Biometrics 65, 1030–1040
December 2009
DOI: 10.1111/j.1541-0420.2008.01181.x

# Bayesian Variable Selection with Joint Modeling of Categorical and Survival Outcomes: An Application to Individualizing Chemotherapy Treatment in Advanced Colorectal Cancer

**Wei Chen,[1,*] Debashis Ghosh,[2] Trivellore E. Raghunathan,[3] and Daniel J. Sargent[4]**

[1]Biostatistics Core, Karmanos Cancer Institute, Wayne State University, Detroit, Michigan 48201, U.S.A.
[2]Departments of Statistics and Public Health Sciences, Penn State University, University Park, Pennsylvania 16802, U.S.A.
[3]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[4]Division of Biostatistics, Mayo Clinic, Rochester, Minnesota 55905, U.S.A.
[*]*email:* chenw@karmanos.org

Summary. Colorectal cancer is the second leading cause of cancer related deaths in the United States, with more than 130,000 new cases of colorectal cancer diagnosed each year. Clinical studies have shown that genetic alterations lead to different responses to the same treatment, despite the morphologic similarities of tumors. A molecular test prior to treatment could help in determining an optimal treatment for a patient with regard to both toxicity and efficacy. This article introduces a statistical method appropriate for predicting and comparing multiple endpoints given different treatment options and molecular profiles of an individual. A latent variable-based multivariate regression model with structured variance covariance matrix is considered here. The latent variables account for the correlated nature of multiple endpoints and accommodate the fact that some clinical endpoints are categorical variables and others are censored variables. The mixture normal hierarchical structure admits a natural variable selection rule. Inference was conducted using the posterior distribution sampling Markov chain Monte Carlo method. We analyzed the finite-sample properties of the proposed method using simulation studies. The application to the advanced colorectal cancer study revealed associations between multiple endpoints and particular biomarkers, demonstrating the potential of individualizing treatment based on genetic profiles.

Key words: Bayesian multivariate regression; Biomarker; Hierarchical model; Interaction; Latent variable; Oncology.

## 1. Introduction

In most cancer clinical trials, some patients respond to chemotherapy very well while others show no sign of response. Likewise, some patients experience more toxicity than others given the same treatment. Recent clinical studies have suggested that patients who possess specific genetic alterations or mutations may respond differently to the same treatment for colorectal cancer (Milano and McLeod, 2000). However, tools for individualizing chemotherapy treatment using genetic profiles are not yet fully developed (McLeod and Murray, 1999).

The objective of a randomized phase III trial (Goldberg et al., 2004), initiated by the Mayo Clinic in 1997, was to compare the effect of combinations of chemotherapy agents in patients with advanced colorectal cancer. At that time, two chemotherapy drugs had been approved by the Food and Drug Administration for treatment of advanced colon cancer: 5-fluorouracil (5-FU) and irinotecan (CPT-11), while oxaliplatin (OXAL), a cisplatinum analogue with activity in colorectal cancer, was an investigational agent in the United States and Canada. Two experimental combinations of regimens, 5-FU+OXAL and OXAL+CPT-11, were compared to the standard regimen, 5-FU+CPT-11, in the trial. We refer to these regimens as arm F, arm G, and the control as arm A, respectively. A total of 1705 patients were included in the study, of which 513 (115 patients in arm A, 292 patients in arm F, and 106 patients in arm G) were genotyped for 23 biomarkers. These biomarkers were selected based on previous reports indicating that they were related to bioactivity of the chemotherapies by direct or indirect mechanisms. Descriptive summaries of the covariates are shown in Table 1.

In this article, we develop a statistical model, given multiple treatment options and characteristics of an individual, which predicts and compares toxicity and efficacy simultaneously. A direct comparison of treatments will compare the probability that the predicted multiple endpoints for an individual patient or a group of patients fall in a favorable region of treatment outcomes ($\Delta$). This process involves two tasks: (1) building a predictive model with appropriate prognostic and predictive factors and (2) for each treatment, predicting the probability of being in region $\Delta$. To be concrete, let $y$ be a $k$-dimensional vector of outcomes ($y = (y_1, \ldots, y_k)$), $x$ a vector of $p$ predictors, and $\theta$ a vector of unknown parameters given model $M$. The goal is to find $\Pr(y \in \Delta \,|\, x, \theta, M)$ through the joint distribution $p(y_1, \ldots, y_k \,|\, x, \theta, M)$.

Methods that involve Bayesian inference for variable selection, motivated by the seminal work of George and

**Table 1**
*Summary of the variables in the colorectal cancer study*

| | | | Mean | | |
|---|---|---|---|---|---|
| | Variables | Description | Arm A ($n = 115$) | Arm F ($n = 292$) | Arm G ($n = 106$) |
| Outcomes | Response | 1: Yes, 0: No | 33.9% | 42.8% | 33% |
| | Toxicity | 1: Grade > 3 | 12.2% | 23.3% | 20.8% |
| | TTP | Median TTP | 188 | 308 | 249 |
| | status | 0: censor, 1: event | 84.3% | 68.2% | 79.2% |
| DGV* | AGE | age | 60.835 | 59.685 | 59.179 |
| | SEX | 0: Female, 1: Male | 61.7% | 59.6% | 53.8% |
| Marker | M1 | abcb1/_12 | 75.7% | 83.2% | 79.2% |
| | M2 | abcb1/_2677 | 73.9% | 68.5% | 72.6% |
| | M3 | abcb1/_3435 | 80.9% | 77.7% | 84.9% |
| | M4 | abcc1/_14008 | 94.8% | 93.2% | 93.4% |
| | M5 | abcc1/_34215 | 35.7% | 28.4% | 34.9% |
| | M6 | abcc2/_24 | 98.3% | 94.5% | 98.1% |
| | M7 | abcc2/_c1515y | 13.9% | 11.3% | 7.5% |
| | M8 | abcc2/_v417i | 39.1% | 38.7% | 39.6% |
| | M9 | cyp3a4 | 13% | 14% | 9.4% |
| | M10 | cyp3a5 | 16.5% | 18.8% | 15.1% |
| | M11 | dpyd/_5 | 13.9% | 18.8% | 8.5% |
| | M12 | dpyd/_6 | 5.2% | 9.6% | 9.4% |
| | M13 | dpyd/_9a | 42.6% | 41.1% | 41.5% |
| | M14 | ercc2/_k751q | 88.7% | 86.6% | 80.2% |
| | M15 | gstm1/_0 | 47.8% | 49% | 45.3% |
| | M16 | gstp1/_I105v | 54.8% | 57.9% | 52.8% |
| | M17 | gstp/_114 | 5.2% | 16.4% | 15.1% |
| | M18 | mthfr | 20.9% | 18.5% | 26.4% |
| | M19 | tyms/_1494del | 56.5% | 50.3% | 53.8% |
| | M20 | tyms/_tser | 73% | 68.8% | 67% |
| | M21 | ugt1a1/_28 | 57.4% | 50.3% | 49.1% |
| | M22 | xrcc1/_399 | 57.4% | 54.5% | 62.3% |
| | M23 | ABCG2Q141K | 19.1% | 24.7% | 17% |

DGV = demographic variables.
For binary variables, percentage of 1 was calculated.

McCulloch (1993), have been applied in many studies (e.g., Brown, Vannucci, and Fearn, 1998; Chen and Dey, 2003; Sha, Tadesse, and Vannucci, 2006). Bayesian methods have the potential to account for small samples and selection of derived covariates, e.g., interaction terms, through a proper specification of the priors. In this article, we propose a statistical method to individualize treatment in the Bayesian framework.

