A Bayesian Subgroup Analysis with a Zero-Enriched Polya Urn Scheme

Siva Sivaganesan [1], Purushottam W. Laud[2], and Peter Muller[3]

[1]University of Cincinnati

[2]Medical College of Wisconsin

[3] MD Anderson Cancer Center

Abstract

We introduce a new approach to inference for subgroups in clinical trials. The main elements of the proposed approach are the use of a priority ordering on covariates that define potential subgroups, the use of Bayesian model selection methodology, and the use of a threshold on posterior model probabilities to identify subgroup effects for reporting. We control for multiplicities by following a predetermined priority order and by using coherent posterior probabilities across competing models. As usual in Bayesian clinical trial design we compute frequentist operating characteristics, and achieve the desired error probability by choosing an appropriate threshold(s) for the posterior probabilities.

Keywords: Subgroup analysis, Bayesian model selection, frequentist operating characteristics.

1 Introduction

In randomized clinical trials designed to investigate the effectiveness of a new treatment, data are also collected on a number of baseline variables or covariates on the subjects who are admitted to the study. Typically, the primary goal of the clinical trial is to determine the overall effectiveness of the treatment. But patient populations are not homogeneous with respect to many covariates. Thus the question arises as to whether the overall conclusion about the effectiveness of the treatment, or the lack of it, is also valid in some sub-population defined by the covariates. Statistical analysis carried out to answer questions about the treatment effects in sub-populations defined by a covariate (or set of covariates) is called subgroup analysis.

It is important for clinicians to know whether an overall finding about a new treatment applies equally to all patients who satisfy the criteria for inclusion in the trial, or only to some subgroups of patients. Furthermore, regulatory guidelines also encourage or require subgroup analysis [2]. While subgroup analysis is important, there are also several concerns. The main concerns relate to the potential for data dredging, to multiple testing, to lack of power, and to the interpretation of findings. When a plethora of subgroups is available, as is the case in many trials, there can be a tendency to test for subgroup effects in a large number of mostly un-planned subgroups. Spurious significance findings are bound to occur as more tests are carried out. Even when only pre-planned subgroups are tested, studies may not adjust for multiple testing, and report p-values for each subgroup. Often, sample sizes are only adequate for detecting possible overall treatment effect, and true subgroup effects, when present, may not be detected due to insufficient power.

Several recent articles have proposed guidelines on the use and interpretation of subgroup analysis, [2, 8, 9]. Most of these guidelines emphasize the following. Subgroups

to be tested must be pre-specified, motivated by biological reasoning or by results from previous studies. Analysis should be limited to a small number of clinically important subgroups; adjustment must be made for multiple testing; separate tests for individual subgroup effects must be avoided and only tests of subgroup-treatment interaction must be done; and all subgroups tested must be reported clearly distinguishing those planned prior to the study from those decided post-hoc. Interpretations of non-significant subgroup effects must be in light of whether there was sufficient power to detect such effects since, often, a false-negative result may occur due to lack of power.

In this paper, we provide a Bayesian approach to subgroup analysis, which addresses some of the common concerns about subgroup analysis. We assume that the subgroups of interest are defined in terms of covariates such as age, gender, treatment history, biomarkers, etc., and that these covariates can be ordered by the investigators according to their clinical importance. We propose to test the overall hypotheses of interest first, namely the overall null hypothesis that treatment is not effective in the study population as a whole, versus the overall alternative that the treatment is effective. At this stage, we also compare the overall effect model with possible subgroup models. Only when the treatment is declared not effective overall, we proceed to test for subgroup effects one covariate at a time in the order of importance. We continue until a subgroup effect is found or all covariates are exhausted. This approach limits the number of subgroups tested to those deemed as clinically more important. We use a Bayesian model selection approach, representing each possible hypothesis of interest by a model. The posterior probabilities of models associated with each covariate are obtained, one covariate at a time, and a determination is made at each step based on specified thresholds for the posterior probabilities. Prior distributions are carefully chosen to allow reasonable probabilities for the models as well as the parameters under each model. The thresholds for the posterior probabilities are set by matching desired (frequentist) operating characteristics, such as the overall Type

I error.

In [4, 11], the authors have provided a useful Bayesian approach, using a subjective prior. Their approach uses a single model that includes both main effects and treatment-subgroup interaction for all covariates simultaneously, and define a suitable prior distribution for the interaction term(s). 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 are obtained. These results are then used to draw conclusions about the subgroup effects. This is a very useful approach and can be used with a variety of generalized linear models, using MLE results and appropriately specified prior distributions. In a related paper,[10], a hierarchical Bayesian approach is used with different variances for different interaction terms. This approach permits each interaction to shrink to zero, or to be estimated by a non-zero quantity, according to the evidence in the data and aided by the choice of priors. While these approaches have similarities with ours, there are also differences: we use a model selection approach to choose from among models with and without interaction effects, based on the posterior probabilities of these models; we focus on one covariate at a time according to their clinical importance in order to control the number of subgroups tested; and we adjust for multiplicity by controlling the overall Type-I error rate.

