Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–15

DOI: 10.1177/0962280216677748
smm.sagepub.com

SAGE

Variable selection for mixture and promotion time cure rate models

Abdullah Masud,[1] Wanzhu Tu[1] and Zhangsheng Yu[2]

Abstract

Failure-time data with cured patients are common in clinical studies. Data from these studies are typically analyzed with cure rate models. Variable selection methods have not been well developed for cure rate models. In this research, we propose two least absolute shrinkage and selection operators based methods, for variable selection in mixture and promotion time cure models with parametric or nonparametric baseline hazards. We conduct an extensive simulation study to assess the operating characteristics of the proposed methods. We illustrate the use of the methods using data from a study of childhood wheezing.

Keywords

Mixture cure rate model, promotion time cure rate model, adaptive least absolute shrinkage and selection operators, expectation-maximization algorithm, Bayesian information criterion, wheeze

1 Introduction

Standard survival models, such as the frequently used Cox regression models, assume that all subjects are susceptible to the event of interest, and that all subjects will eventually experience the outcome if the follow-up is long enough.[1] Data from some applications, however, contradict the notion that all subjects are at risk. In practice, analysts deal with the situation by treating the risk-free subjects as "cured". Compared to the non-cured, the cured tend to have much extended survival times, as indicated by long flat tails and heavy right censoring in Kaplan-Meier curves.[2]

Data with such characteristics are abundant in clinical studies. For instance, childhood wheezing, an airway symptom defined by a coarse or whistling breathing sound, tends to occur only in certain children, while others never exhibit wheezing symptoms in early years of life.[3] Data from our own studies showed that Kaplan-Meier curves of the onset age of wheezing essentially flattened after the first 48 and 32 months of life in girls and boys, and thus confirming the existence of risk-free subgroups (see Figure 1). Data with similar features are also seen in immuno-oncological studies.[4]

Cure rate models are standard techniques for such data. Traditional cure rate models assume that the population consists of both cured and non-cured subjects.[5] The standard formulation is a mixture of logistic regression and survival analysis, with the former quantifying the cured portion and the latter depicting the event time distribution of the non-cured.[6] This mixture has been the basis of several model extensions.[2,7,8] A more biology-motivated approach is the promotion time cure model, proposed by Yakovlev et al. in the context of cancer recurrence.[9] Briefly, Yakovlev's model assumes that cancer recurrence is promoted by carcinogenic cells that remain active after treatment. So the unobserved number of carcinogenic cells is incorporated into the analysis through a Poisson model. This line of models has been further extended by others, mostly in the Bayesian framework (Chen et al.,[10–12] Chen and Ibrahim,[13] and Tsodikov et al.[14]). The two different modeling approaches have been compared by a number of authors.[15,16]

[1]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, USA
[2]Department of Bioinformatics and Biostatistics, Shanghai Jiaotong University, Shanghai, China

Corresponding author:
Zhangsheng Yu, Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, USA.
Email: yuzhangsheng@sjtu.edu.cn

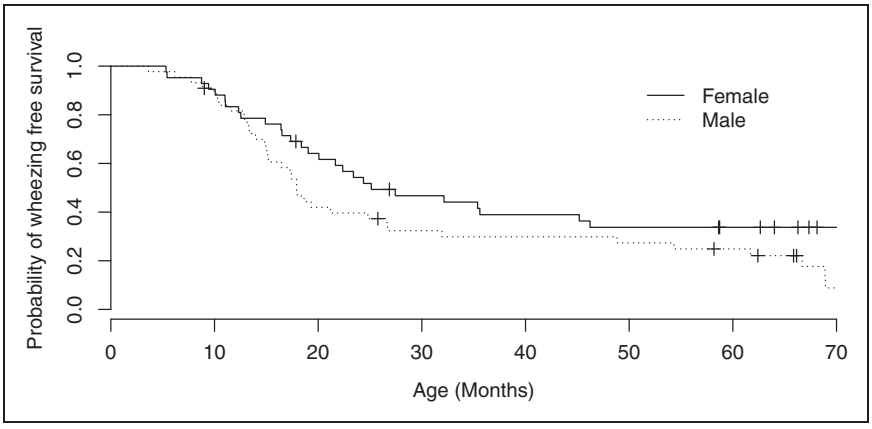

Figure 1. Kaplan-Meier estimates of wheezing-free probabilities in male and female subjects.

Regardless of one's modeling preference, a common challenge faced by analysts is to select the right independent variables for the intended model. With the complex structures of cure rate models, variable selection is certainly not a trivial exercise. Among other things, traditional stepwise procedures often lack the desired stability.[17] Following Tibshirani's works on the least absolute shrinkage and selection operators (LASSO),[18,19] penalize likelihood-based regularization methods have been developed for variable selection in frequently used statistical models, including the traditional Cox regression models.[1] Theoretically, some of these methods have been shown to possess the oracle properties.[20,21] Most recently, attempts have been made to extend the LASSO-based selection approach to joint models of longitudinal and survival outcomes.[22] The successful use of LASSO in complex models points to the plausibility of a similar application in the cure rate models.

Literature on variable selection in cure rate models is relatively sparse. One notable piece of work in this field is by Liu et al.[23] who proposed to use LASSO with a smoothly clipped absolute deviation (SCAD) penalty to select variables for the mixture cure rate model (MCM). The non-convex form of the SCAD penalty, however, tends to increase the difficulty of parameter estimation. As a result, estimators often lack numerical stability.[21] Alternatively, Zou proposed an adaptive LASSO method with L_1 penalty, which is computationally more stable in comparison with SCAD.[24]

In this research, we discuss variable selection in mixture and promotion time cure models using LASSO and adaptive LASSO. To the best of our knowledge, this is the first study of its kind, especially for the promotion time cure model. We compare the selection performance of LASSO and adaptive LASSO. The methods are easily implementable using an expectation-maximization (EM) algorithm, with generally consistent performance. An extensive simulation study is conducted to evaluate the operational characteristics of the procedures in both modeling settings. Finally, we apply the methods to select variables for a mixture cure model using data from a study of childhood wheezing.

2 Models and estimations

2.1 Mixture cure rate model

2.1.1 Model

Let $\widetilde{T}_i$ and C_i be the respective failure time and censoring time for the ith subject, $i = 1, 2, \ldots, n$. The observed time is $T_i = \min(\widetilde{T}_i, C_i)$. We assume that the censoring time C_i is random and noninformative. We define the failure time indicator as $\delta_i = 1$ if $\widetilde{T}_i \leq C_i$ (T_i is observed), and $\delta_i = 0$ otherwise. Let $Y_i = 1$ be a binary indicator for the non-cured, and $P(Y_i = 1) = \theta(\cdot)$. We write the independent variable vectors for the logistic and survival components as $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{z}_i \in \mathbb{R}^q$, respectively; and vectors $\mathbf{x}_i$ and $\mathbf{z}_i$ may share common elements. Under such a notation, the population survival function $S_p(t)$ can be written as

$$S_p(t) = \{1 - \theta(\mathbf{x}_i)\} + \theta(\mathbf{x}_i) S_{nc}(t|\mathbf{z}_i) \tag{1}$$

where $S_{nc}(t|\mathbf{z}_i)$ is the survival function of the non-cured, given $\mathbf{z}_i$. As t increases, $S_p(t) \to \{1 - \theta(\mathbf{x}_i)\} > 0$. We note that $S_p(.)$ may not be a proper survival function.

