Supplementary Material for "Personalized Screening Intervals for Biomarkers using Joint Models for Longitudinal and Survival Data"

Dimitris Rizopoulos[1,*], Jeremy M.G. Taylor[2], Joost van Rosmalen[1], Ewout W. Steyerberg[3] and Johanna J.M. Takkenberg[4]

[1]Department of Biostatistics, Erasmus University Medical Center
[2]Department of Biostatistics, University of Michigan - Ann Arbor, USA
[3]Department of Public Health, Erasmus University Medical Center
[4]Department of Cardiothoracic Surgery, Erasmus University Medical Center

1 Joint Modeling Framework

In this section we present a general definition of the framework of joint models for longitudinal and survival data that will be used later on for planning the optimal visit schedule. Let $\mathcal{D}_n = \{T_i, \delta_i, \boldsymbol{y}_i; i = 1, \ldots, n\}$ denote a sample from the target population, where T_i^* denotes the true event time for the i-th subject, C_i the censoring time, $T_i = \min(T_i^*, C_i)$ the corresponding observed event time, and $\delta_i = I(T_i^* \leq C_i)$ the event indicator, with $I(\cdot)$ being the indicator function that takes the value 1 when $T_i^* \leq C_i$, and 0 otherwise. In addition, we let $\boldsymbol{y}_i$ denote the $n_i \times 1$ longitudinal response vector for the i-th subject, with element y_{il} denoting the value of the longitudinal outcome taken at time point t_{il}, $l = 1, \ldots, n_i$.

To accommodate different types of longitudinal responses in a unified framework, we postulate a generalized linear mixed effects model. In particular, the conditional distribution

[*]Correspondance at: Department of Biostatistics, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, the Netherlands. E-mail address: d.rizopoulos@erasmusmc.nl.

of $\boldsymbol{y}_i$ given a vector of random effects $\boldsymbol{b}_i$ is assumed to be a member of the exponential family, with linear predictor given by

$$k\big[E\{y_i(t) \mid \boldsymbol{b}_i\}\big] = \eta_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i, \qquad (1)$$

where $k(\cdot)$ denotes a known one-to-one monotonic link function, and $y_i(t)$ denotes the value of the longitudinal outcome for the i-th subject at time point t, $\boldsymbol{x}_i(t)$ and $\boldsymbol{z}_i(t)$ denote the time-dependent design vectors for the fixed-effects $\boldsymbol{\beta}$ and for the random effects $\boldsymbol{b}_i$, respectively. The random effects are assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{D}$. For the survival process, we assume that the risk for an event depends on a function of the subject-specific linear predictor $\eta_i(t)$ and/or the random effects. More specifically, we have

$$\begin{aligned} h_i(t \mid \mathcal{H}_i(t), \boldsymbol{w}_i) &= \lim_{\Delta t \to 0} \Pr\{t \leq T_i^* < t + \Delta t \mid T_i^* \geq t, \mathcal{H}_i(t), \boldsymbol{w}_i\}\big/\Delta t \\ &= h_0(t) \exp\big[\boldsymbol{\gamma}^\top \boldsymbol{w}_i + f\{\mathcal{H}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\}\big], \quad t > 0, \qquad (2) \end{aligned}$$

where $\mathcal{H}_i(t) = \{\eta_i(s), 0 \leq s < t\}$ denotes the history of the underlying longitudinal process up to t, $h_0(\cdot)$ denotes the baseline hazard function, $\boldsymbol{w}_i$ is a vector of baseline covariates with corresponding regression coefficients $\boldsymbol{\gamma}$. Function $f(\cdot)$, parameterized by vector $\boldsymbol{\alpha}$, specifies which components/features of the longitudinal outcome process are included in the linear predictor of the relative risk model. Some examples, motivated by the literature (Brown 2009; Rizopoulos and Ghosh 2011; Rizopoulos 2012; Taylor et al. 2013; Rizopoulos et al. 2014), are:

$$\begin{aligned} f\{\mathcal{H}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\} &= \alpha \eta_i(t), \\ f\{\mathcal{H}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\} &= \alpha_1 \eta_i(t) + \alpha_2 \eta_i'(t), \text{ with } \eta_i'(t) = \frac{d\eta_i(t)}{dt}, \\ f\{\mathcal{H}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\} &= \alpha \int_0^t \eta_i(s)\, ds, \\ f\{\mathcal{H}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\} &= \boldsymbol{\alpha}^\top \boldsymbol{b}_i. \end{aligned}$$

