Variable Selection and the Assessment of Predictive Accuracy with Interval-Censored Responses

Richard Cook

Statistics and Actuarial Science

University of Waterloo

Statistical Issues in Biomarker and Drug Co-Development

Toronto, Ontario

November 8, 2014

Joint work with Ying Wu and Ker-Ai Lee

PART I

VARIABLE SELECTION WITH INTERVAL-CENSORED RESPONSES

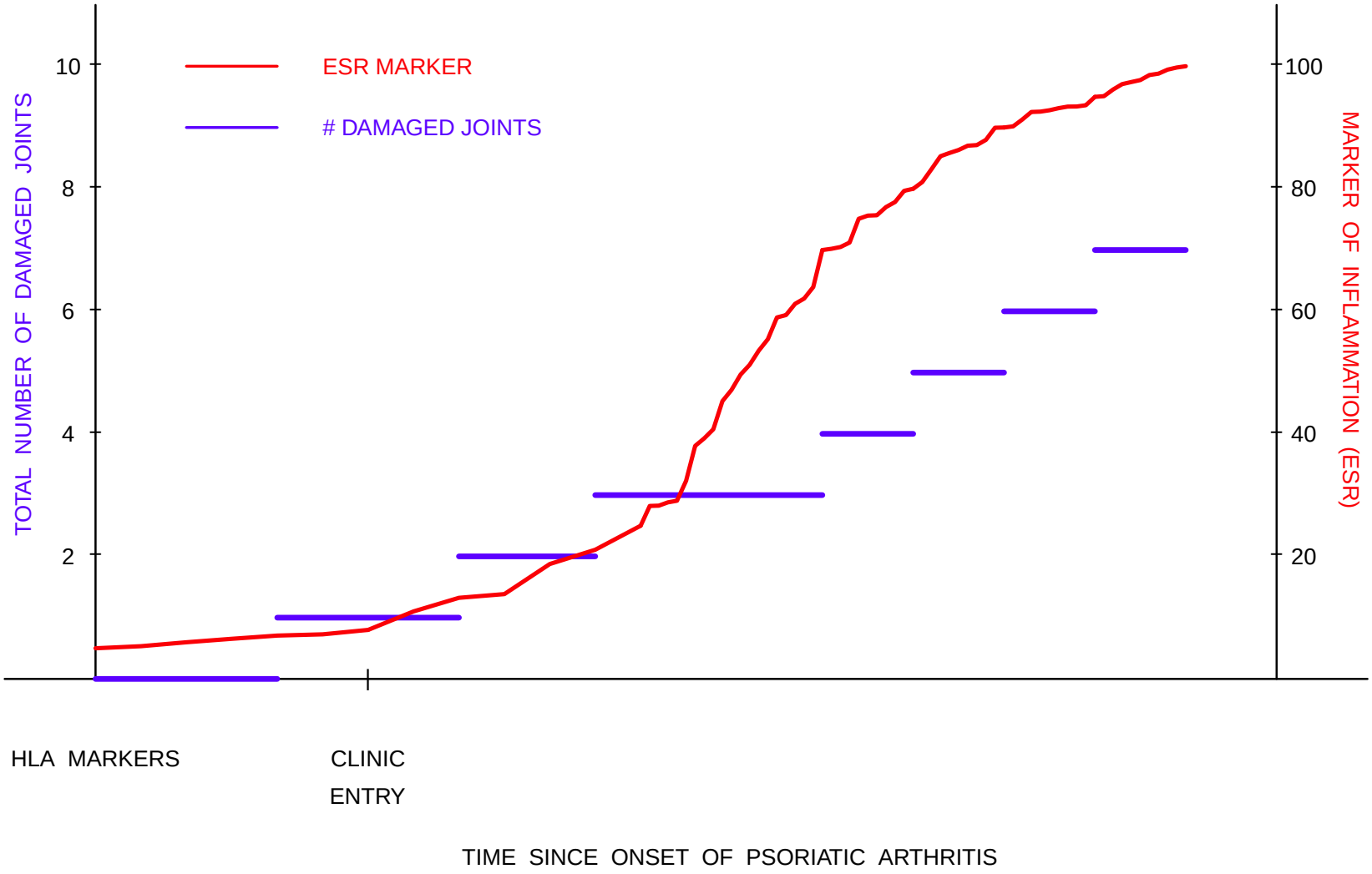
Prognostic Human Leukocyte Antigens in Psoriatic Arthritis

- The University of Toronto Psoriatic Arthritis Clinic is a tertiary referral clinic comprised of 1300 patients with extensive longitudinal follow-up on disease progression and collection of genetic and serum samples.
- Patients with psoriatic arthritis are classified as suffering from *arthritis mutilans* if they have *5 or more damaged joints*
- Patients are scheduled to be *radiologically assessed every two years*.
- The time for the development of arthritis mutilans is unknown because it is subject to interval-censoring.

