathbf{G}_{\bar{E}}^T \mathbf{G}_{\bar{E}})^{-1} \mathbf{G}_{\bar{E}} \mathbf{y}$ be the least squares estimate after regressing $\mathbf{y}$ against $\mathbf{G}_{\bar{E}}$, with covariance matrix $\boldsymbol{\Sigma}_{\bar{E}}$. Then, under the modeling assumptions (4) and (5), the likelihood obtained by conditioning the law of $\widehat{\boldsymbol{\beta}}_{\bar{E}}$ upon the observed event of selection in (7) agrees with:*

$$\left\{\mathbb{P}[\widehat{E}(\mathbf{Y}, \mathbf{R}) = E \mid \boldsymbol{\beta}_{\bar{E}}]\right\}^{-1} \cdot \rho(\widehat{\boldsymbol{\beta}}_{\bar{E}}; \boldsymbol{\beta}_{\bar{E}}, \boldsymbol{\Sigma}_{\bar{E}}) \quad (8)$$

up to a proportionality constant in $\boldsymbol{\beta}_{\bar{E}}$.

Underscored in Section 2, the selection-aware posterior that appends the conditional likelihood in Proposition 1 with a prior for $\boldsymbol{\beta}_{\bar{E}}$ is obtained by applying a multiplicative correction term to the usual posterior. Basing inference on the selection-aware posterior is a formidable challenge, because the value of the posterior, due to conditioning, involves the probability of selection

$$\mathbb{P}[\widehat{E}(\mathbf{Y}, \mathbf{R}) = E \mid \boldsymbol{\beta}_{\bar{E}}^{(d)}], \quad (9)$$

which must be computed in each new draw $\boldsymbol{\beta}_{\bar{E}}^{(d)}$ sampled from the posterior. The event of selection in (9) as we characterize next is associated with an intricate geometry and the probability for the event involves integrating out several variables to render an exact value. Circumventing geometric and analytic impediments to selection-aware inference, our solution in the following development casts the core step as an easy-to-solve, low-dimensional, convex optimization problem [**OP**].

4.2 | A simplified conditional likelihood

Before stating our solution, Proposition 2 characterizes the event of selection $\{(\mathbf{y}, \mathbf{r}) : \widehat{E}(\mathbf{y}, \mathbf{r}) = E\}$ as a union of polyhedral regions determined by

$$\left\{\mathbf{U}_{\mathbf{s}_E}, \mathbf{V}_{\mathbf{s}_E}, \mathbf{W}_{\mathbf{s}_E}, \mathbf{t}_{\mathbf{s}_E} : \mathbf{s}_E = \{-1, 1\}^{|E|}\right\}. \quad (10)$$

Detailed expressions for (10) are included in the Supporting Information.

Proposition 2. *After solving (2), the selection event $\{(\mathbf{y}, \mathbf{r}) : \widehat{E}(\mathbf{y}, \mathbf{r}) = E\}$ is equivalent to*

$$\bigcup_{\mathbf{s}_E = \{-1,1\}^{|E|}} \left\{\mathbf{U}_{\mathbf{s}_E} \widehat{\boldsymbol{\beta}}_{\bar{E}} + \mathbf{V}_{\mathbf{s}_E} \begin{pmatrix} \mathbf{r}_E^T & \mathbf{r}_{E^c}^T \end{pmatrix}^T + \mathbf{W}_{\mathbf{s}_E} \begin{pmatrix} (\widehat{\boldsymbol{\beta}}_E^{\perp})^T & (\widehat{\boldsymbol{\beta}}_{E^c}^{\perp})^T \end{pmatrix}^T > \mathbf{t}_{\mathbf{s}_E}\right\},$$

where $\widehat{\boldsymbol{\beta}}^{\perp} = \mathbf{G}_{\bar{F}}^T \mathbf{y} - \mathbf{G}_{\bar{F}}^T \mathbf{G}_{\bar{E}} \widehat{\boldsymbol{\beta}}_{\bar{E}}$ and $E^c = \bar{F} \setminus E$.

By recognizing next a careful conditioning event, we reduce the seemingly complicated probability of selection, equivalent to the probability of a union of polyhedral regions, to that of an orthant based on very simple sign restrictions on our data variables. This results in a considerably simpler conditional likelihood function, which we formalize in Theorem 1. Establishing some more notations, the stationary equation at the solution of (2) is given by:

$$\begin{pmatrix} \mathbf{r}_E^T & \mathbf{r}_{E^c}^T \end{pmatrix}^T + \begin{pmatrix} (\mathbf{G}_E^T \mathbf{y})^T & (\mathbf{G}_{E^c}^T \mathbf{y})^T \end{pmatrix}^T = \begin{bmatrix} \mathbf{G}_E^T \mathbf{G}_E + \epsilon \cdot \mathbf{I} & \mathbf{G}_E^T \mathbf{G}_{E^c} \end{bmatrix}^T \widehat{\boldsymbol{\beta}}_E^{\text{LASSO}} + \begin{pmatrix} (\boldsymbol{\Lambda}_E \mathbf{s}_E)^T & (\boldsymbol{\Lambda}_{E^c} \mathbf{z})^T \end{pmatrix}^T, \quad (11)$$

where the active solution, $\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}$, and the inactive part of the subgradient, $\mathbf{z}$, satisfy the constraints

$$\text{sign}(\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}) = \mathbf{s}_E \; ; \; \|\mathbf{z}\|_\infty < 1; \tag{12}$$

$\mathbf{s}_E$ is the vector of signs for the active (nonzero) LASSO solution. We use $\widehat{\mathbf{B}}^{\text{LASSO}}$, $\widehat{\mathbf{Z}}$, and $\widehat{\mathbf{B}}^{\perp}$ to represent the random variables that assume the realizations $\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}$, $\mathbf{z}$, and $\widehat{\boldsymbol{\beta}}^{\perp}$, respectively, upon solving (2). We defer the explicit forms for the matrices $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{o}$, $\mathbf{K}$, $\mathbf{l}$, and $\boldsymbol{\Theta}_{\bar{E}}$ in the next result to the Supporting Information.

