Journal of Machine Learning Research 17 (2016) 1-15 Submitted 8/15; Revised 3/16; Published 12/16

Structure-Leveraged Methods in Breast Cancer Risk Prediction

Jun Fan JUNFAN@STAT.WISC.EDU
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue, Madison, WI 53706, United States
Yirong Wu YWu@UWHEALTH.ORG
Department of Radiology
University of Wisconsin-Madison
600 Highland Avenue, Madison, WI 53792, United States
Ming Yuan MYUAN@STAT.WISC.EDU
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue, Madison, WI 53706, United States
David Page PAGE@BIOSTAT.WISC.EDU
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
600 Highland Avenue, Madison, WI 53792, United States
Jie Liu LIU6@UW.EDU
Department of Genome Sciences
University of Washington-Seattle
3720 15th Avenue, Seattle, WA 98105, United States
Irene M. Ong ONG@CS.WISC.EDU
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
600 Highland Avenue, Madison, WI 53792, United States
Peggy Peissig PEISSIG.PEGGY@MCRF.MFLDCLIN.EDU
Marshfield Clinic Research Foundation
1000 North Oak Avenue, Marshfield, WI 54449, United States
Elizabeth Burnside EBURNSIDE@UWHEALTH.ORG
Department of Radiology
University of Wisconsin-Madison
600 Highland Avenue, Madison, WI 53792, United States

Editor: Benjamin M. Marlin and Suchi Saria

Abstract

Predicting breast cancer risk has long been a goal of medical research in the pursuit of precision medicine. The goal of this study is to develop novel penalized methods to improve breast cancer risk prediction by leveraging structure information in electronic health records. We conducted a retrospective case-control study, garnering 49 mammography descriptors and 77 high-frequency/low-penetrance single-nucleotide polymorphisms (SNPs) from an existing personalized medicine data repository. Structured mammography reports and breast imaging features have long been part of a standard electronic health record (EHR), and genetic markers likely will be in the near future. Lasso and its variants are widely used approaches to integrated learning and feature selection, and our methodological contribution is to incorporate the dependence structure among the features into these approaches. More specifically, we propose a new methodology by combining group penalty and ℓ^p $(1 \leq p \leq 2)$ fusion penalty to improve breast cancer risk prediction, taking into account structure information in mammography descriptors and SNPs. We demonstrate that our method provides benefits that are both statistically significant and potentially significant to people's lives.

Keywords: structure information, breast cancer risk prediction, mammography descriptors, genetic variants, personalized medicine

1. Introduction

Breast cancer is the most common non-skin malignancy affecting women, with approximately 1.67 million cases diagnosed annually worldwide (Ferlay et al., 2013). If an individual's risk of breast cancer could be predicted, then screening, prevention, and treatment strategies could be targeted toward those women to maximize survival benefit and minimize harm. Risk prediction models are important tools to improve breast cancer care by leveraging multi-dimensional electronic health data. Traditional breast cancer risk prediction models use demographic risk factors to estimate breast cancer risk, but they demonstrate only limited discriminatory power. In clinical practice, mammography is the most common breast cancer screening test, and the only imaging modality supported by randomized trials demonstrating reduction in mortality rate. However, its effectiveness is not universally accepted (Freedman et al., 2004). Recent advances in genome-wide association studies (GWAS) have revitalized the quest for genetic variants (single-nucleotide polymorphisms—SNPs) in risk prediction. However, the optimism of these studies has been tempered by disappointment and caution (Gail, 2008, 2009; Wacholder et al., 2010).

Although many breast cancer risk prediction models have been developed, current applications of these models are inadequate in the following respects: (1) due to the rare occurrence of breast cancer, many seemingly 'large' studies have small effective sample size to adequately model a large number of variables; (2) even for large studies, investigators often fail to systematically model risk factor interactions to avoid overly complicated models which are hard to interpret; and (3) they do not take available structure information into consideration. For example, there are five descriptors for mass margins in mammogram: circumscribed, microlobulated, obscured, indistinct, and spiculated, with an order of increasing probability of malignancy. However, few models utilize this structure information (group structure and dependence structure) to improve predictive performance. The quest for novel breast cancer risk prediction models is motivated to address these shortcomings.

In this paper, we propose to develop novel penalized methods to improve breast cancer risk prediction by incorporating unique structure information embedded in electronic health record data. Regularization is a common technique used in regression and classification problems. The lasso (Tibshirani, 1996) is one of the most popular penalized method and has achieved great success in various fields. However, lasso does not take into account the prior structure information among features. The group lasso (Yuan and Lin, 2006) is a natural extension of the lasso by taking advantage of the underlying group structure of features. It leads to the selection for groups of features and can improve the predictive performance in many real applications such as microarray data analysis (Ma et al., 2007) and GWAS (Liu et al., 2013). To incorporate the dependence structure of features, fused lasso (Tibshirani et al., 2005; Tibshirani and Wang, 2008) is introduced by penalizing the ℓ^1 norm of both the coefficients and their successive differences. To the best of our knowledge, no breast cancer prediction models utilize group penalty and within-group ℓ^p fusion penalty simultaneously to improve risk prediction by leveraging structure information.