Another challenge to conducting an appropriate analysis is that $\Delta$ is a complicated multidimensional space. For example, a patient with a confirmed tumor response may have a longer survival. A multivariate analysis is a reasonable choice when outcomes are not independent, as information is borrowed from the potentially correlated outcomes for variable selection. Furthermore, by estimating the posterior predictive probability of multiple outcomes in region $\Delta$, one can provide a single score to compare treatments from multiple perspectives which may be related, e.g., time to progression, and overall survival. This is of critical importance, because an informed therapeutic decision is often based on consideration of multiple endpoints. When outcomes follow different type of distributions the coefficients of the same regressor across the different outcomes are not directly comparable. In the context of a multivariate regression model, properly scaled coefficients are needed. This can been done using latent variables; in addition, the variance–covariance matrix of such a multivariate regression model is often structured, which dictates additional care when specifying a prior. To address these issues, we propose a method called multivariate Bayesian selection of interactions (MBSI). The MBSI method features joint modeling of categorical and survival responses, a nonconjugate prior for structured variance components, selection of interactions with limited sample size, and a variable selection rule. Each issue has been addressed separately in literature; nevertheless, the particular combination has not been studied.

In Section 2, we describe the hierarchical structure of our MBSI model, and we will show how the decision rules for variable selection are derived. The performance of the MBSI is studied through simulations in Section 3. In Section 4, the method will be illustrated in an analysis of the large phase III colorectal cancer study. Finally, a short discussion is given in Section 5.

## 2. Method

### 2.1 *Multivariate Regression Model with Latent Variables*

We consider the problem of modeling the relationship between $K$ multiple responses (including categorical and survival) and $p$ predictor variables with sample size $n$ in a Bayesian framework. In the motivating example, we are interested in $K \equiv 3$ outcomes. Let $y_1$ and $y_2$ be the binary outcomes of toxicity and tumor response, $y_3$ be the survival outcome, $x_i$ be the $p \times 1$ vector of predictors for the $i$th sample, and $\boldsymbol{X} = (x_1, \ldots, x_n)^T$. We assume in this study that the columns of $\boldsymbol{X}$ have been standardized by subtracting their column means and dividing by their column standard deviations. Each response is assumed to follow its own regression model; let $\beta_{(k)}$ be a $p \times 1$ vector that is a column of the regression coefficients and $\alpha_k$ a scalar that is an intercept corresponding to the $k$th outcome, $k = 1, \ldots, 3$.

We relate the regressors with the binary responses $y_1$ and $y_2$ through a probit link, which yields

$$Pr(y_k = 1 \,|\, \boldsymbol{X}, \alpha_k, \beta_{(k)}) = \Phi(\alpha_k + \boldsymbol{X}\beta_{(k)}),$$

where $\Phi$ is the normal cumulative distribution function and $k = 1, 2$. We introduce a $n \times 1$ vector of latent variable $z_k$ (Albert and Chib, 1993). For each individual ($i = 1, \ldots, n$),

$$y_{ik} = \begin{cases} 1 & \text{if } z_{ik} > 0 \\ 0 & \text{if } z_{ik} \leqslant 0 \end{cases}.$$

The latent variable has a linear model form: $z_k = \alpha_k + \boldsymbol{X}\beta_{(k)} + \varepsilon_k$, where $\varepsilon_k \overset{iid}{\sim} N(0,1)$. For reasons of identifiability, we set the scale parameter of the normal distribution to be 1.

With respect to the survival outcome $y_3$, denote the censoring indicator by a $n \times 1$ vector $c$, where $c_i = 0$ if $y_{i3}$ is right censored and $c_i = 1$ otherwise ($i = 1, \ldots, n$). We introduce a $n \times 1$ vector of complete variable $z_3$, defined as

$$\log(y_{i3}) \begin{cases} = z_{i3} & \text{if } c_i = 1 \\ < z_{i3} & \text{if } c_i = 0 \end{cases}.$$

The complete variable has a normal distribution as $z_3 \sim N(\alpha_3 + \boldsymbol{X}\beta_{(3)}, \sigma^2)$. Note that when $c = 0$, the variable $z_3$ is unobservable. Our approach is same as the data augmentation Gibbs sampling method used in Bayesian variable selection with log-normal accelerated failure time model in Sha et al. (2006). This results in explicit full conditional distributions of $\boldsymbol{\beta} \equiv \left[\beta_{(1)}, \beta_{(2)}, \beta_{(3)}\right]$ and $\boldsymbol{Z} \equiv [z_1, z_2, z_3]$, and leads to a more efficient Markov chain Monte Carlo (MCMC) implementation. If a proportional hazards model is preferred, we can specify a conditional hazard function $h(y_3 \,|\, z_3) = h_0(y_3) \exp(z_3)$, where $z_3 \sim N(\alpha_3 + \boldsymbol{X}\beta_{(3)}, \sigma^2)$. This setting changes the relationship between $y_3$ and $z_3$ to a stochastic one. With $h_0(y_3) = \lambda$ we obtain the exponential regression model; with $h_0(y_3) = \varphi y_3^{(\varphi-1)}$, where $\varphi > 0$, we obtain the Weibull regression model. Consequently, the full conditional distributions of $\boldsymbol{\beta}$ and $\boldsymbol{Z}$ are no longer in explicit forms. Thus, the MCMC sampling efficiency decreases greatly for variable selection in a trivariate regression setting.

Having introduced the latent variables ($z_1$, $z_2$) for binary responses and the complete variable ($z_3$) for survival response, we can rewrite the preceding model in a multivariate regression form. Letting $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]^T$ and $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \varepsilon_3]$, we have

$$\boldsymbol{Z} = \mathbf{1}_n \boldsymbol{\alpha}^T + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{and} \quad \text{vec}(\boldsymbol{\varepsilon}^T) \sim N(\mathbf{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}),$$
$$\text{where} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_1 & \rho_2\sigma \\ \rho_1 & 1 & \rho_3\sigma \\ \rho_2\sigma & \rho_3\sigma & \sigma^2 \end{bmatrix}. \tag{1}$$

Note that $\otimes$ denotes the Kronecker product, and vec ($\boldsymbol{\varepsilon}^T$) is the vector obtained by stacking the columns of $\boldsymbol{\varepsilon}^T$ on top of each other. Let $\boldsymbol{\rho} = [\rho_1, \rho_2, \rho_3]$ be the vector of correlation coefficients resulting from the assumption that the three outcomes are mutually correlated.

### 2.2 *Bayesian Variable Selection*

We introduce a $p \times 3$ matrix of binary latent variables $\boldsymbol{\gamma}$ with components $\gamma_{jk} = 1$ ($j = 1, \ldots, p$ and $k = 1, 2, 3$) if the $j$th regressor $\boldsymbol{X}_{(j)}$ is included in the $k$th model and $\gamma_{jk} = 0$ otherwise. We assume that the prior distribution of $\boldsymbol{\beta}$ is a multivariate mixture normal that depends on $\boldsymbol{\gamma}$. It takes the form $\text{vec}(\boldsymbol{\beta}^T) \,|\, \boldsymbol{\gamma}, \sigma \sim N_{3p}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ a block diagonal matrix with blocks $\Sigma_{\beta j} =$

$$\begin{bmatrix} \left(1-\gamma_{j1}+\gamma_{j1}c^2\right)\tau^2 & 0 & 0 \\ 0 & \left(1-\gamma_{j2}+\gamma_{j2}c^2\right)\tau^2 & 0 \\ 0 & 0 & \left(1-\gamma_{j3}+\gamma_{j3}c^2\right)\tau^2\sigma^2 \end{bmatrix}. \tag{2}$$

Here $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ corresponds to an a priori independence assumption for the coefficients. Note that for the survival outcome in equation (2), the variance is adjusted by a quantity $\sigma^2$. Because the outcomes have different scales, the unadjusted $\boldsymbol{\beta}$ will be incomparable across the multiple outcomes.

