An Adaptive version of the Metropolis Adjusted Langengevin Algorithm for Survival prediction in a high dimensional framework

Gabriele Tinè[1], Rosalba Miceli[1]

[1] Biostatistics for Clinical Research Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy

Abstract

The objective is to construct a prognostic index that incorporates radiomic information with the validated prognostic index (Sarculator) provided by the Fondazione IRCCS Istituto Nazionale dei Tumori di Milano. A Bayesian approach was employed, utilising a Weibull model. Vague prior distributions were elicited for the shape parameter, the intercept, and the Sarculator. A multivariate Gaussian prior was elicited for the 2,144 radiomic parameters, incorporating a penalty factor, λ. A total of 100 penalty values were considered. A new, ad hoc adaptive version of the pre-conditioned Metropolis adjusted Langevin algorithm (A-MALA) was proposed for sampling. Bayesian Model Averaging (BMA) was employed to yield a composite of the 100 models. A Bayesian hypothesis test was constructed to evaluate the superiority of the BMA prognostic index relative to the Sarculator. The five-year AUC posterior mean was 0.809, with a 95% credible interval (CI) of (0.768, 0.851). The posterior mean of the C-index was 0.804 (95% CI, 0.764, 0.845) for the BMA, 0.743 (95% CI, 0.713, 0.771) for the best model $\log \lambda = 10.39$ and 0.735 (95% CI, 0.674, 0.761) for the Sarculator. The results suggest that radiomic variables should be included in the model.

Keywords

Bayesian computation, Survival Analysis, Shrinkage Prior in Survival Analysis, Metropolis Adjusted Langevin Algorithm, Adaptive Metropolis, Bayesian Model Averaging, Hypotesis test construction, Omics data.

1. Background

The combination of radiomic variables with clinical variables has been extensively utilized in the literature for the differentiation of benign and malignant lesions and for tumor grading differentiation [1], [2]. In contrast, the application of radiomic variables in constructing prognostic indices remains a relatively unexplored area of research. Among the few studies that address the topic of soft tissue sarcomas (STS), prognosis, and radiomics, Spraker et al. conducted an analysis of T1-weighted MRI sequences [3]. A notable limitation of the study is the relatively small number of radiomic variables considered, with only 30 included. This limitation is not merely a matter of quantity; it also concerns the selection process itself. The rationale behind the selection of this specific subset of variables is not clearly articulated. The exclusion of a more extensive range of radiomic attributes inherent to MRI images increases the likelihood that an inadequate variable set has been selected, which may result in the omission of valuable predictive factors. Furthermore, the method used to select the variables is a cause for concern. Spraker et al. employed a Cox proportional hazards model with LASSO penalization for variable selection. However, LASSO regression does not provide inferential guarantees and can suffer from selection bias due to the uncertainty associated with the selection process itself. The reliability of the selected predictors may be limited in the absence of control for the false discovery rate (FDR). Other studies [4] put forth the proposition of constructing a prognostic index that incorporates radiomic variables through a joint analysis of T1- and T2-weighted MRI sequences. This resulted in the extraction of a

HC@AIxIA 2024: 3rd AIxIA Workshop on Artificial Intelligence For Healthcare

gabriele.tine@istitutotumori.mi.it (G. Tinè); rosalba.miceli@istitutotumori.mi.it (R. Miceli)

total of 1,394 radiomic variables, which appears to be an adequate number for MRI images. Nevertheless, analogous constraints are evident in their study. Moreover, the authors utilized Cox-LASSO regression for variable selection but did not provide inferential guarantees on the selected predictors, as there was no control for the FDR. In a previous study [5], the authors employed machine learning algorithms on computed tomography (CT) images. However, a comprehensive analysis revealed that the FDR control measure was not employed, which may have resulted in an inadequate level of inferential assurance regarding the selected predictors. Additionally, no statistically significant difference was found between the prognostic accuracy of the models that were based exclusively on clinical variables and those that incorporated both clinical and CT-derived radiomic variables. This result could suggest that the CT-derived radiomic variables do not contribute additional prognostic value beyond that of the clinical factors. However, the absence of inferential guarantees concerning the selection process could lead to misleading conclusions. Similarly, [6] extracted 103 radiomic variables from diffusion-weighted imaging (DWI) MRI sequences but faced comparable limitations. They applied Cox-LASSO regression without inferential guarantees on the predictors, and the number of radiomic variables was relatively small. As with the approach taken by [3], no explanation was provided for the choice of variables extracted. It is notable that the aforementioned studies not only exhibit a similar range in sample sizes but also appear to utilise Cox-LASSO regression in a manner that is somewhat unconventional from a statistical perspective. Indeed, the analyses were conducted on an identical dataset in two stages. Initially, variable selection was conducted using Cox-LASSO regression. Subsequently, an unpenalized Cox model was constructed with the selected predictors, enabling the extraction of p-values for each variable. This methodology introduces several biases that must be considered. Firstly, hypothesis tests on the coefficients and their associated p-values are inherently unreliable due to the inherent bias introduced by LASSO, which affects the coefficients. Therefore, the testing of the coefficients in the newly constructed model is also affected by this inherent bias, resulting in the invalidation of the tests performed on the resulting model coefficients [7]. Secondly, the selection process is biased due to the utilization of the same data set for both variable selection and model estimation. This recycling of data can result in overfitting and an underestimation of true variability, which may compromise the generalizability of the results. Furthermore, the absence of FDR control means that the selection lacks inferential guarantees, leaving uncertainty about the reliability of the selected predictors [8]. It seems that the aforementioned studies, which aimed to construct a valid and generalizable prognostic indicator, may have faced some challenges in terms of generalizability and potential biases that could have arisen from the use of the same data for variable selection and model estimation. Although the relatively small sample sizes are to be expected given the rarity of the disease, they nevertheless limit the robustness of the findings. Furthermore, to date, no study has attempted to integrate radiomic variables into a prognostic index constructed from a significantly larger dataset and validated across four distinct patient cohorts, such as the Sarculator. In view of these shortcomings, the aim of the present study is to employ methodologies that overcome the generalizability issues observed in previous studies and to enhance the prognostic accuracy of the Sarculator by integrating radiomic variables. We propose a formal Bayesian test with inferential guarantees to determine whether radiomic features provide meaningful prognostic information beyond that captured by clinical variables alone. This test is constructed with an innovative and unbiased method to ensure rigorous statistical inference and control for the false discovery rate. Furthermore, we introduce an innovative Bayesian algorithm to improve the estimation process. Specifically, we have developed an adaptive version of the Metropolis adjusted Langevin algorithm (MALA) to accelerate the convergence of our Bayesian sampling [9, 10]. This adaptation enhances computational efficiency and allows for more effective exploration of the parameter space, thereby facilitating the generation of more reliable and robust estimates. By focusing exclusively on radiomic data, our approach does not sacrifice degrees of freedom for additional clinical variables; rather, these are encapsulated within the Sarculator, ensuring inferential reliability through the application of proper statistical controls. Furthermore, our methodology avoids the biases associated with unsuitable variable selection techniques.

Dealing with clinical data

The clinical omic data presents several challenges due to the limited sample size (consisting of 91 patients) and the high dimensionality of the feature space (comprising 2145 variables). Consequently, it may prove challenging to conduct an objective inference that incorporates the inherent uncertainty associated with the estimated parameters. Furthermore, the outcome to be predicted is a survival outcome. The Fondazione IRCCS Istituto Nazionale dei Tumori di Milano developed a Cox model on 1,452 patients affected by soft tissue sarcomas, which was validated on three independent cohorts [11], [12].