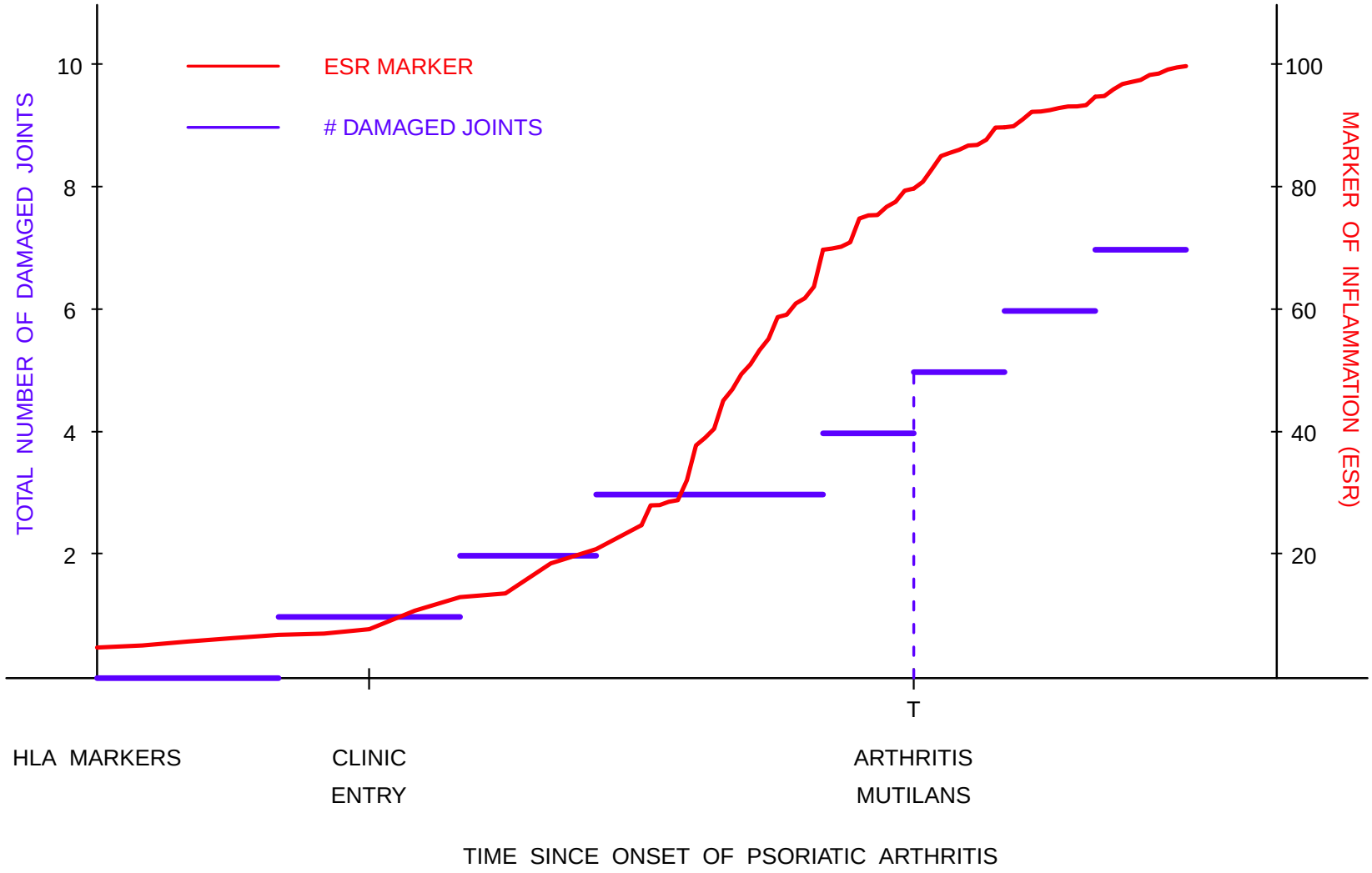
Immediate Goal

Interest lies in identifying HLA markers that predict onset of arthritis mutilans.

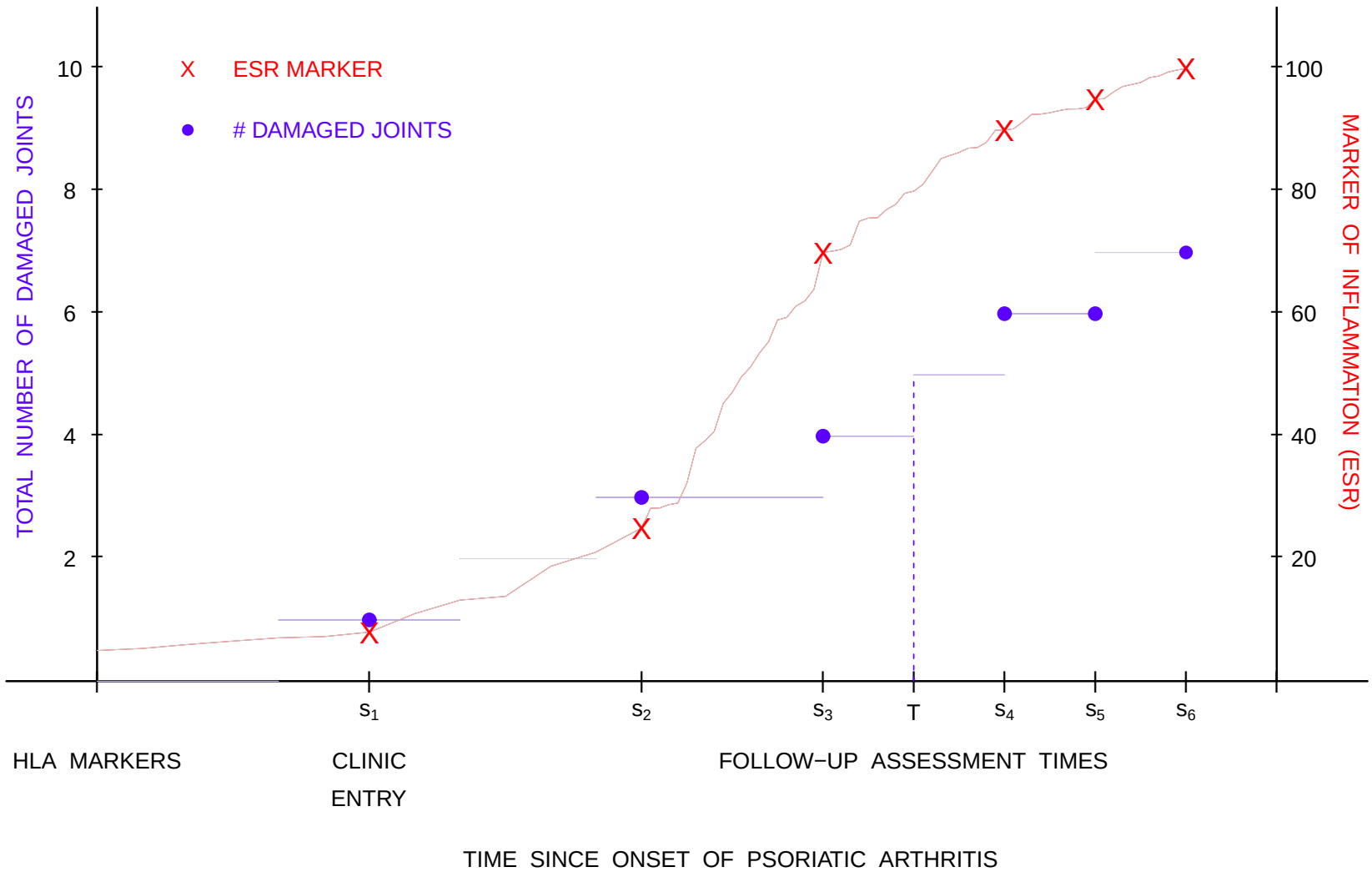
Joint Damage and Marker Values in Continuous Time



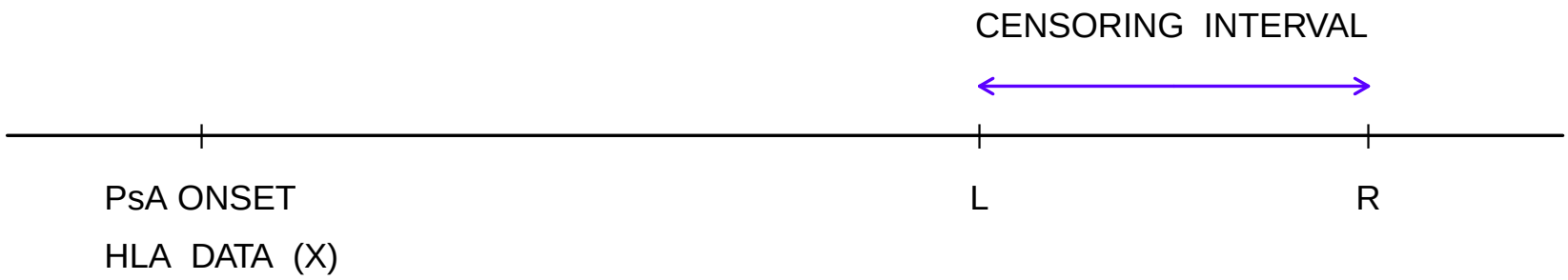
Joint Damage and Marker Values in Continuous Time