In this prior setting, the variable selection problem is formulated in terms of making inferences regarding $\boldsymbol{\gamma}$. A value of $c$ determines the magnitude of the difference between the two mixture normal distributions. George and McCulloch (1993) suggested that choosing $c$ between 10 and 100 tends to work well when implementing MCMC and that computational difficulties can be avoided whenever $c \leqslant 100$. In addition, $\tau$ shall be small enough so that $\beta_{jk}$ is close to zero when $\gamma_{jk} = 0$.

We assume that all three intercepts are always included in the model. Hence, selection of the intercepts is not performed. We assume a simple diffuse multivariate normal prior for $\boldsymbol{\alpha}$ independent of $\boldsymbol{\gamma}$:

$$\boldsymbol{\alpha} \,|\, \sigma \sim N_3(\mathbf{0}, \boldsymbol{\Sigma}_\alpha), \qquad \text{where} \quad \boldsymbol{\Sigma}_\alpha = \begin{bmatrix} \sigma_\alpha^2 & 0 & 0 \\ 0 & \sigma_\alpha^2 & 0 \\ 0 & 0 & \sigma^2\sigma_\alpha^2 \end{bmatrix}.$$

Note that $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ is adjusted for the survival outcome in the same manner as in equation (2).

In a Bayesian regression framework, a Wishart distribution is often used as a prior for the variance matrix $\boldsymbol{\Sigma}$ for

convenience. However, in our case the Wishart distribution is not a proper prior due to the constraint on the variance of the binary outcomes $y_1$ and $y_2$. Therefore, we modeled $\boldsymbol{\Sigma}$ using the separation strategy of Barnard, McCulloch, and Meng (2000).

The prior distributions for $\sigma$ and $\boldsymbol{\rho} = [\rho_1, \rho_2, \rho_3]$ in equation (1) are specified separately. For $\sigma$, an inverse gamma prior is used $\sigma \sim$ IG ($\nu/2$, $\nu\lambda/2$). We assume that the correlations $\rho_1$, $\rho_2$, $\rho_3$ are a priori exchangeable. To ensure the positive definiteness of $\boldsymbol{\Sigma}$, we need a constraint

$$1 - \rho_1^2 - \rho_2^2 - \rho_3^2 + 2\rho_1\rho_2\rho_3 > 0. \tag{3}$$

Hence, a joint uniform prior for $\rho_1$, $\rho_2$, $\rho_3$ is specified as

$$p(\rho_1, \rho_2, \rho_3) = \frac{2}{\pi^2} I_{(1-\rho_1^2-\rho_2^2-\rho_3^2+2\rho_1\rho_2\rho_3>0, \rho_1,\rho_2,\rho_3\in(-1,1))}$$

with informative marginal distribution

$$p(\rho_k) = \frac{2}{\pi}\sqrt{1-\rho_k^2}, \quad \rho_k \in (-1, 1), \quad k = 1, 2, 3.$$

Given the other two elements of $\boldsymbol{\rho}$, say $\rho_2$ and $\rho_3$, $\rho_1$ has a uniform conditional distribution $\rho_1 \,|\, \rho_2,\ \rho_3 \sim \mathrm{U}(L_1, U_1)$, where $L_1 = \rho_2\rho_3 - \sqrt{(1-\rho_2^2)(1-\rho_3^2)}$ and $U_1 = \rho_2\rho_3 + \sqrt{(1-\rho_2^2)(1-\rho_3^2)}$. The lower and upper bounds $L_1$ and $U_1$ are the roots of the inequality (3). The marginal densities of $\boldsymbol{\rho}$ are symmetric and have more mass close to zero than the tails, which is a plausible assumption because the correlations are rarely very large in real applications.

### 2.3 *Prior for* $\boldsymbol{\gamma}$

We first assume that the selection of regressors is a priori independent across the outcomes. Second, within each outcome, we assume that the selection of a main effect is dependent on the selection of all its interaction terms. Because interaction terms represent deviation from an additive model, we adopt the convention that a model containing an interaction term should also contain the corresponding main effects (Neter et al., 1996).

Define a $p \times 3$ matrix of indicator variables $\boldsymbol{\xi}$ with components $\xi_{jk} = 0$, ($j = 1, \ldots, p$, $k = 1, 2, 3$) if the corresponding regressor is an interaction, and $\xi_{jk} = 1$ if a main effect. Let $\Omega_{jk}$ be the set of all the latent variables ($\gamma_{j'k}, j' \neq j$) for the interaction terms that are related to the $j$th main effect of $k$th model. Following the above assumptions, the prior for $\gamma_{jk}$ is a Bernoulli distribution and takes the form

$$p(\gamma_{jk} = 1 \,|\, \Omega_{jk}, \pi_{jk}) = \begin{cases} 1 & \text{if } \xi_{jk} = 1 \text{ and } \sum_{\gamma_{j'k}\in\Omega_{jk}} \gamma_{j'k} > 0 \\ \pi_{jk} & \text{if } \xi_{jk} = 1 \text{ and } \sum_{\gamma_{j'k}\in\Omega_{jk}} \gamma_{j'k} = 0 \\ \pi_{jk} & \text{if } \xi_{jk} = 0 \end{cases}$$

$\pi_{jk} \sim \mathrm{beta}(a, b)$.

To favor parsimonious models or when $n < p$, the parameters ($a$, $b$) in the beta prior can be chosen to force $\pi_{jk}$ to be small.

### 2.4 *Decision Rules for Variable and Model Selection*

It is commonly perceived that the optimal predictive model is the model with highest joint posterior probability, but this is not necessarily the case as discussed in Barbieri and Berger (2004). Our simulation studies also showed that a method based on marginal posterior probabilities was better than that of the highest joint posterior probability method regarding the prediction performance (Table 2 in Section 3.2 and Web Figure 1).

The major differences between choosing a model (using highest joint posterior probability) and choosing important predictors (using marginal posterior probabilities) are the following: (1) The selection domain is $p$ for predictors in contrast to $2^p$ for models, assuming no restriction. For a fixed number of MCMC iterations, it is desirable to search in a small parameter domain space. Determining marginal probabilities is computationally simpler than determining highest joint probabilities (Barbieri and Berger, 2004). (2) The joint prior probabilities are affected by the size of $p$; whereas the marginal prior probabilities are invariant. (3) The "dilution effect" (George, 1999), caused by multicollinearity, reduces the marginal posterior probabilities allocated to highly correlated variables. This may lead to ruling out potential predictors. To alleviate this problem, if a nearly perfect Pearson's coefficient of correlation (0.9 or greater) between any two variables is observed, the one with smaller sample variance is suggested to be removed when there is no scientific reason to suggest removing the other. (4) When variable selection is performed, a decision of selecting how many predictors is required. We will briefly describe a false discovery rate (FDR)-based decision rule developed by Chen et al. (2008) at the end of this section.