Given the accuracy and calibration of the clinical prognostic index, it was adopted for the prediction of survival probability in patients with sarcomas. However, the prognostic index (referred to as "Sarculator") considers only clinical variables [13]. The incorporation of radiomic data may enhance the prognostic performance of the index. However, the challenge lies in verifying whether the radiomic information can effectively augment the efficacy of the Sarculator and in developing a novel prognostic index that incorporates both the existing index and the novel radiomic variables. To address these challenges, a Bayesian approach has been selected for both the construction of the new prognostic index and the assessment of its potential improvement.

2. Methods

The Weibull distribution was chosen to model the time-to-event data. Let $\mathbf{X}$ be the standardized matrix of the predictors of dimension $n \times p$, where $n = 91$ and $p = 2145$. Let $\boldsymbol{t} = (t_1, \ldots, t_n)$ be the vector of the observed time of event (or time of censoring) and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ the vector of event or censor indicators, which is equal to 1 if the patient developed the event and 0 otherwise.

Let $T_i \sim \text{Weibull}(\alpha, \epsilon)$ be the time-to-event random variable, where $\alpha > 0$ is the shape parameter and $\gamma_i = \epsilon^{-\alpha} = \exp(\mathbf{x}_i' \boldsymbol{\beta})$. The likelihood model can be written as:

$$\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}; \boldsymbol{\beta}, \alpha) = \prod_{i=1}^{n} \left(\alpha \gamma_i t_i^{\alpha-1}\right)^{\delta_i} \exp\left(-\gamma_i t_i^{\alpha}\right), \tag{1}$$

where $\boldsymbol{\beta}$ is a 2146-dimensional vector: 2144 are the radiomic coefficients, β_0 is the intercept, and β_1 is the coefficient of the linear predictor derived from the Sarculator. For further details on the parametrization adopted, see [14].

In order to make regression over the 2145 variables feasible, it is necessary to elicit a prior on $\boldsymbol{\vartheta}$ that shrinks the coefficients. However, it is not appropriate to shrink all the coefficients since we know that the coefficient β_1 associated with the Sarculator is surely relevant, as demonstrated by validation studies. There is no reason to penalize the intercept either. Hence, β_0 and β_1 should not be penalized. For the other coefficients, there is a need to elicit a prior that shrinks them toward zero. Considering that radiomic variables are highly correlated and usually characterized by low signal and poor informativeness, the data does not match the sparsity hypothesis. Therefore, a normal prior equivalent to a Ridge penalty seems to be the better choice, as discussed in [15].

The parameterization of the Weibull model presented here differs from the more traditional Accelerated Failure Time (AFT) models. In standard AFT models, the logarithm of the time-to-event is typically modeled as a linear function of the covariates, expressed as:

$$\log(T_i) = \mathbf{x}_i' \boldsymbol{\beta} + \sigma W_i, \tag{2}$$

where W_i follows a specific distribution (e.g., standard extreme value, normal). This formulation emphasizes the multiplicative effect of covariates on the survival time, effectively accelerating or decelerating the event process.

In contrast, our parameterization directly models the scale parameter $\gamma_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$ within the Weibull distribution framework. This approach aligns more closely with the proportional hazards paradigm by specifying the hazard function as:

$$\lambda(t|\mathbf{x}_i) = \alpha \gamma_i t^{\alpha - 1}, \tag{3}$$

allowing covariates to have a multiplicative effect on the hazard rate. This distinction is crucial, especially given the high dimensionality and correlation structure of the radiomic predictors in our study. By employing this parameterization, we facilitate the application of shrinkage priors, such as the Ridge-type normal prior, which effectively manages multicollinearity and enhances model interpretability. Additionally, this approach respects the known importance of specific coefficients like β_1 associated with the Sarculator without imposing undue penalization.

The choice between these parameterizations hinges on the underlying assumptions about how covariates influence the survival process. While AFT models are advantageous when the primary interest lies in understanding the acceleration or deceleration of survival times, the proportional hazards-based Weibull model offers flexibility in modeling hazard functions directly, which is beneficial in high-dimensional settings with correlated predictors.

For a more in-depth comparison and methodological details, readers may refer to [14] and standard texts on survival analysis that discuss the nuances between different Weibull parameterizations and their relationship to AFT models.

Prior and Hyperprior elicitation

Regarding α remind that $\alpha > 0$, consequently, a prior on the positive real has to be specified. A lognormal prior seems to be the better choice since it's reparametrization $\xi = \log(\alpha)$ it's a normal distribution and in the calculation of the posterior there is no need to calculate the Jacobian. The hyperparameters of the prior distribution for α were elicited such that the prior mean of α was set to one, reflecting a neutral position regarding the hazard function specification. By setting a high variance, σ_α^2, the log-normal prior becomes effectively non-informative, allowing the parameter to update primarily based on the likelihood. To ensure $\mathbb{E}(\alpha) = 1$, the log-normal hyperparameters are specified as follows: $\mu = -\log(\sigma_\alpha^2 + 1)/2$ and $\sigma^2 = \log(\sigma_\alpha^2 + 1)$. In fact, recall that if $\alpha \sim \text{lognorm}(\mu, \sigma^2)$, then:

$$\mathbb{E}(\alpha) = \exp\left(\mu + \frac{\sigma^2}{2}\right) = 1, \quad \text{Var}(\alpha) = \exp\left(2\mu + \sigma^2\right)\left(\exp\left(\sigma^2\right) - 1\right) = \sigma_\alpha^2.$$

Therefore:

$$\mu = -\frac{\log(\sigma_\alpha^2 + 1)}{2}, \quad \sigma^2 = \log(\sigma_\alpha^2 + 1).$$

Finally, since the density function of the log-normal distribution is given by:

$$f(\alpha) = \frac{1}{\alpha\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(\alpha) - \mathbb{E}(\alpha)^2}{2\sigma^2}\right),$$

and the logarithm of a log-normal is normally distributed with the same parameters, it is possible to reparameterize by setting $\xi = \log(\alpha)$ and considering the prior distribution for ξ, with

$$\xi \sim \mathcal{N}\left(-\frac{\log(\sigma_\alpha^2 + 1)}{2}, \log(\sigma_\alpha^2 + 1)\right). \tag{4}$$

Regarding $\boldsymbol{\beta}$, let σ^2_β be a higher variance, used to imposed a vague normal prior on β_0, β_1 centered in zero. Let λ identified the precision parameter. The prior distribution for $\boldsymbol{\beta}$ is:

$$\boldsymbol{\beta} \sim \mathcal{N}_{p+1}\left[\mathbf{0}, \begin{pmatrix} \sigma^2_{\boldsymbol{\beta}}\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\lambda}\mathbf{I}_{p-1} \end{pmatrix}\right], \tag{5}$$

where $p = 2146$ and $\mathbf{I}$ is the identity matrix. The prior distribution imposed on the $p-1 = 2144$ radiomic parameters is the Bayesian version of the Ridge penalty.

Finally, regarding the precision parameter, λ, it is not uncommon to assign it a Half-Cauchy hyperprior [16]. However, in this case, a sequence of λ values was evaluated to allow for the specification of a model combining coefficients associated with different precision parameter values.

