5-1-2023

Identifying regions of interest in mammogram images

Shu Jiang

Jiguo Cao

Graham A. Colditz

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Original Research Article

Statistical Methods in Medical Research
2023, Vol. 32(5) 895–903
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802231160551
journals.sagepub.com/home/smm
Sage

Identifying regions of interest in mammogram images

Shu Jiang[1], Jiguo Cao[2] and Graham A. Colditz[1]

Abstract

Screening mammography is the primary preventive strategy for early detection of breast cancer and an essential input to breast cancer risk prediction and application of prevention/risk management guidelines. Identifying regions of interest within mammogram images that are associated with 5- or 10-year breast cancer risk is therefore clinically meaningful. The problem is complicated by the irregular boundary issue posed by the semi-circular domain of the breast area within mammograms. Accommodating the irregular domain is especially crucial when identifying regions of interest, as the true signal comes only from the semi-circular domain of the breast region, and noise elsewhere. We address these challenges by introducing a proportional hazards model with imaging predictors characterized by bivariate splines over triangulation. The model sparsity is enforced with the group lasso penalty function. We apply the proposed method to the motivating Joanne Knight Breast Health Cohort to illustrate important risk patterns and show that the proposed method is able to achieve higher discriminatory performance.

Keywords

Bivariate splines, Cox proportional hazards model, imaging predictor, group lasso, triangulation

1 Introduction

Breast cancer (BC) is the leading cancer diagnosis among women worldwide, accounting for 1 in 4 cancers diagnosed in women, with more than 2.1 million new cases identified each year.[1] Mammography is currently the primary BC screening strategy. Guidelines recommend screening mammography from the age of 45 or 50 in the US. Making full use of mammogram images is key for risk stratification to implement risk reduction strategies in BC prevention. In this particular setting, a predictive model approach, where mammogram images serve as predictors in predicting the onset of BC over 5 or 10 years (survival-valued response variable) is well-suited to exploit the dependencies between spatial regions and BC risk. The personalized predictions on the future BC risk bring potential clinical utility as demonstrated by current guidelines based on 5- and 10-year risk.[2] However, prediction with imaging data is challenging due to the high dimensionality of the observation space (∼13 million pixels per mammogram).

Regression models incorporating high-dimensional imaging predictors have been extensively studied in the literature. Zipunnikov et al.[3] explored a connection between the singular value decomposition and functional principal component analysis (FPCA) for analyzing brain images by transferring images to a long vector. Other approaches for functional or smooth principal component analysis for tensor data have been proposed; see Huang et al.,[4] Allen,[5] and Lin et al.[6] for example. Reiss and Ogden[7] developed a class of generalized scalar-on-image regression models. Wang et al.[8] considered the scalar-on-image regression model via total variation. Bayesian approaches using spatial priors have also been developed; see Huang et al.[9] and Guo et al.[10] for example.

[1] Division of Public Health Sciences, Washington University School of Medicine, St Louis, MO, USA
[2] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

Corresponding author:
Jiguo Cao, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada.
Email: jiguo_cao@sfu.ca

Despite the rich literature on incorporating imaging data as a predictor, most approaches that use the whole image without pre-specifying summary-level features are based on the assumption that the image is bounded within a rectangular domain. The breast area within mammogram images, however, is bounded within a semi-circular region. For imaging data bounded in a complex domain, many conventional smoothing methods in the literature, such as tensor product, kernel, and wavelet smoothing, all suffer from the problem of boundary leakage.[11] Boundary leakage causes the erroneous selection of background information because sharp changes from the image to the background will dominate the information conveyed from the image; see Ramsay[12] for a few of such examples. Accommodating the irregular boundary is especially crucial when our interest lies in identifying regions of interest (ROIs) for future risk within mammogram images, as the true signal comes only from the semi-circular domain of the breast region, and noise elsewhere.

With the prior belief that a few ROIs are “important” to predict abnormalities or evolving premalignant tissue proliferation within the breast that will more accurately classify risk, we are interested in spatial location selection to identify a few important locations and perform risk stratification based on the selected locations. We aim to encourage ROI selection in a contiguous manner to enable structured and interpretable solutions for potential implementation in the clinical setting. In this paper, we show how the bivariate smooth piecewise polynomial function over triangulations[13,14] can be used to accommodate the irregular domain in the mammogram imaging data. We show how the ROI selection over triangulation can be incorporated into a predictive model. In a translational context for BC, this would mean that the focus can be shifted to the selected ROI to investigate biology and evolution of tissue to cancer over time as recently applied in the NCI Precancer Atlas.[15]

The main contributions of this article are as follows. First, we propose a unified framework incorporating group-wise lasso penalty to identify ROIs within mammogram imaging data that are predictive to 5-year BC risk. Second, we incorporate the bivariate spline basis functions defined on triangulations to capture the variations and connections among values at different locations for mammograms bounded in an irregular domain. Third, we apply the proposed framework to the Joanne Knight Breast Health Cohort at Siteman Center to illustrate important risk patterns and show that we can achieve higher prediction accuracy than benchmark models.