These formulations of $f(\cdot)$ postulate that the hazard of an event at time t may be associated with the underlying level of the biomarker at the same time point, the slope of the longitudinal profile at t, the accumulated longitudinal process up to t, or the random effects alone. Finally, the baseline hazard function $h_0(\cdot)$ is modeled flexibly using a B-splines approach, i.e.,

$$\log h_0(t) = \gamma_{h_0,0} + \sum_{q=1}^{Q} \gamma_{h_0,q} B_q(t, \boldsymbol{v}), \tag{3}$$

where $B_q(t, \boldsymbol{v})$ denotes the q-th basis function of a B-spline with knots $v_1, \ldots, v_Q$ and $\boldsymbol{\gamma}_{h_0}$ the vector of spline coefficients. To avoid the task of choosing the appropriate number and position of the knots, we include a relatively high number of knots (e.g., 15 to 20) and appropriately penalize the B-spline regression coefficients $\boldsymbol{\gamma}_{h_0}$ for smoothness using the differences penalty (Eilers and Marx 1996).

For the estimation of joint model's parameters we use a Bayesian approach based on Markov chain Monte Carlo (MCMC) algorithms. The expression for the posterior distribution of the model parameters given the observed data is derived under the assumptions that given the random effects, both the longitudinal and event time process are independent, and the longitudinal responses of each subject are independent. Formally we have,

$$\begin{aligned} p(\boldsymbol{y}_i, T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) &= p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}), \quad (4)\\ p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) &= \prod_l p(y_{il} \mid \boldsymbol{b}_i, \boldsymbol{\theta}), \quad (5) \end{aligned}$$

where $\boldsymbol{\theta}$ denotes the full parameter vector, and $p(\cdot)$ denotes an appropriate probability density function. Under these assumptions the posterior distribution is given by:

$$p(\boldsymbol{\theta}, \boldsymbol{b}) \propto \prod_{i=1}^{n} \prod_{l=1}^{n_i} p(y_{il} \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(\boldsymbol{b}_i \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{6}$$

where

$$p(y_{il} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \exp\Bigl\{ \bigl[y_{il}\psi_{il}(\boldsymbol{b}_i) - c\{\psi_{il}(\boldsymbol{b}_i)\} \bigr] \big/ a(\varphi) - d(y_{il}, \varphi) \Bigr\},$$

with $\psi_{il}(\boldsymbol{b}_i)$ and φ denoting the natural and dispersion parameters in the exponential family, respectively, $c(\cdot)$, $a(\cdot)$, and $d(\cdot)$ are known functions specifying the member of the exponential family, and for the survival part

$$p(T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = h_i(T_i \mid \mathcal{H}_i(T_i, \boldsymbol{b}_i))^{\delta_i} \exp\Bigl\{ - \int_0^{T_i} h_i(s \mid \mathcal{H}_i(s, \boldsymbol{b}_i))\, ds \Bigr\},$$

with $h_i(\cdot)$ given by (2). The integral in the definition of the survival function

$$S_i(t \mid \mathcal{H}_i(t), \boldsymbol{b}_i, \boldsymbol{w}_i) = \exp\Bigl\{ - \int_0^t h_0(s) \exp\bigl[\boldsymbol{\gamma}^\top \boldsymbol{w}_i + f\{\mathcal{H}_i(s), \boldsymbol{b}_i, \boldsymbol{\alpha}\}\bigr] ds \Bigr\}, \tag{7}$$

does not have a closed-form solution, and thus a numerical method must be employed for its evaluation. Standard options are the Gauss-Kronrod and Gauss-Legendre quadrature rule.