Available Data Due to Intermittent Assessments



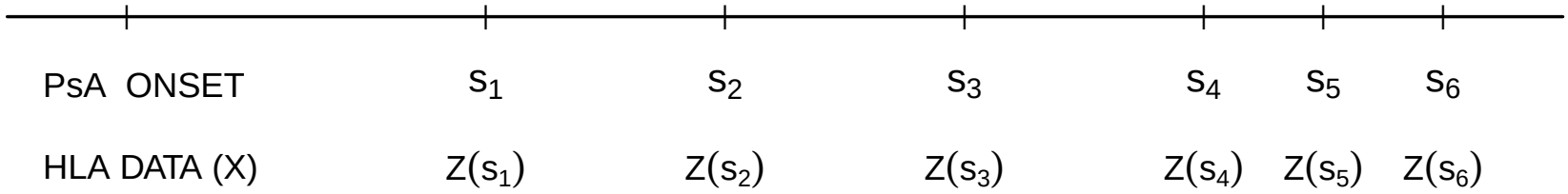
DATA FOR RESPONSE MODEL



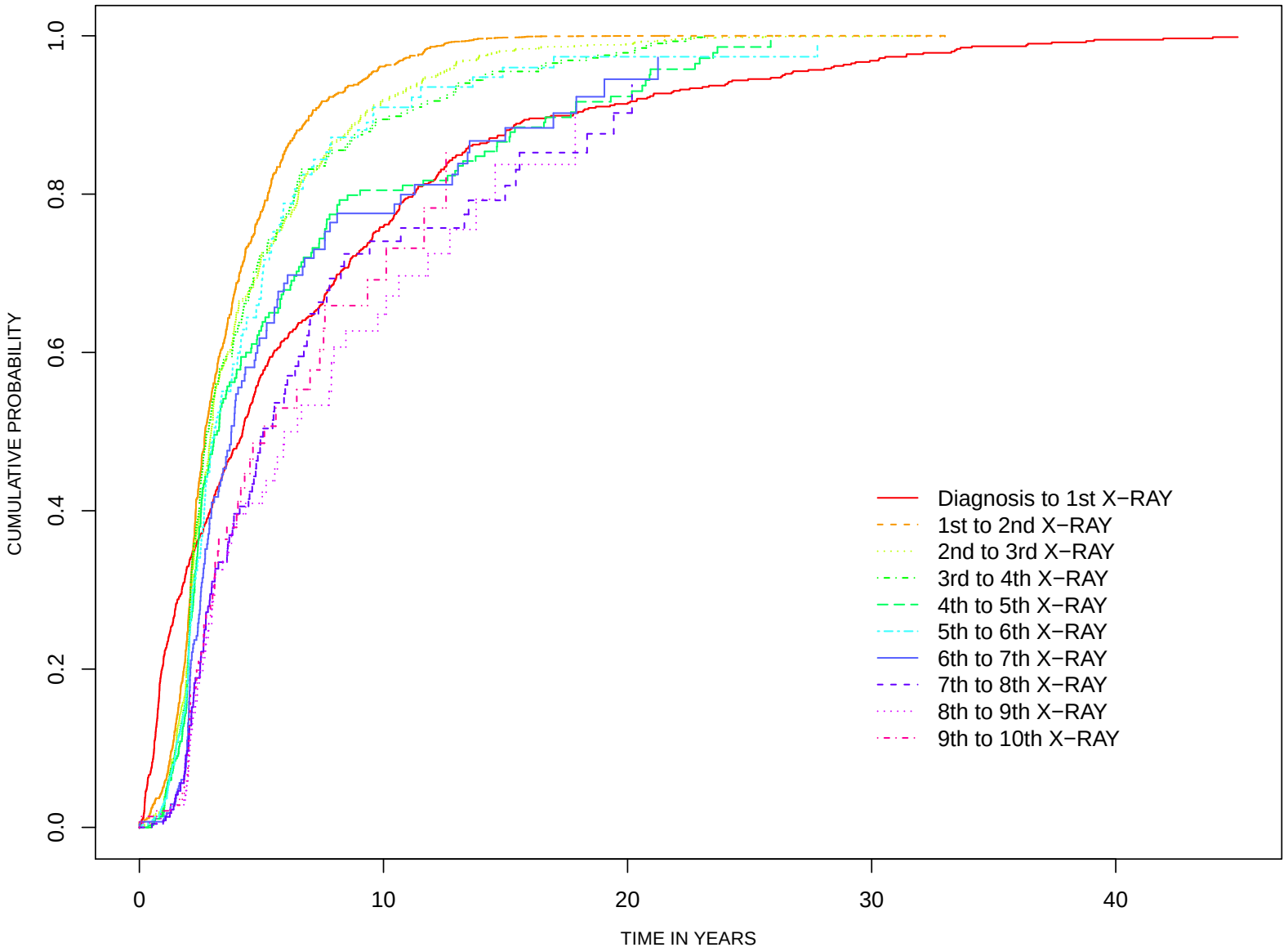
DATA FOR ASSESSMENT PROCESS

$Z(s_j)$ denotes marker of inflammation

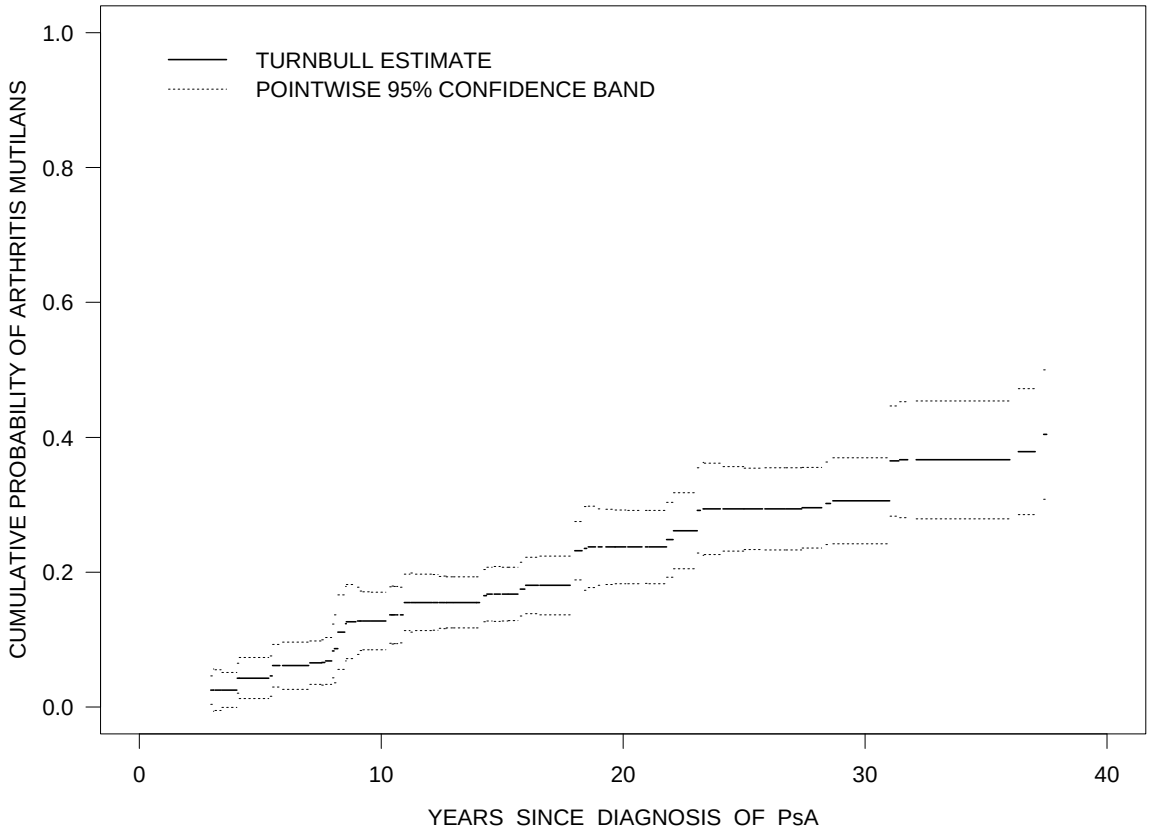
$w_j = s_j - s_{j-1}$, $j = 1, 2, \ldots$ are waiting times



Semi-Parametric Estimates of Waiting Time Distributions



ESTIMATE [1] OF DISTRIBUTION OF TIME TO ARTHRITIS MUTILANS



[1]Turnbull BW (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, Journal of the Royal Statistical Society. Series B (Methodological) 38, 290-295.

