Web-based Supplementary Materials for "Joint Partially Linear Model for Longitudinal Data with Informative Drop-outs" by

Sehee Kim*, Donglin Zeng and Jeremy M. G. Taylor

*email: seheek@umich.edu

A. Proof of Asymptotic Properties

This section proves Theorems 1-3 explained in Section 3.4 by using techniques from empirical process theory. Suppose the study duration is $\mathbb{T} = [0,\, \tau]$. Let $(\theta_0,\, \alpha_0,\, \Lambda_0)$ denote the true parameter values of $(\theta,\, \alpha,\, \Lambda)$, and $(\hat{\theta},\, \hat{\alpha},\, \hat{\Lambda})$ denote the MLEs. To establish the asymptotic properties of the MLEs, we impose the following regularity conditions:

(A1) The true parameter value θ_0 belongs to the interior of a compact set Θ within the domain of θ.

(A2) With probability 1, $X(t)$ and $Z(t)$ is left-continuous with uniformly bounded left and right derivatives in $[0, \tau]$.

(A3) For some constant c_0, $P(C \geq \tau \,|\, \mathcal{X}, \mathcal{Z}) > c_0 > 0$ with probability 1.

(A4) For some positive constant M_0, $M_0^{-1} < \sigma_{0e}^2 < M_0$ and $M_0^{-1} < c^T \Sigma_{0b}\, c < M_0$ for any $\|c\| = 1$.

(A5) The transformation function $H(\cdot)$ is four-times differentiable with $H(0) = 0$ and $H'(0) > 0$. In addition, there exist positive constants μ_0 and κ_0 such that

$$(1 + x)\, H'(x) \exp\{-H(x)\} \leq \mu_0 (1 + x)^{-\kappa_0}.$$

Furthermore, there exists a constant $\rho_0 > 0$ such that

$$\sup_x \left\{ \frac{|H''(x)| + |H^{(3)}(x)| + |H^{(4)}(x)|}{H'(x)\,(1+x)^{\rho_0}} \right\} < \infty,$$

where $H^{(3)}$ and $H^{(4)}$ are the third and fourth derivatives.

(A6) For some $t \in [0, \tau]$, if there exist a deterministic function $c(t)$ and v such that $c(t) + v^T X(t) = 0$ with probability 1, then $c(t) = 0$ and $v = 0$.

(A7) With some positive probability, $\boldsymbol{Z}_1^T \boldsymbol{Z}_1$ has full rank, where $\boldsymbol{Z}_1$ denotes a matrix with each row equal to the observed covariate $Z_1(t)^T$ at the time of each measurement.

(A8) The potential observation process of $Y(t)$ has a continuous intensity over $[0, \tau]$.

(A9) For a fixed integer $r \geq 2$, $\alpha_0(t)$ lies in $W^{r,\infty}(R)$, where $W^{r,\infty}(R)$ is a Sobolev space consisting of the functions with bounded rth derivatives.

(A10) For a fixed constant r_0 such that $1/(4r) < r_0 < 1/3$, (K_n, M_n) satisfies

$$\begin{aligned} M_n &= O(\log \log n), \\ K_n &= O(n^{r_0}). \end{aligned}$$

(A11) A partition of $\mathbb{T}$, $\{0 = s_0 < s_1 < \cdots < s_{K_n+1} = \tau\}$, satisfies

$$\bar{s}_{max} / \min_k \bar{s}_k \leq c_1, \quad \text{and} \quad \max_{1 \leq k \leq K_n} |\bar{s}_{k+1} - \bar{s}_k| = o(K_n^{-1}),$$

where $\bar{s}_k = s_k - s_{k-1}$, $\bar{s}_{max} = \max_{1 \leq k \leq K_n+1} \{\bar{s}_k\}$, and c_1 is a positive constant.

Conditions (A1)-(A3) are the standard assumptions in survival analysis. Condition (A4) is necessary to prove the existence of the NPMLEs. It can be easily verified that Condition (A5) holds for all transformations commonly used, including the logarithmic transformations described in Section 2. Condition (A6) implies that the columns of design matrix, say X, for fixed effects should be linearly independent. When it holds, the inverse of $X^T X$ exists. Condition (A7) implies that a similar condition is required for random effects. These conditions are used to prove identifiability of the parameters, following arguments similar to those given by Kim et al. (2012). Condition (A8) prescribes that some subjects have sufficient repeated measures. Finally, Condition (A9) grants sufficient smoothness of α_0, and Condition (A10) determines the size of the sieve space in terms of the number of knots and the upper bound of the sieve functions. Condition (A11) is the restriction on the length of subintervals of knots, which obviously holds for the equally spaced knots. For the percentile-based knots, $\min_k\{s_k - s_{k-1}\}$ should be away from zero to satisfy the condition.

Under the above conditions, the following theorem shows the consistency of the MLEs.

Theorem 1 *Under Conditions (A1) - (A11),*

$$\hat{\theta} \rightarrow_p \theta_0, \quad \|\hat{\alpha}(t) - \alpha_0(t)\|_{W^{1,\infty}(\mathbb{T})} \rightarrow_p 0, \quad \|\hat{\Lambda}(t) - \Lambda_0(t)\|_{L_\infty(\mathbb{T})} \rightarrow_p 0,$$

where $\|\cdot\|_{W^{1,\infty}(\mathbb{T})}$ *is the Sobolev norm on* $\mathbb{T}$ *and* $\|\cdot\|_{L_\infty(\mathbb{T})}$ *is the supremum norm on* $\mathbb{T}$*.*

Here, the Sobolev norm $\|\alpha(t)\|_{W^{r,\infty}(\mathbb{T})}$ is defined as $\|\alpha(t)\|_{L_\infty(\mathbb{T})} + \|\nabla_t^r \alpha(t)\|_{L_\infty(\mathbb{T})}$, where ∇_t^r is the rth derivative of a function with respect to t.

Proof of Theorem 1

Proof The whole proof can be divided into three steps: first, we construct some functions in a sieve space, which approximate the true parameters; then by using empirical process theory, we obtain one key inequality; finally, this inequality is used to obtain the consistency result. Without loss of generality, we assume $\mathbb{T} = [0, 1]$.

Step 1 We construct some functions in $S_n(m, K_n, M_n)$ to approximate the true parameter $\alpha_0(t)$. From the properties of B-spline functions (Schumaker, 2007, Chap. 4), we can define a linear operator $\mathcal{Q}$ mapping $W^{r,\infty}(\mathbb{T})$ to the sieve space. That is, for any $g \in W^{r,\infty}(\mathbb{T})$,

$$\mathcal{Q}[g] = \sum_{k=1}^{m+K_n} \Psi_k[g] B_k^m(t),$$

where Ψ_k are the linear functionals in $L_\infty(\mathbb{T})$. Moreover, according to Theorem 4.41 and Corollary 6.26 of Schumaker (2007),

$$|\Psi_k[g]| \leq (2m+1)9^{m-1}\|g\|_{L_\infty(\mathbb{T})},$$

and there exists a constant $C(m)$ depending only on m such that

$$\|g - \mathcal{Q}[g]\|_{L_\infty(\mathbb{T})} \leq C(m)\bar{\Delta}^r\|g\|_{W^{r,\infty}(\mathbb{T})},$$

where $\bar{\Delta} = \max\{s_{k+1} - s_k\}$. The above inequality holds even for non-equally-spaced knots because we have $\tau/c_1 \leq \bar{s}_{max}(K_n+1) \leq c_1\tau$ from Condition (A11). Now, we define $\alpha_n(t) = \mathcal{Q}[\alpha_0]$, then the following boundness holds

$$\|\alpha_n - \alpha_0\|_{L_\infty(\mathbb{T})} \leq O(K_n^{-r}).$$

Step 2 We obtain a key inequality based on empirical process theory. Let $\mathbb{P}_n$ be the empirical measure determined by n *iid* subjects, let $\boldsymbol{P}$ be its expectation, and let $\mathbb{G}_n$ be the empirical process given by $\sqrt{n}(\mathbb{P}_n - \boldsymbol{P})$. For simplicity of notation, we denote $\ell(\alpha, \theta, \Lambda)$ as the log-likelihood function from a single observation. Since $(\hat{\alpha}, \hat{\theta}, \hat{\Lambda})$ maximizes $\mathbb{P}_n[\ell(\alpha, \theta, \hat{\Lambda})]$ over the sieve space, it follows that

$$\mathbb{P}_n[\ell(\hat{\alpha},\, \hat{\theta},\, \hat{\Lambda})] \geq \mathbb{P}_n[\ell(\alpha_n, \theta_0, \hat{\Lambda})].$$

