Published in final edited form as:
Lifetime Data Analysis (2018), **24(1)**(Issue): 72-93
DOI: https://doi.org/10.1007/s10985-017-9392-5

Variable selection and prediction in biased samples with censored outcomes

YING WU

Institute of Statistics, Nankai University, Tianjin, China,

RICHARD J. COOK

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: rjcook@uwaterloo.ca

Summary

With the increasing availability of large prospective disease registries, scientists studying the course of chronic conditions often have access to multiple data sources, with each source generated based on its own entry conditions. The different entry conditions of the various registries may be explicitly based on the response process of interest, in which case the statistical analysis must recognize the unique truncation schemes. Moreover, intermittent assessment of individuals in the registries can lead to interval-censored times of interest. We consider the problem of selecting important prognostic biomarkers from a large set of candidates when the event times of interest are truncated and right- or interval-censored. Methods for penalized regression are adapted to handle truncation via a Turnbull-type complete data likelihood. An expectation-maximization algorithm is described which is empirically shown to perform well. Inverse probability weights are used to adjust for the selection bias when assessing predictive accuracy based on individuals whose event status is known at a time of interest. Application to the motivating study of the development of psoriatic arthritis in patients with psoriasis in both the psoriasis cohort and the psoriatic arthritis cohort illustrates the procedure.

Keywords: Expectation-maximization algorithm, Inverse probability weighted estimator, Truncation, Penalized regression, Prediction error, ROC curve

1 Introduction and Motivating Problem

The availability of large disease registries with longitudinal follow-up has lead to increased interest in utilizing such data for scientific inquiry about the genetic basis for disease onset, disease progression, and the development of co-morbidities. In disease processes with multiple stages, some registries may recruit individuals in an early phase of a disease process, while others may sample individuals in a more advanced stage. Synthesis of data from registries with stage dependent recruitment criteria requires suitable handling of the selection mechanisms.

We consider the problem of identifying human leukocyte antigens (HLA) associated with the rapid onset of psoriatic arthritis (PsA) in psoriasis patients. The motivation of this work is to ensure that psoriasis patients at high risk for PsA are closely monitored, onset of PsA is detected promptly, and to ensure treatments geared toward the prevention of joint damage from arthritis are administered in a timely fashion. The develeopment of predictive models for PsA can also help guide the selection of high risk patients for inclusion in clinical trials of experimental prophylactic treatments. Previous work in this vein was based on a binary classification of the disease status of individuals (psoriatic arthritis versus psoriasis) with cross-sectional logistic regression analyses carried out to identify factors associated with PsA versus PsC (Eder et al., 2015); nested-case-control designs (Julian et al., 2002) have also been employed. A preferred approach, however, is to consider the temporal aspects of the disease process and to model the time from the development of psoriasis to psoriatic arthritis. To carry out this analysis, data from a registry of patients with psoriasis and a registry of patients with psoriatic arthritis are utilized. The registries are described briefly in what follows.

Researchers at the Centre for Prognosis Studies in Rheumatic Diseases at the Toronto Western Hospital created the University of Toronto Psoriasis Clinic (UTPC) registry in 2008 to study the course of psoriasis (Ps), a chronic inflammatory skin condition which affects up to 3% of the population (Schafer, 2006). Screened patients identified as having psoriasis are recruited to this clinical registry and upon entry they undergo a detailed clinical examination, provide samples for genetic testing, are then followed prospectively according to a standardized protocol; clinical assessments are planed every 6 months; the actual timing of the assessments is quite variable however. Approximately 30% of psoriasis patients develop psoriatic arthritis (PsA), a rheumatological disorder featuring inflammatory psoriatic disease as well as inflammation and damage in and around the joints of several areas including the wrists, hands, knees, ankles, lower back, and neck (Chandran et al., 2010).

The University of Toronto Psoriatic Arthritis Clinic (UTPAC) registry was launched much earlier in 1977 to study this complex disease (Gladman et al., 2008). A primary method of recruitment of patients is through the use of a population-based screening tool in the form of a 10 item questionnaire (Tom et al., 2015). Individuals suspected of having psoriatic arthritis based on this tool are invited to attend the clinic for a more definitive diagnosis, and those found to have the disease are invited to join the UTPAC. Upon entry to the UTPAC, as in the UTPC, a detailed history is taken, patients undergo a thorough clinical and radiological examination, and samples are collected for genetic testing. Patients are then scheduled to undergo detailed annual clinical examinations and biannual radiological examinations.

The genotypes of HLA-A, HLA-B, HLA-C, HLA-DR and HLA-DQ alleles were collected in both clinic registries, and a total of 96 HLA markers were identified as of interest *a prior*; 20 of these markers had a frequency in the sample of less than 1% and so were excluded from further consideration. This problem of finding key HLA variables in this setting can then be characterized as one of variable selection in the context of a two-stage disease process, while utilizing data from registries formed based on different disease-related truncation schemes. To address this challenge we propose an expectation-maximization algorithm to deal with truncated event times through specification of a Turnbull-type (Turnbull, 1976) complete data likelihood which involves an augmentation term corresponding to pseudo-individuals in the population who did not satisfy the respective truncation conditions. We then penalize the complete data likelihood by the introduction of the LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006) or SCAD (Fan and Li, 2001, Zou and Li, 2008) penalty functions.

The remainder of the article is organized as follows. In Section 2.1 we define the notation, formulate the model for the waiting time with a piecewise-constant baseline hazard, give the form of the augmented and penalized complete data likelihood and discuss the variable selection algorithm. Details on the design and results of simulation studies are also provided in Section 2.2. In Section 2.3 we apply the proposed algorithm to a dataset which involves left and right truncated samples with right-censored responses. Section 3 discusses the challenges associated with assessing predictive accuracy of models with interval-censored responses, where inverse weighting method is adopted to address the need to restrict attention to individuals whose status is known. Another example involving left-truncated and interval-censored outcomes is discussed here. Concluding remarks and topics for future research are given in Section 4.

2 Penalized Regression for Truncated and Censored Data

2.1 Notation and the Penalized Observed Data Likelihood

Figure 1 contains two Lexis diagrams characterizing the selection criteria for patients into the UTPC and UTPAC cohorts for a hypothetical individual; the horizontal axis represents the timing of events in calendar time while the vertical axis conveys the times since the development of psoriasis. We let E_{i0} denote the calendar time of the onset of psoriasis and E_{i1} denote the calendar time psoriatic arthritis developed for individual i . The time from the onset of Ps to the onset of PsA is denoted $T_i = E_{i1} - E_{i0}$.

The calendar time at which individuals are screened is denoted by A_0. For the UTPC cohort, individuals are required to have psoriasis at the time of screening but cannot have developed PsA, so patients are recruited to this registry subject to the constraint $E_{i0} < A_0 < E_{i1}$ (left panel Figure 1). Given E_{i0}, this can be equivalently expressed as the constraint $T_i \geq \mathcal{L}_i$ where $\mathcal{L}_i = A_0 - E_{i0}$ is the left-truncation time for T_i. For the PsA cohort, only screened subjects who are determined to have PsA are included in the registry, so in this cohort, subjects are sampled subject to the constraint $E_{i1} < A_0$, or equivalently given E_{i0} subject to $T_i \leq \mathcal{R}_i$ where $\mathcal{R}_i = A_0 - E_{i0}$ is the right-truncation time for T_i (right panel Figure 1). To unify the notation for the two cohorts we let $\mathcal{A}_i = [\mathcal{L}_i, \mathcal{R}_i)$ denote the truncation interval for individual i, such that $0 < \mathcal{L}_i < \mathcal{R}_i = \infty$ for individuals in the UTPC, and $0 = \mathcal{L}_i < T_i < \mathcal{R}_i$ for individuals in the UTPAC.

Upon recruitment to each cohort patients are examined intermittently and we let $A_{i1} < A_{i2} < \cdots < A_{in_i}$ denote the calendar times of n_i follow-up assessments for individual i realized over $[A_0, A]$ where A is the date the databases are locked for analysis. If $E_{i1} \in [A_{i,j-1}, A_{ij}]$ for some $j = 1, \ldots, n_i$, then PsA is known to have developed, but it is subject to interval-censoring. We let $\mathcal{C}_i = [L_i, R_i)$ denote the interval containing T_i where $L_i = A_{i,j-1} - E_{i0}$ and $R_i = A_{ij} - E_{i0}$. When T_i is interval-censored $0 < L_i < R_i < \infty$, if it is right-censored $R_i = \infty$, and if T_i is observed then $L_i = R_i = T_i$. We take the dates of diagnosis of psoriatic arthritis in medical records as known; with respect to the onset time of PsA only the retrospective data are used from the UTPAC. If $Z_i = (Z_{i1}, \ldots, Z_{ip})'$ denotes a $p \times 1$ covariate vector associated with individual i, the observed data from individual i are denoted by $D_i = (\mathcal{A}_i, \mathcal{C}_i, Z_i)$ and the observed data for a pooled sample of size m is $D = \{D_i, i = 1, \ldots, m\}$. Interest lies in the relation between the covariates and the time of interest and we assess this by means of a proportional hazards model with $h(t|Z_i;\theta) = h_0(t;\alpha)\exp(Z_i'\beta)$ where α parameterizes the baseline hazard, $\beta = (\beta_1, \ldots, \beta_p)'$, and $\theta = (\alpha', \beta')'$. The survivor function is then $\mathcal{F}(t|Z_i;\theta) = \exp\{-H(t|Z_i;\theta)\}$ where $H(t|Z_i;\theta) = \int_0^t h(s|Z_i;\theta)ds$.