The rest of the paper is organized as follows. Section 2 describes our data, proposed methods, and study design. Section 3 presents the results. The conclusions are described in Section 4.

2. Materials and Methods

The main purpose of this paper is to take into account both the group structure and the dependence structure within each group of features by imposing both group penalty and ℓ^p fusion penalty simultaneously.

2.1 Data

The Marshfield Clinic Institutional Review Board approved the use of Marshfield Clinic's Personalized Medicine Research Project (PMRP) (McCarty et al., 2005) cohort in our study.

2.1.1 Subjects

The population-based PMRP cohort, details of which have been previously published (McCarty et al., 2005), was used in this study. Though the details of this population have been described previously (Burnside et al., 2015), we will summarize here, in brief, for the convenience of the reader. Women with an available DNA sample, a mammogram, and a breast biopsy within 12 months after the mammogram were included in the study. For this case/control study, cases were defined as women having a confirmed diagnosis of breast cancer obtained from the institutional cancer registry. Controls were confirmed through the Marshfield Clinic electronic medical records as never having had a breast cancer diagnosis (and absence from cancer registry).

We identified 362 cases and 376 controls (738 in total) who have both genetics and mammogram data available. The majority of mammograms were performed between 1993 and 2005 (Burnside et al., 2015). The age range for the subjects in this study was 29 to 90 years of age, with mean 62 and standard deviation 12.8. Among the cases, there were 358 Caucasians, three non-Caucasians and one case whose race information was unknown. Among the controls, there were 372 Caucasians and four non-Caucasians. These race distributions are consistent with that of the general population in this area. For the family history of breast cancer, we observed a considerably larger proportion of people with family history in the case group (45.30%) than in the control group (33.51%), which demonstrated the family aggregation of breast cancer (Table 1).

Family history	Cases	Controls	All
Yes	164(45.30%)	126(33.51%)	290(39.30%)
No	188(51.93%)	236(62.77%)	424(57.45%)
N/A	10(2.77%)	14(3.72%)	24(3.25%)

Table 1: Family aggregation of breast cancer.

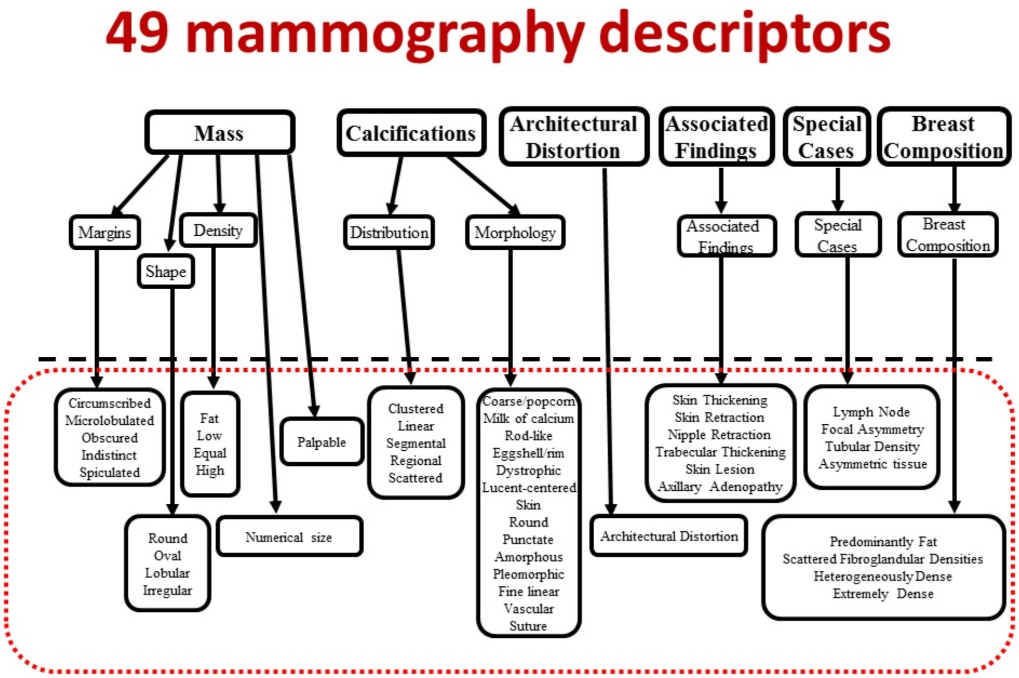


Figure 1: Mammography descriptors described in BI-RADS lexicon.

2.1.2 Mammography features

Mammography features are recorded in the Breast Imaging Reporting and Data System (BI-RADS) lexicon (BI-RADS, 2014) developed by the American College of Radiology. The BI-RADS lexicon consists of 49 descriptors, including the characteristics of masses and microcalcifications, breast composition and other associated findings (Figure 1). In this study, mammography data was recorded as free text reports in the electronic health record, from which we used a parser to extract these mammography features (Nassif et al., 2009).

