Biometrics 000, 000–000 DOI: 000
000 0000

Supporting Information for "Integrative Bayesian models using Post-selective Inference: a case study in Radiogenomics" by Snigdha Panigrahi, Shariq Mohammed, Arvind Rao, and Veerabhadran Baladandayuthapani

Summary: The supplementary material contains proofs for our technical results and additional details for our empirical analyses including comparisons of the finite-sample performance of our methods with other inferential approaches. The file is organized as follows:

(1) Section A details out the proof for the theory in the paper.

(2) Section B provides details of the approximation we apply to obtain a working version of the selection-aware posterior.

(3) Section C provides additional comparisons of our selection-aware methods with alternatives, and details out the construct of the radiogenomic features in our application.

This paper has been submitted for consideration for publication in *Biometrics*

A. Proofs of main results

Proof. Proposition 1. Using the following notations

$$\boldsymbol{\mathcal{I}}_i = \begin{pmatrix} I_{i,1} & \cdots & I_{i,L} \end{pmatrix}, \quad \boldsymbol{\mu}_i(\boldsymbol{\alpha}_{\bar{F}}) = \begin{pmatrix} \mathbf{G}_{i;F_1}\boldsymbol{\alpha}_{F_1} & \cdots & \mathbf{G}_{i;F_L}\boldsymbol{\alpha}_{F_L} \end{pmatrix},$$

and using the independence between the clinical outcome, the intermediary outcome and the randomization variable, we observe that the selection-ignorant (unconditional) likelihood under (4) and (5) is proportional to

$$\begin{aligned} &\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{G}_{\bar{E}}\boldsymbol{\beta}_{\bar{E}})^T(\mathbf{y}-\mathbf{G}_{\bar{E}}\boldsymbol{\beta}_{\bar{E}})\right)\cdot\exp\left(-\frac{1}{2\eta^2}\mathbf{r}^T\mathbf{r}\right) \\ &\qquad\times\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\mathcal{I}}_i-\boldsymbol{\mu}_i(\boldsymbol{\alpha}_{\bar{F}}))^T\boldsymbol{\Sigma}_I^{-1}(\boldsymbol{\mathcal{I}}_i-\boldsymbol{\mu}_i(\boldsymbol{\alpha}_{\bar{F}}))\right). \end{aligned} \tag{16}$$

Fix the following sets:

$$\mathcal{F}_0 := \left\{\boldsymbol{\mathcal{I}} : \widehat{F}_l(\boldsymbol{\mathcal{I}}_l) = F_l \text{ for } l \in \{1,\cdots,L\}\right\};$$

$$\mathcal{E}_0 := \left\{(\mathbf{y},\mathbf{r}) : \widehat{E}(\mathbf{y},\mathbf{r}) = E\right\}.$$

Truncating the law for the random variables $\mathbf{Y}$, $\mathbf{I}$, $\mathbf{R}$ in (16) to the event resulting in the selected sets (7) yields us a joint conditional law proportional to:

$$\begin{aligned} &\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{G}_{\bar{E}}\boldsymbol{\beta}_{\bar{E}})^T(\mathbf{y}-\mathbf{G}_{\bar{E}}\boldsymbol{\beta}_{\bar{E}})\right)\cdot\exp\left(-\frac{1}{2\eta^2}\mathbf{r}^T\mathbf{r}\right) \\ &\qquad\cdot\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\mathcal{I}}_i-\boldsymbol{\mu}_i(\boldsymbol{\alpha}_{\bar{F}}))^T\boldsymbol{\Sigma}_I^{-1}(\boldsymbol{\mathcal{I}}_i-\boldsymbol{\mu}_i(\boldsymbol{\alpha}_{\bar{F}}))\right)\cdot 1_{\mathcal{F}_0}(\boldsymbol{\mathcal{I}})\cdot 1_{\mathcal{E}_0}(\mathbf{y},\mathbf{r}). \end{aligned}$$

Note, the normalizing constant for the likelihood function derived from the above conditional law is given by

$$\mathbb{P}[\widehat{E}(\mathbf{Y},\mathbf{R}) = E \mid \boldsymbol{\beta}_{\bar{E}}] \times \mathbb{P}[\widehat{F}_l(\mathbf{I}_l) = F_l \text{ for } l \in \{1,2,\cdots,L\} \mid \boldsymbol{\alpha}_{F_1},\cdots,\boldsymbol{\alpha}_{F_L}].$$

This is because the probability

$$\mathbb{P}[\widehat{E}(\mathbf{Y},\mathbf{R}) = E,\ \widehat{F}_l(\mathbf{I}_l) = F_l \text{ for } l \in \{1,2,\cdots,L\} \mid \boldsymbol{\beta}_{\bar{E}},\ \boldsymbol{\alpha}_{F_1},\cdots,\boldsymbol{\alpha}_{F_L}]$$

decouples due to the independence we assume between the intermediary outcomes and the primary outcome as well as the randomization variable . Lastly, observe that the probability involving exclusively the auxiliary parameters

$$\boldsymbol{\alpha}_{F_1},\cdots,\boldsymbol{\alpha}_{F_L}$$

contributes to a constant in our likelihood, a function of $\boldsymbol{\beta}_{\bar{E}}$. Our conditional likelihood is therefore proportional to

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

which agrees with the expression in Proposition 1.

Define the following matrices that we use to characterize the selection event of interest as a union of polyhedral regions in Proposition 2:

$$\mathbf{U}_{\mathbf{s}_E} = \begin{bmatrix} \mathrm{diag}(\mathbf{s}_E)(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1}\mathbf{G}_E^T\mathbf{G}_{\bar{E}} \\ \mathbf{G}_{E^c}^T\mathbf{G}_{\bar{E}} - \mathbf{G}_{E^c}^T\mathbf{G}_E(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1}\mathbf{G}_E^T\mathbf{G}_{\bar{E}} \\ -\mathbf{G}_{E^c}^T\mathbf{G}_{\bar{E}} + \mathbf{G}_{E^c}^T\mathbf{G}_E(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1}\mathbf{G}_E^T\mathbf{G}_{\bar{E}} \end{bmatrix},$$

$$\mathbf{V}_{\mathbf{s}_E} = \begin{bmatrix} \mathrm{diag}(\mathbf{s}_E)(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1} & \mathbf{0} \\ -\mathbf{G}_{E^c}^T\mathbf{G}_E(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1} & \mathbf{I} \\ \mathbf{G}_{E^c}^T\mathbf{G}_E(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1} & -\mathbf{I} \end{bmatrix}, \mathbf{W}_{\mathbf{s}_E} = \begin{bmatrix} \mathrm{diag}(\mathbf{s}_E)(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1} & \mathbf{0} \\ -\mathbf{G}_{E^c}^T\mathbf{G}_E(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1} & \mathbf{I} \\ \mathbf{G}_{E^c}^T\mathbf{G}_E(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1} & -\mathbf{I} \end{bmatrix},$$

$$\mathbf{t}_{\mathbf{s}_E} = \begin{bmatrix} \mathrm{diag}(\mathbf{s}_E)(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1}\boldsymbol{\Lambda}_E\mathbf{s}_E \\ -\boldsymbol{\lambda}_{E^c} - \mathbf{G}_{E^c}^T\mathbf{G}_E(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1}\boldsymbol{\Lambda}_E\mathbf{s}_E \\ -\boldsymbol{\lambda}_{E^c} + \mathbf{G}_{E^c}^T\mathbf{G}_E(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1}\boldsymbol{\Lambda}_E\mathbf{s}_E \end{bmatrix}.$$

