

Chapter 16

Subgroup Analysis from Bayesian Perspectives



Yang Liu, Lijiang Geng, Xiaojing Wang, Donghui Zhang, and Ming-Hui Chen

Abstract Identifying the sub-population structures along with the tailored treatments for all groups plays a critical rule for assigning the best available treatment to an individual patient. Subgroup analysis, a key to develop personalized medicine, becomes increasingly important over the past decade. Besides frequentist methods, there are a spectrum of methods developed from Bayesian perspectives to identify subgroups. In this chapter, we provide a comprehensive overview of Bayesian methods and discuss their properties. We further examine empirical performance of the two Bayesian methods via simulation studies and a real data analysis.

16.1 Introduction

In order to provide the best available treatment for individual patients, it is critical to examine whether heterogeneous treatment effect exists among the patient population. Many exploratory methods are developed in the literature to identify subgroups. Among them there are a variety of frequentist approaches, for instance, recursive tree based methods such as Interaction Trees (Su et al. 2009), Virtual Twins (Foster et al. 2011), Subgroup Identification based on Differential Effect Search (SIDES) (Lipkovich et al. 2011), Qualitative Interaction Trees (Dusseldorp and Van Mechelen 2014) and Generalized Unbiased Interaction Detection and Estimation (GUIDE) (Loh et al. 2015). Some optimization-oriented optimal treatment regime methodologies (Zhao et al. 2012, 2015; Tian et al. 2014; Chen et al. 2017) are also developed within the frequentist framework.

Y. Liu · L. Geng · X. Wang · M.-H. Chen (✉)
Department of Statistics, University of Connecticut, Storrs, CT, USA
e-mail: ming-hui.chen@uconn.edu

D. Zhang
Global Biostatistics and Programming, Sanofi US, Bridgewater, NJ, USA

Meanwhile, many Bayesian methods are proposed from different perspectives to identify subgroups. In Sect. 16.2, we give an overview of some recently developed Bayesian methods for subgroup analysis. Simulation studies are conducted in Sect. 16.3. Section 16.4 presents a real data analysis. We conclude this chapter with a brief discussion in Sect. 16.5.

16.2 Bayesian Subgroup Analysis Methods

In subgroup analysis, a nonparametric mean structure $E(Y|X, trt) = g(X, trt)$ is often considered for the data Y , where $g(\cdot)$ is a multivariate function representing an underlying mechanism of the signal, trt indicates the treatment option, and X is a vector of potential covariates used to identify subgroups. In a commonly investigated scenario, there are two treatment options, placebo or treatment, i.e., $trt = 0$ or $trt = 1$. Then, the difference of treatment effects $\Delta(X)$ between these two options can be defined as

$$\Delta(X) = E(Y | X, trt = 1) - E(Y | X, trt = 0) = g(X, 1) - g(X, 0).$$

Therefore, we can equivalently model the nonparametric mean structure as

$$E(Y | X, trt) = g(X, 0) + \Delta(X)trt = a(X) + \Delta(X)trt. \quad (16.2.1)$$

The first term $a(X)$ in Eq. (16.2.1) is usually referred as the *prognostic effect*, since it affects the response at the same amount regardless of the treatment assignment. $\Delta(X)$ is often called the *predictive effect* or *predictive subgroup effect*, as $\Delta(X)trt$ affects the response differently under the different treatment assignment trt .

Tracing back to the literature in the twentieth century, Dixon and Simon (1991) proposed a linear regression model

$$E(Y | X, trt) = \mu + \tau trt + X\beta + \gamma Xtrt \quad (16.2.2)$$

with the first-order term γX serving as $\Delta(X)$, and a linear function of X serving as the prognostic effect $a(X)$, assuming the covariate X has two possible values. The parameters $(\mu, \tau, \beta, \gamma)$ are estimated using a Bayesian approach. Jones et al. (2011) extended the previous linear regression framework of Dixon and Simon (1991) by allowing second-order and third-order interaction terms for the predictive effects. These two methods are not directly applicable when there are other types of covariates, and may not work well when there are a large number of candidate variables.

Many other Bayesian subgroup analysis methods have been proposed from various perspectives. Below we introduce several recently developed Bayesian methods grouped by their similarity.

16.2.1 Tree-Based Bayesian Subgroup Analysis Methods

There are a few Bayesian subgroup analysis approaches which are linked to tree structures. The advantage of a tree structure is that it can handle interactions and nonlinear relationships between covariates and responses in an implicit way.

Berger et al. (2014) used a tree-splitting process to construct the treatment (subgroup) submodels, i.e., $\Delta(X)$ and baseline (prognostic) submodels, i.e., $a(X)$, which simultaneously incorporate the predictive effects and prognostic effects in the modeling. The tree-splitting process is randomly bisecting the covariate space recursively and leads to an allowable partition of the entire population arising from terminal nodes of a tree based on covariate splits, with possible zero treatment or baseline effects. There are several key steps in stochastically splitting a tree: (1) randomly select an ordering of covariates for splitting; (2) randomly determine the existence of a zero effect node at each level, and then randomly choose one of the nodes at that level to be the zero effect, which is a terminal node; (3) randomly decide non-zero effect nodes at each level to be further split by the corresponding covariate at that level; if not it becomes a terminal node. The detailed elaboration of the tree constructions is discussed in Wang (2012). The advantage of this tree splitting process is the elimination of possible partitions of the entire population without scientific meaning in comparison of treatment or baseline effects, which dramatically reduces the total number of models considered in the model space for the outcome.