PENALIZED REGRESSION FOR FAILURE TIME DATA

- $\log L(\beta)$ is the log likelihood or log partial likelihood
- Consider a penalized "likelihood" function

$$\log L_{\text{PEN}}(\beta) = \log L(\beta) - \sum_{j=1}^{p} \pi_{\gamma,\lambda}(\beta_j) \tag{1.1}$$

- $\pi_{\gamma,\lambda}(\cdot)$ is a *penalty function*
- (γ, λ) are *tuning parameters*
- $\lambda = (\lambda_1, \ldots, \lambda_p)'$ if we use different penalties for each variable

Some Particular Penalty Functions

The L_2 penalty $\pi_\lambda(|\beta|) = \lambda|\beta|^2$ gives *ridge regression* [2]

The L_1 penalty $\pi_\lambda(|\beta|) = \lambda|\beta|$ yields the *LASSO* [3]

Smoothly Clipped Absolute Deviation (SCAD) Penalty

The *smoothly clipped absolute deviation* (SCAD) [4] penalty has the form

Adaptive LASSO

The adaptive LASSO [5] with penalty has the form

$$\pi_\lambda(|\beta_j|) = \lambda|\beta_j|\tau_j \ ,$$

with small weights τ_j chosen for large coefficients and large weights for small

[2] Hoerl AE and Kennard RW (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12 (1), 55–67.

[3] Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.

[4] 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.

[5] Zou H (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101 (476), 1418–1429.

Penalized Regression with Interval-Censored Data

- For individual i, $D_i = \{(L_i, R_i), X_i\}$, where X_i is a $p \times 1$ covariate vector
- Data consists of $D = \{D_i, i = 1, 2, \ldots, m\}$

Observed Data Log-Likeliood

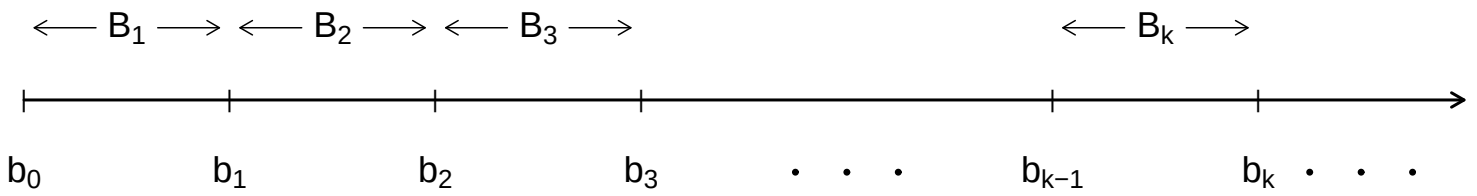
$$\log L \propto \sum_{i=1}^{m} \log \left[\mathcal{F}(L_i|X_i) - \mathcal{F}(R_i|X_i)\right]$$

where $\mathcal{F}(s|X)$ is the survivor function

Penalized Observed Data Log-Likeliood

$$\log L_{\text{penalized}} \propto \sum_{i=1}^{m} \log \left[\mathcal{F}(L_i|X_i) - \mathcal{F}(R_i|X_i)\right] - \sum_{j=1}^{p} \pi_{\gamma,\lambda}(\beta_j)$$

Penalized Regression with Interval Censored Data



Breakpoints $0 = b_0 < \cdots < b_K = \infty$ define $\mathcal{B}_k = [b_{k-1}, b_k), k = 1, \ldots, K.$

If $I_k(u) = I(u \in \mathcal{B}_k)$ and $S_k(u) = \int_0^u I(v \in \mathcal{B}_k)dv$ then

$$h(s;\theta) = \prod_{k=1}^{K} \left(\rho_k \ \exp\left(x_i'\beta\right)\right)^{I_k(u)}$$

where $\theta = (\rho', \beta')'$, $\rho = (\rho_1, \ldots, \rho_K)'$ and $\beta = (\beta_1, \ldots, \beta_p)'$

Complete Data Likelihood

$$\log L_c(\theta) = \sum_{i=1}^{m}\sum_{k=1}^{K} \left\{ I_k(u_i) \left[\log(\rho_k) + X_i'\beta\right] - S_k(u_i)\rho_k \exp(X_i'\beta)\right\}$$

AN EM ALGORITHM [6] WITH PENALIZED REGRESSION

THE EXPECTATION STEP

Take the conditional expectation of penalized complete data log-likelihood

$$Q(\theta;\theta^{r-1}) = E\left[\log L_c(\theta)|D;\theta^{r-1}\right] - \sum_{j=1}^{p}\pi_{\alpha,\lambda}(\beta_j)$$

If

$$\begin{aligned}\hat{g}_{ik}^r &= E\left[I_k(u_i)|D_i;\theta^{r-1}\right]\\ \hat{S}_{ik}^r &= E\left[S_k(u_i)|D_i;\theta^{r-1}\right]\end{aligned}$$

then

$$Q(\theta;\theta^{r-1}) = \sum_{i=1}^{m}\sum_{k=1}^{K}\left[\hat{g}_{ik}^r(\log(\rho_k)+X_i'\beta) - \hat{S}_{ik}^r\rho_k\exp(X_i'\beta)\right] - \sum_{j=1}^{p}\pi_{\gamma,\lambda}(\beta_j)$$

[6] Dempster AP, Laird NM and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1), 1–38.

Maximization Step

Let

- $Z_{ij} = I(j = k), j = 2, \ldots, K$, $Z_{ik} = (1, Z_{i2}, \ldots, Z_{iK})'$
- $\alpha_1 = \log(\rho_1)$, $\alpha_j = \log(\rho_j) - \log(\rho_1)$, $j = 2, \ldots, K$

Then $Q(\theta; \theta^{r-1})$ is

$$\sum_{i=1}^{m} \sum_{k=1}^{K} \left[\hat{g}_{ik}^{r} (Z_{ik}' \alpha + X_i' \beta) - \hat{S}_{ik}^{r} \exp(Z_{ik}' \alpha + X_i' \beta) \right] - \sum_{j=1}^{p} \pi_{\gamma,\lambda}(\beta_j)$$

With a *pseudo dataset* we can maximize $Q(\theta; \theta^{r-1})$ *using standard software for penalized regression* (e.g. glmnet(.), SIS(.))