This paper is organized as follows. In Section 2, we define notation and discuss the bivariate spline basis approximation over triangulation to overcome the irregular boundary issue. We show that the spatial location selection over triangulation can be done with the penalized regression with a group-wise penalty. In Section 3, we investigate the finite sample performance of the proposed method via intensive simulation studies. The proposed method is then applied to the motivating Joanne Knight Breast Health Cohort at Siteman Cancer Center in Section 4 to leverage new insights. We conclude the paper with a discussion and future remarks in Section 5.

2 Model and estimation method

2.1 A Cox proportional hazards model with the imaging predictor

Suppose there are n independent individuals in the cohort. For an individual i, we let the pair (T_i, δ_i) denote the observed survival outcome, where T_i is the minimum of failure and censoring time, and $\delta_i = 1$ indicate that the observed T_i is the failure time. We then let $\widetilde{T}_1 < \widetilde{T}_2 < \ldots < \widetilde{T}_n$ be the distinct failure times and define the risk set at time τ_i as $R_i = \{r : T_r \geq \tau_i\}$.

For the imaging predictors, we let Ω be a bounded two-dimensional domain of arbitrary shape, and $\mathbf{s} = (s_1, s_2)$ represent a particular point $\in \Omega \subset \mathbb{R}^2$. We allow the imaging observations $y_i(\mathbf{s})$ for $\mathbf{s} \in \Omega$ to be accompanied with some measurement error,

$$y_i(\boldsymbol{s}) = x_i(\boldsymbol{s}) + \epsilon_i(\mathbf{s}) \tag{1}$$

where $x_i(\boldsymbol{s})$ is the true image with spatial correlated pixels, $\epsilon_i(\mathbf{s})$ are i.i.d. with mean zero and covariance function $\mathrm{cov}\{\epsilon_i(\mathbf{s}), \epsilon_i(\mathbf{s}')\} = I(\mathbf{s} = \mathbf{s}')\sigma^2$, and $x_i(\mathbf{s})$ and $\epsilon_i(\mathbf{s})$ are mutually independent. Without loss of generality, we will assume that $x_i(\mathbf{s})$ is de-meaned, that is, $E\{x_i(\boldsymbol{s})\} = 0$, from here on.

Because the mammogram images are observed on a semi-circular region with an irregular boundary, we approximate the imaging data by the bivariate spline that are piecewise polynomial functions over a two-dimensional triangulated domain.[14] We first define some notations as follows. Let τ be a triangle that is the convex hull of three points that are not collinear. A collection $\triangle = \{\tau_1, \ldots, \tau_J\}$ of triangles is called a triangulation of $\Omega = \bigcup_{j=1}^{J} \tau_j$ if any nonempty intersection between a pair of triangles in $\triangle$ is either a common vertex or a common edge. An example of the triangulation with $J = 115$ for the mammogram imaging data is shown in Figure 1. The triangulation plot can be created using `Matlab` function `delaunay`. We now define a degree d and smoothness r spline space over triangulation $\triangle$: $\mathcal{T}_d^r(\triangle) = \{t \in C^r(\Omega) : t|_\tau \in \mathbb{P}_d, \tau \in \triangle\}$,

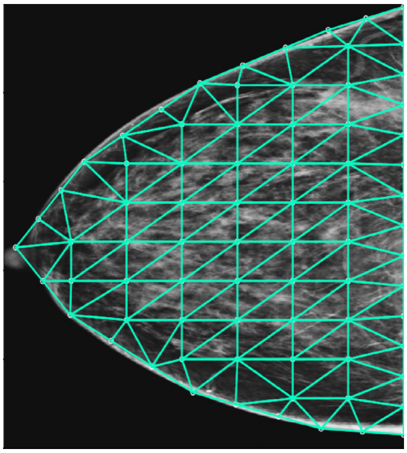

Figure 1. The triangulation grid with 115 triangles.

where $C^r(\Omega)$ is the collection of all rth continuously differentiable functions over Ω, $r \geq 0$. The space of all polynomials with degree $\leq d$ is denoted with $\mathbb{P}_d$ and thus $t|_\tau$ is the polynomial restricted on triangle τ.

We adopt the Bernstein polynomial basis function. For an arbitrary point $\mathbf{s} \in \Omega$ and the triangle $\tau \in \triangle$, we define b_1, b_2, and b_3 to be the barycentric coordinates of the point $\mathbf{s}$ relative to the triangle τ. The barycentric coordinates of a point $\mathbf{s}$ can be interpreted as mass placed at the vertices of the triangle τ. The masses are all positive if and only if the point is inside of the triangle. The Bernstein basis polynomial of degree d for a point $\mathbf{s}$ relative to a triangle τ can then be written as $(d_1!d_2!d_3!)^{-1}d!b_1^{d_1}b_2^{d_2}b_3^{d_3}$, for $d_1 + d_2 + d_3 = d$.