To discuss conditional independence assumptions regarding truncation we consider the setting of random left truncation times $\mathcal{L}_i, i = 1, \ldots, m$. Independent left truncation (Keiding and Moeschberger, 1992) implies

$$\lim_{\Delta t \downarrow 0} \frac{P(t \leq T_i < t + \Delta t | t \leq T_i, \mathcal{L}_i = \ell_i, Z_i)}{\Delta t} = h(t|Z_i;\theta) \ , \quad \ell_i < t \ ;$$

we assume this in what follows and comment briefly on when this might be violated in the Discussion. If $g_\ell(\ell_i)$ is the density of the left truncation time, then $\mathcal{L}_i$ is non-informative if the parameters of its density are functionally independent of θ; as a consequence modeling $\mathcal{L}_i$ would not offer any parametric information regarding θ and so we may omit likelihood contributions pertaining to the model for $\mathcal{L}_i$, $i = 1, \ldots, m$. Analogous independence conditions are assumed for the right truncated setting with these perhaps more naturally expressed in terms of the density function. The independence assumptions regarding the observation process that we make are given by Grüger et al. (1991)

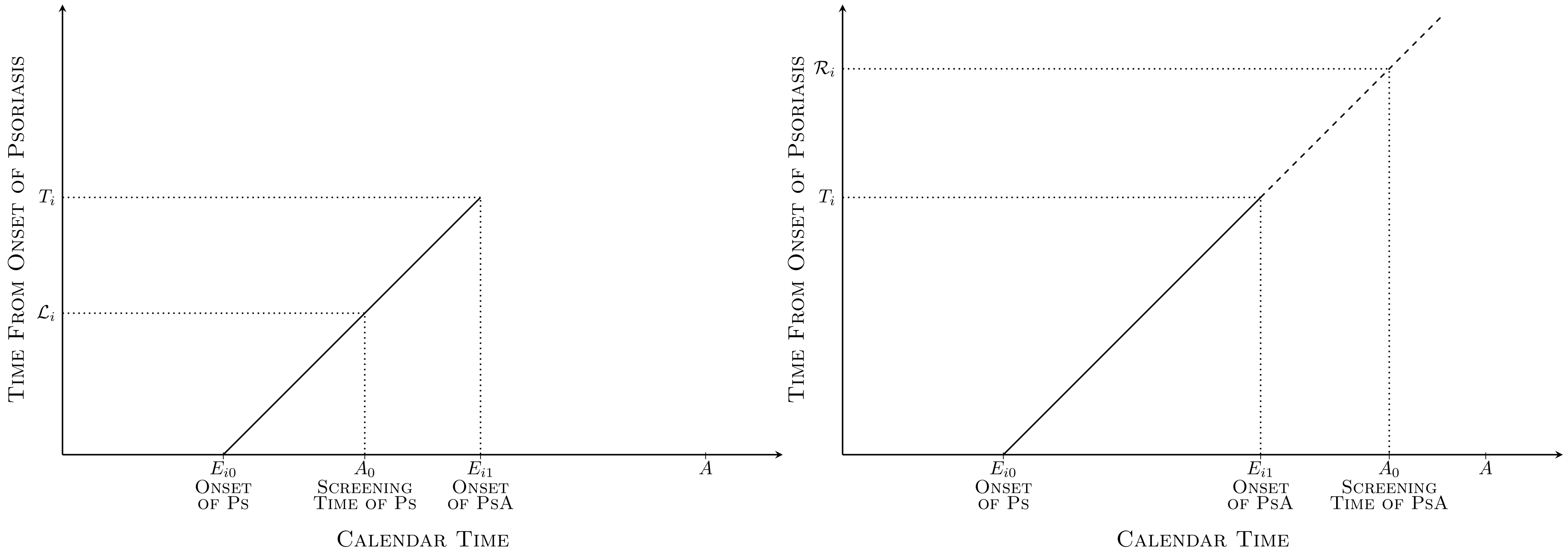

Figure 1: Lexis diagrams of the calendar times of onset of psoriasis (E_{i0}) and onset of psoriatic arthritis (E_{i1}) for individual i, along with screening times (A_0) for UTPC (left panel) and the UTPAC (right panel).

and the further assumption that this inspection process is non-informative justifies use of the observed (partial) likelihood

$$L(\theta) \propto \prod_{i=1}^{m} \frac{\mathcal{F}(L_i|Z_i;\theta) - \mathcal{F}(R_i|Z_i;\theta)}{\mathcal{F}(\mathcal{L}_i|Z_i;\theta) - \mathcal{F}(\mathcal{R}_i|Z_i;\theta)} \; .$$

When the dimension p is large it is customary to adopt some form of penalty for model complexity to help in the selection of important variables for further investigation. Most such penalized log-likelihoods can be written in the form

$$\ell_{\text{PEN}}(\theta) = \frac{1}{m} \log L(\theta) - p_{\gamma,\lambda}(\beta) \; , \tag{1}$$

where γ and λ are tuning parameters that determine the nature and extent of the penalty for complexity. Ridge regression (Hoerl and Kennard, 1970) is implemented with the L_2 penalty $p_{\gamma,\lambda}(\beta) = \lambda \sum_{j=1}^{p} \beta_j^2$ and the LASSO (Tibshirani, 1996) uses the L_1 penalty $p_{\gamma,\lambda}(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|$; there is no tuning parameter γ in these penalty functions. The value of the scalar λ is typically found by cross-validation (Shao, 1993) or generalized cross-validation (Golub et al., 1979). The adaptive LASSO uses adaptively weighted L_1 penalties of the form

$$p_{\gamma,\lambda}(\beta) = \sum_{j=1}^{p} \lambda_j |\beta_j| \; , \tag{2}$$

with small penalties λ_j chosen for large coefficients to reduce their shrinkage, and large penalties for small coefficients to address the selection objective (Zou, 2006). One option is to set $\lambda_j = \lambda/|\widetilde{\beta}_j|$, where $\widetilde{\beta} = (\widetilde{\beta}_1, \widetilde{\beta}_2, \ldots, \widetilde{\beta}_p)'$ is the maximum likelihood estimate (Zou, 2006, Zhang and Lu, 2007). Alternatively, the penalties can be updated iteratively. In this case, at the $(\ell+1)$st implementation, λ_j is set to $\lambda_j^{(\ell)} = \lambda/|\widetilde{\beta}_j^{(\ell)}|$ where $\widetilde{\beta}^{(\ell)}$ is obtained on the ℓth iteration; when $\ell = 0$, we set $\lambda_j^{(0)} = \lambda/|\widetilde{\beta}_j|$ as in the first implementation (Fan and Lv, 2010). We investigate the iterative implementation of the adaptive LASSO in the next section.

The smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001) is defined by

$$p'_{\gamma,\lambda}(\beta) = \lambda \sum_{j=1}^{p} \left\{ I(|\beta_j| \le \lambda) + \frac{(\gamma\lambda - |\beta_j|)_+}{(\gamma-1)\lambda} I(|\beta_j| > \lambda) \right\} \; ,$$

where $\gamma > 2$ and $y_+ = I(y \ge 0) \times y$. This penalty function is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$, but singular at 0 with its derivatives zero outside the range $[-\gamma\lambda, \gamma\lambda]$. Therefore, the SCAD penalty results in "small" coefficients being set to zero, "moderate" coefficients being shrunk towards zero, and "large" coefficients retained as they are. In principle, the optimal pair (γ, λ) could be obtained using a two dimensional grid search by cross validation or generalized cross validation. From empirical work, Fan and Li (2001) suggest $\gamma = 3.7$ is a reasonable choice for a variety of problems and we use this in what follows and select λ by (generalized) cross validation.

2.2 A Complete Data Likelihood for Truncated and Censored Data

We let J_i denote the number of "missing" individuals who have the same characteristics as the ith sampled individual except they did not satisfy the selection criteria (i.e. their event times fall in $\mathcal{A}_i^c$). We further let $T_{ij} \in \mathcal{A}_i^c$ be the event time of the jth unselected individual corresponding to individual i, so a Turnbull-type (Turnbull, 1976) complete data likelihood is

$$L_C(\theta) \propto \prod_{i=1}^{m} \left\{ h(T_i|Z_i;\theta) \exp(-H(T_i|Z_i;\theta)) \prod_{j=1}^{J_i} h(T_{ij}|Z_i;\theta) \exp(-H(T_{ij}|Z_i;\theta)) \right\} \; .$$

The reason for considering this form is that by introducing the unobserved failure times and adopting a weakly parametric piecewise constant baseline hazard model via an EM algorithm (Dempster et al., 1977), the maximization step of the complete data likelihood will be simplified.