The simplest way to model the outcome is to combine the treatment and baseline submodels with additive effects. Then, the model space for the outcome Y includes all possible distinct combinations of these two submodels. Next, the prior probabilities of the outcome models are assigned according to the stochastic scheme to generate trees. Once the prior specification is complete, the Bayesian model average techniques are utilized for subgroup analysis and, as a byproduct, the yielded results provide individual probabilities of treatment effect that might be useful for personalized medicine.

Here, we briefly discuss their main idea of defining an outcome model and specifying the priors. Let Ω denote the set of covariates in the study. Let X_{ij} be the j -th binary covariate for the i -th person, where $j \in \Omega$ and $i = 1, \dots, n$. If we allow at most one covariate to split the treatment submodel, we are going to have five different types of models, i.e., $S_i^{1,0} = 0$, $S_i^{2,0} = trt_i \mu_2$, $S_i^{3,j} = trt_i \mu_{3j} \mathbf{1}_{\{X_{ij}=0\}} + trt_i \mu'_{3j} \mathbf{1}_{\{X_{ij}=1\}}$, $S_i^{4,j} = trt_i \mu_{4j} \mathbf{1}_{\{X_{ij}=0\}}$, $S_i^{5,j} = trt_i \mu_{5j} \mathbf{1}_{\{X_{ij}=1\}}$, where μ_2 is the mean overall treatment effect (if present), μ_{3j} , μ'_{3j} , μ_{4j} and μ_{5j} are the potential treatment (predictive) effects in the subgroups defined by the covariate j , trt_i is the treatment indicator, and $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Similarly, there are two possible types of baseline submodels via splitting one factor. That is, $B_i^{1,0} = \mu_1$ and $B_i^{2,k} = \mu_1 + \beta_k \mathbf{1}_{\{X_{ik}=0\}}$, where μ_1 is the overall mean and β_k is the mean baseline effect for covariate $k \in \Omega$.

Then, the outcome model in Berger et al. (2014) is

$$Y_i = S_i^{h,j} + B_i^{\ell,k} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad (16.2.3)$$

$i = 1, \dots, n$, $h = 1, \dots, 5$, $\ell = 1, 2$ and $j, k \in \{0, \Omega\}$. Let m be the number of covariates considered, then the total number of models for at most one covariate splitting is $2 + 5m + 3m^2$, which is a huge reduction from 2^{m+1} possible models when m is large.

The method developed in Berger et al. (2014) automatically takes account of multiplicity adjustment in the prior specification for the model space. The prior probability is computable via specifying three interpretable prior inputs, which are: (1) specifying the prior probability that an individual has no treatment (predictive) effect and no baseline (prognostic) effect, respectively; (2) assigning relative effect odds for a covariate i has an effect compared to the first covariate; (3) defining the ratio of the sum of the prior probabilities of the submodels with $i - 1$ split and the sum of the prior probabilities of the submodels with i splits. An advantage for this prior specification is that the experts can easily incorporate pre-experimental preference to specific subgroups. See Section 3 of Berger et al. (2014) for more details of computing the prior probability for each outcome model based on the three interpretable inputs.

Once the prior specification for the outcome model and the unknown parameters in the model is complete, then we can summarize the posterior quantity we are interested in. In Berger et al. (2014), they summarized the posterior quantity of interest using the Bayesian model averaging idea. Two interesting posterior summaries discussed in their paper are:

1. *Individual Treatment Effects*: first, the probability for an individual to have treatment effects is given by $P_i = \sum_{\mathcal{M}_\kappa \in \mathcal{M}} \mathbb{P}(\mathcal{M}_\kappa \mid Y_1, \dots, Y_n) \mathbf{1}_{\{\mu_{i,\kappa} \neq 0\}}$, for any $i = 1, \dots, n$, where \mathcal{M} denotes the entire model space for the outcome model, \mathcal{M}_κ is a specific outcome model in the model space, $\mu_{i,\kappa}$ is the subgroup treatment effect associated with the i th individual in the given model \mathcal{M}_κ and $\bar{\mu}_{i,\kappa}$ is the posterior mean of $\mu_{i,\kappa}$. Then, the *individual treatment effect size* for each individual is defined as weighted average of $\bar{\mu}_{i,\kappa}$, i.e., $\Lambda_i = \sum_{\kappa} \mathbb{P}(\mathcal{M}_\kappa \mid Y_1, \dots, Y_n) \bar{\mu}_{i,\kappa} \mathbf{1}_{\{\mu_{i,\kappa} \neq 0\}} / P_i$.
2. *Subgroup Treatment Effects*: based on individual posterior probability for the treatment effects, the posterior probability of a nonzero treatment effect for Subgroup j (denoted as S_j) is defined as an average of P_i over the subgroups that individual j belongs to (using the symbol $\{\#i \in S_j\}$, i.e., $Q_j = \sum_{i \in S_j} P_i / \{\#i \in S_j\}$). Similarly, the subgroup treatment effect size for S_j is calculated via $\Delta_j = \sum_{i \in S_j} P_i \Lambda_i / \sum_{i \in S_j} P_i$.

The Bayesian approach described in Berger et al. (2014) can be generally extended to allow more than one covariate used in splitting. However, when more than two covariates are utilized in tree-splitting process, the total number of models that we need to consider will be increasing and the model enumeration scheme in Berger et al. (2014) becomes impossible.