Theorem 1. *Consider the modeling assumptions in (4) and (5). Define*

$$I(\boldsymbol{\beta}_{\bar{E}}) = \int_{sign(w)=\mathbf{s}_E} \rho(b; \mathbf{K}\boldsymbol{\beta}_{\bar{E}} + \mathbf{l}, \boldsymbol{\Theta}_{\bar{E}}) \cdot \rho(w; \mathbf{P}b + \mathbf{o}, \eta^{-2}\mathbf{Q}^T\mathbf{Q}) dw db. \tag{13}$$

Then, conditional upon

$$\left\{ \widehat{E} = E,\ sign(\widehat{\mathbf{B}}^{\text{LASSO}}) = \mathbf{s}_E,\ \widehat{\mathbf{Z}} = \mathbf{z},\ \widehat{\mathbf{B}}^{\perp} = \widehat{\boldsymbol{\beta}}^{\perp} \right\},$$

the likelihood of the least squares estimate $\widehat{\boldsymbol{\beta}}_{\bar{E}}$ in Proposition 1 is proportional to

$$(I(\boldsymbol{\beta}_{\bar{E}}))^{-1} \cdot \rho(\widehat{\boldsymbol{\beta}}_{\bar{E}}; \mathbf{K}\boldsymbol{\beta}_{\bar{E}} + \mathbf{l}, \boldsymbol{\Theta}_{\bar{E}}). \tag{14}$$

To further avoid carrying out the integration to calculate $I(\boldsymbol{\beta}_{\bar{E}})$ that lacks a value in closed form, we apply the Laplace technique (Kass & Raftery, 1995) for approximating:

$$\log I(\boldsymbol{\beta}_{\bar{E}}) \approx -\inf_{b,w} \left\{ \frac{1}{2}(b - \mathbf{K}\boldsymbol{\beta}_{\bar{E}} - \mathbf{l})^T \boldsymbol{\Theta}_{\bar{E}}^{-1} (b - \mathbf{K}\boldsymbol{\beta}_{\bar{E}} - \mathbf{l}) + \frac{1}{2\eta^2}(w - \mathbf{P}b - \mathbf{o})^T \mathbf{Q}^T\mathbf{Q}(w - \mathbf{P}b - \mathbf{o}) + \text{Barr}_{\mathbf{s}_E}(w) \right\} + C; \tag{15}$$

C is a constant, and w_j and $s_{j;E}$ are the values of the j-th coordinate of the respective vectors and $\text{Barr}_{\mathbf{s}_E}(w) = \sum_{j=1}^{|E|} \text{Barr}_{j;\mathbf{s}_{j,E}}(w_j)$ is a barrier penalty that encodes the sign constraints on each coordinate of w through a smooth function $\text{Barr}_{j;\mathbf{s}_{j;E}}(w_j) = \log(1 + \delta(\mathbf{s}_{j;E} w_j)^{-1})$. In conjunction with our prior, the approximate value of $I(\boldsymbol{\beta}_{\bar{E}})$ plugged into the likelihood in (1) leads us to a working version for the (log-) posterior:

$$[\mathbf{SaP}] : \log \rho(\widehat{\boldsymbol{\beta}}_{\bar{E}}; \mathbf{K}\boldsymbol{\beta}_{\bar{E}} + \mathbf{l}, \boldsymbol{\Theta}_{\bar{E}}) + \inf_{b,w} \left\{ \frac{1}{2}(b - \mathbf{K}\boldsymbol{\beta}_{\bar{E}} - \mathbf{l})^T \boldsymbol{\Theta}_{\bar{E}}^{-1} (b - \mathbf{K}\boldsymbol{\beta}_{\bar{E}} - \mathbf{l}) + \frac{1}{2\eta^2}(w - \mathbf{P}b - \mathbf{o})^T \mathbf{Q}^T\mathbf{Q}(w - \mathbf{P}b - \mathbf{o}) + \text{Barr}_{\mathbf{s}_E}(w) \right\} + \log \pi(\boldsymbol{\beta}_{\bar{E}}) \tag{16}$$

after ignoring constants.

4.3 | Reparameterization mapping

We develop a reparameterization mapping to enable tractable updates from (16) when gradient-based Markov chain Monte Carlo (MCMC) sampling is deployed for inference. With no additional cost, we can easily reconstruct using the same mapping our original targets, the parameters in the radiogenomic clinical model.

Consider the optimization:

$$[\mathbf{OP}] : \mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}) = \underset{w}{\text{argmin}} \ \frac{1}{2\eta^2}(w - \mathbf{P}\boldsymbol{\zeta}_{\bar{E}} - \mathbf{o})^T \mathbf{Q}^T\mathbf{Q}(w - \mathbf{P}\boldsymbol{\zeta}_{\bar{E}} - \mathbf{o}) + \text{Barr}_{\mathbf{s}_E}(w), \tag{17}$$

denoting the optimal value by $V^*(\boldsymbol{\zeta}_{\bar{E}})$. Based on the solution of (17), fix

$$\Psi(\boldsymbol{\zeta}_{\bar{E}}) = \left(\mathbf{I} + \eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\mathbf{P}\right)\boldsymbol{\zeta}_{\bar{E}} + \eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}(\mathbf{o} - \mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}})). \tag{18}$$

Then, we define the reparameterization $\boldsymbol{\beta}_{\bar{E}} \to \boldsymbol{\zeta}_{\bar{E}}$ through $\mathbf{K}$ and $\mathbf{l}$ and the mapping $\Psi(\cdot)$ as follows:

$$[\mathbf{RP}] : \mathbf{K}\boldsymbol{\beta}_{\bar{E}} + \mathbf{l} = \Psi(\boldsymbol{\zeta}_{\bar{E}}). \tag{19}$$

Applying (19) to the working version for the (log-) posterior in (16), the next theorem provides the value of a transformed analog for the working posterior $\widetilde{\pi}(\boldsymbol{\zeta}_{\bar{E}}|\widehat{\boldsymbol{\beta}}_{\bar{E}})$ and the corresponding gradient in terms of the variables $\boldsymbol{\zeta}_{\bar{E}}$.