With a logit link function in the MCM, Farewell[6] described the effects of independent variables $\mathbf{x}$ on the probability of not being cured as

$$\theta(\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients for $\mathbf{x}_i$. For the non-cured, the Cox proportional hazards (PH) model can be written as $\lambda_{nc}(t|\mathbf{z}_i) = \lambda_{nc,0}(t)e^{\gamma^T\mathbf{z}_i}$, where $\boldsymbol{\gamma}$ is the coefficient vectors for $\mathbf{z}_i$, and $\lambda_{nc,0}(t)$ is the baseline hazard. The cumulative baseline hazard function is $\Lambda_{nc,0}(t) = \int_0^t \lambda_{nc,0}(u)du$. The independent variable effects for the non-cured in Model (1) are interpreted in a way similar to that in the traditional Cox models.

2.1.2 Variable selection and estimation

For simplicity, we denote the observed data from the ith subject as $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$. The likelihood function of model (1) is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_{nc,0}) = \prod_{i=1}^{n}\left\{\left\{\theta(\mathbf{x}_i)\lambda_{nc,0}(t_i)e^{\gamma^T\mathbf{z}_i}S_{nc,0}(t_i)^{e^{\gamma^T\mathbf{z}_i}}\right\}^{\delta_i} \times \left\{1-\theta(\mathbf{x}_i)\left(1-S_{nc,0}(t_i)^{e^{\gamma^T\mathbf{z}_i}}\right)\right\}^{1-\delta_i}\right\} \tag{2}$$

Estimation of the nonparametric baseline hazard is needed to maximize (2). Here we use an EM algorithm to maximize the complete likelihood based on $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i, y_i)$, by treating y_i as a latent binary variable. The complete likelihood includes a logistic component for the cured, and a PH component for the non-cured. We write

$$L_C(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_{nc,0}; y) = \prod_{i=1}^{n}\left[\theta(\mathbf{x}_i)^{y_i}(1-\theta(\mathbf{x}_i))^{1-y_i}\right]\prod_{i=1}^{n}\left[\left\{\lambda_{nc,0}(t_i)exp(\boldsymbol{\gamma}^T\mathbf{z}_i)\right\}^{\delta_i}S_{nc,0}(t_i)^{e^{\gamma^T\mathbf{z}_i}}\right]^{y_i}$$

The log-likelihood is

$$\begin{aligned} l_C(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_{uc,0}; y) = &\sum_{i=1}^{n}\left\{y_i\boldsymbol{\beta}^T\mathbf{x}_i - \log\{1+\exp(\boldsymbol{\beta}^T\mathbf{x}_i)\}\right\} \\ &+ \sum_{i=1}^{n}\left\{y_i\delta_i\{\log\{\lambda_{nc,0}(t_i)\} + \boldsymbol{\gamma}^T\mathbf{z}_i\} + y_i\exp(\boldsymbol{\gamma}^T\mathbf{z}_i)log\{S_{nc,0}(t_i)\}\right\} \end{aligned} \tag{3}$$

For simplicity, we write the first term of the above equation as $l_{C,1}(\boldsymbol{\beta}; y)$, and second term as $l_{C,2}(\boldsymbol{\gamma}, \lambda_{nc,0}; y)$. To allow for sparse estimation, we use an adaptive LASSO and impose an L_1 norm penalty on the log likelihood:

$$pl_c(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_{nc,0}; y) = \left\{l_{C,1}(\boldsymbol{\beta}; y) - \tau_1\sum_{j=1}^{p}\frac{|\boldsymbol{\beta}_j|}{|\rho_{1,j}|}\right\} + \left\{l_{C,2}(\boldsymbol{\gamma}, \lambda_{nc,0}; y) - \tau_2\sum_{k=1}^{q}\frac{|\gamma_k|}{|\rho_{2,k}|}\right\} \tag{4}$$

where $\rho_{1,j}$ and $\rho_{2,k}$ are the weight parameters, and τ_1 and τ_2 are the tuning parameters controlling the amount of penalty. Values of the tuning parameters can be determined either by cross-validation or by the Bayesian information criteria (BIC). We discuss the selection of tuning parameters later in the section.

Following Zou,[24] we use consistent estimators of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ as the weight parameters (ρ_1, ρ_2). The closer the true estimate to 0, the greater the penalty. As a result, factors with smaller coefficients are more likely to be excluded from the model. The adaptive LASSO essentially shrinks the less important effects to zeros, and thus achieving a more parsimonious model. When ρ_1 and ρ_2 are set to the 1, the method leads to LASSO estimators proposed by Liu et al.[23] In this research, we estimate ρ_1 and ρ_2 by maximizing (3).

Computation: For computation, we use adaptive LASSO estimates $(\hat{\beta}, \hat{\gamma})$ and the quadratic approximation algorithm.[20]

E-step: Let $(\boldsymbol{\beta}^{(m)}, \boldsymbol{\gamma}^{(m)}, \lambda_{nc,0}^{(m)})$ be the parameter estimates in the mth iteration. Given the observed data $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, in the $(m+1)$ th iteration, we replace y_i in (3) with $y_i^{(m+1)}$

$$y_i^{(m+1)} = \delta_i + (1-\delta_i)\frac{\theta(\mathbf{x}_i)^{(m)}S_{nc,0}^{(m)}(t_i)^{e^{\gamma^{(m)T}\mathbf{z}_i}}}{1-\theta(\mathbf{x}_i)^{(m)}\left\{1-S_{nc,0}^{(m)}(t_i)^{e^{\gamma^{(m)T}\mathbf{z}_i}}\right\}}$$

M-step: With $y^{(m+1)}$ plugged in, we maximize (4) with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_{nc,0})$. The M-step involves the following sub-steps

1. Estimate the cumulative baseline hazard function $\Lambda_{nc,0}(t)$ using a Breslow type estimator.[25] Specifically, the nonparametric estimate for the $(m+1)$ th iteration is

$$\Lambda_{nc,0}^{(m+1)}(t) = \sum_{t_l \le t} \frac{d_l}{\sum_{k^* \in R_l} y_{k^*}^{(m+1)} \exp(\boldsymbol{\gamma}^{(m)T} \mathbf{z}_{k^*})}$$

   where d_l is the number of events at the earliest time point t_l, and R_l is the number of individuals at risk at t_l.
2. Solve the penalized score equation for $\beta^{(m+1)}$ in the logistic model

$$\begin{aligned} 0 = U(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left[y_i^{(m+1)} - \frac{exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \right] \mathbf{x}_i^T - \tau_1 \sum_{j=1}^{p} \frac{\boldsymbol{\beta}_j / |\boldsymbol{\beta}_j^{(m)}|}{|\rho_{1,j}|} \\ &= \nabla l_{C,1}\left(\boldsymbol{\beta}; y^{(m+1)}\right) - \tau_1 \boldsymbol{\beta}^T \psi(\boldsymbol{\beta}^{(m)}, \rho_1) \end{aligned}$$

   where $\psi(\boldsymbol{\beta}^{(m)}, \rho_1) = diag\left\{\frac{1/|\boldsymbol{\beta}_j^{(m)}|}{|\rho_{1,j}|}\right\}, j = 1, 2, \ldots, p,$ and $\nabla l_{C,1}(\boldsymbol{\beta}; y^{(m+1)}) = \frac{\partial}{\partial \boldsymbol{\beta}} l_{C,1}(\boldsymbol{\beta}; y^{(m+1)})$. We obtained the penalty term $\sum_{j=1}^{p} \frac{\boldsymbol{\beta}_j / |\boldsymbol{\beta}_j^{(m)}|}{|\rho_{1,j}|}$ by using a quadratic approximation of the penalized likelihood. The penalized Hessian matrix $H_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is given by $H_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} U(\boldsymbol{\beta})$.
3. Solve the penalized score equation for the survival model with respect to $\boldsymbol{\gamma}^{(m+1)}$ with given $\Lambda_{nc,0}^{(m+1)}(t)$, $\boldsymbol{\beta}^{(m+1)}$