There are two types of errors in the variable selection problem: (1) selecting a variable that in truth is not a predictor (false discovery); and (2) not selecting a variable that in truth is a predictor (false negative). These two errors can be quantified by two complementary Bayesian losses: posterior expected FDR ($\overline{\mathrm{FDR}}$) and posterior expected FNR ($\overline{\mathrm{FNR}}$). For the sake of simplicity in this section only, we introduce notation omitting subscript $k$ for the $k$th outcome, because the definition is the same for all three outcomes. Let $d_j$ denote the decision of inclusion ($d_j = 1$) or exclusion ($d_j = 0$) of $j$th predictor given data, $D = \sum_{j=1}^{p} d_j$, and $p$ the total number of regressors in consideration. The $\overline{\mathrm{FDR}}$ and $\overline{\mathrm{FNR}}$ are

$$\overline{\mathrm{FDR}} = \begin{cases} E_{\boldsymbol{\gamma}|\mathrm{Data}}\left(\dfrac{\sum d_j(1-\gamma_j)}{D}\,\middle|\,\mathrm{Data}\right) & \\ \quad = \dfrac{\sum d_j(1-\nu_j)}{D} & \text{if } D > 0 \\ 0 & \text{if } D = 0 \end{cases}$$

and (4)

$$\overline{\mathrm{FNR}} = \begin{cases} E_{\boldsymbol{\gamma}|\mathrm{Data}}\left(\dfrac{\sum (1-d_j)\gamma_j}{p-D}\,\middle|\,\mathrm{Data}\right) & \\ \quad = \dfrac{\sum (1-d_j)\nu_j}{m-D} & \text{if } D < p \\ 0 & \text{if } D = p, \end{cases}$$

where $\nu_j = P(\gamma_j = 1 \,|\, \text{Data})$ is the marginal posterior probability of unobserved truth ($\gamma_j$) of $j$th predictor. It is estimated by the proportion of times that $\gamma_j = 1$ through the MCMC iterations. In other words, selection of a variable is based on its frequency of occurrences in any model. This is similar to the variable importance measures in the random forest approach, where a variable is considered important if it appears in many trees. The posterior expected losses defined in equation (4) utilize the simple zero-one loss function. To minimize a two-dimensional loss, one can minimize one dimension while controlling the other.

Accounting for the relationship between main effects and interactions, the $\overline{\text{FDR}}$ and $\overline{\text{FNR}}$ will be calculated separately within the set of main effects ($s_1$) and the set of interactions terms ($s_2$). Hence, given $\alpha_L$ for main effects and $\alpha_H$ for interactions, we have $\min\left\{\overline{\text{FNR}}_{s_1}, \overline{\text{FNR}}_{s_2} \mid \overline{\text{FDR}}_{s_1} \leqslant \alpha_L, \overline{\text{FDR}}_{s_2} \leqslant \alpha_H\right\}$. The total $\overline{\text{FDR}}_{(s_1,s_2)}$ is controlled at $\max(\alpha_L, \alpha_H)$ (see Lemma 1 in Chen et al. [2008]).

As the decision in $s_1$ is affected by the decision of their higher-order terms in $s_2$, we start with minimizing the $\overline{\text{FNR}}_{s_2}$ followed by minimizing the $\overline{\text{FNR}}_{s_1}$. After the decisions are made for all the terms in $s_2$, a subset (denoted as $s_1'$) of the lower-order terms in $s_1$ is forced to be included in the model due to the constraint. Hence, those terms are not involved in the decision to minimize the $\overline{\text{FNR}}_{s_1}$. Decisions will be made for the remaining terms in the complement of $s_1'$ (denoted as $\overline{s_1'}$). The following algorithm illustrates the steps to reach these decisions. First, if $t$ is a threshold such that the decision $d_j = I(\nu_j \geqslant t)$, $j \in s_2$ satisfies $\overline{\text{FDR}}_{s_2} \leqslant \alpha_H$, the optimal threshold $t_H^* \equiv \min\{t : \overline{\text{FDR}}_{s_2} \leqslant \alpha_H\}$ minimizes $\overline{\text{FNR}}_{s_2}$. Hence, the decisions for the terms in $s_2$ is $d_j = I(\nu_j \geqslant t_H^*)$, $j \in s_2$. Second, within $s_1$, we identify a subset $s_1'$ so that its related higher-order term is selected in step 1. Set $d_j = 1$, $j \in s_1'$. We force these terms to be selected. Third, for the terms in $\overline{s_1'}$ (complement of $s_1'$), we find an optimal threshold $t_L^*$ such that the decision $d_j = I(\nu_j \geqslant t_L^*), j \in \overline{s_1'}$ minimizes $\overline{\text{FNR}}_{s_1}$ in the same fashion as in step 1.

The choice of $\alpha_L$ and $\alpha_H$ is different from the conventional significance level for the family-wise error rate. While a high level of ($\alpha_L$, $\alpha_H$) results in more falsely selected regressors, it would not necessarily cause a worse prediction to the outcome but may lead to a less parsimonious model. Therefore, $\alpha_L$ and $\alpha_H$ can be set according to one's degree of tolerance in the number of the false predictors. In the case of $n < p$, however, the prior dominates the posterior distribution. Hence, any variable selection method based on the marginal posterior distributions must take the prior assumptions into account. We consider the average posterior to prior probability ratio as the strength of evidence from the data. Similar to Jeffreys' interpretation of Bayes factor (Jeffreys, 1961), we use a ratio larger than 3 (e.g., $\alpha_H = 1 - 3 \times$ mean of beta hyperprior) as moderate-to-strong evidence. Because the marginal posterior probabilities for main effects are inflated due to the constrained prior structure, a small value of $\alpha_L$ should be used to reduce the false discovery among main effects. Yet, $\alpha_L$ does not affect the decision of main effects whose interactions are selected.

### 2.5 *Prediction*

Let $M_l (l = 1, \ldots, L)$ be a model among $L$ models obtained by varying $\alpha_L$ and $\alpha_H$, and $\boldsymbol{\theta}_l = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}, \sigma\}$ be the vector of parameters for model $M_l$. To predict the probability that $y_{\text{new}}$, given treatment and patient characteristics data ($x_{\text{new}}$), belongs to a predefined desirable outcome region $\Delta$, we use the posterior predictive probability $\Pr(y_{\text{new}} \in \Delta \,|\, \text{Data}, x_{\text{new}}) = \sum_{l=1}^{L} \Pr(y_{\text{new}} \in \Delta \,|\, M_l, \text{Data}, x_{\text{new}}) \Pr(M_l \,|\, \text{Data})$. The conditional posterior probability $\Pr(y_{\text{new}} \in \Delta \,|\, M_l, \text{Data}, x_{\text{new}})$ in the summation is given by $\Pr(y_{\text{new}} \in \Delta \,|\, M_l, \text{Data}, x_{\text{new}}) = \int \Pr(y_{\text{new}} \in \Delta \,|\, \boldsymbol{\theta}_l, x_{\text{new}}) p(\boldsymbol{\theta}_l \,|\, M_l, \text{Data}) \mathrm{d}\boldsymbol{\theta}_l$.

Analytically evaluating the summation and integral is difficult due to the large model space. An immediate Monte Carlo estimate can be calculated as follows. We obtain draws of the parameters $\boldsymbol{\theta}_l^{(1)}, \ldots, \boldsymbol{\theta}_l^{(T_l)}$ from model $M_l$. For each set of parameters $\boldsymbol{\theta}_l^{(t)}$, $t = 1, \ldots, T_l$, we calculate the $\Pr(y_{\text{new}} \in \Delta \,|\, \boldsymbol{\theta}_l^{(t)}, x_{\text{new}})$ from the multivariate regression model in equation (1). The estimated probability of a new observation $y_{\text{new}}$ belongs to set $\Delta$ is $\Pr(y_{\text{new}} \in \Delta \,|\, \text{Data}, x_{\text{new}}) \approx \frac{1}{L} \sum_{l=1}^{L} \frac{1}{T_l} \sum_{t=1}^{T_l} \Pr(y_{\text{new}} \in \Delta \,|\, \boldsymbol{\theta}_l^{(t)}, x_{\text{new}})$. Note that if one model is preferred over Bayesian model averaging, $L$ can be set to 1.