Under a weakly parametric piecewise constant baseline hazard function, the number and location of break-points at which the baseline hazard changes value must be specified. If $0 = b_0 < b_1 < \cdots < b_{K-1} < b_K = \infty$ denote K break-points, we let $h_0(s; \alpha) = \exp(\alpha_k)$, for $s \in \mathcal{B}_k = [b_{k-1}, b_k)$, $k = 1, \ldots, K$. Let $D_k(u) = I(u \in \mathcal{B}_k)$ denote whether or not the time u is in the interval $\mathcal{B}_k$ and $W_k(u) = \int_0^u D_k(s)ds$ denote the duration of $[0, u)$ over interval k, $k = 1, \ldots, K$. Then under the piecewise constant model and given a covariate vector Z_i, the complete data log-likelihood would be

$$\log L_C(\theta) \propto \sum_{i=1}^{m} \sum_{k=1}^{K} \Bigg\{ D_k(T_i)\,(\alpha_k + Z_i'\beta) - W_k(T_i)\exp(\alpha_k + Z_i'\beta) \\ + \sum_{j=1}^{J_i} \left[D_k(T_{ij})\,(\alpha_k + Z_i'\beta) - W_k(T_{ij})\exp(\alpha_k + Z_i'\beta)\right] \Bigg\} . \tag{3}$$

If $X_{ik\ell} = I(k = \ell)$ and $X_{ik} = (X_{ik1}, \ldots, X_{ikK})'$ denotes the corresponding vector of indicator functions, $k = 1, \ldots, K$; thus $X_{i1} = (1, 0, \ldots, 0)'$, $X_{i2} = (0, 1, \ldots, 0)'$, ..., $X_{ik} = (0, 0, \ldots, 1)'$. Then if $\alpha = (\alpha_1, \ldots, \alpha_K)'$ and $\theta = (\alpha', \beta')'$, we can write

$$\log L_C(\theta) = \sum_{i=1}^{m} \log L_{Ci}(\theta) ,$$

where upon letting $\bar{Z}_{ik} = (X_{ik}', Z_i')'$ we can write $\log L_{Ci}(\theta)$ as

$$\sum_{k=1}^{K} \left\{ D_k(T_i)\bar{Z}_{ik}'\theta - W_k(T_i)\exp(\bar{Z}_{ik}'\theta) + \sum_{j=1}^{J_i} \left[D_k(T_{ij})\bar{Z}_{ik}'\theta - W_k(T_{ij})\exp(\bar{Z}_{ik}'\theta)\right] \right\} .$$

At the E-step of the EM algorithm, the conditional expectation of the penalized complete data log-likelihood function at the $(r + 1)$st iteration is evaluated as

$$Q_{\text{PEN}}(\theta; \theta^{(r)}) = \sum_{i=1}^{m} Q_i(\theta; \theta^{(r)}) - p_{\gamma,\lambda}(\beta) , \tag{4}$$

where $Q_i(\theta; \theta^{(r)}) = E\left\{\log L_{Ci}(\theta)|D; \theta^{(r)}\right\}$ and $\theta^{(r)}$ is estimated by maximizing $Q_{\text{PEN}}(\theta; \theta^{(r-1)})$. The required conditional expectations are therefore $\widehat{\Delta}_{ik}^{(r)} = E[D_k(T_i)|D_i; \theta^{(r)}]$, $\widehat{\mathcal{S}}_{ik}^{(r)} = E[W_k(T_i)|D_i; \theta^{(r)}]$, $\widehat{\iota}_{ik}^{(r)} = E[D_k(T_{ij})|D_i; \theta^{(r)}]$, $\widehat{\omega}_{ik}^{(r)} = E[W_k(T_{ij})|D_i; \theta^{(r)}]$ and $\mathcal{J}_i^{(r)} = E[J_i|D_i; \theta^{(r)}]$.

Let $\mathcal{C}_{ik} = \mathcal{C}_i \cap \mathcal{B}_k = [L_{ik}, R_{ik})$ denote the sub-interval of the censoring interval $\mathcal{C}_i$ contained within $\mathcal{B}_k$. When $\mathcal{C}_{ik} = \emptyset$, the required expectations are relatively easy to compute since, for instance, it is clear that $D_k(t_i) = 0$ and $\widehat{\Delta}_{ik}^{(r)} = 0$. Moreover, if $b_k < L_i$, then it is known that individual i was at risk for the entire interval $\mathcal{B}_k$ so $W_k(t_i) = \widehat{\mathcal{S}}_{ik}^{(r)} = b_k - b_{k-1}$, and if $R_i < b_{k-1}$, then $W_k(t_i) = \widehat{\mathcal{S}}_{ik}^{(r)} = 0$ since they are known to have failed prior to the start of interval $\mathcal{B}_k$. If $\mathcal{C}_{ik} \neq \emptyset$,

$$\widehat{\Delta}_{ik}^{(r)} = \frac{\mathcal{F}(L_{ik}|Z_i; \theta^{(r)}) - \mathcal{F}(R_{ik}|Z_i; \theta^{(r)})}{\mathcal{F}(L_i|Z_i; \theta^{(r)}) - \mathcal{F}(R_i|Z_i; \theta^{(r)})} , \tag{5}$$

$$\widehat{\mathcal{S}}_{ik}^{(r)} = \max(L_i - b_{k-1}, 0) + \int_{L_{ik}}^{R_{ik}} \frac{\mathcal{F}(s|Z_i; \theta^{(r)})}{\mathcal{F}(L_i|Z_i; \theta^{(r)}) - \mathcal{F}(R_i|Z_i; \theta^{(r)})} \mathrm{d}s , \tag{6}$$

where $\mathcal{F}(t|Z_i; \theta) = \exp\left\{-\left(\sum_{k=1}^{K} \exp(\alpha_k)W_k(t)\right)\exp(Z_i'\beta)\right\}$.

Let $\mathcal{A}_{ik} = \mathcal{A}_i^c \cap \mathcal{B}_k = [\mathcal{L}_{ik}, \mathcal{R}_{ik})$ be the sub-interval of the complement of the truncation interval $\mathcal{A}_i^c$ contained within $\mathcal{B}_k$, if $\mathcal{A}_{ik} = \emptyset$, then $\widehat{\iota}_{ik}^{(r)} = 0$. Moreover, if $\mathcal{A}_i^c = [0, \mathcal{L}_i)$, then $\widehat{\omega}_{ik}^{(r)} = 0$ since they are known to have failed prior to the start of interval, and if $\mathcal{A}_i^c = (\mathcal{R}_i, \infty)$, then the individual i was at risk for the entire interval $\mathcal{B}_k$ so $\widehat{\omega}_{ik}^{(r)} = b_k - b_{k-1}$. If $\mathcal{A}_{ik} \neq \emptyset$,

$$\widehat{\iota}_{ik}^{(r)} = \frac{\mathcal{F}(\mathcal{L}_{ik}|Z_i;\theta^{(r)}) - \mathcal{F}(\mathcal{R}_{ik}|Z_i;\theta^{(r)})}{1 - \mathcal{F}(\mathcal{L}_i|Z_i;\theta^{(r)}) + \mathcal{F}(\mathcal{R}_i|Z_i;\theta^{(r)})}\ , \tag{7}$$

$$\widehat{\omega}_{ik}^{(r)} = \mathcal{L}_{ik} - b_{k-1} + \int_{\mathcal{L}_{ik}}^{\mathcal{R}_{ik}} \frac{\mathcal{F}(s|Z_i;\theta^{(r)})}{1 - \mathcal{F}(\mathcal{L}_i|Z_i;\theta^{(r)}) + \mathcal{F}(\mathcal{R}_i|Z_i;\theta^{(r)})} \mathrm{d}s\ . \tag{8}$$

Also

$$\widehat{\mathcal{J}}_i^{(r)} = E[J_i|D_i;\theta^{(r)}] = \frac{1 - \mathcal{F}(\mathcal{L}_i|Z_i;\theta^{(r)}) + \mathcal{F}(\mathcal{R}_i|Z_i;\theta^{(r)})}{\mathcal{F}(\mathcal{L}_i|Z_i;\theta^{(r)}) - \mathcal{F}(\mathcal{R}_i|Z_i;\theta^{(r)})}\ . \tag{9}$$

Given these results, (4) can be written more explicitly as

$$\sum_{i=1}^{m}\sum_{k=1}^{K}\left\{\left[\widehat{\Delta}_{ik}^{(r)}\bar{Z}'_{ik}\theta - \widehat{\mathcal{S}}_{ik}^{(r)}\exp(\bar{Z}'_{ik}\theta)\right] + \widehat{\mathcal{J}}_i^{(r)}\left[\widehat{\iota}_{ik}^{(r)}\bar{Z}'_{ik}\theta - \widehat{\omega}_{ik}^{(r)}\exp(\bar{Z}'_{ik}\theta)\right]\right\} - p_{\gamma,\lambda}(\beta)\ . \tag{10}$$

Since (10) has the form of a penalized Poisson likelihood, the M-step can be carried out using software for penalized Poisson regression. This can be implemented by creating an augmented pseudo-dataset with individual i contributing up to K lines with weight 1 and K lines (for the corresponding unselected individuals) with weight $\widehat{\mathcal{J}}_i^{(r)}$, $i = 1, \ldots, m$.

Classical variable selection methods are often based on the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), while more recently cross-validation (CV) and generalized cross-validation (GCV) techniques have been advocated. The traditional G-fold CV statistic is defined as $\widehat{CV}(\lambda) = \sum_{g=1}^{G}[\log L(\widehat{\theta}_{-g}(\lambda)) - \log L_{-g}(\widehat{\theta}_{-g}(\lambda))]$ where L_{-g} is the observed data likelihood for the gth training dataset and $\widehat{\theta}_{-g}(\lambda)$ is the estimate for the gth training data, obtained through the EM algorithm; the optimal λ maximizes $\widehat{CV}(\lambda)$.