$$\begin{aligned} 0 = U(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} \left[y_i^{(m+1)} \delta_i - y_i^{(m+1)} \exp(\boldsymbol{\gamma}^T \mathbf{z}_i) \Lambda_{nc,0}^{(m+1)}(t_i) \right] \mathbf{z}_i^T - \tau_2 \sum_{k=1}^{q} \frac{\boldsymbol{\gamma}_k / |\boldsymbol{\gamma}_k^{(m)}|}{|\rho_{2,k}|} \\ &= \nabla l_{C,2}\left(\boldsymbol{\gamma}; \lambda_{nc,0}^{(m)}, y^{(m+1)}\right) - \tau_2 \boldsymbol{\gamma}^T \psi\left(\boldsymbol{\gamma}^{(m)}, \rho_2\right) \end{aligned}$$

   where $\psi(\boldsymbol{\gamma}^{(m)}, \rho_2) = diag\left\{\frac{|1/\boldsymbol{\gamma}_k^{(m)}|}{|\rho_{2,k}|}\right\}, k = 1, 2, \ldots, q,$ and $\nabla l_{C,2}\left(\boldsymbol{\gamma}; \lambda_{nc,0}^{(m)}, y^{(m+1)}\right) = \frac{\partial}{\partial \boldsymbol{\gamma}} l_{C,2}\left(\boldsymbol{\gamma} \lambda_{nc,0}^{(m)}, y^{(m+1)}\right)$. We obtained the penalty term $\sum_{k=1}^{q} \frac{\boldsymbol{\gamma}_k / |\boldsymbol{\gamma}_k^{(m)}|}{|\rho_{2,k}|}$ by using a quadratic approximation for the penalized likelihood. The penalized Hessian matrix $H_{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma}$ in the $(m+1)$ th iteration is obtained by $H_{\boldsymbol{\gamma}} = \frac{\partial}{\partial \boldsymbol{\gamma}} U(\boldsymbol{\gamma})$.

The M-step iterates through the above sub-steps until convergence is achieved. The final maximum likelihood (ML) estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ are achieved by iterating between the E and the M steps.

Alternatively, one could use a parametric function for the baseline hazard $\lambda_{nc,0}$ in (3) to simplify process. For a finite partition of follow-up time intervals $0 < s_1 < s_2 < \ldots < s_G$ with $s_G > \max\{t_i : i = 1, 2, \ldots, n\}$ for a prespecified G, one could assume a constant hazard rate $\lambda_{nc,0}(t) = \alpha_g$ for the gth interval. The estimate $\alpha^{(m+1)}$ of α is obtained by maximizing $l_{C,2}(.)$ with respect to α. For $g = 1, 2, \cdots, G$, it is easy to show that $\alpha_g^{(m+1)} = \left[\sum_{s_{g-1} < t_i \le s_g} \delta_i y_i^{(m+1)}\right] \times \left[\left\{\sum_{s_{g-1} < t_i \le s_g} y_i^{(m+1)}(t_i - s_{g-1}) + \sum_{y_i > s_g} y_i^{(m+1)}(s_g - s_{g-1})\right\} \exp(\boldsymbol{\gamma}^{(m+1)} \mathbf{z}_i)\right]^{-1}$. We later evaluate the selection performance of the nonparametric and parametric baseline hazard function in our simulation study.

In summary, the key steps of the EM algorithms are:

Step 1: Fix the tuning parameter $\tau = (\tau_1, \tau_2)$ and initialize $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)}, \lambda_{nc,0}^{(0)}(t))$
Step 2: Execute the E-step and compute $\lambda_{nc,0}(t)$
Step 3: Update the estimates as $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} - H^{-1}(\boldsymbol{\beta}^{(0)}) U(\boldsymbol{\beta}^{(0)})$ for logistic regression and $\boldsymbol{\gamma}^{(1)} = \boldsymbol{\gamma}^{(0)} - H^{-1}(\boldsymbol{\gamma}^{(0)}) U(\boldsymbol{\gamma}^{(0)})$ for survival model
Step 4: Repeat step 2 and 3 until $|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}| \to 0$ and $|\boldsymbol{\gamma}^{(1)} - \boldsymbol{\gamma}^{(0)}| \to 0$

Tuning/regularization parameter selection: Choosing appropriate tuning parameters $\tau = (\tau_1, \tau_2)$ is essential for variable selection. As τ increases, more coefficients shrink to zero.[24] At the same time, estimates of non-zero coefficients are likely to have increased biases.[21] Nishii adopted a generalized information criterion (GIC) to select τ.[26] The GIC type regularization parameter selector takes the form

$$GIC(\tau) = \frac{1}{n}\{l_C + \kappa df_\tau\} \tag{5}$$

where df_τ is the degree of freedom associated with Model (3). We select the combination of τ_1, τ_2 that minimizes equation (5) for a given κ. As κ increases, the size of selected model decreases. When $\kappa = \log(n)$, the GIC-type selector reduces to the traditional BIC[27] selector. To solve for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we use the BIC regularization parameter selector. The BIC selector has been shown to identify the true model consistently,[28] and is asymptotic efficient.[29]

Post-selection inference: Making valid inference in the selected models poses a new set of challenges, which are beyond the scope of the current paper. First, LASSO penalty could introduce biases to parameter estimation. An obvious way to minimize the bias is to fit the selected model without the penalty term. Such a two-stage approach is consistent with the current practice where inferences are based on the selected models, as advocated by standard textbooks.[30] What left unsaid is the conditional nature of the inference. The validity of such inference is clearly contingent upon the goodness of the selected model. Recently, Berk et al. prescribed an attractive solution.[31] They argued that in linear models, one could treat the post-selection inference as one in a multiple comparison situation, by properly accounting for the errors associated with *all* possible sub-models. While the idea is intuitively appealing, its validity in nonlinear models remains to be validated.

Another issue that affects the inference is the estimation of standard errors of the model parameters. Traditionally, asymptotic standard errors are derived from the Hessian matrix of the observed likelihood. With the use of EM algorithm for estimation, one could simply plug in the model parameter estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ into the Hessian matrix. The following formulae are typically used to approximate the covariance estimators of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, respectively, given $\hat{\Lambda}_{nc,0}(t_i)$:

$$H(\hat{\boldsymbol{\beta}}) = -\mathbf{x}_i\left\{\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i})^2}\right\}\mathbf{x}_i^T + \mathbf{x}_i\left\{(1-\delta_i)\times\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i}exp\left\{-\hat{\Lambda}_{nc,0}exp(\hat{\boldsymbol{\gamma}}^T\mathbf{z}_i)\right\}}{\left(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i}exp\left\{-\hat{\Lambda}_{nc,0}(t_i)exp(\boldsymbol{\gamma}^T\mathbf{z}_i)\right\}\right)^2}\right\}\mathbf{x}_i^T;$$

$$H(\hat{\boldsymbol{\gamma}}) = -\mathbf{z}_i\left\{\delta_i\times e^{\hat{\boldsymbol{\gamma}}^T\mathbf{z}_i}\hat{\Lambda}_{nc,0}(t_i)\right\}\mathbf{z}_i^T - \mathbf{z}_i\left\{(1-\delta_i)\times\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i}exp\left\{-\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\boldsymbol{\gamma}}^T\mathbf{z}_i}\right\}\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\boldsymbol{\gamma}}^T\mathbf{z}_i}}{1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i}exp\left\{-\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\boldsymbol{\gamma}}^T\mathbf{z}_i}\}^2\right\}\mathbf{z}_i^T}\right\}$$
$$-\mathbf{z}_i\left\{(1-\delta_i)\times\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i}exp\left\{-\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\boldsymbol{\gamma}}^T\mathbf{z}_i}\right\}\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\boldsymbol{\gamma}}^T\mathbf{z}_i}\times\left(1-e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i}exp\left\{-\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\boldsymbol{\gamma}}^T\mathbf{z}_i}\right\}\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\boldsymbol{\gamma}}^T\mathbf{z}_i}\right)}{1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i}exp\left\{-\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\boldsymbol{\gamma}}^T\mathbf{z}_i}\right\}}\right\}\mathbf{z}_i^T$$

Alternatively, one could resort to resampling methods to ascertain the standard error estimates. An advantage of bootstrap standard error estimates is their non-reliance of distributional assumptions. To implement, we resample the observations for a finite number of times with replacement. The resamples are all of size n, the size of the original sample. We then estimate the parameters for each of the bootstrap samples; bootstrap standard errors are calculated from the parameter estimates. With the use of EM algorithm, we use this resampling procedure to obtain the appropriate standard errors of $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\gamma}}$.