To decide which treatment is best for a given patient with covariates $x_{\text{new}}$, we computed posterior predictive probabilities of superiority for the three treatments (A, F, and G). For each posterior sample of the model parameters $\boldsymbol{\theta}_A^{(t)}$, $t = 1, \ldots, T$, compute $I_A^{(t)} = \{\Pr(y_{\text{new}} \,|\, \boldsymbol{\theta}_A, x_{\text{new}}) > \Pr(y_{\text{new}} \,|\, \boldsymbol{\theta}_F, x_{\text{new}})$ and $\Pr(y_{\text{new}} \,|\, \boldsymbol{\theta}_A, x_{\text{new}}) > \Pr(y_{\text{new}} \,|\, \boldsymbol{\theta}_G, x_{\text{new}})\}$, and analogously compute $I_F^{(t)}$ and $I_G^{(t)}$. Averaging these indicators over the $T$ posterior samples would yield relative probabilities for superiority $p_A = \frac{1}{T} \sum_{t=1}^{T} I_A^{(t)}$ (and likewise, $p_F$ and $p_G$).

## 3. Simulation Studies

### 3.1 *Designs*

To investigate the performance of the variable selection, the prediction of multiple non-Gaussian outcomes, and the effect of sample size on our proposed method, we performed a series of simulation studies. For each sample size ($n = 30$ or 200), we set the number of potential predictors $p \equiv 55$, which consisted of 10 main effects and 45 pairwise interaction terms. We assume that all of the 10 main effects are independent and from a $N(0, 1)$ distribution.

Three responses were simulated: two binary outcomes (with probit link) and one survival outcome (with a lognormal link). Among the 55 potential predictors, the true predictors and their coefficients were set as $0.3 + X_1 + X_2 + X_3 + X_4 + X_1 X_2 + X_1 X_4$ for the first binary outcome, $-0.3 + X_2 + X_3 + X_4 + X_5 + X_2 X_3 + X_4 X_5$ for the second binary outcome, and $0.3X_3 + 0.3X_4 + 0.3X_5 - 0.2X_3 X_4 + 0.2X_4 X_5$ for the survival outcome. We drew a $n \times 3$ matrix of latent variables $\boldsymbol{Z}$ from the trivariate normal density (1) given these true predictors. The parameters in the variance–covariance matrix of $\boldsymbol{Z}$ in equation (1) were set as $\rho_1 = 0.1$, $\rho_2 = 0.3$, $\rho_3 = 0.7$, and $\sigma = 0.3$. For each $i$th simulated observation ($i = 1, \ldots, n$), if $z_{ik} > 0$, $k = 1, 2$, we set $y_{ik} = 1$ and $y_{ik} = 0$ otherwise. This setting generated $y_1 = 1$ approximately 50% and $y_2 = 1$

**Table 2**
*True and estimated* $\Pr(y_{\text{test}} \in \Delta)$ *and 90% CPI coverage*

| | | Highest joint posterior* | | | | $\overline{\text{FDR}}$** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $n = 200$ | | $n = 30$ | | $n = 200$ | | $n = 30$ | |
| ID | True | Est.(SE) | Cvg. | Est.(SE) | Cvg. | Est.(SE) | Cvg. | Est.(SE) | Cvg. |
| 1 | 0.056 | 0.074 (0.026) | 0.85 | 0.137 (0.096) | 0.71 | 0.055 (0.023) | 0.88 | 0.079 (0.088) | 0.80 |
| 2 | 0.311 | 0.284 (0.063) | 0.86 | 0.229 (0.134) | 0.76 | 0.278 (0.074) | 0.82 | 0.227 (0.200) | 0.67 |
| 3 | 0.524 | 0.499 (0.122) | 0.86 | 0.402 (0.213) | 0.73 | 0.514 (0.139) | 0.82 | 0.481 (0.302) | 0.70 |
| 4 | 0.757 | 0.744 (0.031) | 0.89 | 0.491 (0.140) | 0.48 | 0.754 (0.028) | 0.94 | 0.606 (0.212) | 0.77 |
| 5 | 0.933 | 0.905 (0.039) | 0.90 | 0.645 (0.170) | 0.42 | 0.919 (0.033) | 0.95 | 0.819 (0.145) | 0.81 |

*Select model with the highest joint posterior probability.
**Select variables satisfying the $\overline{\text{FDR}}$ criteria.

approximately 40%. For the survival outcome $y_3$, we assume log-normal distribution and noninformative censoring. Random deviates were generated independently from the Unif (0, 4.5) distribution. If $y_3 = \exp(z_3)$ was larger than the corresponding uniform deviate, then that patient was censored at the time of that uniform deviate. This yielded approximately 30% censoring, which matches the percentage of censoring in the colorectal cancer data. We generated 100 replications for each sample size ($n = 30$ and 200).

For all the simulated data sets, we applied the MBSI method with $c = 10$ and $\tau = 0.05$. For the hyperparameters in the beta prior, the mode was set to $3/p$ and the mean was 20% larger than the mode. The length of the MCMC chain was set to be 20,000, from which the first 5000 iterations were discarded. The significance levels ($\alpha_L$, $\alpha_H$) were set to (0.1, 0.2) and (0.3, 0.8), respectively, for the sample sizes 200 and 30. When there was a sufficiently large sample size ($n > p$), the posterior distribution was not sensitive to the prior assumptions. The setting ($\alpha_L = 0.1$, $\alpha_H = 0.2$) led to at most 20% falsely discovered predictors. However, when the sample size was small ($n < p$), the marginal posterior distribution was flattened and sensitive to the prior assumptions. We assumed that on average the prior probability to be a true predictor is 0.065 (the mean chosen for the beta prior). The setting $\alpha_L = 0.3$ and $\alpha_H = 0.8$ when $n = 30$ resulted in an average posterior probability of at least 0.2 among the selected regressors. The average posterior to prior probability ratio for the selected regressors was about 3.

### 3.2 *Prediction Evaluation*

We evaluated prediction using the posterior predictive distributions as follows. Let $y_{test}$ and $x_{test}$ be the simulated outcomes and covariates independent from the training set. We defined $\Delta$ as requiring both binary outcomes be equal to 1 and the survival outcome be larger or equal to 12 months. Let $\Pr(y_{test} \in \Delta)$ (calculated under model (1)) be the true probability. For each replication, we obtained the posterior predictive distribution $p(y_{test} \in \Delta \mid x_{test}, \boldsymbol{\theta}_l)$. If the $\Pr(y_{test} \in \Delta)$ is within 90% of central posterior interval (CPI), we set coverage = 1. The average coverage across all the 100 replications gives the estimate for the 90% coverage for that test case. Five test cases that were scattered in the space of design points were simulated to demonstrate the performance of proposed method.

The true probability $\Pr(y_{test} \in \Delta)$, the estimated mean and standard error of posterior predictive probability, and the 90% CPI are shown in Table 2. Two model selection approaches are compared in the table. One selects a model with the highest joint posterior probability, while the other selects a model that satisfies the $\overline{\text{FDR}}$ criteria. Graphical illustration of the comparison between the two modeling approaches can be found in Web Figure 1.