Proof. Proposition 2. To see a proof for Proposition 2, the selection of active variables E with signs $\mathbf{s}_E$ is equivalent to

$$\mathrm{sign}(\widehat{\boldsymbol{\beta}}_E^{\mathrm{LASSO}}) = \mathbf{s}_E \ , \ \|\mathbf{z}\|_\infty < 1,$$

where $\widehat{\boldsymbol{\beta}}_E^{\mathrm{LASSO}}$ and $\mathbf{z}$ are obtained from the stationary equation at the solution of (2), given by:

$$\begin{aligned} &\begin{pmatrix} \mathbf{r}_E^T & \mathbf{r}_{E^c}^T \end{pmatrix}^T + \begin{pmatrix} (\mathbf{G}_E^T\mathbf{y})^T & (\mathbf{G}_{E^c}^T\mathbf{y})^T \end{pmatrix}^T \\ &= \begin{bmatrix} \mathbf{G}_E^T\mathbf{G}_E + \epsilon\cdot\mathbf{I} & \mathbf{G}_E^T\mathbf{G}_{E^c} \end{bmatrix}^T \widehat{\boldsymbol{\beta}}_E^{\mathrm{LASSO}} + \begin{pmatrix} (\boldsymbol{\Lambda}_E\mathbf{s}_E)^T & (\boldsymbol{\Lambda}_{E^c}\mathbf{z})^T \end{pmatrix}^T. \end{aligned} \tag{17}$$

Using the stationary equation (17) for realizations $(\mathbf{y}, \mathbf{r})$ and based on the decomposition $\mathbf{G}_{\bar{F}}^T\mathbf{y} = \mathbf{G}_{\bar{F}}^T\mathbf{G}_{\bar{E}}\widehat{\boldsymbol{\beta}}_{\bar{E}} + \widehat{\boldsymbol{\beta}}^{\perp}$, we first note that the sign constraints are equivalent to

$$\mathrm{diag}(\mathbf{s}_E)(\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I})^{-1}(\mathbf{r}_E + \mathbf{G}_E^T\mathbf{G}_{\bar{E}}\widehat{\boldsymbol{\beta}}_{\bar{E}} + \widehat{\boldsymbol{\beta}}_E^{\perp} - \boldsymbol{\Lambda}_E\mathbf{s}_E) > 0.$$

Next observe the inactive coordinates of the stationary equation (10) yield the equality:

$$\mathbf{r}_{E^c} + \mathbf{G}_{E^c}^T \mathbf{G}_{\bar{E}} \widehat{\boldsymbol{\beta}}_{\bar{E}} + \widehat{\boldsymbol{\beta}}_{E^c}^{\perp} - \mathbf{G}_{E^c}^T \mathbf{G}_E \widehat{\boldsymbol{\beta}}_E^{\text{LASSO}} = \boldsymbol{\Lambda}_{E^c} \mathbf{z}.$$

Coupled with the equation

$$\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}} = (\mathbf{G}_E^T \mathbf{G}_E + \epsilon \mathbf{I})^{-1} (\mathbf{r}_E + \mathbf{G}_E^T \mathbf{G}_{\bar{E}} \widehat{\boldsymbol{\beta}}_{\bar{E}} + \widehat{\boldsymbol{\beta}}^{\perp} - \boldsymbol{\Lambda}_E \mathbf{s}_E),$$

the restriction on the ℓ_∞-norm of inactive subgradient vector $\mathbf{z}$ is equivalent to the following constraints

$$\mathbf{r}_{E^c} - \mathbf{G}_{E^c}^T \mathbf{G}_E (\mathbf{G}_E^T \mathbf{G}_E + \epsilon \mathbf{I})^{-1} \mathbf{r}_E + \left(\mathbf{G}_{E^c}^T \mathbf{G}_{\bar{E}} - \mathbf{G}_{E^c}^T \mathbf{G}_E (\mathbf{G}_E^T \mathbf{G}_E + \epsilon \mathbf{I})^{-1} \mathbf{G}_E^T \mathbf{G}_{\bar{E}} \right) \widehat{\boldsymbol{\beta}}_{\bar{E}}$$
$$+ \widehat{\boldsymbol{\beta}}_{E^c}^{\perp} - \mathbf{G}_{E^c}^T \mathbf{G}_E (\mathbf{G}_E^T \mathbf{G}_E + \epsilon \mathbf{I})^{-1} \widehat{\boldsymbol{\beta}}_E^{\perp} > -\boldsymbol{\lambda}_{E^c} - \mathbf{G}_{E^c}^T \mathbf{G}_E (\mathbf{G}_E^T \mathbf{G}_E + \epsilon \mathbf{I})^{-1} \boldsymbol{\Lambda}_E \mathbf{s}_E,$$

and

$$-\mathbf{r}_{E^c} + \mathbf{G}_{E^c}^T \mathbf{G}_E (\mathbf{G}_E^T \mathbf{G}_E + \epsilon \mathbf{I})^{-1} \mathbf{r}_E - \left(\mathbf{G}_{E^c}^T \mathbf{G}_{\bar{E}} - \mathbf{G}_{E^c}^T \mathbf{G}_E (\mathbf{G}_E^T \mathbf{G}_E + \epsilon \mathbf{I})^{-1} \mathbf{G}_E^T \mathbf{G}_{\bar{E}} \right) \widehat{\boldsymbol{\beta}}_{\bar{E}}$$
$$- \widehat{\boldsymbol{\beta}}_{E^c}^{\perp} + \mathbf{G}_{E^c}^T \mathbf{G}_E (\mathbf{G}_E^T \mathbf{G}_E + \epsilon \mathbf{I})^{-1} \widehat{\boldsymbol{\beta}}_E^{\perp} > -\boldsymbol{\lambda}_{E^c} + \mathbf{G}_{E^c}^T \mathbf{G}_E (\mathbf{G}_E^T \mathbf{G}_E + \epsilon \mathbf{I})^{-1} \boldsymbol{\Lambda}_E \mathbf{s}_E.$$

The union of polytopes now follows from considering all possible signs for the active coefficients, which varies over the set $\{-1, 1\}^{|E|}$.

Fixing some matrices to derive Theorem 1, we let

$$\mathbf{Q} = \left[\mathbf{G}_E^T \mathbf{G}_E + \epsilon \cdot \mathbf{I} \quad \mathbf{G}_E^T \mathbf{G}_{E^c} \right]^T, \ \mathbf{S} = -\mathbf{G}_{\bar{F}}^T \mathbf{G}_{\bar{E}}, \ \ \mathbf{s} = \left((\boldsymbol{\Lambda}_E \mathbf{s}_E)^T \quad (\boldsymbol{\Lambda}_{E^c} \mathbf{z})^T \right)^T - \widehat{\boldsymbol{\beta}}^{\perp}$$

$$\mathbf{P} = -(\mathbf{Q}^T \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{S}, \ \mathbf{o} = -(\mathbf{Q}^T \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{s}.$$

Then, we set the following notations based on the above matrices:

$$\boldsymbol{\Theta}_{\bar{E}} = \left(\boldsymbol{\Sigma}_{\bar{E}}^{-1} + \eta^{-2} \mathbf{S}^T \mathbf{S} - \eta^{-2} \mathbf{P}^T \mathbf{Q}^T \mathbf{Q} \mathbf{P} \right)^{-1};$$

$$\mathbf{K} = \boldsymbol{\Theta}_{\bar{E}} \boldsymbol{\Sigma}_{\bar{E}}^{-1}, \ \mathbf{l} = \boldsymbol{\Theta}_{\bar{E}} (\eta^{-2} \mathbf{P}^T \mathbf{Q}^T \mathbf{Q} \mathbf{o} - \eta^{-2} \mathbf{S}^T \mathbf{s}).$$