SELECTION OF OPTIMAL PENALTY λ_{OPT}

- The criterion for selecting the optimal λ is similar to the traditional cross-validation.
- We partition the dataset into R subsamples $T^1, \ldots, T^R$.
- T^r and $T - T^r$ are rth testing and training sets.
- For a given λ, the *cross-validation statistic* is

$$\widehat{CV}(\lambda) = \sum_{r=1}^{R} \log L(\theta_{-r}(\lambda)) - \log L_{-r}(\theta_{-r}(\lambda)).$$

- L_{-r} is the *observed likelihood* for the rth training dataset.
- $\theta_{-r}(\lambda)$ is the estimate for the rth training data.
- The optimal λ maximizes $\widehat{CV}(\lambda)$.

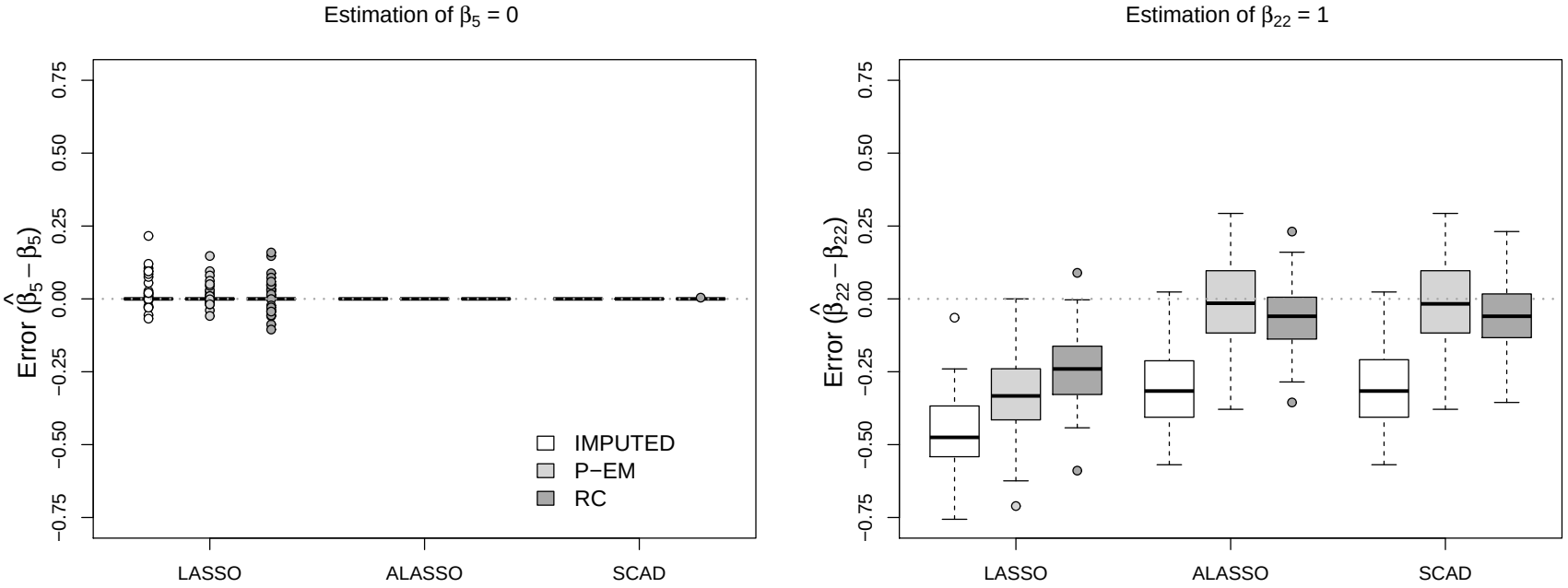
EMPIRICAL STUDIES – NORMAL COVARIATES $m = 1000, p = 100$

Method		$\mu = 10$ TP(10)	FP(90)	MSE (SD)	$\mu = 20$ TP(10)	FP(90)	MSE (SD)
				Shape parameter: $\kappa = 1$			
LASSO	P-EM	10.00	14.80	0.312 (0.126)	10.00	14.83	0.261 (0.105)
	MID	10.00	13.05	1.346 (0.286)	10.00	12.05	0.912 (0.251)
ALASSO	P-EM	10.00	0.12	0.057 (0.047)	10.00	0.07	0.047 (0.040)
	MID	9.69	0.30	0.953 (0.328)	10.00	1.57	0.499 (0.201)
SCAD	P-EM	9.98	0.36	0.059 (0.073)	9.99	0.24	0.050 (0.048)
	MID	9.39	0.96	0.946 (0.354)	9.91	1.01	0.521 (0.213)
FORWARD		10.00	9.17	0.218 (0.088)	10.00	9.50	0.201 (0.082)
BACKWARD		10.00	15.35	0.322 (0.130)	10.00	14.80	0.289 (0.099)
				Shape parameter: $\kappa = 1.25$			
LASSO	P-EM	10.00	14.88	0.291 (0.118)	10.00	14.13	0.245 (0.109)
	MID	10.00	15.28	1.037 (0.271)	10.00	12.94	0.685 (0.216)
ALASSO	P-EM	9.99	0.23	0.055 (0.050)	10.00	0.08	0.045 (0.031)
	MID	9.75	0.29	0.724 (0.327)	10.00	1.25	0.314 (0.160)
SCAD	P-EM	9.98	0.29	0.055 (0.052)	9.99	0.13	0.044 (0.036)
	MID	9.53	0.76	0.741 (0.336)	9.97	0.91	0.317 (0.167)
FORWARD		10.00	8.66	0.324 (0.089)	10.00	8.81	0.313 (0.089)
BACKWARD		10.00	14.35	0.383 (0.092)	10.00	14.17	0.363 (0.092)