By following arguments similar to those given by Zeng (2005), we can show that the $\epsilon-$bracketing numbers covering the constructed sieve space $\mathcal{L}_n$, defined by

$$\mathcal{L}_n = \left\{\ell(\tilde{\alpha}_n, \theta, \hat{\Lambda}) - \ell(\alpha_n, \theta_0, \hat{\Lambda});\ \tilde{\alpha}_n \in S_n(m, K_n, M_n),\ \|\theta\| \leq M\right\},$$

is of order $O(e^{2M_n}/\epsilon)^{m+K_n+d}$, where $\|\theta\| = \sqrt{\theta^T\theta}$ for $\theta \in \mathbb{R}^d$ and M is a constant. According to the Theorem 19.35 of van der Vaart (1998), Theorem 1 of Zeng and Cai (2005) and by the identifiability conditions of Λ, which can be shown by following similar arguments in Kim et al. (2012), we obtain

$$\|\ell(\hat{\alpha},\, \hat{\theta},\, \hat{\Lambda}) - \ell(\alpha_0, \theta_0, \Lambda_0)\|^2_{L_2(P)} \leq O_p(1)\, B_n(n, K_n, M_n), \tag{A.1}$$

where $B_n(n, K_n, M_n) = e^{c_1 M_n} K_n^{1/2} \log K_n/\sqrt{n} + e^{c_1 M_n}/K_n^r$ and c_1 is a constant.

Step 3 We obtain the L_2-convergence of the estimators. Suppose we select K_n and M_n satisfying Assumption (A10). From the boundedness (A.1) and Assumptions (A6) and (A7) (i.e., identifiability conditions of the parameters), we can obtain that $\hat{\theta} \to_p \theta_0$, $\|\hat{\alpha} - \alpha_0\|_{L_2(P)} \to_p 0$ and $\|\hat{\Lambda} - \Lambda_0\|_{L_2(P)} \to_p 0$. Moreover, $B_n(n, K_n, M_n)^{1/2}$ is the convergence rate of $(\hat{\theta},\, \hat{\alpha},\, \hat{\Lambda})$.

To obtain the convergence of $\hat{\alpha}$ in $W^{1,\infty}$-space, we notice from Theorem 4.22 of Schumaker (2007) that

$$\|\nabla_t^r\, \hat{\alpha}(t)\|_{L_\infty(\mathbb{T})} \leq c_2\, K_n^r \sum_{k=1}^{m+K_n} |\zeta_k| \leq O(M_n K_n^r),$$

where ∇_t^r is the rth derivative of a function with respect to t. Hence, according to the Sobolev interpolation inequality (Adams and Fournier, 1975), we obtain

$$\|\hat{\alpha}(t) - \alpha_0(t)\|_{W^{1,\infty}(\mathbb{T})} \leq O_p(1)\, (M_n K_n^r)^{\delta_0} B_n(n, K_n, M_n)^{(1-\delta_0)/2}, \tag{A.2}$$

where $\delta_0 = 1/(2r)$. By Assumption (A10), the right side of (A.2) converges to zero. Therefore, Theorem 1 holds.

Now to achieve the asymptotic normality and the semiparametric efficiency of $\hat{\theta}$, we need a tighter bound for the convergence rate of the estimators, which is stated in Theorem 2.

Theorem 2 *Under Conditions (A1) - (A11),*

$$\|\hat{\alpha}(t) - \alpha_0(t)\|^2_{L_2(P)} + \|\hat{\Lambda}(t) - \Lambda_0(t)\|^2_{L_2(P)} \leq O_p(K_n^{-2r}) + o_p(n^{-1/2}),$$

where $\|\cdot\|_{L_2(P)}$ *is the* L_2*-norm with measure* P.

Proof of Theorem 2

Proof Using the results of Theorem 1, we repeat the Step 2 of Theorem 1. Since $\hat{\alpha}$ is within a $W^{1,\infty}$-neighborhood of α_0, based on the parameter identifiability conditions, the left-hand side of (A.1) can be further bounded from above by the $L_2(P)$-norm of $\|\hat{\Lambda} - \Lambda_0\|$ and $\|\hat{\alpha} - \alpha_0\|$. That is,

$$\|\ell(\hat{\alpha},\, \hat{\theta},\, \hat{\Lambda}) - \ell(\alpha_0, \theta_0, \Lambda_0)\|^2_{L_2(P)} \leq o_p(1/\sqrt{n}) + O_p(1/K_n^{2r}).$$

Thus, Theorem 2 holds.

Theorem 3 *Under Conditions (A1) - (A11),* $n^{1/2}(\hat{\theta} - \theta_0)$ *weakly converges to a zero-mean Gaussian process in* $\mathbb{R}^{d_\theta}$*, where* d_θ *is the dimension of* θ*. Furthermore, the asymptotic covariance matrix of* $n^{1/2}(\hat{\theta} - \theta_0)$ *achieves the semiparametric efficiency bound.*

Proof of Theorem 3

Proof We will prove Theorem 3 by writing $\sqrt{n}\,(\hat{\theta} - \theta_0)$ as a linear functional of the empirical process $\mathbb{G}_n$. Let $\ell(\alpha,\, \Lambda,\, \theta)$ be the log-likelihood function from a single subject, and let $\ell_0 = \ell(\alpha_0, \Lambda_0, \theta_0)$.

Step 1 We define a least favorable direction for θ_0. We treat $\psi = (\alpha,\, \Lambda)$ as the vector of nuisance parameters with $\psi_0 = (\alpha_0,\, \Lambda_0)$, and then the tangent space for ψ is given by $\mathcal{H} = \{h(t) = (h_1(t),\, h_2(t));\ h(t) \in L_2(\mathbb{T}^2)\}$. Let $\ell_\psi(\psi_0, \theta_0)[h]$ be the derivative of ℓ with respect to ψ along with the direction h_1 for α and the direction h_2 for Λ, and let $\ell_\theta(\psi_0, \theta_0)$

be the derivative of ℓ_0 with respect to θ. Then, a least favorable direction for θ_0 is defined as a tangent function $h(t) \in \mathcal{H}$ for ψ that satisfies

$$\ell_\psi^*(\psi_0, \theta_0)\ell_\psi(\psi_0, \theta_0)[h] = \ell_\psi^*(\psi_0, \theta_0)\ell_\theta(\psi_0, \theta_0) \quad a.s.,$$

where $\ell_\psi^*(\psi_0, \theta_0)$ is the adjoint operator of $\ell_\psi(\psi_0, \theta_0)$ in the Hilbert space $L_2(P)$.

Step 2 We prove the existence and smoothness of the least favorable direction. The existence can be shown by proving the operator $\ell_\psi^*(\psi_0, \theta_0)\ell_\psi(\psi_0, \theta_0)$ is invertible based on the Lax-Milgram theorem. The details of proofs are the same as in Zeng (2005).