The imaging predictor $x_i(\boldsymbol{s})$ is expanded with a set of bivariate Bernstein polynomial basis functions as:

$$x_i(\boldsymbol{s}) = \sum_{j=1}^{J}\sum_{k=1}^{K} c_{ijk}\phi_k^{(j)}(\boldsymbol{s}) \tag{2}$$

where j is the index for jth triangle, $j = 1, \ldots, J$, and k is the kth basis function within the jth triangle, $k = 1, \ldots, K$. Note that K is a function of d which is the degree of polynomial that does not vary between groups.

In addition to the imaging data, we assume that there are some non-functional demographic variables z_i of length $q \times 1$. We consider a Cox proportional hazards model with the hazard function,

$$h_i(t) = h_0(t)\exp\left\{\alpha^{\mathrm{T}} z_i + \int_{s\in\Omega} \beta(\boldsymbol{s})x_i(\boldsymbol{s})\mathrm{d}\boldsymbol{s}\right\} \tag{3}$$

where $h_0(t)$ is the unspecified baseline hazards function, $\alpha = (\alpha_1, \ldots, \alpha_q)^{\mathrm{T}}$ is the coefficient vector for z_i. We expand $\beta(\boldsymbol{s})$ using the set of bivariate Bernstein polynomial basis functions as:

$$\beta(\boldsymbol{s}) = \sum_{j=1}^{J}\sum_{k=1}^{K} b_{jk}\phi_k^{(j)}(\boldsymbol{s}) \tag{4}$$

Then, we can rewrite the equation (3) as:

$$\begin{aligned} h_i(t) &= h_0(t)\exp\left\{\alpha^{\mathrm{T}} z_i + \int_{s\in\Omega}\sum_{j=1}^{J}\sum_{k=1}^{K} b_{jk}\phi_k^{(j)}(\boldsymbol{s})\sum_{j'=1}^{J}\sum_{k'=1}^{K} c_{ij'k'}\phi_{k'}^{(j')}(\boldsymbol{s})\mathrm{d}\boldsymbol{s}\right\} \\ &= h_0(t)\exp\left\{\alpha^{\mathrm{T}} z_i + \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{k'=1}^{K} b_{jk}\langle\phi_k^{(j)}, \phi_{k'}^{(j)}\rangle c_{ijk'}\right\} \\ &= h_0(t)\exp\left\{\alpha^{\mathrm{T}} z_i + \sum_{j=1}^{J} b_j^{\mathrm{T}} W^{(j)} c_{ij}\right\} \end{aligned} \tag{5}$$

where $b_j = (b_{j,1}, \ldots, b_{j,K})^{\mathrm{T}}$, $W^{(j)} = \langle\{\phi^{(j)}\}^{\mathrm{T}}, \phi^{(j)}\rangle$, and $c_{ij} = (c_{ij,1}, \ldots, c_{ij,K})^{\mathrm{T}}$.

In matrix notation, we can rewrite (5) as

$$h_i(t) = h_0(t)\exp\{\alpha^{\mathrm{T}} z_i + b^{\mathrm{T}} W c_i\} \tag{6}$$

where $b = (b_1, b_2, \dots, b_J)^{\mathrm{T}}$, $W = \mathrm{diag}(W^{(1)}, \dots, W^{(J)})$ is a block-diagonal and positive-definite matrix, and $c_i = (c_{i,1}, \dots, c_{i,J})^{\mathrm{T}}$. Given the formulation of the Cox proportional hazards model, we can express the negative partial log-likelihood as,

$$l_p(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \delta_i \left[\alpha^{\mathrm{T}} z_i + b^{\mathrm{T}} W c_i - \log\left\{\sum_{r \in R_i} \exp(\alpha^{\mathrm{T}} z_r + b^{\mathrm{T}} W c_r)\right\}\right] \tag{7}$$

where R_i is the set of indices of the patients who are at risk at time τ_i, $R_i = \{r : T_r \geq \tau_i\}$, $\theta = (\alpha, b)$.