Sivaganesan et al. (2017) restricted the scope from searching for subgroup effects among all possible subgroups, to searching for subgroup effects among only a few pre-determined candidate groups. More specifically, the authors focus on identifying subgroup effects related to certain pre-specified covariates and shapes of subgroups. Any center regions in the covariate space will be excluded from consideration, for instance, a subgroup defined as $\{a < X_1 < b, c < X_2 < d\}$ would be excluded. For any subgroup A , the amount by which its predictive effect $\Delta(A) = E(Y|X \in A, trt = 1) - E(Y|X \in A, trt = 0)$ exceeds the predictive effects of entire patient population $\Delta(C) = E(Y|trt = 1) - E(Y|trt = 0)$, that is, $\delta(A) = \Delta(A) - \Delta(C)$, is used as the primary measure for identifying any potential enhanced subgroup effects. The author defined a utility function to compare potential subgroups:

$$U(A) = \begin{cases} \frac{[(|A|-N)_+]^d}{(1+c)^{nvar(A)-1}} [\delta(A) - T_s], & \text{if } \emptyset \subset A \subset C, \\ 0, & A = \emptyset, \end{cases} \quad (16.2.4)$$

where $|A|$ is the number of observations in A , and $nvar(A)$ is the number of covariates used to define A , $\{x\}_+ = \max(0, x)$, N is the pre-specified minimum subgroup size, $c, d > 0$ are constants to control the “reward” for the subgroup size and the “penalty” for complex subgroups, respectively. T_s is the minimum threshold for $\delta(A)$ which corresponds to the clinically meaningful effect magnitude. Bayesian Additive Regression Trees (BART) (Chipman et al. 2010) approach is used to fit the response Y on the combined covariate space (X, trt) as a nonparametric function, to get the predicted value of $\delta(A)$ for each subgroup A . Subgroups with larger positive expected utility are preferred. Since the candidate subgroups are pre-specified, the process of exploring from the entire covariate space is omitted, which makes this approach differ from many other exploratory subgroup analysis methods.

Zhao et al. (2018) proposed another BART-based subgroup analysis approach to identify important biomarkers. They modeled the predictive effects $\Delta(X)$ with a single tree for better interpretability, and impose an additive tree structure on the prognostic effects $a(X)$ to enhance model fitting. Such an additive tree structure allows more flexibility for the prognostic effect comparing to the commonly assumed linear structures in Dixon and Simon (1991), Jones et al. (2011), and Schnell et al. (2016), which may lead to better estimation performance for the predictive effects $\Delta(X)$ at the same time. However, the computation time will also increase quickly when sample size and number of candidate variables get larger.

Similar to BART, the posterior sampling procedure is carried out using Bayesian backfitting algorithm (Hastie et al. 2000). The posterior probability that a biomarker served as a splitting variable in the predictive tree will be used to determine whether any covariate has notable predictive subgroup effect. In order to reduce “type I error”, i.e., claiming irrelevant covariates as predictive biomarkers, no specific subgroup will be declared when such posterior probabilities for all biomarkers are less than a certain threshold. Several simulation scenarios with at most two biomarkers are considered in the paper, and the estimated probabilities for the true predictive biomarker(s) to rank as the top two predictive variables are reported. The

method seems to identify the predictive biomarkers well when there is only one predictive biomarker, despite presence of some prognostic effects. Meanwhile it appears to be underpowered when there are two predictive biomarkers in the model, especially for the purpose of identifying both predictive variables as the top two candidate biomarkers.

16.2.2 ANOVA-Based Bayesian Subgroup Analysis Methods

Sivaganesan et al. (2011) developed a Bayesian approach from model selection perspective by considering each covariate separately and constructed the model space by enumerating the possible cases for different levels of treatment-subgroup effects. First denote M_{00} and M_{10} the overall null and the overall effect model, representing no treatment effect and homogeneous treatment effect in the whole population, and the model space of “overall effect”, i.e., the model space of no treatment-subgroup interaction models, is $\mathcal{M}_0 = \{M_{00}, M_{01}\}$. Then for each covariate, define models in the model space by introducing the cluster membership indicator $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_S)$, where the elements in $\boldsymbol{\gamma}$ range from 0 to number of distinct non-zero treatment-subgroup effects and represent the order of appearances of distinct treatment-subgroup effects, and S is the number of levels of the covariate. To demonstrate this setting more clearly, Table 16.1 shows an example of models defined by a covariate of two levels.

Use the zero-enriched Polya urn scheme as the probability distribution on the model space $\mathcal{M}_{\mathcal{X}}$. After getting the posterior model probabilities, the authors proposed a decision-making algorithm, comparing the posteriors of models in \mathcal{M}_0 with the models defined by each covariate to determine whether notable subgroup effects should be reported. In the algorithm, two threshold values c_0 and c_1 are used for comparing model posterior probabilities. The model selected is the most likely model, and also beats the overall null model M_{00} and the overall effect model M_{01} as its posterior probability odds exceeding c_0 and c_1 . Therefore, c_0 represents the threshold for the posterior probability odds of the overall or a subgroup effect model against the overall null model, and c_1 represents the threshold for the posterior probability odds of a subgroup effect model against the overall effect model. When

Table 16.1 Example of model space $\mathcal{M}_{\mathcal{X}}$ defined by covariate X