Step 3 We construct the projection of $h_1(t)$ on the tangent space of the sieve space. The tangent function for ψ at $\hat{\psi} = (\hat{\alpha}, \hat{\Lambda})$ in the sieve space can be chosen by $h_n = (h_{1n}(t), h_2 d\hat{\Lambda})$ in $L_2(\mathbb{T}^2)$ such that

$$\|h_n - h\|_{L_2(P)}^2 \leq O(K_n^{-2r}) + o_p(n^{-1/2}).$$

Step 4 We derive the empirical process for $\sqrt{n}\,(\hat{\theta} - \theta_0)$. Since $(\hat{\psi}, \hat{\theta})$ maximizes the log-likelihood over the sieve space, the score along the path $(\hat{\psi} + \nu h_n, \hat{\theta} + \nu)$ is zero when $\nu = 0$. Thus, it holds that

$$\mathbb{G}_n\{\ell_\psi(\hat{\psi}, \hat{\theta})[h_n] + \ell_\theta(\hat{\psi}, \hat{\theta})\} = -\sqrt{n}\,\boldsymbol{P}\{\ell_\psi(\hat{\psi}, \hat{\theta})[h_n] + \ell_\theta(\hat{\psi}, \hat{\theta})\}. \tag{A.3}$$

Since the function in the left side of (A.3), indexed by both $(\hat{\psi}, h_n) \in W_{1,\infty}$ and $\hat{\theta} \in \Theta$, belongs to a P-Donsker class, we apply Theorem 2.11.23 of van der Vaart and Wellner (1996). By linearizing the right side of (A.3) at the true parameters and approximating h_n

to h, we obtain that

$$
\begin{aligned}
&-\boldsymbol{P}\{\ell_{\psi\theta}(\psi_0,\theta_0)[h]+\ell_{\theta\theta}(\psi_0,\theta_0)\}\sqrt{n}\,(\hat{\theta}-\theta_0)\\
&\quad=\mathbb{G}_n\{\ell_\psi(\psi_0,\theta_0)[h]+\ell_\theta(\psi_0,\theta_0)\}+\sqrt{n}\,O_p(\|\hat{\psi}-\psi_0\|^2_{L_2(P)}+\|h_n-h\|^2_{L_2(P)}+\|\hat{\theta}-\theta_0\|^2)\\
&\quad\leq\mathbb{G}_n\{\ell_\psi(\psi_0,\theta_0)[h]+\ell_\theta(\psi_0,\theta_0)\}+O_p(\sqrt{n}/K_n^{2r})+o_p(1).
\end{aligned}
$$

Since the second term in the right side of the above equation is $o_p(1)$ by Theorem 2 and Assumption (A10) and $-\boldsymbol{P}\{\ell_{\psi\theta}(\psi_0,\theta_0)[h]+\ell_{\theta\theta}(\psi_0,\theta_0)\}>0$, the asymptotic normality of $\sqrt{n}\,(\hat{\theta}-\theta_0)$ holds. Moreover, the influence function of $\hat{\theta}$ is given by

$$[-\boldsymbol{P}\{\ell_{\psi\theta}(\psi_0,\theta_0)[h]+\ell_{\theta\theta}(\psi_0,\theta_0)\}]^{-1}\{\ell_\psi(\psi_0,\theta_0)[h]+\ell_\theta(\psi_0,\theta_0)\}.$$

Clearly, the above influence function is contained in the tangent space, therefore, we conclude that $\hat{\theta}$ is semiparametrically efficient.

B. Web Tables and Figures

Web Tables 1-3 present the simulation results based on 1000 replications with $n=500$ and $\alpha(t)=(t+0.5)^{-1.5}+(t/2+1)^3-4$. The Web Tables 1-3 report the average of the differences between the true parameter and the estimates (Bias), the sample standard deviation of the parameter estimates (SD), and the average of the standard error estimates (SEE), and the coverage probability of 95% confidence intervals (CP). The confidence intervals for variances are constructed based on the the Satterthwaite approximation. Web Tables 1 and 2 summarize the performance of the proposed $\hat{\alpha}(t)$, in terms of Bias, SD, the mean square error (MSE), and the ratio of the MSE for $\alpha(t)$ estimates in the joint model to the counterpart in the marginal model (MSER). Individual trajectories of PSA are illustrated in Web Figure 1. The analysis results of the prostate cancer data with the AIC-based knot selection procedure are presented in Web Table 4, and Web Figures 3-4.

Web Table 1: Simulation results for true $\phi = 0.5$ (i.e., missing not at random) with $K_n \in \{k_{AIC}, k_{BIC}, 2, 4, 8\}$ interior knots of B-spline approximation. True values are $\alpha(\tau_{20}) = -1.10$, $\alpha(\tau_{40}) = -0.58$, and $\alpha(\tau_{80}) = 2.16$, where τ_p represents $p\%$ of study duration τ.

		Joint Model					**Marginal Model**			
K_n		*Bias*	*SD*	*SEE*	*CP*	*MSE*	*Bias*	*SD*	*MSE*	*MSER*
						$H(x) = x$				
k_{AIC}	$\alpha(\tau_{20})$	-0.007	0.035	0.034	0.941	0.001	-0.039	0.036	0.003	0.448
	$\alpha(\tau_{40})$	-0.008	0.043	0.041	0.936	0.002	-0.048	0.043	0.004	0.453
	$\alpha(\tau_{80})$	-0.008	0.080	0.076	0.931	0.006	-0.060	0.080	0.010	0.638
k_{BIC}	$\alpha(\tau_{20})$	-0.010	0.032	0.033	0.950	0.001	-0.042	0.033	0.003	0.398
	$\alpha(\tau_{40})$	-0.009	0.040	0.039	0.942	0.002	-0.049	0.041	0.004	0.415
	$\alpha(\tau_{80})$	-0.005	0.078	0.075	0.937	0.006	-0.057	0.078	0.009	0.648
2	$\alpha(\tau_{20})$	-0.010	0.032	0.033	0.948	0.001	-0.042	0.033	0.003	0.394
	$\alpha(\tau_{40})$	-0.009	0.040	0.039	0.943	0.002	-0.049	0.041	0.004	0.411
	$\alpha(\tau_{80})$	-0.004	0.078	0.075	0.935	0.006	-0.056	0.078	0.009	0.651
4	$\alpha(\tau_{20})$	-0.006	0.032	0.033	0.961	0.001	-0.038	0.033	0.003	0.423
	$\alpha(\tau_{40})$	-0.007	0.041	0.040	0.940	0.002	-0.047	0.042	0.004	0.442
	$\alpha(\tau_{80})$	-0.009	0.078	0.075	0.932	0.006	-0.061	0.078	0.010	0.625
8	$\alpha(\tau_{20})$	-0.005	0.035	0.037	0.959	0.001	-0.037	0.036	0.003	0.477
	$\alpha(\tau_{40})$	-0.008	0.044	0.042	0.939	0.002	-0.048	0.045	0.004	0.465
	$\alpha(\tau_{80})$	-0.011	0.080	0.078	0.932	0.007	-0.063	0.081	0.010	0.632
						$H(x) = log(1 + x)$				
k_{AIC}	$\alpha(\tau_{20})$	-0.007	0.035	0.034	0.937	0.001	-0.031	0.037	0.002	0.551
	$\alpha(\tau_{40})$	-0.007	0.041	0.039	0.936	0.002	-0.037	0.044	0.003	0.523
	$\alpha(\tau_{80})$	-0.007	0.071	0.067	0.937	0.005	-0.049	0.076	0.008	0.623
k_{BIC}	$\alpha(\tau_{20})$	-0.011	0.033	0.032	0.929	0.001	-0.035	0.035	0.002	0.497
	$\alpha(\tau_{40})$	-0.005	0.038	0.038	0.951	0.001	-0.035	0.041	0.003	0.505
	$\alpha(\tau_{80})$	-0.004	0.069	0.066	0.940	0.005	-0.046	0.074	0.008	0.630
2	$\alpha(\tau_{20})$	-0.012	0.033	0.032	0.929	0.001	-0.036	0.034	0.002	0.490
	$\alpha(\tau_{40})$	-0.005	0.038	0.038	0.953	0.001	-0.035	0.041	0.003	0.506
	$\alpha(\tau_{80})$	-0.004	0.069	0.066	0.941	0.005	-0.046	0.074	0.008	0.629
4	$\alpha(\tau_{20})$	-0.007	0.033	0.033	0.948	0.001	-0.030	0.035	0.002	0.529
	$\alpha(\tau_{40})$	-0.008	0.039	0.038	0.945	0.002	-0.038	0.041	0.003	0.491
	$\alpha(\tau_{80})$	-0.007	0.069	0.067	0.937	0.005	-0.049	0.074	0.008	0.612
8	$\alpha(\tau_{20})$	-0.006	0.035	0.035	0.949	0.001	-0.029	0.037	0.002	0.564
	$\alpha(\tau_{40})$	-0.008	0.041	0.040	0.938	0.002	-0.038	0.043	0.003	0.513
	$\alpha(\tau_{80})$	-0.009	0.072	0.069	0.938	0.005	-0.051	0.076	0.008	0.618

