Journal of Biopharmaceutical Statistics, 24: 110–129, 2014

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2013.856026

Taylor & Francis
Taylor & Francis Group

A BAYESIAN APPROACH TO SUBGROUP IDENTIFICATION

James O. Berger[1], Xiaojing Wang[2], and Lei Shen[3]
[1]*Department of Statistical Science, Duke University, Durham, North Carolina, USA*
[2]*Department of Statistics, University of Connecticut, Storrs, Connecticut, USA*
[3]*Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana, USA*

This article discusses subgroup identification, the goal of which is to determine the heterogeneity of treatment effects across subpopulations. Searching for differences among subgroups is challenging because it is inherently a multiple testing problem with the complication that test statistics for subgroups are typically highly dependent, making simple multiplicity corrections such as the Bonferroni correction too conservative. In this article, a Bayesian approach to identify subgroup effects is proposed, with a scheme for assigning prior probabilities to possible subgroup effects that accounts for multiplicity and yet allows for (preexperimental) preference to specific subgroups. The analysis utilizes a new Bayesian model selection methodology and, as a by-product, produces individual probabilities of treatment effect that could be of use in personalized medicine. The analysis is illustrated on an example involving subgroup analysis of biomarker effects on treatments.

Key Words: Bayesian analysis; Model uncertainty; Multiplicity; Subgroup analysis.

1. INTRODUCTION

Traditionally, clinical trials have primarily been concerned with comparing treatments on an entire population. But often it is also important to determine whether there are differential treatment effects on subpopulations. For instance, it may be of interest to determine whether a drug has a different effect on young patients than on old patients, or whether patients who carry a genetic mutation in a cytochrome enzyme would respond more actively to a new therapy than other patients.

Temple and Ellenberg (2000) mentioned that approved treatments for many conditions often fail to be effective in follow-up clinical trials and attribute this, in part, to the heterogeneity of effectiveness among subgroups of patients, who often have distinctive genetic makeup, clinical profile (e.g., medical history, disease severity), demographics (e.g., sex, age), and social or environmental factors (e.g., smoking habits). In such situations, determining the subgroups for which there are substantial benefits might salvage some investigational therapeutic agents and

Received September 10, 2013; Accepted September 20, 2013
Address correspondence to James O. Berger, Department of Statistical Science, Duke University, 221 Old Chemistry Building, Durham, NC 27708-0251, USA; E-mail: berger@stat.duke.edu

improve overall success in drug development (Woodcock, 2007). Moreover, even if the treatment does exhibit overall significance, there may be considerable differences of effectiveness in subpopulations (Wang et al., 2007) that would be important to understand for clinical practice. Indeed, there is an increasing trend (or at least wish) to consider tailored treatments in the pharmaceutical industry. The key to tailored therapeutics is to identify subgroups that are more likely to respond to a drug (Lipkovich et al., 2011).

While the appeal of subgroup analyses is undeniable, there are a number of concerns with the approach. The main concerns are (e.g., Berry, 1990; Pocock et al., 2002; Cui et al., 2002; Lagakos, 2006; Wang et al., 2007):

1. *Multiplicity*: When multiple subgroup analyses are performed, the issue of multiple testing arises. Standard adjustments for multiplicity, such as the Bonferroni correction, can be utilized to correct for multiple testing, but can give up too much power when the test statistics are highly dependent, as is virtually always the situation in subgroup analyses. There are, of course, more powerful methods based on the bootstrap and other techniques.
2. *Post hoc analyses (unplanned analyses or data dredging)*: Given a plethora of baseline covariates and the tendency not to have prespecified subgroups, there are a large number of possible subgroups that may be analyzed in a post hoc manner. Proper multiplicity adjustment for post hoc subgroup analyses is extremely difficult due to uncertainty in the number of tests that were actually performed.
3. *Lack of power*: Most studies recruit just enough participants to ensure sufficient statistical power to detect an overall main effect, if one is present. The power for detecting differential effects in subgroups is then inherently lower, and this power is further reduced by the needed multiplicity corrections.

In addition to these primarily statistical concerns, there are also issues concerning the interpretation of findings from subgroup analyses. Pocock et al. (2002) advocated assessing biological plausibility along with consideration of the statistical strength of evidence.

A Bayesian approach to subgroup analysis is developed herein and illustrated on a simulated clinical trial data set from Eli Lilly and Company. A two-group randomized clinical trial for the treatment of schizophrenia is considered. The outcome variable, which is continuous, is the reduction from baseline in the Positive and Negative Syndrome Scale (PANSS) Total Score that is commonly used in these studies. The treatment groups are placebo ($T = 0$) and active treatment by an investigational agent ($T = 1$). In an effort to develop tailored therapies, a 32-dimensional covariate vector X of dichotomous biomarkers is measured for each patient at baseline. There is interest in determining whether these covariates have any predictive value related to the disease progression while on placebo (their prognostic effects). But the primary interest centers on understanding whether there are differential treatment effects for subgroups defined by values of these covariates (their predictive effects): that is, whether patients in a subgroup derive a larger benefit from the new treatment compared to the rest of the patients. Using the same X matrix, four distinct outcome variables (y_a, y_b, y_c, and y_d) were generated to represent different scenarios.

Before turning to our approach, two relevant exploratory approaches to identifying subgroup effects are worth mentioning. One common approach is based on fitting a single model, which includes both main effects and interactions for all covariates simultaneously. For a continuous response variable, a general form of this regression-based model is

$$\mathrm{E}(Y) = \mu_0 + \eta T + \boldsymbol{\beta} h(\mathbf{X}) + \boldsymbol{\theta} T \omega(\mathbf{X}) \tag{1}$$

where $h(\mathbf{X})$ reflects a possible baseline (prognostic) effect of the covariates and the main interest would be in the response increment for the treatment-by-covariate interaction, that is, the term $\omega(\mathbf{X})$ in the treatment group. Although an interaction test based on the model just shown partially overcomes the multiplicity concerns, such tests are likely to be underpowered (Pocock et al., 2002). Dixon and Simon (1991) and Simon (2002) overcame some of those difficulties by providing a Bayesian approach to analyze an analog of equation (1) but use the first-order interactions as the last term. They defined a suitable prior distribution for the parameters of the interaction terms and summarized the point and interval estimates for the subgroup-specific treatment effects, as well as posterior probabilities about the effect size of the treatment in each subgroup. Jones et al. (2011) extended Dixon and Simon (1991) and Simon (2002) by including second-order and higher order interaction terms. Hodges et al. (2007) used different variances for different interaction terms and permitted each interaction to shrink to zero. One deficiency of the approach (1) is that only covariates or combinations of covariates included in the function $\omega(\mathbf{X})$ are considered as interactions in the model, which might rule out some important situations.

Another popular approach relies on tree-based methods to identify subgroups in which the outcome significantly differs. It is well known that the tree-based method is an excellent tool for exploring heterogeneous structure, and can be used to find complex and nonlinear relationships among predictors. Ruberg et al. (2010) also argued that clinicians can better understand and more easily apply the results of such methods in clinical practice. Negassa et al. (2005), Su et al. (2008), and Su et al. (2009), among others, utilized the CART algorithm or random forests to recursively partition the data into two subgroups that show the greatest heterogeneity in the treatment. Unlike these approaches aiming at partitioning of the entire covariate space, Ruberg et al. (2010), Foster et al. (2011), and Lipkovich et al. (2011) developed search tools to identify "interesting" regions in the covariate space (e.g., a region of the covariate space where the treatment effect on the response is substantially better than the average treatment), but ignored the rest of covariate space as "uninteresting" in their recursive partitioning algorithm to define subgroups.

Enlightened by these two approaches, we develop a Bayesian model selection approach to identify subgroup effects based on defining treatment models and baseline models using tree-based methods. Each terminal node of the tree used to construct a treatment model defines a subgroup with possible treatment effect. A similar idea is applied to construct a baseline model to indicate whether or not a covariate has any prognostic effect. Enumerating all the possible distinct combinations of treatment models and baseline models constitutes the model space.