2.3 Design and Interpretation of Simulation Studies

We consider a sample size of $m = 1200$ with $m_1 = 400$ of the subjects left-truncated and $m_2 = 800$ right-truncated with the number of covariates is set to $p = 100$. We consider binary covariates with $P(Z_{ij} = 1) = 0.5$, $i = 1, \ldots, m, j = 1, \ldots, p$. There are eight covariates specified to have coefficients not equal to zero and all other covariate effects were set to zero, that is $\beta_j = \log(2) = 0.6931$, $j = 1, 2, 9, 10$ and $\beta_j = \log(0.5) = -0.6931$, $j = 17, 18, 19, 20$ and $\beta_j = 0$, otherwise. The conditional hazard for T_i is based on a Weibull regression model where

$$h(t|Z_i;\theta) = \kappa\eta(\eta t)^{\kappa-1}\exp(Z_i'\beta)\ ,$$

where $\kappa = 1.25$. We consider a study with median event time equal to 1, thus for each of $\kappa = 1$ and 1.25, we solve for η so that

$$P(T_i < 1;\theta) = E_Z\left[P(T_i < 1|Z;\theta)\right] = 0.5\ .$$

Let t_{Q25}, t_{Q50} and t_{Q75} be the quartiles of the marginal distribution of T_i and the truncation times are drawn from these quartiles with equal probabilities. For each subject i, it has either a left-truncated right-censored event time (Ps cohort) or a right-truncated event time (PsA cohort).

For the ith subject, $i = 1, \ldots, m_1$, which are subject to left truncation, we generate the left truncation time $\mathcal{L}_i$ which is randomly drawn from the quartiles with equal probabilities. To ensure the sample covariate distribution is compatible with the truncation scheme, we generate Z_i using the conditional distribution $P(Z_i|T_i > \mathcal{L}_i)$. We then generate $U_i \sim \text{Uniform}(0,1)$ and solve for the event time T_i that satisfies $P(T \geq T_i|T \geq \mathcal{L}_i, \mathcal{L}_i = l_i, Z_i; \theta) = U_i$. For the ith subject, $i = m_1 + 1, \ldots, m$, whose times are subject to right truncation, we generate the right truncation time $\mathcal{R}_i$ uniformly from the quartiles, Z_i is generated from $P(Z_i|T_i < \mathcal{R}_i)$, solve for T_i in the constraint $P(T < T_i|T < \mathcal{R}_i, \mathcal{R}_i = r_i, Z_i; \theta) = U_i$ where $U_i \sim \text{Uniform}(0,1)$. We consider this study with duration of follow-up planned to be $\tau = A - A_0$, where τ is obtained from $P(T \geq \mathcal{L}_i + \tau|T \geq \mathcal{L}_i; \theta) = 0.5$. For simplicity, we consider a fixed number of inspections $n_i = 5$, $i = 1, \ldots, m$, and the follow-up inspection times are generated uniformly from $[\mathcal{L}_i, \mathcal{L}_i + \tau]$, $j = 1, \ldots, 5$, $i = 1, \ldots, m$.

For each dataset, variable selection was carried out based on the penalized EM (P-EM) algorithm of Section 2.2 with the LASSO, adaptive LASSO (ALASSO) and SCAD ($\gamma = 3.7$) penalty functions. The tuning parameter was selected in each case using the AIC, the BIC or using a 5-fold cross-validation statistic. Analyses were conducted based on proportional hazards models with a piecewise constant baseline hazards; hazard functions with four pieces (PWC-4) where the break-points were based on the quantiles of the baseline survival function.

Table 1 displays the performance of LASSO, ALASSO and SCAD for each method of selecting the tuning parameter in the setting with some trend in the baseline hazard and for a time homogeneous model. The probability that an important variable is appropriately selected is generally very high for all methods, but false positive rates are quite high under the LASSO penalty regardless of how the tuning parameter is selected; all methods have high false positive rates when AIC is used for the selection of the tuning parameter. The ALASSO and SCAD penalty functions perform very well when the tuning parameter is selected by BIC or 5-fold cross-validation; the performance is slightly better for the CV than with the BIC criterion.

2.4 HLA Markers for the Development of PsA in Individuals with Ps

The data from the UTPC and UTPAC are comprised of 338 and 603 individuals with left- and right-truncated PsA onset times respectively along with data on 76 human leukocyte antigen (HLA) markers. Among the 338 individuals in the UTPC cohort 38 yielded onset dates for psoriatic arthritis. Given the high false positive selection rate of the LASSO and of all methods when the tuning parameter is selected based on the AIC criterion, in this application we use the ALASSO and SCAD penalty functions and select the tuning parameter based on the BIC and 5-fold CV statistic. The basic model involves a piecewise (4-piece) constant baseline hazard and all models control for age and gender.

The findings based on the BIC suggest HLA-DRB-16 is protective for the development of PsA with coefficients estimated as -0.9284 for the ALASSO and -0.9317 for the SCAD penalty functions. When the 5-fold CV statistic is used to select the tuning parameter, we find HLA-DRB1-10 and HLA-DRB-16 are both identified using ALASSO (coefficient estimates of -0.7144 and -0.9749 respectively) and SCAD (-0.7160 and -0.9771 respectively).

3 Estimating Predictive Accuracy for Interval-Censored Data

3.1 Estimating Predictive Accuracy

Here we consider the problem of estimating the predictive error with interval-censored untruncated data, wherein $\mathcal{L}_i = 0$ and $\mathcal{R}_i = \infty$, $i = 1, \ldots, m$; we omit the subscript i in the discussion that follows. The scientific goal is to identify which, among the 76 human leukocyte antigen markers, are associated with the development of arthritis mutilans. While there is no clinical agreement on how to precisely define arthritis mutilans, it represents a state of significant joint damage arising from an extreme form of the disease; here we define it as present if an individual has 4 or more joints with the advanced stage of damage according to the modified Steinbrocker score. Data from

Table 1: Empirical results for dataset (33.33% left truncation and 66.67% right truncation) with binary covariates ($p = 100$, $P(Z_{ij} = 1) = 0.5$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE). The tuning parameter is selected by AIC, BIC or CV. Sample size $m = 1200$, and $nsim = 100$ simulations.

	AIC			*BIC*			*CV*		
Method	TP(8)	FP(92)	MSE (SD)	TP(8)	FP(92)	MSE (SD)	TP(8)	FP(92)	MSE (SD)
Shape parameter $\kappa = 1.25$									
Interval-Censored Left-Truncated and Right-Truncated Data									
LASSO	8.00	15.95	0.120 (0.042)	7.94	3.02	0.224 (0.087)	8.00	16.37	0.120 (0.041)
ALASSO	8.00	10.40	0.146 (0.051)	8.00	0.52	0.034 (0.022)	7.98	0.41	0.029 (0.032)
SCAD	8.00	10.44	0.149 (0.051)	8.00	0.54	0.040 (0.021)	7.98	0.34	0.029 (0.030)
Right-Censored Left-Truncated and Right-Truncated Data									
LASSO	8.00	17.60	0.113 (0.039)	7.98	2.94	0.224 (0.078)	8.00	16.06	0.119 (0.039)
ALASSO	8.00	10.86	0.152 (0.054)	8.00	0.51	0.035 (0.021)	7.96	0.27	0.026 (0.031)
SCAD	8.00	10.61	0.155 (0.052)	8.00	0.50	0.036 (0.020)	7.98	0.26	0.028 (0.025)
Shape parameter $\kappa = 1$									
Interval-Censored Left-Truncated and Right-Truncated Data									
LASSO	8.00	19.69	0.098 (0.030)	8.00	3.27	0.181 (0.065)	8.00	17.09	0.105 (0.032)
ALASSO	8.00	11.42	0.165 (0.063)	8.00	0.61	0.030 (0.028)	8.00	0.47	0.023 (0.032)
SCAD	8.00	11.37	0.166 (0.062)	8.00	0.65	0.032 (0.030)	7.97	0.53	0.023 (0.042)
Right-Censored Left-Truncated and Right-Truncated Data									
LASSO	8.00	19.21	0.098 (0.030)	8.00	3.38	0.180 (0.070)	8.00	17.09	0.102 (0.033)
ALASSO	8.00	11.49	0.163 (0.063)	8.00	0.64	0.031 (0.029)	7.97	0.48	0.023 (0.042)
SCAD	8.00	11.42	0.160 (0.063)	8.00	0.63	0.029 (0.029)	7.99	0.46	0.024 (0.036)

604 members of the UTPAC are used here; the median time from the diagnosis of PsA to the last assessment is 12.48 years (lower quartile = 5.06, upper quartile = 21.49). A total of 109 are known to have developed arthritis mutilans between two follow-up assessments. The quartiles of the length of the closed censoring intervals are 2.48, 8.06 and 14.34 years respectively, which are wider than one might expect from a protocol in which radiological assessments are to be scheduled every two years, but this is due to the variation between individuals in their tendency to attend the clinic. We begin with a development of the methods for assessing prediction accuracy.

In the context of time to event data, we obtain flexible prediction models and often evaluate their predictive values on the same set of data, or a validation data. The purpose of assessing the predictive accuracy of a regression model is often to establish whether a prognostic model can be used to reliably predict patients survival status and provide a basis for clinical decision making. Predictive accuracy can also be used as a strategy for model selection. The prediction of event times and survival status at a particular time has been considered by many authors, and we focus on the latter. Loss functions measure the distance between predicted and true values such that predictive values with small loss are good. The absolute error loss function is $L(Y, \widehat{Y}) = |Y - \widehat{Y}|$ and the squared error loss function is $L(Y, \widehat{Y}) = (Y - \widehat{Y})^2$ assigns different losses to particular errors. An overall measure of prediction error is obtained by averaging the loss function over all possible values of the data. An optimal predictor is defined to be the predictor that minimizes the predictor error.