Proof. Theorem 1. For any fixed set $\bar{E}$, the independence between $\widehat{\boldsymbol{\beta}}_{\bar{E}}$, $\widehat{\boldsymbol{\beta}}^{\perp}$ and $\mathbf{r}$ leads us to their unconditional likelihood:

$$\rho(\widehat{\boldsymbol{\beta}}_{\bar{E}}; \boldsymbol{\beta}_{\bar{E}}, \boldsymbol{\Sigma}_{\bar{E}}) \cdot \rho(\mathbf{r}; \mathbf{0}, \eta^2 \mathbf{I}) \cdot \rho(\widehat{\boldsymbol{\beta}}^{\perp}; \mathbf{0}, \boldsymbol{\Sigma}^{\perp}),$$

where $\boldsymbol{\Sigma}^{\perp}$ denotes the covariance for $\widehat{\boldsymbol{\beta}}^{\perp}$. To derive an expression for the conditional

likelihood, we use a change of measure:

$$\mathbf{R} \to \left(\widehat{\mathbf{B}}^{\text{LASSO}},\ \ \widehat{\mathbf{Z}}\right)$$

defined through the stationary equation of the randomized LASSO:

$$\mathbf{R} = -\mathbf{G}_{\bar{F}}^T\mathbf{G}_{\bar{E}}\widehat{\boldsymbol{\beta}}_{\bar{E}} - \widehat{\boldsymbol{\beta}}^{\perp} + \begin{bmatrix}\mathbf{G}_E^T\mathbf{G}_E + \epsilon\mathbf{I}\\ \mathbf{G}_{E^c}^T\mathbf{G}_E\end{bmatrix}\widehat{\mathbf{B}}^{\text{LASSO}} + \begin{pmatrix}\boldsymbol{\Lambda}_E\mathbf{s}_E\\ \boldsymbol{\Lambda}_{E^c}\widehat{\mathbf{Z}}\end{pmatrix} \tag{18}$$

such that $\widehat{\mathbf{B}}^{\text{LASSO}}$ denotes the active LASSO solution and $\widehat{\mathbf{Z}}$ denotes the subgradient from the ℓ_1 penalty evaluated at the inactive coordinates. This mapping coincides with (10) for our observed data, when $\widehat{\mathbf{B}}^{\text{LASSO}}$ and $\widehat{\mathbf{Z}}$ assume the values $\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}$ and $\mathbf{z}$ respectively. Applying the change of measure (18), the likelihood for the new variables $\widehat{\boldsymbol{\beta}}_{\bar{E}}$, $\widehat{\boldsymbol{\beta}}^{\perp}$, $\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}$ and $\mathbf{z}$ agrees with

$$\rho(\widehat{\boldsymbol{\beta}}_{\bar{E}};\boldsymbol{\beta}_{\bar{E}},\boldsymbol{\Sigma}_{\bar{E}})\cdot\rho(\mathbf{S}\widehat{\boldsymbol{\beta}}_{\bar{E}} + \mathbf{Q}\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}} + \mathbf{s};0,\eta^2\mathbf{I})\cdot\rho(\widehat{\boldsymbol{\beta}}^{\perp};\mathbf{0},\boldsymbol{\Sigma}^{\perp})$$

up to constants. We use the fact that the Jacobian for this change of measure is free of $\widehat{\mathbf{B}}^{\text{LASSO}}$ and $\widehat{\mathbf{Z}}$ and therefore dissolves as a constant. Conditioning upon the event

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

observed after solving (2), yields us the following truncated likelihood for the variables $\widehat{\boldsymbol{\beta}}_{\bar{E}}$ and $\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}$:

$$\begin{aligned}
&\left(\int \rho(b;\boldsymbol{\beta}_{\bar{E}},\boldsymbol{\Sigma}_{\bar{E}})\cdot\rho(\mathbf{S}b+\mathbf{Q}w+\mathbf{s};\mathbf{0},\eta^2\mathbf{I})\cdot\rho(\widehat{\boldsymbol{\beta}}^{\perp};\mathbf{0},\boldsymbol{\Sigma}^{\perp})\cdot 1_{\{\text{sign}(w)=\mathbf{s}_E\}}dwdb\right)^{-1}\\
&\times\rho(\widehat{\boldsymbol{\beta}}_{\bar{E}};\boldsymbol{\beta}_{\bar{E}},\boldsymbol{\Sigma}_{\bar{E}})\cdot\rho(\mathbf{S}\widehat{\boldsymbol{\beta}}_{\bar{E}}+\mathbf{Q}\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}+\mathbf{s};\mathbf{0},\eta^2\mathbf{I})\cdot\rho(\widehat{\boldsymbol{\beta}}^{\perp};\mathbf{0},\boldsymbol{\Sigma}^{\perp})\cdot 1_{\left\{\text{sign}(\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}})=\mathbf{s}_E\right\}}\\
&=\left(\int \rho(b;\boldsymbol{\beta}_{\bar{E}},\boldsymbol{\Sigma}_{\bar{E}})\cdot\rho(\mathbf{S}b+\mathbf{Q}w+\mathbf{s};\mathbf{0},\eta^2\mathbf{I})\cdot 1_{\{\text{sign}(w)=\mathbf{s}_E\}}dwdb\right)^{-1}\\
&\qquad\times\rho(\widehat{\boldsymbol{\beta}}_{\bar{E}};\boldsymbol{\beta}_{\bar{E}},\boldsymbol{\Sigma}_{\bar{E}})\cdot\rho(\mathbf{S}\widehat{\boldsymbol{\beta}}_{\bar{E}}+\mathbf{Q}\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}+\mathbf{s};\mathbf{0},\eta^2\mathbf{I})\quad\cdot 1_{\left\{\text{sign}(\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}})=\mathbf{s}_E\right\}}.
\end{aligned}$$

Integrating out $\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}$ from the likelihood, we are left with the marginal likelihood for $\widehat{\boldsymbol{\beta}}_{\bar{E}}$ which is proportional to:

$$\left(\int_{\text{sign}(w)=\mathbf{s}_E}\rho(b;\mathbf{K}\boldsymbol{\beta}_{\bar{E}}+\mathbf{l},\boldsymbol{\Theta}_{\bar{E}})\cdot\rho(w;\mathbf{P}b+\mathbf{o},\eta^{-2}\mathbf{Q}^T\mathbf{Q})dwdb\right)^{-1}$$
$$\times\rho(\widehat{\boldsymbol{\beta}}_{\bar{E}};\mathbf{K}\boldsymbol{\beta}_{\bar{E}}+\mathbf{l},\boldsymbol{\Theta}_{\bar{E}}).$$

This completes the derivation of our conditional likelihood.

Proof. Theorem 2. (i). We begin by computing the Jacobian associated with

the reparameterization mapping; this Jacobian is equal to:

$$\mathbf{K}^{-1}\left(\mathbf{I}+\eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\mathbf{P}\right)-\eta^{-2}\mathbf{K}^{-1}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\frac{\partial}{\partial\boldsymbol{\zeta}_{\bar{E}}}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}})). \tag{19}$$

In order to compute $\dfrac{\partial}{\partial\boldsymbol{\zeta}_{\bar{E}}}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}))$, we note that $\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}})$ satisfies the estimating equation:

$$\eta^{-2}\mathbf{Q}^T\mathbf{Q}\left(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}})-\mathbf{P}\boldsymbol{\zeta}_{\bar{E}}-\mathbf{o}\right)+\nabla\mathrm{Barr}_{\mathbf{s}_E}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}))=0.$$

Taking a derivative of the estimating equation with respect to $\boldsymbol{\zeta}_{\bar{E}}$ yields us:

$$\left(\eta^{-2}\mathbf{Q}^T\mathbf{Q}+\nabla^2\mathrm{Barr}_{\mathbf{s}_E}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}))\right)\frac{\partial}{\partial\boldsymbol{\zeta}_{\bar{E}}}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}))=\eta^{-2}\mathbf{Q}^T\mathbf{Q}\mathbf{P},$$

which implies

$$\frac{\partial}{\partial\boldsymbol{\zeta}_{\bar{E}}}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}))=\left(\eta^{-2}\mathbf{Q}^T\mathbf{Q}+\nabla^2\mathrm{Barr}_{\mathbf{s}_E}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}))\right)^{-1}\eta^{-2}\mathbf{Q}^T\mathbf{Q}\mathbf{P}.$$

Plugging the value of $\dfrac{\partial}{\partial\boldsymbol{\zeta}_{\bar{E}}}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}))$ into (19), the Jacobian assumes the expression

$$\mathbf{K}^{-1}\left(\mathbf{I}+\eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\mathbf{P}\right)-\eta^{-4}\mathbf{K}^{-1}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\left(\eta^{-2}\mathbf{Q}^T\mathbf{Q}+\nabla^2\mathrm{Barr}_{\mathbf{s}_E}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}))\right)^{-1}\mathbf{Q}^T\mathbf{Q}\mathbf{P}.$$