EMPIRICAL STUDIES – BINARY COVARIATES, $m = 1000, p = 100$

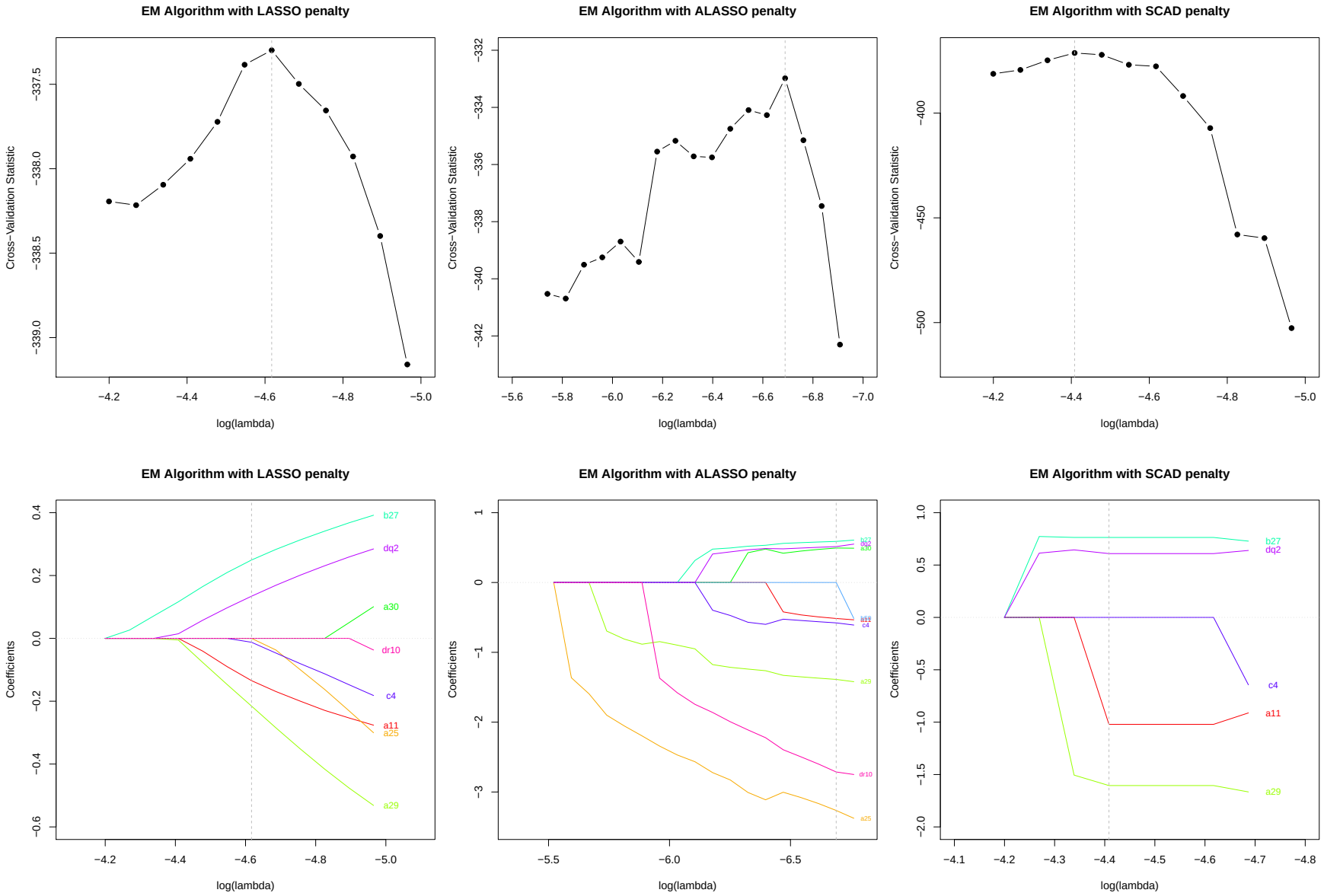
		$\mu = 10$			$\mu = 20$		
Method		TP(10)	FP(90)	MSE (SD)	TP(10)	FP(90)	MSE (SD)
				Shape parameter: $\kappa = 1$			
LASSO	P-EM	10.00	12.49	0.304 (0.068)	10.00	15.30	0.201 (0.052)
	MID	10.00	17.64	0.690 (0.117)	10.00	19.01	0.436 (0.086)
ALASSO	P-EM	9.88	0.82	0.071 (0.067)	9.98	0.26	0.039 (0.033)
	MID	9.18	0.78	0.491 (0.149)	9.83	0.49	0.255 (0.097)
SCAD	P-EM	9.94	0.54	0.063 (0.063)	10.00	0.10	0.038 (0.031)
	MID	9.02	0.96	0.505 (0.166)	9.79	0.40	0.254 (0.102)
FORWARD		10.00	11.14	0.244 (0.078)	10.00	11.09	0.183 (0.057)
BACKWARD		10.00	15.18	0.299 (0.083)	10.00	14.64	0.231 (0.064)
				Shape parameter: $\kappa = 1.25$			
LASSO	P-EM	10.00	12.04	0.277 (0.064)	10.00	15.65	0.186 (0.053)
	MID	9.99	18.15	0.609 (0.100)	10.00	17.91	0.374 (0.074)
ALASSO	P-EM	9.98	0.59	0.051 (0.042)	10.00	0.22	0.034 (0.023)
	MID	9.59	0.60	0.404 (0.116)	9.97	0.26	0.186 (0.064)
SCAD	P-EM	10.00	0.48	0.053 (0.038)	10.00	0.16	0.033 (0.021)
	MID	9.54	0.93	0.414 (0.118)	9.95	0.42	0.186 (0.064)
FORWARD		10.00	10.86	0.198 (0.060)	10.00	10.81	0.180 (0.045)
BACKWARD		10.00	14.49	0.233 (0.064)	10.00	13.76	0.195 (0.052)

Box plots of the error for the estimated regression coefficients $\widehat{\beta}_k - \beta_k$, $k = 5, 22, 95, 96$, for each penalty function for datasets with correlated binary covariates ($p = 100$) with $\kappa = 1.25$, $\mu = 20$.



Application to University of Toronto PsA Cohort

	LASSO				ALASSO				SCAD			
	P-EM		MID		P-EM		MID		P-EM		MID	
HLA Marker	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)
HLA-A11	-0.135	0.199	-0.280	0.263	-0.516	0.629	-0.556	0.836	-1.021	0.746	-0.922	0.947
HLA-A25			-0.232	0.288	-3.265	0.707	-3.229	1.529				
HLA-A29	-0.216	0.254	-0.502	0.353	-1.388	1.284	-1.385	1.440	-1.605	2.376	-1.658	2.482
HLA-A30			0.101	0.260	0.494	0.417	0.494	0.525				
HLA-B27	0.249	0.232	0.397	0.272	0.588	0.356	0.595	0.547	0.763	0.312	0.725	0.425
HLA-C04	-0.012	0.134	-0.170	0.233	-0.578	0.492	-0.569	1.086			-0.637	0.611
HLA-DQB1-02	0.134	0.164	0.270	0.205	0.514	0.307	0.503	0.540	0.609	0.276	0.623	0.415
HLA-DRB1-10					-2.713	1.007	-2.714	1.725				


EM Algorithm with LASSO penalty
Cross−Validation Statistic
log(lambda)
EM Algorithm with ALASSO penalty
Cross−Validation Statistic
log(lambda)
EM Algorithm with SCAD penalty
Cross−Validation Statistic
log(lambda)
EM Algorithm with LASSO penalty
Coefficients
log(lambda)
b27
dq2
a30
dr10
c4
a11
a25
a29
EM Algorithm with ALASSO penalty
Coefficients
log(lambda)
EM Algorithm with SCAD penalty
Coefficients
log(lambda)
b27
dq2
c4
a11
a29

Findings

Some old (HLA-B27, HLA-DQB1-02) and some new markers identified for future study.

Next Steps - Validation

There are three other cohorts in which we can validate this predictive model including registries in :

Ireland [7]

Spain [8]

Newfoundland [9]

Issues of *variation in the genetic composition of these cohorts* may affect accuracy of predictive model

[7] Winchester R, Minevich G, Steshenko V, Kirby B, Kane D, Greenberg DA, FitzGerald O. (2012). HLA associations reveal genetic heterogeneity in psoriatic arthritis and in the psoriasis phenotype. Arthritis Rheum. 64(4), 1134-44.

[8] Queiro R, Torre JC, González S, López-Larrea C, Tinturé T, López-Lagunas I (2003). HLA antigens may influence the age of onset of psoriasis and psoriatic arthritis. J Rheumatol. 30(3), 505-5077.

[9] Rahman P, Roslin NM, Pellett FJ, Lemire M, Greenwood CM, Beyene J, Pope A, Peddle L, Paterson AD, Uddin M, Gladman DD (2011). High resolution mapping in the major histocompatibility complex region identifies multiple independent novel loci for psoriatic arthritis. Ann Rheum Dis. 70(4), 690-694.