Web Table 2: Simulation results for true $\phi = 0$ (i.e., missing at random) with $K_n \in \{k_{AIC}, k_{BIC}, 2, 4, 8\}$ interior knots of B-spline approximation. True values are $\alpha(\tau_{20}) = -1.10$, $\alpha(\tau_{40}) = -0.58$, and $\alpha(\tau_{80}) = 2.16$, where τ_p represents p% of study duration τ.

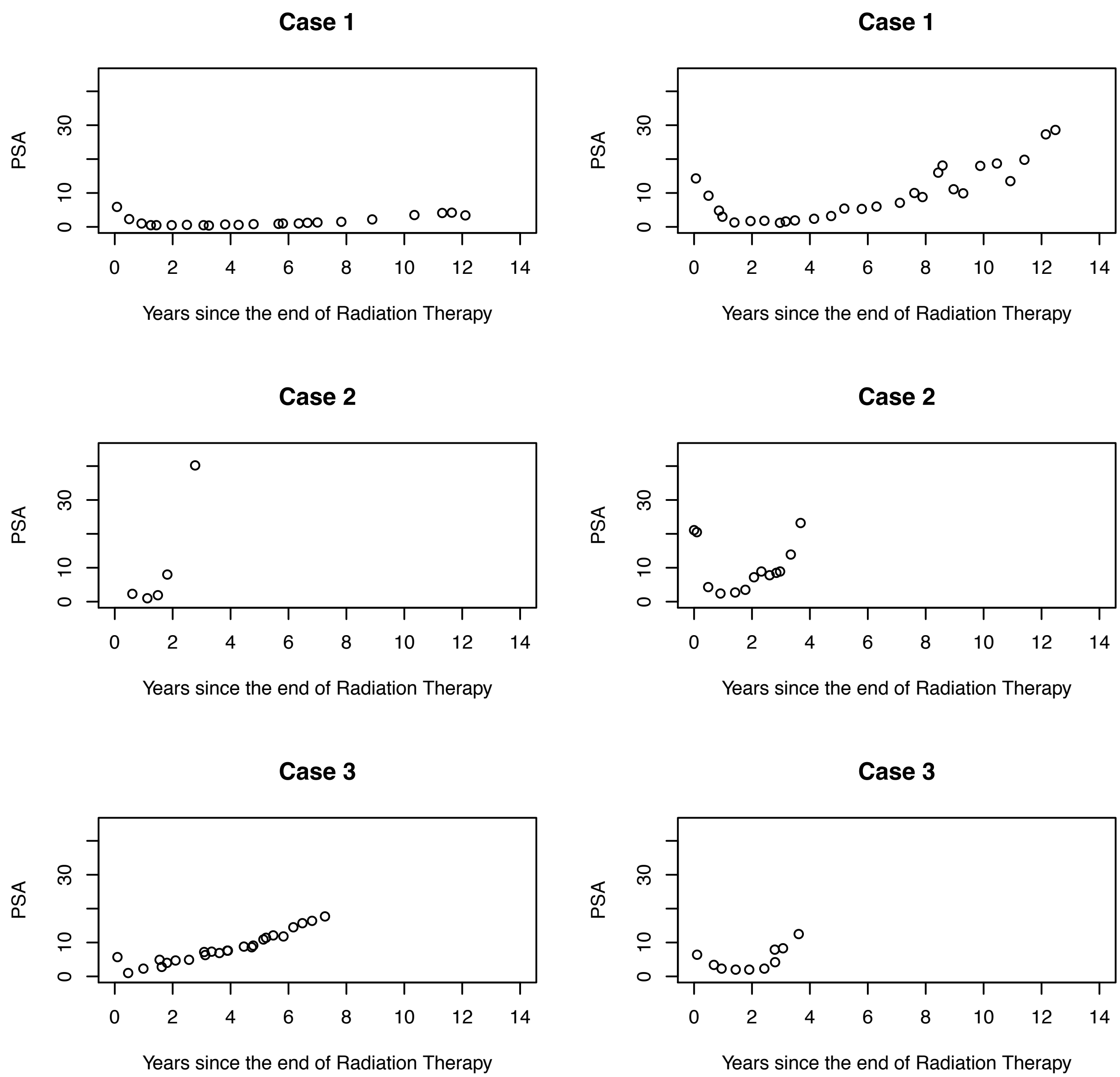
		Joint Model					**Marginal Model**			
K_n		*Bias*	*SD*	*SEE*	*CP*	*MSE*	*Bias*	*SD*	*MSE*	*MSER*
						$H(x) = x$				
k_{AIC}	$\alpha(\tau_{20})$	0.000	0.035	0.034	0.950	0.001	0.000	0.036	0.001	0.916
	$\alpha(\tau_{40})$	0.001	0.041	0.040	0.942	0.002	0.001	0.044	0.002	0.878
	$\alpha(\tau_{80})$	0.001	0.077	0.075	0.941	0.006	0.002	0.083	0.007	0.855
k_{BIC}	$\alpha(\tau_{20})$	-0.003	0.032	0.032	0.958	0.001	-0.003	0.034	0.001	0.901
	$\alpha(\tau_{40})$	0.000	0.039	0.039	0.945	0.001	0.000	0.042	0.002	0.862
	$\alpha(\tau_{80})$	0.006	0.074	0.074	0.944	0.006	0.006	0.081	0.007	0.852
2	$\alpha(\tau_{20})$	-0.003	0.031	0.032	0.958	0.001	-0.003	0.031	0.001	0.998
	$\alpha(\tau_{40})$	0.000	0.038	0.039	0.946	0.001	0.000	0.038	0.001	0.998
	$\alpha(\tau_{80})$	0.006	0.074	0.074	0.942	0.006	0.006	0.074	0.006	1.000
4	$\alpha(\tau_{20})$	0.001	0.032	0.033	0.962	0.001	0.001	0.032	0.001	0.997
	$\alpha(\tau_{40})$	0.003	0.039	0.040	0.946	0.002	0.003	0.039	0.002	0.998
	$\alpha(\tau_{80})$	0.001	0.075	0.075	0.949	0.006	0.001	0.075	0.006	1.000
8	$\alpha(\tau_{20})$	0.002	0.036	0.036	0.955	0.001	0.002	0.036	0.001	0.998
	$\alpha(\tau_{40})$	0.002	0.042	0.042	0.952	0.002	0.002	0.042	0.002	0.998
	$\alpha(\tau_{80})$	-0.001	0.078	0.077	0.947	0.006	-0.001	0.078	0.006	1.000
						$H(x) = log(1+x)$				
k_{AIC}	$\alpha(\tau_{20})$	-0.002	0.034	0.034	0.943	0.001	-0.002	0.036	0.001	0.890
	$\alpha(\tau_{40})$	0.000	0.040	0.039	0.950	0.002	0.000	0.044	0.002	0.828
	$\alpha(\tau_{80})$	0.001	0.067	0.067	0.943	0.005	0.000	0.075	0.006	0.814
k_{BIC}	$\alpha(\tau_{20})$	-0.005	0.032	0.032	0.944	0.001	-0.005	0.034	0.001	0.879
	$\alpha(\tau_{40})$	0.002	0.038	0.038	0.954	0.001	0.002	0.042	0.002	0.812
	$\alpha(\tau_{80})$	0.004	0.065	0.066	0.948	0.004	0.003	0.073	0.005	0.806
2	$\alpha(\tau_{20})$	-0.006	0.032	0.032	0.943	0.001	-0.006	0.032	0.001	0.993
	$\alpha(\tau_{40})$	0.003	0.038	0.038	0.954	0.001	0.003	0.038	0.001	0.992
	$\alpha(\tau_{80})$	0.004	0.065	0.066	0.947	0.004	0.004	0.065	0.004	1.000
4	$\alpha(\tau_{20})$	-0.001	0.032	0.033	0.940	0.001	-0.001	0.032	0.001	0.993
	$\alpha(\tau_{40})$	-0.000	0.038	0.038	0.957	0.001	-0.000	0.038	0.001	0.993
	$\alpha(\tau_{80})$	0.001	0.066	0.066	0.948	0.004	0.001	0.066	0.004	1.000
8	$\alpha(\tau_{20})$	0.000	0.034	0.035	0.953	0.001	0.000	0.034	0.001	0.993
	$\alpha(\tau_{40})$	-0.000	0.040	0.040	0.960	0.002	-0.000	0.040	0.002	0.995
	$\alpha(\tau_{80})$	-0.001	0.068	0.069	0.952	0.005	-0.001	0.068	0.005	1.000