Model index	$\boldsymbol{\gamma}$	Treatment-subgroup effects
M_0	(0, 0)	$\Delta(X = 0) = \Delta(X = 1) = 0$
M_1	(1, 0)	$\Delta(X = 0) \neq 0, \Delta(X = 1) = 0$
M_2	(0, 1)	$\Delta(X = 0) = 0, \Delta(X = 1) \neq 0$
M_3	(1, 1)	$\Delta(X = 0) = \Delta(X = 1) \neq 0$
M_4	(1, 2)	$\Delta(X = 0) \neq \Delta(X = 1), \Delta(X = 0), \Delta(X = 1) \neq 0$

X has two levels

no subgroup or overall effect models satisfy the comparing conditions, the overall null model will be selected at last. A characteristic of this algorithm is that, when selecting subgroup models, it only compares models within the model space of each covariate, and in the end reports either models in \mathcal{M}_0 , or one or more subgroup models defined by different covariates. Therefore, this method cannot discover subgroups defined by interactions of multiple covariates, unless data transformation is done. However, an advantage of this method is that it does not only discover subgroups, but also detect orders of subgroup effect sizes.

Liu et al. (2017) extended Sivaganesan et al. (2011) by considering two variables at a time and enumerated all possible situations for the mean levels to construct the model space. The authors elaborated on the case that there are two covariates of interest and each has two levels, which are specified a priori by the investigators. Similar to the decision algorithm introduced in Sivaganesan et al. (2011), a stepwise procedure is adopted based on posterior model probabilities to determine potential subgroup effects. The model space grows quickly when more covariates are considered and/or there are more than two levels for each covariate.

Both of these two ANOVA-based methods do not model prognostic effects as a function of the covariates, and the results may be biased when there exist covariate-dependent prognostic effects.

16.2.3 Other Types of Bayesian Subgroup Analysis Methods

Schnell et al. (2016) also used a linear combination of the covariates to model both prognostic effects and predictive effects similar to Dixon and Simon (1991). Denote the predictive effects as $\Delta(x) = x'\gamma$ for any covariate vector x , and define the beneficial subgroup as $B_\gamma = \{x : \Delta(x) > \delta, \delta > 0\}$ for a pre-specified threshold δ . This method aims to find a credible subgroup pair (D, S) satisfying $D \subseteq B_\gamma \subseteq S$, where D , defined as the “exclusive credible subgroup”, is the region such that the posterior probability of $\Delta(x) > \delta$ for all $x \in D$ is no less than $1 - \alpha$. The “inclusive credible subgroup” S is defined as the region such that the posterior probability of including all x , s.t. $\Delta(x) > \delta$ for all patients in S is no less than $1 - \alpha$. The highest posterior density method is applied to find the $1 - \alpha$ credible region G_α for the posterior distribution of γ . Therefore (D, S) can be constructed as: $D = \{x : x'\gamma > \delta \text{ for all } \gamma \in G_\alpha\}$, and $S^c = \{x : x'\gamma \leq \delta \text{ for all } \gamma \in G_\alpha\}$. There are two other ways of constructing (D, S) discussed in the paper.

This approach may work well when the dimension of the parameter space is low, while the computational costs increase quickly when the number of candidate covariates increases. When the dimension of parameter space is high, it is also difficult to interpret the credible subgroup pair (D, S) and characterize the patient population within it.

Gu et al. (2013) applied a two-stage Bayesian lasso approach to time-to-event responses and also used the first-order terms of X to model the predictive effects. Three different treatment options are considered. In the first stage, linear

combinations of main effects, overall treatment effects and first-order treatment-covariate interactions are considered to model the predictive effects, and shrinkage priors are specified on the parameters, and a distance-based criterion is implemented to help screen the unimportant biomarkers. In the second stage, all the biomarkers retained after the first stage will be included in the model, and the Bayesian adaptive lasso approach is deployed to perform further biomarker selection. The authors only considered the case when true predictive effects are linear structured in the simulation study, and the robustness of this method remains unknown when the predictive effect model is misspecified.

This method does not consider higher-order treatment subgroup interactions and the variable selection step does not extend further to split point selections to identify a potential subgroup such as $\{X_1 > 0.5\}$. Also, the sure screening property for the first stage has not been established yet for this method, we shall be wary of the fact that certain important biomarkers may be missed since the variables excluded from the first stage will never enter the second stage.

16.3 Simulation Studies

In this section, we carry out simulation studies to examine the empirical performance of two aforementioned methods (Berger et al. 2014; Sivaganesan et al. 2011). Both methods consider subgroups defined by one binary variable. We focus on testing of the scenarios listed below

- (a) $y_i = 2 + \epsilon_i$,
- (b) $y_i = 2 + 2trt_i + \epsilon_i$,
- (c) $y_i = 2 + 2trt_i I(X_{i1} = 0) + \epsilon_i$,
- (d) $y_i = 2 + 2trt_i I(X_{i1} = 0) + I(X_{i1} = 0) + 2I(X_{i2} = 0) - 3I(X_{i3} = 0) + \epsilon_i$,
- (e) $y_i = 2 + 2trt_i \{I(X_{i1} = 0) + I(X_{i2} = 0)\} + \epsilon_i$,
- (f) $y_i = 2 + 2trt_i \{I(X_{i1} = 0) - I(X_{i2} = 0)\} + \epsilon_i$,

where y_i is the i -th univariate response. The treatment variable $trt_i \stackrel{i.i.d}{\sim}$ Bernoulli(0.5). Ten independent binary covariates are considered: $X_{ij} \stackrel{i.i.d}{\sim}$ Bernoulli(0.5), $i = 1, \dots, n$, $j = 1, \dots, 10$. The random error is set at two levels, $\epsilon_i \stackrel{i.i.d}{\sim} N(0, 1)$ and $N(0, 4)$. We assume $\{\epsilon_i\}'s$, $\{X_{ij}\}'s$, and $\{trt_i\}'s$ are mutually independent, for $i = 1, \dots, n$, $j = 1, \dots, J$. The indicator function $I(E)$ takes a value of 1 if the event E is true and 0 otherwise. Here we set $n = 100$ for all cases. We illustrate scenarios (c), (e), and (f) with tree diagrams in Fig. 16.1, where X_i denotes the i -th covariate.