The difficulty in assessing predictive accuracy due to censoring has been studied by several authors. Korn and Simon (1990) proposed a bounded loss function to be used for predicting survival time, whereas Graf et al. (1999), Hothorn et al. (2006), Gerds and Schumacher (2006), Lawless and Yuan (2010) model the censoring distribution and use inverse probability weighting (IPW) to deal with censored outcomes. If we focus on the binary indicator of the event status at a particular time t_0, then we can consider predictors in the following class

$$\widehat{Y}(X; \theta) = I(P(T > t_0|Z) > c; \theta) \ ,$$

where optimal binary predictor uses $c = 0.5$; again the inverse probability weights can be used based on models for the censoring distribution.

When individuals are only assessed intermittently it is also necessary to model the visit process. We consider a single individual and let $0 = V_0 < V_1 < \cdots < V_n$ denote the time of assessments since the onset of PsA, and let $N(u) = \sum_{r=1}^{\infty} I(V_r \leq u)$ count the number of assessments at time u. Let C be a random drop-out time and $C(u) = I(u \leq C)$ indicate whether this individual is in cohort or not at time u. We also let T be the event time, Z be a $p \times 1$ fixed covariate vector, $\{X(s), 0 < s\}$ be a time-dependent covariate process, $\bar{X}(u) = \{X(a_1), \ldots, X(a_{N(u^-)})\}$ be the history of the observed value at time $u > 0$, and $\bar{W}(u) = \{W(a_1), \ldots, W(a_{N(u^-)})\}$ denote the recorded event status at the process assessment here where $W(u) = I(T < u)$. The complete history observed at time s is then $\mathcal{H}(s) = \{(dN(u), C(u)), 0 < u < s, Z, \bar{X}(s), \bar{W}(s)\}$. Since the goal is to use genetic data to predict the development of PsA, it is inappropriate to control for time-varying markers in the casual pathway in the model for the response process. Here we adopt a simple hazard function of the form

$$\lim_{\Delta t \downarrow 0} \frac{P(T < t + \Delta t | T \geq t, Z)}{\Delta t} = I(t \leq T) h(t|Z) \ .$$

The intensities for the inspection and censoring processes are meant to provide good representations of the data and so conditioning on all available data is appropriate; we let

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta N(t) = 1 | \mathcal{H}(t))}{\Delta t} = C(t) \lambda(t | \mathcal{H}(t)) \ ,$$

$$\lim_{\Delta t \downarrow 0} \frac{P(C < t + \Delta t | \mathcal{H}(t))}{\Delta t} = C(t) \lambda^c(t | \mathcal{H}(t))$$

represent the inspection and censoring intensities.

We define a multistate process $\{Z(s), 0 < s\}$ with a state space $\mathfrak{S} = \{\mathbb{V}_0, \mathbb{V}_1, \dots, \mathbb{V}_1^E, \mathbb{V}_2^E, \dots,$ $\mathbb{C}_1, \mathbb{C}_2, \dots, \mathbb{C}_1^E, \mathbb{C}_2^E, \dots, \mathbb{E}\}$ for joint consideration of event, inspection and censoring processes, shown as Figure 2. Here $\mathbb{E}$ denotes the event, $\mathbb{V}_r$ denotes the rth assessment without having disease, $\mathbb{V}_r^E$ denotes the rth assessment after having disease, $\mathbb{C}_r$ denotes the random dropout after $(r-1)$th assessments and without having disease, and $\mathbb{C}_r^E$ denotes the random dropout after $(r-1)$th assessments and after having disease, note here we use a superscript E to denote the states after $\mathbb{E}$.

Following the occurrence of any event in $\mathfrak{S}$, the next event to occur is governed by a competing risk process. The cause-specific intensities are shown in the figure. We give a subscript E for the intensity post-disease, that is, the intensity for inspection post-disease is $\lambda_E(t)$ and the intensity for random dropout post-disease is $\lambda_E^c(t)$. If the event process is independent of the inspection process, then $\lambda_E(t) = \lambda(t)$, otherwise, we may assign a different intensity, such as $\lambda_E(t) = \lambda(t)\exp(\alpha^s)$. Similarly, the event process is independent of the censoring process, then $\lambda_E^c(t) = \lambda^c(t)$, otherwise, we may assign a different intensity, such as $\lambda_E^c(t) = \lambda^c(t)\exp(\alpha^c)$.

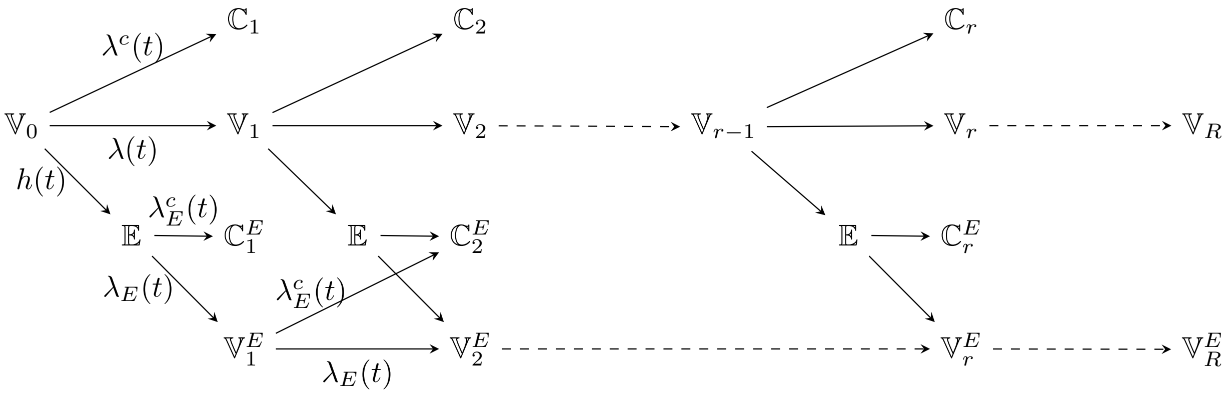

Figure 2: A multistate diagram for joint consideration of event, random drop-out and assessment times.

Figure 3 shows all the possible combinations of (Y, Ψ). The IPW estimator of the prediction error is

$$\widehat{\text{PE}}(t_0) = \frac{1}{m}\sum_{i=1}^{m} \frac{\Psi_i}{E(\Psi_i|Y_i, Z_i, D)} \left\{Y_i - \widehat{Y}_i(Z_i; \widehat{\theta})\right\}^2 ,$$

where $\Psi_i = I(Y_i \text{ is known}) = I(t_0 \notin [L_i, R_i])$ and the weight is $E(\Psi_i|Y_i, Z_i, D)$ which is the conditional expectation of Ψ_i given (Y_i, Z_i, D). Here Ψ_i depends on the inspection process, the censoring process and the event process. The weight can then be written as

$$E(\Psi_i|Y_i, Z_i, D) = E_{dN,C,T|Y,X,D}\left[\Psi_i\right] = P(\Psi_i = 1|Y_i, Z_i, D) \tag{11}$$

Expectations Under the Condition $T \leq t_0$

After the occurrence of disease $\mathbb{E}$ (i.e. entry to an $\mathbb{E}$ state), the next event to occur can be a visit (i.e. entry to a $\mathbb{V}_r^E$ state) or censoring (i.e. entry to a $\mathbb{C}_r^E$ state). If Y is known, then a post-disease assessment must be observed before t_0, so $P(\Psi_i = 1|Y_i = 0, Z_i, D) = P(\Psi_i = 1|T_i \leq t_0, Z_i, D)$ can be written as

$$\begin{aligned}\int_0^{t_0} \left[\int_t^{t_0} \lambda_E(u|\mathcal{H}(u)) \exp\left\{-\int_t^u \lambda_E(v|\mathcal{H}(v)) + \lambda_E^c(v|\mathcal{H}(v))\mathrm{d}v\right\} \mathrm{d}u\right] \\ \times f(t|T_i \leq t_0, Z_i) \exp\left\{-\int_0^t \lambda^c(s|\mathcal{H}(s))\mathrm{d}s\right\} \mathrm{d}t \,. \end{aligned} \tag{12}$$