Web Table 3: Simulation results by varying ϕ and $H(\cdot)$ when the interior knots of B-spline approximation are selected by AIC (i.e., $K_n = k_{AIC}$). τ_p represents $p\%$ of study duration τ.

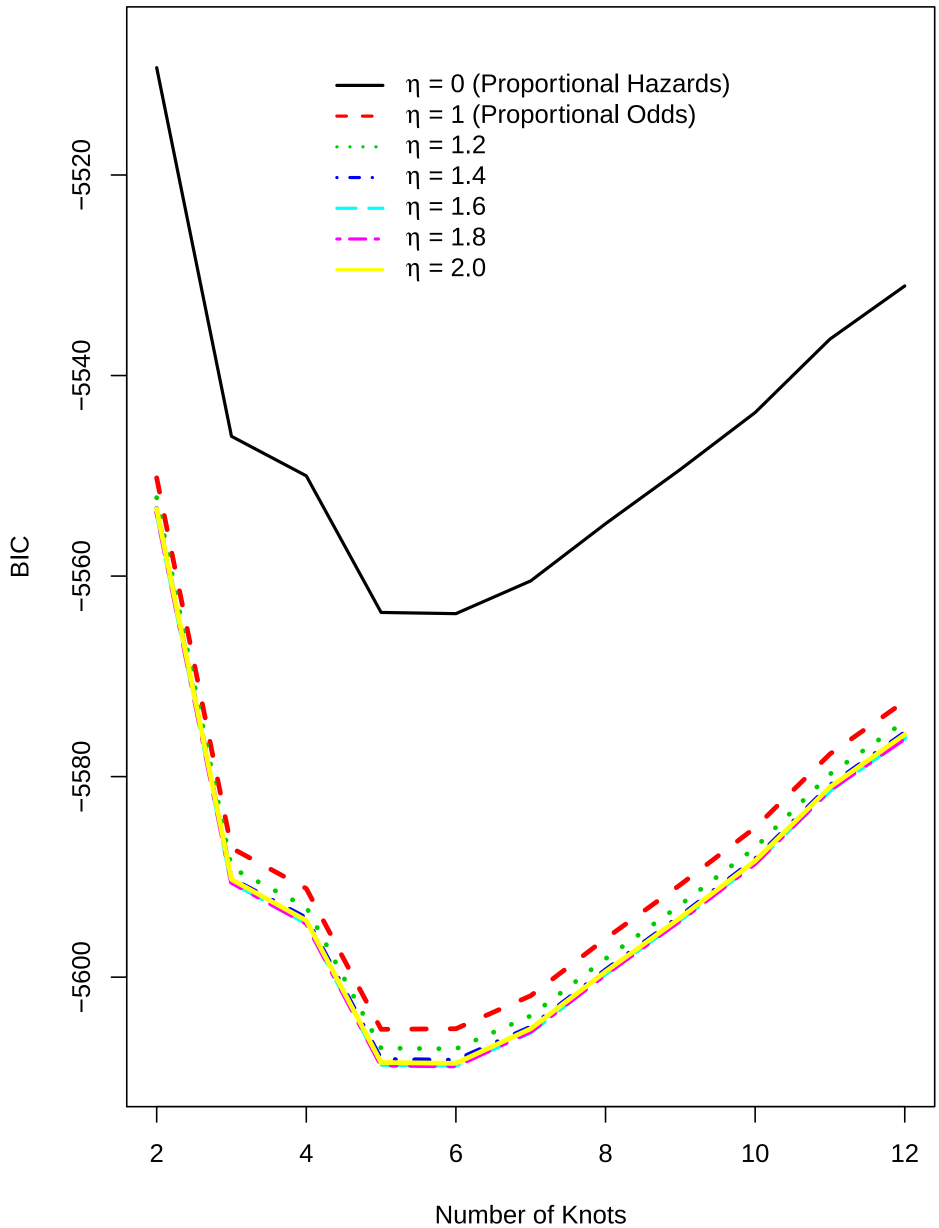
		$H(x) = x$				$H(x) = \log(1 + x)$			
Parameter	*Target*	*Bias*	*SD*	*SEE*	*CP*	*Bias*	*SD*	*SEE*	*CP*
					$\phi = 0.0$				
β	0.5	-0.000	0.031	0.029	0.938	-0.000	0.028	0.029	0.961
σ_e^2	0.1	-0.000	0.003	0.003	0.940	-0.000	0.003	0.003	0.951
σ_{b1}^2	0.4	-0.003	0.028	0.028	0.949	-0.002	0.028	0.028	0.959
σ_{b2}^2	0.2	-0.002	0.030	0.030	0.955	-0.002	0.026	0.027	0.952
ρ	-0.1	0.002	0.087	0.085	0.946	0.003	0.076	0.079	0.961
γ	0.5	0.020	0.065	0.066	0.951	0.020	0.095	0.095	0.947
ϕ	0.0	-0.002	0.106	0.105	0.945	0.001	0.151	0.154	0.959
$\Lambda(\tau_{20})$	0.3	-0.007	0.029	0.029	0.948	-0.011	0.035	0.035	0.943
$\Lambda(\tau_{40})$	0.6	-0.003	0.050	0.050	0.940	-0.006	0.065	0.067	0.954
$\Lambda(\tau_{80})$	1.3	0.023	0.116	0.115	0.950	0.011	0.165	0.165	0.954
					$\phi = 0.5$				
β	0.5	-0.001	0.029	0.030	0.950	-0.001	0.030	0.029	0.939
σ_e^2	0.1	-0.000	0.003	0.003	0.944	-0.000	0.003	0.003	0.937
σ_{b1}^2	0.4	-0.002	0.028	0.028	0.957	-0.003	0.028	0.028	0.947
σ_{b2}^2	0.2	-0.001	0.031	0.030	0.943	-0.002	0.028	0.027	0.953
ρ	-0.1	0.002	0.090	0.087	0.942	0.003	0.082	0.080	0.948
γ	0.5	0.021	0.068	0.068	0.947	0.023	0.099	0.097	0.955
ϕ	0.5	0.019	0.108	0.108	0.946	0.014	0.159	0.158	0.957
$\Lambda(\tau_{20})$	0.3	-0.008	0.031	0.030	0.936	-0.011	0.034	0.036	0.951
$\Lambda(\tau_{40})$	0.6	-0.005	0.052	0.051	0.945	-0.009	0.066	0.068	0.958
$\Lambda(\tau_{80})$	1.3	0.023	0.124	0.119	0.942	0.005	0.161	0.167	0.964
					$\phi = 1.7$				
β	0.5	0.000	0.030	0.030	0.948	-0.000	0.030	0.029	0.947
σ_e^2	0.1	-0.000	0.003	0.003	0.946	-0.000	0.003	0.003	0.952
σ_{b1}^2	0.4	-0.001	0.030	0.029	0.939	-0.001	0.029	0.029	0.948
σ_{b2}^2	0.2	-0.000	0.029	0.030	0.957	-0.003	0.027	0.027	0.957
ρ	-0.1	-0.001	0.101	0.102	0.952	0.001	0.089	0.087	0.944
γ	0.5	0.026	0.087	0.086	0.941	0.026	0.112	0.111	0.947
ϕ	1.7	0.091	0.152	0.149	0.920	0.086	0.192	0.195	0.937
$\Lambda(\tau_{20})$	0.3	-0.015	0.036	0.035	0.935	-0.015	0.040	0.041	0.946
$\Lambda(\tau_{40})$	0.6	-0.009	0.067	0.065	0.942	-0.014	0.078	0.080	0.952
$\Lambda(\tau_{80})$	1.3	0.025	0.151	0.153	0.959	-0.000	0.190	0.191	0.952