Then, in each model, the response is assumed to have a linear regression structure similar to (1), replacing $\omega(\mathbf{X})$ and $h(\mathbf{X})$ accordingly.

The stochastic structure by which the trees are generated defines the prior probabilities of the resulting models, and is the device by which multiplicity is controlled (cf. Scott and Berger, 2010; Westfall et al., 1997). Expert knowledge can be incorporated into the process through elicitation of the prior probabilities, allowing factors believed more likely to be predictive to have greater weight and hence increasing the power to detect them; this still provides valid multiplicity control as long as the elicitation is done preexperimentally. Sivaganesan et al. (2011) and Laud et al. (2013) develop a Bayesian model selection approach to subgroup analyses from a decision-theoretic viewpoint in which the problem is viewed as a series of decisions concerning whether or not to delve more deeply into subgroups. In our approach, we simultaneously consider all possible subgroups to be tested.

This article is organized as follows. Section 2 describes the the model space and response variable. Section 3 presents an approach for specification of the prior probabilities of the models and the prior distributions of parameters in the each model. Section 4 discusses relevant posterior inferences, with illustration on a simulated clinical trial dataset from Eli Lilly and Company. Section 5 presents conclusions.

2. THE OUTCOME MODELS AND MODEL SPACE

2.1. Notation for Models

Bayesian analysis of the subgroup problem will proceed by considering all models arising from a tree splitting process for factors defining subgroups, computing their marginal likelihoods, and determining their posterior probabilities. Each marginal likelihood computation is separate, internal to the specified model.

The general procedure for constructing the subgroup models and baseline models using tree splitting are given in Wang (2012). Here we illustrate the process by considering splitting on just one factor, with some discussion later of splitting on two factors.

Let $\mathcal{F}$ denote the set of factors in the study; for example, sex, age, and genetic biomarkers could be factors. Let $x_{i,j}$ be the value of the jth factor for the ith individual, where $j \in \mathcal{F}$ and $i = 1, \ldots, n$, with n being number of individuals in the sample. In this article we assume that the $x_{i,j}$ take only the values 0 or 1. If at most one factor split is allowed for baseline effect and treatment effect, the linear regression models we consider are of the form

$$Y_i = S_i^{h,j} + B_i^{\ell,k} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

for $i = 1, \ldots, n$, $h = 1, \ldots, 5$, $\ell = 1, 2$, and $j, k \in \{0, \mathcal{F}\}$, where the S and B submodels are defined as follows.

The treatment submodels are

$$S_i^{1,0} = 0,$$
$$S_i^{2,0} = T_i \mu_2,$$

$$S_i^{3,j} = T_i \mu_{3,j} \mathbf{1}_{\{x_{i,j}=0\}} + T_i \mu'_{3,j} \mathbf{1}_{\{x_{i,j}=1\}},$$
$$S_i^{4,j} = T_i \mu_{4,j} \mathbf{1}_{\{x_{i,j}=0\}},$$
$$S_i^{5,j} = T_i \mu_{5,j} \mathbf{1}_{\{x_{i,j}=1\}}, \tag{3}$$

where μ_2 is the mean overall treatment effect (if present), $\mu_{3,j}$, $\mu'_{3,j}$, $\mu_{4,j}$, and $\mu_{5,j}$ are the potential treatment (predictive) effects in the corresponding subgroups defined by factor j, T_i is the indicator of treatment, and $\mathbf{1}_{\{a=c\}}$ equals 1 if a is the same as c.

Similarly, the two possible baseline or prognostic submodels for factor $k \in \mathcal{F}$ are

$$B_i^{1,0} = \mu_1$$
$$B_i^{2,k} = \mu_1 + \beta_k \mathbf{1}_{\{x_{i,k}=0\}}, \tag{4}$$

where μ_1 is the overall mean and β_k is the mean baseline effect for factor k. (The submodel with $\{x_{i,k} = 1\}$ need not be considered, as it can be transformed into the last submodel.)

The total number of models is $2 + 5m + 3m^2$, where m is the number of factors. This follows simply by counting all the possible combinations of models that can be formed as in equation (2) from m factors, accounting for duplicates. In the testbed application introduced in the next section, there are $m = 32$ factors, and thus there are $2 + 5 \times 32 + 3 \times 32^2 = 3234$ distinct models.

2.2. Four Testbed Data Sets

For illustration, the proposed methods will be applied to Eli Lilly and Company testbed data sets. Since the data-generating models are known for these created data sets, it is possible to compare the answers with known truth. The data sets were created to mimic data from clinical trials concerning the development of new therapies for schizophrenia, for which there have been efforts in pharmacogenomics research to identify subgroups that respond well to a treatment (Lavedan et al., 2008; Liu et al., 2012). In all, four data sets were created to resemble what might be observed in such pharmacogenomic studies.

The four data sets share the same predictors, which are 32 single-nucleotide polymorphisms (SNPs). A two-level structure was used for correlations among the predictors. Specifically, the SNPs belong to four genes, with no correlation between SNPs from different genes. The SNPs are named "g1snp01, . . . , g1snp08,g2snp01, . . . , g4snp08" to reflect this structure. Within each gene, the eight SNPs are further broken into four blocks, with two SNPs in the same block more highly correlated than SNPs belonging to different blocks. To fit the notation for the Bayesian methodology, the SNPs "g1snp01, . . . ,g1snp08, g2snp01, . . . ,g4snp08" are ordered and coded as factors from 1 to 32. The response variable is the amount of reduction in PANSS Total Score, the most commonly used endpoint in developing treatments for schizophrenia. In total, 200 subjects were randomized equally to two arms, treatment, and placebo.

The four data sets differ in how the response variables (named "y_a," "y_b," "y_c," and "y_d") were generated, given the predictors. Recall that each model

can have predictive biomarkers, those that indicate subgroups that will exhibit a differential treatment effect, and/or prognostic biomarkers, which indicate a differential baseline response to the disease, but not a differential treatment effect.

- y_a: No predictive biomarker and Factor 31 is a prognostic biomarker, with an overall treatment mean, so the true model was

$$Y_i = \mu_1 + T_i\mu_2 + \beta_{31}\mathbf{1}_{\{x_{i,31}=0\}} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2);$$

  the parameters used were $\mu_1 = 1.85$, $\mu_2 = 6.3$, $\beta_{31} = 6.3$, and $\sigma = 20.76$.
- y_b: Factor 13 is a predictive biomarker and Factor 7 is a prognostic biomarker, so the true model is of the form

$$Y_i = \mu_1 + T_i\mu_{5,13}\mathbf{1}_{\{x_{i,13}=1\}} + \beta_7\mathbf{1}_{\{x_{i,7}=7\}} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2);$$

  the parameters used were $\mu_1 = 1.85$, $\mu_{5,13} = 12.60$, $\beta_7 = 6.30$, and $\sigma = 20.76$.
- y_c: Factors 23 and 27 are both predictive biomarkers, with no prognostic biomarker, so the true model is

$$Y_i = \mu_1 + T_i\mu\mathbf{1}_{\{x_{i,23}=1\}}\mathbf{1}_{\{x_{i,27}=1\}} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2);$$

  the parameters used were $\mu_1 = 5$, $\mu = 25.2$, and $\sigma = 21$. Note that this is only in the model space if two factor splittings are allowed.
- y_d: No predictive or prognostic biomarker; thus, the true model was just the overall effect model

$$Y_i = \mu_1 + T_i\mu_2 + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2);$$

  the parameters used were $\mu_1 = 5$, $\mu_2 = 6.3$, and $\sigma = 21$.