Under scenarios (a) and (b), there is actually no subgroup with heterogeneous treatment effects. Under scenarios (c) and (d), there are heterogeneous treatment effects, between group $\{i : I(X_{i1} = 0), 1 \leq i \leq n\}$ and the rest of the population. Since the subgroup is defined by a single covariate, these two approaches are expected to detect X_1 with a high probability. In scenario (d), there are three

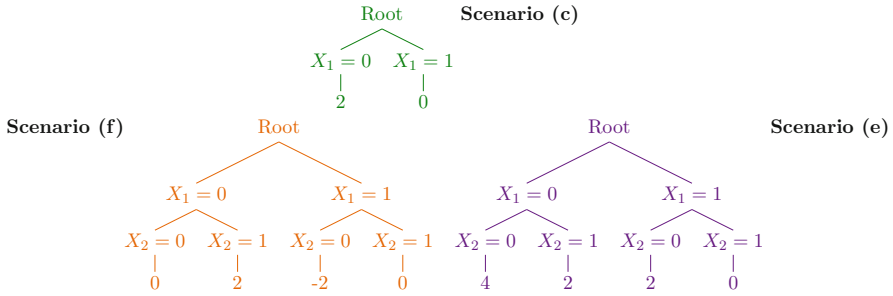


Fig. 16.1 Tree diagrams of scenarios (c), (e), and (f) along with the treatment effect size for each terminal node

prognostic variables X_1 , X_2 , and X_3 , and it is desirable to test how these two methods perform when there are prognostic effects. For scenario (e), there are indeed three subgroups with heterogeneous treatment effects: $g_1 = \{i : I(X_{i1} = 0 \cap X_{i2} = 0), 1 \leq i \leq n\}$, $g_2 = \{i : I[(X_{i1} = 0 \cap X_{i2} = 1) \cup (X_{i1} = 1 \cap X_{i2} = 0)], 1 \leq i \leq n\}$, $g_3 = \{i : I(X_{i1} = 1 \cap X_{i2} = 1), 1 \leq i \leq n\}$, among which only g_3 is the subgroup with zero treatment effect. Under scenario (f), there is also qualitative treatment-subgroup interaction (i.e., there exists both subgroups with positive treatment effects and negative treatment effects), and there are two subgroups, $\{i : I(X_{i1} = 0 \cap X_{i2} = 1), 1 \leq i \leq n\}$ and $\{i : I(X_{i1} = 1 \cap X_{i2} = 0), 1 \leq i \leq n\}$, having non-zero treatment effects. We repeat each simulation independently for 200 times.

For the Bayesian tree method (Berger et al. 2014), the posterior probabilities P_i of having a non-zero treatment effect for individual $i = 1, \dots, n$ are extracted as the major outcome for analysis. The simulations results are reported in Table 16.2. \bar{P}^0 represents the mean of P_i 's of the patients whose treatment effects are 0. On the contrary, \bar{P}^1 represents the mean of P_i 's of the patients whose treatment effects are non-zero. The medians of \bar{P}^0 across different scenarios are shown as the point estimate and the 95% confidence intervals for \bar{P}^0 are displayed below the point estimates correspondingly in the table. Similar results are shown for \bar{P}^1 . Ideally, we should have \bar{P}^0 being close to 0, while \bar{P}^1 being close to 1 if a particular method performs well in distinguishing the patient group with non-zero treatment effects and the subgroup with the zero treatment effect.

For scenarios (c)–(f), since there are both subgroups with zero treatment effects and non-zero treatment effects, we can construct the receiver operating characteristic (ROC) curve and obtain the area under curve (AUC) for each scenario. To construct the ROC curve, we compare P_i with each of these threshold values 0, 0.01, 0.02, \dots , 1.00 to classify individuals to two groups: the group with non-zero treatment effect and the group with zero treatment effect. The AUC values are given in Table 16.2.

From Table 16.2, we see that for scenario (a), the mean of the patient's posterior probability of getting a non-zero treatment effect, \bar{P}^0 , is 0.25, which is noticeably

Table 16.2 Simulation results under scenarios (a)–(f) obtained by the method (Berger et al. 2014)

	Scenario					
	(a)		(b)		(c)	
σ	1	2	1	2	1	2
\bar{P}^0	0.25	0.25	–	–	0.17	0.35
	(0.18, 0.59)	(0.18, 0.59)	–	–	(0.11, 0.59)	(0.18, 0.92)
\bar{P}^1	–	–	1.00	0.997	1.00	0.93
	–	–	(1.00, 1.00)	(0.82, 1.00)	(1.00, 1.00)	(0.27, 1.00)
AUC	–	–	–	–	1.00	0.89