Posterior distribution kernel

Let's denoted with $\boldsymbol{\vartheta} = (\xi, \boldsymbol{\beta})$. Let's also considered 100 values of λ belonging to the interval $[e^5, e^{10.5}]$. The interval has been established so that the lower value generated a identifiable model and the upper one corresponded to a high shrinkage. Assuming independence between ξ and $\boldsymbol{\beta}$, and combining the relations (1), (**??**) and (5), is possible to define the kernel of the λ-th posterior distributions such as:

$$\begin{aligned}\pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) &\propto \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\beta}, \xi)\pi(\xi)\pi(\boldsymbol{\beta}) \\ &\propto \exp\left(\sum_{i=1}^{n} \delta_i\xi + \delta_i\boldsymbol{x}_i'\boldsymbol{\beta} - e^{\boldsymbol{x}_i'\boldsymbol{\beta}}t_i^{e^\xi}\right)\prod_{i=1}^{n} t_i^{\delta_i(e^\xi-1)} \times \\ &\quad\times \exp\left(-\frac{\xi^2}{2\log(\sigma^2_\alpha+1)} - \frac{\xi}{2} - \lambda\sum_{j=2}^{p}\beta_j^2 - \frac{\beta_0^2+\beta_1^2}{2\sigma^2_{\boldsymbol{\beta}}}\right).\end{aligned} \tag{6}$$

The posterior distribution lacks a known, tractable representation that would allow us to leverage (marginal) conjugacy. This renders Gibbs sampling infeasible. It is necessary to adopt a Metropolis algorithm, but in a high-dimensional framework, the standard Metropolis-Hastings is typically ineffective in exploring the posterior distribution efficiently [17]. Therefore we need to move towards more complex algorithm such as the Metropolis Adjusted Langevin algorithm (MALA). The mixing of such an algorithm is contingent upon the covariance matrix. It is therefore crucial to precondition the algorithm to an appropriate covariance matrix [18].

In order to construct an appropriate covariance matrix for each of the 100 proposal distributions, a reasonable strategy is to derive the covariance from the observed information matrix, which is calculated on the log posterior distribution:

$$\begin{aligned}\log\pi(\boldsymbol{\beta}, \xi \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \sum_{i=1}^{n}\Big(\delta_i\xi + \delta_i\boldsymbol{x}_i'\boldsymbol{\beta} - e^{\boldsymbol{x}_i'\boldsymbol{\beta}}t_i^{e^\xi} + \delta_i\left(e^\xi - 1\right)\log t_i\Big) + \\ -\frac{\xi^2}{2\log(\sigma^2_\alpha+1)} - \frac{\xi}{2} - \lambda\sum_{j=2}^{p}\beta_j^2 - \frac{\beta_0^2+\beta_1^2}{2\sigma^2_{\boldsymbol{\beta}}} + c,\end{aligned} \tag{7}$$

where c is a constant term including all the log-additional terms which do not depend from parameters. Now we can derive the pre-conditioning matrix. Formally, called $\mathcal{I}\left(\tilde{\boldsymbol{\vartheta}}\right)$ the observed matrix in $\tilde{\boldsymbol{\vartheta}}$, where $\tilde{\boldsymbol{\vartheta}}$ is the initialization vector for the generic proposal distribution, the observed covariance matrix Σ for the pre-conditioning can be computed as follow [19]:

$$\Sigma \approx \mathcal{I}^{-1}\left(\tilde{\boldsymbol{\vartheta}}\right) = \left[-\mathcal{H}\left(\tilde{\boldsymbol{\vartheta}}\right)\right]^{-1}, \tag{8}$$

where $\mathcal{H}\left(\tilde{\boldsymbol{\vartheta}}\right)$ is the hessian of the log posterior evaluated in $\tilde{\boldsymbol{\vartheta}}$. Hence, for a sequence of optimal initialization vectors $\left(\tilde{\boldsymbol{\vartheta}}_0, \ldots \tilde{\boldsymbol{\vartheta}}_\lambda \ldots \tilde{\boldsymbol{\vartheta}}_{100}\right)$ it is possible to derive the 100 pre-conditioning matrices.

In order to establish the initialization vectors for each proposal distribution $(\tilde{\boldsymbol{\vartheta}}_0, \dots \tilde{\boldsymbol{\vartheta}}_\lambda \dots \tilde{\boldsymbol{\vartheta}}_{100})$ it is possible to maximize the log posterior so that, for the generic $\tilde{\boldsymbol{\theta}}_\lambda$ we have:

$$\tilde{\boldsymbol{\vartheta}}_\lambda = \underset{\boldsymbol{\vartheta}_\lambda}{\arg\max}\ \log \pi\left(\boldsymbol{\vartheta}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right). \tag{9}$$

From equation (9), it is possible to derive the generic initialization vector $\tilde{\boldsymbol{\vartheta}}_\lambda$ for the initialization of the λ-th pre-conditioning matrix $\mathcal{I}^{-1}\left(\tilde{\boldsymbol{\vartheta}}_\lambda\right)$ for the λ-th proposal distribution as follows:

$$\begin{cases}
\frac{\partial \log \pi(\cdot|\mathbf{X},\mathbf{t},\boldsymbol{\delta})}{\partial \xi} = 0 & \Rightarrow \sum_{i=1}^{n} \left[\delta_i + \log t_i \left(e^{\xi \delta_i} - e^{\boldsymbol{x}_i' \boldsymbol{\beta} + \xi} t_i^{e^{\xi}}\right)\right] - \frac{\xi}{\log(\sigma_\alpha^2 + 1)} - \frac{1}{2} = 0 \\
\frac{\partial \log \pi(\cdot|\mathbf{X},\mathbf{t},\boldsymbol{\delta})}{\partial \beta_0} = 0 & \Rightarrow \sum_{i=1}^{n} \left(\delta_i x_{i0} - x_{i0} e^{\boldsymbol{x}_i' \boldsymbol{\beta}} t_i^{e^{\xi}}\right) - \frac{\beta_0}{\sigma_{\beta_0}^2} = 0 \\
\frac{\partial \log \pi(\cdot|\mathbf{X},\mathbf{t},\boldsymbol{\delta})}{\partial \beta_1} = 0 & \Rightarrow \sum_{i=1}^{n} \left(\delta_i x_{i1} - x_{i1} e^{\boldsymbol{x}_i' \boldsymbol{\beta}} t_i^{e^{\xi}}\right) - \frac{\beta_1}{\sigma_{\beta_1}^2} = 0 \\
\frac{\partial \log \pi(\cdot|\mathbf{X},\mathbf{t},\boldsymbol{\delta})}{\partial \beta_2} = 0 & \Rightarrow \sum_{i=1}^{n} \left(\delta_i x_{i2} - x_{i2} e^{\boldsymbol{x}_i' \boldsymbol{\beta}} t_i^{e^{\xi}}\right) - 2\lambda \beta_2 = 0 \\
 & \vdots \\
\frac{\partial \log \pi(\cdot|\mathbf{X},\mathbf{t},\boldsymbol{\delta})}{\partial \beta_j} = 0 & \Rightarrow \sum_{i=1}^{n} \left(\delta_i x_{ij} - x_{ij} e^{\boldsymbol{x}_i' \boldsymbol{\beta}} t_i^{e^{\xi}}\right) - 2\lambda \beta_j = 0 \\
 & \vdots \\
\frac{\partial \log \pi(\cdot|\mathbf{X},\mathbf{t},\boldsymbol{\delta})}{\partial \beta_p} = 0 & \Rightarrow \sum_{i=1}^{n} \left(\delta_i x_{ip} - x_{ip} e^{\boldsymbol{x}_i' \boldsymbol{\beta}} t_i^{e^{\xi}}\right) - 2\lambda \beta_p = 0
\end{cases}$$

The system cannot be solved analytically, so it is necessary to use a numerical approximation for which an adaptive nonlinear least squares algorithm was chosen [20]. The procedure was repeated over a sequence of 100 distinct values of λ, and the vectors $\hat{\boldsymbol{\vartheta}}$ minimizing $-\log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \mathbf{t}, \boldsymbol{\delta})$ were selected as initialization values for the sampling algorithm .