Using the reparameterization mapping where

$$\mathbf{K}\boldsymbol{\beta}_{\bar{E}}+\mathbf{l}=\Psi(\boldsymbol{\zeta}_{\bar{E}}),$$

the log-posterior is given by:

$$\log\widetilde{\pi}(\boldsymbol{\zeta}_{\bar{E}}|\widehat{\boldsymbol{\beta}}_{\bar{E}})=\log|\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}})|+\log\pi(\mathbf{K}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}})-\mathbf{K}^{-1}\mathbf{l}\mid\widehat{\boldsymbol{\beta}}_{\bar{E}})$$

where $\pi(\cdot|\,\widehat{\boldsymbol{\beta}}_{\bar{E}})$ is the working version of the selection-aware posterior in (12). Letting $\mathbf{b}^*$, $\mathbf{w}^*$ be the optimal solutions for the problem:

$$U^*(\Psi(\boldsymbol{\zeta}_{\bar{E}}))=\inf_{b,w}\Big\{\frac{1}{2}(b-\Psi(\boldsymbol{\zeta}_{\bar{E}}))^T\boldsymbol{\Theta}_{\bar{E}}^{-1}(b-\Psi(\boldsymbol{\zeta}_{\bar{E}}))+\frac{1}{2\eta^2}(w-\mathbf{P}b-\mathbf{o})^T\mathbf{Q}^T\mathbf{Q}(w-\mathbf{P}b-\mathbf{o})+\mathrm{Barr}_{\mathbf{s}_E}(w)\Big\},$$

we observe that $\log \widetilde{\pi}(\boldsymbol{\zeta}_{\bar{E}}|\widehat{\boldsymbol{\beta}}_{\bar{E}})$ is given by

$$\begin{aligned}
&\log|\det(\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}}))| + \widehat{\boldsymbol{\beta}}_{\bar{E}}^T\boldsymbol{\Theta}_{\bar{E}}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}}) - \frac{1}{2}\Psi(\boldsymbol{\zeta}_{\bar{E}})^T\boldsymbol{\Theta}_{\bar{E}}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}}) + U^*(\Psi(\boldsymbol{\zeta}_{\bar{E}})) \\
&\qquad\qquad + \log\pi(\mathbf{K}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}}) - \mathbf{K}^{-1}\mathbf{l}) \\
&= \log|\det(\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}}))| + \frac{1}{2}\mathbf{b}^{*T}\boldsymbol{\Theta}_{\bar{E}}^{-1}\mathbf{b}^* + \widehat{\boldsymbol{\beta}}_{\bar{E}}^T\boldsymbol{\Theta}_{\bar{E}}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}}) - \mathbf{b}^{*T}\boldsymbol{\Theta}_{\bar{E}}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}}) \qquad (20) \\
&\quad + \frac{1}{2\eta^2}(\mathbf{w}^* - \mathbf{P}\mathbf{b}^* + \mathbf{o})^T\mathbf{Q}^T\mathbf{Q}(\mathbf{w}^* - \mathbf{P}\mathbf{b}^* + \mathbf{o}) + \mathrm{Barr}_{\mathbf{s}_E}(\mathbf{w}^*) \\
&\qquad\qquad + \log\pi(\mathbf{K}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}}) - \mathbf{K}^{-1}\mathbf{l}).
\end{aligned}$$

Our claim then follows by observing:

$$\begin{aligned}
\Psi(\boldsymbol{\zeta}_{\bar{E}}) &= \mathbf{b}^* + \eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\left(\frac{\partial}{\partial b}\mathbf{w}^*(b)\Big|_{\mathbf{b}^*} - \mathbf{P}\right)^T\mathbf{Q}^T\mathbf{Q}(\mathbf{w}^*(\mathbf{b}^*) - \mathbf{P}\mathbf{b}^* + \mathbf{o}) \\
&\quad + \boldsymbol{\Theta}_{\bar{E}}\left(\frac{\partial}{\partial b}\mathbf{w}^*(b)\Big|_{\mathbf{b}^*}\right)^T\nabla\mathrm{Barr}_{\mathbf{s}_E}(\mathbf{w}^*(b)) \\
&= \mathbf{b}^* - \eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}(\mathbf{w}^*(\mathbf{b}^*) - \mathbf{P}\mathbf{b}^* + \mathbf{o}) \\
&= \mathbf{b}^* + \eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\mathbf{P}\mathbf{b}^* + \eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}(\mathbf{o} - \mathbf{w}^*(\mathbf{b}^*)),
\end{aligned}$$

where we use the below optimality conditions for $\mathbf{w}^*$ to deduce the second equality

$$\eta^{-2}\mathbf{Q}^T\mathbf{Q}(\mathbf{w}^*(\mathbf{b}^*) - \mathbf{P}\mathbf{b}^* + \mathbf{o}) + \nabla\mathrm{Barr}_{\mathbf{s}_E}(\mathbf{w}^*(\mathbf{b}^*)) = 0.$$

From the definition of $\Psi(\boldsymbol{\zeta}_{\bar{E}})$, we have

$$\mathbf{b}^* = \boldsymbol{\zeta}_{\bar{E}}.$$

The expression for $\log \widetilde{\pi}(\boldsymbol{\zeta}_{\bar{E}}|\widehat{\boldsymbol{\beta}}_{\bar{E}})$ then immediately follows from (20).

(ii). Observe that $\nabla \log \widetilde{\pi}(\boldsymbol{\zeta}_{\bar{E}}|\widehat{\boldsymbol{\beta}}_{\bar{E}})$ equals

$$\begin{aligned}
&(\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}}))^T\cdot\left(\nabla\log\pi(\mathbf{K}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}}) - \mathbf{K}^{-1}\mathbf{l}) + \mathbf{K}^T\boldsymbol{\Theta}_{\bar{E}}^{-1}(\widehat{\boldsymbol{\beta}}_{\bar{E}} - \boldsymbol{\zeta}_{\bar{E}})\right) \\
&\quad - \boldsymbol{\Theta}_{\bar{E}}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}}) + \boldsymbol{\Theta}_{\bar{E}}^{-1}\boldsymbol{\zeta}_{\bar{E}} - \eta^{-2}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}) - \mathbf{P}\boldsymbol{\zeta}_{\bar{E}} - \mathbf{o}) + \nabla\log|\det(\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}}))| \\
&= (\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}}))^T\left(\nabla\pi(\mathbf{K}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}}) - \mathbf{K}^{-1}\mathbf{l}) + \mathbf{K}^T\boldsymbol{\Theta}_{\bar{E}}^{-1}(\widehat{\boldsymbol{\beta}}_{\bar{E}} - \boldsymbol{\zeta}_{\bar{E}})\right) + \nabla\log|\det(\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}}))|.
\end{aligned}$$

In obtaining the first expression, we use the estimating equation for the solution of the optimization problem (13):

$$\eta^{-2}\mathbf{Q}^T\mathbf{Q}\left(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}) - \mathbf{P}\boldsymbol{\zeta}_{\bar{E}} - \mathbf{o}\right) + \nabla\mathrm{Barr}_{\mathbf{s}_E}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}})) = 0,$$

and that

$$\frac{\partial}{\partial\boldsymbol{\zeta}_{\bar{E}}}(\Psi(\boldsymbol{\zeta}_{\bar{E}})) = \left(\mathbf{I} + \eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\mathbf{P}\right) - \eta^{-2}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\frac{\partial}{\partial\boldsymbol{\zeta}_{\bar{E}}}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}})) = \mathbf{K}\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}}).$$

The second equality uses the definition of the reparameterization map (14), that is:

$$\boldsymbol{\Theta}_{\bar{E}}^{-1}\boldsymbol{\zeta}_{\bar{E}} - \eta^{-2}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}(\mathbf{w}^*(\boldsymbol{\zeta}_{\bar{E}}) - \mathbf{P}\boldsymbol{\zeta}_{\bar{E}} - \mathbf{o}) - \boldsymbol{\Theta}_{\bar{E}}^{-1}\Psi(\boldsymbol{\zeta}_{\bar{E}}) = 0.$$