	Scenario					
	(d)		(e)		(f)	
σ	1	2	1	2	1	2
\bar{P}^0	0.97	0.71	0.98	0.52	0.27	0.29
	(0.35, 1.00)	(0.19, 1.00)	(0.46, 1.00)	(0.24, 0.99)	(0.12, 0.97)	(0.15, 0.84)
\bar{P}^1	1.00	0.96	0.97	0.57	0.77	0.74
	(0.70, 1.00)	(0.59, 1.00)	(0.65, 1.00)	(0.24, 0.99)	(0.64, 0.99)	(0.45, 0.93)
AUC	0.77	0.73	0.54	0.50	0.83	0.82

Symbol “–” is deployed when the criterion is not applicable for the cell

smaller comparing to \bar{P}^1 in scenario (b). Results from these two extreme scenarios give us some ideas about the “benchmark value” of P_i , regarding to patients with zero or non-zero treatment effects. Under scenario (c), there is only one binary predictive variable and no prognostic variable, this method performs very well in terms of AUC, and AUC drops a little when the noise level increases from $N(0, 1)$ to $N(0, 4)$. When adding prognostic variables to the model, we see from the results of scenario (d) that the AUCs are much smaller comparing to those under scenario (c). The point estimates of \bar{P}^0 are much closer to 1, which indicates that the method is not able to distinguish the group with zero treatment effect from the others. Since the method only considers up to one prognostic variable, when the prognostic effect structure is more complicated, it will affect the estimates of P_i 's. For scenarios (e)–(f), there is no prognostic variable, while there are subgroups with non-zero treatment effects defined by more than one variable. Since we use the algorithm that allows at most one factor for split in Berger et al. (2014), the performance is not very good as expected. Results from scenario (f) are better comparing to those from (e), since there are more patients with zero treatment effect and it is easier to distinguish this “null group” from others.

Under the model space setting in Sivaganesan et al. (2011), the true models for scenarios (a) and (b) are M_{00} and M_{01} , namely, the overall null and the overall effect model. For scenarios (c) and (d), the true model is M_{11} indicating two subgroups defined by X_1 , and the treatment effects in these two subgroups are zero and non-zero. For scenarios (e) and (f), based on the decision making algorithm, the models expected to be reported are M_{13} and M_{23} , representing there are heterogenous non-zero treatment effects defined by both X_1 and X_2 . For the comparing threshold in

the decision making algorithm, we set $c_0 = c_1 = c$ for simplicity, c varying from 0 to $\exp(25)$. Figures 16.2 and 16.3 show the probabilities of models reported under scenarios (a)–(f) for different values of c when $\sigma = 1, 2$. Note that c is chosen when type I error (TIE) is controlled and power is reached as big as possible. Under scenario (a) where $TIE = 1 - \Pr(M_{00} \text{ is reported} | M_{00})$, we observe from Figs. 16.2 and 16.3 that TIE is controlled at 0.1 for $\log(c) > 1.5$ and TIE is controlled at 0.05 for $\log(c) > 2$ approximately. From Fig. 16.3 when $\sigma = 2$, we notice that under scenarios (c)–(f), the probabilities of reporting true models are obviously lower than the probabilities when $\sigma = 1$. This indicates that selection accuracy of the method (Sivaganesan et al. 2011) is easily affected by data noise. In general, $\log(c) = 2$ controls TIE and achieves relatively high rates of reporting true models,

