Variable Selection of a Bayesian hierarchical changepoint Model for longitudinal biomarkers of Prostate Cancer

Wonsuk Yoo$^{(1)}$**, Elizabeth H. Slate**$^{(2)}$

$^{(1)}$ Department of Mathematical Sciences
New Jersey Institute of Technology
Newark, NJ 07102

$^{(2)}$ Department of Biostatistics, Bioinformatics and Epidemiology
Medical University of South Carolina
Charleston, SC 29425

CAMS Report 0405-35, Spring 2005

Center for Applied Mathematics and Statistics

NJIT

Variable Selection of a Bayesian hierarchical changepoint Model for longitudinal biomarkers of Prostate Cancer DRAFT

Wonsuk Yoo[1] and Elizabeth H. Slate[2]

1. Department of Mathematical Sciences, New Jersey Institute of Technology

2. Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina [*†]

Abstract

Prostate specific antigen (PSA) is a common biomarker used to aid detection of prostate cancer (PCa). This research aims to develop and implement models for longitudinal biomarkers for prostate cancer, and to use these models to develop a diagnostic rule for early detection. We generalize a fully Bayesian hierarchical changepoint model, similar to that proposed by Slate and Clark (1999), by incorporating risk factors for prostate cancer as covariates. The changepoint, which is specific to each individual, represents the age of PCa onset. Our model permits the covariates to affect an individual's PSA three ways: the overall level, the age at which cancer initiates (changepoint), and the growth rate following the changepoint. We use Markov chain Monte Carlo (MCMC) to estimate all model parameters, including, especially, the subject-specific changepoints. Using data obtained from the Nutritional Prevention of Cancer Trial (Clark et al, 1996), we investigate the effects of smoking status, alcohol consumption and body mass index (BMI) on PSA growth. Moreover, we consider whether PSA velocity varies with the stage of prostate cancer at diagnosis. We select the most useful combination of covariates in the model by examining the Bayesian credible intervals for the associated parameters, by computing conditional predictive ordinate (CPO) values (Gelfand et al., 1992) and pseudo Bayes factors. A retrospective receiver operating characteristic (ROC) curve method is applied to verify a potential best model.

1 Introduction

In medical and biological researches, it is very common to observe response measurements depending on time. We call a series of those time-dependent response longitudinal series of observations. A biomarker is a biological quantity that can be indicative of underlying disease status. A longitudinal biomarker can be monitored over time for changes that may be associated with changes in disease status. The cholesterol or hyper-pressure levels can be biomarkers for cardiovascular disease and the number of tumors or the size can be also promising biomarkers for skin cancer. This research focuses on developing Bayesian hierarchical models to assess prostate disease status. Currently, prostate specific antigen (PSA) is the most useful biomarker for prostate cancer. Several researchers have investigated the relationship between age and PSA level. This research will extend the Bayesian hierarchical changepoint model with several possible risk factors of prostate cancer such as smoking status, alcohol usage, BMI, and prostate cancer progressing stage etc. This generalization for Bayesian model is one of contributions of this research. There has been a fairly simple way for

[*] This manuscript is a draft and is being revised for publication

[†] E. Slate acknowledges DAMD Grant 17-02-1-0138, NIH Grant CA077789, NIH COBRE Grant RR17696, and NIH Grant 1 U24 DE016508. W. Yoo gratefully acknowledges the support from DAMD Grant 17-02-1-0138 awarded to E. Slate.

assessing prostate status that typically a single reading of prostate specific antigen is taken through the blood test and interpreted as a positive test if elevated (usually more than 4 ng/ml), or the annual rate of increase since the previous PSA reading may be computed and interpreted as yielding a positive result if large (greater than 0.75 ng/ml/yr). Since this research eventually have longitudinal PSA readings from the NPCT data, it will enable us to better understand the relationship between PSA and prostate cancer.

Consideration of any related covariate into the interested model has been a practically important problem in statistical application research. We expect that an additive model including covariates into the Byaesian hierarchical changepoint model can be more effective to detect prostate cancer status. There have been so many effective researches using Markov chain Monte Carlo technique in Bayesian variable selection criteria for a last ten years. In this paper, we focus on cross-validation approach, and receiver operation characteristic (ROC) curve approach.

2 A Generalization from Baysian Changepoint Model

A model used in this research can be called *Bayesian Hierarchical Changepoint Model*. This is a changepoint model which has a changeover point indicating an onset time of a disease under the assumption that every man should have a disease(changepoint) eventually. In other word, if he lives long enough, he get the disease during his lifetime. This model is also a *hierarchical* model since a Bayesian model is used with several different levels of the prior distribution.

There have been several researches that interested in relationship between PSA level and its diagnostic role, "what probability is there that a subject can have the prostate cancer". Similarly, our interest focuses on estimating the subject-specific parameters such as subjects' changepoint. A Bayesian model can be good alternative to figure out this problem since the method permits the "borrowing of strength" which can bring population information into the model. Even though our model is somewhat complicated with hierarchical structure for prior information, we can fit the changepoint easily using Markov chain Monte Carlo (MCMC) techniques. Thus, a fully Bayesian hierarchical changepoint model may be thought of as a Bayesian generalization of the segmented mixed effects model.

Longitudinal data are a series of successive observations (often biomarkers in the medical research) on each observational unit (often subjects in the medical research). The model can be used to describe the apparent differences in the PSA trajectories (growth patterns) among individuals who have and have not been diagnosed with disease (cases and controls). Slate and Clark (1999) proposed a linear mixed effects model which used $\ln(PSA+1)$, the transformed PSA on the log-scale which had used by Whittemore et al. (1995) and fit the above model to the NPCT data. All coefficients are estimated as statistically significant. The assumption that the cases were cancer-free at one time naturally leads to building a changepoint model in the PSA trajectories.

2.1 A Bayesian hierarchical changepoint model

We consider a fully Bayesian hierarchical changepoint model, similar to that proposed by Slate and Clark (1999). The changepoint, which is specific to each individual, represents the age of PCa onset. We call this the **Model 0** which can be viewed as a three-stage model with prior and hyper-prior distributions. Prior and hyper-prior information come from previous literature (Cronin et al. 1994, Slate and Clark 1999) and the Nutritional Prevention of Cancer Trial (Clark et al, 1996). Like these previous authors, and also Whittemore et al. 1995, we consider the log-transformed PSA level, $\ln(PSA+1)$, as the longitudinal response measurements, denoted by $Y_{ij}'s$. The model can be appropriate for any biological marker Y_{ij} which can be indicative of underlying disease status and have potential to improve early detection. The Bayesian hierarchical changepoint model with no covariate is

$$Y_{ij} = a_{0i} + a_{1i}x_{ij} + (b_i - a_{1i})(x_{ij} - t_i)^+ + \varepsilon_{ij} \tag{1}$$

where i indexes the subject and j indexes the reading within a subject. The superscript " $+$ " is an indicator function with $(x_{ij} - t_i)^+ = (x_{ij} - t_i)$ for $x_{ij} > t_i$ and zero otherwise. The variable x_{ij} denotes the age at the j^{th} visit of a subject i. The interpretation of all random effects for the trajectory for subject i is that a_{0i} is the intercept, a_{1i} is slope before the changepoint, the b_i is slope after the changepoint and t_i is the changepoint. Our model is a three-stage hierarchical changepoint model with prior and hyper-prior distribution given below:

$$\begin{pmatrix} a_{0i} \\ a_{1i} \end{pmatrix} | \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \sim \text{MVN}\left\{ \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}, \Omega_a \right\} \tag{2}$$

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \sim \text{MVN}\left\{ \begin{pmatrix} 1 \\ 0.02 \end{pmatrix}, \begin{pmatrix} 100 & 0 \\ 0 & 10000 \end{pmatrix} \right\}$$

$$\Omega_a \sim \textsc{Wishart}\left\{ \left[5 \begin{pmatrix} 0.1 & 0 \\ 0 & 0.0001 \end{pmatrix} \right]^{-1}, 5 \right\}$$

$$b_i | \beta, \tau_b \sim \text{N}(\beta, \tau_b)$$

$$\beta \sim \text{N}(0.15, 3600) \;\;\&\;\; \tau_b \sim \textsc{Gamma}(48, 0.0133)$$

$$t_i | \mu_t, \tau_t \sim \text{N}(\mu_t, \tau_t) \;\;\&\;\; \mu_t \sim \text{N}(80, 0.10) \;\;\&\;\; \tau_t \sim \textsc{Gamma}(47, 4700)$$

$$\varepsilon_{ij} | \tau_i \sim \text{N}(0, \tau_i) \;\;\&\;\; \tau_i \sim \textsc{Gamma}(5, 0.25)$$

Our model is a Bayesian hierarchical changepoint model because it is a changepoint model with a three-stage hierarchical structure and prior and hyper-prior distributions. All normal distributions are parameterized in terms of a mean and a precision; thus τ_i, τ_b, τ_t, σ_0, and σ_1 are all precisions. The Gamma(α, β) distribution has mean α/β where shape parameter $\alpha > 0$ and inverse scale $\beta > 0$. The prior information of the random effects for the trajectory for subject i, $(a_{0i}, a_{1i})'$, is assumed to follow the multivariate normal distribution with mean vector $(\alpha_0, \alpha_1)'$ and precision matrix Ω_a. The model indicates that the hyper-prior for mean vector $(\alpha_0, \alpha_1)'$ follows also multivariate normal distribution with known mean vector and precision matrix, and the precision matrix Ω_a is assumed to follow Wishart distribution. The Wishart distribution, a multivariate generalization of the gamma distribution, is the conjugate prior distribution for the precision matrix Ω_a which is symmetric and positive definite in a multivariate normal distribution. When we considered the random effects model with a unknown precision matrix Ω_a, it's natural to specify a Wishart distribution for Ω_a. Therefore, all prior and hyper-prior distributions are conjugate in the full model.