Theorem 2. *Consider the reparameterization mapping in (19). Fix*

$$\mathcal{J}(\boldsymbol{\zeta}_{\bar{E}}) = \mathbf{K}^{-1}\left(\mathbf{I} + \eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\mathbf{P}\right) - \eta^{-4}\mathbf{K}^{-1}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\mathbf{N}^{-1}\mathbf{Q}^T\mathbf{Q}\mathbf{P}, \tag{20}$$

$$\boldsymbol{\mathcal{M}} = vec(\boldsymbol{\mathcal{M}}_1, \ldots, \boldsymbol{\mathcal{M}}_q),$$

where $\mathbf{N} = \eta^{-2}\mathbf{Q}^T\mathbf{Q} + \nabla^2 \text{Barr}_{\mathbf{s}_E}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}))$ and $\boldsymbol{\mathcal{M}}_j$ takes the value

$$Trace\left(\frac{1}{\eta^4}\mathcal{J}^{-1}(\boldsymbol{\zeta}_{\bar{E}})\mathbf{K}^{-1}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\mathbf{N}^{-1}\left(diag(\nabla^3 Barr_{1;\mathbf{s}_{E;1}}(w_1^*(\boldsymbol{\zeta}_{\bar{E}})), \ldots, \nabla^3 Barr_{|E|;\mathbf{s}_{E;|E|}}(w_E^*(\boldsymbol{\zeta}_{\bar{E}}))\, diag(\eta^{-2}[\mathbf{N}^{-1}\mathbf{Q}^T\mathbf{Q}\mathbf{P}]_j)\right)\mathbf{N}^{-1}\mathbf{Q}^T\mathbf{Q}\mathbf{P}\right) \tag{21}$$

for $j \in \{1, \dots, \bar{E}\}$. *Then, we have the following:*

(1) The value of $\log \widetilde{\pi}(\boldsymbol{\zeta}_{\bar{E}} | \widehat{\boldsymbol{\beta}}_{\bar{E}})$, *up to an additive constant, is equal to*

$$\log |\, det(\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}}))| + (\widehat{\boldsymbol{\beta}}_{\bar{E}} - \boldsymbol{\zeta}_{\bar{E}})^T \boldsymbol{\Theta}_{\bar{E}}^{-1} \Psi(\boldsymbol{\zeta}_{\bar{E}}) + \frac{1}{2} \boldsymbol{\zeta}_{\bar{E}}^T \boldsymbol{\Theta}_{\bar{E}}^{-1} \boldsymbol{\zeta}_{\bar{E}} + V^*(\boldsymbol{\zeta}_{\bar{E}}) + \log \pi(\mathbf{K}^{-1} \Psi(\boldsymbol{\zeta}_{\bar{E}}) - \mathbf{K}^{-1} \mathbf{l}). \tag{22}$$

(2) The gradient for $\log \widetilde{\pi}(\boldsymbol{\zeta}_{\bar{E}} | \widehat{\boldsymbol{\beta}}_{\bar{E}})$ *is equal to*

$$(\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}}))^T \left(\nabla \log \pi(\mathbf{K}^{-1} \Psi(\boldsymbol{\zeta}_{\bar{E}}) - \mathbf{K}^{-1} \mathbf{l}) + \mathbf{K}^T \boldsymbol{\Theta}_{\bar{E}}^{-1} (\widehat{\boldsymbol{\beta}}_{\bar{E}} - \boldsymbol{\zeta}_{\bar{E}}) \right) + \mathcal{M}(\boldsymbol{\zeta}_{\bar{E}}). \tag{23}$$

With a sample $\boldsymbol{\zeta}_{\bar{E}}^{(d)}$ from $\widetilde{\pi}(\cdot | \widehat{\boldsymbol{\beta}}_{\bar{E}})$, we obtain the corresponding draw for our target parameter $\boldsymbol{\beta}_{\bar{E}}^{(d)}$ using the relation (19). Notice, a reconstruction of our original target parameters does not involve an additional cost, because the primary cost is incurred in solving for $\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}^{(d)})$, which we compute to generate samples from the transformed posterior. Viewing this from the perspective of inferential efficiency, the optimization at every update of the transformed posterior is only $|E|$ dimensional, orders smaller in magnitude than p, the size of the initial set of explanatory variables.

5 | SIMULATION ANALYSIS

5.1 | Simulation design

Generating a sparse model with signals of varying amplitudes and random signs, we draw in each round of simulation an outcome from the primary model (4). The signal vector $\boldsymbol{\beta}$ is generated from a mixture of centered Laplace distributions, the true underlying prior. That is, each coordinate for $\boldsymbol{\beta}$ is drawn as follows:

$$\beta_j \sim \Pi_j(\cdot) := \pi \cdot \text{Laplace}(0, 0.10) + (1 - \pi) \cdot \text{Laplace}(0, s) \text{ for } j \in \{1, 2, \dots, r\}; \tag{24}$$

$r = |\bar{F}|$. Changing the scale of one of these Laplace distributions, s and the mixing proportion, π results in different signal regimes. In particular, the mixing proportion controls the sparsity levels of our signal vector.