PART II

ESTIMATING ACCURACY OF PREDICTIVE MODELS WITH INTERVAL-CENSORED RESPONSE TIMES

Assessing Predictive Accuracy with Censored Data

There has been much work on measuring predictive performance with right-censored survival data [10] [11] [12] [13] [14] [15]

One can focus on survival time or survival status at t_0

Measures can be based on explained variation, misclassification rate, etc.

Censoring makes validation assessment challenging since some individuals will not be possible to classify with respect to the response in the validation sample

[10] Rosthoj S, Keiding N (2004). Explained variation and predictive accuracy in general parametric statistical models: the role of model misspecification. Lifetime Data Analysis 10, 461–472.

[11] Gerds TA, Schumacher M (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biometrical Journal 48, 1029–1040.

[12] Efron B (2004). The estimation of prediction error: covariance penalties and cross-validation. Journal of the American Statistician Association 99, 619–632.

[13] Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21, 3301–3307.

[14] Korn EL, Simon R (1990). Measures of explained variation for survival data. Statistics in Medicine 9, 487–503.

[15] Lawless JF, Yuan Y (2010). Estimation of prediction error for survival models. Statistics in Medicine, 16, 262-274.

ESTIMATING PREDICTIVE ACCURACY

We aim to predict $Y(t_0) = I(T \leq t_0)$, the event status at a time t_0

Let $\tilde{Y}(\theta) = I(F(t_0|X;\theta) > 0.5)$ be the prediction

Predictive accuracy can be measured by the mean squared error loss

$$PE = E\left\{\left(Y - \tilde{Y}(X;\theta)\right)^2\right\} \tag{2.1}$$

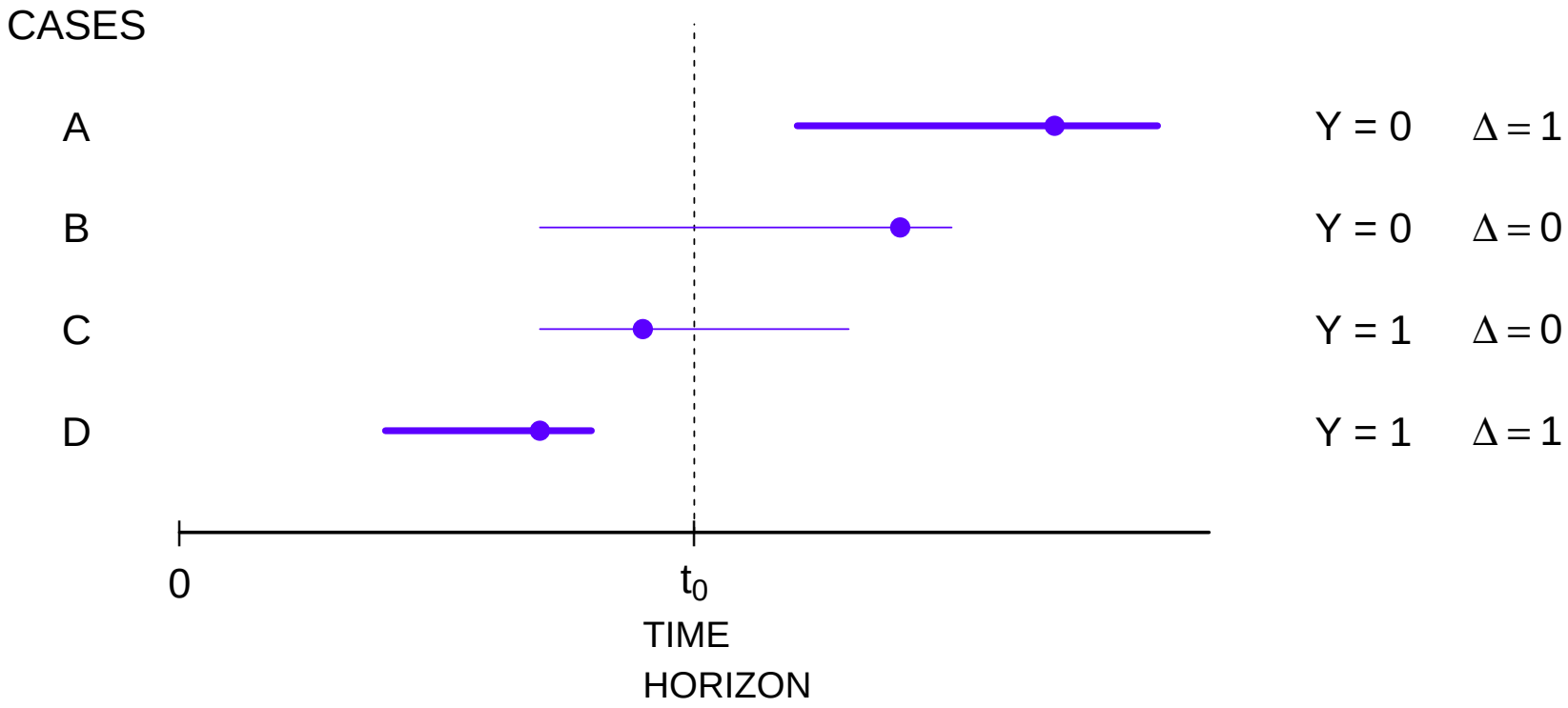
With a sample of size m this is normally estimated as

$$\frac{1}{m}\sum_{i=1}^{m}(Y_i - \tilde{Y}_i(X_i;\theta))^2 \ .$$

If $\Delta_i = I(Y_i \text{ is known})$,

$$\widehat{PE} = \frac{1}{m}\sum_{i=1}^{m} \Delta_i \cdot (Y_i - \tilde{Y}_i(X_i;\theta))^2 \ .$$

POSSIBLE COMBINATIONS OF (Y, Δ)



INVERSE PROBABILITY OR CENSORING WEIGHTS

Note

$$E\left\{\frac{1}{m}\sum_{i=1}^{m}\Delta_i\cdot(Y_i-\tilde{Y}_i(X_i;\theta))^2\right\} \tag{2.2}$$

$$= \frac{1}{m}\sum_{i=1}^{m}E_{Y_i,X_i}\left[E_{\Delta_i|Y_i,X_i}\left\{\Delta_i\cdot(Y_i-\tilde{Y}_i(X_i;\theta))^2|Y_i,X_i\right\}\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m}E_{Y_i,X_i}\left[P(\Delta_i=1|Y_i,X_i)\,(Y_i-\tilde{Y}_i(X_i;\theta))^2\right]$$

So we consider an inverse probability weighted version

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\Delta_i}{P(\Delta_i=1|?)}\left(Y_i-\tilde{Y}_i(X_i;\theta)\right)^2 . \tag{2.3}$$

The challenges is now to specify and estimate $P(\Delta=1|?)$.