All prior information used in this research came from previous literatures including Carter et. al. (1992) and Whittemore et. al. (1995). Slate et al. (2000) constrained the slope after the changepoint to be greater than 0.08 in order for this restriction to facilitate the model's distinction between a_{1i} and b_{1i}. Here we do not constrain the slope. This model assumes that all men eventually reach their changepoint if they live long enough. Since a large fraction of men do not encounter their changepoint, the population mean of the changepoints, μ_t, tends to be quite large. Table 1 shows a model structure of **Model 0**. The stage 1 shows the segmented random effects model with one changepoint, with of prior information at the stage 2 and hyper-parameters in the stage 3. All prior and hyper-prior information have informative prior distributions which are priors not dominated by the likelihood, and has an impact on the posterior density. Gelman et al. (1995) indicates that experimenters can have approximate estimate of the population distribution using the historical(past)

1. first stage : $Y_{ij} = a_{0i} + a_{1i}x_{ij} + c_i\gamma + (b_i - a_{1i})(x_{ij} - t_i)^+ + \varepsilon_{ij}$
   where $\varepsilon_{ij}|\tau_i \sim \mathrm{N}(0, \tau_i)$

2. second stage : prior information for random effects parameters
   $$\begin{pmatrix} a_{0i} \\ a_{1i} \end{pmatrix} | \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \sim \mathrm{MVN}\left\{ \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}, \Omega_a \right\}$$
   $b_i|\beta, \tau_b \sim \mathrm{N}(\beta, \tau_b)$ and $t_i|\mu_t, \tau_t \sim \mathrm{N}(\mu_t, \tau_t)$

3. thirs stage : hyper-prior information for population parameters
   $$\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \sim \mathrm{MVN}\left\{ \begin{pmatrix} 1 \\ 0.02 \end{pmatrix}, \begin{pmatrix} 100 & 0 \\ 0 & 10000 \end{pmatrix} \right\}$$
   $$\Omega_a \sim \textsc{Wishart}\left\{ \left[5 \begin{pmatrix} 0.1 & 0 \\ 0 & 0.0001 \end{pmatrix} \right]^{-1}, 5 \right\}$$
   $\beta \sim \mathrm{N}(0.15, 3600)$ and $\tau_b \sim \mathrm{Gamma}(48, 0.0133)$
   $\mu_t \sim \mathrm{N}(80, 0.10)$ and $\tau_t \sim \mathrm{Gamma}(47, 4700)$
   $\gamma \sim \mathrm{N}(0, 100)$
   $\tau_i \sim \mathrm{Gamma}(5, 0.25)$

Table 1: The Model description of hierarchical structure in the Model 1

data. Considering historical data or papers, experimenters can expect that the interest parameter follows an approximate known distribution. We consider conjugate priors for those prior distributions in Model 0. The property that the posterior distribution follows the same parametric form as the prior distribution is called *conjugacy*; the beta prior distribution is a conjugate family for the binomial distribution. Thus, a *conjugate* prior for a family of distributions can be applied if the prior and posterior distribution are of the same family. The random effects in the Model 0 (1) has conjugate prior distributions of normal and gamma distributions at the prior second stage for mean and variance of the stage 1 normal densities respectively. The Wishart distribution is the conjugate prior distribution for the precision matrix Ω_a which is symmetric and positive definite in a multivariate normal distribution.

2.2 A generalized additive model

A changepoint model assumes that a subject will eventually experience a changepoint in his trajectory of response measurements. In other words, every man will eventually get the disease. We examine several possible risk factors that may affect PSA level directly or indirectly, through examination of the changepoint or the slope after the changepoint. The risk factors such as smoking status, drinking usages, BMI, and family history of specific disease are considered as covariates because they are most common risk factors of cancer, and can be detected by continuous or discrete scales. The covariates must be permitted to influence the dependent observation's trajectory both before and after the changepoint, and also to influence the timing of the changepoint. There are be three possible ways to consider covariates into the Bayesian model. We extend **Model 0**, a segmented linear regression model with fully Bayesian consideration by adding covariates which affect PSA directly to the model, adding effects to the change in slope after a changepoint, or considering an effect on the location/timing of the changepoint.

The **type I** will extend the segmented linear regression model (1) by adding the covariates into the prior to changepoint. The model (5) shows that c_i does not depend on the observation time j.

Thus, it is believed that the covariate affects the readings (response measurement levels) prior to the changepoint without affecting the changepoint itself or the post-changepoint slope:

$$Y_{ij} = a_{0i} + a_{1i}x_{ij} + c_i\gamma + (b_i - a_{1i})(x_{ij} - t_i)^+ + \varepsilon_{ij} \tag{3}$$

Since we do not know whether the covariates ultimately affect PSA levels, we use a flat prior for γ with zero mean and large variance under a normal distribution assumption. As an illustration, if smoking is believed to increase observed PSA levels in men both with and without a cancer, we can consider a smoking indicator as the fixed effect c_i in the model (5). Then, c_i has binary values: 1 if subject i is a smoker and 0 if not a smoker. However, the c_i can also be discrete variables or continuous variables. If a researcher believe that BMI level increases response measurement levels, we can consider the BMI levels as a continuous fixed effect c_i in the model. Furthermore, when we have more than two levels for discrete variable of interest, we can consider it an ordinal variable if we know the relationship between levels. But since we do not know the exact relationship between levels, we use dummy variables. If we have a variable with three levels, we use two dummy variables which are dichotomous:

$$Y_{ij} = a_{0i} + a_{1i}x_{ij} + c_{1i}\gamma_1 + c_{2i}\gamma_2 + (b_i - a_{1i})(x_{ij} - t_i)^+ + \varepsilon_{ij} \tag{4}$$

Model (6) shows the addition to the model of a variable with three levels. Assume that a covariate smoking status has three levels such as none, former smokers and current smokers, and we want to know the effects of smoking on PSA level. Since the smoking indicator has more than two levels, we should have multiple indicator fixed effect terms. Specially the smoking effect has three levels, as described above, so the term, c_{1i} should take the binary values 1 (for former smokers) and 0 (for everyone else), and the term c_{2i} should also take binary values of 1 for current smokers and 0 for everyone else. We consider a joint prior for $\gamma=\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix}$ with a zero mean vector and a variance-covariance matrix including large variances. Then, a sensitivity analysis is needed for examining the correlation between γ_1 and γ_2. It is our goal to obtain the Bayesian posterior estimate for the covariate effect γ, and we expect that the effect will be positive.

The **type II** model arises when the covariate is believed to affect the change in slope after a changepoint occurs. This is represented in the model shown in (7). We would use this type of model if a risk factor is believed to affect the rate of tumor growth once a tumor is present.

$$Y_{ij} = a_{0i} + a_{1i}x_{ij} + ((b_i + c_i\gamma) - a_{1i})(x_{ij} - t_i)^+ + \varepsilon_{ij} \tag{5}$$

For example, if the continuous variable BMI and the dichotomous variable smoking habit are believed to affect the slope after a changepoint, we can consider both variables in the model. Equation (8) indicates a generalized model with both continuous and discrete variables.

$$Y_{ij} = a_{0i} + a_{1i}x_{ij} + ((b_i + c_{1i}\gamma_1 + c_{1i}\gamma_1) - a_{1i})(x_{ij} - t_i)^+ + \varepsilon_{ij} \tag{6}$$

For this model, we also consider a vague normal prior for γ with zero mean and large variance; however, if we expect positive effects from those variables or have some indication of positive effects from previous research and papers, we can also use a gamma distribution. Furthermore, if we have any obvious correlation among different levels, we can construct the variance-covariance matrix.

Model (9) can be applied if the covariate is believed to affect the time when a changepoint occurs. This **type III** model is appropriate if it is believed that a risk factor actually affects when a changepoint occurs. This model is equivalent to shifting the mean of the prior distribution of t_i for the subset of the population with a value of $c_i \neq 0$.

$$Y_{ij} = a_{0i} + a_{1i}x_{ij} + (b_i - a_{1i})(x_{ij} - (t_i + c_i\gamma))^+ + \varepsilon_{ij} \tag{7}$$

We also consider a normal prior for a covariate effect γ with zero mean and large variance. Similarly if there are two covariates believed to affect a changepoint, the model is

$$Y_{ij} = a_{0i} + a_{1i}x_{ij} + (b_i - a_{1i})(x_{ij} - (t_i + c_{1i}\gamma_1 + c_{2i}\gamma_2))^+ + \varepsilon_{ij} \tag{8}$$

An additive model can contain any combination of covariates for all three addition types. Let's assume that a covariate c_1 of smoking status affects the slope after the changepoint, and that a covariate c_2 of body mass index is expected to affect the changepoint. Then we can consider an additive model:

$$Y_{ij} = a_{0i} + a_{1i}x_{ij} + ((b_i + c_{1i}\gamma_1) - a_{1i})(x_{ij} - (t_i + c_{2i}\gamma_2))^+ + \varepsilon_{ij} \tag{9}$$

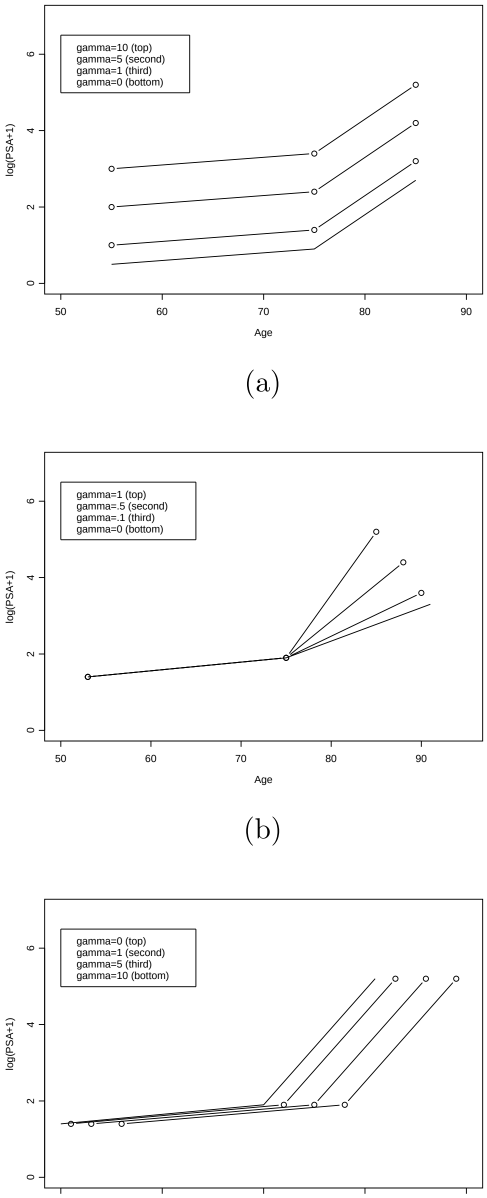
We can propose any combination of covariates with the three different types of model described in (5), (7) and (9).

Figure (6) shows how the γ's affect the full model when we give three different values for γ for each of the three types. Since type I is assumed for a covariate which only directly affects the response level, the first graph (a) shows that graphs depending on different γ values have different log-transformed PSA levels. As the level of γ increases, we see that the risk factor has a positive effect on the log-transformed PSA levels. The graph (b) for type II indicates that the different γ values lead to three different slopes after the changepoint. Under the type III assumption that a covariate only affects the changepoint, graph (c) shows variation of a changepoint. If a risk factor is significant to a changepoint, the changepoint moves horizontally.