The proof is complete by calculating $\nabla \log|\det(\boldsymbol{\mathcal{J}}(\boldsymbol{\zeta}_{\bar{E}}))|$. Let us the diagonal matrices

$$\text{diag}(\nabla^3\text{Barr}_{1;\mathbf{s}_{E;1}}(w_1^*(\boldsymbol{\zeta}_{\bar{E}})), \cdots, \nabla^3\text{Barr}_{|E|;\mathbf{s}_{E;|E|}}(w_E^*(\boldsymbol{\zeta}_{\bar{E}}))$$

and $\text{diag}(\eta^{-2}[\mathbf{N}^{-1}\mathbf{Q}^T\mathbf{Q}\mathbf{P}]_j)$ by $\mathbf{D}_1$ and $\mathbf{D}_2$ respectively. Then, the j-th coordinate of this gradient vector is equal to:

$$\text{Trace}\left(\eta^{-4}\boldsymbol{\mathcal{J}}^{-1}(\boldsymbol{\zeta}_{\bar{E}})\mathbf{K}^{-1}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\mathbf{N}^{-1}\left(\frac{\partial}{\partial\zeta_{j,\bar{E}}}(\mathbf{N}(\boldsymbol{\zeta}_{\bar{E}}))\right)\mathbf{N}^{-1}\mathbf{Q}^T\mathbf{Q}\mathbf{P}\right)$$

$$\text{Trace}\left(\eta^{-4}\boldsymbol{\mathcal{J}}^{-1}(\boldsymbol{\zeta}_{\bar{E}})\mathbf{K}^{-1}\boldsymbol{\Theta}_{\bar{E}}\mathbf{P}^T\mathbf{Q}^T\mathbf{Q}\mathbf{N}^{-1}\mathbf{D}_1\mathbf{D}_2\mathbf{N}^{-1}\mathbf{Q}^T\mathbf{Q}\mathbf{P}\right),$$

which completes the proof of our claim.

B. Working version of selection-aware posterior

For the sake of completeness, we provide the probabilistic motivation behind the working selection-aware posterior (12) based on a Laplace approximation for a multivariate Gaussian integral. The next Proposition derives the approximation for $I(\boldsymbol{\beta}_{\bar{E}})$ as an upper bound on the probability of selection. We use $\widehat{\mathbf{B}}_{\bar{E}}$ and $\widehat{\mathbf{B}}^{\text{LASSO}}$ to denote the random variables assuming the instances $\widehat{\boldsymbol{\beta}}_{\bar{E}}$ and $\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}$ respectively and use C to denote a constant free of $\boldsymbol{\beta}_{\bar{E}}$.

PROPOSITION 1: Suppose the joint multivariate Gaussian likelihood for $\widehat{\boldsymbol{\beta}}_{\bar{E}}$ and $\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}$ is proportional to

$$\rho(\widehat{\boldsymbol{\beta}}_{\bar{E}}; \mathbf{K}\boldsymbol{\beta}_{\bar{E}} + \mathbf{l}, \boldsymbol{\Theta}_{\bar{E}}) \cdot \rho(\widehat{\boldsymbol{\beta}}_E^{\text{LASSO}}; \mathbf{P}b + \mathbf{o}, \eta^{-2}\mathbf{Q}^T\mathbf{Q}).$$

For an arbitrary convex, compact region K, we have the following upper bound for $\mathbb{P}\left[\begin{pmatrix}\widehat{\mathbf{B}}_{\bar{E}} & \widehat{\mathbf{B}}^{\text{LASSO}}\end{pmatrix} \in K \mid \boldsymbol{\beta}_{\bar{E}}\right]$:

$$C \cdot \exp\Big(-\inf_{(b,w)\in K}\Big\{\frac{1}{2}(b - \mathbf{K}\boldsymbol{\beta}_{\bar{E}} - \mathbf{l})^T\boldsymbol{\Theta}_{\bar{E}}^{-1}(b - \mathbf{K}\boldsymbol{\beta}_{\bar{E}} - \mathbf{l}) + \frac{1}{2\eta^2}(w - \mathbf{P}b - \mathbf{o})^T\mathbf{Q}^T\mathbf{Q}(w - \mathbf{P}b - \mathbf{o})\Big\}\Big).$$

Proof. Denoting the MGF of the random variables $\widehat{\mathbf{B}}_{\bar{E}}$ and $\widehat{\mathbf{B}}^{\text{LASSO}}$ at $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ as $\Lambda(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ with respect to their Gaussian law, we observe the following:

$$\begin{aligned}
&\log \mathbb{P}\left[\begin{pmatrix}\widehat{\mathbf{B}}_{\bar{E}} & \widehat{\mathbf{B}}^{\text{LASSO}}\end{pmatrix} \in K \mid \boldsymbol{\beta}_{\bar{E}}\right] \\
&\leqslant \log \mathbb{E}\left[\exp\left(\boldsymbol{\eta}_1^T \widehat{\mathbf{B}}_{\bar{E}} + \boldsymbol{\eta}_2^T \widehat{\mathbf{B}}^{\text{LASSO}} - \inf_{(b,w)\in K}\{\boldsymbol{\eta}_1^T b + \boldsymbol{\eta}_2^T w\}\right) \;\middle|\; \boldsymbol{\beta}_{\bar{E}}\right] \\
&= \sup_{(b,w)\in K} -\boldsymbol{\eta}_1^T b - \boldsymbol{\eta}_2^T w + \log \Lambda(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2).
\end{aligned}$$

Since the above inequality holds for any $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$, optimizing over the parameters $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ and using a minimax equality for a convex and compact subset K yields the following bound on the log-selection probability after ignoring constants:

$$\begin{aligned}
&- \sup_{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2} \inf_{(b,w)\in K} \left\{\boldsymbol{\eta}_1^T b + \boldsymbol{\eta}_2^T w - \log \Lambda(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)\right\} \\
&= - \inf_{(b,w)\in K} \sup_{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2} \left\{\boldsymbol{\eta}_1^T b + \boldsymbol{\eta}_2^T w - \log \Lambda(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)\right\} \\
&= - \inf_{(b,w)\in K} \Big\{ \frac{1}{2}(b - \mathbf{K}\boldsymbol{\beta}_{\bar{E}} - \mathbf{l})^T \boldsymbol{\Theta}_{\bar{E}}^{-1} (b - \mathbf{K}\boldsymbol{\beta}_{\bar{E}} - \mathbf{l}) \\
&\qquad\qquad + \frac{1}{2\eta^2}(w - \mathbf{P}b - \mathbf{o})^T \mathbf{Q}^T \mathbf{Q}(w - \mathbf{P}b - \mathbf{o})\Big\}.
\end{aligned}$$

Notice, the selection probability in Theorem 1 is calculated over the region $\{(b, w) : \text{sign}(w) = s_E\}$, which is convex but not compact. We remark however that the approximate posterior obtained by appending a prior to the conditional likelihood and plugging in the bound in Proposition 1 works well under a large enough compact subset K' of the selection region, for all $\boldsymbol{\beta}_{\bar{E}}$, in a bounded set of probability close to 1 under $\pi(\cdot)$. Rigorous asymptotic justification of the approximate selection-aware posterior is based on consistency guarantees aligned along the moderate deviations scale in (Panigrahi and Taylor, 2019); we refer interested readers to this previous work for the asymptotic guarantees behind the Laplace-type approximation.

C. Supplementary details for empirical experiments

C.1 *Simulation analysis*

In this section, we provide comparisons of our methods with other strategies for inference. First, we showcase how our methods recover the target coverage as opposed

to a commonly prescribed single shot approach at selection and inference taken by the Bayesian LASSO (Park and Casella, 2008) under misspecified models. Next, we carry out comparisons of our method with an adaptation of the *post-double-selection* method in Belloni et al. (2014). The post-double-selection method in accounts for the presence of nuisance parameters (effects of confounding variables) by partialing out their effect from both the outcome and treatment variables, and avoids bias from their estimation by the use of auxiliary samples (also called cross-fitting)

We highlight the accuracy of inference achieved with our Bayesian selection-aware pipeline against the Bayesian LASSO (using the *Bayesreg* R package)— a state-of-the-art single shot approach to both selection and inference. We investigate an application of the Bayesian LASSO in two stages: (1) by using all the 1283 gene pathway measurements (the set of explanatory variables) before any selection and by using a Bayesian LASSO prior in matching dimensions; (2) by using 357 gene pathway measurements (the set of imaging-informed explanatory variables) and a corresponding 357 dimensional Bayesian LASSO prior; that is, we use the explanatory variables associated with the intermediary outcomes when we solve the Bayesian LASSO regression using the outcome variable in the primary model. We call the former "Bayes LASSO (full)" and the latter "Bayes LASSO" in our Figure depicting the accuracy of inference.