3. SPECIFICATION OF PRIORS WHEN AT MOST ONE FACTOR SPLITTING IS ALLOWED

The outcome variables Y_i, $i = 1, \ldots, n$, can be viewed as arising from a three-stage hierarchical mixture model: (1) the model M_κ is drawn from the model space $\mathcal{M} = \{M_1, \ldots, M_{\mathcal{K}}\}$ according to probabilities $\mathrm{P}(M_1), \ldots, \mathrm{P}(M_{\mathcal{K}})$; (2) the unknown parameter vector of M_κ, denoted $\boldsymbol{\alpha}_\kappa$, is generated from $\mathrm{P}(\boldsymbol{\alpha}_\kappa \mid M_\kappa)$; (3) the y_i's, that is, the realizations of the Y_i's, are generated from $\mathrm{P}(Y_i \mid \boldsymbol{\alpha}_\kappa, M_\kappa)$ given in equation (2). This section is concerned with specification of the priors for stages 1) and 2) when models are restricted to having at most one factor split.

3.1. Specifying Priors on the Model Space

Bayesian multiple testing adjustments take place through choice of the $\mathrm{P}(M_i)$ (see, e.g., Berry, 1990; Scott and Berger, 2010). For the subgroup problem, however, it is not particularly natural to think in terms of these probabilities directly; more natural is to elicit probabilities that factors (e.g., biomarkers) are likely to lead to subgroups with treatment effects. Thus, with m denoting the number of possible

factors in the study, view Factor 1 as the base factor, and define o_i, the *effect odds* of Factor i to Factor 1, to be the prior relative odds that Factor i has an effect compared to Factor 1. Thus if $o_3 = 5$, Factor 3 is viewed as being five times more likely to lead to define a subgroup with a differential treatment effect than would Factor 1. The default choice is, of course, $o_i = 1$. It will be assumed that the same odds apply to both baseline and treatment effects. We also assume here that prior beliefs about these odds are independent, so that the odds can be multiplied when considering splits over two or more factors. More generally, one could provide odds o_{ij} corresponding to pairs of factors, or introduce other dependency structures. For notational convenience define $O_1 = \sum_{j=1}^{m} o_j$ and $O_2 = \sum_{j=1}^{m} o_j^2$.

Also important is what could be called "null control." For objective Bayesian testing of a single hypothesis, this control is commonly exerted by assigning prior probability of $1/2$ to the null hypothesis of no treatment effect. In subgroup analysis, however, there are many models that have "no treatment effect" for some individuals. In equation (3), for instance, the first model is the overall model of no treatment effect, while the fourth and fifth have no treatment effect for the individuals with $X_j = 1$ and $X_j = 0$, respectively. The natural prior probability of "no treatment effect" to consider is thus the overall probability of "no treatment effect" for an individual. Assuming symmetry in the prior probability assignments so that, for example, the fourth and fifth models in equation (3) have equal prior probability, it can be shown that each individual will have the same prior probability of no treatment effect, which will be denoted p_0. Similarly, one can define the prior probability of no baseline effect for an individual; this will be denoted q_0. Natural default choices would be $q_0 = p_0 = 0.5$.

When allowing at most one split for the tree, one must also choose r, the ratio of the prior probability of the overall treatment model to the sum of the prior probabilities of the treatment models with a split. A default choice is $r = 1$. That completes the prior specification, and it can be shown that the implied prior probabilities of baseline submodels and treatment submodels are as follows:

3.1.1. Baseline Submodels. Recall that $B_{\cdot}^{1,0}$ is the null baseline model and $B_{\cdot}^{2,k}$ is the baseline model with a prognostic effect for Factor k. Then the prior probability of $B_{\cdot}^{1,0}$ is equal to q_0 (default being 0.5) and the prior probability of $B_{\cdot}^{2,k}$ is $(1 - q_0)o_k/O_1$ (default is $0.5/m$).

3.1.2. Treatment Submodels. Recall that $S_{\cdot}^{1,0}$ is the null treatment model, $S_{\cdot}^{2,0}$ is the model with a common nonzero treatment effect, and $S_{\cdot}^{3,j}$, $S_{\cdot}^{4,j}$, and $S_{\cdot}^{5,j}$ are the three possible treatment models after splitting on Factor j. Assuming the three treatment models after splitting on j all have equal prior probability, it follows from the elicitations that:

$$\text{The prior probability of } S_{\cdot}^{1,0} = \frac{3(1+r)p_0 - 1}{2 + 3r} \quad \text{(default is 0.4).}$$

$$\text{The prior probability of } S_{\cdot}^{2,0} = \frac{3r(1-p_0)}{2+3r} \quad \text{(default is 0.3).}$$

$$\text{The prior probability of } S_{\cdot}^{h,j} = \frac{(1-p_0)o_j}{(2+3r)O_1} \quad \text{(default is 0.1/m),} \quad h = 3, 4, 5.$$

The elicitations must satisfy the constraint $p_0 > 1/[3(1+r)]$ for these to be proper probabilities. Finally, the prior probability, $\mathrm{P}(M_i)$, of a model will simply be the prior probability of the baseline submodel multiplied by the prior probability of the treatment submodel.

3.2. Priors for Parameters in the Outcome Model

A reasonable choice of priors for unknown parameters in the outcome model (2), for the at most one factor splitting case, is

$$\begin{aligned} \mu_2, \mu_{4,j}, \text{ and } \mu_{5,j} &\sim N(0, v^2); \\ \beta_k &\sim \mathcal{N}(0, \tau^2); \\ \pi(\mu_1, \sigma^2) &\propto \frac{1}{\sigma^2}; \\ \text{for } S_{\cdot}^{3,j}: \quad (\mu_{3,j}, \mu'_{3,j})' &\sim \mathcal{MVN}\left(\mu_g \begin{pmatrix}1\\1\end{pmatrix}, v^2 \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right), \quad \mu_g \sim \mathcal{N}(0, \omega^2), \end{aligned} \tag{5}$$

where ω^2, v^2, and τ^2 are hyperparameters. Note that μ_1 and σ^2 are common to all models and hence can be assigned the usual objective prior. However, the other parameters are not common to all models (and, in particular, none occur in the null model of no treatment effect with zero baseline effect), and thus they must be assigned proper priors. A natural hierarchical exchangeable prior is used for $(\mu_{3,j}, \mu'_{3,j})'$.

There are several possible approaches to dealing with ω^2, v^2, and τ^2. They can, of course, be subjectively specified. Alternatively, they could be viewed as unknown, and dealt with through either a hierarchical Bayes or an empirical Bayes analysis; while this is feasible in the at most one factor split situation, the computations become quickly overwhelming as the number of allowed factor splits increases.

A third option, which could be called "local empirical Bayes," is to choose ω^2, v^2, and τ^2 to maximize each models marginal likelihood

$$\mathrm{P}(\boldsymbol{y} \mid \omega^2, v^2, \tau^2, M_\kappa) = \int P(\boldsymbol{y} \mid \boldsymbol{\alpha}_\kappa M_\kappa)\mathrm{P}(\boldsymbol{\alpha}_\kappa \mid M_\kappa) d\boldsymbol{\alpha}_\kappa, \tag{6}$$

$\boldsymbol{\alpha}_\kappa$ refers to the other parameters in the model, subject to the constraint that they cannot be smaller than σ^2. (The reason for the constraint is to prevent treatment models from collapsing to null models, and is related to the common "unit information" choice of prior variances in testing.) This is computationally simple since (for our problem) a very accurate closed form approximation to equation (6) is available (see Wang, 2012). The result of these computations will be the $P(\boldsymbol{y} \mid M_i)$, used herein to compute the posterior probabilities.

A noteworthy feature of this approach is that it allows each model to maximize its marginal likelihood over a sensible range of prior variances. The resulting model probabilities are thus arguably the largest probabilities in favor of treatment effects that can be obtained while accounting for multiple testing.