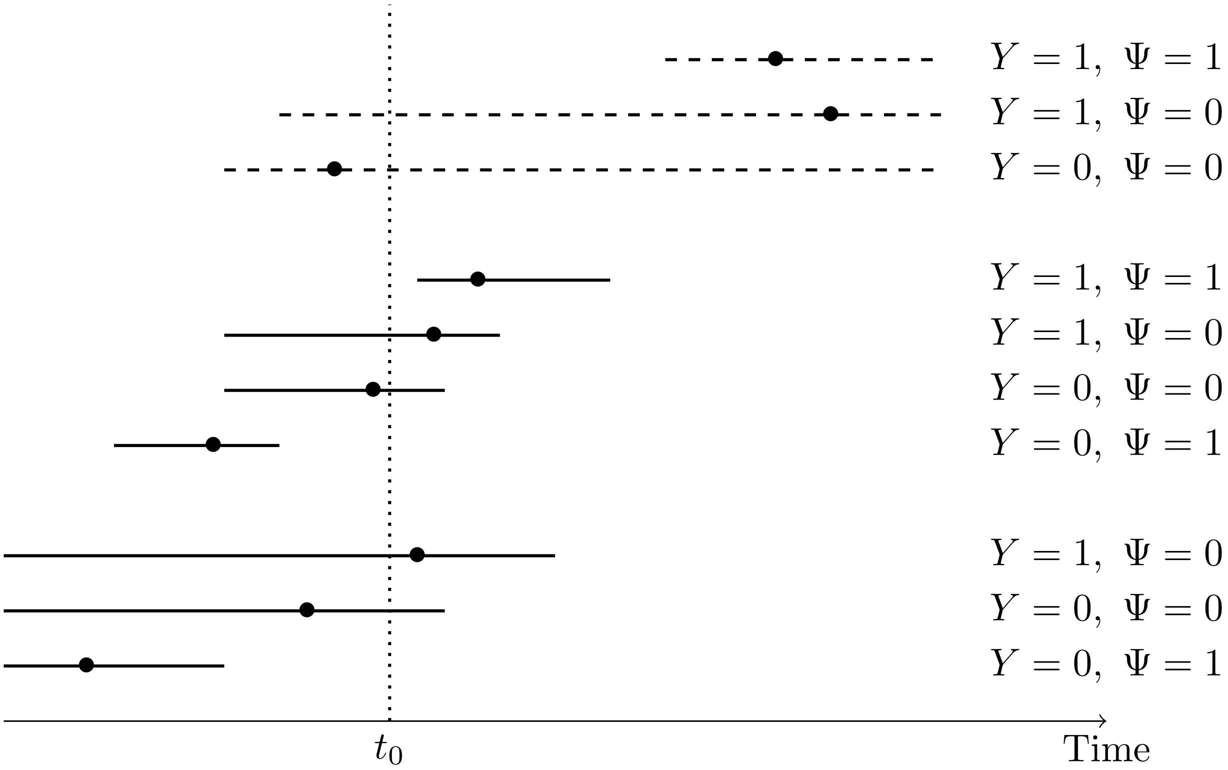

Figure 3: All Possible Combinations of (Y, Ψ), the solid lines denote observing the event, that is, either left-censored or interval-censored, the dashed lines denote right-censoring without observing the occurrence of event. The solid dots denote the (unobserved) exact event times.

If we assume $\lambda_E(t) = \lambda(t)$ and $\lambda^c_E(t) = \lambda^c(t)$, then (12) becomes

$$\int_0^{t_0} \left[\int_t^{t_0} \lambda(u|\mathcal{H}(u)) \exp\left\{ -\int_t^u \lambda(v|\mathcal{H}(v))\ \mathrm{d}v - \int_0^u \lambda^c(v|\mathcal{H}(v))\mathrm{d}v \right\} \mathrm{d}u \right] \\ \times f(t|T_i \leq t_0, Z_i)\ \mathrm{d}t\ .$$

Expectations Under the Condition $T > t_0$
If Y is known to be one, then there must be an assessment without disease after t_0, which can be represented by an entry to a $\mathbb{V}_r$ state, $r = 1, 2, \ldots$. Therefore $Z(t_0^-) = \mathbb{V}_{r-1}$ and the next event to occur can be $\mathbb{V}_r$, $\mathbb{C}_r$ or $\mathbb{E}$. In this case, $P(\Psi_i = 1|Y_i = 1, Z_i, D) = P(\Psi_i = 1|T_i > t_0, Z_i, D)$ is

$$\int_{t_0}^{\infty} \lambda(u|\mathcal{H}(u)) \exp\left[-\int_{t_0}^u \{\lambda(v|\mathcal{H}(v)) + h(v|Z)\}\ \mathrm{d}v - \int_0^u \lambda^c(v|\mathcal{H}(v))\ \mathrm{d}v \right] \mathrm{d}u \qquad (13)$$

Another approach to examining the performance of a classification scheme is to use a receiver operating characteristic (ROC) curve. When the response is a binary indicator of the survival status at a specific time t_0, a key component of assessing the predictive performance is the ability to correctly classify individuals with respect to their status at time t_0, which can then be quantified through construction of a receiver operating characteristic (ROC) curve, which plots the true positive rate (sensitivity) against the false positive rate (1 - specificity). Akritas (1994) proposed an estimator based on a nearest neighbor estimator for the bivariate distribution function $P(\widehat{Y}, Y)$, which can guarantee the monotonicity of sensitivity and specificity; an alternative simple estimator based on the sensitivity and specificity using the Kaplan-Meier estimate do not satisfy the necessary condition of the monotonicity (Heagerty et al., 2000). Lawless and Yuan (2010) discussed an estimator based on the inverse probability weighting approach when the event time is right-censored and this IPW approach guarantees monotonicity.

The true positive rate (TPR) and false positive rate (FPR) are defined as

$$\begin{aligned} \text{TPR}(c) &= P(\widehat{Y}=1|Y=1) = \frac{P(\mathcal{F}(t_0|Z;\theta) > c, T > t_0)}{P(T > t_0)} , \\ \text{FPR}(c) &= P(\widehat{Y}=1|Y=0) = \frac{P(\mathcal{F}(t_0|Z;\theta) > c, T \le t_0)}{P(T \le t_0)} . \end{aligned} \tag{14}$$

Similarly, we can estimate those probabilities using inverse weighting, as for example

$$\widehat{P}(\mathcal{F}(t_0|Z;\theta) > c, T > t_0) = \frac{1}{m}\sum_{j=1}^{m} \frac{\Psi_j}{E(\Psi_j|Y_j, Z_j, D)} I(\mathcal{F}(t_0|Z_j;\widehat{\theta}) > c, T_j > t_0) .$$

The ROC curve is obtained by plotting TPR(c) against FPR(c) for values of c increasing from 0 to 1. The best possible prediction method would yield a point in the upper left corner at coordinate (0,1) of the ROC space (representing 100% sensitivity and 100% specificity). While a point along a diagonal line (the so-called line of no-discrimination) corresponds to a prediction scheme no better than a random guess. The area under curve (AUC) is a summary measure of ROC curve, which is equal to the probability that a predictor will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative').

3.2 Simulation Studies

We consider the setting with three covariates denoted Z_{i1}, Z_{i2} and Z_{i3}. In one scenario they have marginal standard normal distributions with $Z_{i1} \perp Z_{i2}$, $Z_{i1} \perp Z_{i3}$, and $\text{corr}(Z_{i2}, Z_{i3}) = 0$ or 0.5. In a second scenario the covariates are binary with $P(Z_{ij} = 1) = 0.5$, $j = 1, 2, 3$. The event time T_i follows a Weibull distribution given (Z_{i1}, Z_{i2}) with $\beta_1 = \log(2)$, $\beta_2 = \log(1.5)$ and shape $\kappa = 1.25$; that is,

$$\mathcal{F}(t|Z_{i1}, Z_{i2};\theta) = \exp\left\{-(\lambda t)^{\kappa} \exp\left(Z_{i1}\beta_1 + Z_{i2}\beta_2\right)\right\} ,$$

where $\theta = (\lambda, \kappa, \beta_1, \beta_2)'$; the value of λ is determined so that $P(T > 1) = 0.5$, where $P(T > 1) = E\{\mathcal{F}(t|Z_{i1}, Z_{i2};\theta)\}$. We consider an administrative censoring time τ such that $\mathcal{F}(\tau) = 0.9$. A time homogeneous Poisson process is used for the inspection process with rate

$$\lambda(s|Z_{i1}, Z_{i3};\gamma) = \exp(\gamma_0 + Z_{i1}\gamma_1 + Z_{i3}\gamma_2) ,$$

where $\gamma_1 = \log(1.1)$ and $\gamma_2 = \log(1.5)$ for the normal covariates and $\gamma_1 = \log(2)$ and $\gamma_2 = \log(2.5)$ for the binary covariates; γ_0 is determined to ensure that the average number of assessments by τ is controlled at $\mu = 10$ where $\mu = E\{\int_0^{\tau} \lambda(s|Z_{i1}, Z_{i3};\gamma)\mathrm{d}s\}$.

Let $0 = v_0 < v_1 < \ldots < v_n \le \tau$ denote the inspection times, then the left and right endpoints of the censoring interval are then $L = \max(v_r \cdot I(v_r < t))$ and $R = \min(v_r \cdot I(v_r > t))$ respectively. In the application there is no recorded right censoring time and so the expressions of (12) and (13) simplify to the following expressions (15) and (16)

$$\int_0^{t_0}\left[\int_t^{t_0} \lambda(u|\mathcal{H}(u)) \exp\left\{-\int_t^{u} \lambda(v|\mathcal{H}(v))\, \mathrm{d}v\right\} \mathrm{d}u\right] \times f(t|T_i \le t_0, Z_i)\, \mathrm{d}t , \tag{15}$$

$$\int_{t_0}^{\tau} \lambda(u|\mathcal{H}(u)) \exp\left[-\int_{t_0}^{u} \{\lambda(v|\mathcal{H}(v)) + h(v|X)\}\, \mathrm{d}v\right] \mathrm{d}u . \tag{16}$$

Thus the weights are estimated by modeling the event and inspection processes as described in the discussion of the simulation study. Datasets with sample sizes of $m = 500$ are simulated 100 times ($nsim = 100$) for each scenario. For each simulated dataset, parametric analyses were carried out to model the event time by using a Weibull distribution and to model the gap times between two consecutive inspection times by an exponential distribution. The empirical bias (EBIAS) and the empirical standard error (ESE) of the unweighted and weighted (IPW) estimators of the prediction error at time t_0 are summarized in Table 2, where t_0 values are taken to be the quartiles of the marginal distribution of T. The proposed IPW estimators have relatively small biases compared to the unweighted estimators, while the variability (in terms of ESE) is greater. The misspecification of inspection model was next investigated by omitting one important covariate. In broad terms we found, as one would expect, that there was a consequent increase in the empirical bias, but that this remains smaller than that of the unweighted estimators for the misspecifications considered here.