SNP	Chr	SNP	Chr
rs616488	1	rs11814448	10
rs11249433	1	rs7072776	10
rs1550623	2	rs7904519	10
rs16857609	2	rs2981582	10
rs2016394	2	rs10995190	10
rs4849887	2	rs2380205	10
rs1045485	2	rs2981579	10
rs13387042	2	rs704010	10
rs17468277	2	rs11820646	11
rs4666451	2	rs3903072	11
rs12493607	3	rs3817198	11
rs6762644	3	rs2107425	11
rs4973768	3	rs614367	11
rs6828523	4	rs12422552	12
rs9790517	4	rs17356907	12
rs10472076	5	rs6220	12
rs1353747	5	rs10771399	12
rs1432679	5	rs1292011	12
rs10941679	5	rs11571833	13
rs889312	5	rs2236007	14
rs30099	5	rs2588809	14
rs981782	5	rs941764	14
rs10069690	5	rs999737	14
rs11242675	6	rs13329835	16
rs204247	6	rs17817449	16
rs2046210	6	rs3803662	16
rs2180341	6	rs12443621	16
rs17530068	6	rs8051542	16
rs3757318	6	rs6504950	17
rs720475	7	rs1436904	18
rs11780156	8	rs527616	18
rs2943559	8	rs3760982	19
rs6472903	8	rs4808801	19
rs9693444	8	rs8170	19
rs13281615	8	rs2284378	20
rs10759243	9	rs2823093	21
rs1011970	9	rs132390	22
rs865686	9	rs6001930	22
rs11199914	10		

Table 2: The 77 SNPs identified to be associated with breast cancer

2.1.3 Genetic variants

We decided to focus on high-frequency/low-penetrance SNPs that affect breast cancer risk as opposed to low frequency SNPs with high penetrance or intermediate penetrance. We consolidated a list of 77 common genetic variants (Table 2) which were identified by recent large-scale GWAS studies or used to generate published predictive models (Liu et al., 2014). The list included 41 SNPs identified by COGS through a meta-analysis of 9 GWAS studies (Michailidou et al., 2013). Recently, a similar set of 77 breast cancer-associated SNPs is also studied for risk prediction (Mavaddat et al., 2015).

2.2 Logistic Regression

Assume that we have independent and identical distributed subjects $\{(x_i, y_i)\}_{i=1}^n$, where the explanatory variable $X \in \mathcal{R}^d$ and the binary response variable $Y \in \{-1, 1\}$. Note that the conditional probability $\eta(x) = \mathbb{P}(Y = 1|X = x)$ plays an important role in the classification problem. Denote $x_i = (x_{i1}, ..., x_{id})^T$, and linear logistic regression model is defined by

$$\log \frac{\eta(x_i)}{1 - \eta(x_i)} = x_i^T \beta, \quad i = 1, ..., n,$$

where $\beta = (\beta_1, ..., \beta_d)^T$ is the slope parameter. And the logistic regression estimator $\hat{\beta}$ is given by the minimizer of the negative log-likelihood function

$$L(\beta) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i \cdot x_i^T \beta)). \tag{1}$$

With $\hat{\beta}$, we then estimate the conditional probability $\eta(x_i)$ by

$$\hat{\eta}(x_i) = \frac{\exp(x_i^T \hat{\beta})}{1 + \exp(x_i^T \hat{\beta})} = \frac{1}{1 + \exp(-x_i^T \hat{\beta})}.$$

Then we should predict $y_i = 1$ if $\hat{\eta}(x_i) \geq 0.5$ and $y_i = -1$ if $\hat{\eta}(x_i) < 0.5$.

2.3 Group Penalty and ℓ^p Fusion Penalty

Note that there exist natural group structure and dependence structure in mammography features (Figure 1), which allows us to include the structure information into our risk prediction models directly. For genetic variants, group structures also exist (Liu et al., 2012, 2013). In this paper, we apply hierarchical clustering to cluster the 77 SNPs based on their dissimilarity matrix obtained by computing Spearman's correlation or Hamming distance among them. More details are provided in Section 2.5.

Suppose that d features are divided into G groups with d_g the number of features in group g. Define $\beta_g \in \mathbb{R}^{d_g}$ to be the corresponding coefficient vector in group g. The group lasso logistic regression (Meier et al., 2008) is defined as the following optimization problem

$$\min_{\beta \in \mathbb{R}^d} \left\{ L(\beta) + \lambda_1 \sum_{g=1}^G \sqrt{d_g} \|\beta_g\|_2 \right\},$$

where $L(\beta)$ is defined by (1) and $\lambda_1 \geq 0$ is the tuning parameter. It includes lasso as a special case with $G = d$.

The fact that there exist dependence structure within each mammography feature group and each SNP group encourages us to propose the following novel method by combining group lasso logistic regression and ℓ^p fusion penalty.

$$\min_{\beta \in \mathbb{R}^d} \left\{ L(\beta) + \sum_{g=1}^{G} \left(\lambda_1 \sqrt{d_g} \|\beta_g\|_2 + \lambda_2 \|D_g \beta_g\|_p^p \right) \right\}, \tag{2}$$

where D_g is a $(d_g - 1) \times d_g$ sparse matrix with only $D[i,i] = 1$ and $D[i, i+1] = -1$, $\lambda_2 \geq 0$ is the tuning parameter, and $1 \leq p \leq 2$ is the shrinkage parameter.