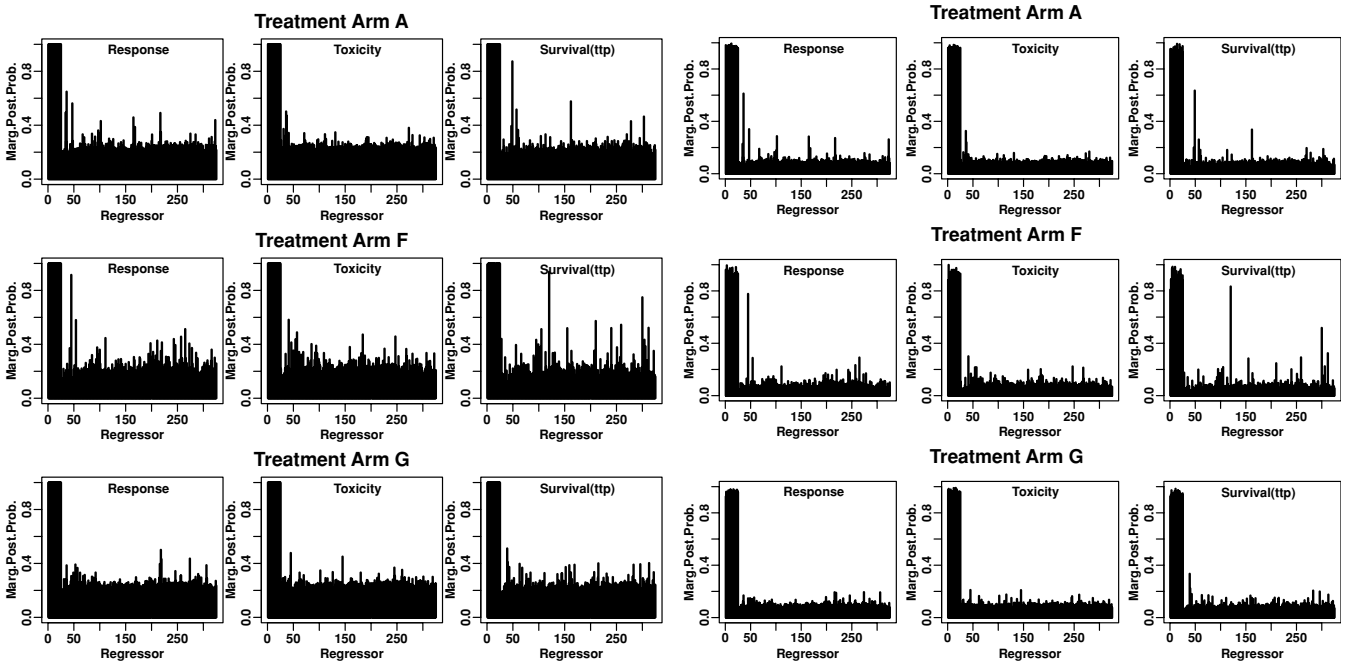
When the sample size equals 200, there was no significant difference between the two model selection methods in terms of estimated mean and coverage. When the sample size dropped to $n = 30$, the prediction uncertainty increased greatly. Test cases 4 and 5 were underestimated using the joint posterior model selection method. The cause of this biased estimation was likely due to the exclusion of the important predictors. The joint posterior probabilities were dominated by the prior distribution, which favors parsimonious models. The predictions for cases 4 and 5 were slightly improved when using the $\overline{\text{FDR}}$ criteria.

## 4. Colorectal Cancer Study

### 4.1 *Preliminary Analysis*

Now we return to the colorectal study described in the Introduction. We dichotomized the 23 biomarkers to mutant (0) or wild type (1). The endpoints observed for each patient in this trial were an indication of toxicities, tumor response, and time to tumor progression (TTP). Several types of toxicities were monitored, e.g., nausea, dehydration, neutropenia, vomiting, diarrhea, febrile neutropenia, and paresthesia. We chose the maximum grade among all types of toxicities and set a binary outcome toxicity (TOX) to 1 if the maximum grade was 4 or 5, which represents life threatening or fatal toxicity on a 5 point scale (National Cancer Institute Common Toxicity Criteria Version 2), and 0 otherwise. If a patient had at least two consecutive complete or partial tumor regressions, the binary outcome tumor response (RSP) was set to 1 and 0 otherwise. The third outcome was the TTP, which was a censored continuous variable. The total number of regressors is 325: 25 main effects (23 biomarkers plus AGE and SEX) plus 300 two-way interactions.

Preliminary investigation of the multiple endpoints (RSP, TOX, and TTP) indicated that they were not independent of each other, as there were positive correlations between RSP and TTP across all the treatment groups. Patients who had a




(a) mean=.25, mode=.2 (b) mean=.1, mode=.05

**Figure 1.** Marginal posterior distributions of the regressors by treatment arms in the colorectal cancer study for the sensitivity analysis of beta priors.

confirmed tumor response consistently had a longer progression time relative to those who had no confirmed tumor response. To check the log-normal assumption for the survival model, we fit a parametric log-normal model to the failure times with no predictors using the `survreg` function in `R`. Based on the fit, we plotted the residuals (See Web Figure 2). In all three arms, log-normality appeared to be a tenable assumption.

### 4.2 *Model Building*

We applied the proposed MBSI method on each treatment arm and chose $c = 10$ and $\tau = 0.05$ for the mixture normal prior. This implied that if a coefficient was less than about $2\tau = 0.1$, we were comfortable excluding that regressor from the model. For the hyperparameters in the beta prior, the mode was set to 0.2 and mean was 0.25, which is 25% larger than the mode.

The estimated marginal posterior distributions of the regressors by treatment arms are shown in Figure 1a. The first 25 regressors are main effects and the rest are interaction terms. All the main effects show very high marginal posterior probabilities. The reason is that the marginal probability of a main effect consists of two sources: one is from itself, the other is from its interactions. The more interaction terms, the more the marginal probabilities for the main effects will be inflated. When $n > p$, there is more separation between signal and noise. See Web Figure 3 for the marginal posterior probabilities in our simulation study. However, when $n < p$, the beta hyperprior plays an important role. When the anticipated proportion of true predictors (the mean of beta hyperprior) was set to be 0.25, the marginal posterior probabilities for interactions were influenced by this prior accordingly (Figure 1a). To analyze the sensitivity to the choice of beta prior, we used another set of hyperparameters with mode equal to 0.1 and mean 0.05. This setting kept the basic shape of beta distribution unchanged, but the distribution is flatter and skewed more toward smaller values, suggesting that a priori there were fewer true predictors and we would be more uncertain about the number of the true predictors. The estimated marginal posterior probabilities of the interaction terms (Figure 1b) indicated that the posterior probabilities are sensitive to the prior, which is not a surprise as $n < p$ in each treatment group. However, for those interaction terms whose estimated posterior probabilities of being selected are above average, the ordering from the highest to the lowest probabilities is very similar from either settings of beta prior. There is much uncertainty about most regressors, yet the strong signals from several of the interactions are worthy of further investigation.

To select the important predictors, we set the $\overline{\text{FDR}}$ threshold $\alpha_L = 0.001$ and $\alpha_H = 0.3$ for the main effects and interactions, respectively. The reason to set a low $\alpha_L$ in this case is to avoid selecting too many main effects, whose marginal probabilities might be inflated by the large number of interaction terms. The variable selection result was not sensitive to the choice of $\alpha_H$ in a range from 0.2 to 0.5. Table 3 shows the estimated coefficients for the selected predictors. The trace plots for the parameters in the model were provided as supplementary material Web Figures 4–6.

Note that the estimated coefficients were not significantly different from 0 for some main effects. For example, in arm F the estimated mean and standard error (SE) of the coefficient were 0.07(0.12) for marker 11 and 0.14(0.14) for marker 23. However, their interaction had estimated mean and SE