3.3 Application to the Psoriatic Arthritis Cohort

Our interest lies in identifying which among the 76 HLA markers are associated with increased risk of developing arthritis mutilans from the time of diagnosis with psoriatic arthritis; we are also interested in assessing the predictive performance of the models obtained by penalized regression through application of the methods in Section 3.1. We adopt a proportional hazards model with a piecewise constant (5-piece) baseline hazard with cut points at years 6.5, 10.5, 18 and 22. All models controlled for age and sex. Given the superior performance of the penalized methods based on the ALASSO and SCAD penalty functions, we focus on these with the tuning parameter selected based on the BIC and 5-fold cross-validation statistic. The findings in Table 3 are slightly more variable than in the previous application; with the ALASSO we find no markers when the BIC is used, but seven are selected when the 5-fold cross-validation statistic is used. For the SCAD penalty function, three HLA markers (HLA-A29, HLA-B27 and HLA-DQB1-02) are selected when the BIC is used. With the tuning parameter selected by 5-fold cross-validation, an additional marker HLA-A11 is selected. The sign of the coefficients are consistent in the various final models.

We next apply the inverse weighting approach to estimate the prediction errors and the discriminative abilities for all the models reported in Table 3. Figure 4 shows both the unweighted and weighted (IPW) estimates of the prediction error against time t_0, ranging from 0 to 40 years after the diagnosis of psoriatic arthritis. It is obvious that the unweighted estimates are greater than the weighted estimates since the unweighted estimators do not account for the unclassified portion in the sample. The ROC curves at time $t_0 = 10$ and 20 years after diagnosis of psoriatic arthritis for all the models with inverse weighting methods are shown in Figure 5; the summary statistic AUC is also given in the legend. From the upper two panels, we can conclude that when using BIC as the method of selecting tuning parameter, the predictive performance of the SCAD penalty is better than the ALASSO, which makes sense because the method of using ALASSO penalty with BIC does not select any HLA markers. The bottom two panels show that when using 5-fold CV to select the tuning parameter, the predictive performance of the ALASSO and SCAD penalties are quite close.

4 Discussion

This article has focussed on methods for synthesizing data from different disease registries with a view to fitting penalized regression models for the selection of important genetic predictors of psoriatic arthritis among patients with psoriasis. A key contribution is the formulation of an expectation-maximization algorithm which enables use of existing optimization software for penalized regression with truncated and interval-censored data; this extends the work of Wu and Cook (2015) to deal with truncation. It is important to quantify the accuracy of any predictive model and this is challenging

Table 2: Empirical performance of PE; sample size $m = 500$, number of simulations $nsim = 100$. The predictor is $\widehat{Y}(X;\widehat{\theta}) = I(P(T > t_0|X;\widehat{\theta}) > 0.5)$.

		$Z_{i2} \perp Z_{i3}$				$Z_{i2} \not\perp Z_{i3}$			
		Unweighted		Weighted		Unweighted		Weighted	
t_0	TRUE	EBIAS	ESE	EBIAS	ESE	EBIAS	ESE	EBIAS	ESE
		Normal							
		Correct* Model Specification							
Q_{25}	0.2269	-0.0973	0.0176	-0.0043	0.0311	-0.0940	0.0193	-0.0023	0.0310
Q_{50}	0.3063	-0.0458	0.0280	-0.0022	0.0317	-0.0464	0.0222	0.0014	0.0273
Q_{75}	0.2129	-0.0223	0.0182	0.0042	0.0203	-0.0282	0.0178	0.0008	0.0209
		Misspecified† Inspection Model							
Q_{25}	0.2269	-0.0973	0.0176	-0.0122	0.0282	-0.0940	0.0193	-0.0046	0.0297
Q_{50}	0.3063	-0.0458	0.0280	-0.0145	0.0306	-0.0464	0.0222	-0.0116	0.0254
Q_{75}	0.2129	-0.0223	0.0182	-0.0035	0.0195	-0.0282	0.0178	-0.0095	0.0192
		Binary							
		Correct* Model Specification							
Q_{25}	0.2500	-0.0820	0.0176	-0.0026	0.0285	-0.0777	0.0189	0.0009	0.0279
Q_{50}	0.3836	-0.0327	0.0241	0.0015	0.0293	-0.0342	0.0249	-0.0026	0.0296
Q_{75}	0.2500	-0.0309	0.0206	0.0021	0.0220	-0.0338	0.0215	0.0018	0.0236
		Misspecified† Inspection Model							
Q_{25}	0.2500	-0.0820	0.0176	-0.0140	0.0237	-0.0777	0.0189	-0.0060	0.0253
Q_{50}	0.3836	-0.0327	0.0241	-0.0156	0.0271	-0.0342	0.0249	-0.0176	0.0273
Q_{75}	0.2500	-0.0309	0.0206	-0.0110	0.0204	-0.0338	0.0215	-0.0141	0.0218

* correct inspection model involves fitting $\lambda(s|Z_{i1}, Z_{i3}; \gamma) = \exp(\gamma_0 + Z_{i1}\gamma_1 + Z_{i3}\gamma_2)$,

† misspecified inspection model involves fitting $\lambda(s|Z_{i1}; \gamma) = \exp(\gamma_0 + Z_{i1}\gamma_1)$.

Table 3: HLA markers selected following variable selection with LASSO, ALASSO or SCAD penalty and AIC, BIC or cross-validation in analysis of interval-censored data in psoriatic arthritis cohort.

	ALASSO		SCAD	
HLA Marker	BIC	CV	BIC	CV
HLA-A11				-0.8095
HLA-A25		-3.2399		
HLA-A29		-1.3471	-1.6215	-1.7063
HLA-A30		0.5912		
HLA-B27		0.4536	0.6578	0.6624
HLA-C04		-0.3492		
HLA-DQB1-02		0.3467	0.5270	0.4928
HLA-DRB1-10		-2.5831		

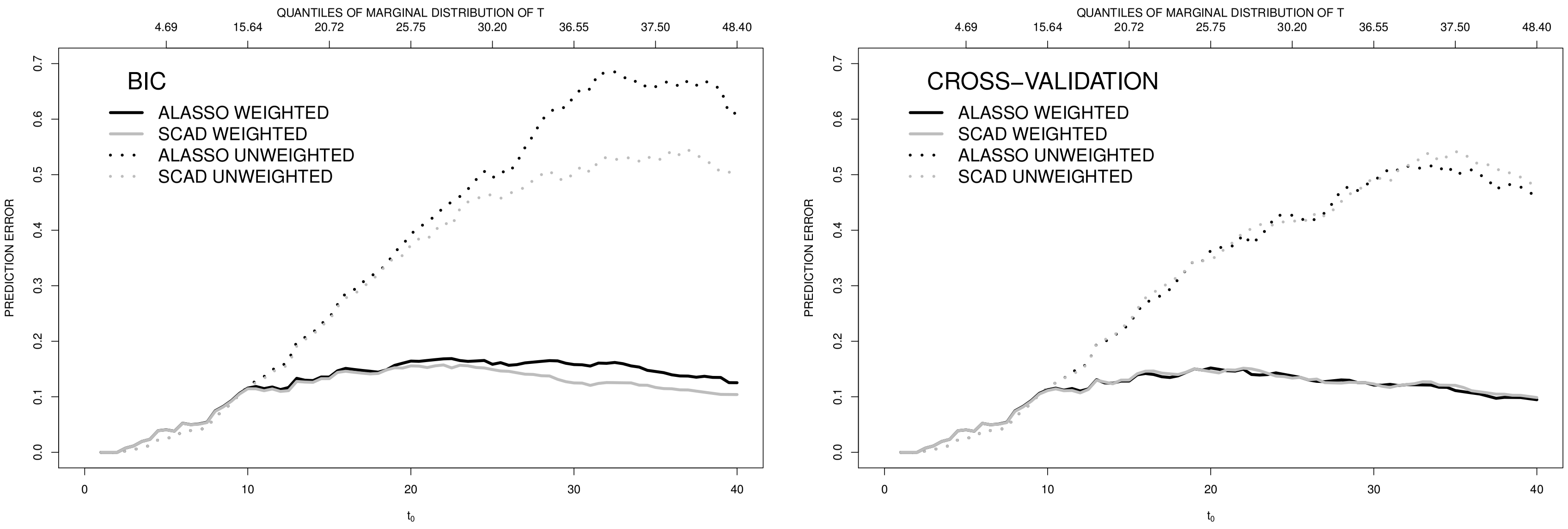

Figure 4: Plots of the estimates of the prediction error against t_0 with a binary predictor $I(P(T > t_0) > 0.5)$, the left panel corresponds to the estimates obtained by selecting tuning parameter with BIC and the right panel corresponds to the estimates obtained by using 5-fold cross-validation as the method of tuning paramter selection.