2.2 Promotion time cure model

Development of the selection method for promotion time cure models parallels that of the MCMs.

2.2.1 Model

Promotion time cure rate model was developed in the context of cancer recurrence led by carcinogenic cells. For example, Chen et al. assume that the ith subject has Y_i carcinogenic cells that could lead to a recurrent disease.[10] They further assume that Y_i follows a Poisson distribution with mean function $\theta(\mathbf{x}_i) = \exp(\boldsymbol{\beta}^T\mathbf{x}_i)$, where $\boldsymbol{\beta}$ is the coefficient vector for independent variables $\mathbf{x}_i \in \mathbb{R}^p$, and that for each cell, time to event ζ follows a distribution $F_1(t)$, or a survival function $S_1(t) = 1 - F_1(t)$. The observed event time $\widetilde{T}_i$ is the time at which the first carcinogenic

source becomes activated. In other words, $\widetilde{T}_i = \min\{\zeta_k\}_{0\le k\le Y_i}$ for the ith subject. The population survival function $S_p(t)$ is defined as the probability of cancer non-detection at time t, which is expressed as

$$S_p(t) = P(Y=0) + P(\zeta_1 > t, \cdots, \zeta_{Y_i} > t;\, Y_i \ge 1) = exp\{-\theta(\mathbf{x}_i)F_1(t)\} \tag{6}$$

The population hazard function corresponding to (6) is $\lambda_p(t) = \theta(\mathbf{x}_i)f_1(t)$, where the density function is $f_1(t) = \frac{d}{dt}F_1(t)$. The cumulative hazard corresponding to (6) is defined as $\Lambda_p(t) = \int_0^t \theta(\mathbf{x}_i)f_1(z)dz = \theta(\mathbf{x}_i)F_1(t)$. We therefore rewrite equation (6) as $S_p(t) = \exp\{-\Lambda_p(t)\}$. As $t \to \infty,\ S_p(t) \to \exp\{-\theta(\mathbf{x}_i)\} > 0$, where $S_p(t)$ is typically not a proper survival function.

Following Tsodikov,[32] we introduced a PH structure into Model (6): $S_p(t|\mathbf{x}_i) = \exp\{-F_1(t)\}^{\exp(\boldsymbol{\beta}^T\mathbf{x}_i)} = S_{p,0}(t)^{\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}$. Suppose $S_{p,0}(t) = \exp\{-F_1(t)\}$, one could regard it as the baseline survival function associated with $F_1(t)$.

Variable selection and estimation: We first introduce an adaptive LASSO method for the promotion time cure model. Let the observed time $T_i = \min(\widetilde{T}_i, C_i)$, where C_i is the non-informative and random censoring time. The censoring indicator $\delta_i = 1$ if $\widetilde{T}_i \le C_i$, and $\delta_i = 0$ otherwise. Model (6) has one set of independent variables $\mathbf{x}_i$ for subject i. For Model (6), the observed likelihood is

$$L(\boldsymbol{\beta}, \alpha) = \prod_{i=1}^{n} \lambda_p(t_i)^{\delta_i} S_p(t_i) = \prod_{i=1}^{n} \Big\{ \{\exp(\boldsymbol{\beta}^T\mathbf{x}_i) f_1(t_i|\alpha)\}^{\delta_i} \exp\{-F_1(t|\alpha)\}^{\exp(\boldsymbol{\beta}^T\mathbf{x}_i)} \Big\} \tag{7}$$

where α is the parameter in $F_1(\cdot)$.

For variable selection and parameter ($\boldsymbol{\beta}$) estimation, we develop an EM-algorithm based on $(t_i, \delta_i, \mathbf{x}_i, y_i)$, where y_i is value of the Poisson cell count, Y_i. The log-likelihood function for the complete data is

$$l_{pc}(\boldsymbol{\beta}, \alpha; y) = \sum_{i=1}^{n} \big\{ \delta_i log(\, y_i f_1(t_i|\alpha)) + (\, y_i - \delta_i) log(1 - F_1(t_i|\alpha)) + y_i \boldsymbol{\beta}^T\mathbf{x}_i - \exp(\boldsymbol{\beta}^T\mathbf{x}_i) - log(\, y_i!) \big\} \tag{8}$$

For variable selection, we use an adaptive LASSO with the following penalized log-likelihood function:

$$pl_{pc}(\boldsymbol{\beta}, \alpha; y) = \left\{ l_{pc}(.) - \tau^* \sum_{j=1}^{p} \frac{|\boldsymbol{\beta}_j|}{|\rho_j|} \right\} \tag{9}$$

As in the case of mixture models, the tuning parameter τ^* determines the amount of penalty in equation (9) and ρ functions as weight. Similarly, we obtain a consistent estimate of $\boldsymbol{\beta}$ by maximizing (8), and use it as the weight. When $\rho = 1$, this penalized function reduces to the familiar LASSO penalized function.

Computation: Let $(\boldsymbol{\beta}^{(m)}, \alpha^{(m)})$ be the parameter estimates in the mth iteration. To maximize equation (9) for given τ^*, the EM algorithm takes the following steps:

E step: In the $(m+1)$ th iteration, we compute $y_i^{(m+1)} = \exp(\boldsymbol{\beta}^{(m)T}x_i)(1 - F_1(t_i|\alpha^{(m)}))$, and replace y_i in (9) with $y_i^{(m+1)}$.

M step: Solve the penalized score equation $U_P(\boldsymbol{\beta})$ for $\boldsymbol{\beta}^{(m+1)}$ of $\boldsymbol{\beta}$ by using quadratic approximation[20]:

$$0 = U_P(\boldsymbol{\beta}) = \sum_{i}^{n} \Big[y_i^{(m+1)} - \exp(\boldsymbol{\beta}^T\mathbf{x}_i) \Big] \mathbf{x}_i^T - \tau^* \sum_{j=1}^{p} \frac{\boldsymbol{\beta}/|\boldsymbol{\beta}^{(m)}|}{|\rho_j|}$$

The penalized Hessian matrix $H^*_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ at $(m+1)$ th iteration is given by $H^*_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} U_P(\boldsymbol{\beta})$.