We vary in our design the ratio between the number of our samples, n, and the number of regressors, r, before selecting the set E that in turn determines the model (4). For the real data analysis, we note $r/n \approx 5$, $n = 60$, $r = 357$. In this case, the matrix of explanatory variables we use is based on the real values of genomic measurements that are significantly associated with the imaging (radiomic) outcomes. We provide a comprehensive background for these data in the next section. Beyond the real setting, we investigate the following sample sizes: $n = 180,\ 360,\ 720$, to match the dimension ratio $r/n = 2, 1, 0.5$, respectively. To generate predictor measurements for sample sizes larger than 60, we append synthetic design values to the real design matrix, in order to achieve the regression dimensions set as per the ratio r/n. Specifically, we draw $x_i \in \mathbb{R}^r, 1 \le i \le n_1$ such that $x_i \sim N(0, \boldsymbol{\Sigma}(\rho))$, for $i = 1, 2, \dots, n_1$ and n_1 is chosen so that $r/(n_1 + n) = 2, 1, 0.5$ in the three case studies of interest; $\boldsymbol{\Sigma}(\rho)$ is an autocorrelation covariance matrix such that the (i, j)-th entry of $\boldsymbol{\Sigma}$ is equal to $\rho^{|i-j|}$ and $\rho = 0.70$.

Our strategy to reflect a realistic data generation process through simulations is aligned along the following principles. First, we note that the variability in the outcome variable is explained by multiple markers, consistent with our expectation of a polygenic response variable. Second, the generative model we use incorporates a mix of weak and strong signals. This enables us to investigate the genuine ability of our inferential methods to adapt to the strength of the signals present in the data and to reconstruct efficiently the corresponding effect sizes. Third, by conducting simulations over a range of regression dimensions, our experiments showcase the necessity of adopting a selection-aware method even in moderate dimensions wherein a severe impact of selection bias is noted.

5.2 | Empirical analysis: Consistent with radiogenomic case study

We apply our methods to the real radiogenomic measurements by simulating $n = 60$ samples that closely agree with the real data. We set the randomization variation η^2 in (2) to be roughly equal to the noise level in the outcome by plugging in an estimate of this value from the data. Setting $\pi = 0.95$ in the generative scheme (24) and varying the scale of the Laplace distribution s to take one of the values in the set $\{0.20, 1, 2, 4\}$, we consider 4 signal regimes—numbered 1–-4 on the x-axis of Figure 2. Using the expressions for the posterior and the gradient in Theorem 2, we update our estimates for $\boldsymbol{\beta}_{\bar{E}}$ from the working version of the selection-aware posterior (16) and construct intervals for these parameters by setting $\bar{E} = E$, the final output of the regularized variable selection algorithms.

As noted in the introduction, a balance in the quality of the model in terms of support recovery and inferential power for the matched parameters is imminent when a finite amount of data must be allocated for deciding a model and inferring for the parameters in it. Illustrating

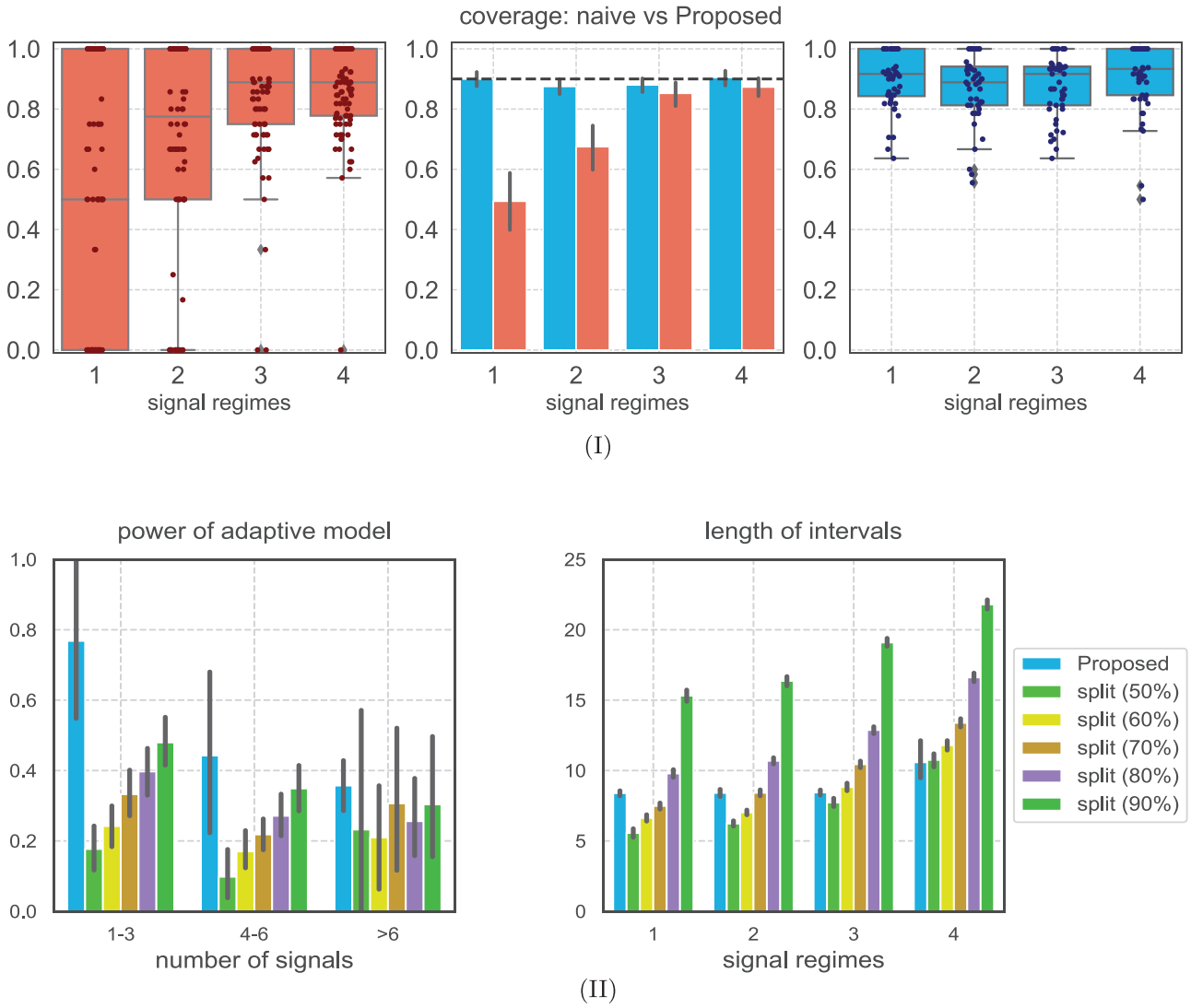


FIGURE 2 (I) Invalidity of naive inference (in orange)—the x-axis represents different signal regimes and the y-axis plots empirical coverage of 90% intervals across the regimes. Left panel shows the distribution of naive estimates that ignore adaptive nature of model; central panel compares the averaged coverages across the regimes; right panel plots empirical distribution of the proposed intervals. The dotted line at 0.90 is the nominal coverage. (II) Trade-off in model selection and inferential power—left panel highlights the quality of our integrative model measured via the screening power of true signals; right panel shows the distribution of averaged lengths of interval estimates in the selected model.