**Table 3**
*Est.(SD) of coefficients for the selected predictors*

| | Arm A ($n = 115$) | | | Arm F ($n = 292$) | | | Arm G ($n = 106$) | | |
|---|---|---|---|---|---|---|---|---|---|
| Var | RSP | TOX | TTP | RSP | TOX | TTP | RSP | TOX | TTP |
| Int. | −0.55 (0.17) | −1.60 (0.37) | 4.95 (0.16) | 0.11 (0.15) | −0.37 (0.11) | 5.85 (0.16) | −0.33 (0.17) | −1.51 (0.53) | 5.41 (0.15) |
| †Age | | | 0.13 (0.09) | −0.21 (0.10) | | | | 0.24 (0.14) | |
| Sex | | | | −0.12 (0.16) | −0.59 (0.15) | | −0.22 (0.23) | −0.49 (0.31) | |
| M1 | | | | | | | | 0.02 (0.30) | |
| M2 | | 0.51 (0.37) | | | | | | | |
| M3 | | | | | | 0.15 (0.14) | | 0.44 (0.38) | |
| M5 | 0.60 (0.23) | | 0.56 (0.21) | −0.98 (0.26) | | | | | |
| M7 | −0.38 (0.30) | 0.05 (0.46) | | | | | | | −0.46 (0.33) |
| M8 | −0.37 (0.24) | | | | | 0.59 (0.23) | | | |
| M10 | 0.35 (0.27) | | | | −0.40 (0.20) | | | −0.94 (0.49) | |
| M11 | | | | | | 0.07 (0.12) | | | |
| M12 | | | | | 0.10 (0.28) | | | | |
| M13 | | | 0.15 (0.23) | | | −0.05 (0.11) | | | |
| M14 | | | | | | | | 0.73 (0.34) | |
| M16 | | | | | | −0.35 (0.13) | | | |
| M17 | | 0.14 (0.55) | | | | | | −0.58 (0.40) | |
| M19 | | | | −0.16 (0.13) | | −0.5 (0.15) | | | |
| M21 | | | | | | | | 0.21 (0.25) | −0.10 (0.21) |
| M23 | | | 0.22 (0.23) | | | 0.14 (0.14) | −0.04 (0.31) | −0.09 (0.39) | |
| Age ∗ M19 | | | | 0.49 (0.13) | | | | | |
| Age ∗ M23 | | | −0.69 (0.23) | | | | | | |
| Sex ∗ M5 | | | | 0.83 (0.32) | | | | | |
| M3 ∗ M8 | | | | | | −0.88 (0.23) | | | |
| M5 ∗ M13 | | | −0.90 (0.30) | | | | | | |
| M8 ∗ M13 | | | | | | 0.33 (0.20) | | | |
| M11 ∗ M23 | | | | | | −0.65 (0.28) | | | |
| M16 ∗ M19 | | | | | | 0.52 (0.17) | | | |

†Variable age was scaled.
The cell is blank if the variable was not selected for an outcome.

of −0.65(0.28), which was significantly different from zero. This demonstrates that even though main effects were not significant, their significant joint effect was detected by the MBSI method.

In other examples (i.e., marker 10 and marker 12 in arm F), the coefficients of main effects were not significantly different from zero, and no interactions were selected. This may have one of the three possible explanations. First, fitting all the important variables that are selected due to their high marginal probabilities into a single multiple regression model does not guarantee that all of the coefficients are still significant. The multicollinearity among some important variables may lead to the nonsignificant coefficients. Second, the threshold for $\overline{\text{FDR}}$ procedure could be too low, leading to a high FDR. In our analysis, the threshold for the main effects was set very close to zero, which rules out this possibility. Third, the estimated high marginal probability of a main effect could be due to the falsely selected interactions during the MCMC procedure. A possible remedy for this problem is to apply the MBSI method on the main effects only and review the marginal probabilities of the main effects in question.

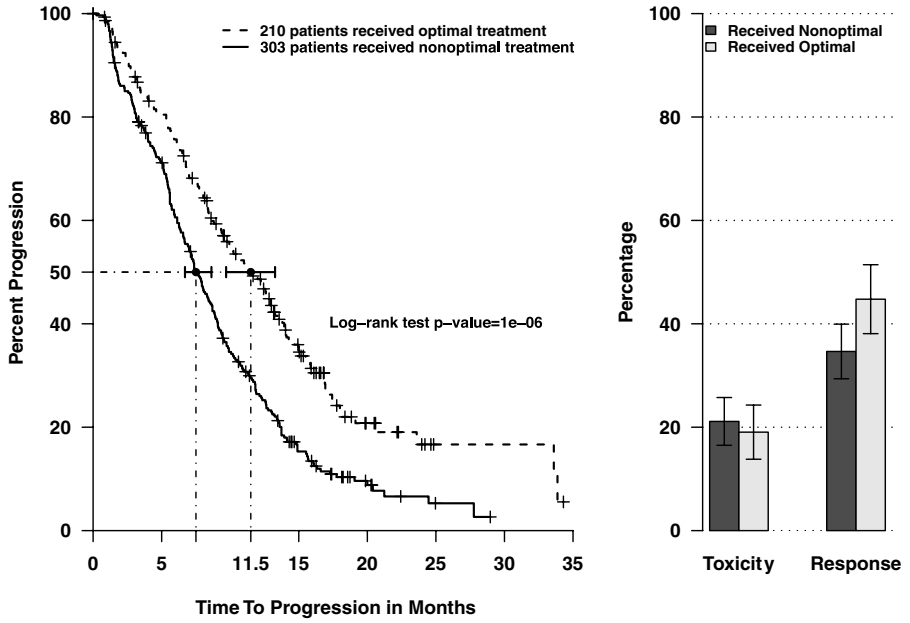
The estimated variance components are presented in Web Table 1. A positive correlation between RSP and TTP ($\rho_2$) was found in all the study arms. Across all three arms, arm F resulted in the largest model size (Table 3), which is defined as the total number of distinctive regressors in a multivariate model allowing different regressors for different outcomes. The model size will tend to increase with the size of sample, because the prediction variance will usually be reduced with increased sample size (Miller, 2002).

### 4.3 *Comparing Treatments*

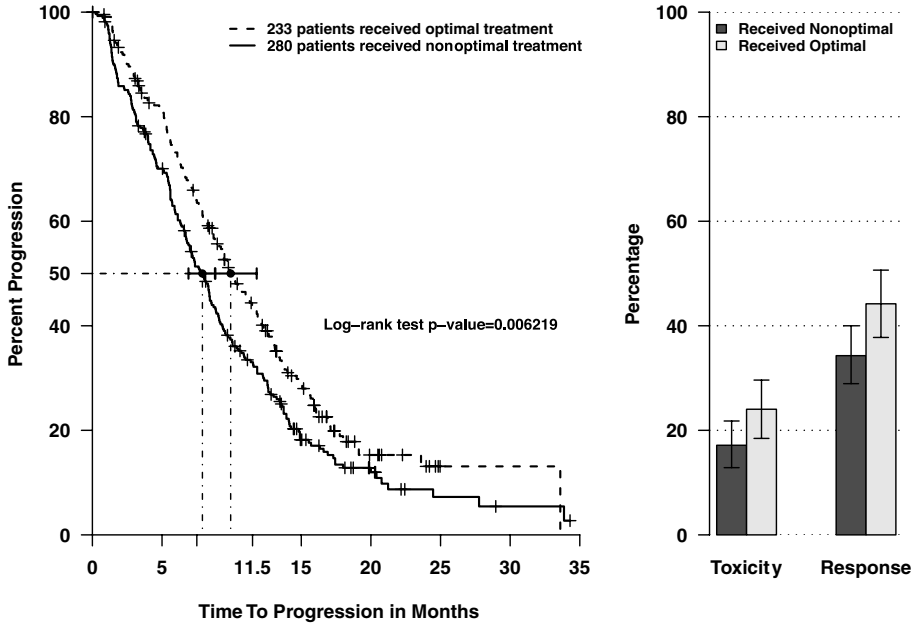
To compare the three treatment regimens, we defined a favorable region ($\Delta$) of treatment outcomes as a union of $\Delta_1$ (RSP = 0 or 1, TOX = 0, and TTP > 365) and $\Delta_2$ (RSP = 1, TOX = 0 or 1, and TTP > 730). The definition of $\Delta$ can be very flexible and can change from study to study. A favorable treatment outcome often represents a trade-off between efficacy and toxicity: if a cancer treatment has high efficacy, sometimes patients are willing to accept more toxicity.

In this study, 18 variables, including AGE, were selected. If we select the design points of the continuous variable AGE at age 50 and 70 (20th and 80th percentile of age distribution), together with the other 17 binary variables, the total possible number of different configurations of patient profiles is $2^{18} = 262{,}144$. For each of these 262,144 hypothetical or future patients, we can predict the probability of outcome in region $\Delta$ given treatment A, F, or G.