In the M-step, to estimate α, we partition the time interval into non-overlapping sub-intervals defined by $0 < s_1 < s_2 < \ldots < s_G$, with $s_G > max\{t_i\}$. We assume that $F_1(t|\alpha)$ follows a piecewise exponential model for which the hazard $\alpha_g (\, g = 1, 2, \cdots, G)$ remains constant for each sub-interval.[13] It can be shown by maximizing (8) with respect to α_g that for $i = 1, 2, \cdots, n$

$$\alpha_g^{(m+1)} = \left[\sum_{s_{g-1} < t_i \le s_g} \delta_i \right] \times \left[\sum_{s_{g-1} < t_i \le s_g} y_i^{(m+1)}(t_i - s_{g-1}) + \sum_{y_i > s_g} y_i^{(m+1)}(s_g - s_{g-1}) \right]^{-1}$$

Alternatively, we can use the empirical distribution of $F_1(t)$ by assigning a point mass at each distinct observed event time so that $\sum f_1(t) = 1$ over the entire range of t. Suppose we have D distinct event times defined by $t_1^* < \cdots < t_D^*$. Let $f_1(t_d^*) = \alpha_d$ for $d = 1, 2, \cdots D$ so that $F_1(t_i|\alpha) = \sum_{t_d^* \le t_i} \alpha_d$. For given values of $\beta^{(m)}$, we maximize (7) as a function of α only. The function to be maximized is

$$L_{\boldsymbol{\beta}^{(m)}}(\alpha_1, \cdots, \alpha_D) \propto \prod_{d=1}^{D} \alpha_d \times exp\left\{-\alpha_d \sum_{i \in R_d} \exp(\boldsymbol{\beta}^{(m)T}\mathbf{x}_i)\right\}$$

The profile ML estimate $\alpha^{(m+1)}$ of α is given by

$$\alpha_d^{(m+1)} = \frac{1}{\sum_{i \in R_d} \exp(\boldsymbol{\beta}^{(m)T}\mathbf{x}_i)}$$

where R_d is the number of individuals at risk at time t_d^*. This yields an estimate of $F_1(t|\alpha)$

$$F_1^{(m+1)}(t|\alpha) = \sum_{t_d^* \le t} \alpha_d^{(m+1)}$$

which is similar to the nonparametric version of the Breslow estimator of the baseline cumulative hazard.

The final estimator is obtained by iterating between the E and M steps until convergence. The EM algorithm has the following key steps:

Step 1: Determine an appropriate value for the tuning parameter τ^*, and initialize $\boldsymbol{\beta}^{(0)}$
Step 2: Execute the E-step and estimate α
Step 3: Update the estimates as $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} - H^{-1}(\boldsymbol{\beta}^{(0)})U(\boldsymbol{\beta}^{(0)})$
Step 4: Repeat steps 2 and 3 until $|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}| \to 0$

For the tuning parameter selection, we use the same equation (5) to derive the BIC criterion for τ^*. Given $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$ we obtain an estimate of the log likelihood $\hat{l_{pc}}(.)$ from the unpenalized likelihood function. Using the BIC formula (5), we select a value of τ^* that minimizes the BIC.

As in mixture cure models, we take a two-step approach for parameter estimation and inference, i.e. independent variable effects are estimated and tested in a model with the selected variables. A standard approach for variance estimate is to use the inverse of the negative Hessian matrix derived from the observed likelihood (7). The covariance estimators for $\hat{\boldsymbol{\beta}}$ given $F_1(t_i|\hat{\alpha})$ is

$$H(\hat{\boldsymbol{\beta}}) = -F_1(t_i|\hat{\alpha})e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i}\mathbf{x}_i\mathbf{x}_i^T(e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i})^T$$

In this research, we use bootstrap standard deviations for inference.

3 Simulation study

We conduct a simulation study to evaluate the selection performance of the two cure rate models. Specifically, we compare the rates of selection accuracy of the proposed LASSO and adaptive LASSO methods, against that of the naïve p value selection method. The significance level for the p value procedure is set at 0.05, i.e. a variable is retained if the corresponding p value is less than 0.05. The R-code can be downloaded at the ftp website using windows file explorer at: *ftp://public.sjtu.edu.cn*. The user name is *yuzhangsheng* and the password is *public*.

3.1 Mixture cure rate models

Data generation. We consider a scenario where $\mathbf{x} = (x_1, \cdots, x_9)^T$ has nine independent variables. Three of the nine, x_5, x_6, x_9, are independent binary variables (1 vs. 0) with probability $P(x_j = 1) = 0.5$, $j = 5$, 6, 9. The other independent variables in $\mathbf{x}$ are standard normally distributed with a pairwise correlation between x_i and x_j of $\rho^{|i-j|} = 0.5$, which reflects a moderately strong correlation. For the logistic component of the mixture cure model,

vectors of regression coefficients are set to $\boldsymbol{\beta} = (0.5, 0.10, -0.25, 0, 0, 0, 0, 0, 0)^T$. For the survival model, we assume without loss of generality that $\mathbf{x} = \mathbf{z}$. Failure times are generated from a Weibull distribution with a survival function $S(t|a, b) = \exp\{-\frac{t}{b}\}^a$. The shape parameter is $a = 1.5$, and scale parameter $b = \exp\{e^{\boldsymbol{\gamma}^T\mathbf{z}}\}^{-1/a}$ with $\boldsymbol{\gamma} = (1, 0, 0.1, 0.25, 0, 0, 0, 0, 0)^T$. We include an intercept for the logistic model, and no intercept for the survival model. The mean cure rate is approximately 30%. Censoring times are generated from Uniform (c, d), where c and d are selected to achieve the desired censoring rate. We considered two different levels of censoring: 20% and 50%.

For each parameter setting, we generated 100 datasets, with sample sizes of 250 and 500. We apply the LASSO, adaptive LASSO, and naïve p value procedures for variable selection. We implement the selection procedure with both parametric and nonparametric estimators for the baseline hazard $\Lambda_{nc,0}(t)$. We apply the penalized methods for variable selection with given values of the tuning parameter $\tau = (\tau_1, \tau_2)$ for the mixture cure rare model. Optimal values of the tuning parameter are selected by minimizing the BIC selector (5).

Simulation results. Table 1 presents the selection results for the mixture cure model. Six elements of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ have zero effects, whereas the other three have non-zero effects. We present the average number of correct exclusion (unimportant effects not being selected) and the average number of incorrect exclusion (important effects not being selected) for the logistic regression coefficients $\boldsymbol{\beta}$ and PH regression coefficients $\boldsymbol{\gamma}$. The table summarizes the results based on 100 simulations.

Briefly, for the logistic regression component in the mixture model, the rate of incorrect exclusion is zero for both LASSO and adaptive LASSO. In other words, both regularization methods have correctly included all three non-zero effects. In comparison, the p value method has on average incorrectly excluded $2.96 - 3$ of the 3 non-zero effects, a very poor performance by any standard. In the meantime, the adaptive LASSO has excellent rates of correct exclusion: On average, it is able to exclude $4.8 - 6$ of the 6 true zero effects. This performance is similar to that of the p value method which consistently excludes $5.8 - 6$ of the zero effects. The LASSO method, on the other hand, tends to exclude fewer zero effects.

For the PH component, all three methods have correctly included the three non-zero effects. The difference is in the exclusion of zero effects. In this regard, the adaptive LASSO has the best performance. It is able to exclude $3.66 - 5.91$ of the 6 zero effects. LASSO has slightly worse but still acceptable performance. The p value method, on the other hand, completely fails to exclude any of the zero effects.

In comparing the selection performance of balancing the two different types of errors, the adaptive LASSO appears to outperform its competitors. Importantly, the superior performance of the adaptive LASSO procedure is consistent across all simulation settings and it does not appear to be greatly influenced by the censoring proportion and how baseline hazards are estimated.

3.2 Promotion time cure rate models

Data generation. For the promotion time cure model, we again consider a situation where $\mathbf{x} = (x_1, \cdots, x_9)^T$ has nine independent variables. Three of the nine, x_5, x_6, x_9, are independent Bernoulli variables with probability $P(x_j = 1) = 0.5$, $j = 5, 6, 9$. The other six variables of $\mathbf{x}$ are standard normally distributed with a pairwise correlation between x_i and x_j of $\rho^{|i-j|} = 0.5$. As in Model (6), we assume that the mean number of cancer cells is $\theta = \exp(\boldsymbol{\beta}^T\mathbf{x})$ with $\boldsymbol{\beta} = (0.5, 0.10, -0.25, 0, 0, 0, 0, 0, 0)^T$. We also assume that $F_1(t)$ follows a Weibull distribution with scale parameter $b = \exp\{\theta\}^{-1/a}$, and the shape parameter $a = 1.5$. Censoring times are generated from a uniform distribution yielding censoring rate of 20% and 50%. We generated 100 datasets for each setting with sample sizes 250 and 500, and censoring percentages 20% and 50%. We fit Model (6) by using both parametric and nonparametric estimates of $F_1(t)$.