4. POSTERIOR INFERENCES FOR SUBGROUP ANALYSES

4.1. Posterior Summaries

Information about unknowns is encoded in the posterior distribution, which consists of the posterior model probabilities

$$\mathrm{P}(M_\kappa \mid \boldsymbol{y}) = \frac{\mathrm{P}(\boldsymbol{y} \mid M_\kappa)\mathrm{P}(M_\kappa)}{\sum_{j=1}^{\mathcal{K}} \mathrm{P}(\boldsymbol{y} \mid M_j)\mathrm{P}(M_j)}, \tag{7}$$

as well as the posterior distributions of unknown parameters in the outcome models. There will typically be a huge number of models and particular subgroups will appear in many models, so it will be necessary to consider marginal features of this posterior.

One marginal posterior quantity of interest is the posterior probability of a treatment effect for each individual, that is, for $i = 1, \ldots, n$,

$$P_i = \sum_\kappa \mathrm{P}(M_\kappa \mid \boldsymbol{y})\mathbf{1}_{\{\mu_{i,\kappa} \neq 0\}}, \tag{8}$$

where $\mu_{i,\kappa}$ is the subgroup mean treatment effect associated with the ith individual in the κth given model. Also available is the individual mean treatment effect, given that the treatment has an effect on an individual, given by

$$\Lambda_i = \sum_\kappa \mathrm{P}(M_\kappa \mid \boldsymbol{y})\bar{\mu}_{\kappa,i}\mathbf{1}_{\{\mu_{\kappa,i} \neq 0\}}/P_i, \tag{9}$$

where $\bar{\mu}_{\kappa,i}$ is the posterior mean of $\mu_{\kappa,i}$. In fact, the entire posterior distribution of each individual's treatment effect is available.

This is of interest in the context of personalized medicine. Indeed, P_i and Λ_i could be termed the *personalized treatment effect probability* and *personalized treatment magnitude mean*, respectively. These are potentially of considerable use by medical professionals in judging the value of the treatment to individual patients.

Another interesting posterior summary is Q_j, the *posterior probability of a nonzero treatment effect for S_j, subgroup j*. The most natural way to define this probability is as the expectation of the individual treatment probabilities over all individuals potentially making up the subgroup. The obvious estimate of this probability is then simply

$$Q_j = \frac{1}{\{\#i \in S_j\}} \sum_{i \in S_j} P_i, \tag{10}$$

the average of the P_i over all individuals in the sample that are in the subgroup. Likewise one can consider Ω_j, the *posterior expected treatment effect size for S_j*, estimated as

$$\Omega_j = \sum_{i \in S_j} \Lambda_i P_i \Big/ \sum_{i \in S_j} P_i. \tag{11}$$

Both Q_j and Ω_j would typically need to be large enough to approve the treatment for use by subgroup j.

Only two of the possible uses of the Bayesian analysis have been highlighted here. From the full posterior distribution that is available, any desired inference could be performed, including construction of confidence sets for unknowns, evaluation of expected losses of decisions, and prediction.

4.2. Analysis of Testbed Data Sets

4.2.1. Posterior Analysis for at Most One Factor Splitting

Posterior model probabilities. Tables 1 through 4 present the model posterior probabilities for the four data sets discussed in section 2.2, assuming at most one factor split is allowed and using the default settings for the prior probabilities. Only the models with posterior probability larger than 0.01 are listed. In each table, the first and third columns indicate whether a factor is used as a predictive or a prognostic factor in the tree splitting process for treatment models and baseline models, respectively. The symbol "–" in those columns implies no factors are used in splitting. The second and fourth columns give the model by specifying the form of the treatment submodel and baseline submodel. The prior and posterior probability for each listed model are given in the last two columns. For convenience, the true model used to generate the data is also indicated in the table caption.

In Table 1, the response data y_a was generated by the 10th model, consisting of the overall treatment effect submodel and the baseline (prognostic) submodel with Factor 31. This model indeed has the largest posterior probability, although it

Table 1 Models with posterior probability > 0.01 for data set y_a

	Predictive factor j	Treatment submodel	Prognostic factor k	Baseline submodel	Prior probability	Posterior probability	Posterior probability using BIC approximation
1	—	$S_{\cdot}^{1,0}$	—	$B_{\cdot}^{1,0}$	0.20000	0.04114	0.10405
2	—	$S_{\cdot}^{2,0}$	—	$B_{\cdot}^{1,0}$	0.15000	0.03364	0.04518
3	16	$S_{\cdot}^{4,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.01190	0.01587
4	31	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.04621	0.06420
5	—	$S_{\cdot}^{1,0}$	15	$B_{\cdot}^{2,k}$	0.00625	0.02839	0.04121
6	—	$S_{\cdot}^{1,0}$	16	$B_{\cdot}^{2,k}$	0.00625	0.02267	0.03270
7	—	$S_{\cdot}^{1,0}$	31	$B_{\cdot}^{2,k}$	0.00625	0.09701	0.14464
8	—	$S_{\cdot}^{2,0}$	15	$B_{\cdot}^{2,k}$	0.00469	0.01777	0.01354
9	—	$S_{\cdot}^{2,0}$	16	$B_{\cdot}^{2,k}$	0.00469	0.01347	0.01017
10	—	$S_{\cdot}^{2,0}$	31	$B_{\cdot}^{2,k}$	0.00469	0.21397	0.17744
11	16	$S_{\cdot}^{3,j}$	31	$B_{\cdot}^{2,k}$	0.00005	0.02181	<0.01
12	31	$S_{\cdot}^{3,j}$	15	$B_{\cdot}^{2,k}$	0.00005	0.01252	<0.01
13	15	$S_{\cdot}^{4,j}$	31	$B_{\cdot}^{2,k}$	0.00005	0.01177	<0.01
14	16	$S_{\cdot}^{4,j}$	31	$B_{\cdot}^{2,k}$	0.00005	0.10689	0.09056
15	31	$S_{\cdot}^{5,j}$	15	$B_{\cdot}^{2,k}$	0.00005	0.06168	0.05106
16	31	$S_{\cdot}^{5,j}$	16	$B_{\cdot}^{2,k}$	0.00005	0.02556	0.02029
17	—	$S_{\cdot}^{1,0}$	4	$B_{\cdot}^{2,k}$	0.00625	<0.01	0.01149

Note. Each row defines a model, indicating whether the model has a predictive and/or prognostic factor, with "–" indicating the absence of such a factor and a number indicating which factor is involved. The definitions of the treatment and baseline submodels are given in equations (3) and (4). (The true model generating the data is the 10th listed here.) BIC = Bayesian Information Criterion.

started with such a small prior probability that the evidence in favor of that model is not conclusive. The model in line 14 also has significant posterior probability, although it is not the correct model, in that factor 16 does not have a predictive effect. The posterior to prior odds here are $0.10689/.00005 = 2138$, indicating that the data (incorrectly) very strongly suggest that factor 16 has a predictive effect; the Bayesian multiplicity correction controls that. Note that when considering a large number of models (here 3234), it is not common for a single model to have posterior probability near 1. Hence of more relevance are the overall posterior summaries discussed in section 4.1. (These are presented later.)

For Table 2 concerning the response data y_b, the correct model is given in line 10 of the first panel. The posterior to prior odds for this model are 593.6, indicating strong data support for this model; it is just not enough to overcome the very small prior probability of the model. For comparison,the large posterior probability models in lines 5 and 2 have posterior to prior odds of 97.6 and 1.17, respectively. The presence of Factor 14 in the table is because Factors 13 and 14 are highly correlated.

Recall that the response data y_c of Table 3 are actually generated from the a model with predictive Factors 23 and 27 and hence cannot be described by a single split model. It is clear from the table that the predictive effect of Factor 27 is dominant, so that Factor 23 does not appear as a predictive factor. Factor 23 does come in rather strongly at line 6 as having a prognostic effect, which is the best that the single factor splitting model can do.

The response data y_d of Table 4 were created using the model in line 2, an overall treatment effect but no predictive or prognostic effects. The data do seem to strongly and incorrectly suggest that Factor 15 is predictive, but again this is controlled by the very small model prior probabilities of the models with that factor being predictive.