A-MALA pre-conditioned

Given that we decided to use the MALA algorithm, the structure of proposal distribution for the generic $\boldsymbol{\vartheta}_\lambda$ can be written as follows:

$$\boldsymbol{\vartheta}_\lambda^* \mid \boldsymbol{\vartheta}_\lambda \sim \mathcal{N}_{p+2}\left(\boldsymbol{\vartheta}_\lambda + \frac{\varepsilon^2}{2(p+2)^{\frac{1}{3}}} \mathcal{I}^{-1}(\tilde{\boldsymbol{\vartheta}}_\lambda) \nabla \log \pi(\boldsymbol{\vartheta}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}_\lambda = \tilde{\boldsymbol{\vartheta}}_\lambda}, \frac{\varepsilon^2}{(p+2)^{\frac{1}{3}}} \mathcal{I}^{-1}(\tilde{\boldsymbol{\vartheta}}_\lambda)\right),$$

where $\nabla \log \pi(\boldsymbol{\vartheta}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$ is the gradient of the log posterior, which incorporates the information of the structure of the posterior distribution; the vector $\boldsymbol{\vartheta}_\lambda = \tilde{\boldsymbol{\vartheta}}_\lambda$ at the first iteration; the quantity ε is the step-size parameter, which regulates the entity of the jumps. Incorporating the structure of the posterior at each iteration through the gradient of the lo posterior in the proposal distribution helps the proposal to generate candidate from region with higher densities and, consequently, with more probability of being accepted [21]. In order to make more computational efficient generating from the proposal without the needing of inverting the covariance matrix at each iteration we used the spectral decomposition such that:

$$\mathcal{I}^{-1}(\tilde{\boldsymbol{\vartheta}}_\lambda) = \mathbf{V}_\lambda \mathbf{D}_\lambda^{-1/2} \mathbf{D}_\lambda^{-1/2} \mathbf{V}'{}_\lambda \tag{10}$$

where $\mathbf{V}_\lambda$ represents the standardized eigenvectors and $\mathbf{D}_\lambda$ represents the diagonal matrix of the eigenvalues of Σ_λ. Hence, setting $\mathbf{A}_\lambda = \mathbf{V}_\lambda \mathbf{D}_\lambda^{-1/2}$, such that $\mathcal{I}^{-1}(\tilde{\boldsymbol{\vartheta}}_\lambda) = \mathbf{A}_\lambda \mathbf{A}_\lambda'$ we can generate the new candidate from the following equation:

$$\boldsymbol{\vartheta}_\lambda^{*(j)} = \boldsymbol{\vartheta}_\lambda^{(j-1)} + \frac{\varepsilon^2}{(p+2)^{\frac{1}{3}}} \nabla \log \pi(\boldsymbol{\vartheta}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}_\lambda = \boldsymbol{\vartheta}_\lambda^{(j-1)}} + \frac{\varepsilon}{(p+2)^{\frac{1}{6}}} \mathbf{A}\mathbf{Z}, \tag{11}$$

where $\mathbf{Z} \sim \mathcal{N}_{p+2}(\mathbf{0}, \mathbf{I}_{p+2})$. Note that using spectral decomposition we only need to deal with $\mathbf{D}_\lambda$ at each iteration, instead of Σ, which is computationally more efficient to manipulate.

Moreover, for computational efficiency, we implemented Singular Value Decomposition (SVD), however since the pre-conditioning matrix is symmetric and squared, SVD and Spectral Decomposition are equivalent.
Algorithm 1 reported the pseudocode used to analytically compute the Hessian in the optimum point. Furthermore, the entire process for the construction of the matrices used to efficiently generate from the proposal distribution is reported.

Algorithm 1: Proposal distributions initialization

```
Analytical derivation of Hessian matrix and optimal point evaluation $-\mathcal{H}\Big|_{\boldsymbol{\mu}=\tilde{\boldsymbol{\mu}}} = \mathcal{I}(\tilde{\boldsymbol{\mu}})$
Function Amatrix($\boldsymbol{\vartheta}$,$\sigma^2_\alpha$, $\sigma^2_\beta$, $l$, $\mathbf{X}$, $\boldsymbol{t}$, $\boldsymbol{\delta}$):
    $H \leftarrow$ empty matrix $(p+2) \times (p+2)$
    $H[1,1] \leftarrow \sum_{i=1}^{n} \left[ e^{\boldsymbol{x}_i' \tilde{\boldsymbol{\vartheta}}_{-0}} \log t_i \left( e_0^{\tilde{\vartheta}} t_i^{e_0^{\tilde{\vartheta}}} + e^{2\tilde{\vartheta}_0} t_i^{e^{\tilde{\vartheta}_0}} \log t_i \right) - e^{\tilde{\vartheta}_0} \delta_i \log t_i \right] + \frac{1}{\log(\sigma^2_\alpha+1)}$
    for $j \leftarrow 2$ to $p+2$ do
        $H[2:(p+1),1] \leftarrow \sum_{i=1}^{n} x_{ij} e_{-0}^{\boldsymbol{x}_i' \tilde{\boldsymbol{\vartheta}}} t_i^{e^{\tilde{\vartheta}_0}} \log t_i$
    end
    for $j \leftarrow 2$ to $p+2$ do
        $\mathrm{diag}(H)[j] \leftarrow \sum_{i=1}^{n} \left( x_{ij}^2 e_{-0}^{\boldsymbol{x}_i' \tilde{\boldsymbol{\vartheta}}} t_i^{e^{\tilde{\vartheta}_0}} \right) + 2f(\tilde{\vartheta}_{-0,j})$
    end
    for $j \leftarrow 2$ to $p+2$ do
        for $k \leftarrow 1$ to $j-1$ do
            $H[j,k] \leftarrow \sum_{i=1}^{n} x_{ij} x_{ik} e_{-0}^{\boldsymbol{x}_i' \tilde{\boldsymbol{\vartheta}}} t_i^{e^{\tilde{\vartheta}_0}}$
        end
    end

    $H_{\boldsymbol{\vartheta}}$ is obtained by symmetry completion of the upper triangular matrix
    $\mathbf{V} \leftarrow \mathrm{SVD}(H_{\boldsymbol{\vartheta}})[u]$  Normalized eigenvectors matrix
    $\mathbf{d} \leftarrow \mathrm{SVD}(H_{\boldsymbol{\vartheta}})[\delta]$  Singular value vector
    $\mathbf{D} = \mathrm{diag}\left(1/\sqrt{\mathbf{d}}\right); \tilde{\circ} \leftarrow \frac{2.38^2}{p+2} \mathbf{V}\mathbf{D}^2\mathbf{V}'$
    Matrix used to efficiently generate from the proposal distribution
    $\mathbf{A} \leftarrow \frac{2.38}{\sqrt{p+2}} \mathbf{V}\mathbf{D}$
    return A
End Function
```

The step-size ε is scaled so that it matches the optimal step that maximizes the diffusion (being the speed of diffusion related to asymptotic variance) and has to be tuned [22]. The strategy implemented was adapting the step-size each 50 iterations so that the optimal acceptance rate of the MALA, which is 57.4% was reached [22]. Besides, in order not to compromise the convergence to the stationary distribution, which is guaranteed by the ergodic theorem, we adapted the step-size according to the assumption of diminishing adaptation [23], [24]. Eventually, we adapted the step-size parameter only within the burn in period, making the number of bur in iterations variable between different λ according to the stabilization of the step-size parameter. In particular, the adaptive strategy we adopted was the following: let $\iota = 0.054$ identifying the tolerance parameter, the stopping rule for the burn in period requires the satisfaction of the following conditions:

1. the *burn-in* iterations are higher than 40000;
2. ε is constant for 500 iterations;
3. if conditions 1 e 2 are not satisfied the *burn-in* is stopped at 150000 iterations.