Simulation results. Table 2 depicts the selection results for promotion time cure model. The simulation shows that the adaptive LASSO outperforms both LASSO and the p value methods in identifying the zero effects, as evidenced by its high correct exclusion rates, while maintaining a perfect rate of including all non-zero effects. The LASSO has respectable performance in achieving a perfect rate of including all non-zero effects, but it is slightly less effective in identifying the zero effects. The p value method tends to incorrectly exclude the true non-zero effects at the unacceptable rates of $1.28 - 1.98$ out of 3.

3.3 Post-selection inference

In the absence of formal theoretical development of post-selection inference, analysts are likely to perform inference based on the selected model. Here we conduct a simulation study to examine the empirical

Table 1. Simulation study. Performance of variable selection results for mixture cure model with 20% and 50% censoring. The average numbers of correct exclusion (exclusion of zero effects) and incorrect exclusion (exclusion of non-zero effects).

		Average number of 0 coefficients			
		$\boldsymbol{\beta}$(logistic)		γ(survival)	
n	Method	Correct exclusion (6)	Incorrect exclusion (3)	Correct exclusion (6)	Incorrect exclusion (3)
20% censoring					
Nonpar $\lambda_0(t)$					
250	Oracle	6	0	6	0
	$p < 0.05$	6	2.96	0	0
	Adaptive LASSO	5.23	0	4.68	0
	LASSO	4.97	0	4	0
500	Oracle	6	0	6	0
	$p < 0.05$	6	2.98	0	0
	Adaptive LASSO	6	0	5.90	0
	LASSO	6	0	4.57	0
Par $\lambda_0(t)$					
250	Oracle	6	0	6	0
	$p < 0.05$	5.98	3	0	0
	Adaptive LASSO	5.76	0	4.34	0
	LASSO	4	0	3.86	0
500	Oracle	6	0	6	0
	$p < 0.05$	5.99	3	0	0
	Adaptive LASSO	6	0	5.91	0
	LASSO	4.52	0	4	0
50% censoring					
Nonpar $\lambda_0(t)$					
250	Oracle	6	0	6	0
	$p < 0.05$	5.80	2.90	0	0
	Adaptive LASSO	5.02	0	4.02	0
	LASSO	3.99	0	3.76	0
500	Oracle	6	0	6	0
	$p < 0.05$	4.92	2.98	0	0
	Adaptive LASSO	6	0	4.99	0
	LASSO	6	0	3.71	0
Par $\lambda_0(t)$					
250	Oracle	6	0	6	0
	$p < 0.05$	5.88	2.94	0	0
	Adaptive LASSO	4.80	0	3.66	0
	LASSO	3.75	0	3.66	0
500	Oracle	6	0	6	0
	$p < 0.05$	6	3	0	0
	Adaptive LASSO	6	0	5.69	0
	LASSO	4.41	0	4	0

performance of the practice. Specifically we examine the 95% coverage probabilities and the average bootstrap standard errors (ASE) for the non-zero coefficients of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Here bootstrap standard errors are obtained based on 100 resamples. Simulation results are presented in Table 3. Briefly, the coverage probabilities are generally good, especially for the promotion time cure rate models, even with 50% censoring. The performance of the mixture model is slightly more variable. Overall, the simulation seems to provide some empirical evidence in support of the two step selection-estimation procedure.

Finally, we conducted a sensitivity analysis examining the selection performance in misspecified models, i.e. data are generated from mixture models when promotion time models are fitted, or vice versa. In the strictest sense, the MCM and promotion cure rate model (PCM) are not directly comparable because of their differences in

Table 2. Simulation study. Performance of variable selection results for promotion time cure model with 20% and 50% censoring. The average numbers of correct exclusion (exclusion of zero effects) and incorrect exclusion (exclusion of non-zero effects).

		Average number of 0 coefficients			
		20% censoring		50% censoring	
		β		β	
n	Method	Correct exclusion (6)	Incorrect exclusion (3)	Correct exclusion (6)	Incorrect exclusion (3)
Nonparametric specification					
250	Oracle	6	0	6	0
	$p < 0.05$	4.42	1.46	5.55	1.83
	Adaptive LASSO	6	0	6	0
	LASSO	4.62	0	4.30	0
500	Oracle	6	0	6	0
	$p < 0.05$	4.85	1.56	5.92	1.98
	Adaptive LASSO	6	0	6	0
	LASSO	5.64	0	5.33	0
Parametric specification					
250	Oracle	6	0	6	0
	$p < 0.05$	4.72	1.48	4.65	1.48
	Adaptive LASSO	6	0	4.45	0
	LASSO	3.98	0	3.45	0
500	Oracle	6	0	6	0
	$p < 0.05$	4.44	1.28	5.48	1.89
	Adaptive LASSO	6	0	6	0
	LASSO	4	0	3.83	0

Table 3. Simulation study. Empirical 95% coverage probability (Coverage prob), and average values of the estimated bootstrap standard errors (ASE) of the estimates in the adaptive LASSO selected models.

			20% censoring		50% censoring	
N	Model	Coefficient	Coverage prob.	ASE	Coverage prob.	ASE
250	MCM–logistic	β_1	0.82	0.090	0.91	0.203
		β_2	0.95	0.072	0.96	0.148
		β_3	0.95	0.036	0.96	0.074
250	MCM–survival	γ_1	0.96	0.024	0.95	0.022
		γ_2	0.96	0.029	0.91	0.024
		γ_3	0.96	0.014	0.91	0.011
500	MCM–logistic	β_1	0.88	0.080	0.89	0.071
		β_2	0.95	0.057	0.96	0.068
		β_3	0.95	0.028	0.96	0.034
500	MCM–survival	γ_1	0.96	0.021	0.98	0.139
		γ_2	0.94	0.022	0.98	0.150
		γ_3	0.94	0.011	0.98	0.075
250	PCM	β_1	0.93	0.002	0.95	0.010
		β_2	0.95	0.004	0.95	0.010
		β_3	0.95	0.002	0.95	0.005
500	PCM	β_1	0.95	0.002	0.96	0.003
		β_2	0.95	0.003	0.96	0.003
		β_3	0.95	0.002	0.96	0.002

MCM: mixture cure rate models, PCM: promotion time cure rate model.

Table 4. Sensitivity analysis on robustness of model misspecification.

		20% censoring		50% censoring	
n	Method	Correct exclusion (6)	Incorrect exclusion (3)	Correct exclusion (6)	Incorrect exclusion (3)
CASE A(true model MCM and fitted as PCM)					
250	Oracle	6	0	6	0
	PCM-Adp. Lasso	5.26	0	5.39	0
	PCM-Lasso	4.58	0	4.19	0
	MCM-Adp. Lasso	4.68	0	4.02	0
	MCM-Lasso	4	0	3.76	0
500	Oracle	6	0	6	0
	PCM-Adp. Lasso	5.83	0	5.90	0
	PCM-Lasso	5.47	0	5.05	0
	MCM-Adp. Lasso	5.90	0	4.99	0
	MCM-Lasso	4.57	0	3.71	0
CASE B(true model PCM and fitted as MCM)					
250	Oracle	6	0	6	0
	PCM-Adp. Lasso	6	0	6	0
	PCM-Lasso	4.62	0	4.30	0
	MCM-Adp. Lasso	4	0	4	0
	MCM-Lasso	4	0	4	0
500	Oracle	6	0	6	0
	PCM-Adp. Lasso	6	0	6	0
	PCM-Lasso	5.64	0	5.33	0
	MCM-Adp. Lasso	4.91	0	4.08	0
	MCM-Lasso	4.98	0	4	0

PCM: promotion time cure model, MCM: mixture cure rate model.

structure and assumption. In case A (MCM as true model) and case B (MCM as misspecified model), Table 4 only shows the survival component of the MCM fits. Simulation shows that selection accuracies in the survival components of the true and misspecified models were generally comparable, which provided some assurance on the robustness of the selection method. But we caution against over-interpretation because the simulation has not taken into account the selection performance of the logistic component in the MCM. Detailed results are included in Table 4.