3 Bayesian Cross-Validation Approach

Cross-validation methods have been well established in frequentist statistics. Typically, the data are split into two sets. The first set is used for model fitting, while the second set is used for model validation. Sometimes the second set has only one observation. This is one complete iteration in the process, and a new partitioned set is chosen and the process repeated hundreds or thousands of times. Finally, we can choose some criterion as a basis for model selection. The predictive distribution has long been used as the correct Bayesian approach to model determination. Box *et al.* (1980) mentioned the complementary roles of the posterior and predictive distributions, arguing that the posterior is used for ”estimation of parameters conditional on the adequacy of the model” while the predictive distribution is used for ”criticism of the entertained model in light of the current data”. Thus, in comparing several models, it is necessary to use the predictive, not the posterior distribution. The cross-validation approach involves predictions of a subset Y_i of data Y when only the complement of Y_i, denoted $Y_{(i)}$ is used to update the prior. Thus the cross-validation approach needs to find $p(y_i|y_{(i)})$, which is called the cross-validatory predictive distribution or conditional predictive density. The quantity $p(y_i|y_{(i)})$ is known as the conditional predictive ordinate (CPO). The conditional predictive density of $Y_i|y_{(i)}$ is

$$\begin{aligned} p(y_i|y_{(i)}) &= \int p(y_i|\theta, y_{(i)})\, p(\theta|y_{(i)})d\theta. \\ &= \int p(y_i|\theta)\, p(\theta|y_{(i)})d\theta \end{aligned} \tag{10}$$



(a)

(b)

(c)

Figure 1: How the covariate affects the full model depending on three values of γ (a) of type I, (b) of type II, (c) of type III .

where $p(\theta|y_{(i)})$ is the posterior distribution of parameters θ given the data $y_{(i)}$ under the prior $p(\theta)$. Because of conditional independence between y_i and $y_{(i)}$, we can disregard the variable $y_{(i)}$. This approach is consistent with the predictive purposes to which the chosen model is often put and has the advantage of remaining feasible where the posterior distributions $p(\theta_i|Y)$ are proper but the prior distributions $p(\theta_i)$ are not. How can we compute the CPO for each y_i? We can directly compute of CPO when using the posterior density omitting y_i:

$$\begin{aligned} p(y_i|y_{(i)}) &= \int p(y_i|\theta)\, p(\theta|y_{(i)})d\theta \\ &= \mathrm{E}_{\theta|y_{(i)}}\left[p(y_i|\theta)\right] \end{aligned} \tag{11}$$

In order to get the estimates of (22), we can use the MCMC technique on $p(\theta|y_{(i)})$ and calculate

$$\bar{p}(y_i|y_{(i)}) = \frac{1}{N-M}\sum_{t=M+1}^{N} p(y_i|\theta^{(t)}) \tag{12}$$

However this direct method often finds it difficult to estimate the above expectation, and needs a separate MCMC running for each y_i. Gelfand, Dey and Chang (1992) and Dey, Chen and Chang (1997) showed a method using Monte Carlo integration to estimate the conditional predictive ordinate(CPO). They observed that

$$\begin{aligned} p(y_i|y_{(i)}) &= \frac{p(y)}{p(y_{(i)})} \\ &= \frac{\int p(y|\theta)\pi(\theta)d\theta}{\int p(y_{(i)}|\theta)\pi(\theta)d\theta} \\ &= \frac{\int \frac{p(y|\theta)\pi(\theta)}{p(y)}\frac{p(\theta|y)}{p(\theta|y)}d\theta}{\int \frac{p(y_{(i)}|\theta)\pi(\theta)}{p(y)}\frac{p(\theta|y)}{p(\theta|y)}d\theta} \\ &= \left\{\int \frac{p(y_{(i)},\theta)}{p(y,\theta)}\pi(\theta|y)d\theta\right\}^{-1} \\ &= \left\{\int \frac{1}{p(y_i|y_{(i)},\theta)}p(\theta|y)d\theta\right\}^{-1} \\ &= \left\{\int \frac{1}{p(y_i|\theta)}p(\theta|y)d\theta\right\}^{-1} \end{aligned} \tag{13}$$

where $y = (y_i, y_{(i)})$ and $\frac{p(y,\theta)}{p(y_{(i)},\theta}= p(y_i|y_{(i)},\theta)$. Then a Monte Carlo integration of CPO_i is given by

$$\begin{aligned} \widehat{CPO}_i &= \left\{E_{\theta|y}\left[\frac{1}{p(y_i|\theta)}\right]\right\}^{-1} \\ &= \left\{\frac{1}{N-M}\sum_{t=M+1}^{N}\frac{1}{p(\mathbf{y}_i|\theta_i^{(t)})}\right\}^{-1} \end{aligned} \tag{14}$$

where N is the total number of iterations, and M is the number of burn-in samples, which are discarded prior to computation of $\widehat{CPO}_i$. Thus the Monte Carlo estimate for the CPO leads to the harmonic mean of the conditional density of y_i given $\theta^{(t)}$, which is determined by the posterior samples. We use the Gibbs sampler to estimate $p(y_i|\theta^{(t)})$. The Bayesian cross-validation approach uses the predictive distribution to assess model adequacy. The value of CPO$_i$ indicates how much the i^{th} observation supports the model, with a large CPO value indicating a good fit to the model.

We now construct a cross-validatory predictive density for longitudinal series of data. Consider a model which includes three additive covariates: a covariate affecting PSA level directly, a slope-effect covariate and a changepoint-effect covariate.

$$Y_{ij} = a_{0i} + a_{1i}x_{ij} + c_{1i}\gamma_1 + \{(b_i + c_{2i}\gamma_2) - a_{1i}\}\{x_{ij} - (t_i + c_{3i}\gamma_3)\}^+ + \varepsilon_{ij} \tag{15}$$

Assume that we want to calculate CPO_i to compare with the no-covariate model. Suppose we have n_i observations on the i^{th} unit for $i = 1, \ldots, I$, and that $N = \sum_{i=1}^{I} n_i$ is the total number of observations. We are interested in the distribution of observed data at the i^{th} unit, given subject-specific and population parameters. We denote θ_i as the complete set of related parameters:

$$\theta_i = (\gamma, a_{0i}, a_{1i}, b_i, t_i, \tau_i, \alpha_0, \alpha_1, \Sigma_a, \beta, \tau_b, \mu_t, \tau_t, \gamma_2, \gamma_3). \tag{16}$$

We know the distribution of y_i:

$$y_i|\gamma, a_{0i}, a_{1i}, b_i, t_i, \tau_i \ \sim \ \mathrm{N}_{n_i}\left(\mu_i, \tau_i I_{n_i}\right) \tag{17}$$

where $\mu_i = a_{0i} + a_{1i}x_{ij} + c_{1i}\gamma_1 + \{(b_i + c_{2i}\gamma_2) - a_{1i}\}\{x_{ij} - (t_i + c_{3i}\gamma_3)\}^+$ and $\tau_i = 1/\sigma_i^2$. Our purpose is to find the marginal predictive density for subject i,

$$\begin{aligned} p(\mathbf{y}_i|\theta_i) &= p[y_{i1}, \ldots, y_{in_i}|\theta_i] \\ &= \prod_{j=1}^{n_i} p(y_{ij}|\theta_i) \\ &= \prod_{j=1}^{n_i} \frac{\tau_i}{\sqrt{2\pi}} \exp\left[-\frac{\tau_i}{2}(y_{ij} - \mu_{ij})^2\right] \\ &= \left(\frac{\tau_i}{\sqrt{2\pi}}\right)^{n_i} \exp\left[-\sum_{j=1}^{n_i} \frac{\tau_i}{2}(y_{ij} - \mu_{ij})^2\right] \end{aligned} \tag{18}$$

where $\mu_{ij} = a_{0i} + a_{1i}x_{ij} + c_{1i}\gamma_1 + \{(b_i + c_{2i}\gamma_2) - a_{1i}\}\{x_{ij} - (t_i + c_{3i}\gamma_3)\}^+$. We obtain samples from MCMC integration, and apply the *ergodic* theorem to get the i^{th} subject's conditional predictive ordinate, $\widehat{CPO}_i$:

$$\widehat{CPO}_i = \left\{\frac{1}{N - M} \sum_{t=M+1}^{N} \frac{1}{p(\mathbf{y}_i|\theta_i^{(t)})}\right\}^{-1} \tag{19}$$

where $\mathbf{y}_i = (y_{ij}, \ldots, y_{in_i})$. How can we use estimated CPO values to choose the most appropriate model? The concept of a pseudo Bayes factor can be used. The product of cross-validation predictive densities, $\prod_{i=1}^{n} p(y_i|y_{(i)})$, which is called the pseudo-predictive likelihood, can lead to the *pseudo-Bayes factor*, the ratio of pseudo-predictive likelihoods. If we denote pseudo-BF$_{12}$ to be pseudo-Bayes factor of Model 1 against Model 2, we have

$$\text{pseudo-BF}_{12} = \frac{\prod p(y_i|y_{(i)}, M_1)}{\prod p(y_i|y_{(i)}, M_2)} \tag{20}$$

With pseudo-Bayes factors for several candidate models, we can apply Jefferys' (1961) scale of evidence. For convergence of calculation, we use the log-scale for the pseudo-Bayes factor to choose a best-fitted model.

Decision Rule :
Choose Model 1 against Model 2 if $\log\left\{\frac{\prod p(y_i|y_{(i)},M_1)}{\prod p(y_i|y_{(i)},M_2)}\right\} > 0$

Dey, Chen & Chang (1997) and Sinha, Chen & Ghosh (1999) showed a Bayesian model selection procedure with several tools for determining a better model, such as the CPO plot and deviance plot. We will use the log-ratio of CPO plot for all cases of both models for model determination, as shown in Sinha, *et al.* (1999). The observation with a larger CPO value under one model will support that model and criticize the other. Thus, a plot of CPOs under both models against observation number should reveal that the better model has the majority of its CPOs above those of the worse one. The sampling based implementation enables straightforward investigation of model choice as well as model fitting.