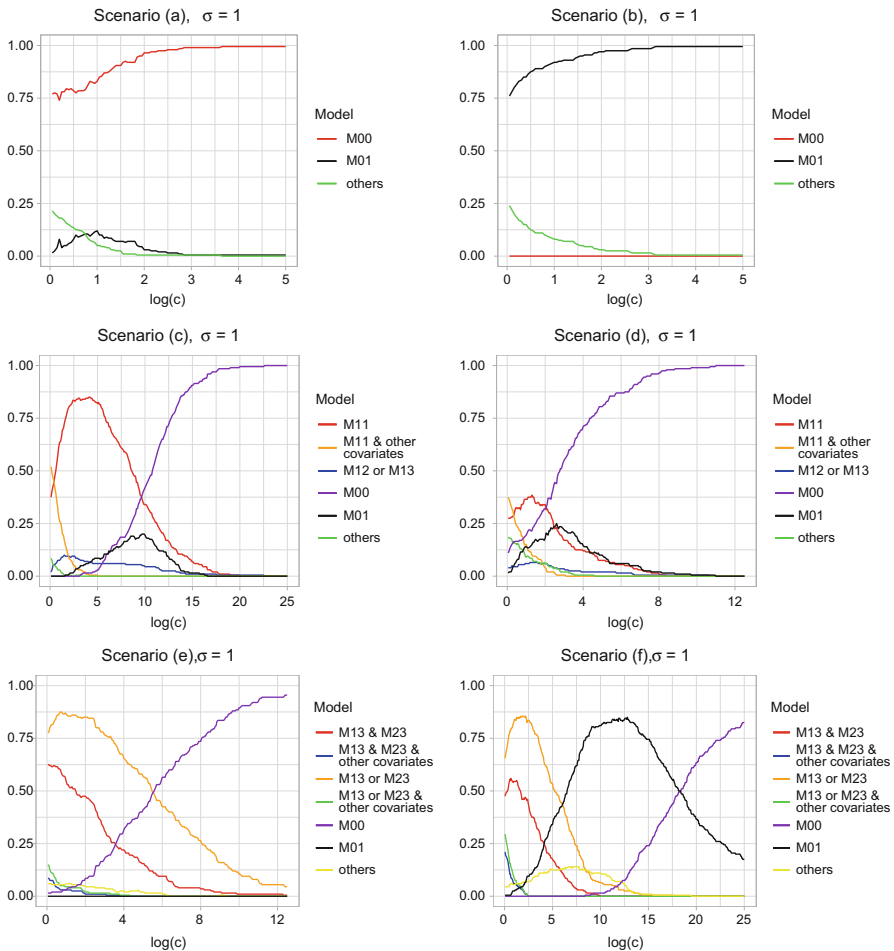


Fig. 16.2 Probabilities of models reported when $\sigma = 1$

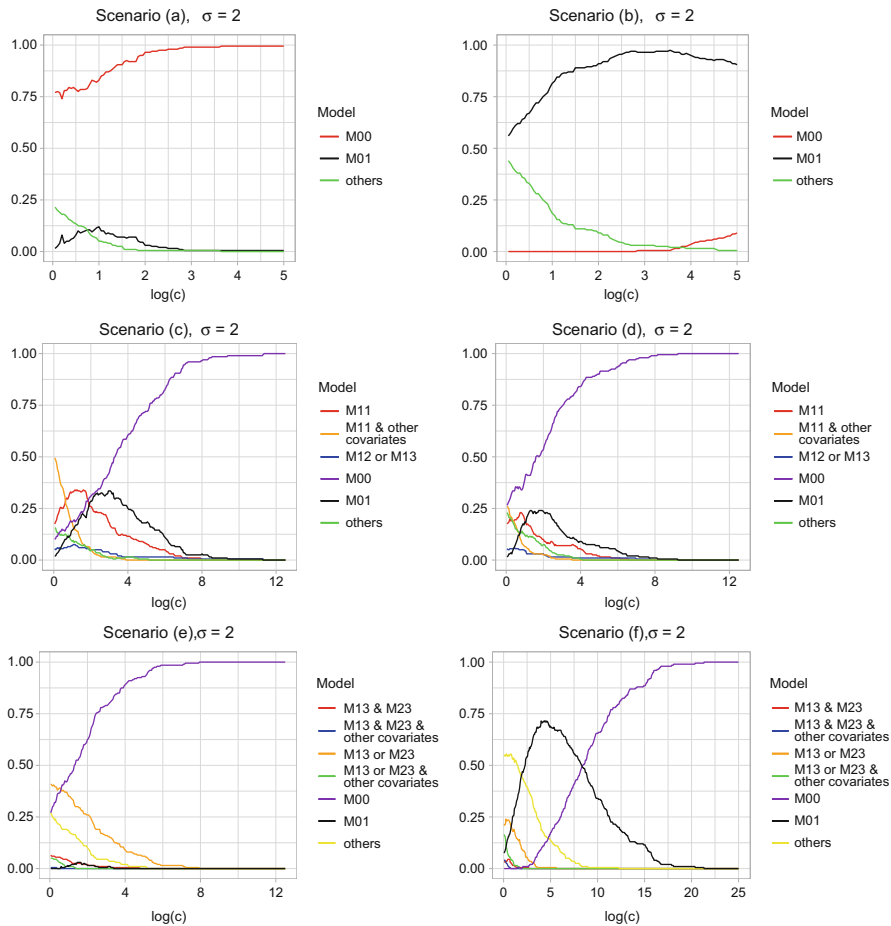


Fig. 16.3 Probabilities of models reported when $\sigma = 2$

therefore we choose $c_0 = c_1 = \exp(2)$ as the comparing threshold in the decision making algorithm, and the simulation results under this threshold value are shown in Table 16.3.

From Table 16.3 it can be seen that the method in Sivaganesan et al. (2011) performs quite well under scenarios (a) and (b) when there are no subgroup treatment effects for $\sigma = 1, 2$. However, by comparing results of scenarios (c) and (d), we find that prognostic variables cause great interference to selection results. This can also be seen from Fig. 16.2 and 16.3, where the reporting probability curve of M_{11} under scenario (d) is always lower than the curve of M_{11} in scenario (c). Under scenarios (e) and (f) when multiple covariates have subgroup effects, though reporting all true subgroup models is difficult, the probabilities of discovering at least one true subgroup model are notably higher.

Table 16.3 Simulation results under scenarios (a)–(f) obtained by the method (Sivaganesan et al. 2011) when $c_0 = c_1 = \exp(2)$

Scenario	(a)		(b)		(c)	
Ture model	M_{00}		M_{01}		M_{11}	
σ	1	2	1	2	1	2
$P(\text{TrueModel})$	0.965	0.965	0.970	0.910	0.805	0.260

Scenario	(d)		(e)		(f)	
Ture model	M_{11}		M_{13}, M_{23}		M_{13}, M_{23}	
σ	1	2	1	2	1	2
$P(\text{TrueModel})$	0.340	0.100	0.475	0.020	0.465	0.005
P^*	–	–	0.850	0.260	0.855	0.090

$P^* = \Pr(\text{At least one true model is reported} \mid \text{multiple subgroup models})$ under scenarios (e)–(f)

16.4 A Real Data Example

We apply the method of Berger et al. (2014), the QUINT (Dusseldorp and Van Mechelen 2014) method (a frequentist approach), and the approach of Sivaganesan et al. (2011) to analyze the Breast Cancer Recovery Project (BCRP) dataset. BCRP dataset is publicly available in the R package *quint*. The test subjects were women with early-stage breast cancer. There were three treatment arms in the randomized trial: a nutrition intervention arm, an education intervention arm, and a standard care arm. We only study the patients from the education intervention (assign $trt = 1$) arm and the standard care arm ($trt = 0$). After removing missing values, we had 146 test subjects left, among which 70 patients were in $trt = 1$ group and 76 patients were in $trt = 0$ group.