The manuscript is organized as follows. Section 2 introduces the question of interest and relevant notation; Section 3 defines the models representing various scenarios of interest concerning the treatment effects; Section 4 describes the assumed sampling model for response variables and the prior distributions for the unknowns, and gives the posterior probabilities of the models; Section 5 introduces the stepwise procedure for the determination of the existence of the overall and subgroup effects; Section 6 addresses error rates and their evaluation, and gives an example using real data; and Section 7 ends with a discussion.

2 Problem Description and Notation

Consider a two-arm clinical trial, where a treatment is compared with a control. Suppose that a continuous response variable is observed on independent random samples of subjects under treatment and control. We let $Y_{0j}, j = 1, ..., J_0$, and $Y_{1j}, j = 1, ..., J_1$ denote the responses under control and treatment, respectively. We assume a $N(\mu_t, \sigma^2)$ distribution as a sampling model for Y_{tj}, $t = 0, 1$.

We make these assumptions without loss of generality. The proposed approach remains valid for any alternative sampling model or primary aim of the trial.

We also assume that the values of I covariates ($X_i,\ i = 1, .., I$) are available on each subject, and that the values of the $i-$th covariate are classified into S_i categories. The covariates $X_1, ..., X_I$ are assumed to be ordered according to their importance, with X_1 (respectively, X_I) being the most (respectively, least) important covariate for which the investigator is interested in determining subgroup effects.

The main question of interest is whether the treatment is effective overall, i.e., $\mu_1 = \mu_0$ or not. In the subsequent subgroup analysis, interest lies in the effectiveness of the treatment among all patients, and among the subgroups of patients defined by each covariate. For instance, suppose that a covariate X is classified into 3 categories. We want to determine whether the treatment is effective based on all the subjects in the study (i.e., full data), and then, determine whether it is effective within any of the three subgroups defined by the three categories of X. If the treatment is found effective in 2 or more of the 3 subgroups, then we ask to determine how the treatment effects differ across these subgroups. We proceed to test for subgroup effects one covariate at a time, and stop when a subgroup effect is found for a covariate. This limits the number of subgroup analyses carried out.

3 Models for Overall and Subgroup Effects

We use a model selection approach to carry out the desired inference about treatment effects, beginning with an overall treatment effect, and then followed by the subgroup effects. First, we consider two competing models to determine if the treatment is effective in the overall population of interest,

$$M_{00} : \delta = \mu_1 - \mu_0 = 0, \quad M_{01} : \delta = \mu_1 - \mu_0 \neq 0.$$

Thus we define the model space at this stage as

$$\mathcal{M}_0 = \{M_{00}, M_{01}\}. \tag{1}$$

3.1 Subgroup effects due to a Single Covariate

We first focus on a single covariate X taking values $1, \ldots, S$. We consider the S subgroups defined by X. Let μ_{0s}, μ_{1s} denote the mean efficacy outcome under control and treatment, respectively, in subgroup s, $s = 1, \ldots, S$, and let

$$\delta_s = \mu_{1s} - \mu_{0s} \text{ for } s = 1, ..., S \tag{2}$$

represent the treatment effect in subgroup s.

Our goal is to identify subgroups which have no treatment effect, and, among those having treatment effects, to characterize how the treatment effects differ across the corresponding levels of the covariate. To this end, we will consider several models representing all such configurations of subgroup effects, and use $\mathcal{M}_{\mathcal{X}}$ to represent the set of all such models.

Indexing Models in $\mathcal{M}_{\mathcal{X}}$

We index the models in $\mathcal{M}_{\mathcal{X}}$ using a vector, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_S)$, of length S. A zero in any position indicates no treatment effect in that subgroup. The nonzero elements are integers ranging from 1 to K, where K is the number of *distinct* nonzero treatment effects among all the subgroups. The integers 1 to K are assigned to elements of $\boldsymbol{\gamma}$ by order of appearance, such that subgroups with common treatment effect values receive the same integer. For example, taking $S = 3$, the model with nonzero and distinct treatment effects in the three subgroups is denoted by $\boldsymbol{\gamma} = (1, 2, 3)$; the model with equal but nonzero effect in the first and third subgroups and a distinct nonzero effect in subgroup 2 is denoted as $\boldsymbol{\gamma} = (1, 2, 1)$. Taking $S = 5$ subgroups, $\boldsymbol{\gamma} = (1, 0, 2, 1, 3)$ represents the case $\delta_2 = 0, \delta_1 = \delta_4 \neq 0, \delta_3 \neq 0, \delta_5 \neq 0, \delta_1 \neq \delta_3, \delta_1 \neq \delta_5, \delta_3 \neq \delta_5$. Thus $\gamma_s, s = 1, .., S$, can be regarded as the cluster membership indicator for the S subgroup effects corresponding to the covariate X.