4 Retrospective ROC verification

There are largely two types of diagnostic rules for detecting prostate cancer, the threshold method and the posterior probability method. The former rule use one single observation while the latter method depends on longitudinal series of biomarkers. Diagnostic rules are commonly evaluated using the concepts of sensitivity and specificity. Sensitivity is the probability that the rule is positive when prostate cancer is present, and specificity is the probability that the rule is negative when prostate cancer is absent. We want to use a receiver-operating characteristic (ROC) curve to compare the two models, the no-covariate model and the best combination of covariates model. A receiver-operating characteristic (ROC) curve is generated by plotting sensitivity versus 1-specificity for the rule R(h) as h varies throughout its range. The ROC curve is a useful summary of the performance of the rule and the point on the curve nearest (0,1) gives an optimal value for h.

When evaluating the rule R derived from repeated PSA measurements, it is apparent that sensitivity is time-dependent. That is, the nearer to prostate cancer onset that the PSA series extends, the greater the expected sensitivity of the rule. Thus sensitivity is a function of not only the cutoff h, but also the timing of the available PSA. This leads to the need for a generalization of sensitivity and specificity to the context of a longitudinal series of tests. For example, evaluation of the rule R leads to not merely an ROC curve, but an ROC surface with third dimension given by the proximity of the last PSA reading to prostate onset. Slate and Cronin (1997) and Slate and Clark (1999) has proposed one such generalization, and this approach will be adapted to each of the rules. Comparison of the rules will be performed according to two criteria: first, for fixed proximity of the PSA series to prostate cancer diagnosis (i.e. fixed slice through the ROC surface), the ROC curves will be compared; second, the proximity required to achieve specified sensitivity and specificity will be compared (the greater the lead time provided, the better the rule).

The diagnostic rule depending on the threshold method is very simple. All we should do is to compare a cut-off value of PSA with the PSA level of the most recent reading in a subject. If the latter exceeds a cut-off value, we declare ”positive” to a disease. With the threshold rule, we can conclude that the subject has a positive result for prostate cancer when he has the j^{th} visit. The threshold gives following diagnostic rule: The posterior probability rule which depends on longitu-

dinal measurements needs to calculate the probability that a change-point of a subject is less than an age at last reading for a pre-scribed cut-off probability. Thus, the most important thing is to choose a cut-off value of probability. Based on the posterior probability rule, we declare a positive result if the calculated probability is greater than a cut-off value and conclude that the subject has a positive

Threshold Diagnostic Rule:
A positive result if the most recent PSA reading exceeds a cut-off ($4ng/ml$).

Table 2: A diagnostic rule depending on the traditional threshold method

result for prostate cancer when he has the j'^{th} visit. The threshold gives following diagnostic rule:
The performance of this rule can be compared to that of other diagnostic rules by examining the

Posterior Probability Diagnostic Rule:
A positive result if $Pr(t_i \leq AGE_{MRR}) \geq h$, where t is a changepoint of the i^{th} subject, h is a cut-off, and MRR stands for ”at most recent PSA reading”.

Table 3: A diagnostic rule depending on the posterior probability method

receiver operating characteristic (ROC) curves. ROC curves plot sensitivity versus (1-specificity) as the cutoff for the given criterion varies. Sensitivity is defined as the proportion of diseased subjects that test positive, and specificity as the proportion of non-diseased subjects that test negative. Since one subject may yield a false positive test result at one visit and a true positive test result at a later visit with a series of longitudinal data, these definitions must be extended from the usual notions. Murtagh et al. (1991, 1995) have discussed longitudinal ROC curves, as have Slate and Cronin (1997).

We consider only retrospective diagnostic using the full series of PSA readings available. Hence, each subject has one test result based on his full series, which contribute to sensitivity or specificity according to whether the subject is a case or a control, respectively. Since diagnostic rules depend on cut-off values, our ROC curves also calculated depending on different cut-off values. The above box briefly shows how to calculate ROC curves. We divide the data (932 subjects including 60 cases and 872 controls) into two groups: the first 500 subjects for estimation and the other 432 subjects for prediction. To estimate the specificity, we only use the control subjects. The test result for a subject has a binary value and can be calculated in different ways for two diagnostic rules. When the probability is greater than a cut-off value, we have a negative value 1 for specificity while we have a positive value 1 for sensitivity. For an illustration, consider calculation of sensitivity based on posterior probability. We prescribe cut-off probabilities from zero to one with an interval of 0.1 or 0.05. Then for a evaluation subject with a case, we have binary values (zero or one) at each cut-off value. With 500 subject data, we do this for all 60 case subjects, and calculate sensitivities at each cut-off value,

$$\text{sensitivity} = \frac{\text{the number of positive tests}}{\text{the total number of cases}}$$

Similarly we calculate specificities at each cut-off values of the threshold method for all 372 controls whether PSA level at last reading is greater than prescribed cut-off values or not.

$$\text{specificity} = \frac{\text{the number of negative tests}}{\text{the total number of controls}}$$

The Table 7 briefly shows how to calculate sensitivities and specificities at each cut-off h to use ROC curves. for only cases

1. $Specificity = \frac{the\ \ number\ \ of\ \ negative\ \ results}{the\ \ number\ \ of\ \ controls}$,
For the posterior probability method,
If $Pr(t_i \leq AGE_{ALR}) \geq h$, positive result = 0; 1 for else, i for control subjects and AGE_{ALR} means an age at last reading. $h's$ are cut-off probabilities.
For the threshold method,
If [most recent PSA reading at each control subject] $> h$, negative result = 0; 1 for else, and where $h's$ are a cut-off PSA levels.
2. $Sensitivity = \frac{the\ \ number\ \ of\ \ positive\ \ tests}{the\ \ number\ \ of\ \ cases}$,
For the posterior probability method,
If $Pr(t_i \leq AGE_{ALR}) \geq h$, positive result = 1; 0 for else, i for case subjects and AGE_{ALR} means an age at last reading. $h's$ are cut-off probabilities.
For the threshold method,
If (most recent PSA reading at each case subject) $> h$, positive result = 1; 0 for else, and where $h's$ are cut-off PSA levels.

Table 4: Calculation of sensitivity and specificity depending on diagnostic rules of threshold method and posterior probability method

5 Application To NPC Trial Data

Our model permits the covariates to affect an individual's PSA three ways: the overall level, the age at which cancer initiates (changepoint), and the growth rate following the changepoint. We use Markov chain Monte Carlo (MCMC) to estimate all model parameters, including, especially, the subject-specific changepoints. We use WinBUGS (Bayesian Inference Using Gibbs Sampler) for posterior estimation. Data is obtained from the Nutritional Prevention of Cancer Trial (Clark et al, 1996), and we investigate the effects of smoking status, alcohol consumption and body mass index (BMI) on PSA growth. Moreover, we consider whether PSA velocity varies with the stage of prostate cancer at diagnosis. We select the most useful combination of covariates in the model by examining Bayesian critical intervals for the associated parameters. We compute conditional predictive ordinate (CPO) values (Gelfand et al., 1992) and pseudo-Bayes factors, and interpret according to Jeffreys' guideline. We provide analytical and graphical evidence for our selection of covariates.

5.1 Data

We have longitudinal PSA data from the Nutritional Prevention of Cancer Trials (NPCT). These trials, initiated in 1983, were designed to investigate whether a nutritional supplement of selenium would reduce skin cancer. The data including baseline information have 932 subjects with 5245 readings for different time points. Among total 932 subjects, 61 subjects with 348 total readings are cases while 871 subjects with 4897 observations are controls. The mean values of PSA and age at baseline are 2.35 ng/ml and 68.05 yrs. Among controls mean values of age at baseline and PSA level are 67.98 and 1.899 respectively, while those of cases are 69.02 and 8.01 respectively. The cases can be divided into two groups: localized and advanced stages.

5.2 Risk factors of prostate cancer

Cancer is a group of diseases characterized by uncontrolled growth and spread of abnormal cells. It can lead to death with uncontrollable spread of abnormal cells. Cancer is caused by both external factors such as tobacco, chemicals, radiation, and infectious organisms and internal factors including inherited mutations, hormones, immune conditions. Causal factor, called risk factors, may act together or in sequence to initiate or promote carcinogenesis. A *risk factor* is anything that increases the chance of developing a disease such as cancer. It can be any kind of external or internal conditions including food habit, life-style, or health condition which can lead higher chance to develop a specific disease.

Even if having a risk factor, or even several, does not mean that you will get the disease, a risk factor has chance to develop a disease such as cancer. Those who have one or more risk factors never develop the disease while other with this disease have no known risk factors. Nonetheless, it is important that we know about risk factors that we can try to change any unhealthy lifestyle behaviors or can choose to have the early detection tests for a potential cancer. There have been many researches to figure out risk factors that increase the risk of developing prostate cancer. Like most diseases, prostate cancer has the most distinct risk factor, age. The incidence of prostate cancer increases with age and more than 70 % of all prostate cancers are diagnosed in men over age 65. Another important risk factor may be race. Black Americans have the highest prostate cancer incidence rates in the world.

American Cancer Society (ACS) considers several risk factors for prostate cancer such as age, race, diet, family history, physical inactivity and others. Age is one of most common risk factors to increase the risk of developing cancers. Those who are over age 50 have more rapidly increased chance to have prostate cancer. Nevertheless more than 70 % of all prostate cancers are diagnosed in men over the age of 65. Prostate cancer occurs almost 70 % more often in African-American men as it does in white American men. African-American men are twice as likely to die of prostate cancer as white men and are more likely to be diagnosed at an advanced stage. Unfortunately there are no known reasons for these racial differences. American Cancer Society informs that a diet can be a risk factor of the prostate cancer. Men who eat a lot of red meat or who have a lot of high-fat dairy products in their diet appear to have a greater chance of developing prostate cancer. Those who have a father or brother with prostate cancer doubles a man's risk of developing the disease. Scientists have identified several inherited genes that seem to increase prostate cancer risk. Thus there seems obvious cause-effect relationship between family history of prostate cancer and disease developing chance.

Smoking, alcohol consumption and body mass index have been investigated as potential risk factors in most cancer researches. It would be substantially important to public health to have any evidence that even a moderate level of smoking or alcohol consumption can increase relative risk(RR) since there might be high incidence between prostate cancer and potential risk factors such as cigarette smoking, use of alcohol, body mass index(BMI). Hiatt, R.A. *et.al* investigated relation between risk factors such as smoking, drinking and BMI and prostate cancer and concluded that there exists the positive relation between prostate cancer and heavy smoking since cigarette smokers had an elevated relative risk of prostate cancer of $1.2(\mathrm{CI} = 1.1 \sim 1.3)$ when compared with that of nonsmokers. It's interesting to see that other cohort studies withy fewer cases did not find any increased prostate cancer risk attributable to smoking. Their conclusion supported a recent finding reported by Hsing *et al*(1991). They could not detect any statistical significance for increased risk of prostate cancer associated with alcohol consumption either for men who reported having ever drank, or for the heaviest consumers when compared with those who never drank. They found no relation between height, weight, or body mass index and prostate cancer risk. These findings are similar to those of other studies.