Web Table 4: Joint analysis results of the prostate cancer data under the best fit of transformation $H(x) = \log(1 + x)$ and 7 interior knots selected by AIC. The 50:50 mixture of χ^2 distributions is used for testing variances. Reference groups for categorical covariates are T-stage=1, Gleason score between 2 and 6, and age<65. τ_p represents $p\%$ of study duration τ.

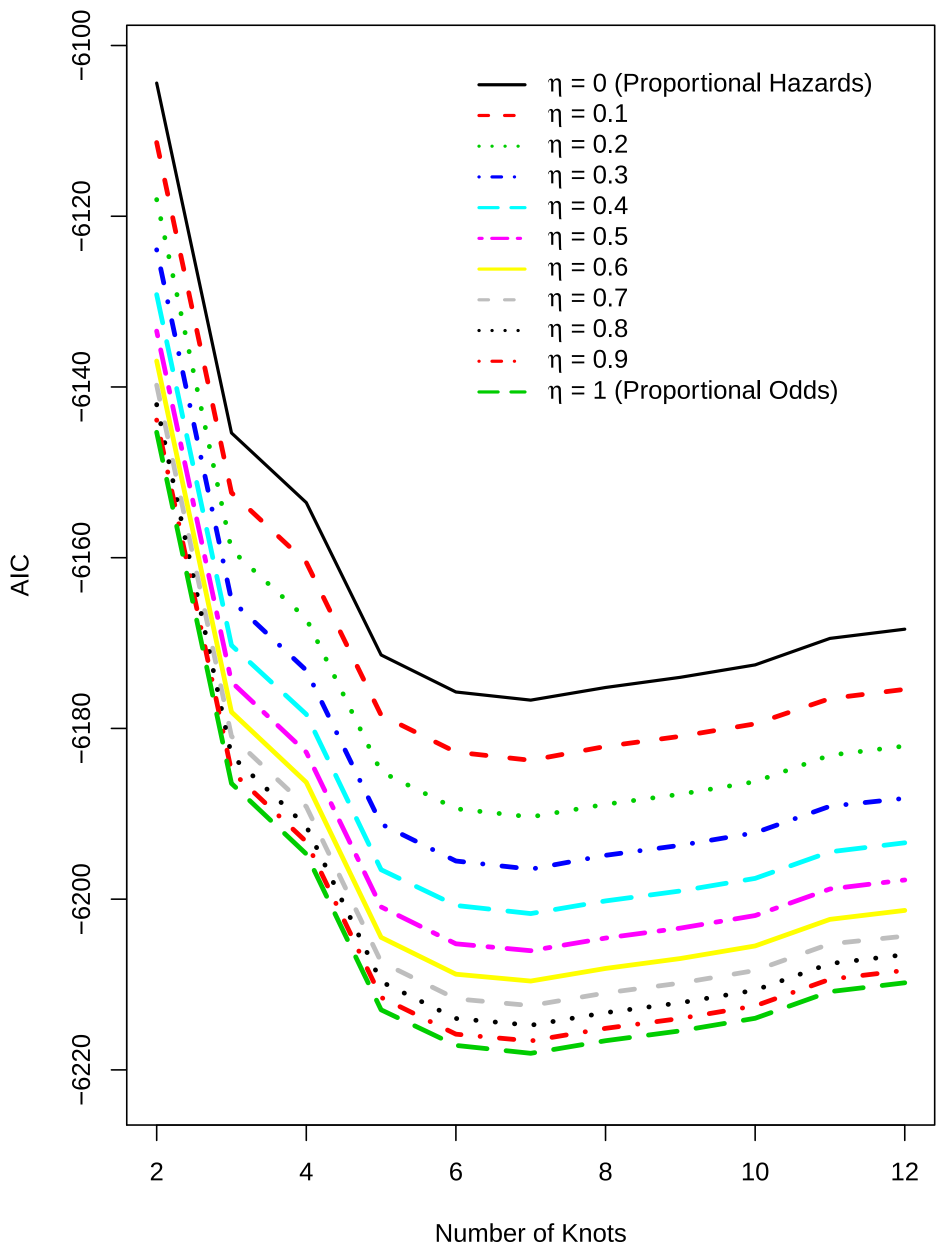
	Joint Model			**Marginal Model**		
Effect	*Est*	*SE*	*p-value*	*Est*	*SE*	*p-value*
			Longitudinal PSA score			
$\alpha(\tau_{20})$	0.331	0.088	.0002	0.546	0.092	$< .0001$
$\alpha(\tau_{40})$	0.960	0.136	$< .0001$	1.368	0.142	$< .0001$
$\alpha(\tau_{60})$	1.791	0.195	$< .0001$	2.390	0.201	$< .0001$
$\alpha(\tau_{80})$	2.628	0.263	$< .0001$	3.419	0.270	$< .0001$
log(baseline PSA+0.1)	0.510	0.034	$< .0001$	0.508	0.034	$< .0001$
T-stage 2	-0.062	0.064	.3334	-0.049	0.064	.4480
T-stage 3 or 4	0.012	0.117	.9154	0.021	0.117	.8602
Gleason score 7 to 9	0.056	0.060	.3462	0.042	0.060	.4793
Age 65-75 years	-0.191	0.069	.0054	-0.203	0.069	.0035
Age > 75 years	-0.117	0.088	.1829	-0.131	0.088	.1362
σ_e^2	0.117	0.003	$< .0001$	0.118	0.003	$< .0001$
			Random effects			
$\sigma_{b_1}^2$	0.373	0.028	$< .0001$	0.371	0.027	$< .0001$
$\sigma_{b_2}^2$	0.207	0.018	$< .0001$	0.196	0.017	$< .0001$
ρ	-0.121	0.055	.0290	-0.113	0.054	.0377
			Informative drop-out			
log(baseline PSA+0.1)	0.530	0.161	.0010			
T-stage 2	1.183	0.320	.0002			
T-stage 3 or 4	1.800	0.467	.0001			
Gleason score 7 to 9	1.090	0.293	.0002			
Age 65-75 years	-1.183	0.329	.0003			
Age > 75 years	-1.546	0.471	.0010			
ϕ	1.730	0.134	$< .0001$			