Moreover, if the within-group dependence structures are different for groups $\{1, ..., G_1\}$ and $\{G_1 + 1, ..., G\}$, we can split the ℓ^p fusion penalty into two parts as

$$\min_{\beta \in \mathbb{R}^d} \left\{ L(\beta) + \lambda_1 \sum_{g=1}^{G} \sqrt{d_g} \|\beta_g\|_2 + \lambda_2 \left(\sum_{g=1}^{G_1} \|D_g \beta_g\|_{p_1}^{p_1} + \sum_{g=G_1+1}^{G} \|D_g \beta_g\|_{p_2}^{p_2} \right) \right\}, \tag{3}$$

where $1 \leq p_1, p_2 \leq 2$ are selected based on cross validation.

The novelty of our method compared to previous works is three-fold: First, it includes within-group fusion penalty in the model and makes the coefficients of features in the same group close to each other, which reflects the dependence structure of features and improves the risk prediction; Second, in breast cancer risk prediction, we find that the dependence structures are different for mammography features and SNPs, which are actually two different views of the same data. And the utilization of method (3) will improve the predictive performance further; At last, we find that genetic variants improve risk prediction on mammography features, which provides some insight regarding personalized breast cancer diagnosis.

2.4 Computational Algorithms

Many algorithms have been proposed in the literatures to solve the logistic regression with fused lasso regularization (Lin, 2015; Yu et al., 2015). In this subsection we adopt the fast iterative shrinkage thresholding algorithm (Beck and Teboulle, 2009) to solve (2) as

$$\beta^{k+1} = \arg\min_{\beta \in \mathbb{R}^d} L(\beta^k) + \langle \beta - \beta^k, \nabla L(\beta^k) \rangle + \frac{\tau}{2} \|\beta - \beta^k\|_2^2 + \sum_{g=1}^{G} \left(\lambda_1 \sqrt{d_g} \|\beta_g\|_2 + \lambda_2 \|D_g \beta_g\|_p^p \right)$$

with $\beta = (\beta_1, \cdots, \beta_d)^T$ and $\tau > 0$ the Lipschitz constant of $L(\cdot)$.

And the iteration step is equivalent to solving

$$\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{2} \|\beta - (\beta^k - \frac{1}{\tau} \nabla L(\beta^k))\|_2^2 + \sum_{g=1}^{G} \left(\frac{\lambda_1 \sqrt{d_g}}{\tau} \|\beta_g\|_2 + \frac{\lambda_2}{\tau} \|D_g \beta_g\|_p^p \right) \right\}. \tag{4}$$

Therefore, it suffices to solve the following optimization problem within each group

$$\min_{\beta_g \in \mathbb{R}^{d_g}} \left\{ \frac{1}{2}\|\beta_g - z\|_2^2 + \rho_1\|\beta_g\|_2 + \rho_2\|D_g\beta_g\|_p^p \right\}, \tag{5}$$

where $z = \beta_g^k - \frac{1}{\tau}\nabla L(\beta_g^k)$, $\rho_1 = \frac{\lambda_1\sqrt{d_g}}{\tau}$ and $\rho_2 = \frac{\lambda_2}{\tau}$.

The proximity operator (Polson et al., 2015) of a function f is defined as

$$P_f(z) = \arg\min_t \left\{ \frac{1}{2}\|t - z\|^2 + \lambda f(t) \right\}.$$

- For $f(t) = |t|$ and $z \in \mathbb{R}$, $P_f(z) := S_1(z, \lambda) = sign(z)\max\{|z| - \lambda, 0\}$, which is also called soft threshold operator.

- For $f(t) = |t|^p$ with $1 < p \leq 2$ and $z \in \mathbb{R}$, $P_f(z) := S_p(z, \lambda) = sign(z)\xi$, where ξ is the unique nonnegative solution to $\xi + p\lambda\xi^{p-1} = |\xi|$. In particular, we have $S_2(z, \lambda) = \frac{z}{2\lambda+1}$, $S_{3/2}(z, \lambda) = z + 9\lambda^2 sign(z)(1 - \sqrt{1 + 16|z|/(9\lambda^2)})/8$ and $S_{4/3}(z, \lambda) = z + \frac{4\lambda}{32^{\frac{1}{3}}}((\chi - z)^{1/3} - (\chi + z)^{1/3})$ with $\chi = \sqrt{z^2 + 256\lambda^3/729}$.

- For $f(t) = \|t\|_2$ and $z \in \mathbb{R}^d$, $P_f(z) := S_{2,1}(z, \lambda) = \max\{1 - \frac{\lambda}{\|z\|_2}, 0\} * z$.