Demark-Wahnefried *et al* suggested that upper-body obesity, measured by ratio of trunk height to total height, may be the critical factor in prostate cancer development, as it is for other hormonerelated forms of cancer such as breast and endometrial cancer. Marital status, sexually transmitted disease history, and religion also have not been associated consistently with prostate cancer. Several epidemiologic studies of risk factors for benign prostatic hyperplasia (BPH) including Gann, PH *et al*(1995) and Platz EA *et al*(1999) have shown an inverse association of BPH with alcohol consumption and smoking. Josept, M.A. *et.al* investigated relationship of Serum Sex-Stroid Hormones and prostate volume in African-American men and reported that BMI was positively corrected with increasing levels of hormones after age adjustment. In this research, we consider four risk factors such as smoking status, drinking habit, BMI and prostate cancer stage as covariates of the model.

5.3 What are our covariates?

The NPCT data has baseline information for risk factors of prostate cancer: smoking habit, alcohol consumption, body mass index. Several investigations have been carried out to examine the relationship between prostate cancer and common risk factors such as smoking, drinking and body mass index. We have another important variable describing prostate cancer stage for case subjects. Smoking is one of the most common risk factors of cancer. Of interest are smoking history, current smoking status, and how many cigarettes are smoked per day. We design three covariates for investigating the relationship between smoking and PSA level.

A covariate **smkever** is used to examine a risk from smoking history, under the assumption that those who have ever smoked may have higher PSA level than those who never smoked. The **smkhvy** variable is designed to reveal the relationship between smoking amount and PSA level. We denote a subject as a heavy smoker if he currently smokes more than twenty cigarettes a day. The **smkstat** variable has three different levels of smoking status: never-smoker, former-smoker, and current-smoker. We treat **smkstat** as a categorical variable, not as an ordinal variable, because we do not know the relationship between former and current smoker. Thus, we create two dummy variables for **smkstat**. We expect positive effects for heavy smokers and ever smokers.

$$\text{smkever} = \begin{cases} 1 & \text{if subject has ever smoked} \\ 0 & \text{else} \end{cases}$$

$$\text{smkstat1} = \begin{cases} 1 & \text{if subject is a former smoker} \\ 0 & \text{else} \end{cases}$$

$$\text{smkstat2} = \begin{cases} 1 & \text{if subject is a current smoker} \\ 0 & \text{else} \end{cases}$$

$$\text{smkhvy} = \begin{cases} 1 & \text{if subject currently smokes more than 20 cigarettes per day} \\ 0 & \text{else} \end{cases}$$

We investigate the baseline PSA and log-transformed PSA levels for smoking related covariates. We observe that there is a slightly higher PSA level for never-smokers and not-heavy-smokers than for ever-smokers and heavy-smokers. The former smokers have the highest PSA level among all the groups.

Body mass index (BMI) is a potential risk factor for most types of cancer, including prostate cancer. BMI is calculated through a formula

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{height}^2 (m^2)}$$

Risk Factors & Covariates	Variable Type	Description
Body Measure Index (BMI)		
bmi	continuous	bmi = kg/m^2
bmi30	dichotomous	bmi30 = 1 if bmi $\geq$ 30
Drinking Consumption		
drkday1	continuous	cups/day
drkever	dichotomous	drkever = 1 if drkday1 $\geq$ 2
Smoking Habit		
smkever	dichotomous	smkever = 1 for ever-smokers
smkstat1	dichotomous	smkstat1 = 1 for former smokers
smkstat2	dichotomous	smkstat2 = 1 for current smokers
smkhvy	dichotomous	smkhvy =1 if cigaday $\geq$ 20

Table 5: Types of Covariates for all data of the NPCT data.

The covariate **bmi** is a continuous variable while **bmi30** is a dichotomous variable defined as:

$$\text{bmi30} = \begin{cases} 1 & \text{if BMI} > 30\ kg/m^2 \\ 0 & \text{else} \end{cases}$$

Alcohol consumption can be a important potential risk factor for prostate cancer. NPCT data has a variable of "drkday" which indicates drink consumption per day, categorized into a value from 0 to 7. We consider the variable as a continuous variable. From this variable, we also consider a covariate **drkever**. The covariate is a dichotomous variable of ever-drinkers and never-drinkers. Since he/she may reply no-drink even though he has one cup of drink a day, we denote ever-drinker as those who have at least two cups of drink a day.

$$\text{drkever} = \begin{cases} 1 & \text{for those who have more than one cup of drink a day} \\ 0 & \text{else} \end{cases}$$

The Table 14 shows a list of covariates and variable types used in this research for all subjects' data. In this research, we consider three risk factors of smoking and drinking habits, and BMI, for a total of eight variables.

5.4 Bayesian posterior estimation

We want to investigate whether or not each covariate is significant when it is added into a generalized Bayesian hierarchical changepoint model. This is an univariate testing hypothesis within full Bayesian framework. Consider a covariate which has multiple levels with three different categories. Since the covariate is a categorical variable, we have several dummy variables which has a dichotomous level. Now we denote the dummy variables c_1 and c_2. We aim to decide which covariates should be included into the generalized Bayesian model. Therefore, we consider all three types for each covariate, and investigate critical regions for the covariates.

Model Robustness : When there are covariates with more than two levels such as "smkstat" and we want to use several dummy variables for the covaruate, it is necessary to investigate whether there may be any relationship among coefficients of related dummy variables. Assume that we want to test a hypothesis that each level from a covariate "smkstat" has a significant effect to slope after a changepoint.

$$\begin{array}{ll} H_0: & \text{a reduced model } (c_{1i}{=}c_{2i}{=}0) \\ & Y_{ij} = a_{0i} + a_{1i}x_{ij} + (b_i - a_{1i})(x_{ij} - t_i)^+ + \varepsilon_{ij} \\ H_1: & \text{a generalized model (At least one } c_{ki} \neq 0, \text{ k=1,2)} \\ & Y_{ij} = a_{0i} + a_{1i}x_{ij} + ((b_i + c_{1i}\gamma_1 + c_{2i}\gamma_2) - a_{1i})(x_{ij} - t_i)^+ + \varepsilon_{ij} \end{array}$$

Before testing a hypothesis, first of all, we want to examine correlation of two dummy variables's coefficients γ_1 and γ_2 of a covariate "**smkstat**" according to correlation between the two. We have same prior and hyper-prior information same as **Model 1** except prior for coefficients of newly added covariate. We consider a conjugate prior distribution for γ_1 and γ_2. We assume multivariate normal distribution with mean zero vector and flat variance with some covariance information between the two γ's.

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \sim \text{MVN}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 100 & \sigma_{\gamma_1\gamma_2} \\ \sigma_{\gamma_1\gamma_2} & 100 \end{pmatrix} \right\}$$

Sensitivity analysis is needed to investigate whether or not the model is robust depending on variation of correlation coefficient ρ of variance-covariance matrix of γ's. We examine five different cases of ρ's: ρ = 0, 0.5, 0.9 and -0.5. For ρ=0, we believe that former smokers(γ_1) does not have correlation with current smokers(γ_2) while it is assumed that those who had been smokers has negative correlation with current smoking for ρ=-0.5. Below table is a summary statistics. Table 9 shows that the critical regions for type I and II are exactly same and those of type III is not little varied. That means that the correlation coefficients ρ is non-sensitive, and we can conclude that the covariance structure is robust to the model. Hence, we have an independence assumption between coefficients γ_1 and γ_2 of a covariate **smkstat**.

Smoking Status		Critical Regions		
Corr. Coef.	γ's	**Type I**	**Type II**	**Type III**
$\rho = 0$	γ_1	(-0.1084, 0.0195)	(-0.0568, 0.0267)	(-3.150, 2.505)
	γ_2	(-0.1427, 0.0122)	(-0.0565, 0.0123)	(-1.575, 5.608)
$\rho = 0.5$	γ_1	(-0.1084, 0.0195)	(-0.0568, 0.0267)	(-2.128, 2.543)
	γ_2	(-0.1427, 0.0122)	(-0.0565, 0.0123)	(-1.026, 6.057)
$\rho = 0.9$	γ_1	(-0.1084, 0.0195)	(-0.0568, 0.0267)	(-3.129, 3.881)
	γ_2	(-0.1427, 0.0122)	(-0.0565, 0.0123)	(-1.703, 6.037)
$\rho = -0.5$	γ_1	(-0.1084, 0.0195)	(-0.0568, 0.0267)	(-2.783, 4.049)
	γ_2	(-0.1427, 0.0122)	(-0.0565, 0.0123)	(-1.005, 5.789)

Table 6: Summary table of Critical Regions of coefficient γ's according to different correlation coefficient ρ's: 0, 0.5, 0.9 and -0.5

Univariate Analysis for Covariates : We have an univariate analysis for each covariate: **smkstat**, **smkever**, **smkhvy**, **bmi**, **bmi30**, **drkever** and **drkday1** as earlier mentioned. This is an important step because we want to obtain covariates which affect significantly to the log PSA level. Table 10 indicates that covariates of bmi30 and smkstat are not significant at all but a covariate "bmi" might have a change-point significant effect, slope-effect for "smkever" and changepoint-effect for "smkhvy", and changepoint-effect for "drkever" and slope-effect for "drkday1". Since covariates "smkever" and "smkhvy" comes from same risk factor "smoking status" like "drkever" and "drkday1", we want to have only one covariate from each risk factor to avoid multicollinearity problem. For a risk factor "SMOKING", there might be two potential significant covariates, smkever-slope

Covariates	Critical Regions		
	Type I	**Type II**	**Type III**
SMOKING			
smkstat 1	(-0.1084, 0.0195)	(-0.0568, 0.0267)	(-3.150, 2.505)
smkstat 2	(-0.1427, 0.0122)	(-0.0565, 0.0123)	(-1.575, 5.608)
smkever	(-0.0983, 0.0115)	(-0.0545, 0.0209)	(-1.306, 3.925)
smkhvy	(-0.0398, 0.0708)	(-0.0208, 0.0036)	(-3.937, 0.271)
Body Measure Index			
bmi	(-0.1171, 0.0033)	(-0.0016, 0.0012)	(-0.295, -0.127)
bmi30	(-0.0398, 0.0708)	(-0.0255, 0.0250)	(-2.099, 7.319)
DRINKING			
drkday1	(-0.0136, 0.0104)	(-0.0065, 4.7E-4)	(-0.2563, 1.058)
drkever	(-0.0707, 0.0509)	(-0.0353, -0.0011)	(0.3827, 5.883)

Table 7: Summary table of Critical Regions for coefficient γ's of each covariate

and smkhvy-changepoint effects. For ”DRINKING”, a covariate ”drkever” may has possible effects on both slope and changepoint effects.