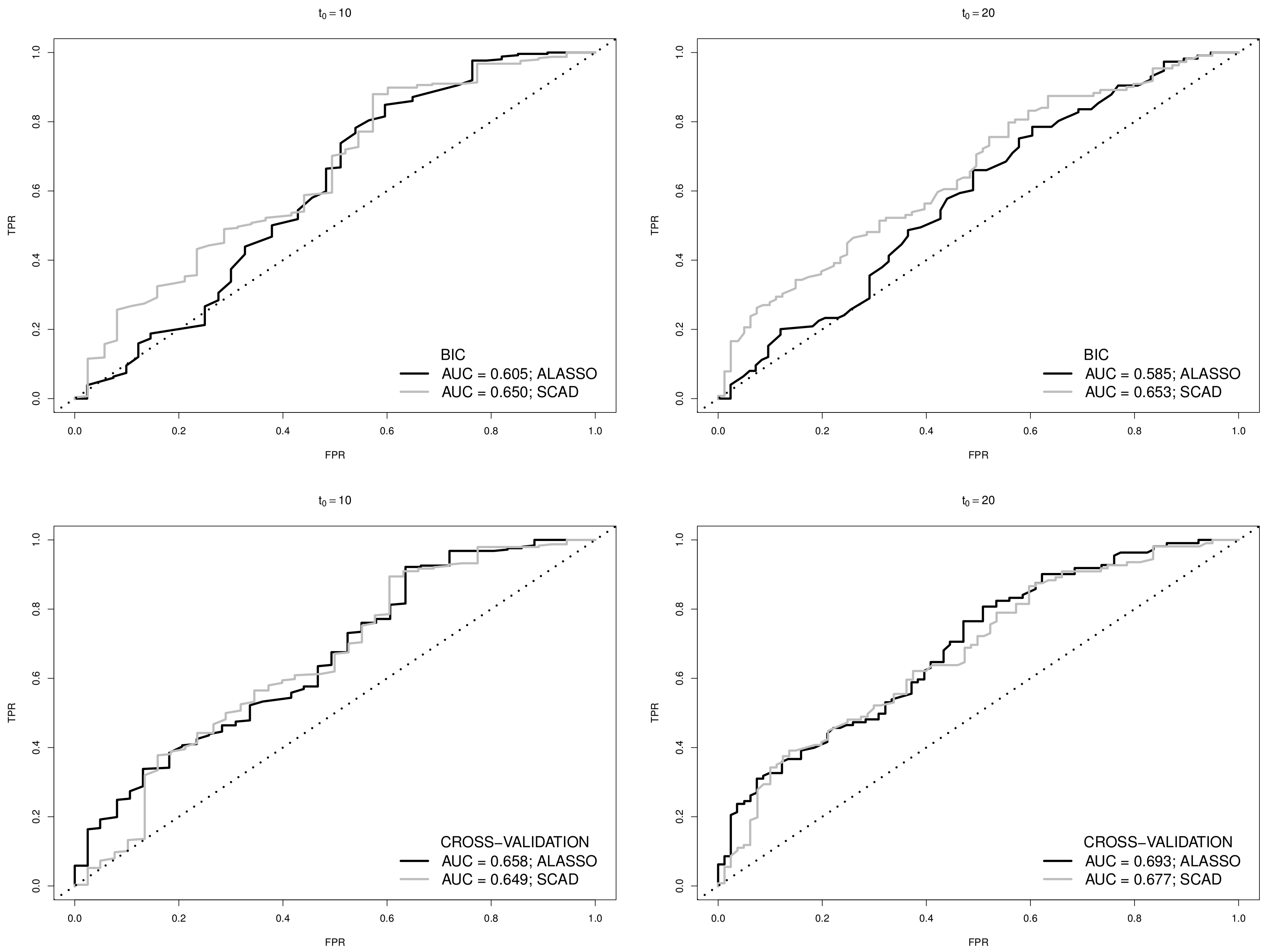

Figure 5: ROC curves at t_0 =10 and 20 years for the models obtained from penalized regression with the ALASSO and SCAD penalties with the tuning parameter selected by the BIC and 5-fold CV.

when data are interval-censored. We extend the methods of inverse probability weighting used for right-censored data to deal with interval-censored data arising from intermittent inspection of individuals; these methods require all assessment times be available so that the observation process can be modeled.

Throughout this manuscript the assumption of conditionally independent truncation has been made. As an illustration of a setting with dependent left truncation, Chaieb et al. (2006) discuss estimation of the survival time distribution based on a sample of individuals from homes for the elderly which offer on-site health care. They point out that individuals admitted earlier to such an institution may benefit from enhanced quality of care, and so live longer that they might otherwise, resulting in a negative association between the left-truncation time and the survival time. The analogous situation in the psoriasis clinic would be if treatment for psoriasis in a tertiary care setting reduced the rate of progression to psoriatic arthritis given covariates; there is no scientific evidence that this is the case, but exploration of the impact of dependent truncation along with methods for correcting for its effect is an area worthy of development in the context of variable selection. Tests of the null hypothesis of independent truncation are available (Tsai, 1990, Martin and Betensky, 2005) and may be possible to adapt, as are methods for dealing with dependent left truncation using copula models linking the failure and truncation times (Chaieb et al., 2006).

It is well-known that truncation schemes greatly reduce parametric information and can introduce identifiability concerns. Samples which feature right truncation seem to be particularly problematic (Kalbfleisch and Lawless, 1991) and it is somewhat unclear what the effective sample size is in some settings; this is more challenging with inspection processes leading to highly variable censoring intervals in the left-truncation setting. Extensive simulation studies (not reported here) suggest that the algorithms described here do not perform well when all observations are subject to right truncation; here the combination of the psoriasis cohort with prospective observation yielding data on the incidence of psoriatic arthritis plays a crucial role in making these analyses possible. As the fraction of right-truncated times increases in the sample, the more frequent the occurence of convergence problems.

A quite different framework for identifying genetic risk factors involves carrying out repeated univariate tests of the association between each genetic marker and the outcome of interest. The goal of this approach is to screen a number of potential markers for ones with an association with the outcome, and selected markers are identified as requiring further study and validation. When data from different cohorts are to be synthesized, or data are available according to another response-dependent observation scheme (Derkach et al., 2015), the selection effects and censoring must again be accounted for and this can be achieved through specification of similar complete data likelihoods and a corresponding self-consistency algorithm. However, score tests have great appeal in this setting as the model need only be fitted under the null hypothesis and so considerable simplifications may be realized.

Acknowledgements

This research was support by Natural Sciences and Engineering Research Council of Canada (RGPIN 155849); Canadian Institutes for Health Research (FRN 13887); Canada Research Chair (Tier 1) - CIHR funded (950-226626)

References

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics*, 22(3):1299–1327.

Chaieb, L. L., Rivest, L. P., and Abdous, B. (2006). Estimating survival under a dependent truncation. *Biometrika*, 93(3):655–669.

Chandran, V., Cook, R. J., Edwin, J., Shen, H., Pellett, F. J., Shanmugarajah, S., Rosen, C. F., and Gladman, D. D. (2010). Soluble biomarkers differentiate patients with psoriatic arthritis from those with psoriasis without arthritis. *Rheumatology*, 49(7):1399–1405.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Derkach, A., Lawless, J. F., and Sun, L. (2015). Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika*, 102(4):988–994.

Eder, L., Chandran, V., and Gladman, D. D. (2015). What have we learned about genetic susceptibility in psoriasis and psoriatic arthritis? *Current opinion in rheumatology*, 27(1):91–98.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Gladman, D. D., Schentag, C. T., Tom, B. D. M., Chandran, V., Brockbank, J., Rosen, C., and Farewell, V. T. (2008). Development and initial validation of a screening questionnaire for psoriatic arthritis: the toronto psoriatic arthritis screen (ToPAS). *Annals of the Rheumatic Diseases*, 68(4):497–501.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Grüger, J., Kay, R., and Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. *Biometrics*, 47:595–605.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, 7(3):355–373.

Julian, T., Kristine, U., Shbeeb, M. I., W Michael, O., Crowson, C. S., Gibson, L. E., Michet, C. J., and Gabriel, S. E. (2002). Risk factors for the development of psoriatic arthritis: a population based nested case control study. *The Journal of Rheumatology*, 29(4):757–762.

Kalbfleisch, J. D. and Lawless, J. F. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica*, 1(1):19–32.

Keiding, N. and Moeschberger, M. (1992). Independent delayed entry. In Klein, J. P. and Goel, P. K., editors, *Survival Analysis: State of the Art*, pages 309–326. Springer Netherlands, Dordrecht.

Korn, E. L. and Simon, R. (1990). Measures of explained variation for survival data. *Statistics in Medicine*, 9(5):487–503.

Lawless, J. F. and Yuan, Y. (2010). Estimation of prediction error for survival models. *Statistics in Medicine*, 29(2):262–274.

Martin, E. C. and Betensky, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional kendall's tau. *Journal of the American Statistical Association*, 100(470):484–492.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tom, B. D., Chandran, V., Farewell, V. T., Rosen, C. F., and Gladman, D. D. (2015). Validation of the toronto psoriatic arthritis screen version 2 (ToPAS 2). *The Journal of Rheumatology*, 42(5):841–846.

Tsai, W. Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika*, 77(1):169–177.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

Wu, Y. and Cook, R. J. (2015). Penalized regression for interval-censored times of disease progression: Selection of HLA markers in psoriatic arthritis. *Biometrics*, 71(3):782–791.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533.