the validity of our interval estimates post selection, we present detailed comparisons for this trade-off of information between our approach and sample-splitting at different resolutions. In the Supporting Information, we showcase how our methods recover the target coverage as opposed to the single shot approach taken by the Bayesian LASSO (Park & Casella, 2008) under misspecified models. We also compare our method with an adaptation of the *post-double-selection* method in Belloni et al. (2014) and report our findings based on the statistical efficiency for the two approaches.

Figure 2 anchors the motivation behind using a selection-aware posterior in panel (I). The distribution of the empirical coverages of naive intervals that ignore the selection-aware nature of integrative models have averaged coverage falling way short of the benchmark target of 90%. The interval estimates furnished by our methods support the validity and necessity of the inferential proposal in the paper. Panel (II) exemplifies a significantly better reconciliation between the recovery of signals from the integrative model and the subsequent inferential power for the matched parameters, when compared against splitting based on varying proportions of data reserved for selection. Specifically, the performance of the rather unconventional randomized query (2) in terms of model selection is evaluated using the number of true signals

screened under different sparse scenarios and the follow-up inferential power is depicted as the lengths of interval estimates averaged across simulations in these signal regimes.

Observe, splitting where 90% of the data are assigned for selecting signals is the best performer among all the split-based methods in terms of model selection. However, this power is clearly dominated by the randomized scheme we adopt for modeling. In an assessment of inferential power, our methods accounting appropriately for the bias from model selection provide interval estimates, which are less than half the length of the 90% split-based intervals. A takeaway from this illustration is the attractive alternative our methods offer in comparison to splitting across a range of resolution in terms of data allocation for the two core tasks in Panel (II). Evidently, our methods allow a distinctly unique yet more efficient trade-off in the use of information for modeling and estimating uncertainties thereof.

5.3 | Inferential results: An illustration of our scope

We next demonstrate how our methods generalize in their application to other data dimensions beyond our focused study. The depiction in Figure 3, through the averaged coverages of naive and proposed intervals across the regression dimensions $r/n = 2, 1, 0.5$, emphasizes the strong need to correct for selection bias. For the signal regimes described in our simulation design, we see a severe shortfall of coverage for the naive interval estimates, ranging as low as $\sim 10\%$ and increasing to a level of only $\sim 70\%$ in the moderate signal-to-noise ratio (SNR) regimes. We remark here that the coverage of the naive intervals worsens in comparison to the case study in the previous discussion. This difference in the behavior of the naive intervals can be attributed to the synthetic predictor values we append to the real radiogenomic observations in the simulations in order to vary the size of regression in this design.

In panel (II) of the same figure, we highlight (1) the sharpness of the selection-aware model in (A), (B), (C) in terms of the number of signals screened by the randomized strategy (2) and split-based schemes, (2) inferential power in (D), (E), (F) measured as the averaged lengths of the interval estimates produced by the Bayesian proposal in the paper when compared to splitting. Coherent with the findings in the preceding discussion, the proposed methods dominate all the split-based methods when assessed for the quality of the selected model; the percentage in the legend indicates the proportion of data samples reserved for model selection.

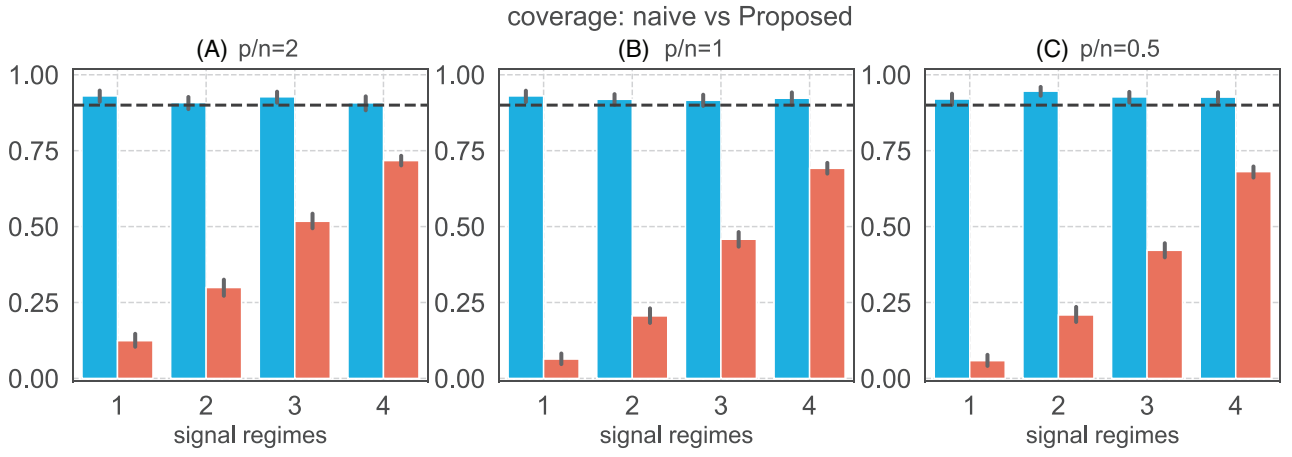
Consistent with the patterns in the preceding discussion, Figure 3 summarizes the advantages our selection-aware techniques enjoy over the common practice of splitting the data into two parts. Applying conditional inference corrects for the bias from a randomized version of model selection and permits an optimal reuse of data samples at the same time.