2.2 Group lasso for identifying ROIs

We consider the vector of K coefficients $b_j = (b_{j,1}, \dots, b_{j,K})^{\mathrm{T}}$ to the bivariate Bernstein polynomial basis functions in the jth triangle as a group, $j = 1, \dots, J$, and apply the group lasso approach[17,16] to minimize

$$l_p(\theta) + \lambda \sum_{j=1}^{J} s(K_j)||b_j||_2 \tag{8}$$

where the demographic variables z_i are not penalized in our setting, $s(K_j)$ is used to rescale the penalty with respect to the dimensionality of the group size, usually taken to be the square root of K, and $\lambda > 0$ is the tuning parameter to control the amount of penalty. The group lasso penalty can be viewed as an intermediate between the l_1 and l_2-type penalty where variable selection is done under the group level. Thus, if triangle j is not penalized, then all values of b_j are non-zero, ensuring that the entire area defined by K basis functions are kept simultaneously. Under the Cox regression, the penalty parameter λ can be determined by cross-validated partial likelihood (CVPL) as in Segal[18] or the model selection criteria such as Akaike information criterion or Bayesian information criterion as in Huang et al.[19] Note that standardization is carried out prior to the optimization algorithm as follows: let the $K \times 1$ vectors $\tau_{ij} = W^{(j)} c_{ij} = (\tau_{ij1}, \dots, \tau_{ijK})^{\mathrm{T}}$, then τ_{ijk} is standardized to have mean 0 and variance 1 across i for each given j and k.[20]

3 Simulation study

We investigate the finite sample performance of the proposed method in this simulation study. The measurements on the images are simulated as follows:

$$y_i(\mathbf{s}) = \mu(\mathbf{s}) + \sum_{j=1}^{J}\sum_{k=1}^{K} c_{ijk}\phi_k^{(j)}(\mathbf{s}) + \epsilon_i(\mathbf{s}) \tag{9}$$

where $\mathbf{s} \in \Omega$, and $\mu(\mathbf{s}) = 0$ without loss of generality, $i = 1, \dots, n$. Specifically, we considered the set of bivariate Bernstein polynomial basis function ϕ of degree 2 with $J = 80$ triangles, 54 vertices, and $J = 144$ triangles, 87 vertices. The corresponding coefficients c_{ijk} associated with the bivariate Bernstein polynomial are assumed to follow $MVN(\mathbf{0}, \Sigma)$, where we set the diagonal of Σ to be 1, the within triangle correlation ρ^W to be 0.4 or 0.8, and we consider the between triangle correlation ρ^B to be 0 or 0.1. We also consider two different settings for the measurement error $\epsilon_i(\mathbf{s})$ that follows a normal distribution with mean 0 and covariance function $\mathrm{cov}\{\epsilon_i(\mathbf{s}), \epsilon_i(\mathbf{s}')\} = I(\mathbf{s} = \mathbf{s}')\sigma^2$ for $\sigma = 0$ and 0.2.

To generate the survival outcome, we consider the following proportional hazards model:

$$h_i(t) = h_0(t)\exp\left\{\int_{\mathbf{s}\in\Omega} \beta(\mathbf{s}) \sum_{j \in \mathcal{S}_j}\sum_{k=1}^{K} c_{ijk}\phi_k^{(j)}(\mathbf{s})\mathrm{d}\mathbf{s}\right\}$$

where $\mathcal{S}_j$ denotes the set of triangles that are associated with the hazard function over time, and $\beta(\mathbf{s}) = b^{\mathrm{T}}\phi(\mathbf{s})$. For each selected triangle j, there are 6 Bernstein basis polynomials, and we set the vector of basis coefficients $b_j = (b_{j1}, \dots, b_{j6})^{\mathrm{T}}$ to take on values of (0.10, 0.28, 0.46, 0.64, 0.82, 1) for all $j \in \mathcal{S}_j$; this is just a random sequence taken from 0.10 to 1 with equal increments. We set the maximum time to the end of the study to be 15 years. The baseline hazard function is assumed to follow a Weibull distribution $h_0(t) = \kappa\omega(\omega t)^{\kappa - 1}$ with increasing risk over time, where we set $\kappa = 2$ and $\omega = 0.16$. The

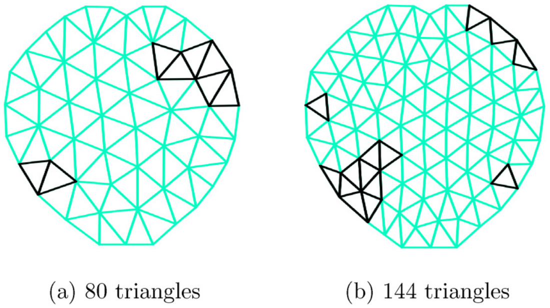

Figure 2. Illustration of the triangulation grid used for the simulation study with irregular boundary; triangles colored in black are the ones that are selected to be predictive of the hazard function. This triangulation plot is created using the `Matlab` function `delaunay`.[21]

survival time T_i is generated from the inverse of the cumulative hazard function $H^{-1}(u)$, where $u \sim \text{unif}(0, 1)$. We have assumed the independent censoring scheme in this simulation study, where $C_i \sim \text{unif}(0, C_{max})$, with C_{max} set at a value such that the % of being censored by the end of the study is approximately 30%.