4 Application: a childhood wheezing study

To illustrate the proposed methods, we consider a real clinical investigation of childhood wheezing. The basic study design was described elsewhere.[3] Briefly, this is an observational study aimed at understanding the risk factors associated with early onset of wheezing. For this purpose, the variable selection methods that we develop provided a logical tool for risk factor screening. The onset age of wheezing symptoms was the main outcome of interest. Onset age was determined from the monthly reports of wheezing episodes during the study period. The study recruited a total of 116 children. Enrolled children were followed prospectively for up to five years. Eighty-six ($n = 86$) children completed the designed follow-up. The current analysis was based on data from these 86 children with complete follow-up.

A total of 13 variables were considered in the current analysis. The demographic and general health variables included race (RACE, 1 = white and 0 = non-white), sex (GENDER, 1 = male, 0 = female), and mother's smoking status during pregnancy (1 = nonsmoker mother during pregnancy, and 0 = otherwise), allergy to food (FOODANT; 1 = yes, 0 = no), egg or milk (EGGMILK, 1 = yes, 0 = no), and use of topical steroids (TOPSTEO, 1 = yes, 0 = no). Continuous variables included: (1) provocative concentration of methacholine corresponding to 30% drop in forced expiratory volume in 1 s (logPC30 (mg/ml)); (2) centralized height (CenHEIGHT (cm)); (3) severity of eczema, a score ranged from 0 to 29 calculated based the levels of body surface involvement, intensity of symptom, and presence of pruritus and insomnia (SCVALUE); (4) logarithmic transformed level of total serum immunoglobulin E (log(ITOTAL)); (5) Z-score of forced vital capacity (ZFVC); (6) Z-score of forced expiratory flow 25% − 75%

(ZFEF2575); (7) Z-score of forced expiratory volume in half a second (ZFEV5). Among these, the last three variables (ZFVC, ZFEF2575, and ZFEV5) were lung function measurements. The average age at enrollment of these children was approximately 10.7 months. The median age at the first wheeze episode was 21.67 months. Summary statistics of the independent variables are reported in Table 5.

Kaplan-Meier estimates of the wheezing free probabilities for boys and girls are presented in Figure 1. The Kaplan-Meier plot for girls flattened after 48 months, with relatively few censoring, suggesting that a portion of the population were not subject to any risk of wheezing. A similar pattern was seen in boys. To accommodate this fraction of the cured, we analyzed the data using a MCM (1). We did not consider promotion time cure models in the absence of a clear biological rationale for that approach. Wheezing, as an airway symptom, does not have a single and specific cause that justifies the use of a promotion time model. We performed variable selection using methods described in the paper. Both LASSO and adaptive LASSO methods were used.

To select the tuning parameters for the logistic regression and PH regression models, for a given set of tuning parameter values we plug in the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ into equation (3). And then we optimize the tuning parameters that minimize the BIC selector (5). Under the LASSO penalty, all 13 variables were retained for the logistic regression model. The adaptive LASSO produced a more parsimonious logistic model with five independent variables: SCVALUE, GENDER, RACE, MONSMOKE, and TOPSTEO. For the PH model, the LASSO penalty selected 11 of the 13 variables: GENDER, RACE, MOMSMOKE, FOODANT, EGGMI LK, TOPSTEO, ZFEF2575, ZFEV5, HEIGHT, SCVALUE, and ITOTAL. The adaptive LASSO selected seven variables: SCVALUE, GENDER, RACE, MONSMOKE, FOODANT, EGGMILK, and TOPSTEO. We present the final model fitting results based on the adaptive LASSO method in Table 6. Of note, the model identified by the adaptive LASSO was more parsimonious, and it included all of the variables identified by the LASSO method.

Table 5. Baseline characteristics of subjects included in the analysis.

Factor		*n*	Variable	Mean	Variance
GENDER	0	42	*ZFVC*	−0.319	1.222
	1	44			
RACE	0	43	*ZFEF2575*	−0.689	0.837
	1	43			
MOMSMOKE	0	9	*ZFEV5*	−0.614	1.108
	1	77			
FOODANT	0	55	*CenHEIGHT*	−0.673	41.935
	1	31			
EGGMILK	0	59	*SCVALUE*	9.547	50.203
	1	27			
TOPSTEO	0	47	*log(ITOTAL)*	2.147	2.700
	1	39			
			logPC30	−0.787	1

Table 6. Summary of parameter estimates with confidence intervals and two sided *p* values for the childhood wheezing study. In the logistic model, OR stands for odds ratio. In the survival model, HR refers to hazard ratio.

Variable	OR (CI)	*p* value	HR (CI)	*p* value
Intercept	1.030 (1.017, 1.042)	0.000		
SCVALUE	1.337 (1.242, 1.439)	0.000	1.275 (1.196, 1.400)	0.000
MOMSMOKE	1.025 (1.014, 1.037)	0.000	1.022 (1.011, 1.034)	0.000
RACE	1.016 (1.010, 1.025)	0.000	1.014 (1.007, 1.022)	0.000
GENDER	1.017 (1.010, 1.025)	0.000	1.016 (1.010, 1.024)	0.000
TOPSTEO	1.013 (1.010, 1.018)	0.000	1.010 (1.005, 1.015)	0.000
FOODANT			1.009 (1.003, 1.016)	0.002
EGGMILK			1.008 (1.002, 1.014)	0.005

A careful examination of the parameter estimates from the selected model revealed that: (1) an estimated $49\% = 1/(1 + 1.03)$ of population was subject to the risk of wheezing if all other factors (SCVALUE, GENDER, RACE, MOMSMOKE, and TOPSTEO) were set to 0; (2) male sex, white race, mother smoked during pregnancy, topical steroid use, and greater eczema severity were associated with increased risk of wheezing. For the children who were at risk, a greater eczema severity, mother smoking during pregnancy, white race, male sex, topical steroid use, and known allergy to food, egg, and milk were associated with early onset of wheezing.

5 Conclusion and discussion

Cure rate model represents an important class of methods for analyzing time-to-event data, in situations where certain individuals are free of the disease risk. Because of the increased complexity in modeling structure, a common challenge that analysts face is the determination of model composition, i.e. what independent variables should be included in or excluded from which modeling components. While fully subjective variable selection by investigators is usually thought to be error-prone, the traditional p value-based selection methods are not always efficient and stable. To alleviate the challenge, we present two selection methods, based on LASSO and adaptive LASSO, to aid variable selection in different types of cure rate models. Built on earlier attempts on the mixture cure model,[23] this work further extends the selection tool to promotion time models. Extensive simulation shows that the adaptive LASSO method has superior performance than the LASSO and p value methods, in terms of selection accuracy. The method appears to have worked well for both mixture and promotion time cure rate models. Making these methods available to practitioners, we hope, would have an impact on how cure rate models are used in analytical practice. The selection of independent variables are of course not limited to main effects, two-way or higher order interactions can be incorporated with modification of the design matrices for the logistic and survival components, with the usual understanding that the main effects are to be included if an interaction involving them is selected. Computationally, as we have demonstrated in the current paper, adaptive LASSO is generally efficient, and it is easily implementable in various computing platforms.