The criteria to assess whether, every 50 iterations, the burn in had to be stopped was the following: let r be the r-th burn in iteration and let $\varepsilon^{(j-1)}$ the ε value deriving from the previous updating step, then

$$\begin{cases} \varepsilon^{(j)} = \varepsilon^{(j-1)} + \min\left(0.01, \frac{1}{\sqrt{r_j}}\right) & \text{if} \quad \varepsilon^{(j)} \notin [0.574, 0.547 + \iota] \\ \varepsilon^{(j)} = \varepsilon^{(j-1)} - \min\left(0.01, \frac{1}{\sqrt{r_j}}\right) & \text{if} \quad \varepsilon^{(j)} \notin [0.547 - \iota, 0.547) \\ \varepsilon^{(j)} = \varepsilon^{(j-1)} & \text{if} \quad \varepsilon^{(j)} \in [0.547 - \iota, 0.547 + \iota]. \end{cases} \tag{12}$$

We introduce an adaptive versione of the MALA to efficiently tune the step-size; we called it Adaptive-MALA (A-MALA). The detailed construction of the A-MALA algorithm we proposed is reported as pseudocode in Algorithm 2.

Algorithm 2: Pre-conditioned A-MALA with *tuning* step-size

```
Function MALA($burnin, R, \tau, \boldsymbol{\vartheta}, \sigma^2_\alpha, \sigma^2_\beta, \boldsymbol{\lambda}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, bat, target = 0.574, tolerance = 0.054$):
    out ← array: 100 matricies: rows= $R/\tau$, coulumns = $p + 2$. $p$ number of features; $\tau$ thinning period
    for $i \leftarrow 1$ to length($\boldsymbol{\lambda}$) do
        $\boldsymbol{\vartheta} = betamat[i,]$
        $\mathbf{A} \leftarrow \mathtt{Amatrix}(\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha, \sigma^2_\beta, l = \lambda_i)$
        $\mathbf{S} \leftarrow \mathbf{AA}'$ ; $\mathbf{S} \leftarrow \mathbf{S}^{-1}$
        $logp \leftarrow \mathtt{logposterior}(\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha, \sigma^2_\beta, l = \lambda_i)$
        $lgrad \leftarrow \mathbf{S}\ \mathtt{logradient}(\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha, \sigma^2_\beta, l = \lambda_i)$
        $accepted \leftarrow 0$ ; $batch \leftarrow 1$ ;
        $index \leftarrow 0$ ; $\varepsilon \leftarrow 1$ ; $j \leftarrow 0$ ; $r \leftarrow 0$
        while $r \leqslant burnin + R$ do
            $r \leftarrow r + 1$
            if $batch = bat$ & $r < burnin + 1$ then
                $j \leftarrow j + 1$
                if $accepted/bat > target + tolerance$ then
                    $\varepsilon \leftarrow \varepsilon + \min\left(0.01, \sqrt{\frac{1}{r}}\right)$
                end
                if $accepted/bat < target - tolerance$ then
                    $\varepsilon \leftarrow \varepsilon - \min\left(0.01, \sqrt{\frac{1}{r}}\right)$
                end
                $\mathtt{adaptivemonitoring}[j,i] \leftarrow \varepsilon$ ; $batch \leftarrow 0$ ; $accepted \leftarrow 0$
                if $r \geqslant 40000$ & $\mathtt{all}(\varepsilon = \mathtt{adaptivemonitoring}[(j - r + 1) : j, i])$= TRUE then
                    $r \leftarrow burnin$
                end
            end
            $batch \leftarrow batch + 1$
            $\boldsymbol{\vartheta}^* \leftarrow \boldsymbol{\vartheta} + \frac{\varepsilon^2}{2(p+2)^{1/3}} lgrad \sqrt{\frac{\varepsilon^2}{2(p+2)^{1/3}}} \mathbf{AZ}$
            $logp_{new} \leftarrow \mathtt{logposterior}(\boldsymbol{\vartheta}^*, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha = 16, \sigma^2_\beta = 600, l = \lambda_i)$
            $lgrad_{new} \leftarrow \mathbf{S}\ \mathtt{logradient}(\boldsymbol{\vartheta}^*, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha = 16, \sigma^2_\beta = 600, l = \lambda_i)$
            $diff_{old} \leftarrow \boldsymbol{\vartheta} - \boldsymbol{\vartheta}^* - \frac{\varepsilon^2}{2(p+2)^{1/3}} lgrad_{new}$
            $diff_{new} \leftarrow \boldsymbol{\vartheta}^* - \boldsymbol{\vartheta} - \frac{\varepsilon^2}{2(p+2)^{1/3}} lgrad$
            $q_{old} \leftarrow diff_{old} \mathbf{S}^{-1} diff_{old} \frac{(p+2)^{1/3}}{\varepsilon^2}$
            $q_{new} \leftarrow diff_{new} \mathbf{S}^{-1} diff_{new} \frac{(p+2)^{1/3}}{\varepsilon^2}$
            $\alpha \leftarrow \min\{1, \exp(logp_{new} - logp + q_{old} - q_{new})\}$
            $u \leftarrow u \sim U(0,1)$
            if $u < \alpha$ then
                $logp \leftarrow logp_{new}$
                $lgrad \leftarrow lgrad_{new}$
                $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*$    value accepted
                $accepted \leftarrow accepted + 1$    update accepted values counter
            end
            if $r > burnin$ & $r \in sequenza$ then
                $index \leftarrow index + 1$
                $out[index,,i] \leftarrow \boldsymbol{\vartheta}$
            end
        end
    end
    return $out$
End Function
```

Bayesian Model Averaging

When sampling stage was succesfully completed $\forall \lambda$ using the A-MALA with 75000 iterations of sampling and a thinning period of 15, we used the Watanable Akaike Criterion Information (WAIC) [25], to select the best model and to build a new model by applying the Bayesian Model Averaging (BMA) [26]. We retained that BMA was more appropriate than just select the best model since selecting one model which has a low posterior probability

(5%) introduces a selection bias since the uncertainty of the selection process is not taken into account [27], [28]. In order to take into account the selection uncertainty seems better to use a mixture of all models, where each of them is weighted with its posterior probability. Formally, let Δ be a quantity of inferential interest. Since we do not have any prior information about the probability of each model $\mathcal{M}_\lambda$ let elicit a uniform prior distribution on each model: $\mathcal{M}_\lambda \sim U(0,1)\ \forall\lambda$. Then we can calculate BMA estimate of Δ as:

$$\tilde{\Delta}_{\text{BMA}} = \frac{1}{S}\sum_{s=1}^{S}\sum_{\lambda=1}^{\Lambda}\Delta_\lambda^s \pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}), \tag{13}$$

where $\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$ represents the posterior probability for the λ-th model.

Approximation of the posterior probability model

In equation (13) the term $\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$ has to be calculated as the ratio of the marginal likelihood of the λ-th model and the sum of the all marginal likelihoods:

$$\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \frac{\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)}{\sum_\lambda \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)}. \tag{14}$$

However, these quantity are unknown ad have to be estimated. We could approximate it using the relationship between the marginal likelihood and the Bayesian Information Criteria (BIC), as [29]:

$$\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \frac{\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)}{\sum_\lambda \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)} \approx \frac{\exp\left(-\frac{1}{2}\text{BIC}_\lambda\right)}{\sum_{\lambda=1}^{\Lambda}\exp\left(-\frac{1}{2}\text{BIC}_\lambda\right)}. \tag{15}$$

Nevertheless, we prefer to use the relation (15) replacing the BIC with the WAIC. WAIC is indeed a fully Bayesian criteria which measures how well the model will perform on new data. Indeed, WAIC approximate the Leave One Out Cross Validation (LOO-CV) [30], [27], [31], [32]. In this way, the posterior probability is related to the model's ability to fit new data. So we approximated (14) with:

$$\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \frac{\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)}{\sum_\lambda \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)} \approx \frac{\exp\left(-\frac{1}{2}\text{WAIC}_\lambda\right)}{\sum_{\lambda=1}^{\Lambda}\exp\left(-\frac{1}{2}\text{WAIC}_\lambda\right)}. \tag{16}$$

Note that the approximation (14) has already been introduced also for other information criteria, such as AIC [33].