With the help of these proximity operators and Bregman splitting algorithm (Ye and Xie, 2011), we can solve (5) by iteratively solving the following procedures:

$$\begin{cases} \beta^{k+1} = \arg\min_{\beta_g} \frac{1}{2}\|\beta_g - z\|_2^2 + \langle u^k, \beta_g - a^k\rangle + \langle v^k, D_g\beta_g - b^k\rangle \\ \qquad\qquad + \frac{\mu}{2}\|\beta_g - a^k\|_2^2 + \frac{\mu}{2}\|D_g\beta_g - b^k\|_2^2 \\ a^{k+1} = \arg\min_a \rho_1\|a\|_2 + \langle u^k, \beta^{k+1} - a\rangle + \frac{\mu}{2}\|\beta^{k+1} - a\|_2^2 \\ b^{k+1} = \arg\min_b \rho_2\|b\|_p^p + \langle v^k, D_g\beta^{k+1} - b\rangle + \frac{\mu}{2}\|D_g\beta^{k+1} - b\|_2^2 \\ u^{k+1} = u^k + \mu(\beta^{k+1} - a^{k+1}) \\ v^{k+1} = v^k + \mu(D_g\beta^{k+1} - b^{k+1}) \end{cases}$$

where μ acts like a step size in this algorithm.

Remark 1 *The minimization over β, a and b can all be solved in closed form.*

- $\beta^{k+1} = [(\mu + 1)I + \mu D_g^T D_g]^{-1}[z + \mu(a^k - u^k/\mu) + \mu D_g^T(b^k - v^k/\mu)]$

- $a^{k+1} = S_{2,1}(\beta^{k+1} + u^k/\mu, \rho_1/\mu)$

- $b^{k+1} = S_p(D_g\beta^{k+1} + v^k/\mu, \rho_2/\mu)$

Note that $(\mu + 1)I + \mu D_g^T D_g$ is a tridiagonal positive definite matrix.

Remark 2 *For $p = 1$, we can solve (5) more efficiently by the algorithm proposed in Zhou et al. (2012) based on the fact*

$$P_{\|\cdot\|_2 + \|D_g(\cdot)\|_1} = P_{\|\cdot\|_2} \circ P_{\|D_g(\cdot)\|_1}.$$

However, we cannot show this equation for $1 < p \leq 2$.

Remark 3 *For $p = 2$, since $\|\cdot\|_2^2$ is Lipschitz continuous, we can rewrite (4) as*

$$\min_{\beta\in\mathbb{R}^d}\left\{\frac{1}{2}\left\|\beta-\left(\beta^k-\frac{1}{\tilde{\tau}}(\nabla L(\beta^k)+2\lambda_2\sum_{g=1}^{G}D_g^T D_g\beta_g^k)\right)\right\|_2^2+\sum_{g=1}^{G}\frac{\lambda_1\sqrt{d_g}}{\tilde{\tau}}\|\beta_g\|_2\right\},$$

where $\tilde{\tau}$ is the Lipschitz constant of $L(\beta)+\lambda_2\sum_{g=1}^{G}\|D_g\beta_g\|_2^2$. Then we can solve it efficiently via the proximity operator of $\|\cdot\|_2$.

2.5 Study Design and Statistical Analysis

We apply the ℓ^p fused group lasso logistic regression algorithm to the Marshfield breast cancer data set. There are 11 groups for 49 mammography features (Figure 1). For SNPs, we compute the Hamming distances (Wang et al., 2015) of 77 SNPs to get the dissimilarity matrix and then apply hierarchical clustering to obtain 10 groups.

We built three prediction models based on different sets of risk factors: the Mammo model developed by using mammography features only, the SNP77 model developed by using 77 SNPs only, and the Combined model developed by using both mammography features and 77 SNPs. We furthermore apply five methods for each model: logistic regression (LR), lasso in logistic regression (LR+Lasso), ℓ^p fused lasso logistic regression (LR+fusedLasso), group lasso logistic regression (LR+groupLasso), and ℓ^p fused group lasso logistic regression (LR+Structure).

The ℓ^p fused group lasso logistic regression method has several parameters. For the tuning parameters λ_1 and λ_2, we let them vary among a given set of values, and the shrinkage parameter p (or p_1 and p_2) among $\{1, 4/3, 3/2, 2\}$. Each combination of these parameters is evaluated using stratified 5-fold cross-validation, and AUC (the area under the receiver operating characteristic (ROC) curve) is used as the performance measure. All 738 samples are randomly partitioned into five equal sized folds with approximately equal proportions of cases and controls. In each iteration (totally five iterations), four folds are used as training set and the rest one as validation set to compute AUC. And the parameters with the best average AUC are selected. At last we repeat this process ten times and report the average AUC. We obtain p-value by performing two-tailed-two-sample t-test when we compare AUCs.

3. Experimental Results

In this section, we demonstrate the performance of the ℓ^p fused group lasso logistic regression method from three aspects: the significant improvement of AUCs by considering the structure information, the predictive performance under different p (or p_1 and p_2), and the detected important mammography features and SNPs.

3.1 Performance of Fused Group Lasso

The result is summarized in Table 3.