Personalized treatment outputs. Table 5 gives the marginal posterior probability of there being a treatment effect and the expected treatment magnitude (given that there is an effect) for the first 20 individuals in the study and for each of the four data sets. The fact that there was some treatment effect in each of the four scenarios means that this probability is never near zero. Only for y_c do we see treatment effect probabilities near 1 for some individuals, presumably those with the right aspect of Factor 27.

Probabilities of subgroup treatment effects. Tables 6 and 7 present the probabilities of subgroup treatment effects, more precisely the marginal posterior probabilities that each factor has a predictive treatment effect. Table 6 shows that the subgroup defined by $X_{13} = 1$ has a probability slightly bigger than 0.7 of having a treatment effect. That $X_{14} = 1$ also seems to have a predictive effect is due to the strong correlation between X_{13} and X_{14}. Table 7 shows that the subgroup defined by $X_{27} = 1$ has probability near 1 of having a treatment effect. Again, the lack of an apparent predictive effect for Factor 23 in Table 7 is likely due to the analysis being based only on single factor splits, and the apparent predictive effect of X_{28} is due to its strong correlation with X_{27}.

Sensitivity analysis. The local empirical Bayes choice of the parameter prior distributions in section 3.2 was rather aggressive, designed to more strongly

Table 2 Models with posterior probability > 0.01 for data set y_b

	Predictive factor j	Treatment submodel	Prognostic factor k	Baseline submodel	Prior probability	Posterior probability	Posterior probability using BIC approximation
1	—	$S_{\cdot}^{1,0}$	—	$B_{\cdot}^{1,0}$	0.20000	0.17341	0.33492
2	—	$S_{\cdot}^{2,0}$	—	$B_{\cdot}^{1,0}$	0.15000	0.17598	0.18152
3	13	$S_{\cdot}^{3,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.02844	0.01152
4	14	$S_{\cdot}^{3,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.01079	<0.01
5	13	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.15224	0.15356
6	14	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.05479	0.05468
7	—	$S_{\cdot}^{1,0}$	7	$B_{\cdot}^{2,k}$	0.00625	0.02293	0.02434
8	—	$S_{\cdot}^{2,0}$	7	$B_{\cdot}^{2,k}$	0.00469	0.02377	0.01352
9	—	$S_{\cdot}^{2,0}$	31	$B_{\cdot}^{2,k}$	0.00469	0.01467	<0.01
10	13	$S_{\cdot}^{5,j}$	7	$B_{\cdot}^{2,k}$	0.00005	0.02968	0.01686
1	—	$S_{\cdot}^{1,0}$	—	$B_{\cdot}^{1,0}$	0.20000	0.13792	0.28170
2	—	$S_{\cdot}^{2,0}$	—	$B_{\cdot}^{1,0}$	0.15000	0.13996	0.15268
3	13	$S_{\cdot}^{3,j}$	—	$B_{\cdot}^{1,0}$	0.00234	0.03393	0.01453
4	14	$S_{\cdot}^{3,j}$	—	$B_{\cdot}^{1,0}$	0.00234	0.01287	<0.01
5	9	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00234	0.01188	0.01203
6	13	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00234	0.18163	0.19374
7	14	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00234	0.06537	0.06899
8	15	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00234	0.01066	0.01063
9	—	$S_{\cdot}^{1,0}$	7	$B_{\cdot}^{2,k}$	0.00938	0.02735	0.03071
10	—	$S_{\cdot}^{2,0}$	7	$B_{\cdot}^{2,k}$	0.00703	0.02836	0.01706
11	13	$S_{\cdot}^{5,j}$	7	$B_{\cdot}^{2,k}$	0.00011	0.05312	0.03191
12	14	$S_{\cdot}^{5,j}$	7	$B_{\cdot}^{2,k}$	0.00011	0.01335	<0.01
1	—	$S_{\cdot}^{1,0}$	—	$B_{\cdot}^{1,0}$	0.20000	0.06660	0.14924
2	—	$S_{\cdot}^{2,0}$	—	$B_{\cdot}^{1,0}$	0.15000	0.06759	0.08089
3	13	$S_{\cdot}^{3,j}$	—	$B_{\cdot}^{1,0}$	0.00852	0.05959	0.02800
4	14	$S_{\cdot}^{3,j}$	—	$B_{\cdot}^{1,0}$	0.00852	0.02260	0.01007
5	13	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00852	0.31895	0.37324
6	14	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00852	0.11479	0.13290
7	—	$S_{\cdot}^{1,0}$	13	$B_{\cdot}^{2,k}$	0.02557	0.01264	<0.01
8	13	$S_{\cdot}^{3,j}$	7	$B_{\cdot}^{2,k}$	0.00029	0.01240	<0.01
9	13	$S_{\cdot}^{5,j}$	7	$B_{\cdot}^{2,k}$	0.00029	0.06784	0.04470
10	14	$S_{\cdot}^{5,j}$	7	$B_{\cdot}^{2,k}$	0.00029	0.01705	0.01088
11	—	$S_{\cdot}^{1,0}$	7	$B_{\cdot}^{2,k}$	0.00682	<0.01	0.01183

Note. Each row defines a model, indicating whether the model has a predictive and/or prognostic factor, with "–" indicating the absence of such a factor and a number indicating which factor is involved. The definitions of the treatment and baseline submodels are given in equations (3) and (4). The first panel is for the default prior probabilities, while the second and third panels are for prior probabilities I and II given in section 4.3. (The true model generating the data is in line 10 of the first panel.)

enable detection of subgroup treatment effects. To see the extent of this effect and to perform a limited sensitivity study, we also did the computations using a standard default Bayesian procedure, namely the Bayesian Information Criterion (BIC) (Schwarz, 1978)

$$\text{BIC} \equiv -2l(\hat{\boldsymbol{\theta}}) + p \log n,$$

where $l(\hat{\boldsymbol{\theta}})$ is the log likelihood of a model M at the parameter mle's, p is the number of parameters in the model, and n is the sample size. As BIC is meant to be an

Table 3 Models with posterior probability > 0.01 for data set y_c

	Predictive factor j	Treatment submodel	Prognostic factor k	Baseline submodel	Prior probability	Posterior probability	Posterior probality using BIC approximation
1	27	$S_{\cdot}^{3,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.12476	0.07242
2	27	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.45455	0.63838
3	27	$S_{\cdot}^{3,j}$	23	$B_{\cdot}^{2,k}$	0.00005	0.03673	0.01269
4	27	$S_{\cdot}^{5,j}$	17	$B_{\cdot}^{2,k}$	0.00005	0.02161	0.01645
5	27	$S_{\cdot}^{5,j}$	18	$B_{\cdot}^{2,k}$	0.00005	0.01750	0.01342
6	27	$S_{\cdot}^{5,j}$	23	$B_{\cdot}^{2,k}$	0.00005	0.13661	0.11232
7	27	$S_{\cdot}^{5,j}$	24	$B_{\cdot}^{2,k}$	0.00005	0.02051	0.01591
8	27	$S_{\cdot}^{5,j}$	25	$B_{\cdot}^{2,k}$	0.00005	0.02926	0.02132
1	27	$S_{\cdot}^{3,j}$	—	$B_{\cdot}^{1,0}$	0.00100	0.06117	0.04119
2	27	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00100	0.22285	0.36310
3	27	$S_{\cdot}^{3,j}$	23	$B_{\cdot}^{2,k}$	0.00002	0.11526	0.04620
4	27	$S_{\cdot}^{3,j}$	24	$B_{\cdot}^{2,k}$	0.00002	0.01564	<0.01
5	27	$S_{\cdot}^{5,j}$	23	$B_{\cdot}^{2,k}$	0.00002	0.42863	0.40887
6	27	$S_{\cdot}^{5,j}$	24	$B_{\cdot}^{2,k}$	0.00002	0.06436	0.05791
7	23	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.01000	<0.01	0.01148

Note. Each row defines a model, indicating whether the model has a predictive and/or prognostic factor, with "–" indicating the absence of such a factor and a number indicating which factor is involved. The definitions of the treatment and baseline submodels are given in equations (3) and (4). The first panel is for the default prior probabilities, while the second is for prior probabilities III, as discussed in section 4.3. (The true model generating the data has predictive effects for both Factor 23 and Factor 27; hence, the true model is not in the set of single split candidates.)

approximation to twice the log of the marginal likelihood, one can convert back to an approximate marginal likelihood via $P(\boldsymbol{y} \mid M) = e^{-\mathrm{BIC}/2}$, and use these in equation (7) to perform the posterior analysis.