Hypotesis test

We compared the best model with the BMA but we also compared the BMA with the Sarculator. To do so, we use exactly the same algorithm to sample the values of the Sarculator excluding the radiomic variables. Called $\mathcal{M}_0$ the Sarculator model, we build a Bayesian in order to evaluate whether the radiomic variables added prognostic power. The test we build can be written as follows:

$$H_0 : \mathcal{M}_0 \text{ is sufficient } vs \quad H_1 : \mathcal{M}_\lambda \text{ is better than } \mathcal{M}_0, \tag{17}$$

where $\pi(\mathcal{M}_0) = 0.5$ and $\pi(\mathcal{M}_\lambda) = 0.005$. If $\pi(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) < 0.5$ then H_0 is rejected. Note that we set the prior probability of the Sarculator much higher than the prior probability of the others model due to the fact that the Sarculator is a validated prognostic index. Formally, the test we built can be specified as:

$$H_0 : \pi(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) \geqslant 0.5 \mid \pi(\mathcal{M}_0) = 0.5 \quad vs \quad H_1 : \pi(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) < 0.5 \mid \pi(\mathcal{M}_0) = 0.5. \tag{18}$$

Combining the equation (14) and the approximation (16) the posterior probability model of $\mathcal{M}_0$ can be computed as:

$$
\begin{aligned}
\pi\left(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right) &= \frac{\mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_0\right) \pi\left(\mathcal{M}_0\right)}{\mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_0\right) \pi\left(\mathcal{M}_0\right) + \sum_{\lambda}^{\Lambda} \mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda\right) \pi\left(\mathcal{M}_\lambda\right)} \\
&\approx \frac{\exp\left(-\frac{1}{2}\mathrm{WAIC}_0\right) \times \frac{1}{2}}{\exp\left(-\frac{1}{2}\mathrm{WAIC}_0\right) \times \frac{1}{2} + \sum_{\lambda}^{\Lambda} \exp\left(-\frac{1}{2}\mathrm{WAIC}_\lambda\right) \times \frac{1}{2\Lambda}}.
\end{aligned}
\tag{19}
$$

3. Results

A total of 91 eSTS patients were included in the study. Demographic and tumor characteristics are detailed in Table 1. The median age of the cohort was 59 years (interquartile range [IQR]: 46, 71), and the median tumor diameter was 8.0 mm (IQR: 5.0, 10.8). The median follow-up period was approximately 55.8 months (IQR: 44.8, 72.9), during which 21 patients (23.1%) succumbed to tumor-related causes.

Table 1
Patients' and Tumor Characteristics

Variable	**N1 = 91**
Age (years)	59 (46, 71)
Diameter (mm)	8.0 (5.0, 10.8)
Observation Time (months)	55.8 (44.8, 72.9)
Grading	
G1	22 (24.2%)
G2	21 (23.1%)
G3	48 (52.7%)
Histology	
Leiomyosarcoma	8 (8.8%)
Dedifferentiated/Pleomorphic Liposarcoma	8 (8.8%)
Myxoid Liposarcoma	17 (18.7%)
Malignant Peripheral Nerve Sheath Tumor	3 (3.3%)
Myxofibrosarcoma	20 (22.0%)
Synovial Sarcoma	4 (4.4%)
Undifferentiated Pleomorphic Sarcoma	24 (26.4%)
Vascular Sarcoma	1 (1.1%)
Others	6 (6.5%)

The results of the Bayesian analysis indicate that the incorporation of radiomic variables into the prognostic model is justified. The Bayesian Model Averaging (BMA) yielded a posterior mean Area Under the Curve (AUC) at five years of 0.809, with a 95% Credible Interval (CI) of (0.768, 0.851). This indicates that the model demonstrates a robust capacity to discriminate over a five-year period. The posterior mean Brier score at five years was 0.277, with a 95% CI of (0.257, 0.304), indicating that the model predictions exhibited acceptable calibration. When evaluated over the entire study period, the posterior mean Brier Score was 0.316, with a 95% CI of (0.291, 0.346). Furthermore, the posterior mean Concordance Index (C-Index) was 0.804, with a 95% CI of (0.764, 0.845), which provides additional evidence in support of the model's predictive accuracy. The posterior mean of the coefficient associated with the previously validated prognostic index (Sarculator) was 1.008, with a 95% CI of (0.989, 1.036). This corresponds to a hazard ratio (HR) of 2.739, indicating that higher scores on the prognostic index are associated with a significantly increased risk of adverse outcomes. The posterior mean of the shape parameter of the Weibull distribution was estimated at 0.963, which suggests a near-constant hazard over time. It is notable that all posterior estimates of the radiomic parameters were close to zero due to the penalty imposed by the prior distribution. This is to be expected due to the model's ability to shrink insignificant coefficients towards zero, mitigating the risk of overfitting without introducing a selection bias. The BMA shows wider distributions and wider credible sets than the best model, as it takes into account the uncertainty associated with the selection process. The posterior mean of the AUC at 5 years and C-index is better for the BMA. With regard to calibration, the Brier Scores of the BMA and the best model are similar (see Tabel 2). The results show that the BMA performs better than the best model.

Table 2
Performance comparison: BMA vs best model

Metric	BMA (95% CI)		Best model ($\log\lambda = 10.39$) (95% CI)		Sarculator (95% CI)	
AUC at 5 years	0.809	(0.768, 0.851)	0.772	(0.741, 0.797)	0.758	(0.712, 0.771)
Brier Score at 5 years	0.277	(0.257, 0.304)	0.274	(0.260, 0.289)	0.283	(0.263, 0.297)
Brier Score overall	0.316	(0.291, 0.346)	0.313	(0.297, 0.329)	0.324	(0.308, 0.340)
C-index	0.804	(0.764, 0.845)	0.743	(0.713, 0.771)	0.740	(0.709, 0.767)

CI: Credible Interval

The Bayesian test implemented gave the following result:

$$\pi(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) \approx 0.0049.$$

Since $\pi(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) < 0.5$, the evidence in favour of H_0 is very low, consequently we reject H_0, propending for $H_1 : \exists\lambda \mid \mathcal{M}_\lambda \quad$ is better than $\quad \mathcal{M}_0$. Consequently, radiomic variables should be included in a prognostic index to increase the prognostic accuracy.

4. Discussion

This study explores aspects of statistics and computational efficiently sampling algorithms rarely fully developed within a single framework. Specifically, high variable dimensionality (omics data) and low sample size are common in today clinical studies.

Typically, high-dimensional data problems assume a large number of observations, even when $p \gg n$. While this increases computational demands, the ample sample size allows for standard approaches like splitting data into training and test sets. The training set is used for variable selection—potentially using cross-validation—and the test set for inference. This works because the sample size overrepresents the population, ensuring the subsample used for selection contains all necessary information.