Under a sparse model with both weak and strong signals of varying amplitudes and random signs, we draw in each round of simulation an outcome from the primary model (Equation 4) where the generating parameter $\boldsymbol{\beta}$ is simulated from a mixture of Laplace distributions, described explicitly in Section 5. Note that the two methods under scrutiny do not use the true generative prior, and therefore provide estimates under potentially misspecified models.

[Figure 1 about here.]

Figure 1 illustrates the comparisons under the four varying regimes of signal-to-noise ratio, which we obtain by varying the scale of one of the mixing Laplace distributions. In contrast to the estimates from the Bayesian LASSO, this depiction

illustrates the robustness of our Bayesian selection-aware inference despite model misspecification. We note that the coverage properties of "Bayes LASSO" and "Bayes LASSO (full)", with and without the dimension reduction from the first step of the integrative pipeline based on the intermediary outcome variables, fall short of the nominal levels (depicted by the horizontal dotted line). Consistent with expectations, "Bayes LASSO" performs worse than the one-shot method on the p predictors, because of ignoring the impact of selection from the first stage. The two-stage filter we apply through our pipeline, first by sifting the signal variables from the noise during selection to obtain a sharp model in terms of support recovery, and second by accounting for the effect of selection during inference, upholds inferential guarantees even under misspecified models.

Turning to the post-double-selection method in Belloni et al. (2014), we construct confidence intervals for the strength of the association between a gene pathway and our clinical outcome by treating all other pathways as potential confounding variables. We use the `hdm` Package in R to implement the post-double-selection approach in our comparisons. The findings of the simulation are summarized in Table 1; our method is called "Proposed" and post-double-selection is highlighted as "PDS". Consistent with our simulations in the main paper, we investigate four different signal regimes (please see details in the main paper). Summaries for the lengths of the interval estimates: averaged length, median length, and 25^{th} and 75^{th} percentiles are reported; the mean run time for the code on a 2.4 GHz, 8-Core, 64 GB Macbook (the same machine is used for both methods) is reported. In a small number of cases ($\sim 0.5\%$), the package `hdm` produced numerically unstable results; we excluded those simulation rounds from the final summaries.

Two key observations stand out in our findings. (i) Our proposed method not only produces shorter intervals (in terms of both mean and median lengths) across all four signal regimes, but also the intervals have smaller variability across all simulations which is evident from the three percentiles (25^{th}, 50^{th}, 75^{th}) in the reported results. For instance, in regime 4, the mean and median length of PDS is longer by $\approx 18\%$

than our method. Furthermore, the interquartile range for lengths produced by PDS is 5.56 as opposed to 3.75 for our proposed method. (ii) PDS is less efficient in terms of computing times than our proposed in simulations that closely resemble the radiogenomic data analysis in terms of problem dimensions. Note, the run time for PDS is almost 45 times longer than our run time on an average over all regimes. For the comparison of run times, we remark that a potential speed up for PDS is possible through parallel computing on multiple nodes. But, we only report the run times of both methods on a single machine for honest comparisons between the two methods when both methods use the same computing resources.

[Table 1 about here.]

C.2 *Radiogenomic feature construction: Pathway scores*

Pathway based methods provide significant benefits by offering interpretability, as gene functions are exerted collectively and may vary based on several factors such as genetic modification, disease state, or environmental stimuli. Using pathways provides an intuitive way and a stable context for assessing the biological activity (Hänzelmann et al., 2013). Pathway scores are computed using the gene-set variation analysis (GSVA), which estimates a value per sample for the variation of pathway activity within an entire gene expression set, using a non-paramteric and unsupervised approach (Hänzelmann et al., 2013). In other words, a pathway score assesses the relative variability of gene expression of the genes in the pathway as compared to expression of genes not in the pathway. We give a brief overview of the analytical procedure for GSVA next.

Let Z be the $p \times n$ matrix of normalized gene expression values corresponding to p genes and n samples ($p \gg n$). Let $G = \{g_1, \ldots, g_m\}$ represent a collection of pathways (also referred to as gene-sets). Each pathway g_k is defined as $g_k \subset \{1, \ldots, p\}$ with $|g_k|$ denoting its cardinality. Let the expression profile for the gene i be given as $z_i = (z_{i1}, \ldots, z_{in})$.

Firstly, in the context of the sample population distribution GSVA evaluates

whether a gene i is highly or lowly expressed in the sample j. To compare distinct expression profiles on the same scale, an expression-level statistic is computed. A non-parametric kernel estimation of the cumulative density function is performed for each z_i using a Gaussian kernel, that is, we compute $\hat{F}_{s_i}(z_{ij}) = \frac{1}{n}\sum_{r=1}^{n}\Phi(\frac{z_{ij}-z_{ir}}{s_i})$. Here s_i is the gene-specific bandwidth parameter controlling the resolution of the kernel estimation. These statistics $\hat{F}_{s_i}(z_{ij})$ are converted to ranks $r_{(i)j}$ for each sample j. The ranks $r_{(i)j}$ are normalized further as $t_{ij} = |\frac{p}{2} - r_{(i)j}|$ so that the tails of the rank distribution are up-weighted while computing the enrichment score. The normalized ranks t_{ij} are used to compute a Kolmogorov-Smirnov (KS) type random walk statistic for $l = 1, \ldots, p$ as

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

Here $I(u_{(i)} \in g_k)$ is an indicator taking the value 1 if the gene corresponding to the rank i expression level statistic belongs to the pathway g_k and τ is the parameter describing the weight of the tail. By identifying if the genes in a pathway are more likely to belong to either tail of the rank distribution, the statistic $\eta_{jk}(l)$ produces a distribution over the genes. The enrichment score for the pathway g_k and the sample j is constructed by converting the corresponding KS-like statistic as $S_{jk} = \max_l(0, \eta_{jk}(l)) - \min_l(0, \eta_{jk}(l))$. S_{jk} has a clear biological interpretation as it emphasizes genes in pathways that are concordantly activated in one direction only, which are either over-expressed or under-expressed relative to the overall population (Hänzelmann et al., 2013). Low enrichment is shown for pathways containing genes strongly acting in both directions. The computations are performed using the GSVA package in R obtained from the Bioconductor package (Gentleman et al., 2004) under the default settings for the choice of parameters.

C.3 *Radiogenomic feature construction: MRI scans and radiomic phenotypes*

We consider four types of MRI sequences which include (i) native (T1), (ii) post-contrast T1-weighted (T1Gd), (iii) T2-weighted (T2), and (iv) T2 fluid attenuated

inversion recovery (FLAIR) volumes. Each of these sequences display different types of tissues with varying contrasts based on the tissue characteristics. Note that the whole brain MRI scans are three dimensional objects and have an array structure. In Figure 2, we show an axial slice from the MRI scan of a LGG subject corresponding to all four imaging sequences with the segmented tumor region indicated by a (red) boundary overlaid on those images. This tumor region is further classified into sub-regions (NC, ED and ET) by GLISTRboost as shown in Figure 2.

[Figure 2 about here.]

Voxel-level features are usually extracted to provide additional insight into the tumor physiology, and have been studied in the context of the progression (or regression) of tumors. Summary statistics such as percentiles, quartiles, skewness, kurtosis etc., are evaluated to represent a region of interest and are used as features/covariates in downstream analysis (Baek et al., 2012; Just, 2014; Song et al., 2013). However, some of their drawbacks include the subjectivity in the choice of number and location of summary features, and limitations in terms of capturing entire information from the histogram of intensity values. To address this, we consider the smoothed density arising from the voxel-level intensity histogram which incorporates granular characteristics of tumor heterogeneity (Saha et al., 2016). The variability in these intensity histograms across the subjects is captured through the scores from a principal component analysis on the space of density functions using a Reimannian-geometric framework. In other words, these principal component scores obtained from the density functions act as the radiomic phenotypes which capture the heterogeneity in the tumor voxels from the MRI scans. We include these details next.