INTRODUCING AND RECALLING SOME NOTATION

- $N(u) = \sum_{j=1}^{\infty} I(s_j \leq u)$ counts the number of assessments
- $C(u) = I(u \leq C)$ indicates in cohort
- X is the set of fixed HLA markers
- $\bar{Z}(u) = \{Z(s_0), Z(s_1), \ldots, Z(s_{N(u^-)})\}$ is time-dependent marker history
- $\mathcal{H}(s) = \{(dN(u), C(u)), 0 < u < s, X, \bar{Z}(s)\}$ is the partial history at s

Model Assumptions

Event Process

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

Intensity for Inspection Process

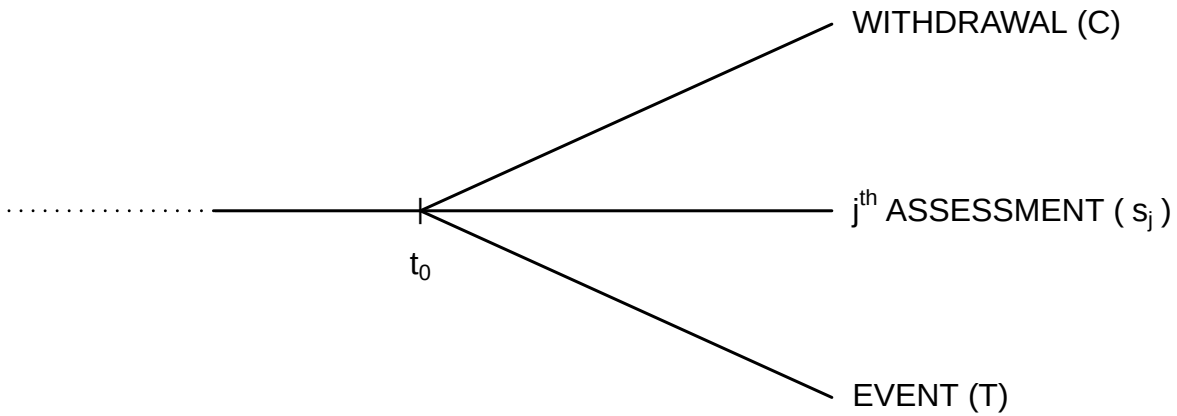
$$\lambda(t|\mathcal{H}(t)) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N(t) = 1|\mathcal{H}(t))}{\Delta t}$$

Intensity for Censoring (Withdrawal) Process

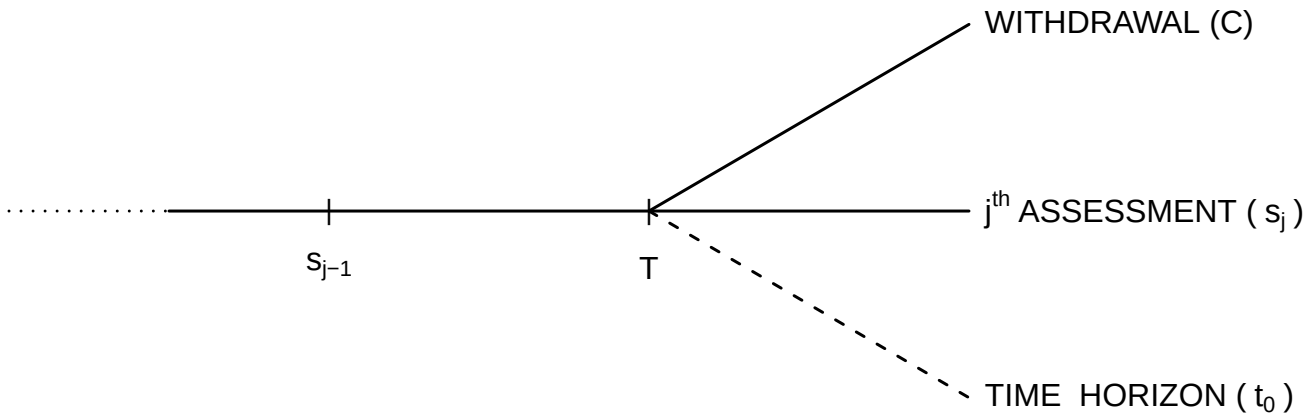
$$\lambda^c(t|\mathcal{H}(t)) = \lim_{\Delta t \downarrow 0} \frac{P(C < t + \Delta t|\mathcal{H}(t))}{\Delta t}$$

Modeling the Censoring and Inspection Process

Competing Risk for Event Occurrence



Competing Risk Following Event



With intermittent inspection (SMAR) [16] and random censoring

$$P(\Delta = 1 | Y(t_0) = 1, \mathcal{H}(t_0), X)$$

$$\int_0^{t_0} \left\{ \int_t^{t_0} \lambda(u|\mathcal{H}(u)) \, \exp\left(- \left[\int_t^u \lambda(v|\mathcal{H}(v)) \, dv + \int_0^u \lambda^c(v|\mathcal{H}(v)) \, dv \right] \right) \, du \right\} \, f(t|T < t_0, X) \, dt$$

$$P(\Delta = 1 | Y(t_0) = 0, \mathcal{H}(t_0), X)$$

$$\int_{t_0}^{\infty} \lambda(u|\mathcal{H}(u)) \, \exp\left(- \left[\int_{t_0}^u \lambda(v|\mathcal{H}(v)) + h(v|X) \, dv + \int_0^u \lambda^c(v|\mathcal{H}(v)) \, dv \right] \right) \, du$$

[16] Hogan JW, Roy J and Korkontzelou C (2004) Handling dropouts in longitudinal studies. Statistics in Medicine, 23, 1455–1497.

Summary of the empirical average of PE; Number of simulations = 100; Number of subjects per simulation = 1000

			Q_{25}			Q_{50}			Q_{75}		
α_0^{-1}	α_1	METHOD	TRUE	BIAS	ESE	TRUE	BIAS	ESE	TRUE	BIAS	ESE
0.10	0	Unweighted	0.2454	-0.0094	0.0140	0.3275	-0.0131	0.0141	0.2460	-0.0327	0.0134
		Weighted	0.2454	-0.0019	0.0144	0.3275	-0.0017	0.0149	0.2460	0.0004	0.0164
0.25	0	Unweighted	0.2454	-0.0173	0.0141	0.3275	-0.0291	0.0147	0.2460	-0.0752	0.0147
		Weighted	0.2454	-0.0020	0.0153	0.3275	-0.0025	0.0176	0.2460	0.0028	0.0212
0.10	$\log 1.1$	Unweighted	0.2454	-0.0093	0.0144	0.3275	-0.0126	0.0161	0.2460	-0.0289	0.0143
		Weighted	0.2454	-0.0016	0.0148	0.3275	-0.0002	0.0168	0.2460	0.0021	0.0166
0.25	$\log 1.1$	Unweighted	0.2454	-0.0144	0.0124	0.3275	-0.0283	0.0169	0.2460	-0.0737	0.0133
		Weighted	0.2454	0.0020	0.0140	0.3275	0.0004	0.0185	0.2460	-0.0002	0.0205

DISCUSSION

- In observational cohorts visit process may be non-ignorable (i.e. the observation process may not been SMAR) and use of inverse intensity weighting [17] could be important for model building.

- Important work to be done in assessing marker effects on progression in cancer trials as progression times are interval censored.

- Methodological work needed for assessing predictive accuracy of models in competing risk settings

- Multistate models are useful for this goal

[17] Lin H, Scharfstein DO, and Rosenheck RA (2004). Analysis of longitudinal data with irregular, outcomedependent followup. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66(3), 791-813.

Thanks and Acknowledgements

Dafna Gladman

Vinod Chandran

Lihi Eder

Shelley Bull

Yan Yuan and Jerry Lawless