We will also use $\{\delta_1^\star, \ldots, \delta_K^\star\}$ to indicate the K non-zero distinct treatment effect sizes. We index the unique levels δ_k^* by order of appearance. For example, when $\boldsymbol{\gamma} = (1, 0, 2, 1, 3)$, then $\{\delta_1^\star, \delta_2^\star, \delta_3^\star\} = \{\delta_1, \delta_3, \delta_5\}$. Models will also be denoted in sequence as $M_0, M_1, \ldots, M_H$ with M_0 corresponding to $\boldsymbol{\gamma} = (0, \ldots, 0)$ and M_H to $\boldsymbol{\gamma} = (1, \ldots, 1)$. Explicit specification of the one-to-one correspondence between M_h and $\boldsymbol{\gamma}$ for $0 < h < H$ is not needed for the purposes of this paper. Finally, a note on notation. We will consistently use j to index subjects, s for subsets (i.e., covariate levels), k for clusters of subsets with equal treatment effect, i for covariates and h for models. We will use corresponding upper case letters J, S, K, I and H to denote the range of each of these indicators. Models are alternatively characterized by their cluster membership indicators $\boldsymbol{\gamma}$'s, i.e., there is a one-to-one correspondence between model index h and $\boldsymbol{\gamma}$. Also we will denote cluster sizes by $N_k, k = 0, \ldots, K$.

Counting Models in $\mathcal{M}_{\mathcal{X}}$

Consider the S subgroups defined by the covariate X. To count all possible models in $\mathcal{M}_{\mathcal{X}}$, note that a model puts the S subgroups into clusters and assigns each cluster a distinct treatment effect δ_k including, possibly, zero. The number of distinct ways S distinguishable objects can be put in K clusters (or indistinguishable cells) equals $\mathrm{St}(S,K)$, the Stirling number of the second kind. Since the value zero has a special meaning, each of the $\mathrm{St}(S,K)$ partitions corresponds to $(K+1)$ models obtained by labeling one of the K clusters with $\delta = 0$, plus the model that does not label any of the K clusters with $\delta = 0$. Thus, the number of models, H, is given by

$$H = \sum_{K=1}^{S}(K+1)\mathrm{St}(S,K) = \sum_{K=1}^{S}\frac{(K+1)}{K!}\sum_{i=0}^{K}(-1)^i\binom{K}{i}(K-i)^S \tag{3}$$

Actual counting can be accomplished more readily by the recursion

$$\mathrm{St}(1,1)=1,\ \ \mathrm{St}(S,K)=0 \text{ if } K<1,\ \text{ and } \mathrm{St}(S,K)=K\mathrm{St}(S-1,K)+\mathrm{St}(S-1,K-1)$$

Straight forward calculations yield

S	2	3	4	5	6
H	5	15	52	203	877

4 Probability Models

We assume a sample of J_0 subjects under control with observations, $Y_{0j}, j = 1, ..., J_0$, and another sample of size J_1 subjects, under treatment, with observations $Y_{1j}, j = 1, ..., J_1$ are available. We assume that the sampling distribution under model $\gamma \in \mathcal{M}_X$ consists of independent normal distributions:

$$\begin{aligned} Y_{0j} &\sim N(\mu_{0s}, \sigma^2), \quad s = x_{0j} \\ Y_{1j} &\sim N(\mu_{1s}, \sigma^2), \quad s = x_{1j} \end{aligned}$$

where $x_{0j} = s$, if the covariate X takes value s ($s = 1, ..., S$), for the j-th subject in the control group and similarly for x_{1j}. Recall that $\mu_{1s} = \mu_{0s} + \delta^*_{\gamma_s}$, where $\delta^*_0 = 0$.

Prior distribution for Parameters in a Model

The unknown parameters under a model M are the vector of s control means, $\boldsymbol{\mu}_0 = (\mu_{01}, ..., \mu_{0s})$, the K non-zero distinct treatment effects vector $\boldsymbol{\delta}^\star = (\delta^\star_1, ..., \delta^\star_K)$, and σ^2. We assign mixture g-priors for the $\delta^\star$'s, and noninformative priors for the other parameters which are common to all models. Mixture g-priors have been found to be reasonable non-informative priors in linear model settings, in addition to being computationally easy to work with, [6]. We will use a generic notation $P(\cdot)$ to denote all probability distributions. Thus, conditionally on g, $\boldsymbol{\mu}_0$ and σ^2,

$$\boldsymbol{\delta}^\star \sim N_K(\underline{0}, g\sigma^2 I_K) \tag{4}$$

and

$$P(g, \boldsymbol{\mu}_0, \sigma^2) \propto \frac{1}{(1+g)^2} \cdot \frac{1}{\sigma^2}, \qquad g > 0.$$

Probability distribution on the Model Space $\mathcal{M}_\mathcal{X}$

The models represent the hypotheses of interest related to the subgroup effects. In particular, a model M designates each subgroup effect as either zero or non-zero, and, when two or more subgroups are non-zero, whether there are ties in the treatment effects. The model space $\mathcal{M}_X$ is the space of all partitions of $\{1, \ldots, S\}$ with additional labeling of one cluster as the special zero cluster with $\delta = 0$.