Models/Methods	LR	Lasso	fusedLasso	groupLasso	Structure	p-value
Mammo	0.700	0.710	0.710	0.716	0.723	< 0.001
SNP77	0.590	0.598	0.676	0.614	0.684	< 0.001
Combined	0.697	0.721	0.754	0.727	0.766	< 0.001

Table 3: Predictive performance of three prediction models by using five different methods. The p-values represent the differences between AUCs of LR and LR+Structure.

1) The fifth column describes the predictive performance of the three prediction models by considering structure information in the logistic regression method. We find that the predictive performance of the three prediction models has been improved respectively, compared to those described in the first column. For each prediction model, the difference of the predictive performance is significant between LR+Structure and LR (p-value < 0.001), which demonstrates that breast cancer prediction models utilizing structure information can improve risk prediction significantly. We also find that mammography descriptors demonstrate a significantly higher predictive performance than 77 SNPs in terms of AUC (0.723 vs. 0.684, p-value < 0.001). The Combined model demonstrates significant improvement of the prediction performance, compared to the Mammo model (0.766 vs. 0.723, p-value < 0.001).

2) The first column describes the predictive performance of the three prediction models by using the logistic regression method. Mammography descriptors demonstrate a significantly higher predictive performance than 77 SNPs in terms of AUC (0.700 vs. 0.590, p-value< 0.001). We find that the difference of predictive performance between the Combined model and the Mammo model is negligible (0.697 vs. 0.700, p-value=0.277).

3) The second column describes the predictive performance of the three prediction models by using lasso in the logistic regression method. The predictive performance of the three prediction models has been improved, compared to those without lasso (using logistic regression method only). Mammography descriptors still demonstrate a significantly higher predictive performance than 77 SNPs in terms of AUC (0.710 vs. 0.598, p-value< 0.001). However, the Combined model demonstrates modest improvement of prediction performance, compared to the Mammo model (0.721 vs. 0.710, p-value=0.0057).

4) The third and fourth columns describe the predictive performance of the three prediction models by considering group structure or dependence structure in the logistic regression method. For the SNP77 model, fused lasso demonstrates a significantly higher performance than group lasso in terms of AUC (0.676 vs. 0.614, p-value< 0.001). For the Mammo model, group lasso plays a more important role than fused lasso (0.716 vs. 0.710, p-value=0.0073). Moreover, both fused lasso and group lasso demonstrate improved prediction performance compared to lasso.

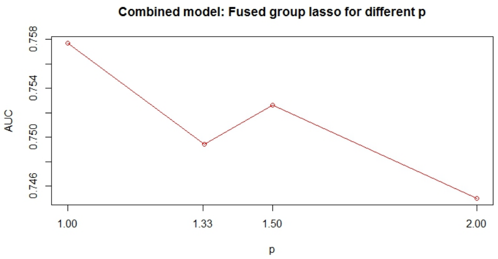
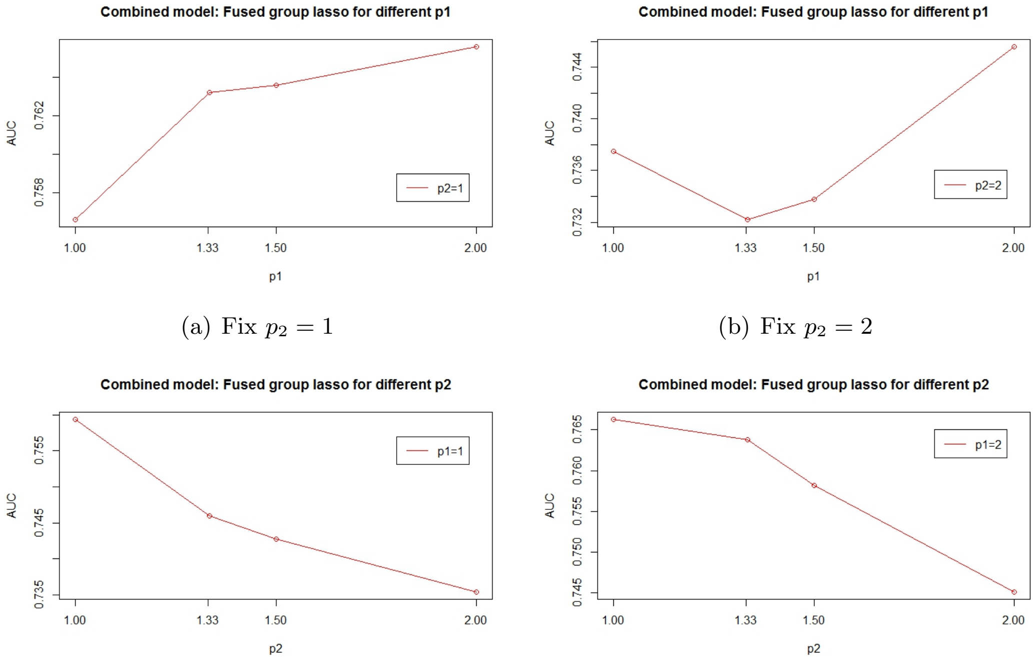


Figure 2: The AUCs under different values of p by using method (2).