The response variable was the improvement in depression score at a 9-month follow-up. There were nine covariates: age, nationality, marital status, weight change, treatment extensiveness index, comorbidities, dispositional optimism, unmitigated communion and negative social interaction, and we dichotomized each continuous or categorical variable by its median value so we can apply the Bayesian methods.

We use the default options to implement QUINT in R, and the final tree is just the trivial tree (i.e., no split is made), which indicates no notable qualitative treatment-subgroup interaction has been found. The posterior probabilities P_i 's of having a non-zero treatment effect for all subjects are between 0.20 and 0.24, which also suggests no findings of subgroup effect. The method in Sivaganesan et al. (2011), where the decision-making is carried out based on $c_0 = c_1 = 2$ from our simulation results, also reports the overall null effect model, which means no subgroup is found. However, QUINT produces a non-trivial tree as reported in Liu et al. (2019) if the variables are not dichotomized. The information loss after dichotomization is also a main disadvantage for methods that are only applicable to binary variables.

16.5 Discussion

Overall, Bayesian subgroup analysis methods add in a lot of varieties and new aspects of thinking to the personalized medicine development. Bayesian methods such as Berger et al. (2014) and Sivaganesan et al. (2011) can provide inference over a model space rather than just one specific model, though it may not be easy to extend and apply these methods to accommodate categorical variables or continuous variables without information loss. In the aforementioned Bayesian tree methods, only the method developed by Zhao et al. (2018) can be applied directly to continuous variables, and considers the issue of splitting point selection implicitly when building the tree. Comparing to frequentist methods, Bayesian methods allow for the incorporation of prior information and expert's inputs as well as account for model uncertainty. Many of the Bayesian methods consider simple prognostic effect structures. When the dimension of the parameter space is high and there are various types of covariates, current Bayesian methods need to be improved to tackle these challenges.

Acknowledgements Mr. Liu's research was partially supported by NIH grant #P01CA142538. Ms. Geng's research was partially supported by NIH grant #GM70335. Dr. Chen's research was partially supported by NIH grants #GM70335 and #P01CA142538.

References

- Berger JO, Wang X, Shen L (2014) A Bayesian approach to subgroup identification. *J Biopharm Stat* 24(1):110–129
- Chen S, Tian L, Cai T, Yu M (2017) A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* 73(4):1199–1209.
- Chipman HA, George EI, McCulloch RE et al (2010) Bart: Bayesian additive regression trees. *Ann Appl Stat* 4(1):266–298
- Dixon DO, Simon R (1991) Bayesian subset analysis. *Biometrics* 47(3):871–881
- Dusseldorp E, Van Mechelen I (2014) Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Stat Med* 33(2):219–237
- Foster JC, Taylor JM, Ruberg SJ (2011) Subgroup identification from randomized clinical trial data. *Stat Med* 30(24):2867–2880
- Gu X, Yin G, Lee JJ (2013) Bayesian two-step lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemp Clin Trials* 36(2):642–650
- Hastie T, Tibshirani R et al. (2000) Bayesian backfitting (with comments and a rejoinder by the authors). *Stat Sci* 15(3):196–223
- Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M (2011) Bayesian models for subgroup analysis in clinical trials. *Clin Trials* 8(2):129–143
- 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. *Stat Med* 30(21):2601–2621
- Liu J, Sivaganesan S, Laud PW, Müller P (2017) A Bayesian subgroup analysis using collections of anova models. *Biom J* 59(4):746–766

- Liu Y, Ma X, Zhang D, Geng L, Wang X, Zheng W, Chen M-H (2019) Look before you leap: Systematic evaluation of tree-based statistical methods in subgroup identification. *J Biopharm Stat*. <https://doi.org/10.1080/10543406.2019.1584204>
- Loh W-Y, He X, Man M (2015) A regression tree approach to identifying subgroups with differential treatment effects. *Stat Med* 34(11):1818–1833
- Schnell PM, Tang Q, Offen WW, Carlin BP (2016) A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics* 72(4):1026–1036
- Sivaganesan S, Laud, PW, Müller P (2011) A Bayesian subgroup analysis with a zero-enriched polya urn scheme. *Stat Med* 30(4):312–323
- Sivaganesan S, Müller P, Huang B (2017) Subgroup finding via Bayesian additive regression trees. *Stat Med* 36(15):2391–2403
- Su X, Tsai C-L, Wang H, Nickerson DM, Li B (2009) Subgroup analysis via recursive partitioning. *J Mach Learn Res* 10:141–158
- Tian L, Alizadeh AA, Gentles AJ, Tibshirani R (2014) A simple method for estimating interactions between a treatment and a large number of covariates. *J Am Stat Assoc* 109(508):1517–1532
- Wang X (2012) Bayesian modeling using latent structures. Ph.D. Thesis, Duke University, Durham
- Zhao Y, Zeng D, Rush AJ, Kosorok MR (2012) Estimating individualized treatment rules using outcome weighted learning. *J Am Stat Assoc* 107(499):1106–1118
- Zhao Y-Q, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR (2015) Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* 102(1):151
- Zhao Y, Zheng W, Zhuo DY, Lu Y, Ma X, Liu H, Zeng Z, Laird G (2018) Bayesian additive decision trees of biomarker by treatment interactions for predictive biomarker detection and subgroup identification. *J Biopharm Stat* 28(3):534–549