We assign probabilities for models using an extension of the Polya urn, [1, 7], extended to allow the labeling of one cluster as the zero cluster. Let $K_s = \max\{\gamma_{s'} \neq 0;\ s' \leq s\}$ denote the number of distinct non-zero treatment effects among the first s covariate

levels, with $K_S = K$ for the full set of γ_s's. Also, when $K_s \geq 1$ and $1 \leq k \leq K_s$, let $N_{sk} = \#\{s' : \gamma_{s'} = k,\ s' \leq s\}$ and $L_s = \#\{s' : \gamma_{s'} \neq 0,\ s' \leq s\}$ denote, respectively, the number of treatment effects that match the k-th distinct non-zero effect and the total number of non-zero treatment effects, among the first s levels of the covariate. Then let

$$\begin{aligned} P(\gamma_{s+1} = 0 \mid \gamma_1, \ldots, \gamma_s) &= p \\ P(\gamma_{s+1} = k \mid \gamma_1, \ldots, \gamma_s) &= (1-p)\,\tfrac{N_{sk}q}{1-q+L_sq} \quad \text{for } k = 1, \ldots, K_s \geq 1 \\ P(\gamma_{s+1} = K_s + 1 \mid \gamma_1, \ldots, \gamma_s) &= (1-p)\,\tfrac{1-q}{1-q+L_sq} \quad \text{for } K_s \geq 0 \end{aligned} \tag{5}$$

Conditional on non-zero treatment effects, the last two lines define a Polya urn with total mass parameter $\alpha = (1-q)/q$. With probability proportional to $q \cdot L_s$ the treatment effect for the $(s+1)$-st level of the covariate is tied with an earlier level. With probability proportional to $1-q$ the treatment effect is distinct.

We will regard p and q as unknowns, and assign independent Beta priors

$$P(p, q) = Beta(p\,; \alpha_1, \delta_1) Beta(q\,; \alpha_2, \delta_2), \tag{6}$$

where $Beta(x\,; a, b)$ indicates the probability density function of a random variable x with parameters a and b.

To specify the prior probability of a model $M \in \mathcal{M}_X$, given p and q, let, for $0 \leq k \leq K \leq S$

$$N_k = \#\{s : \gamma_s = k, 1 \leq s \leq S\} \tag{7}$$

so that $\sum_{k=0}^{K} N_k = S$ and $N_k = N_{Sk}$. For example, N_0 is the number of subgroups with zero treatment effect, and N_1 is the number of subgroups with treatment effects that are equal to the first non-zero subgroup effect, and so on. The process of specifying the probability of a model M indexed by $\boldsymbol{\gamma}$ can be thought of as filling in each of the S positions in $\boldsymbol{\gamma}$ by 0 with probability p, or by a positive integer with probability $1-p$. The positive integers, which identify the configuration of the non-zero subgroup effects, are chosen successively according to the probability specification in (5).

Then prior probability of a model M, given p and q is

$$P(M \mid p, q) = c(p, q)\, P(N_0, \ldots, N_K \mid p, q) \tag{8}$$

where

$$P(N_0, \ldots, N_K \mid p, q) = p^{N_0}(1-p)^{S-N_0} \frac{\alpha^K \prod_{k=1}^{K} [N_k - 1]!}{\prod_{s=1}^{S-N_0} \{\alpha + (s-1)\}}, \tag{9}$$

$[x] = \max\{x, 0\}$, N_k's are as in (7), $c(p, q)$ is the normalizing constant, and a product over an empty set is equal to 1.

The probability model (8) is a zero-enriched Polya urn scheme. Each integer in the index vector $\boldsymbol{\gamma}$ of a model M is allowed to be zero with probability p ; and each non-zero integer is either equal to a previously selected value or to the subsequent (hitherto unselected) value.

The normalizing constant $c(p, q)$ is determined by counting the number of different models corresponding to a given K and $(N_0, ..., N_K)$,

$$c(p, q)^{-1} = \sum_{K=0}^{S} \sum \binom{S}{N_0} P(N_0, \ldots, N_K \mid p, q)$$

where the inside summation spans over all integers $N_0 \geq 0, N_1 > 0, ..., N_K > 0$ satisfying the condition $N_0 + N_1 + ... + N_K = S$ and $P(N_0, \ldots, N_K \mid p, q)$ is as in (9).

Posterior Probability of Models

Let $\underline{y}$ denote the combined sample data. The conditional independence given the model implies $P(\underline{y}|M, p, q) = P(\underline{y}|M)$. Hence, the posterior model probability of a model M is

$$P(M \mid \underline{y}) = \frac{P(\underline{y}|M)P(M)}{\sum_{M' \in \mathcal{M}} P(\underline{y}|M')P(M')}$$

where

$$P(M) = \int P(M|p, q)P(p, q)dpdq \tag{10}$$

In the above,

$$P(\underline{y}|M) = \int P(\underline{y}|M, g)P(g)dg,$$

where

$$P(\underline{y}|M, g) = \int P(\underline{y} \mid \boldsymbol{\mu}_0, \boldsymbol{\delta}^*, \sigma^2)P(\boldsymbol{\delta}^*|g, \sigma^2)P(\boldsymbol{\mu}_0, \sigma^2)d\boldsymbol{\mu}_0 d\boldsymbol{\delta}^* d\sigma^2.$$

Although closed form expression for $P(\underline{y}|M)$ is not available, it can be expressed as a one-dimensional integral over g.