C.4 *Radiogenomic feature construction: Principal Component Scores*

Next, we discuss the construction of the principal component scores when we have probability density functions (pdfs) as data objects corresponding to n samples. Without loss of generality, let us consider densities on $[0, 1]$ and let $\mathcal{F}$ denote the

Banach manifold of such pdfs defined as $\mathcal{F} = \{f : [0,1] \to \mathbb{R}_+ | \int_0^1 f(x)dx = 1\}$. A non-parametric Fisher-Rao Reimannian metric which is invariant to reparameterizations can be defined, however it is computationally challenging to compute the geodesic paths and distances using this metric (Srivastava and Klassen, 2016).

An equivalent representation of the space $\mathcal{F}$ via the square-root transformation (SRT) representation (Bhattacharyya, 1943) simplifies computations. The SRT is defined as a function $h = +\sqrt{f}$ (we omit the + sign hereafter for notational convenience). Also, the inverse mapping is unique and is simply given by $f = h^2$ (Kurtek and Bharath, 2015). Space of SRTs is given by $\mathcal{H} = \{h : [0,1] \to \mathbb{R}^+ | \int_0^1 h(x)^2 dx = 1\}$ and represents the postive orthant of a unit Hilbert sphere (Lang, 2012). The $\mathbb{L}^2$ Riemmanian metric on $\mathcal{H}$ can be defined as $\langle\langle \delta h_1, \delta h_2 \rangle\rangle = \int_0^1 \delta h_1(t)\delta h_2(t)dt$, where $\delta h_1, \delta h_2 \in T_h(\mathcal{H})$ and $T_h(\mathcal{H}) = \{\delta h : [0,1] \to \mathbb{R} | \int_0^1 h(x)\delta h(x)dx = 0\}$. The geodesic paths and lengths can now be analytically computed due to the Reimannian geometry of $\mathcal{H}$ equipped with the $\mathbb{L}^2$ metric. The geodesic distance between $h_1, h_2 \in \mathcal{H}$ is simply given by $d(h_1, h_2) = \theta = \cos^{-1}\big(\int_0^1 h_1(x)h_2(x)dx\big)$.

The geometry of the space of SRTs can be used to define an average (or mean) density corresponding to a sample of density functions. This allows us to efficiently summarize and visualize a sample of densities. The average pdf can be computed using a generalized version of mean on a metric space called the *Karcher* mean (Karcher, 1977). Suppose we have n pdfs $f_1, \ldots, f_n$ and the corresponding SRTs as $h_1, \ldots, h_n$. The sample Karcher mean $\bar{h}$ on $\mathcal{H}$ is defined as the minimizer of the Karcher variance $\rho(\bar{h}) = \sum_{i=1}^{n} d(h_i, \bar{h})^2_{\mathbb{L}^2}$, that is, $\bar{h} = \text{argmin}_{h \in \mathcal{H}} d(h_i, \bar{h})^2_{\mathbb{L}^2}$. Algorithm 1 presents a gradient-based approach to compute the Karcher mean on $\mathcal{H}$ (Dryden and Mardia, 1998).

Note that the Karcher mean of the sample pdfs is an intrinsic average that is computed directly on $\mathcal{H}$ (or equivalently $\mathcal{F}$). Hence we have a mean which is an actual pdf (Karcher mean) and a distance function. Here the inverse exponential map, denoted by $\exp_{h_1}^{-1} : \mathcal{H} \mapsto T_{h_1}(\mathcal{H})$, is given by $\exp_{h_1}^{-1}(h_2) = (\theta/\sin(\theta))(h_2 - h_1\cos(\theta))$.

Algorithm 1 Sample Karcher mean of densities

1: $\bar{h}_0$ (initial estimate for the Karcher mean) $\leftarrow$ any one of the densities in the sample OR the extrinsic average. Set $j \leftarrow 0$ and $\epsilon_1, \epsilon_2 > 0$ be small.
2: For $i = 1, \ldots, n$ compute $u_i = \exp^{-1}_{\bar{h}_j}(h_i)$.
3: Compute the average direction in the tangent space $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$.
4: **if** $||\bar{u}||_{L^2} < \epsilon_1$ **then**
5: **return** $\bar{h}_j$ as the Karcher mean.
6: **else**
7: $\bar{h}_{j+1} = \exp_{\bar{h}_j}(\epsilon_2 \bar{u})$.
8: Set $j \leftarrow j + 1$.
9: Return to step 2.

The exponential map at a point $h_1 \in \mathcal{H}$, denoted by $\exp : T_{h_1}(\mathcal{H}) \mapsto \mathcal{H}$, is defined as $\exp_{h_1}(\delta h) = \cos(\|\delta h\|)h_1 + \sin(\|\delta h\|)(\delta h/\|\delta h\|)$, where $\|\delta h\| = \left(\int_0^1 \delta h(x)^2 dx\right)^{1/2}$.

Under the standard settings, visualizing the space of pdfs intuitively is not straight forward. Principal component analysis (PCA) is an effective method to explore the variability in the pdfs through their primary modes of variation in the data. Note that the tangent space is a vector space (Euclidean), hence PCA can be implemented, as in standard problems. Algorithm 2 describes the computation of PCA on the space generated by the pdfs $f_1, \ldots, f_n$ and the corresponding SRTs $h_1, \ldots, h_n$.

Algorithm 2 PCA on $\mathcal{P}$

1: Compute the Karcher mean of $h_1, \ldots, h_n$ as $\bar{h}$.
2: **for** $i = 1, \ldots, n$ **do**
3: Compute projections ($v_i = \exp^{-1}_{\bar{h}}(h_i)$) of h_i onto $T_{\bar{h}}(\mathcal{H})$.
4: Evaluate sample covariance matrix $K = \frac{1}{n-1}\sum_{i=1}^{n} v_i v_i^\top \in \mathbb{R}^{m \times m}$.
5: Compute the SVD of $K = U\Sigma U^\top$.

Note that the first r columns of U (denoted as $\tilde{U} \in \mathbb{R}^{m \times r}$) span the r-dimensional principal subspace. We can compute the principal coefficients as $X = V\tilde{U}$, where $V^\top = [v_1\ v_2\ \ldots\ v_n] \in \mathbb{R}^{m \times n}$. These principal coefficients X (principal component

scores) act as Euclidean coordinates corresponding to densities f_i and can be used as predictors for downstream modeling.

In Figure 3, we represent the workflow to construct the principal component scores using the tumor intensity values for the T1 MRI sequence. The same workflow is followed for the other three MRI sequences to compute the principal component scores. The number of principal components to include could be chosen using a threshold on the percent variance explained.

[Figure 3 about here.]

C.5 *A direct modeling approach*

Our integrative model in the paper assumes the following sequential flow of information:

$$\text{Genomics} \longrightarrow \text{Imaging Outcomes} \longrightarrow \text{Clinical outcomes.}$$

As indicated by a referee, the above modeling assumption may be prohibitive if there are important genomic variable(s) with a strong impact on the clinical outcome, but is not associated with imaging outcomes (or, is culled out in the regression with the imaging outcomes). To address the above limitation, we explore a direct modeling approach by regressing the clinical outcome against all the 1289 gene pathways through the LASSO in the first step. In the next step, we apply our method of post-selection inference to ascertain uncertainties for the effects of the selected variables.

On the radiogenomic dataset, we find that only two gene pathways— "GNF2 MATK", "MODULE 80"— have significant associations with the clinical outcome after a post-selection correction for their effect sizes. The first of these associations is present in the set of imaging-informed gene pathways (called radiogenomic variables) and is selected in our analysis under Section 6.3. The second significant association is in the complement set of radiogenomic variables. We however find that none of the significant radiogenomic associations reported in Section 6 (with an upstream regression with the imaging outcomes) are discovered in the present analysis. This is along expected lines since we have only 61 samples, but, we begin with a set of

$p = 1289$ potential explanatory variables ($p \sim 20n$) if we eliminate the upstream regression with the imaging outcomes. The reduction in regression dimension when we regress the clinical outcome only against radiogenomic variables, as is done in Section 6, is particularly significant because it brings down $p \sim 5n$. We believe that a significant inflation in the dimension of the regression problem, combined with correlations between some of the gene pathways, might have led to some loss of accuracy and power in terms of support recovery in the current situation. Some of this loss could potentially be salvaged if more sample observations were accrued over time for the presented case study.