However, with small sample sizes, dividing data into subsets can result in samples that no longer represent the population, leading to biased and suboptimal variable selection, in our case, for example, we have only one patient affected by Vascular Sarcomas. In such cases, the classical data-splitting paradigm fails to provide inferential guarantees. This issue is often overlooked, with methods like LASSO applied despite lacking inferential assurances. Moreover, using LASSO on the same data for both selection and inference introduces bias due to selection and the absence of inferential guarantees. In this study, we developed a method that, considering the data type, sample size, and number of variables, predicts the OS without performing variable selection, thus avoiding associated bias. An ad hoc model was constructed by eliciting prior distributions, which did not penalize the Sarculator. We incorporated prior knowledge of relevant variables, applying penalization only to radiomic variables. A Bayesian approach is inherently suitable for ensuring inferential guarantees, which are assurances that the conclusions drawn from a statistical model are both reliable and valid, particularly concerning parameter estimation, prediction, and inference about data relationships. Bayesian inference provides a robust framework for accounting for uncertainty through probability distributions, incorporating prior knowledge, and updating it with observed data. This allows for the generation of credible intervals for parameter estimates and naturally balances model complexity with data support, thereby maintaining inferential guarantees even in complex, high-dimensional feature spaces or when sample sizes are limited. Moreover, Bayesian credible intervals offer direct probabilistic interpretations of parameter uncertainty, which is especially advantageous in sparse data contexts compared to the frequentist freamwork. The inherent nature of the Bayesian approach, incorporating uncertainty through prior and posterior distributions, mitigates the risk of overestimation when compared to frequentist methods. The application of Bayesian Model Averaging (BMA) further addresses uncertainty by considering less probable models, thereby reducing overestimation risk.

To specify suitable prior distributions and initialize proposal distributions based on solid theory, we built our own A-MALA within a MCMC framework, implemented ad hoc for optimal performance. While this introduced theoretical and practical complexities, it allowed us to thoroughly explore each relevant step within the Bayesian framework without relying on inflexible pre-existing algorithms. The limited literature on Bayesian methods for survival data with high dimensionality and low sample sizes necessitated a detailed analysis of theoretical options, involving significant effort to construct a robust approach.

Given the high dimensionality of the data, it is reasonable to question why the Ridge prior was chosen among available shrinkage priors. Addressing this requires understanding the application domain, data characteristics, and Bayesian shrinkage mechanisms. Morever, using a Bayesian Ridge penalty was a choice that we made after a comprehensive theoretical considerations. In fact several Bayesian Variable Selection approaches could be adopted however, the alternatives does not adapt properly to the nature od the data. The Laplace distribution, a Bayesian analogue of LASSO regression, could effectively set coefficients to zero in the frequentist context. However, this property does not hold in Bayesian settings when considering the posterior mean. In high-dimensional Bayesian contexts, using a Laplace or Normal prior yields similar results. Although the Laplace distribution concentrates more probability mass at zero, it cannot set the posterior mean of coefficients to exactly zero. Achieving this requires using the maximum a posteriori (MAP) estimator, which introduces selection bias by ignoring uncertainty around the mode. Moreover, the penalization induced by the Laplace distribution also results in the loss of the oracle property [34]. The Horseshoe prior, introduced by [35], and its variation, the Regularized Horseshoe [36], could be useful for understanding different types of penalization. The Horseshoe prior uses a hierarchical model with hyperpriors on variance, allowing minimal penalization on certain coefficients. It has heavier tails than Laplace or Normal priors, and the combination of local and global shrinkage parameters helps balance penalization. However, in small samples, global shrinkage can dominate local effects, limiting the prior's effectiveness in high-dimensional settings so it did not appear the appropriate choice for our data structure [37]. In contexts like radiomics, where many variables are correlated, it is not reasonable to assume sparsity. Radiomic variables often capture similar effects, leading to redundancy rather than sparsity. The Horseshoe prior may still require arbitrary criteria for variable selection, which introduces bias. The Hyperlasso prior, though addressing Laplace's limitations, also assumes sparsity [38].

Spike-and-slab priors could be another potential alternative, it can set coefficients exactly to zero but face practical challenges in scenarios with low sample sizes and many variables. In such cases, the distribution may not properly characterize the signal [39, 40]. To avoid bias, it is more appropriate to consider all penalized coefficients sampled from the posterior and account for uncertainty across models.

Considering all these factors, the Ridge prior was selected because it penalizes all coefficients without assuming sparsity, making it suitable for a small sample with many variables. Performing variable selection with non-sparse priors inevitably introduces bias. Using the same data for model estimation and variable selection can lead to errors due to multiple uses of the data. From a Bayesian perspective, selecting variables arbitrarily disregards the uncertainty of excluded variables.

This study presents several notable strengths that address the limitations of prior research in radiomics and prognostic modeling for limb soft tissue sarcomas. Through the implementation of an innovative and unbiased formal test with inferential guarantees, we have rigorously assessed the additional prognostic value that radiomic features contribute beyond established clinical predictors, such as those encapsulated by the Sarculator. The formal Bayesian hypothesis test we built allow us to assess whether the Sarculator model alone was sufficient or whether adding radiomic features provides additional prognostic power. To do so, we assign prior probabilities to reflect the Sarculator's status as a validated model, ensuring a conservative stance on adding complexity. The posterior probability of the Sarculator model was then computed. This approach provides a clear, probabilistic criterion for model sufficiency, favoring interpretability by focusing on posterior probabilities rather than relative evidence measures like the Bayes Factor. In clinical context providing posterior probabilities rather than Bayes Factor is particularly useful, as it allows for a cautious, directly interpretable evaluation of whether adding radiomic features meaningfully improves the Sarculator model's utility.

Furthermore in our Bayesian Model Averaging (BMA) framework, overfitting is unlikely due to several key methodological safeguards. First, the Bayesian approach inherently incorporates uncertainty, utilizing prior and posterior distributions that allow for regularization, particularly essential in high-dimensional, small-sample contexts. The use of WAIC (Widely Applicable Information Criterion) to evaluate posterior probabilities further ensures that model complexity is balanced with data fit, prioritizing models that generalize well rather than simply fitting the estimation data closely. Note that the pWAIC of the best model (which can be seen as the number of effective parameters [21] resulted lower than 1, (0.88). This results further limits the risk of overfitting. This can be derived from the influence of the prior distributions, where, given the limited sample size, the prior information on the 2,144 radiomic parameters dominates the effect detected by the pWAIC. This indicates that the model is not excessively complex relative to the available data. By imposing a regularizing structure, the priors mitigate the risk of fitting noise rather than true signal, promoting a more parsimonious model that enhances generalizability. Moreover, BMA mitigates overfitting risk by averaging across multiple models rather than selecting a single, potentially overfitted model. By including less probable models in the averaging process, BMA reduces sensitivity to any one model's idiosyncratic fit, providing a robust estimate that accounts for model uncertainty and minimizes reliance on any single set of coefficients. Consequently, BMA offers a stable, interpretable approach that enhances model generalizability, providing a more reliable prognostic tool and addressing overfitting concerns effectively within the Bayesian framework.

Nevertheless, this methodology has certain limitations that require careful consideration. First, the high computational demands associated with processing an extensive feature space using A-MALA necessitate substantial resources, which may pose constraints in some research settings. Furthermore, given the relatively small sample size compared to the large number of features, external validation is essential to rigorously assess and quantify the improvement in model performance, ensuring reliable generalizability and mitigating potential biases introduced by the high-dimensional feature space in a limited sample context.

To conclude, this study demonstrates the potential of integrating radiomic features with established clinical predictors to improve prognostic modeling for limb soft tissue sarcomas. The use of an adaptive Bayesian approach, coupled with rigorous inferential guarantees, offers a promising framework for enhancing patient stratification and individualized treatment planning. Despite the challenges posed by computational demands and the need for external validation, our findings lay the groundwork for future research aimed at validating and refining these methods to ultimately improve patient outcomes and quality of care.