4.1 A Special Case : $S = 2$

Suppose that X is categorized into two levels defining two subgroups. We are interested in whether the treatment is effective in either subgroup, and when it is in both subgroups, whether the effect sizes are equal. Here, the model space $\mathcal{M}_{\mathcal{X}}$ consists of five models. These are listed in Table 1 along with their indices and prior probabilities.

Model		Index $\boldsymbol{\gamma}$	$P(M\|p, q)$
M_0 :	$\delta_{11} = \delta_{12} = 0$	$(0, 0)$	p^2
M_1 :	$\delta_{11} \neq 0, \delta_{12} = 0$	$(1, 0)$	$p(1-p)$
M_2 :	$\delta_{11} = 0, \delta_{12} \neq 0$	$(0, 1)$	$p(1-p)$
M_3 :	$\delta_{11} \neq \delta_{12} \neq 0$	$(1, 2)$	$(1-p)^2(1-q)$
M_4 :	$\delta_{11} = \delta_{12} \neq 0$	$(1, 1)$	$(1-p)^2 q$

Table 1: Models and their prior probabilities, given p and q, for the case $S = 2$.

4.2 Multiple Covariates

Suppose that there are I covariates of interest X_1,...,X_I, and that X_i is classified into S_i categories. As in the previous section, we define a model space $\mathcal{M}_{X_i} = \mathcal{M}_i$, for each covariate X_i. The number of models, say H_i, in $\mathcal{M}_i$, and their probabilities under the model space $\mathcal{M}_i$, are defined as in (3) and (8), with S replaced by S_i. In summary, we define I distinct probability models, one for each family $\mathcal{M}_i$ of models.

Note that our primary hypotheses of interest are indicated by the overall null and alternative models M_{00} and M_{01}, which constitute the space $\mathcal{M}_0$ in (1). This overall effect model space $\mathcal{M}_0$, along with $\mathcal{M}_i, i = 1, ..., I$ define the collection of models that are of interest in the subgroup analysis.

We find it useful to list the models in $\mathcal{M}_i$ in sequence, labeling the h^{th} model by M_{ih}, $0 \leq h \leq H_i$. When the index i is understood from the context we write simply H. Each model is characterized by a vector $\boldsymbol{\gamma}$ of length S_i, as in Section 3.1. The labeling of the models in $\mathcal{M}_i$ is done so that the first model ($h = 0$) corresponds to the absence of subgroup effects in all s_i subgroups and is represented by $\boldsymbol{\gamma} = (0, ..., 0)$, and the last model ($h = H$) corresponds to the presence of subgroup effects of equal size in all subgroups and is represented by $\boldsymbol{\gamma} = (1, ..., 1)$. These two models are equivalent to the overall null and alternative models M_{00} and M_{01}, respectively. Thus, each of the remaining models in $\mathcal{M}_i$ represents the presence of a "bona-fide" subgroup effect. For instance, when $S_1 = 2$, M_{10} and M_{14}, respectively, represent the overall null and alternative hypotheses, and are equivalent to M_{00} and M_{01}. Table 1 lists these omitting the first subscript $i = 1$ since $I = 1$ in this context. The remaining three models, namely, M_{11}, M_{12} and M_{13}, represent presence of subgroup effects.

Prior and Posterior Probabilities

When calculating the posterior probability of a model $M \in \mathcal{M}_i$, we regard $\mathcal{M}_i$ as the model space, and use the prior probabilities as defined in Section 4. We will use the notation $P_i(M) \equiv P(M \mid \mathcal{M}_i)$ and $P_i(M \mid y) \equiv P(M \mid y, \mathcal{M}_i)$ for the prior and posterior probabilities of M, conditional on $M \in \mathcal{M}_i$, $i = 0, ..., I$. Thus, for $M^* \in \mathcal{M}_i$,

$$P_i(M^* \mid y) = \frac{P(y|M^*)P_i(M^*)}{\sum_{M \in \mathcal{M}_i} P(y|M)P_i(M)} \tag{11}$$

5 A Stepwise Procedure to Decide on Subgroup Effects

Recall that our goal is to first compare the overall null and overall effect models (i.e., models in $\mathcal{M}_0$), and to determine if a subgroup effect is present. Also recall that the covariates are labeled according to their importance to the investigator, in descending order.