The number of selected triangles $|\mathcal{S}_j|$ has been set to be 2.5%, 5%, and 10% of the total number of triangles J within the two settings of the triangulation. We have rounded down the decimal place if the percentage did not result in a whole number for the triangles selected. The percentages are guided by the clinical setting in BC where abnormalities within the breast are usually contiguous and localized in a small patch. We constrain the images to be on a rectangular grid of size 40 × 40, with 921 points falling inside an irregular circular bounded region in a similar manner as our motivating mammogram study. Illustration of the triangulation grid and the selected triangles to be associated with the hazard under the two settings are displayed in Figure 2. Note that we have picked a portion of triangles located on the boundary of the circular region to be associated with the hazard function to illustrate the particular importance to apply the triangulation basis functions.

We show the simulation results in Tables 1 and 2. For each setting, we have simulated 100 datasets with $n = 1000$ and 2000 individuals per dataset. For each dataset, a 10-fold cross-validation was carried out to select the unknown tuning parameter using the CVPL criteria. The percentage of variables selected was recorded. Among those that were truly associated with the hazard, the % of selected triangles across all simulated datasets is reported as the true positive (TP) selections. Among the triangles that have no association with the event time, the % selected for each dataset was averaged and reported as the false positive (FP) selections. We have additionally simulated 500 individuals for each dataset that is used as a validation set to record the prediction performance of the proposed method to avoid over-optimism. The prediction performance is recorded as the integrated area under the receiver operating characteristic curve (AUC).[22]

From Tables 1 and 2, we can observe a decreasing trend in the TP rate with an increase in $|\mathcal{S}_j|$. Within the same $|\mathcal{S}_j|$, we also observe an increase in TP with an increase in sample size. This tells us that a larger sample size is needed to have the power for detecting a greater number of triangles. The FP rates stayed low in all settings which is satisfactory. Additionally, we see that an increase in the within-triangle correlation ρ^W results in better TP rates and that an increase in σ results in slightly worse TP rates. We have also investigated the performance when the triangle correlation ρ^B is set to 0 and 0.1, shown in Tables 1 and 2, respectively. It is expected that we would have a higher TP and lower FP when the triangle correlation is lower and our results support that. The out-of-sample prediction performance reported as the AUC values are satisfactory under all settings.

Software in the form of `R` code,[23] together with a sample input data set and complete documentation is now available on our GitHub repository https://github.com/jj113/roi.

4 The Joanne Knight Breast Health Cohort

This study is motivated by the Joanne Knight Breast Health Cohort at Siteman Cancer Center. The cohort was established to link BC risk factors, mammographic breast density, and blood markers for women from varying socioeconomic and racial backgrounds in the St. Louis region undergoing routine mammographic screening. Women were recruited from November 2008 to April 2012, and have been followed through October 2020. We treat the conversion from baseline (BC free) to the onset of BC as the time-to-event outcome and focus on 785 women within the nested case-control cohort at the baseline. We have excluded women who have been diagnosed with BC within the first 6 months of entry to the cohort and those without a valid craniocaudal view mammogram. Of those who were included in the study, 246 have been diagnosed with

Table 1. Empirical results summarized by the % of correctly (TP) and incorrectly (FP) selected triangles along with the integrated AUC; $\rho^B = 0$.

				$n = 1000$			$n = 2000$		
ρ^W	σ	J	$\lvert S_j \rvert$	TP	FP	AUC	TP	FP	AUC
0.4	0	80	2.5%	0.995	0.102	0.894	1.000	0.117	0.898
			5%	0.803	0.110	0.888	0.930	0.129	0.885
			10%	0.641	0.090	0.902	0.709	0.104	0.907
		144	2.5%	0.878	0.066	0.852	0.953	0.072	0.858
			5%	0.711	0.058	0.801	0.780	0.066	0.802
			10%	0.522	0.070	0.886	0.617	0.075	0.885
	0.2	80	2.5%	0.995	0.103	0.890	1.000	0.118	0.895
			5%	0.803	0.114	0.888	0.903	0.128	0.888
			10%	0.640	0.091	0.901	0.721	0.093	0.905
		144	2.5%	0.853	0.061	0.832	0.980	0.074	0.860
			5%	0.700	0.058	0.797	0.797	0.068	0.800
			10%	0.513	0.067	0.891	0.610	0.079	0.881
0.8	0	80	2.5%	1.000	0.107	0.900	1.000	0.130	0.902
			5%	0.788	0.115	0.893	0.928	0.133	0.893
			10%	0.653	0.090	0.905	0.701	0.108	0.909
		144	2.5%	0.883	0.072	0.865	0.963	0.085	0.866
			5%	0.714	0.061	0.803	0.789	0.071	0.805
			10%	0.530	0.071	0.892	0.614	0.082	0.891
	0.2	80	2.5%	0.990	0.108	0.899	1.000	0.123	0.900
			5%	0.750	0.118	0.893	0.920	0.131	0.892
			10%	0.638	0.091	0.903	0.705	0.096	0.907
		144	2.5%	0.865	0.062	0.861	0.985	0.076	0.864
			5%	0.701	0.059	0.802	0.787	0.067	0.801
			10%	0.533	0.067	0.885	0.611	0.082	0.886

TP: true positive; FP: false positive; AUC: area under the receiver operating characteristic curve.