In general, the experimental treatment F had a slightly higher probability of achieving an outcome in the favorable region $\Delta$ compared to the standard treatment A, whereas the treatment G had a lower probability (See Web Figure 7). Under treatment A, patients with younger age, mutant markers 5 and 23, and wild-type marker 13 had greater than 50%




(a) 10-fold CV with $c = 10$




(b) Nested 10-fold CV

**Figure 2.** Comparing treatment outcomes of patients who received optimal treatment versus those who received nonoptimal treatment in the colorectal cancer study.

chance to achieve the outcome region $\Delta$. Patients with mutant marker 7 have worse outcome than those with wild-type marker 7 given treatment G.

### 4.4 *Choice of Tuning Parameter c*

Because we are dealing with the $n < p$ situation in this application, further sensitivity analyses of model size were done using different settings of the tuning parameters $c$ and $\tau$ (Web Table 2). The range of $c$ where the predictive performance is invariant is narrower in our application (with $n < p$) than that recommended by George and McCulloch (1993). Possible explanations of this phenomenon are the small transition probabilities and sample size, discussed in Web Appendix A.

Because of the strong influence of tuning parameter $c$, we used a 10-fold cross-validation (CV) approach to choose $c$. We randomly divided data into 10 parts, stratified on treatment arms. In each of 10 subsets, one was set aside as testing data and the remaining nine were used as training set for modeling with different values of $c$. Optimal treatments for the testing set were predicted (see Section 2.5 for the calculation) using the model trained from the training set. We then separated the 513 patients into two groups: group 1 consisted of patients who received nonoptimal treatment; group 2 consisted of patients who received optimal treatment. When $c = 10$, the survival curve and percentages of toxicities and responses of each group are shown in Figure 2a. The median time to progression of group 2 was 11.5 months which was 53% greater than that of group 1 (7.5 months). Figures with other $c$ values are shown in Web Figures 8–10. The predictive performance with $c = 5$ (Web Figure 8) was less than that of $c = 10$. This is likely due to the smaller model size, which left out a few important predictors. The predictive performance with $c = 20$ (Web Figure 9) was less than that of $c = 10$. This is probably due to the large model size, which led to model overfitting. Hence, $c = 10$ was chosen to have the best prediction performance.

### 4.5 *Model Validation*

To further validate whether or not the proposed individualized therapy model results in any difference in treatment effect, the 10-fold method in Section 4.4 is not appropriate because it used all the data for selecting $c$. Therefore, to obtain an assessment of prediction, a nested 10-fold CV was performed. We again divided data into 10 parts. For the 9/10 parts that were used for model building, we did inner 9-fold CV to select $c$ – building models for a grid of $c$ values ($c = 5$, 10, 15, 20) on the 8/9 parts and choosing the one that performed best in predicting the held-out part. Then, whichever $c$ was the best, it was used on predicting the 1/10 part that was not used in selecting $c$. This was repeated for each of the ten 9/10 splits resulting in a valid assessment of prediction shown in Figure 2b.

The statistical testing and confidence intervals in Figure 2 are for illustration only, and the significance has to be interpreted with caution. Lusa et al. (2007) reported that the CV process results in inflated testing type I error rates. As pointed out by one referee, the only way to *truly* assess the value of individualized therapy would be to perform a prospective trial in which the patients were treated based on an individualized plan.

## 5. Concluding Comments

In this work, we proposed a multivariate Bayesian regression model for individualizing cancer treatments. We have addressed three issues: comparing the treatments using a quantitative score derived from predicting the multiple endpoints, modeling jointly non-Gaussian outcomes, and building a regression model with the selection of interactions. The endpoints considered were categorical and censored continuous variables, which are very common in most phase III clinical trial settings. The MBSI procedure which incorporates the selection rule for interaction terms by controlling posterior expected $\overline{\text{FDR}}$ was implemented. This particular selection rule increased the power of detecting interactions, which becomes more and more important in defining useful comprehensive models for complex diseases.

Our simulation study suggests the feasibility of predicting the multiple non-Gaussian outcomes simultaneously. Although the categorical outcomes in the current model were assumed to be binary outcomes, it is straightforward to extend the proposed approach to multilevel outcomes. The survival outcome in the colorectal cancer study was assumed to follow a log-normal accelerated failure time model for the simplicity of computation. We suggest checking this assumption before implementation in the data analysis.

Regarding the priors, although spike and slab mixture priors for variable selection were proposed and applied in the literature, we consider the narrow normal priors for nonsignificant regressors as proposed by George and McCulloch (1993) to be more practical. An important issue is the choice of $c$. We used CV type of methods to choose $c$ in this application. These are all fairly computationally intensive methods. Further theoretical development is needed in this area.

When applying the MBSI method in the colorectal cancer study, we modeled each arm separately. Because different treatments work through different mechanisms, the possibility of treatment-biomarker interactions exists. Separate models eliminate those interaction terms and result in a less complex model structure and faster computation. As pointed out by one referee, it would be better if the modeling was not completely separate, but rather borrowed strength across treatments in estimating the covariance parameters. A hierarchical model component could be introduced that allowed each treatment to have its own covariance parameters. Nonetheless, the increased model complexity due to the additional level is beyond the scope of this article.

## 6. Supplementary Materials

Web Tables and Figures referenced in Sections 3.2 and 4.1–4.4 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

## Acknowledgements

The authors thank the editor, the associate editor, and the referees for their comments that considerably improved this article. Part of the work was supported by the National Institutes of Health through Karmanos Cancer Institute Support Grant (5 P30 CA022453).

## References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88,** 669–679.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics* **32,** 870–897.

Barnard, J., McCulloch, R., and Meng, X. L. (2000). Modelling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica* **10,** 1281–1311.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B* **60,** 627–641.

Chen, M. H. and Dey, D. K. (2003). Variable selection for multivariate logistic regression models. *Journal of Statistical Planning and Inference* **111,** 37–55.

Chen, W., Ghosh, D., Raghunathan, T. E., and Sargent, D. J. (2008). A false-discovery-rate-based loss framework for selection of interactions. *Statistics in Medicine* **27,** 2004–2021.

George, E. I. (1999). Discussion of Bayesian model averaging and model search strategies by M.A. Clyde. In *Bayesian Statistics*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds), vol. 6, 175–177. Oxford: Oxford University Press.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88,** 881–889.

Goldberg, R. M., Sargent, D. J., Morton, R. F., Fuchs, C. S., Ramanathan, R. K., Williamson, S. K., Findlay, B. P., Pitot, H. C., and Alberts, S. R. (2004). A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *Journal of Clinical Oncology* **22,** 23–30.

Jeffreys, H. (1961). *The Theory of Probability*, 3rd edition. Oxford: Oxford University Press.

Lusa, L., McShane, L. M., Radmacher, M. D., Shih, J. H., Wright, G. W., and Simon, R. (2007). Appropriateness of some resampling-based inference procedures for assessing performance of prognostic classifiers derived from microarray data. *Statististics in Medicine* **26,** 1102–1113.

McLeod, H. L. and Murray, G. I. (1999). Tumor markers of prognosis in colorectal cancer. *British Journal of Cancer* **79,** 191–203.

Milano, G. and McLeod, H. L. (2000). Can dihydropyrimidine dehydrogenase impact 5FU-based treatment? *European Journal of Cancer Prevention* **36,** 37–42.

Miller, A. J. (2002). *Subset Selection in Regression*, 2nd edition. London: Chapman & Hall.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models*. New York: McGraw-Hill.

Sha, N., Tadesse, M. G., and Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22,** 2262–2268.

*Received March* 2007. *Revised August* 2008.
*Accepted September* 2008.