A few practical issues deserve some discussion: (1) Determination of the initial sets of independent variables going into the logistic and survival components is generally guided by subject science, and it typically reflects the investigators' understanding of the cure and survival processes. In the absence of strong scientific reasons for including and/or excluding certain variables into the initial sets of independent variables, analysts typically use the same set of variables for both components, so $\mathbf{x} = \mathbf{z}$ is a rather common practice. (2) Estimation of the unknown baseline hazard functions. Previously, different authors have explored various approaches. Among the published methods, for MCM Sy and Taylor[2] used a Breslow type estimator and a product limit estimator, Farewell[6] considered a parametric (Weibull) model, Corbiere et al.[33] attempted the use of nonparametric spline functions, and Chen and Ibrahim[13] used a piecewise exponential model for hazard function for promotion time cure model. In this research, we constructed a nonparametric step-function for baseline hazard under the promotion time cure model. For MCM we utilized a piecewise constant hazard function for baseline hazard approximation. We compared the performance of variable selection of adaptive LASSO and LASSO using the Breslow type estimator and piecewise exponential model for the baseline hazard function in the simulation. Our simulation shows that different choices of baseline hazard estimators produced generally comparable selection results. Consider the simplicity of our approach, we conclude that the choice of baseline hazard estimation methods is not as consequential as previously thought, at least for the purpose of variable selection. (3) Determination of the weights for adaptive LASSO. Ideally, the weights need to be data-dependent and consistent with the oracle properties.[20] When the number of variables is larger and many of them are correlated, the consistent estimates may be difficult to obtain. Thus the issue requires further investigation. (4) Estimation of standard errors. Standard error estimates are important for the purpose of inference. For linear models Tibshirani[18] and Fan and Li[20] provided Hessian matrix-based standard error estimates, while Zou[24] advocated the use of bootstrap estimates. For nonlinear models, penalized variable selection methods tend to introduce biases in the estimation of model parameters. The magnitude of the bias is influenced by the choice of weights or tuning parameters. As a result, Hessian matrix-based standard error estimates do not work well for inference, at least in our modeling setting (data not shown). So in this research, we chose to use bootstrap standard error estimates in the selected model, to minimize the impact of tuning parameters and thus alleviating the risk of estimation bias. (5) Post-selection inference. As stated earlier, this paper has primarily focused on variable selection and not on post-selection inference. The two-stage estimation process is somewhat an *ad hoc* way to obtain the approximation of standard errors, but it is generally consistent with the current biostatistical practice of

making inference based on the final selected models.[30] Most recently, Berk et al.[31] have suggested that one could reframe the post-selection testing in the context of simultaneous inference, which takes into account the multiplicity associated with all sub-models (all linear functions of estimates) instead of the selected model, in hoping that the inference no longer depends on correct selection of the true model. Berk's approach was discussed in a linear model setting. Extension of this approach to nonlinear settings remains to be developed. In the absence of rigorous methodological development, we opted for the standard two-step approach. The simulation study seems to support the notion of a generally good selection performance, at least in tested settings. On balance, use of resampling in a two-step process, in our opinion, represents a sensible compromise between accurate standard error estimation and valid inference performance. We have shown in a previous work that such a method works well in complex modeling settings.[22]

Acknowledgements

The authors thank the Editor, the Associate Editor, and three reviewers for their many insightful comments.

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: partially by the National Institutes of Health grants RO1HL095086, P30HS024384, and U54 CA190151, National Natural Science Foundation of China 11671256 and also 2016YFC0902403 of Chinese Ministry of Science and Technology.

References

1. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B* 1972; **34**: 187–220.
2. Sy JP and Taylor JMG. Estimation in a Cox proportional hazards cure model. *Biometrics* 2000; **56**: 227–236.
3. Tepper RS, Llapur CJ, Jones MH, et al. Expired nitric oxide and airway reactivity in infants at risk for asthma. *Am Acad Allergy Asthma Immunol* 2008; **122**: 760–765.
4. Chen T. Statistical issues and challenges in immuno-oncology. *J Immunother Cancer* 2013; **11**: 1–18.
5. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J R Stat Soc* 1949; **11**: 15–53.
6. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982; **38**: 1041–1046.
7. Taylor JMG. Semi-parametric estimation in failure time mixture models. *Biometrics* 1995; **51**: 899–907.
8. Peng Y and Dear KBG. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000; **56**: 237–243.
9. Yakovlev AY, Asselain B, Bardou VJ, et al. A nonparametric mixture model for cure rate estimation. *Biometr Anal Dormees Spatio-Temporelles* 1993; **12**: 66–82.
10. Chen MH, Ibrahim JG and Sinha D. A Bayesian approach to survival data with a cure fraction. *J Am Stat Assoc* 1999; **94**: 909–919.
11. Ibrahim JG, Chen MH and Sinha D. *Bayesian survival analysis*. New York, NY: Springer, 2002.
12. Chen MH, Ibrahim JG and Lipsitz SR. Bayesian methods for missing covariates in cure rate models. *Lifetime Data Anal* 2002; **8**: 117–146.
13. Chen MH and Ibrahim JG. Maximum likelihood methods for cure rate models with missing covariates. *Biometrics* 2001; **57**: 43–52.
14. Tsodikov AD, Ibrahim JG and Yakovlev AY. Estimating cure rates from survival data: an alternative to two-component mixture models. *Lifetime Data Anal* 2003; **98**: 1063–1078.
15. Broët P, De Rycke Y, Tubert-Bitter P, et al. A semiparametric approach for the two-sample comparison of survival times with long-term survivors. *Biometrics* 2001; **57**: 844–852.
16. Yin G and Ibrahim JG. Cure rate models: a unified approach. *Can J Stat* 2005; **57**: 559–570.
17. Breiman L. Heuristics of instability and stabilization in model selection. *Ann Stat* 1996; **24**: 2350–2383.
18. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B* 1996; **56**: 267–288.
19. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med* 1997; **16**: 385–395.
20. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.

21. Zhang HH and Lu W. Adaptive lasso for Cox's proportional hazard model. *Biometrika* 2007; **94**: 691–703.
22. He Z, Tu W, Wang S, et al. Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics* 2015; **71**: 178–187.
23. Liu X, Peng Y, Tu D, et al. Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trails. *Stat Med* 2012; **31**: 2882–2891.
24. Zou H. The adaptive LASSO and its oracle properties. *J Am Stat Assoc* 2006; **101**: 1418–1429.
25. Klein JP. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* 1982; **48**: 795–806.
26. Nishii R. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann Stat* 1984; **12**: 758–765.
27. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978; **19**: 461–464.
28. Zou H and Li R. One step sparse estimates in nonconcave penalized likelihood models. *Ann Stat* 2008; **36**: 1509–1533.
29. Zhang Y, Li R and Tsai CL. Regularization parameters selection via generalized information criterion. *J Am Stat Assoc* 2010; **105**: 312–323.
30. Moore D and McCabe GP. *Introduction to the practice of statistics*. New York, NY: Freeman and Company, 2009.
31. Berk R, Brown L, Buja A, et al. A penalized likelihood approach in mixture cure models. *Ann Stat* 2013; **41**: 802–837.
32. Tsodikov D. A proportional hazards model taking account of long term survivors. *Biometrics* 1998; **54**: 1508–1516.
33. Corbiere F, Commenges D and Taylor JMC. A penalized likelihood approach in mixture cure models. *Stat Med* 2009; **28**: 510–524.