BC prior to the end of follow-up. The set of mammogram images was pre-processed prior to any analytical procedures to ensure that the breast area on the mammogram is approximately aligned across a woman in the cohort; see Jiang et al.,[24,25] for more details on the registration procedure. The pixel intensities between the two mammograms within a woman (left and right) are then averaged in a similar fashion for breast density which is well accepted in the BC literature.[26]

In this analysis, we consider a proportional hazards function with the demographic variables z being age, weight, and height at the baseline. Note that the three demographic variables are not penalized in the proportional hazards model. We considered $J = 115$ triangles as shown in Figure 2 with a degree 3 polynomial Bernstein basis function. We illustrate the locations selected in Figure 3 in the cohort as a function of different penalty values (shown as the horizontal dotted lines). A specific location is selected if it exceeds the dotted line. For example, we find that when $\lambda \geq 0.031$, triangles at locations 7, 12, 78, and 112 are selected. As λ increases, fewer locations are selected as a function of the penalty; in contrast, when $\lambda = 0$, all locations are selected. Given this property, the order of selection from larger λ values to smaller λ values suggests the importance of the locations to the hazard function. For example, we can see that the spatial locations above the 0.031 thresholds are more important as compared to all other spots. This is particularly useful in clinical settings when evaluating potential abnormalities associated with each spot.

The optimal penalty value as a function of a 5-fold cross-validation using the CVPL criteria is shown in Figure 3 with the blue horizontal dotted line. As a result, all spatial locations that exceed the blue dotted line were selected in this study. We further illustrate these specific locations mapped onto the original mammogram in Figure 4. Additionally, we have conducted a 10-fold internal cross-validation to avoid over-optimism of the prediction performance in estimating the 5-year BC risk. The prediction performance is assessed using the integrated AUC. Figure 5 shows the estimated AUC averaged over the 10-fold cross-validation as a function of different penalty values. Of note, the blue vertical dotted line in Figure 5 is in correspondence with the optimal λ value $\lambda = 0.025$, which is also shown in Figure 3. As we can see, there is a notable gain in applying the group-lasso on the proportional hazards model as compared to without using any penalty.

Additional sensitivity analysis with decreasing (87) and increasing (150) number of triangles instead of the presented 115 are shown in the supplementary document for interested readers. The locations selected with 87 and 150 triangles are

Table 2. Empirical results summarized by the % of correctly (TP) and incorrectly (FP) selected triangles along with the integrated AUC; $\rho^B = 0.1$.

				$n = 1000$			$n = 2000$		
ρ^W	σ	J	$\lvert\mathcal{S}_j\rvert$	TP	FP	AUC	TP	FP	AUC
0.4	0	80	2.5%	1.000	0.110	0.928	1.000	0.128	0.929
			5%	0.750	0.127	0.945	0.855	0.144	0.944
			10%	0.625	0.100	0.965	0.733	0.089	0.967
		144	2.5%	0.723	0.083	0.881	0.920	0.080	0.882
			5%	0.604	0.080	0.908	0.747	0.089	0.909
			10%	0.434	0.075	0.950	0.511	0.083	0.972
	0.2	80	2.5%	1.000	0.114	0.927	1.000	0.126	0.927
			5%	0.750	0.128	0.948	0.883	0.135	0.942
			10%	0.605	0.090	0.964	0.748	0.095	0.967
		144	2.5%	0.723	0.072	0.878	0.913	0.076	0.881
			5%	0.587	0.081	0.906	0.726	0.087	0.907
			10%	0.449	0.075	0.949	0.502	0.081	0.970
0.8	0	80	2.5%	1.000	0.110	0.921	1.000	0.126	0.923
			5%	0.835	0.119	0.931	0.900	0.136	0.931
			10%	0.674	0.080	0.953	0.750	0.092	0.957
		144	2.5%	0.805	0.071	0.851	0.968	0.075	0.853
			5%	0.750	0.081	0.903	0.831	0.082	0.901
			10%	0.500	0.070	0.958	0.571	0.080	0.960
	0.2	80	2.5%	0.990	0.112	0.920	1.000	0.127	0.921
			5%	0.818	0.127	0.931	0.905	0.135	0.930
			10%	0.664	0.079	0.953	0.736	0.094	0.956
		144	2.5%	0.801	0.070	0.852	0.940	0.072	0.852
			5%	0.733	0.071	0.901	0.829	0.081	0.900
			10%	0.503	0.068	0.958	0.571	0.077	0.958

TP: true positive; FP: false positive; AUC: area under the receiver operating characteristic curve.