5.5 Variable selection procedure

In order to choose only one covariate from same risk factor, we use cross-validation approach in Bayesian framework. We calculate CPO's for all 932 subjects from competing two models for ”SMOKING”, a model with a covariate ”SMKEVER” as a slope-effect and with ”SMKEVER” as a changepoint-effect as well as for ”DRINKING”, a model with a covariate ”DRKEVER”-slope effect and ”DRKEVER”-changepoint model. Consider ”SMKEVER” covariate to determine which effect is more significant to the model from the two, slope-effect and changepoint-effect. Let CPO$_{iM_{s1}}$ denote conditional predictive ordinate of i^{th} subjects from a model with a covariate ”SMKEVER” as a slope effect and CPO$_{iM_{s2}}$ for a model with a ”SMKEVER” as a changepoint effect. Then the conditional predictive ordinate for a model with slope-effect ”SMKEVER” is

$$\widehat{\text{CPO}}_{i,M_{s1}} = \left\{ \frac{1}{N-M} \sum_{t=M+1}^{N} \frac{1}{\prod_{j=1}^{n_j} p(y_{ij}|\theta^{(t)}, M_{s1})} \right\}^{-1}$$

where

$$p(y_{ij}|\theta^{(t)}) = a_{0i}^{(t)} + a_{1i}^{(t)} x_{ij} + ((b_i + c_{1i}\gamma_1^{(t)}) - a_{1i}^{(t)})(x_{ij} - t_i^{(t)})^+ + \varepsilon_{ij}$$

and

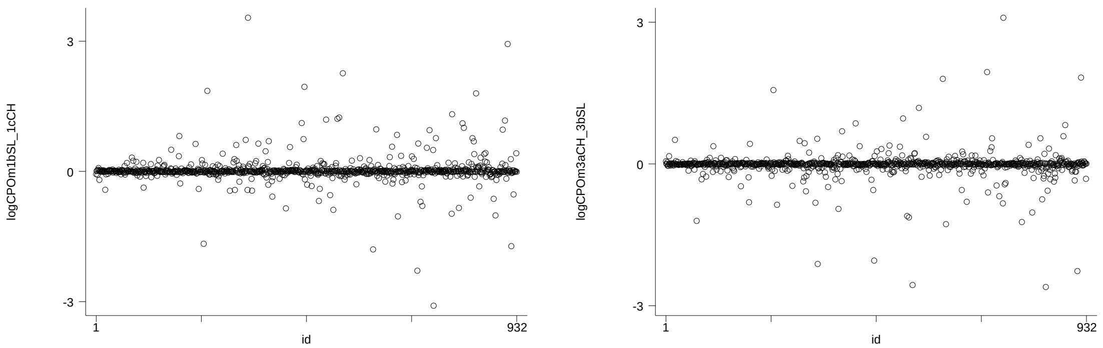
$$c_i = \begin{cases} 1 & \text{if he has ever smoked} \\ 0 & \text{else} \end{cases}$$

The conditional predictive ordinate for a model with changepoint-effect ”SMKEVER” is

$$\widehat{\text{CPO}}_{i,M_{s2}} = \left\{ \frac{1}{N-M} \sum_{t=M+1}^{N} \frac{1}{\prod_{j=1}^{n_j} p(y_{ij}|\theta^{(t)}, M_{s2})} \right\}^{-1}$$

where

$$p(y_{ij}|\theta^{(t)}) = a_{0i}^{(t)} + a_{1i}^{(t)} x_{ij} + (b_i - a_{1i}^{(t)})(x_{ij} - (t_i^{(t)} + c_{2i}\gamma_1^{(t)}))^+ + \varepsilon_{ij}$$



(a) (b)

Figure 2: CPO Plots for (a) Model with SMKEVER-slope vs with SMKHVY-chgpt (b) Model with DRKEVER-slope vs with DRKEVER-chgpt

t is a current iteration, N is total number of MCMC iterations with M burn-in iterations. We have 20,000 total iterations with 5000 burn-in. Since each iteration has one value from a joint density of

$$\begin{aligned} p(\mathbf{y}_i|\theta^{(t)}, M_k) &= p(y_{i1}, \ldots, y_{in_i}|\alpha_0^{(t)}, \alpha_1^{(t)}, \beta^{(t)}, \ldots, \tau_T^{(t)}, t_i^{(t)}, M_k) \\ &= \prod_{j=1}^{n_i} p(y_{ij}|\alpha_0^{(t)}, \alpha_1^{(t)}, \beta^{(t)}, \ldots, \tau_T^{(t)}, t_i^{(t)}, M_k) \end{aligned}$$

MCMC integration produces 15000 values for each subject. The conditional predictive ordinate of i^{th} subject from a model k is the inverse of average of 15000 joint densities of model k. Now we calculate log-transformed CPO ratio's for a covariate "SMKEVER" of slope-effect versus changepoint-effect (i.e, M_{s1} vs M_{s2}). ,

$$\text{logCPO}_{i,M_{s1s2}} = \log\left\{\frac{CPO_{i,M_{s1}}}{CPO_{i,M_{s2}}}\right\}$$

Then we have a plot of $logCPO_{i,M_{s1s2}}$ against subject's id. We have done similarly $logCPO_{i,M_{d1d2}}$ when we denote $CPO_{1,M_{d1}}$ as a value of conditional predictive ordinate of i^{th} subject for a model with a slope-effect of covariate "DRKEVER" and $CPO_{1,M_{d2}}$ for a model with changepoint-effect of "DRKEVER". The y-axis is log-scaled CPO ratio and the x-axis is subjects' id. Figure 1 shows (a) a plot of log-scaled CPO ratios for risk factors of "SMOKING" and (b) for "DRINKING". Our decision to choose a covariate effect from each risk factor "SMKEVER" and "DRKEVER" is simple. It depends on proportion of the log-transformed CPO ratio. Since the value of "CPO_{i,M_k}" shows how much the i_{th} observation supports the model, the large CPO value stands for a good fit to the model. For "SMOKING", the number of CPO values above zero is 498, that of below zero is 426 and 8 subjects' CPO values are zero's. Hence we make a conclusion that a model with a slope effect "SMKEVER" covariate has better fit to the model rather than a model with a changepoint effect "SMKEVER". Therefore, we add a covariate of "SMKEVER" with a slope effect for our potential model. Similarly a slope effect of "DRKEVER" has more effective effect. Since the number of CPO values above zero for preferring a slope-effect model of "DRKEVER" is 476, we can have an evidence to choose a slope-effect "DRKEVER" covariate for potential generalized model. Through the univariate analysis, we have three potential significant covariates: a slope effect of "SMKEVER", a changepoint effect of "BMI" and a slope effect of "DRKEVER". With investigation of our covariates from risk factors of prostate cancer, we have three significant covariates: SMKEVER-slope effect, BMI-changepoint effect and DRKEVER-slope effect. We aim to determine a best generalized model of Bayesian hierarchical changepoint models including these covariates. Any combination of three covariates can be potential best models. Let denote c_1, c_2 and c_3 as covariates of SMKEVER, BMI and DRKEVER respectively, and γ_1, γ_2 and γ_3 denote related coefficients. Also assume flat normal priors for $\gamma' s$.

$$\begin{array}{ll}
\text{[Model 0]} & y_{ij} = a_{0i} + a_{1i}x_{ij} + (b_i - a_{1i})(x_{ij} - t_i)^+ + \epsilon_{ij} \\
\text{[Model 1]} & y_{ij} = a_{0i} + a_{1i}x_{ij} + ([b_i + c_1\gamma_1] - a_{1i})(x_{ij} - t_i)^+ + \epsilon_{ij} \\
\text{[Model 2]} & y_{ij} = a_{0i} + a_{1i}x_{ij} + (b_i - a_{1i})(x_{ij} - [t_i + c_2\gamma_2])^+ + \epsilon_{ij} \\
\text{[Model 3]} & y_{ij} = a_{0i} + a_{1i}x_{ij} + ([b_i + c_3\gamma_3] - a_{1i})(x_{ij} - t_i)^+ + \epsilon_{ij} \\
\text{[Model 4]} & y_{ij} = a_{0i} + a_{1i}x_{ij} + ([b_i + c_1\gamma_1] - a_{1i})(x_{ij} - [t_i + c_2\gamma_2])^+ + \epsilon_{ij} \\
\text{[Model 5]} & y_{ij} = a_{0i} + a_{1i}x_{ij} + ([b_i + c_1\gamma_1 + c_3\gamma_3] - a_{1i})(x_{ij} - t_i)^+ + \epsilon_{ij} \\
\text{[Model 6]} & y_{ij} = a_{0i} + a_{1i}x_{ij} + ([b_i + c_3\gamma_3] - a_{1i})(x_{ij} - [t_i + c_2\gamma_2])^+ + \epsilon_{ij} \\
\text{[Model 7]} & y_{ij} = a_{0i} + a_{1i}x_{ij} + ([b_i + c_1\gamma_1 + c_3\gamma_3] - a_{1i})(x_{ij} - [t_i + c_2\gamma_2])^+ + \epsilon_{ij}
\end{array}$$

Model 0 is no-covariate model. Model 1 through 3 are generalized models with one covariate from SMKEVER, BMI and DRKEVER whereas Model 4-6 are generalized models with any two combinations. The model 7 is a generalized model with all three significant covariates. We choose a best model according to analytical as well as graphical consideration.