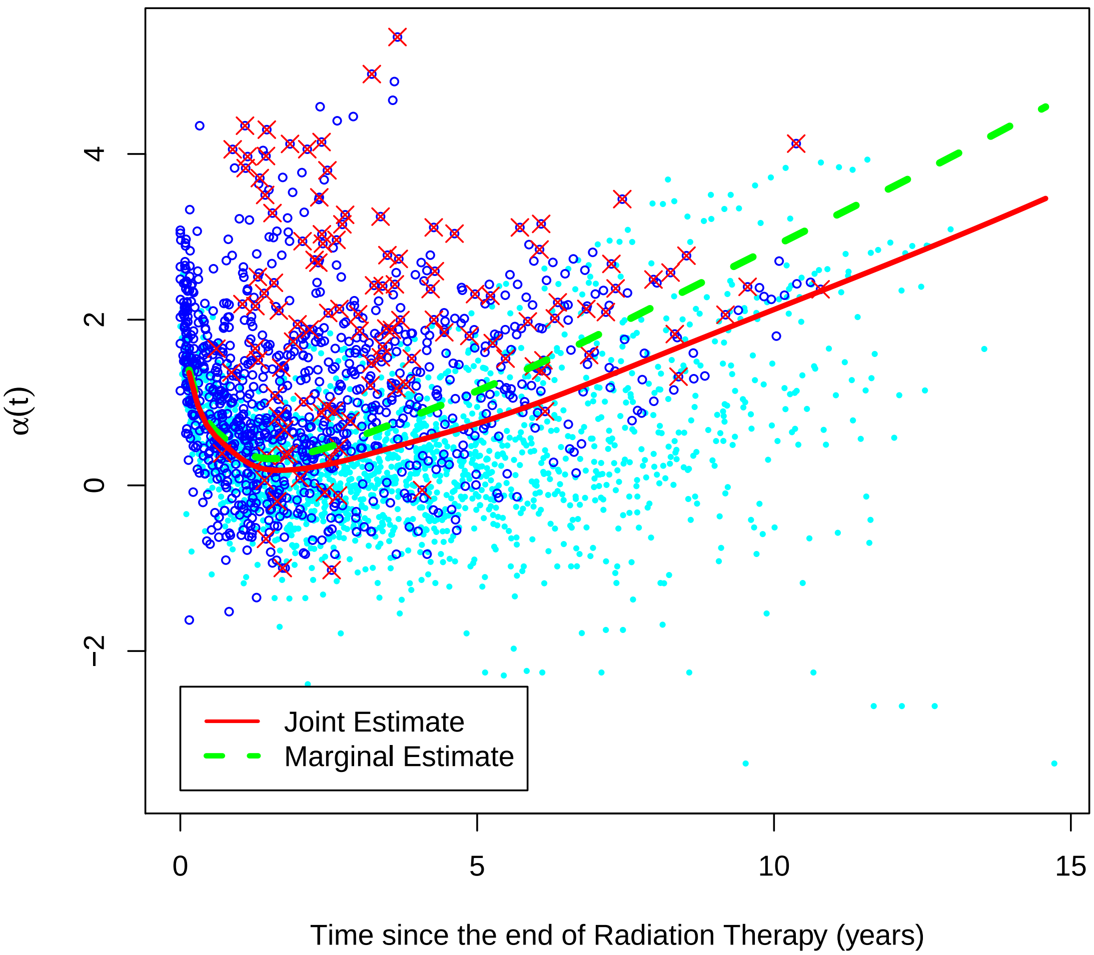
Web Figure 1: Individual trajectory of PSA. Case 1 are sample profiles for patients who had complete follow-up. Case 2 are sample profiles for patients who dropped out of the study due to the initiation of salvage hormone therapy. Case 3 are sample profiles for patients who dropped out of the study due to prostate cancer recurrence.



Web Figure 2: Bayesian information criterion (BIC) plotted for different transformations $H(x) = \log(1 + \eta x)/\eta$ and different numbers of interior knots (K_n). Expanding the range of transformation parameter to $\eta > 1$ led to the smallest BIC at $\eta = 1.6$ and $K_n = 6$.



Web Figure 3: Akaike information criterion (AIC) by the transformation $H(x) = \log(1 + \eta x)/\eta$ and the number of interior knots (K_n).



Web Figure 4: Coefficient function of log PSA score, adjusted by T-stage, gleason score, and age, under the best fit of transformation $H(x) = \log(1 + x)$ and 7 control points. The solid curve is an estimate from the joint model, and the dashed curve is an estimate from the marginal model. The circles and dots present the full history of all post-radiation PSA values for patients whose follow-up was informatively and non-informatively censored, respectively. The mark '×'s on some circles indicate the last observation of PSA before the informative censoring occurred.

C. Assessing the Fit of Joint Model to Observed Data

In this section, we illustrate a graphical tool to assess the proposed joint model's fit to the observed data. Overall fit and model assumptions on observed covariates will be examined based on residual plots. Recall that the final models fitted to the prostate cancer data were

$$\begin{aligned} y_i(t_{ij}) &= \alpha(t_{ij}) + \beta^T X + b_{1i} + b_{2i} t_{ij} + \epsilon(t_{ij}), \\ \Lambda(t \mid X, b) &= H\left(\int_0^t \exp\{\gamma^T X + \phi(b_{1i} + b_{2i} u)\} \, d\Lambda(u) \right), \end{aligned}$$