6 | RADIOGENOMIC ANALYSIS FOR LGG

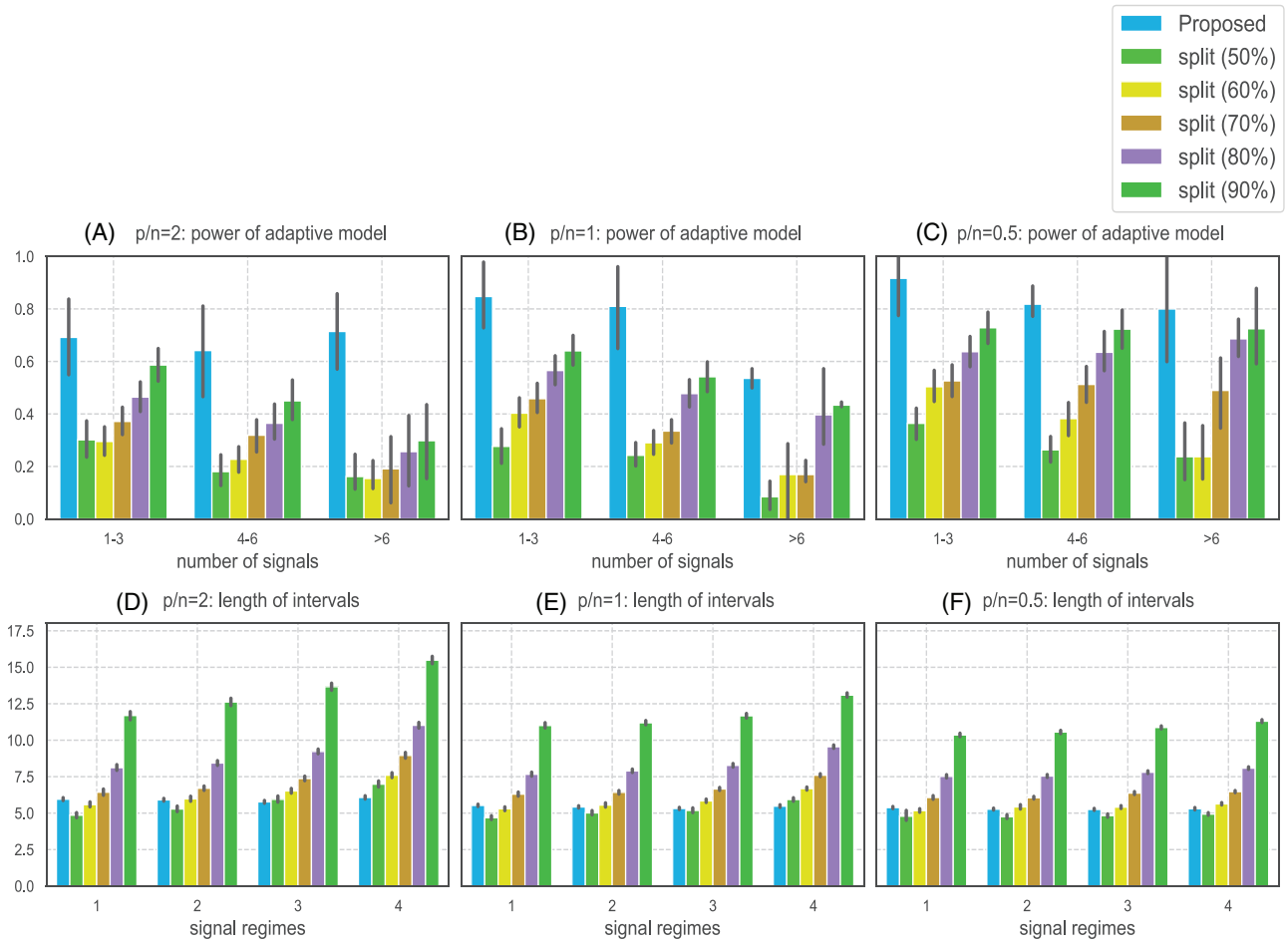
6.1 | Selection-aware pipeline with radiogenomic characteristics

In our radiogenomic case study, the imaging outcomes, also called radiomic phenotypes, are collectively harnessed in integrative models with the genomic measurements to assess associations with overall survival. Data acquisition and preprocessing steps for both the radiomic phenotypes and genomic measurements are deferred to the Supporting Information. The first step of the integrative selection pipeline identifies a set of promising pathways associated with the radiomic-based intermediary (imaging) outcomes. To this end, we solve (1) with the 143 principal component scores across the 12 groups of tumor voxels as responses, regressed against 1289 pathways from the four pathway collections (Hallmark, KEGG, C4, and C6). The output of this step is a set of 369 gene pathways, each of which is associated with one or more of the radiomic phenotypes. Of these 369 pathways selected, we note that the multiplicity of each pathway, defined as the number of LASSO queries, which selects this potential predictor, ranges between 1 and 9.

Using the log-transformed values of overall survival times as our clinical outcome, the second step solves a randomized version of LASSO (2) to partition the information within our data toward selecting a model and ascertaining strengths of these selected associations. The penalty weights in the LASSO are set to be inversely proportional to the multiplicity of a pathway to reflect an importance weight for that feature in terms of its association with the imaging outcomes. We select 15 pathways from this step; these pathways are the imaging informed explanatory variables associated with survival. Letting $\bar{E} = E$ in (4) and using the prior in (6), we use the optimization-based expressions for the selection-aware posterior to adjust for bias from the integrative selection pipeline. Inference for the adaptively determined parameters $\boldsymbol{\beta}_E$ gives us effect size estimates for the pathways indexed by E. In Figure 4 we showcase the bounds for 50%, 80%, and 95% credible intervals based on the MCMC samples for the 15 selected pathways.