References

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

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review* **107,** 261–65.

Dryden, I. and Mardia, K. (1998). *Statistical analysis of shape*. Wiley.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5,** R80.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics* **14,** 7.

Just, N. (2014). Improving tumour heterogeneity mri assessment with histograms. *British journal of cancer* **111,** 2205.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics* **30,** 509–541.

Kurtek, S. and Bharath, K. (2015). Bayesian sensitivity analysis with the fisher–rao metric. *Biometrika* **102,** 601–616.

Lang, S. (2012). *Fundamentals of differential geometry*, volume 191. Springer Science & Business Media.

Panigrahi, S. and Taylor, J. (2019). Approximate selective inference via maximum likelihood. *arXiv preprint:1902.07884* .

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

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

Saha, A., Banerjee, S., Kurtek, S., Narang, S., Lee, J., Rao, G., Martinez, J., Bharath, K., Rao, A. U., and Baladandayuthapani, V. (2016). Demarcate: Density-based magnetic resonance image clustering for assessing tumor heterogeneity in cancer. *NeuroImage: Clinical* **12,** 132–143.

Song, Y. S., Choi, S. H., Park, C.-K., Yi, K. S., Lee, W. J., Yun, T. J., Kim, T. M., Lee, S.-H., Kim, J.-H., Sohn, C.-H., et al. (2013). True progression versus pseudoprogression in the treatment of glioblastomas: a comparison study of normalized cerebral blood volume and apparent diffusion coefficient by histogram analysis. *Korean journal of radiology* **14,** 662–672.

Srivastava, A. and Klassen, E. P. (2016). *Functional and shape data analysis*. Springer.

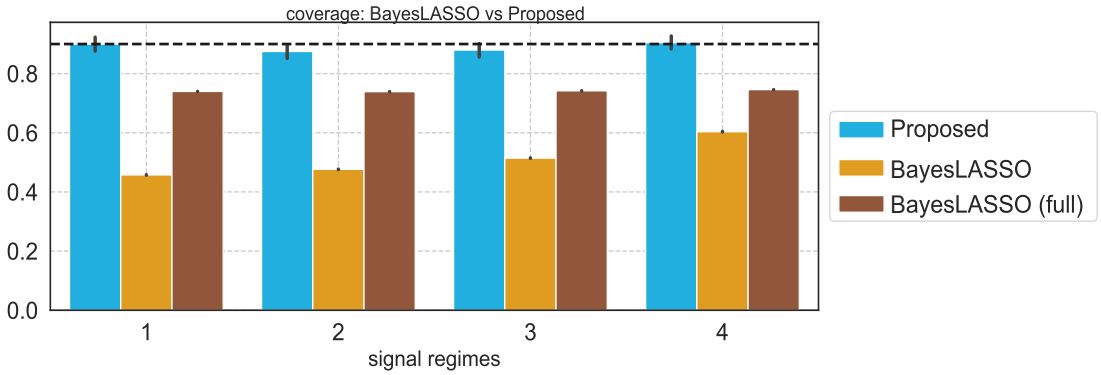
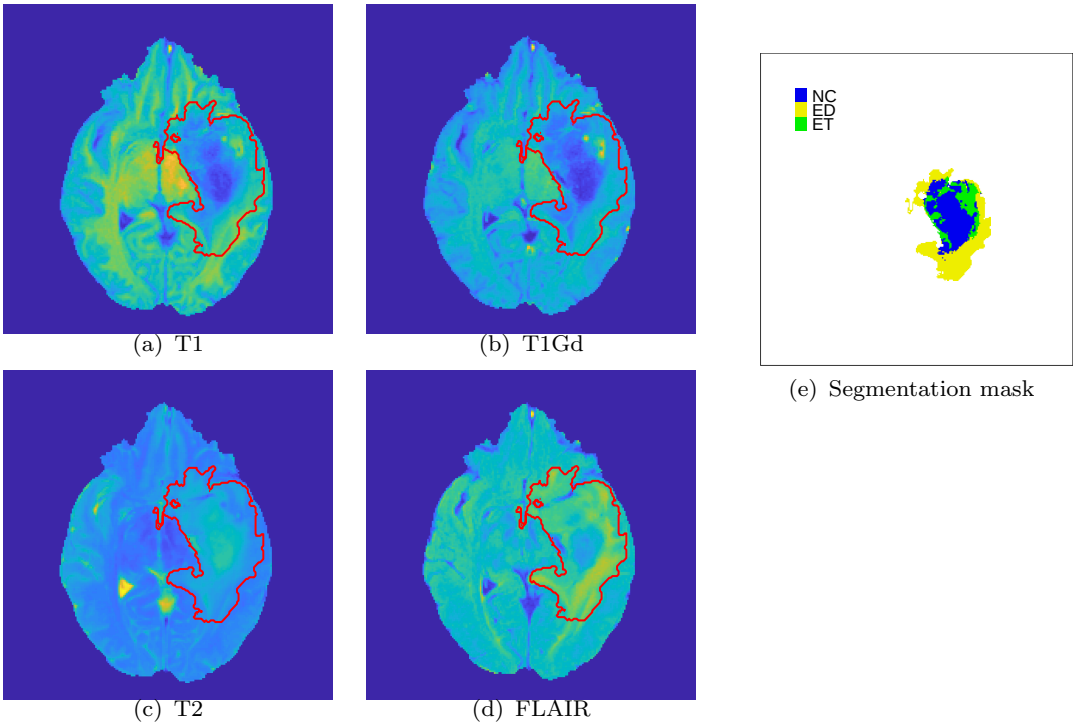


Figure 1. Comparison of inferential accuracy between a selection-aware Bayesian and the Bayesian LASSO.



(a) T1 (b) T1Gd (c) T2 (d) FLAIR (e) Segmentation mask

Figure 2. An axial slice of a brain MRI from four modalities: T1, T1Gd, T2 and FLAIR for a LGG subject. The segmented tumor region is shown with an (red) overlaid boundary. Segmentation mask indicates the necrotic and non-enhancing tumor core (NC), the peritumoral edema (ED) and the enhancing tumor (ET) regions.

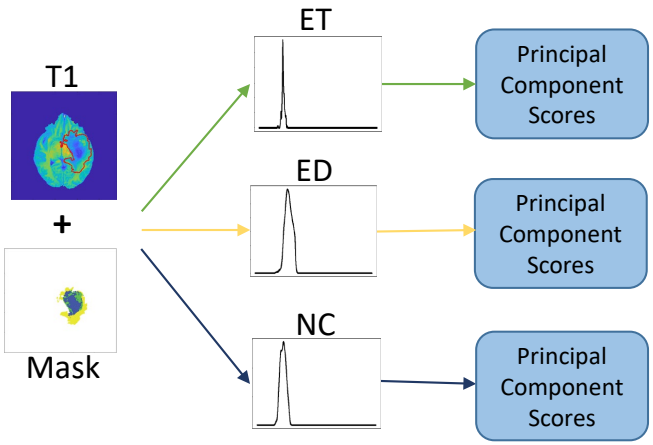


Figure 3. Workflow to obtain the principal component scores from the tumor subregions (ND, ET, and ED) in the T1 MRI scan.

		Lengths				Run Time (in s)
Regime	Method	Mean	25^{th} Percentile	Median	75^{th} Percentile	Mean
1	Proposed	8.43	6.40	7.96	9.61	6.99
1	PDS	9.21	6.21	8.23	10.89	367.71
2	Proposed	8.23	6.48	7.93	9.51	9.96
2	PDS	9.29	6.20	8.26	10.95	396.27
3	Proposed	9.41	6.84	8.27	10.02	9.49
3	PDS	9.84	6.55	8.73	11.67	417.69
4	Proposed	9.23	6.95	8.10	10.70	12.28
4	PDS	10.81	7.19	9.60	12.75	524.79

Table 1
Simulation results under four signal regimes. Proposed dominates PDS by producing shorter interval estimates with smaller variability across all simulations. At the same time, Proposed is computationally more efficient than PDS by orders of magnitude.