We calculate CPO values of all 932 subjects for each seven model by implementing BUGS to generate Gibbs samplers of all related subject-specific and population parameters. BUGS produces estimates of CPO values for i^{th} subject with amount of t iterations. Now consider Model 7 for our calculation.

$$\begin{aligned}
\widehat{CPO}_{i,M_7} &= \hat{p}(y_i|y_{(i)}, M_7) \\
&= \left\{ \frac{1}{N-M} \sum_{t=M+1}^{N} \frac{1}{\prod_{j=1}^{n_j} p(y_{ij}|\theta^{(t)}, M_{s1})} \right\}^{-1}
\end{aligned}$$

where $p(y_{ij}|\theta^{(t)}, M_7) = a_{0i} + a_{1i}x_{ij} + ([b_i + c_1\gamma_1 + c_3\gamma_3] - a_{1i})(x_{ij} - [t_i + c_2\gamma_2])^+ + \epsilon_{ij}$. Now we consider *pesudo-Bayes Factor* to obtain information for decision-making. The pseudo-Bayes factor is the ratio of pseudo-predictive likelihood of two competing models when we denote pesudo-predictive likelihood as production of cross-predictive densities. Then we can calculate $2 \cdot \log(pseudo - BF)$ which can lead to application of Jeffreys' guideline.

$$\text{log-pseudoBF}_{k0} = 2 \cdot \sum_{i=1}^{N} \log \left\{ \frac{\hat{p}(y_i|y_{(i)}, M_k)}{\hat{p}(y_i|y_{(i)}, M_0)} \right\}$$

Table 11 show summary of log-pseudo Bayes Factor of each seven models against Model 0. The decision-making is simple so that we choose a Model with the largest value among seven log-pseudo Bayes Factors. Finally we choose Model 4 with covariates of SMKEVER and BMI. We show our

decision with log-ratio of CPO plot for Model k against Model 0. The log-ratio of CPO_i for Model K against Model 0 is calculated

$$\begin{aligned}
\text{logCPO}_{i,k0} &= \log \left\{ \frac{\hat{p}(y_i|y_{(i)}, M_k)}{\hat{p}(y_i|y_{(i)}, M_0)} \right\} \\
&= \log \left\{ \frac{\left[\frac{1}{N-M} \sum_{t=M+1}^{N} \frac{1}{\prod_{j=1}^{n_j} p(y_{ij}|\theta^{(t)}, M_k)} \right]^{-1}}{\left[\frac{1}{N-M} \sum_{t=M+1}^{N} \frac{1}{\prod_{j=1}^{n_j} p(y_{ij}|\theta^{(t)}, M_0)} \right]^{-1}} \right\}
\end{aligned}$$

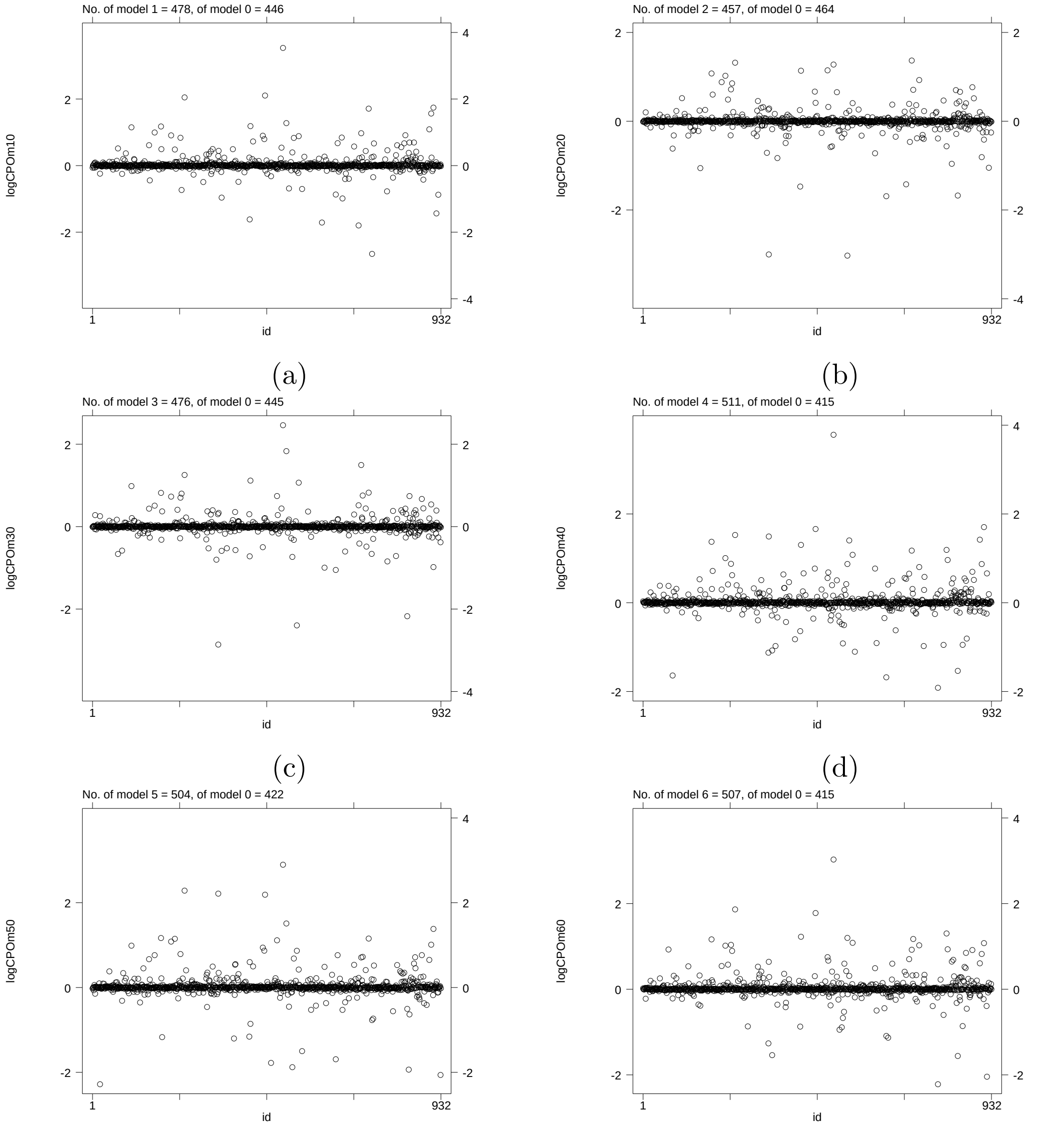
Model	Covariates	Values	Interepretation
1	SMOKING	0.055	Weak support to Model 1
2	BMI	-0.008	Support to Model 0
3	DRINKING	0.012	Weak support to Model 3
4	SMOKING & BMI	0.059	Weak support to Model 4
5	SMOKING & DRINKING	0.045	Weak support to Model 5
6	BMI & DRINKING	0.051	Weak support to Model 6
7	SMOKING, BMI & DRINKING	0.029	Weak support to Model 7

Table 8: Summary table for values of pseudo-Bayes factors for each seven models against no-covariate model

where k = 1, . . . , 7. We can interpret that a better model has the majority of the CPO's above those of the worse one. Below figures shows selective five plots of log-ratio CPO for model k against model 0. The number of CPO_i's of a better model k for above the worse model are on Table 12. We compare the number of subjects on each side for each model. With Figures 2 through 6, we can see that Model 4 which is a generalized model with covariates of SMOKING and BMI is a best model among all seven models.

Plot	Model vs Model	Related Covariates	Better model (the number of CPO's)
1	Model 1 vs Model 0	SMOKING	Model 1 (478 vs 446)
2	Model 2 vs Model 0	BMI	Model 0 (457 vs 464)
3	Model 3 vs Model 0	DRINKING	Model 3 (476 vs 445)
4	Model 4 vs Model 0	SMOKING & BMI	Model 4 (511 vs 415)
5	Model 5 vs Model 0	SMOKING & DRINKING	Model 5 (504 vs 422)
6	Model 6 vs Model 0	BMI & Drinking	Model 6 (507 vs 415)
7	Model 7 vs Model 0	SMOKING, BMI & DRINKING	Model 7 (492 vs 432)
8	Model 4 vs Model 5	ALL Three	Model 4 (484 vs 443)
9	Model 4 vs Model 6	ALL Three	Model 4 (479 vs 445)
10	Model 4 vs Model 7	ALL Three	Model 4 (487 vs 439)

Table 9: Summary table for the number of subjects on the majority indicating a better model against a worse one

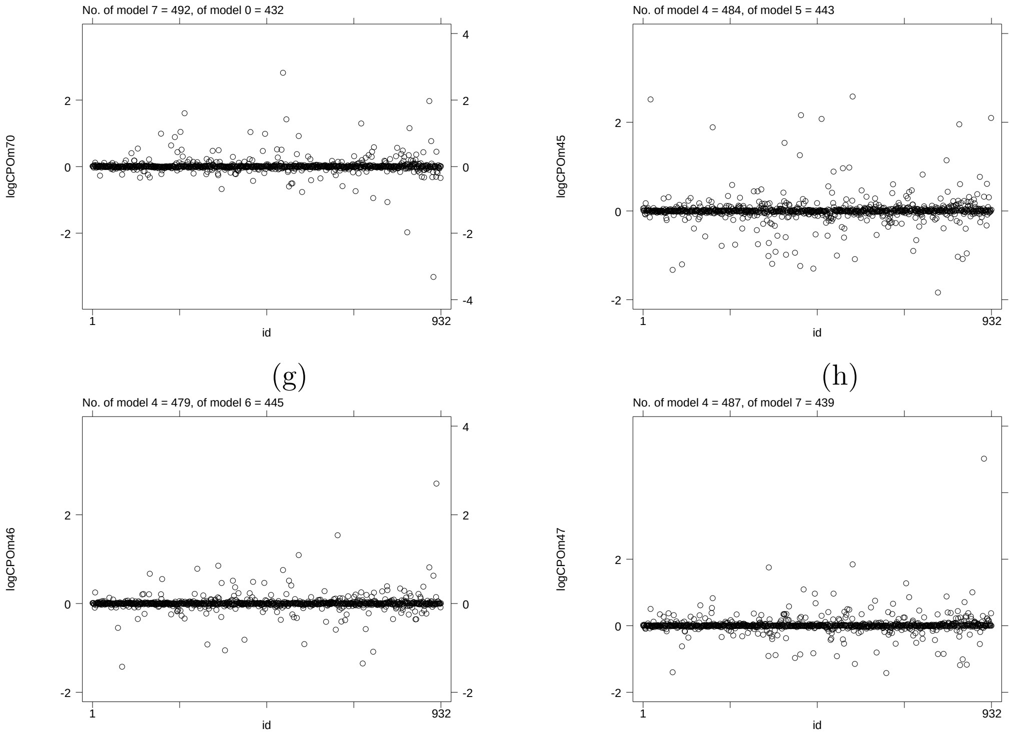


(a) (b)

(c) (d)

(e) (f)

Figure 3: log-ratio CPO Plots for (a) Model 1 with SMOKING effect vs Model 0 without covariates (b) Model 2 with BMI-effect vs Model 0 (c) Model 3 with DRIKING-effect vs Model 0 without covariates (d) Model 4 with SMOKING and BMI effects vs Model 0 (e) Model 5 with SMOKING and DRIKING effects vs Model 0 without covariates (f) Model 6 with BMI and DRIKING effects vs Model . The number of subjects on each side are 478 vs 446 for (a), 457 vs 464 for (b), 476 vs 445 for (c), 511 vs 415 for (d), 504 vs 422 for (e), and 507 vs 415 for (f).