The penalized version of the B-spline approximation to the baseline hazard can be fitted by specifying for $\boldsymbol{\gamma}_{h_0}$ the improper prior (Lang and Brezger 2004):

$$p(\boldsymbol{\gamma}_{h_0} \mid \tau_h) \propto \tau_h^{\rho(K)/2} \exp\Big(-\frac{\tau_h}{2}\boldsymbol{\gamma}_{h_0}^\top \boldsymbol{K}\boldsymbol{\gamma}_{h_0}\Big),$$

where τ_h is the smoothing parameter that takes a Gamma$(1, \tau_{h\delta})$ prior distribution, with a hyper-prior $\tau_{h\delta} \sim$ Gamma$(10^{-3}, 10^{-3})$, which ensures a proper posterior distribution for $\boldsymbol{\gamma}_{h_0}$ (Jullion and Lambert 2007), $\boldsymbol{K} = \Delta_r^\top \Delta_r + 10^{-6}\mathbf{I}$, with Δ_r denoting the r-th difference penalty matrix, and $\rho(\boldsymbol{K})$ denotes the rank of $\boldsymbol{K}$.

2 Aortic Valve Dataset

- Figure 1 shows the longitudinal trajectories of aortic gradient for Patients 7 and 81 from the Aortic Valve dataset (main paper Section 4).

- Tables 1 and 2 present the posterior means and 95% credible intervals of the parameters of the joint models to the Aortic Valve data (main paper Section 4).