The results for the four testbed data sets are given in the last columns of Tables 1 through 4. Overall, the BIC results have similar features to the earlier results, suggesting mostly the same findings. The biggest difference seems to be that BIC significantly increases (almost doubles) the posterior probability of the null

Table 4 Models with posterior probability >0.01 for data set y_d

	Predictive factor j	Treatment submodel	Prognostic factor k	Baseline submodel	Prior probability	Posterior probability	Posterior probability using BIC approximation
1	—	$S_{\cdot}^{1,0}$	—	$B_{\cdot}^{1,0}$	0.20000	0.23624	0.41497
2	—	$S_{\cdot}^{2,0}$	—	$B_{\cdot}^{1,0}$	0.15000	0.18709	0.17440
3	15	$S_{\cdot}^{3,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.03089	0.01109
4	15	$S_{\cdot}^{4,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.16359	0.15599
5	5	$S_{\cdot}^{5,j}$	—	$B_{\cdot}^{1,0}$	0.00156	0.01470	0.01327
6	—	$S_{\cdot}^{1,0}$	2	$B_{\cdot}^{2,k}$	0.00625	0.01133	0.01066
7	—	$S_{\cdot}^{1,0}$	15	$B_{\cdot}^{2,k}$	0.00625	0.01199	0.01130

Note. Each row defines a model, indicating whether the model has a predictive and/or prognostic factor, with "–" indicating the absence of such a factor and a number indicating which factor is involved. The definitions of the treatment and baseline submodels are given in equations (3) and (4). (The true model is the second listed model.)

Table 5 For each of the four data sets y_a, y_b, y_c, and y_d, and for individual i within each data set, P_i is the posterior probability of a nonzero treatment effect and Λ_i is the posterior expected treatment effect magnitude

	y_a		y_b		y_c		y_d	
	P_i	Λ_i	P_i	Λ_i	P_i	Λ_i	P_i	Λ_i
1	0.612	8.550	0.699	9.465	0.996	15.343	0.613	8.972
2	0.737	9.521	0.706	9.483	0.223	−2.760	0.389	5.573
3	0.425	5.834	0.394	5.514	0.993	15.346	0.375	5.355
4	0.438	5.938	0.729	9.533	0.992	15.344	0.411	5.830
5	0.599	8.448	0.412	5.715	0.997	15.339	0.394	5.685
6	0.579	7.971	0.395	5.514	0.220	−3.041	0.598	8.861
7	0.619	8.549	0.405	5.613	0.223	−2.769	0.367	5.253
8	0.714	9.493	0.695	9.465	0.220	−3.022	0.604	8.925
9	0.570	7.906	0.400	5.572	0.994	15.343	0.637	9.039
10	0.738	9.532	0.400	5.564	0.223	−2.765	0.619	8.935
11	0.579	7.962	0.402	5.615	0.220	−3.026	0.618	8.915
12	0.590	8.397	0.411	5.688	0.993	15.343	0.364	5.199
13	0.579	7.975	0.478	6.881	0.996	15.341	0.639	9.034
14	0.755	9.574	0.394	5.494	0.996	15.341	0.627	8.977
15	0.619	8.577	0.703	9.492	0.220	−3.017	0.632	9.022
16	0.754	9.564	0.696	9.447	0.225	−2.675	0.582	8.807
17	0.732	9.520	0.702	9.479	0.997	15.339	0.613	8.927
18	0.603	8.463	0.704	9.476	0.993	15.344	0.373	5.344
19	0.616	8.532	0.472	6.808	0.220	−3.023	0.380	5.421
20	0.587	8.004	0.406	5.656	0.221	−2.947	0.653	9.078

Note. Only the first 20 individuals are listed.

model or models that involve a treatment effect null. And the posterior probabilities of the true models decline using BIC. As BIC is typically viewed as being overly conservative, the suggestion is that the prior parameter choices in section 3.2 are not unreasonable.

4.2.2. Posterior Analysis for at Most Two Factor Splittings. We do not give a full description here of the case in which up to two factor splittings are allowed, but do present the results of the analysis for the situation of data set y_c.

Table 6 For data y_b, posterior probabilities of nonzero treatment effects for subgroups formed by splits of the 32 factors; if a factor is not listed, its posterior probabilities are similar to factor 1

	Default prior		With prior info. I		With prior info. II	
Predictive factor j	$X_j = 1$	$X_j = 0$	$X_j = 1$	$X_j = 0$	$X_j = 1$	$X_j = 0$
1	0.561	0.545	0.567	0.546	0.583	0.551
13	0.707	0.412	0.749	0.380	0.881	0.280
14	0.678	0.407	0.713	0.375	0.821	0.271

Note. The posterior probabilities are given for the default choice of prior probabilities and for prior probabilities I and II given in section 4.3.

Table 7 For data y_c, posterior probabilities of nonzero treatment effects for subgroups formed by splits of the 32 factors; if a factor is not listed, its probabilities are similar to factor 1

	Default prior		With prior info. III	
Predictive factor j	$X_j = 1$	$X_j = 0$	$X_j = 1$	$X_j = 0$
1	0.561	0.629	0.559	0.625
23	0.614	0.574	0.616	0.566
27	0.994	0.222	0.989	0.221
28	0.928	0.307	0.923	0.306

Note. The posterior probabilities are given for the default choice of prior probabilities and for prior probabilities III given in section 4.3.

In particular, we allowed any combination of the splitting of two factors for the predictive submodels, but allowed only up to one factor splits for the prognostic models.

For data set y_c, the resulting model posterior probabilities (exceeding 0.01) are given in Table 8. There are in total 30 distinct predictive submodels for each pair of predictive factors. The treatment submodels $S_{\cdot}^{3,j}$ and $S_{\cdot}^{5,j}$ in the table are the single split submodels discussed earlier. Further, the models in lines 1, 2, 12, 13, and 14 of the table are the same models as in lines 1, 2, 3, 6, and 8 of Table 3. The models in lines 3 through 11 of Table 8 have predictive submodels involving both indicated

Table 8 For data y_c, posterior probabilities of outcome models formed by allowing up to two splittings of the 32 factors in defining treatment submodels

	Predictive factor j	Predictive factor ℓ	Prognostic factor k	Treatment submodel	Baseline submodel	Prior probability	Posterior probability
1	27	—	—	$S_{\cdot}^{3,j}$	—	0.0011097	0.0434030
2	27	—	—	$S_{\cdot}^{5,j}$	—	0.0011097	0.1581320
3	23	27	—	$S_{\cdot}^{11,j,\ell}$	—	0.0000086	0.0275689
4	23	27	—	$S_{\cdot}^{18,j,\ell}$	—	0.0000086	0.0163023
5	23	27	—	$S_{\cdot}^{19,j,\ell}$	—	0.0000086	0.0124966
6	17	27	—	$S_{\cdot}^{21,j,\ell}$	—	0.0000086	0.0100329
7	18	27	—	$S_{\cdot}^{21,j,\ell}$	—	0.0000086	0.0256135
8	23	27	—	$S_{\cdot}^{21,j,\ell}$	—	0.0000086	0.0655428
9	23	27	—	$S_{\cdot}^{26,j,\ell}$	—	0.0000086	0.0137529
10	23	27	—	$S_{\cdot}^{28,j,\ell}$	—	0.0000086	0.0526748
11	23	27	—	$S_{\cdot}^{29,j,\ell}$	—	0.0000086	0.0170835
12	27	—	23	$S_{\cdot}^{3,j}$	$B_{\cdot}^{2,k}$	0.0000347	0.0127792
13	27	—	23	$S_{\cdot}^{5,j}$	$B_{\cdot}^{2,k}$	0.0000347	0.0475241
14	27	—	25	$S_{\cdot}^{5,j}$	$B_{\cdot}^{2,k}$	0.0000347	0.0101776