(a) Fix $p_2 = 1$

(b) Fix $p_2 = 2$

(c) Fix $p_1 = 1$

(d) Fix $p_1 = 2$

Figure 3: The AUCs under different values of p_1 and p_2 by using method (3).

3.2 Performance under Different Values of p

Figure 2 and Figure 3 describe the the pattern of predictive performance for ℓ^p fused group lasso logistic regression over the shrinkage parameter p (or p_1 and p_2) in terms of AUC.

1)The Combined model demonstrates a higher predictive performance for $p = 1$ compared to $p = 2$ in terms of AUC (0.757 vs. 0.745, p-value< 0.001), see Figure 2.

2) Figure 3 describes the prediction performance of method (3) under different values of p_1 for mammography descriptors and p_2 for 77 SNPs.

- Fix $p_2 = 1$ or $p_2 = 2$, the fused group lasso with $p_1 = 2$ demonstrates higher predictive performance compared to $p_1 = 1$, see Figure 3(a) and 3(b).
- Fix $p_1 = 1$ or $p_1 = 2$, the predictive performance of the fused group lasso logistic regression decreases as p_2 increases, see Figure 3(c) and 3(d).
- The fused group lasso logistic regression with $p_1 = 2$ and $p_2 = 1$ demonstrates higher predictive performance than $p_1 = p_2 = 1$ (0.766 vs. 0.757, p-value=0.0053) and $p_1 = p_2 = 2$ (0.766 vs. 0.745, p-value$<$ 0.001).

3.3 Important Features Detected by Fused Group Lasso

To take into account both group and dependence structure information in mammography features and SNPs, two penalty terms (group penalty and fusion penalty) are introduced into the logistic regression model. The idea of group penalty is to force the coefficients of features in the same group to be all zero or nonzero in order to achieve the goal of selecting features within a group simultaneously. The idea of fusion penalty is to shrink the successive difference of coefficients of features in the same group in order to take advantage of the dependence structure information. Applying fusion penalty with $p = 1$ tends to result in zero successive difference of coefficients, while $p = 2$ tends to small but nonzero successive difference of coefficients.

From a feature selection point of view, we can get the order of feature groups selected by fused group lasso via choosing the tuning parameters appropriately. We list below the feature groups selected from high to low in terms of predictive performance.

1) For mammography descriptors, the following features are predictive of malignancy (from most to least): “Mass Size”, “Mass Margins”, “Mass Shape”, “Architectural Distortion” and “Mass Palpability”, consistent with the literature (BI-RADS, 2014).

2) For 77 SNPs, three groups are selected in order, see Table 4.

Feature Group	SNPs
Group 1	rs2016394 rs1432679, rs13281615, rs4666451 rs981782, rs1292011, rs1436904, rs527616
Group 2	rs11249433 rs13387042, rs4973768, rs10069690 rs7904519, rs8051542, rs3760982
Group 3	rs2981579, rs2981582

Table 4: SNP groups selected by fused group lasso.

Remark 4 *It verifies that “Mass size”, “Mass Margins” and “Mass Shape” are the most important mammography descriptors in breast cancer diagnosis. These results are consistent with previous studies about comparing the importance of mammography features and SNPs in breast cancer risk prediction(Wu et al., 2013, 2014).*

4. Discussion and Conclusions

This study demonstrates that models utilizing the novel combination of clinically relevant structure and ℓ^p fused group lasso logistic regression can improve breast cancer risk prediction significantly. Our study also shows that both mammography features and SNPs contribute to this improvement.

The structure information of the mammography features is derived from the BI-RADS lexicon, which is used widely in breast imaging practice. Thus, our model would likely be generalizable to other practices. On the other hand, we extracted the structure information of SNPs by computing Hamming distances (Wang et al., 2015). This method may not perform as well in small sample sizes, which may affect our results perhaps making our predictive performance results conservative.

Our methods for SNPs may not take advantage of biological knowledge that currently exists. For example, it may be possible to utilize the biological information available in HapMap (which encodes linkage disequilibrium) to more accurately emulate the patterns or dependence structure of SNPs, as in (Liu et al., 2012). Furthermore, we realize that taking into account more complicated structure information such as graph or tree structure (Sun and Wang, 2012) may further improve predictive performance of risk prediction models. We leave these promising directions for future work.

In conclusion, our results demonstrate that including structure information in the computational methods we test improves breast cancer risk prediction. Our models use diverse breast cancer risk factors including demographics, genetics, and imaging and leverage structure found in a standardized lexicon that is universally captured in electronic health records (EHRs) throughout the US. This information will increasingly be combined in complex ways. Merging imaging features, clinical notes and genetic data with models that accurately predict disease risk has the potential to provide powerful knowledge to practicing physicians.