(g) (h)

(g) (h)

Figure 4: log-ratio CPO Plots for (g) Model 7 with SMOKING, BMI and DRIKING effects vs Model 0 without covariates (h) Model 4 with SMOKING and BMI effects vs Model 5 with SMOKING and DRIKING (i) Model 4 with SMOKING and BMI effects vs Model 6 with BMI and DRINKING effects (j) Model 4 vs Model 7 with SMOKING, BMI and DRIKING. The number of subjects on each side are 492 vs 432 for (g), 484 vs 443 for (h), 479 vs 445 for (i), and 487 vs 439 for (j).

From Figure 7, we can see that the posterior probability method is better than the threshold method in early detection procedure of prostate cancer, but do not make a conclusion that for the posterior probability method, a diagnostic rule based on a best model has better chance for early detection than that on no-covariate model.

6 Discussion

We execute variable selection procedure to find a most appropriate model which can lead a best detection model of prostate cancer. To do that, we use several techniques of Markov chain Monte Carlo including Bayesian cross-validation and pseudo Bayes factor. A best combination of covariate model has slightly better chance for early detection of prostate cancer rather than no-covariate model but the difference between two models was not far away each other. NTC Trial data has some information about prostate cancer stage: localized or advanced. If we consider the covariate into our model, we might have a more positive result.

For future research, we would like to extend a Bayesian hierarchical change point model by considering more than two change points in the model. There are literatures on the multiple change-point problem. Stephens (1994) emphasizes retrospective change point identification through several examples of discrete and continuous multiple-changepoint with the use of a sampling-based technique,

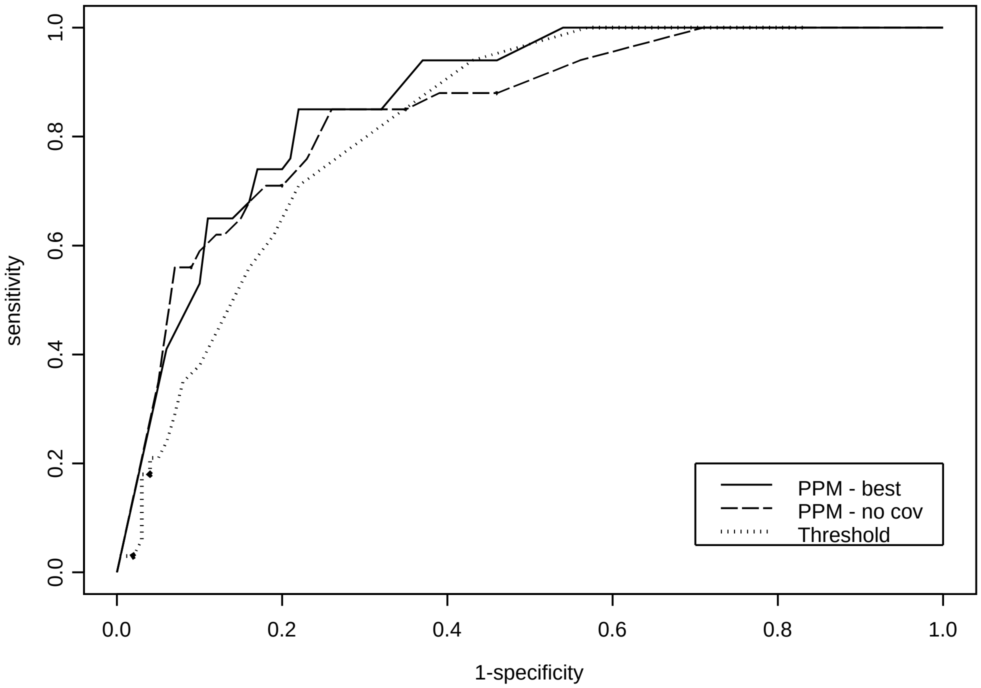


Figure 5: Comparison of Retro-spective ROC curves: (a) threshold method (b) No-covariate model based on posterior probability method (c) Best model based on posterior probability model

the Gibbs sampler, and Slate and Cronin (1997) considered a segmented random-effect model with two changepoints within a fully Bayesian framework, providing a time-dependent sensitivity to show how this prospective diagnostic detection progress is so effective. We also would like to continue to extend the model by applying possible non-linear models.

References

[1] American Cancer Society. ”Cancer Facts and Figures 2004” and ”Prostate Cancer”, Cancer Reference Information, www.cancer.org, 2004.

[2] Burnham, K.P. and Anderson, D.R. (1998). Model Selection and Inference: A Practical Information-Theoretical Approach. Springer.

[3] Carlin, B.P., Gelfand, A.E. and Smith, A.F.M. (1992). Hierarchical Bayesian Analysis of Changepoint Problems. *Applied Statistics,* 41(2). 389-405.

[4] Carter, H.B., Morrell, C.H., Pearson, J.D., Brant, L.J., Plato, C.C., Metter, E.J., Chan, D.W., Fozard, J.L. and Walsh, P.C. (1992). Estimation of prostatic growth using serial prostate-specific antigen measurements in men with and without prostate disease. *Cancer Rsearch,* 52,

[5] Catalona, W.J., Smith, D.S. and Ornstein, D.K. (1997). Prostate cancer incidence in men with serum PSA concentration of 2 to 4 ng/ml and benign prostate examination. *Journal of the American Medical Association,* 277, 1452-1455.

[6] Chen, M.H., Shao, Q., and Ibrahim, J.G. (2000). Monte Carlo Methods in Bayesian Computation. Springer-Verlag.

[7] Chen, M.H. and Shao, Q. (1999). Monte Carlo Estimation of Bayesian Credible and HPD Intervals. *Journal of Computational and Graphical Statistics,* Vol.8, No.1, 69-92

[8] Clark, L.C., Combs, G.F., Turnbull, B.W., Slate, E.H., Alberts, D.S., Abele, D., Allison, R.J., Chalker, D.K., Gross, E.G., Hendrix, J.D., Herlong, J.H., Hixson, L.J., Kight, F., Krongrad, A., Lesher, J.L., Moore, J., Park, H.K., Rice, J.S., Rogers, A.I., Sanders, B.B., Schuman, B., Smith, C.L., Smith, E.H., Tayor, J.R., and Woodard, J.C. (1996). The nutritional prevention of cancer with selenium 1983-1993: A randomized clinical trial. *Journal of the American Medical Association,* 276, 1957-1963.

[9] Cronin, K.A. (1995). Detection of Changepoints in Longitudinal Data, Phd. Thesis. School of Operations Reasearch, Cornell University.

[10] Dey, D.K., Kuo, L. and Sahu, S.K. (1995) A Bayesian predictive Approach to determining the number of components in a mixture distribution. *Statistics and Computing,* **5,** 297-305.

[11] Gelfand A. E. (1995) Model determination using sampling-Based methods. In *Markov Chain Monte Carlo in Practice*(eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 145-151. Lodon: Chapman & Hall.

[12] Gelfand A.E. and Smith, A.F. (1990) Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association,* **85,** No.410, 398-409.

[13] Gelfand A.E. and Smith, A.F. (1991) Gibbs SAmpling for marginal posterior expectations. *Communication in Statistics,* **20(5 & 6),** 1747-1766.

[14] Gelman, A., Carlin, J.B. and Stern, H.S. (1997) *Bayesian Data Analysis.* Boca Raton: Chapman & Hall/CRC.

[15] Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996). Markov Chain Monte Carlo In Practice. Chpaman & Hall

[16] Hiatt, R.A., Armstrong, M.A., Klatsky, A.L. and Sidney, S. (1994). Alcohol consumption, smoking, and other risk factors and prostate cancer in a large health plan cohort in California (United States). *Cancer Cases and Control,* 5, 66-72.

[17] Kass, R.E., and Raftery, A.E. (1993). Bayes Factors. Technical report no. 254, Departement of Statistics, University of Washington.

[18] Kiuchi, A.S., Hartigan, J.A., Holford, T.R., Rubinstein, P. and Stevens, C.E. (1995). Change Points in the Series of T4 Counts Prior to AIDS. *Biometrics,* 51. 236-248.

[19] Lange, N., Carlin, B.P., Gelfand, A.E. (1992). Hierarchical Bayes Models for the Progression of HIV Infection Using Longitudinal CD4 T-Cell Numbers. *Journal of the American Statistical Association,* 87, 615-632. School of Operations Reasearch, Cornell University.

[20] McCullagh, P. and Nelder, J.H. (1989). *Generalized Linear Models.* London: Chapman and Hall.

[21] Murtaugh, P.A. and Schaid, D.J. (1991). Application of R.O.C. curve methodology when markers are repeated measures. *Procedding of the International Biometric Society (ENAR),* Houston, TX, March 24-27.

[22] Sinha, D., Chen, M-H and Ghosh, S.K. (1999) Bayesian Analysis and Model Selection for Interval-Censored Survival Data. *Biometrics,* **55,** 585-590.

[23] Slate, E.H. and Clack, L.C. (1999). Using PSA to detect prostate cancer onset: An application of Bayesian retrospective and prospective changepoint identification. *Case Studies in Bayesian Statistics IV,* eds. 511-534.

[24] Slate, E.H. and Cronin, K.A. (1997). Changepoint modeling of longitudinal PSA as a biomarker for prostate cancer. *Case Studies in Bayesian Statistics III,* eds. C. Gatsonis, J.S. Hodges, R.E. Kass, R. McCulloch, P. Rossi and N.D. Singpurwall, (eds), , Springer-Verlag, New York, pp. 444-456.

[25] Slate, E.H. and Turnbull, B.W. (2000). Statistical methods for longitudinal biomarkers of disease onset. *Statistics in Medicine,* 19(4), 617-637.

[26] Spiegelhalter, D.J., Thomsa, A., Best. N.G. and Gilks, W.R. (1996) *BUGS: Bayesian Inference using Gibbs Sampling, Version 0.5.* Cambridge: Medical Research Council Biostatistics Unit.

[27] Whittemore, A.S., Lele, C., Friedman, G.C., Stamey, T., Vogelman, J.J., and Orentreich, N. (1995). Prostate-specific antigen as predictor of prostate cancer in black men and white men. *Journal of the National Cancer Institute,* 87(5), 354-360

[28] Yoo, W. and Slate, E.H. A simulation study of a Bayesian hierarchical changepoint model with covariates (in prep).