The proposed procedure is carried out in steps. In the initial step (Step 0), we first compare the models in $\mathcal{M}_0$. If the overall effect model is preferred over the null model by a sufficient margin, the former is further compared with subgroup models. Then, if the overall effect model is not selected, we continue to focus on model spaces representing the subgroup effects, $\mathcal{M}_1$, $\mathcal{M}_2$,..., and $\mathcal{M}_I$, in that order. We stop either when the overall effect model (at Step 0) or a model with subgroup effect (at Step $i \geq 1$) is selected, or when all such models are exhausted. The overall null model is chosen at Step I when we fail to choose any of the models representing subgroup effects. At Step i, $i \geq 1$, the determination of the presence (or the absence) of a subgroup effect is made based on the posterior probabilities of the models in the model

space $\mathcal{M}_i$. Specifically, we use two threshold values c_0 and c_1, $0.5 < c_1 \leq c_0 < 1$, for the posterior probabilities. We select a model representing a subgroup effect if it is the most likely model and, in head to head competition with the null model M_{i0} (resp. with the overall effect model M_{iH}), its posterior probability exceeds c_0 (resp. c_1). Recall that M_{i0} and H_{iH} are the representations in $\mathcal{M}_i$ of the overall null and the overall effect model, respectively. Below, we give an algorithmic description of the proposed stepwise procedure.

Step 0 :

Choose M_{01} and stop if

$$\begin{aligned} P_0(M_{01}|y) &> c_0 \text{ and} \\ P_i(M_{ih}|M_{ih} \cup M_{iH_i}, y) &< c_1 \text{ for } 1 \leq i \leq I \text{ and } 0 < h < H, \end{aligned}$$

else continue to next step.

Step i : for $i = 1, ..., I-1$

Choose M_{ih} for $0 < h < H_i$, and stop if

$$\begin{aligned} P_i(M_{ih} \mid y) &= \max_{0<h'<H_i} P_i(M_{ih'}|y) \text{ and} \\ P_i(M_{ih} \mid M_{ih} \cup M_{i0}, y) &> c_0, \text{ and} \\ P_i(M_{ih}|M_{ih} \cup M_{iH_i}, y) &> c_1 \end{aligned}$$

else continue to next step.

Step I :

Choose M_{Ih} and stop if

$$\begin{aligned} P_I(M_{Ih} \mid y) &= \max_{0<h'<H_I} P_I(M_{Ih'}|y) \text{ and} \\ P_I(M_{Ih}|M_{Ih} \cup M_{I0}, y) &> c_0, \end{aligned}$$

else choose M_{00} and stop.

The following is a schematic description when there are two covariates, each at two levels, i.e., $I = 2$, $S = 2$, and the number of models at each stage is 5, i.e., $H = 4$.

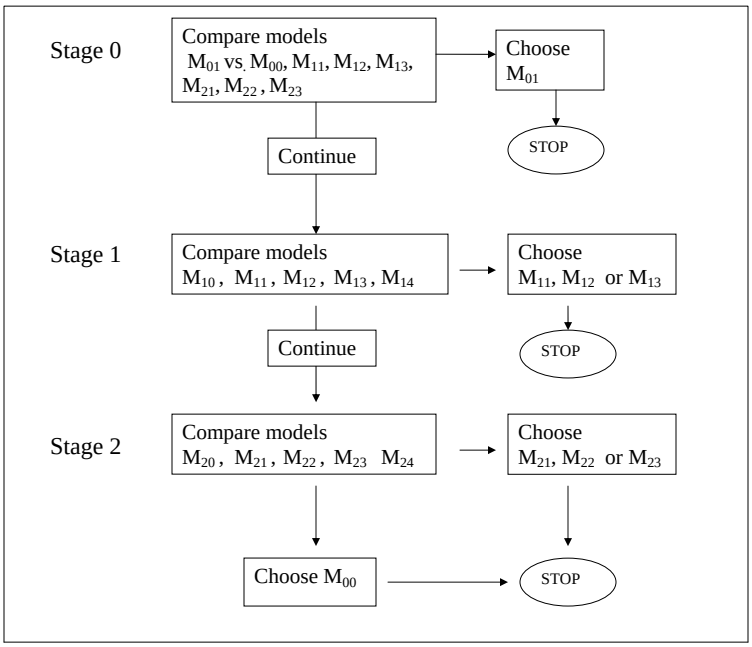


Figure 1: The stepwise procedure, when $k = 2$ and $N = 4$

5.1 Important Aspects of the Stepwise Procedure

The above procedure is designed to incorporate two important features. One is the adjustment for multiplicity via control of the important error rates, namely Type I error and the error of choosing a subgroup effect model when the overall effect model is true. The other is to maintain good error rates by controlling the number of subgroups tested.

At Step 0, a determination of whether or not to choose the overall effect model is made. The overall effect model is chosen, only if there is strong evidence in its favor

in comparison to the overall null model, and if there is no strong evidence against it in comparison to any single subgroup effect model. This serves to control two types of errors. When a model representing a subgroup effect (in $\mathcal{M}_i, i > 0$) is true, simply comparing M_{00} and M_{01} alone could result in a very high posterior probability for M_{01} (if this is the one "closer" to the true model) and lead to choosing the wrong model M_{01}. Comparing the overall effect model with the subgroup effect models (in addition to the null model) protects against this error. On the other hand, it is also important to control the error of choosing a subgroup effect model when in fact the overall effect model is true; this can be achieved by requiring sufficiently large evidence against the overall effect model, i.e., by choosing an appropriately large value for c_1.