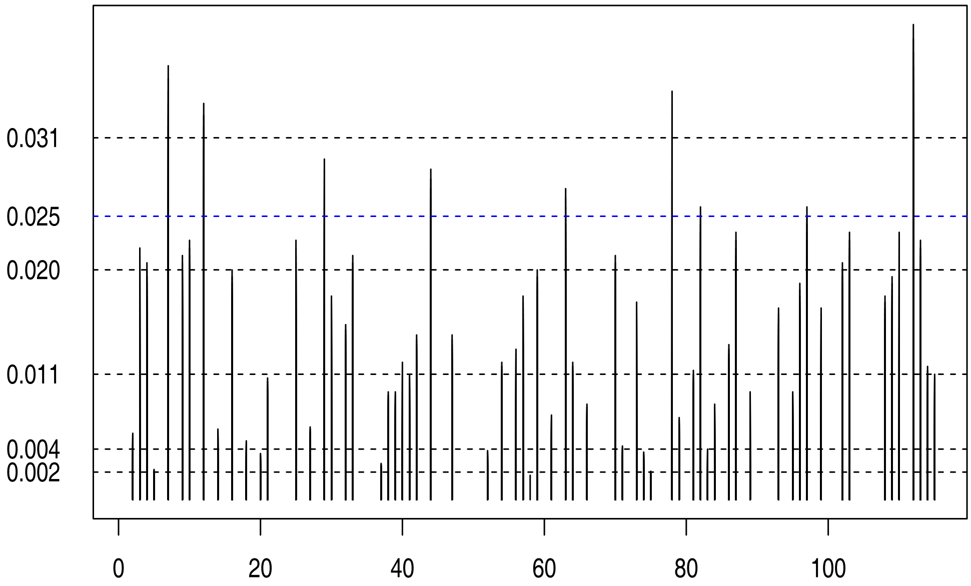

Figure 3. The selected spatial locations (115 triangles) at different penalty λ values (dotted lines); the blue dotted line indicates the optimal penalty value selected using the cross-validated partial likelihood (CVPL) criteria.

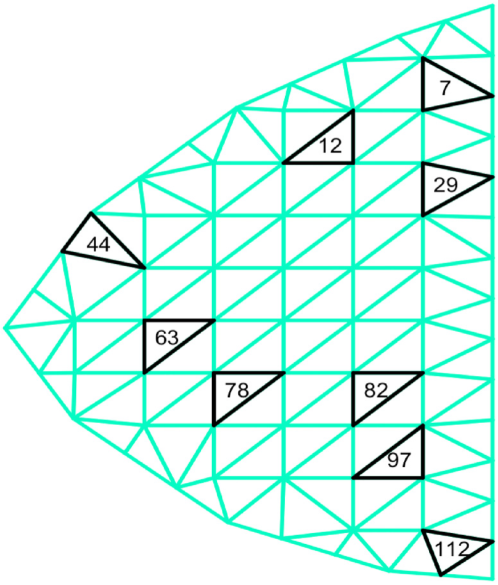

Figure 4. The selected locations with the tuned penalty value mapped onto the mammogram image.

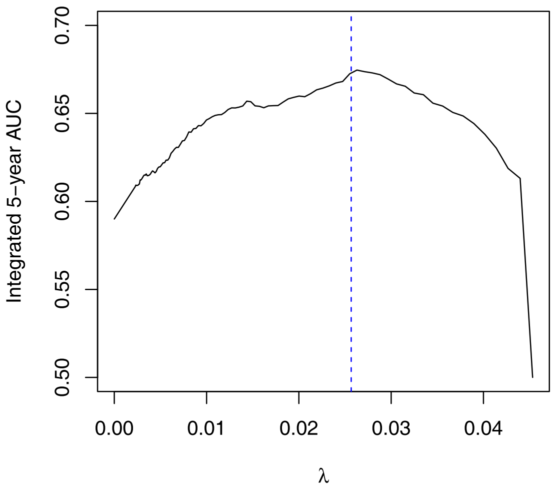

Figure 5. The average 5-year integrated AUC across the 10-fold cross-validation at different penalty values; the blue dashed line represents the selected optimal value using the CVPL criteria. AUC: area under the receiver operating characteristic curve; CVPL: cross-validated partial likelihood.

similar to what we observe in Figure 4. The average AUC value was superior for the setting with 115 triangles, thus, we present the results under the current setting. Finally, we would like to comment on the promising computational time for the proposed analysis. The entire estimation, applied to our cohort dataset, took approximately 30 min on a standard laptop (2.9 GHz, 16 GB RAM) without parallel computing.