(I)



(II)

FIGURE 3 (I) The x-axis represents different signal regimes and the y-axis plots empirical coverage of 90% intervals across the regimes with the dotted line at 0.90 as the target coverage. The panels depict the averaged coverages of naive and proposed intervals across the regression dimensions $r/n = 2, 1, 0.5$, highlighting the severely flawed nature of inferential decisions when the selection-aware nature of models is ignored in moderate dimensional regimes. (II) Assessment of the screening power for the selected models in Panels (A), (B), (C), followed by the matching inferential power in Panels (D), (E), (F)—showcasing a trade-off between selection and inference

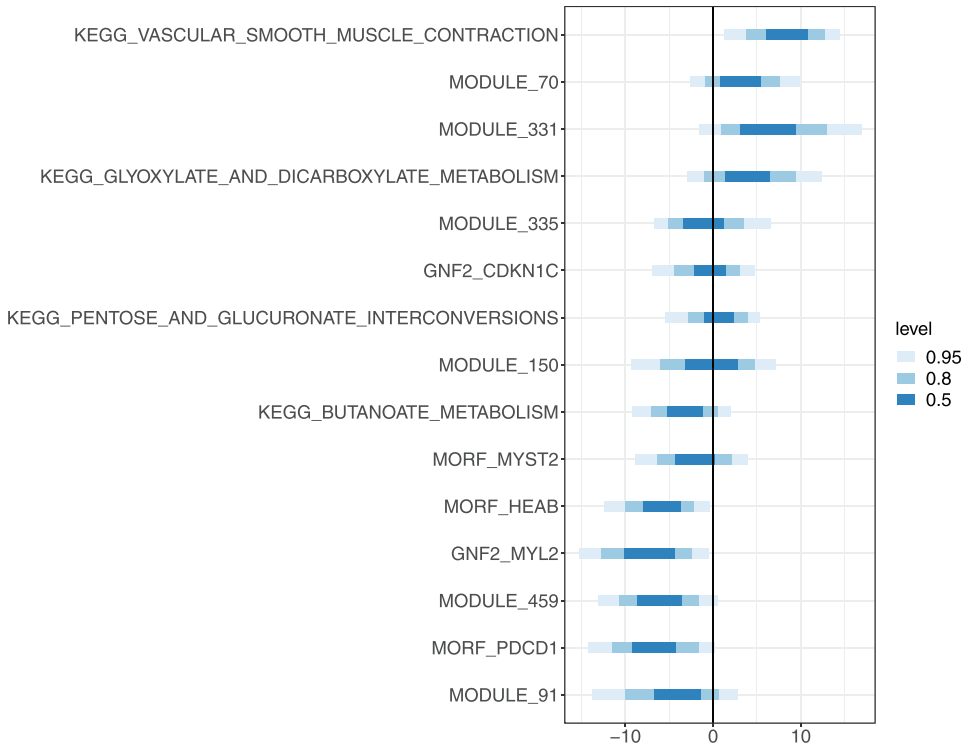


FIGURE 4 Bounds of the credible intervals for the gene pathways associated with the overall survival and with the radiomic characteristics

6.2 | Biological interpretations

We now focus on some of our findings, providing their biological implications and interpreting the same in the context of existing clinical knowledge in this domain.

(1) We see that the gene pathway, vascular smooth muscle contraction, from the KEGG collection has significant association with overall survival. Vascular smooth muscle cell (VSMC) is a highly specialized cell whose principal function is contraction. These cells shorten on contraction, consequently decreasing the diameter of a blood vessel to regulate the blood flow and pressure. Moreover, in a clinically relevant mouse model of glioma, it was found that the glioma cells disrupt the VSMCs as they populate the perivascular space of preexisting vessels, causing a focal breach in the blood brain barrier (Watkins et al., 2014). It has been demonstrated that endothelial-specific growth factor, such as vascular endothelial growth factor (VEGF), can interact with nonendothelial cells and play a role in modulating the response of VSMCs. VEGF expression levels were associated with the presence of ringlike tumor contrast enhancement, which present phenotypically as variable contrast on T1Gd MRI scan and were jointly associated with progression-free survival in glioblastoma (Wang et al., 2016).

(2) The gene pathway MORF PDCD1, which includes the genes in the neighborhood of the gene PDCD1, is seen to have a significant association with the overall survival. Röver et al. (2018) indicate that PDCD1 promoter methylation is a prognostic factor in LGG with isocitrate dehydrogenase (IDH) mutations. It is known that high expression of PDCD1 on the immune cells infiltrating the LGG is a marker for immune evasion and associated with survival.

(3) Another significant association in our analysis is the gene pathway GNF2 MYL2. This is a group of genes in the neighborhood of MYL2 (myosin light chain II). Previous studies (Beadle et al., 2008) show that myosin II plays a significant role in glioma invasion in vivo, where it regulates the deformation of the nucleus as well as the membrane of glioma cells. This has been further validated through mathematical modeling in recent literature (Lee et al., 2017).

Some of the other pathways associated with the overall survival include metabolic pathways from KEGG such as (1) pentose and glucuronate interconversions, (2) glyoxylate and dicarboxylate metabolism, and (3) butanoate metabolism. In light of the significant role that metabolic reprogramming plays in glioma pathogenesis (Strickland & Stoll, 2017), it is encouraging to see that a number of metabolic pathways are identified to be significantly

associated with patient prognosis. From the pathways for cancer gene neighborhoods we see significant associations with the gene pathways such as neighborhoods of (1) MYST2, a histone acetyltransferase that plays crucial functions in transcription, DNA replication, and repair; and (2) CDKN1C, which is known to regulate several of the hallmark properties of cancer (Kavanagh & Joseph, 2011). A follow-up validation of these pathways will illuminate nuances in understanding the tumor etiology under study.