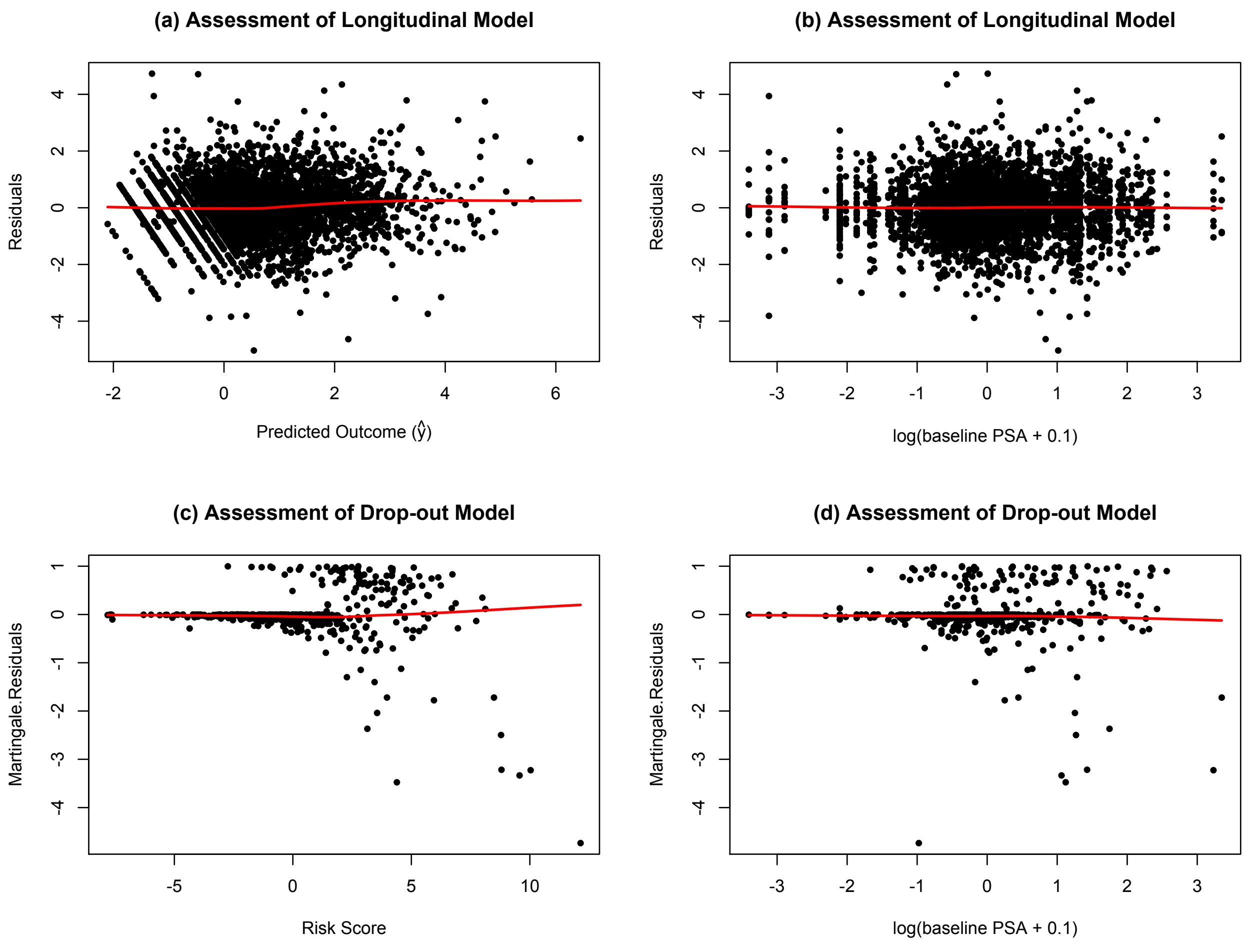
where $H(x) = \log(1 + x)$. For our purposes, we consider two types of residuals for each of the two processes. First, the residuals for the longitudinal model are defined as

$$r_i(t_{ij}) = \{y_i(t_{ij}) - \hat{\alpha}(t_{ij}) - \hat{\beta}^T X_1(t_{ij}) - \hat{b}_{1i} - \hat{b}_{2i} t_{ij}\} / \hat{\sigma}_e,$$

which is the difference between the observed and fitted values, conditioning on the empirical Bayes estimates $\hat{b}_{1i}$ and $\hat{b}_{2i}$ given the model. Second, the martingale-based residuals (Therneau et al., 1990) for the survival model are defined as $M_i(\infty)$, where

$$M_i(t) = N_i(t) - H\left(\int_0^t I(V_i \geq u) \exp\{\hat{\gamma}^T X + \hat{\phi}(\hat{b}_{1i} + \hat{b}_{2i} u)\} \, d\hat{\Lambda}(u) \right)$$

is the difference between the observed number of events over $[0,\ t]$ and the expected counterpart given the fitted model. For each individual in the data set we have computed both residuals and the predicted outcomes $\hat{y}_i(t_{ij}) = \hat{\alpha}(t_{ij}) + \hat{\beta}^T X_1(t_{ij}) + \hat{b}_{1i} + \hat{b}_{2i} t_{ij}$ and the risk score $\hat{\gamma}^T X + \hat{\phi}(\hat{b}_{1i} + \hat{b}_{2i} V_i)$ at the observed survival time V_i. In Web Figure 5 (a) and (c), we observed that the fitted loess curves in the plots of the residuals versus the fitted values showed no systematic trends, suggesting a good overall fit for both longitudinal and dropout models. In Web Figure 5 (b) and (d), no systematic trends in residual plots against log(baseline PSA + 0.1) values, the only continuous covariate in the fitted model, were found. It appeared to support the linearity on baseline PSA effect.



Web Figure 5: Residuals for the prostate cancer data.

D. Sensitivity Analysis to the Dropout Mechanism

In this section, we suggest a simple way of examining sensitivity to the assumptions about the missing data mechanism. As stated in the National Research Council (2010) report, it is important to note:

> *With incomplete data, inference about the treatment arm means requires two types of assumptions: (i) untestable assumptions about the distribution of missing outcomes data, and (ii) testable assumptions about the distribution of observed outcomes.* (Chapter 5, page 85)

Strategies to check type (ii) assumptions have been extensively discussed in Section 6 and Web Section C. Type (i) assumptions, however, are not testable with missing outcomes; therefore, it is essential to conduct a sensitivity analysis under different type (i) assumptions. In the proposed joint partially linear model (1), the distribution of the outcomes after drop-out is nonidentifiable, and thereby we assume that it remains the same as before the drop-out. One simple approach to exploring sensitivity to this assumption is to 1) introduce a sensitivity parameter as the difference between the mean of observed and unobserved responses, and then 2) examine how sensitive the results are over a clinically plausible range of the sensitivity parameter.

Suppose an individual dropped out of the study at time d, and we focus on inference about the mean $\mu(s) = E[Y(s)|X]$ of the intended outcome at time s. A sensitivity parameter $\delta(s)$ can be formulated in our partially linear model with the following mean function:

$$E[Y(s)|T = d, X] \quad = \quad \alpha(s) + \beta^T X + \delta(s)I(s > d), \tag{A.4}$$

where the temporal trend of Y is assumed to change after dropout time d from $\alpha(s)$ to $\alpha(s) + \delta(s)$. With a known $\delta(s)$, the conditional mean of $Y(s)$ is estimated as

$$\hat{E}[Y(s)|T = d, X] = \hat{\alpha}(s) + \hat{\beta}^T X + \delta(s)I(s > d),$$

where $\hat{\alpha}(\cdot)$ and $\hat{\beta}$ are the proposed MLEs. For each $\delta(s)$, then $\mu(s)$ can be obtained by

$$\tilde{\mu}(s) = \int_u \hat{E}[Y(s)|T = u, X]\, \hat{P}\{T = u|X\}\, du,$$

where $\hat{P}\{T = u|X\}$ is the estimate of the conditional density $E_b[P\{T = u|X, b\}]$ of drop-out time T given the covariates X. These inference procedures about $\mu(s)$ will be repeated for a user-specified set of sensitivity parameters. In our prostate cancer analysis, it is reasonable to assume a log-transformed PSA level linearly increases after drop out, i.e., $\delta(s) =$ constant $\times(s - d)$. In general, any user-specified sensitivity function $\delta(s)$ can be posited, which may depend on drop-out time d and covariates as well.

References

Adams, R. and Fournier, J. (1975). *Sobolev spaces.* New York: Academic Press.

Kim, S., Zeng, D., Chambless, L., and Li, Y. (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in Biosciences* **4,** 262–281.

National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials.* Panel on Handling Missing Data in Clinical Trials. Washington, DC: National Academies Press.

Schumaker, L. (2007). *Spline functions: basic theory.* Cambridge Univ. Press.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77,** 147–160.

van der Vaart, A. (1998). *Asymptotic statistics.* Cambridge Univ. Press.

van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes.* New York: Springer-Verlag.

Zeng, D. (2005). Likelihood approach for marginal proportional hazards regression in the presence of dependent censoring. *Annals of Statistics* **33,** 501–521.

Zeng, D. and Cai, J. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Annals of Statistics* **33,** 2132–2163.