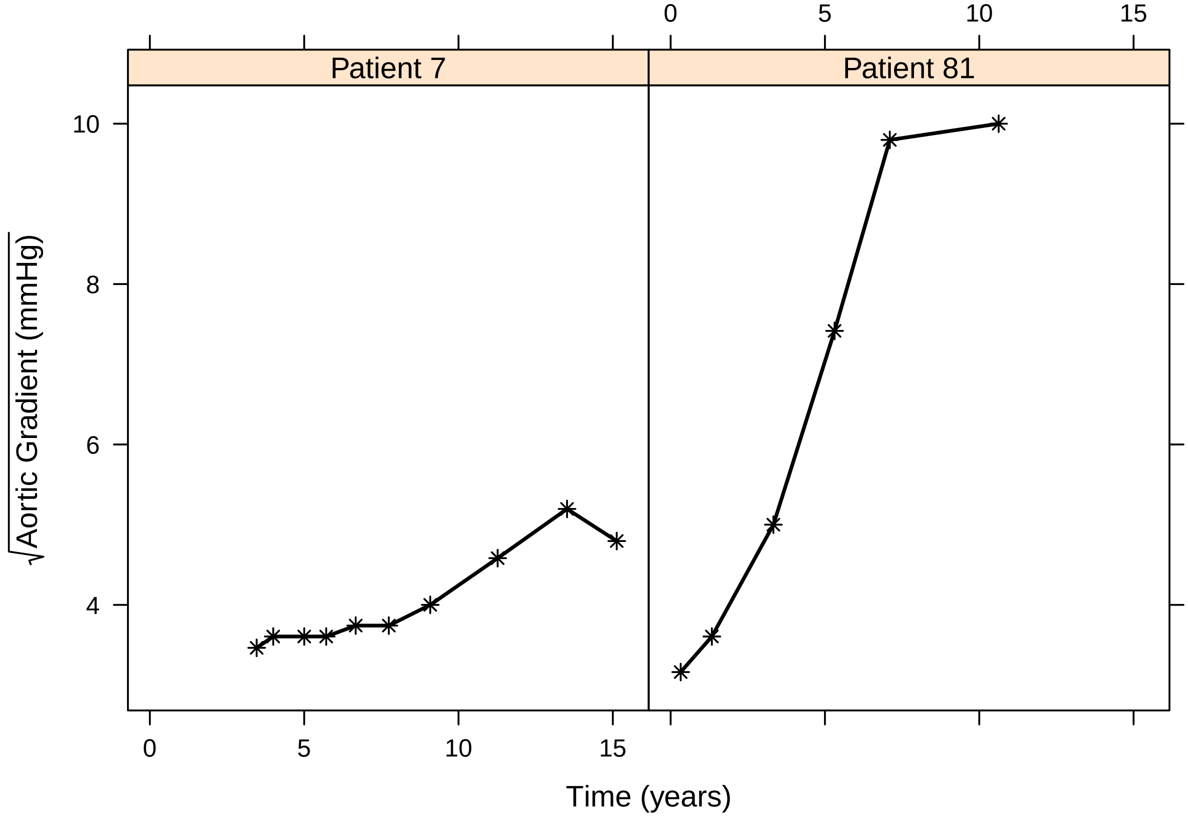

Figure 1: Longitudinal trajectories of the square root aortic gradient for Patients 7 and 81.

Table 1: Estimated coefficients and 95% credibility intervals for the parameters of the longitudinal submodels fitted to the Aortic Valve dataset. The top part of the table refers to the results for aortic gradient and the bottom part for aortic regurgitation. d_{ij} denotes the ij-th element of the corresponding covariance matrix of the random effects.

	M_1	M_2	M_3	M_4	M_5
	Value (95% CI)	Value (95% CI)	Value (95% CI)	Value (95% CI)	Value (95% CI)
Intercept	3.67 (3.33; 4.00)	3.66 (3.32; 4.01)	3.68 (3.34; 4.03)	3.67 (3.34; 4.02)	3.67 (3.32; 4.02)
B-spln1	3.28 (2.79; 3.83)	3.37 (2.84; 3.89)	3.31 (2.81; 3.84)	3.26 (2.76; 3.75)	3.34 (2.85; 3.83)
B-spln2	2.70 (2.33; 3.13)	2.77 (2.34; 3.21)	2.72 (2.35; 3.13)	2.68 (2.28; 3.08)	2.72 (2.31; 3.15)
Age	-0.02 (-0.02; -0.01)	-0.02 (-0.02; -0.01)	-0.02 (-0.02; -0.01)	-0.02 (-0.02; -0.01)	-0.02 (-0.02; -0.01)
Female	0.17 (-0.06; 0.40)	0.17 (-0.06; 0.39)	0.17 (-0.04; 0.41)	0.18 (-0.05; 0.41)	0.18 (-0.04; 0.41)
σ	0.61 (0.58; 0.65)	0.62 (0.58; 0.65)	0.62 (0.58; 0.65)	0.61 (0.58; 0.65)	0.62 (0.58; 0.66)
d_{11}	0.64 (0.48; 0.83)	0.62 (0.47; 0.80)	0.63 (0.48; 0.84)	0.64 (0.47; 0.82)	0.61 (0.46; 0.81)
d_{21}	-0.69 (-1.44; -0.06)	-0.51 (-1.19; 0.10)	-0.64 (-1.38; -0.04)	-0.69 (-1.42; -0.05)	-0.53 (-1.23; 0.12)
d_{31}	-0.43 (-1.05; 0.09)	-0.40 (-0.97; 0.15)	-0.43 (-1.00; 0.13)	-0.46 (-1.10; 0.11)	-0.34 (-0.96; 0.28)
d_{22}	13.61 (9.66; 18.74)	13.87 (9.79; 19.14)	13.50 (9.76; 18.14)	13.91 (9.85; 19.09)	13.06 (9.26; 17.83)
d_{32}	8.14 (4.40; 13.62)	9.46 (5.84; 14.20)	8.05 (4.55; 12.82)	8.70 (4.61; 13.66)	8.01 (4.09; 13.47)
d_{33}	7.04 (3.11; 12.99)	8.08 (4.62; 13.25)	6.63 (3.32; 11.71)	7.80 (3.55; 14.06)	7.31 (3.06; 15.57)

Table 2: Estimated coefficients and 95% credibility intervals for the parameters of the survival submodels fitted to the Aortic Valve dataset.