7 | CONCLUDING REMARKS

We conclude by remarking that there is certainly room for future directions. Integrative models may be prohibitive if there are important genomic variable(s) with a strong impact on the clinical outcome, but are culled out in the regression with the imaging outcomes. Models attempting to link genomic variables directly with clinical endpoints might, however, be less viable, especially if the number of explanatory variables is many times larger than the number of available samples and subsets of these variables share substantial correlations. In these situations, eliminating the upstream regression with the imaging outcomes may lead to a loss of accuracy as well as power in terms of support recovery. Some discussion around a direct modeling approach (without the use of intermediary outcomes) is provided for our radiogenomic study in the Supporting Information. Recently, work by Panigrahi et al. (2020) proposes selection-aware methods for regression with the Group LASSO penalty. A generalization of our methodology for regression with other structured penalties in the integrative domain and on radiogenomic studies are left as promising directions for future work.

ACKNOWLEDGMENTS

S.P. acknowledges support through NSF-DMS 1951980 and NSF-DMS 2113342. S.M., A.R, and V.B. were supported through CCSG P30 CA046592, Precision Health Scholar award (to S.M.), Institutional Research Grants from The University of Michigan, NCI R37CA214955-01A1, and a Research Scholar Grant from the American Cancer Society (RSG-16-005-01). V.B. was also supported by NIH grants R01-CA160736, R21-CA220299, NSF grant 1463233, and start-up funds from the U-M Rogel Cancer Center and School of Public Health.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available in the Supporting Information of this paper.

ORCID

Snigdha Panigrahi https://orcid.org/0000-0001-8869-207X

Shariq Mohammed https://orcid.org/0000-0002-3107-2969

Veerabhadran Baladandayuthapani https://orcid.org/0000-0001-9107-3157

REFERENCES

Beadle, C., Assanah, M.C., Monzo, P., Vallee, R., Rosenfeld, S.S. & Canoll, P. (2008) The role of Myosin II in glioma invasion of the brain. *Molecular Biology of the Cell*, 19, 3357–3368.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014) Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81, 608–650.

Benjamini, Y. & Yekutieli, D. (2005) False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100, 71–81.

Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. (2013) Valid post-selection inference. *The Annals of Statistics*, 41, 802–837.

Hurvich, C.M. & Tsai, C. (1990) The impact of model selection on inference in linear regression. *The American Statistician*, 44, 214–217.

Kass, R.E. & Raftery, A.E. (1995) Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Kavanagh, E. & Joseph, B. (2011) The hallmarks of cdkn1c (p57, kip2) in cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1816, 50–56.

Lee, J.D., Sun, D.L., Sun, Y. & Taylor, J.E. (2016) Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44, 907–927.

Lee, W., Lim, S. & Kim, Y. (2017) The role of Myosin II in glioma invasion: a mathematical model. *PLoS One*, 12, e0171312.

Panigrahi, S., MacDonald, P.W. & Kessler, D. (2020) Approximate post-selective inference for regression with the group lasso. *arXiv preprint arXiv:2012.15664*.

Panigrahi, S. & Taylor, J. (2018) Scalable methods for Bayesian selective inference. *Electronic Journal of Statistics*, 12, 2355–2400.

Panigrahi, S. & Taylor, J. (2022) Approximate selective inference via maximum likelihood. *Journal of the American Statistical Association*. 1–24 Just accepted.

Panigrahi, S., Taylor, J. & Weinstein, A. (2021) Integrative methods for post-selection inference under convex constraints. *The Annals of Statistics*, 49, 2803–2824.

Panigrahi, S., Zhu, J. & Sabatti, C. (2019) Selection-adjusted inference: an application to confidence intervals for cis-eQTL effect sizes. *Biostatistics*. kxz024.

Park, T. & Casella, G. (2008) The Bayesian lasso. *Journal of the American Statistical Association*, 103, 681–686.

Röver, L.K., Gevensleben, H., Dietrich, J., Bootz, F., Landsberg, J., Goltz, D. & Dietrich, D. (2018) Pd-1 (pdcd1) promoter methylation is a prognostic factor in patients with diffuse lower-grade gliomas harboring isocitrate dehydrogenase (IDH) mutations. *EBioMedicine*, 28, 97–104.

Strickland, M. & Stoll, E.A. (2017) Metabolic reprogramming in glioma. *Frontiers in Cell and Developmental Biology*, 5, 43.

Tian, X., Panigrahi, S., Markovic, J., Bi, N. & Taylor, J. (2016) Selective sampling after solving a convex problem. *arXiv preprint arXiv:1609.05609*.

Tian, X. & Taylor, J. (2018) Selective inference with a randomized response. *The Annals of Statistics*, 46, 679–710.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

Wang, K., Wang, Y., Wang, J., Ma, J., Jiang, T. & Dai, J. (2016) Radiologic features and expression of vascular endothelial growth factor stratify survival outcomes in patients with glioblastoma. *American Journal of Neuroradiology*, 37, 629–635.

Watkins, S., Robel, S., Kimbrough, I.F., Robert, S.M., Ellis-Davies, G. & Sontheimer, H. (2014) Disruption of astrocyte–vascular coupling and the blood–brain barrier by invading glioma cells. *Nature Communications*, 5, 1–15.

Yekutieli, D. (2012) Adjusted Bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 515–541.

Zhang, Y., Morris, J.S., Aerry, S.N., Rao, A.U. & Baladandayuthapani, V. (2019) Radio-ibag: radiomics-based integrative Bayesian analysis of multiplatform genomic data. *The Annals of Applied Statistics*, 13, 1957–1988.

SUPPORTING INFORMATION

Figures and supplementary material referenced in Sections 2, 3, 4, 5 and 6 are available with this paper at the Biometrics website on Wiley Online Library. Code for computing a working version of our selection-aware posterior, and provide updates using a gradient-based prototype sampler is available with our paper.

Supporting Information

How to cite this article: Panigrahi, S., Mohammed, S., Rao, A. & Baladandayuthapani, V. (2023) Integrative Bayesian models using Post-selective inference: A case study in radiogenomics. *Biometrics*, 79, 1801–1813. https://doi.org/10.1111/biom.13740