Note. Each row defines a model, indicating whether the model has predictive and/or prognostic factors, with "–" indicating the absence of such a factor and numbers indicating the factors involved. The definitions of the treatment and baseline submodels are given in equations (3) and (4). (The true model is the third listed model.)

predictive factors and are of the form

$$
\begin{aligned}
S_i^{11,j,\ell} &= T_i \mu_{11,j,\ell} \mathbf{1}_{\{x_{i,j}=1\}} \mathbf{1}_{\{x_{i,\ell}=1\}}, \\
S_i^{18,j,\ell} &= T_i \mu_{18,\ell} \mathbf{1}_{\{x_{i,\ell}=0\}} + T_i \mu_{18,j,\ell} \mathbf{1}_{\{x_{i,j}=1\}} \mathbf{1}_{\{x_{i,\ell}=1\}} + T_i \mu'_{18,j,\ell} \mathbf{1}_{\{x_{i,j}=0\}} \mathbf{1}_{\{x_{i,\ell}=1\}}, \\
S_i^{19,j,\ell} &= T_i \mu_{19,\ell} \mathbf{1}_{\{x_{i,\ell}=0\}} + T_i \mu_{19,j,\ell} \mathbf{1}_{\{x_{i,j}=0\}} \mathbf{1}_{\{x_{i,\ell}=1\}}, \\
S_i^{21,j,\ell} &= T_i \mu_{21,j,\ell} \mathbf{1}_{\{x_{i,j}=1\}} \mathbf{1}_{\{x_{i,\ell}=1\}} + T_i \mu'_{21,j,\ell} \mathbf{1}_{\{x_{i,j}=0\}} \mathbf{1}_{\{x_{i,\ell}=1\}}, \\
S_i^{26,j,\ell} &= T_i \mu_{26,j,\ell} \mathbf{1}_{\{x_{i,j}=1\}} \mathbf{1}_{\{x_{i,\ell}=1\}} + T_i \mu'_{26,j,\ell} \mathbf{1}_{\{x_{i,j}=1\}} \mathbf{1}_{\{x_{i,\ell}=0\}} \\
&\quad + T_i \mu''_{26,j,\ell} \mathbf{1}_{\{x_{i,j}=0\}} \mathbf{1}_{\{x_{i,\ell}=1\}} + T_i \mu'''_{26,j,\ell} \mathbf{1}_{\{x_{i,j}=0\}} \mathbf{1}_{\{x_{i,\ell}=0\}}, \\
S_i^{28,j,\ell} &= T_i \mu_{28,j,\ell} \mathbf{1}_{\{x_{i,j}=1\}} \mathbf{1}_{\{x_{i,\ell}=1\}} + T_i \mu'_{28,j,\ell} \mathbf{1}_{\{x_{i,j}=0\}} \mathbf{1}_{\{x_{i,\ell}=1\}} + T_i \mu''_{28,j,\ell} \mathbf{1}_{\{x_{i,j}=0\}} \mathbf{1}_{\{x_{i,\ell}=0\}}, \\
S_i^{29,j,\ell} &= T_i \mu_{29,j,\ell} \mathbf{1}_{\{x_{i,j}=1\}} \mathbf{1}_{\{x_{i,\ell}=1\}} + T_i \mu'_{29,j,\ell} \mathbf{1}_{\{x_{i,j}=1\}} \mathbf{1}_{\{x_{i,\ell}=0\}} + T_i \mu''_{29,j,\ell} \mathbf{1}_{\{x_{i,j}=0\}} \mathbf{1}_{\{x_{i,\ell}=0\}}.
\end{aligned}
$$

The models in lines 3 through 5 and 8 through 11 all have predictive effects for both Factors 23 and 27 (which was how the data was generated), and the sum of the prior probabilities and posterior probabilities of these models is 0.0000602 and 0.205, respectively. Thus, the posterior to prior odds of these two factors having a predictive effect (based on this collection of models) is 3405.

Table 9 presents the marginal posterior probability that each factor has a predictive treatment effect. There is compelling evidence that Factor 27 has a predictive treatment effect, but only moderate evidence for Factor 23. Factor 28 appears to have a strong predictive effect because it is highly correlated with Factor 27.

Table 10 presents the marginal posterior probability of a predictive treatment effect for subgroups formed by pairs of factors. The correct model has treatment subgroup $\{X_{27} = 1, X_{23} = 1\}$, which does have the largest posterior probability of 0.935. This table does reveal that Factor 28 is not predictive given Factor 27. For instance, given $X_{27} = 1$, there is virtually no difference between the treatment effect probabilities when $X_{28} = 0$ and $X_{28} = 1$; on the other hand, given $X_{28} = 1$, there is a huge difference between the treatment effect probabilities when $X_{27} = 0$ and $X_{27} = 1$. In contrast, given $X_{27} = 1$, there is a reasonably large difference between the

Table 9 For data y_c, marginal posterior probabilities of nonzero treatment effects for subgroups formed by a single factor, with models allowing splitting by up to two of the 32 factors; if a factor is not listed, its probabilities are similar to factor 1

	Default prior	
Predictive factor j	$X_j = 1$	$X_j = 0$
1	0.607	0.643
23	0.667	0.584
27	0.870	0.396
28	0.830	0.448

Table 10 For data y_c, marginal posterior probabilities of nonzero treatment effects for subgroups formed by splitting with two of the 32 factors; if a factor is not listed, its probabilities are similar to those involving factor 1

Predicted i	Predicted j	$\{X_i = 1, X_j = 1\}$	$\{X_i = 1, X_j = 0\}$	$\{X_i = 0, X_j = 1\}$	$\{X_i = 0, X_j = 0\}$
1	23	0.596	0.617	0.706	0.624
1	27	0.908	0.370	0.906	0.383
1	28	0.839	0.424	0.886	0.444
23	27	0.935	0.341	0.878	0.403
23	28	0.897	0.401	0.828	0.459
27	28	0.908	0.898	0.381	0.375

treatment effect probabilities when $X_{13} = 0$ and $X_{13} = 1$, suggesting that Factor 13 does have a predictive effect (along with Factor 27).

4.3. Introducing Prior Knowledge

It is quite common that when subgroup identification is performed for clinical trials, relevant information preexists that inform the plausibility of each variable being a predictive biomarker. Such information may be provided by data from earlier—and often smaller sized—studies for the same compound, studies of similar compounds, and general understanding of the disease. It is possible that such information has not been accounted for in subgroup identification as much as it should, due to the lack of a coherent and elegant approach to utilize it in frequentist analyses. Furthermore, if one considers candidate gene association studies, upon which the current example was built, genetic markers are selected based on biological knowledge of the mechanism through which the compound is expected to modify the disease to be treated. The same information that leads to the selection of these markers often also informs the investigator of their relative plausibility.

In this subsection, the effect of adding prior knowledge to the elicitation process is illustrated. Since there are true predictive effects only in the response data y_b and y_c, the discussion here focuses on those two data sets.

For the situation of data set y_b, suppose prior knowledge, based on genetic theory or previous experiments, suggests that the first 16 biomarkers are more likely to be predictive than the last 16. Consider two variants of this.

Prior information I: The first 16 biomarkers are assigned effect odds of 1, while the last 16 biomarkers are assigned effect odds of $1/3$, so the first 16 are viewed as being 3 times more likely to have predictive effects.