	$M_{1,1}$	$M_{2,1}$	$M_{3,1}$	$M_{4,1}$	$M_{5,1}$
	Value (95% CI)	Value (95% CI)	Value (95% CI)	Value (95% CI)	Value (95% CI)
`Age`	0.02 (0.01; 0.04)	0.02 (0.01; 0.04)	0.02 (0.01; 0.04)	0.02 (0.00; 0.03)	0.02 (0.01; 0.04)
`Female`	-0.09 (-0.48; 0.29)	-0.02 (-0.43; 0.36)	-0.09 (-0.49; 0.30)	-0.10 (-0.51; 0.29)	0.01 (-0.43; 0.45)
α_1	0.19 (0.09; 0.30)		0.16 (0.02; 0.30)	0.02 (0.01; 0.03)	-0.03 (-0.59; 0.51)
α_2		2.40 (1.07; 3.82)	0.78 (-1.76; 2.99)		0.43 (0.26; 0.65)
α_3					-0.02 (-0.75; 0.81)

3 Simulation Study

The data in the simulation study presented in Section 5 of the main paper have been simulated under the joint model:

$$\left\{\begin{array}{rcl} y_i(t) & = & \eta_i(t) + \varepsilon_i(t), \\ \eta_i(t) & = & (\beta_0 + b_{i0}) + (\beta_1 + b_{i1,1})B_n(t,1) + (\beta_2 + b_{i2})B_n(t,2) + \\ & & \qquad\qquad (\beta_3 + b_{i3})B_n(t,3), \\ \varepsilon_i(t) & \sim & \mathcal{N}(0, \sigma^2), \\ \boldsymbol{b}_i & \sim & \mathcal{N}(0, \boldsymbol{D}), \\ & & \\ h_i(t) & = & h_0(t)\exp\bigl\{\gamma_0 + \gamma_1 \texttt{Group}_i + \alpha_1 \eta_i(t) + \alpha_2 \eta_i'(t)\bigr\}, \\ h_0(t) & = & \phi t^{\phi - 1}, \end{array}\right.$$

where $\boldsymbol{B}_n(t, \{1,2,3\})$ denotes the B-spline basis for a natural cubic spline with boundary knots at baseline and 19.5 years and two internal knots placed at 2.1 and 5.5 years, and `Group` denotes a dummy variable for the type of operation. The parameter values that we used are:

* Fixed effects: $\beta_0 = 2.94$, $\beta_1 = 1.30$, $\beta_2 = 1.84$, and $\beta_3 = 1.82$;

* Random effects covariance matrix:

$$\boldsymbol{D} = \begin{bmatrix} 0.71 & & & \\ 0.33 & 2.68 & & \\ 0.07 & 3.81 & 7.62 & \\ 1.26 & 4.35 & 5.40 & 8.00 \end{bmatrix};$$

* Measurement error standard deviation: $\sigma = 0.60$;

* Baseline covariates relative risk model: $\gamma_0 = -6.70$ and $\gamma_1 = 0.50$;

* Association parameters: $\alpha_1 = 0.19$ and $\alpha_2 = -1.06$;

* Baseline hazard: $\phi = 2$;

* For each subject longitudinal measurements were planned to be taken at baseline, six months, one year, and biannually thereafter up to year 19;

* The censoring mechanism was based on a uniform distribution in the interval $[0, 28]$.

4 R Code

In this document we provide the relevant R code that can be used to fit joint models and compute $\text{cvD}\widehat{\text{CL}}(t)$ and $\widehat{\text{U}}(u \mid t)$ which have been implemented in functions `cvDCL()` and `dynInfo()`, respectively, available in package **JMbayes** (`version >= 0.7-2`; http://cran.r-project.org/package=JMbayes).