Acknowledgments

The authors acknowledge the support of the Wisconsin Genomics Initiative, NCI grant R01CA127379-01 and its ARRA supplement R01CA127379-03S1, NIGMS grant R01GM097618-01, NLM grant R01LM011028-01, NIEHS grant 5R01ES017400-03, NIH grant 1U54AI117924-01, NIH grant K24CA194251 and NSF FRG grant DMS-1265202. We also acknowledge support from the Clinical and Translational Science Award (CTSA) program through the NIH National Center for Advancing Translational Sciences (NCATS) grant UL1TR000427 and the University of Wisconsin Carbone Cancer Center Cancer Center Support Grant P30CA014520. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We thank the anonymous reviewers for their valuable comments and suggestions.

References

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183-202, 2009.

Breast Imaging Reporting And Data System (BI-RADS). 5th ed. Reston VA: American College of Radiology, 2014.

E. S. Burnside, J. Liu, Y. Wu, A. A. Onitilo, C. A. McCarty, C. D. Page, P. Peissig, A. Trentham-Dietz, T. Kitchner, J. Fan, and M. Yuan. Comparing Mammography Abnormality Features and Genetic Variants in the Prediction of Breast Cancer in Women Recommended for Breast Biopsy. *Academic Radiology*, 23(1):62-9, 2016.

J. Ferlay, I. Soerjomataram, M. Ervik, et al. *GLOBOCAN 2012 cancer incidence and mortality worldwide: IARC cancerbase No. 11.* Lyon, France: International Agency for Research on Cancer, 2013.

D. Freedman, D. Petitti, and J. Robins. On the efficacy of screening for breast cancer. *Int J Epidemiol.*, 33(1):43-55, 2004.

M. Gail. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J Natl Cancer Inst.*, 100(14):1037-41, 2008.

M. Gail. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *J Natl Cancer Inst.*, 101(13):959-63, 2009.

T. Lin, S. Ma, and S. Zhang. An extragradient-based alternating direction method for convex minimization. *Found Comput Math*, DOI 10.1007/s10208-015-9282-8, 2015.

J. Liu, J. Huang, S. Ma, and K. Wang. Incorporating group correlations in genome-wide association studies using smoothed group lasso. *Biostatistics*, 14:205-219, 2013.

J. Liu, C. D. Page, P. L. Peissig, et al. New genetic variants improve personalized breast cancer diagnosis. *AMIA Summit on Translational Bioinformatics (AMIA-TBI)*, 2014.

J. Liu, C. Zhang, C. McCarty, P. L. Peissig, E. S. Burnside, and D. Page. Graphical-model based multiple testing under dependence, with applications to genome-wide association studies. In *Proceedings of the 28th conference on uncertainty in artificial intelligence*, 2012.

S. Ma, X. Song, and J. Huang. Supervised group Lasso with applications to microarray data analysis. *BMC bioinformatics*, 8:60, 2007.

N. Mavaddat, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst*, 107(5):djv036, 2015.

C. McCarty, R. Wilke, P. Giampietro, S. Wesbrook, and M. Caldwell. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Personalized Med.*, 2(1):49-79, 2005.

L. Meier, S. Van De Geer, and P. Bohlmann. The Group Lasso for logistic regression. *J. R. Statist. Soc. B*, 70:53-71, 2008.

K. Michailidou, P. Hall, A. Gonzalez-Neira, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet.*, 45(4):353-61, 2013.

H. Nassif, R. Woods, E. S. Burnside, M. Ayvaci, J. Shavlik, C. D. Page. Information extraction for clinical data mining: a mammography case study. *IEEE International Conference on Data Mining Workshops*, 2009.

N. G. Polson, J. G. Scott, and B. T. Willard. Proximal Algorithms in Statistics and Machine Learning. *Statistical Science*, 30(4):559-581, 2015.

H. Sun and S. Wang. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics*, 28(10):1368-1375, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267-288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, 67:91-108, 2005.

R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18-29, 2008.

S. Wacholder, P. Hartge, R. Prentice, et al. Performance of common genetic variants in breast-cancer risk models. *N Engl J Med.*, 362(11):986-93, 2010.

C. Wang, W. H. Kao, and C. K. Hsiao. Using Hamming distance as information for SNP-sets clustering and testing in disease association studies. *PLoS One*, 10(8), 2015.

Y. Wu, O. Alagoz, M. Ayvaci, A. M. del Rio, D. J. Vanness, R. Woods, and E. S. Burnside. A comprehensive methodology for determining the most informative mammographic features. *J Digit Imaging*, 26(5):941-947, 2013.

Y. Wu, J. Liu, C. D. Page, P. L. Peissig, C. A. McCarty, A. A. Onitilo, and E. S. Burnside. Comparing the Value of Mammographic Features and Genetic Variants in Breast Cancer Risk Prediction. *AMIA Annu Symp Proc.*, 1228-1237, 2014.

G. Ye and X. Xie. Split Bregman method for large scale fused Lasso. *Computational Statistics and Data Analysis*, 55(4):1552-1569, 2011.

D. Yu, S. Lee, W. Lee, S. Kim, J. Lim, and S. Kwon. Classification of spectral data using fused lasso logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 142:70-77, 2015.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68:49-67, 2006.

J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In *KDD*, pages 1095-1103, 2012.