Prior information II: Keep the effect odds as in I, but change the effect odds of 13 and 14 to 5, corresponding to, say, a prior experiment that suggested that these biomarkers are particularly likely to be active.

The changes in the posterior model probabilities are given in the second and third panels of Table 2. The changes in the subgroup predictive probabilities are given in the indicated parts of Table 6. In Table 2, the change to Prior information I does not seem to have a major effect on the posterior model

probabilities. In contrast Prior information II has a clear effect, significantly increasing the probabilities of models involving a predictive effect for Factor 13. The same behavior is shown in Table 6 for the subgroup predictive probabilities. The indication here is that changing the prior effect odds for, say, half of the factors will have only minor influence, while changing the prior effect odds for small groups of factors can have a major influence.

For data set y_c, consider the following change in prior information:

Prior information III: Let all factors have effect odds of 1, except set the effect odds of Factors 23 and 24 to 10. The interest is in seeing whether this is enough of a change to allow Factor 23 to come out from under the shadow of Factor 27.

The results are given in the second panel of Table 3. There is a dramatic increase in the probabilities of the models that have Factor 23 as a prognostic variable (the closest the single-split models can come towards use of 23). Interestingly, however, Table 7 shows that the probability that Factor 23 has a subgroup predictive effect remains negligible, as it should be since this is being classified as having a prognostic effect in the single-split models.

5. CONCLUSION AND DISCUSSION

The Bayesian approach to subgroup analysis exerts quite strong control on multiple testing, through the assignment of prior probabilities to the models. When there are many possible factors defining subgroups, we have seen that the data must provide very strong evidence for a predictive effect in order to yield a high posterior probability for a predictive effect. In addition to the strong multiplicity control, this apparent conservatism is due to the fact that posterior probabilities of "null" models are typically much larger than non-Bayesian quantities such as p-values (see, e.g., Sellke et al., 2001) and more appropriately reflect the actual evidence concerning effects.

The Bayesian analysis yields the full posterior distribution over all unknowns in the problem. From this posterior distribution, any statistical inference (e.g., estimation, testing, confidence sets, prediction) and any decision analysis can be performed. It is not within the scope of the article to recommend any particular methods of utilizing this posterior—that is for the stakeholders in application to determine—but it is useful to review some of the most important features and properties of this posterior distribution.

Subgroup treatment effects: Since there are typically many models that correspond to a predictive effect for a particular factor, overall posterior summaries, such as the overall posterior probability that a factor has a predictive effect and the overall expected effect size, are the key quantities for any decisions regarding subgroup treatment.

Personalized medicine: The analysis yields the posterior probability that an individual with specific factor values will have a treatment effect and also yields the expected magnitude of the individual's treatment effect. For personalized medicine, these would be key in deciding on a treatment.

Potential discoveries: If one is primarily interested in discovering interesting factors for future study, the posterior to prior odds (the ratio of the posterior probability of a predictive effect to the prior probability of a predictive effect) indicates the strength of the effect as suggested by the data alone. Factors with very high posterior to prior odds may well be worth future investigation, although the posterior probability of a predictive effect should also be considered in the decision as to whether or not it is worthwhile to engage in such investigation.

Objective Bayes option: A reasonable default choice of all prior probabilities and prior distributions was presented, allowing automatic use of the methodology.

Power: With the Bayesian approach, there is no loss in power due to dependent test statistics, as commonly happens with many non-Bayesian efforts to control multiplicity. Also, the power for factors/subgroups that are a priori believed to be most likely to exhibit a treatment effect can be enhanced through (preexperimental) choice of their prior probabilities; as long as this is done preexperimentally, there is no loss of multiplicity control.

The main negative with the Bayesian approach discussed here is potential difficulty with the computation. If only one factor splitting for baseline and treatment models is allowed, computation is easy; for instance, computations for the 32 factor testbed examples only took about 2 minutes on a laptop. Allowing up to two factor splittings greatly increases the computation time, however, as the model space grows enormously. For the two factor splitting situation in section 4.2.2, for instance, computation took about 4 hours on a laptop. If there were a much larger number of factors or if more than two factor splits were allowed, the number of potential models becomes too large for enumeration, and schemes for sampling from or searching model space must be utilized. Such schemes will be explored elsewhere.

FUNDING

This work was supported by Eli Lilly and Company through the Lilly Research Award Program (LRAP) and by National Science Foundation grant DMS-1007773.

REFERENCES

Berry, D. A. (1990). Subgroup analyses. *Biometrics* 46:1227–1230.

Cui, L., James Hung, H. M., Wang, S. J., Tsong, Y. (2002). Issues related to subgroup analysis in clincial trials. *Journal of Biopharmaceutical Statistics* 12:347–358.

Dixon, D. O., Simon, R. (1991). Bayesian subset analysis. *Biometrics* 47:871–881.

Foster, J. C., Taylor, J. M., Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 30:2867–2880.

Hodges, J. S., Cui, Y., Sargent, D. J., Carlin, B. P. (2007). Smoothing balanced single-error-term analysis of variance. *Technometrics* 49:12–25.

Jones, H. E., Ohlssen, D. I., Neuenschwander, B., Racine, A., Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. *Clinical Trials* 8:129–143.

Lagakos, S. (2006). The challenge of subgroup analyses–reporting without distorting. *New England Journal of Medicine* 354:1667–1669.

Laud, P. W., Sivaganesan, S., Müller, P. (2013). Subgroup analysis. In: Damien, P., Dellaportas, P., Polson, N., Stephens, D., eds. *Bayesian Theory and Applications*. Oxford, UK: Oxford University Press, pp. 576–592.

Lavedan, C., Volpi, S., Polymeropoulos, M. H., Wolfgang, C. D. (2008). Effect of a ciliary neurotrophic factor polymorphism on schizophrenia symptom improvement in an iloperidone clinical trial. *Pharmacogenomics* 9:289–301.

Lipkovich, I., Dmitrienko, A., Denne, J., Enas, G. (2011). Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine* 30:2601–2621.

Liu, W., Downing, A. C. M., Munsie, L. M., Chen, P., Reed, M. R., Ruble, C. L., Landschulz, K. T., Kinon, B. J., Nisenbaum, L. K. (2012). Pharmacogenetic analysis of the mGlu2/3 agonist LY2140023 monohydrate in the treatment of schizophrenia. *The Pharmacogenomics Journal* 12:246–254.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., Boivin, J.-F. (2005). Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Statistics and Computing* 15:231–239.

Pocock, S. J., Assmann, S. E., Enos, L. E., Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine* 21:2917–2930.

Ruberg, S. J., Chen, L., Wang, Y. (2010). The mean does not mean as much anymore: Finding sub-groups for tailored therapeutics. *Clinical Trials* 7:574–583.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6:461–464.

Scott, J. G., Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics* 38:2587–2619.

Sellke, T., Bayarri, M. J., Berger, J. (2001). Calibration of *p*-values for testing precise null hypotheses. *The American Statistician* 55:62–71.

Simon, R. (2002). Bayesian subset analysis: Application to studying treatment-by-gender interactions. *Statistics in Medicine* 21:2909–2916.

Sivaganesan, S., Laud, P. W., Müller, P. (2011). A Bayesian subgroup analysis with a zero-enriched Polya urn scheme. *Statistics in Medicine* 30:312–323.

Su, X., Zhou, T., Yan, X., Fan, J., Yang, S. (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics* 4:1–26.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* 10:141–158.

Temple, R., Ellenberg, S. S. (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: Ethical and scientific issues. *Annals of Internal Medicine* 133:464–470.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., Drazen, J. M. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine* 357:2189–2194.

Wang, X. (2012). *Bayesian Modeling Using Latent Structures*. Ph.D. thesis, Duke University.

Westfall, P. H., Johnson, W. O., Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* 84:419–427.

Woodcock, J. (2007). The prospects for "personalized medicine" in drug development and drug therapy. *Clinical Pharmacology and Therapeutics* 81:164–169.