4.1 Fit Joint Models

```
# Fit the joint models with the different association structures
library("JMbayes")

lmeFit <- lme(sqrt(AoGradient) ~ ns(time, 2) + Age + sex, data = AoValv,
              random = ~ ns(time, 2) | id)
coxFit <- coxph(Surv(Time, event) ~ Age + sex, data = AoValv.id, x = TRUE)

# value
jointFit1 <- jointModelBayes(lmeFit, coxFit, timeVar = "time",
                             n.iter = 100000, n.burnin = 10000)

# slope & value + slope
dForm <- list(fixed = ~ 0 + dns(time, 2), random = ~ 0 + dns(time, 2),
              indFixed = 2:3, indRandom = 2:3)
jointFit2 <- update(jointFit1, param = "td-extra", extraForm = dForm)
jointFit3 <- update(jointFit1, param = "td-both", extraForm = dForm)

# integral
iForm <- list(fixed = ~ 0 + time + ins(time, 2) + I(time * Age) +
                  I(time * (sex == ’Female’)),
              random = ~ 0 + time + ins(time, 2),
              indFixed = 1:5, indRandom = 1:3)
jointFit4 <- update(jointFit1, param = "td-extra", extraForm = iForm)

# random effects
jointFit5 <- update(jointFit1, param = "shared-RE",
                        n.iter = 200000, n.adapt = 5000)
```

4.2 Use of cvDCL()

```
# Which model is best at different time points
models <- list(jointFit1, jointFit2, jointFit3, jointFit4, jointFit5)
times <- c(5, 7, 9, 11, 13)

cvDCLs <- matrix(0, length(models), length(times))
for (i in seq_along(times)) {
    cvDCLs[, i] <- sapply(models, cvDCL, newdata = AoValv, Tstart = times[i],
                          M = 1000)
}
dimnames(cvDCLs) <- list(paste0("jointFit", 1:5), paste("t =", times))
cvDCLs
```

4.3 Use of dynInfo()

```
# Data of Patient 81
ND <- AoValv[AoValv$id == 81, ]

# function to simulate longitudinal responses (the longitudinal
# submodel had as response the square root Aortic Gradient, hence to
# simulate on the original scale we need this function)
sfun <- function (eta, scale) {x <- rnorm(length(eta), eta, scale); x * x}

# loop over the visits of Patient 81
nn <- nrow(ND)
res81 <- vector("list", nn)
tup <- numeric(nn)
for (i in seq_len(nn)) {
    # data up to visit i
    ND.i <- ND[1:i, ]

    # conditional survival probabilities
    sfit <- survfitJM(jointFit2, newdata = ND.i)$summaries[[1]]
    # upper limit of the time interval to search for the optimal time
    tup[i] <- min(5, max(sfit[sfit[, "Mean"] > 0.8, "times"] - max(ND.i$time)))

    # find the optimal time
    v <- dynInfo(jointFit2, newdata = ND.i, Dt = tup[i], simulateFun = sfun)
    res81[[i]] <- v$summary
}
res81
```

References

Brown, E. (2009), “Assessing the association between trends in a biomarker and risk of event with an application in pediatric HIV/AIDS,” *The Annals of Applied Statistics*, 3, 1163–1182.

Eilers, P. and Marx, B. (1996), “Flexible smoothing with B-splines and penalties,” *Statistical Science*, 11, 89–121.

Jullion, A. and Lambert, P. (2007), “Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models,” *Computational Statistics and Data Analysis*, 51, 2542–2558.

Lang, S. and Brezger, A. (2004), “Bayesian P-splines,” *Journal of Computational and Graphical Statistics*, 13, 183–212.

Rizopoulos, D. (2012), *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*, Boca Raton: Chapman & Hall/CRC.

Rizopoulos, D. and Ghosh, P. (2011), “A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event,” *Statistics in Medicine*, 30, 1366–1380.

Rizopoulos, D., Hatfield, L., Carlin, B., and Takkenberg, J. (2014), “Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging,” *Journal of the American Statistical Association*, 109, 1385–1397.

Taylor, J., Park, Y., Ankerst, D., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T., and Sandler, H. (2013), “Real-time individual predictions of prostate cancer recurrence using joint models,” *Biometrics*, 69, 206–213.