References

[1] V. D. Corino, E. Montin, A. Messina, P. G. Casali, A. Gronchi, A. Marchianò, L. T. Mainardi, Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions, Journal of Magnetic Resonance Imaging 47 (2018) 829–840.

[2] C. Fanciullo, S. Gitto, E. Carlicchi, D. Albano, C. Messina, L. M. Sconfienza, Radiomics of musculoskeletal sarcomas: a narrative review, Journal of Imaging 8 (2022) 45.

[3] M. B. Spraker, L. S. Wootton, D. S. Hippe, K. C. Ball, J. C. Peeken, M. W. Macomber, T. R. Chapman, M. N. Hoff, E. Y. Kim, S. M. Pollack, et al., Mri radiomic features are independently associated with overall survival in soft tissue sarcoma, Advances in radiation oncology 4 (2019) 413–421.

[4] J. C. Peeken, M. B. Spraker, C. Knebel, H. Dapper, D. Pfeiffer, M. Devecka, A. Thamer, M. A. Shouman, A. Ott, R. von Eisenhart-Rothe, et al., Tumor grading of soft tissue sarcomas using mri-based radiomics, EBioMedicine 48 (2019) 332–340.

[5] J. C. Peeken, M. Bernhofer, M. B. Spraker, D. Pfeiffer, M. Devecka, A. Thamer, M. A. Shouman, A. Ott, F. Nüsslin, N. A. Mayr, et al., Ct-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy, Radiotherapy and Oncology 135 (2019) 187–196.

[6] S. Zhao, Y. Su, J. Duan, Q. Qiu, X. Ge, A. Wang, Y. Yin, Radiomics signature extracted from diffusion-weighted magnetic resonance imaging predicts outcomes in osteosarcoma, Journal of Bone Oncology 19 (2019) 100263.

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

[8] P. M. Lukacs, K. P. Burnham, D. R. Anderson, Model selection bias and freedman's paradox, Annals of the Institute of Statistical Mathematics 62 (2010) 117–125.

[9] P. Langevin, Sur la théorie du mouvement brownien, Compt. Rendus 146 (1908) 530–533.

[10] M. Girolami, B. Calderhead, Riemann manifold langevin and hamiltonian monte carlo methods, Journal of the Royal Statistical Society Series B: Statistical Methodology 73 (2011) 123–214.

[11] D. Callegaro, R. Miceli, S. Bonvalot, P. Ferguson, D. C. Strauss, A. Levy, A. Griffin, A. J. Hayes, S. Stacchiotti, C. Le Pechoux, et al., Development and external validation of two nomograms to predict overall survival and occurrence of distant metastases in adults after surgical resection of localised soft-tissue sarcomas of the extremities: a retrospective analysis, The Lancet Oncology 17 (2016) 671–680.

[12] D. Callegaro, R. Miceli, S. Bonvalot, P. C. Ferguson, D. C. Strauss, V. V. van Praag, A. Levy, A. M. Griffin, A. J. Hayes, S. Stacchiotti, et al., Development and external validation of a dynamic prognostic nomogram for primary extremity soft tissue sarcoma survivors, EClinicalMedicine 17 (2019).

[13] R. Miceli, D. Callegaro, F. Barretta, A. Gronchi, R. Vergani, Sarculator 2.1.2, https://apps.apple.com/na/app/sarculator/id1052119173, https://play.google.com/store/apps/details?id=it.digitalforest.sarculator&hl=it&gl=US, 2022.

[14] A. J. Hallinan Jr, A review of the weibull distribution, Journal of Quality Technology 25 (1993) 85–93.

[15] T. Hsiang, A bayesian view on ridge regression, Journal of the Royal Statistical Society Series D: The Statistician 24 (1975) 267–268.

[16] S. Van Erp, D. L. Oberski, J. Mulder, Shrinkage priors for bayesian penalized regression, Journal of Mathematical Psychology 89 (2019) 31–50.

[17] A. Beskos, A. Stuart, Computational complexity of metropolis-hastings methods in high dimensions, 2009.

[18] G. O. Roberts, J. S. Rosenthal, Examples of adaptive mcmc, Journal of computational and graphical statistics 18 (2009) 349–367.

[19] A. O'Hagan, J. Forster, Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian

Inference, volume 2B of *Kendall's Library of Statistics*, 2nd ed., Arnold Publishers, London, 2004. Chapters 4: "Asymptotic Approximations" and 5: "The Posterior Distribution and the Information Matrix".

[20] J. E. Dennis Jr, D. M. Gay, R. E. Walsh, An adaptive nonlinear least-squares algorithm, ACM Transactions on Mathematical Software (TOMS) 7 (1981) 348–368. doi:10.1145/355958.355965.

[21] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian data analysis, Chapman and Hall/CRC, 1995.

[22] G. O. Roberts, J. S. Rosenthal, Optimal scaling for various metropolis-hastings algorithms, Statistical science 16 (2001) 351–367.

[23] G. O. Roberts, J. S. Rosenthal, Coupling and ergodicity of adaptive markov chain monte carlo algorithms, Journal of applied probability 44 (2007) 458–475.

[24] H. Haario, E. Saksman, J. Tamminen, An adaptive metropolis algorithm, Bernoulli (2001) 223–242.

[25] S. Watanabe, A widely applicable bayesian information criterion, The Journal of Machine Learning Research 14 (2013) 867–897.

[26] A. E. Raftery, D. Madigan, J. A. Hoeting, Bayesian model averaging for linear regression models, Journal of the American Statistical Association 92 (1997) 179–191.

[27] L. Wasserman, Bayesian model selection and model averaging, Journal of mathematical psychology 44 (2000) 92–107.

[28] T. M. Fragoso, W. Bertoli, F. Louzada, Bayesian model averaging: A systematic review and conceptual classification, International Statistical Review 86 (2018) 1–28.

[29] T. Ando, Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models, Biometrika 94 (2007) 443–458.

[30] S. Watanabe, M. Opper, Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory., Journal of machine learning research 11 (2010).

[31] A. Vehtari, J. Lampinen, Bayesian model assessment and comparison using cross-validation predictive densities, Neural computation 14 (2002) 2439–2468.

[32] A. Vehtari, A. Gelman, J. Gabry, Practical bayesian model evaluation using leave-one-out cross-validation and waic, Statistics and computing 27 (2017) 1413–1432.

[33] K. P. Burnham, D. R. Anderson, Multimodel inference: Understanding AIC and BIC in model selection, Sociological Methods & Research 33 (2004) 261–304. doi:10.1177/0049124104265511.

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

[35] C. M. Carvalho, N. G. Polson, J. G. Scott, Handling sparsity via the horseshoe, Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 5 (2009) 73–80.

[36] J. Piironen, A. Vehtari, Sparsity information and regularization in the horseshoe and other shrinkage priors, Electronic Journal of Statistics 11 (2017) 5018–5051. doi:10.1214/17-EJS1337SI.

[37] A. Bhadra, J. Datta, N. G. Polson, B. Willard, Lasso meets horseshoe, Statistical Science 34 (2019) 405–427.

[38] J. E. Griffin, P. J. Brown, Inference with normal-gamma prior distributions in regression problems, Bayesian Analysis 5 (2010) 171–188. doi:10.1214/10-BA507.

[39] E. I. George, R. E. McCulloch, Approaches for bayesian variable selection, Statistica Sinica 7 (1997) 339–373.

[40] H. Ishwaran, J. S. Rao, Spike and slab variable selection: frequentist and bayesian strategies, The Annals of Statistics 33 (2005) 730–773. doi:10.1214/009053604000001147.