The values of c_0 and c_1 can be chosen so that the overall Type-I error is equal to a pre-determined value, and the average probability of choosing a wrong subgroup effect model when the overall effect model is true is also small. Values of c_0 and c_1 can be determined via simulation. Typically, the value of c_1 has no bearing on the Type-I error rate. It relates to the amount of evidence required for choosing a subgroup effect model over the overall effect model. A reasonable value for c_1 would be between 0.5 and c_0, with a larger value of c_1 indicating a higher threshold for evidence required for choosing a subgroup effect model, or higher penalty for choosing a subgroup effect model when the overall effect model is true. As discussed in Section 1, adjusting for multiple testing and controlling the error of choosing a subgroup effect when there is an overall effect are important goals in a subgroup analysis; the approach here is specifically geared to achieving these goals.

The procedure proposed above stops when the overall effect model or a subgroup effect model is selected, and continues only when a model representing a treatment effect is not selected. This precludes the procedure from searching for additional subgroup effects at lower levels (i.e., with less important covariates) when a subgroup effect is found at a higher level covariate, thus limiting the number of subgroups tested.

6 Operating Characteristics and Example

In this Section we define error rates relevant to subgroup analysis and study the operating characteristics of the proposed procedure. We do this in the context of an example.

6.1 Introduction to STI Example

Kovach et al.(2006) reported on a double-blinded randomized experiment to study the effectiveness of Serial Trial Intervention (STI), the treatment, on comfort and behavior. The study was conducted in 14 nursing homes on 114 subjects with late-stage dementia. Serial Trial Intervention (STI) is an innovative clinical protocol for assessment and management of unmet needs in people with late-stage dementia. Outcome variable of interest was the difference, pre and post intervention, in Discomfort-DAT (Discomfort-Dementia of the Alzheimer's Type scale), a measure of discomfort felt by the subjects. The sample sizes for the treatment and control were 55 and 57, respectively. The investigators were interested in the subgroups (personal communication) defined by two covariates, Functional Assessment Staging of Dementia (FAST) score (covariate X_1) and Presence/Absence of Vocalization in Behavioral Symptoms (MVOCAL) initiating treatment (covariate X_2). Two subgroups were of interest based on X_1, defined by $X_1 = 1$ when FAST Score ≥ 7, and, =0 if FAST score < 7. The presence ($X_2 = 1$) and absence ($X_2 = 0$) of vocalization in behavioral symptoms initiating treatment also define two subgroups of interest. Here we rank the covariate X_1 as more important than X_2 for finding subgroup effects. Subgroup sample sizes are shown in Table 2.

Thus we have $I = 2$ covariates, each at $S = 2$ levels. There are eight models,

	$X_1 = 0$	$X_1 = 1$	Total	$X_2 = 0$	$X_2 = 1$
Control	31	26	57	19	38
Treated	35	20	55	19	36
Total	66	46	112	38	74

Table 2: Subgroup Sample Sizes

altogether, the overall null model M_{00}, the overall effect model M_{01}, three subgroup effect models M_{11}, M_{12}, and M_{13} corresponding to FAST (as defined in Section 4.1), and three subgroup effect models M_{21}, M_{22}, and M_{23} corresponding to MVOCAL.

6.2 Error Rates

Since the proposed procedure selects a model from many competing models, several error rates of interest can be defined. Of prime importance among these is the Type I Error (TIE) defined as the probability, in repeated experiments, of rejecting the overall null (M_{00}) by selecting any other model when the true model is M_{00}. Other error rates pertain to probabilities under the overall effect model and each subgroup model. While there are many possibilities, we focus on the following definitions, where P_f represents probability under repeated experiments.

TIE	:	Type I Error	$P_f(M_{00}$ not selected$\|M_{00})$
FNR	:	False Negative Rate	$P_f(M_{00}$ selected$\|M_{01})$
TPR	:	True Positive Rate	$P_f(M_{01}$ selected$\|M_{01})$
FSR	:	False Subgroup Rate	$P_f(\text{some } M_{ih}, i \neq 0, h \neq 0, h \neq H_i \text{ selected}\|M_{01})$
TSR	:	True Subgroup Rate	$P_f(M_{ih}$ selected$\|M_{ih}, i \neq 0, h \neq 0, h \neq H_i)$
FPR	:	False (overall) Positive Rate	$P_f(M_{01}$ selected$\|M_{ih}, i \neq 0, h \neq 0, h \neq H_i)$