5 Discussion

In this article, we proposed a modeling framework for incorporating mammogram images bounded in an irregular domain. Based on this framework, a partial data log-likelihood based on the proportional hazards model with group lasso penalty is formulated to (1) enforce sparsity and identify ROIs that are associated with the survival outcome; and (2) predict the 5-year risk of BC. The finite sample performance of the proposed method has been evaluated from intensive simulation studies. We have applied the proposed framework to the motivating dataset from the Joanne Knight Breast Health Cohort at Siteman Cancer Center. We have identified ROI within mammograms that are associated with 5-year BC risk. This is particularly meaningful in practice, as the focus could potentially be shifted to the selected ROI to investigate the biology and evolution of tissue to BC over time. The importance of these ROIs has been illustrated with the barplot. Finally, the prediction performance is superior with the optimal penalty parameter tuned from a nested cross-validation as compared to without including any penalty at all. Although our method is motivated by the mammogram imaging data, it can be naturally applied in other settings, such as brain image and lung computed tomography, where they are also bounded within an irregular domain.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is partially supported by the NCI (R37 CA256810), Breast Cancer Research Foundation (BCRF 20-028), and the NSERC Discovery grant (RGPIN-2018-06008).

ORCID iD

Jiguo Cao https://orcid.org/0000-0001-7417-6330

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018; **68**: 394–424.
2. Bevers TB, Ward JH, Arun BK, et al. Breast cancer risk reduction, version 2.2015. *J Natl Compr Canc Netw* 2015; **13**: 880–915.
3. Zipunnikov V, Caffo B, Yousem DM, et al. Functional principal component model for high-dimensional brain imaging. *NeuroImage* 2011; **58**: 772–784.
4. Huang JZ, Shen H and Buja A. The analysis of two-way functional data using two-way regularized singular value decompositions. *J Am Stat Assoc* 2009; **104**: 1609–1620.
5. Allen GI. Multi-way functional principal components analysis. In: *2013 5th IEEE international workshop on computational advances in multi-sensor adaptive processing (CAMSAP)*, pp.220–223. New York: IEEE, 2013.
6. Lin N, Jiang J, Guo S, et al. Functional principal component analysis and randomized sparse clustering algorithm for medical image analysis. *PLoS ONE* 2015; **10**: e0132945.
7. Reiss PT, Ogden RT. Functional generalized linear models with images as predictors. *Biometrics* 2010; **66**: 61–69.
8. Wang X, Zhu H and Initiative ADN. Generalized scalar-on-image regression models via total variation. *J Am Stat Assoc* 2017; **112**: 1156–1168.
9. Huang L, Goldsmith J, Reiss PT, et al. Bayesian scalar-on-image regression with application to association between intracranial DTI and cognitive outcomes. *NeuroImage* 2013; **83**: 210–223.
10. Guo C, Kang J and Johnson TD. A spatial Bayesian latent factor model for image-on-image regression. *Biometrics* 2020; **78**: 72–84.
11. Wood SN, Bravington MV and Hedley SL. Soap film smoothing. *J R Stat Soc B Stat Methodol* 2008; **70**: 931–955.
12. Ramsay T. Spline smoothing over difficult regions. *J R Stat Soc B Stat Methodol* 2002; **64**: 307–319.
13. Lai M-J, Schumaker LL. *Spline functions on triangulations*. **110**. Cambridge: Cambridge University Press, 2007.
14. Lai M-J, Wang L. Bivariate penalized splines for regression. *Stat Sin* 2013; 1399–1417.
15. Risom T, Glass DR, Averbukh I, et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* 2022; **185**: 299–310.
16. Meier L, Van De Geer S and Bühlmann P. The group lasso for logistic regression. *J R Stat Soc B Stat Methodol* 2008; **70**: 53–71.
17. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc B Stat Methodol* 2006; **68**: 49–67.
18. Segal MR. Microarray gene expression data with linked survival phenotypes: diffuse large-b-cell lymphoma revisited. *Biostatistics* 2006; **7**: 268–285.
19. Huang J, Liu L, Liu Y, et al. Group selection in the cox model with a diverging number of covariates. *Stat Sin* 2014; **24**: 1787–1810.
20. Simon N, Tibshirani R. Standardization and the group lasso penalty. *Stat Sin* 2012; **22**: 983.
21. MATLAB. *Version: R2022b*. Natick, MA: The MathWorks Inc., 2022.
22. Uno H, Cai T, Tian L, et al. Evaluating prediction rules for t-year survivors with censored regression models. *J Am Stat Assoc* 2007; **102**: 527–537.
23. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022.
24. Jiang S, Cao J, Rosner B, et al. Supervised two-dimensional functional principal component analysis with time-to-event outcomes and mammogram imaging data. *Biometrics* 2021. DOI: 10.1111/biom.13611.
25. Jiang S, Cao J, Colditz GA, et al. Predicting the onset of breast cancer using mammogram imaging data with irregular boundary. *Biostatistics* 2021: kxab032.
26. Brentnall AR, Harkness EF, Astley SM, et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UKscreening cohort. *Breast Cancer Res* 2015; **17**: 1–10.