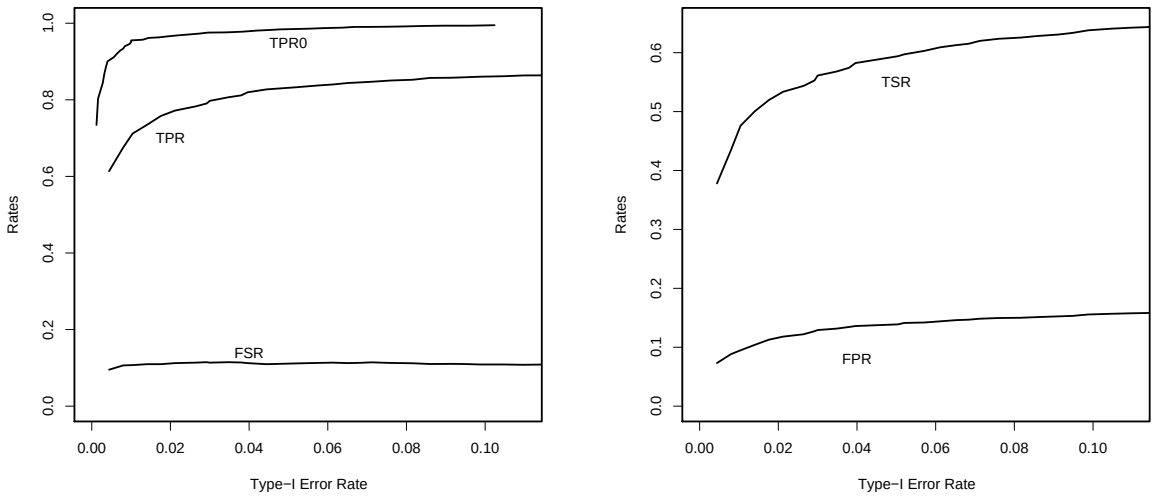
Rates		Truth		
		Overall Null	Subgroup Effect	Overall Alternative
		Model M_{00}	Model M_{ih}	Model M_{01}
Decision	M_{00}	1-TIE		FNR
	M_{ih}		TSR_{ih}	FSR
	M_{01}		FPR_{ih}	TPR

Table 3: Error Rates

It is important to note that all but the TIE require additional specifications for proper definitions. An effect size is required for FNR, TPR and FSR. Moreover, TSR and FPR further depend on i, h. These definitions lead to Table 3 summarizing the various possibilities.

These error rates and the operating characteristics of the procedure can be evaluated via simulation over repeated datasets. For the STI example in the previous subsection, we used the sample sizes in Table 2 when generating simulated datasets. For each of the many settings needed to construct Figures 2 and 3, 1000 datasets were generated. In all cases we used independent $Beta(.5, .5)$ priors for p and q.

The two lower curves in Figure 2(a) show the estimated True Positive Rate(TPR) and the False Subgroup Rate(FSR), both calculated by simulating data from a specific overall effect (true) model. The standardized effect size for this model was set at 0.8. The uppermost curve marked TPR0 results from a simple procedure without subgroup analysis, i.e., this shows the traditional power at various levels of significance. The Type I Error (TIE) was varied by choosing $c_1 = 0.7$ and c_0 in the range from 0.78 to 0.99. At TIE of 0.05, TPR is 0.83, FSR is 0.12 and TPR0 is 0.98. These quantify the tradeoffs involved in planning a subgroup analysis as opposed to a simple overall effect analysis without considering subgroups.

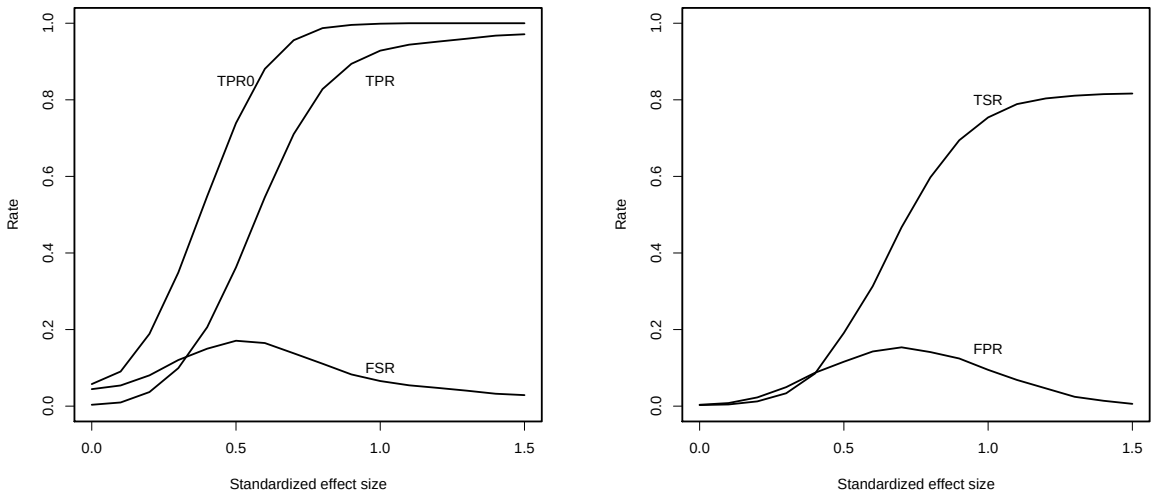


(a) Simulation under M_{01} (b) Simulation under M_{11}, $\boldsymbol{\gamma} = (1, 0)$

Figure 2: Operating Characteristic Curves for STI Example

The curves in Figure 2(b) were obtained with data from a particular subgroup model, namely, $M_{11} = (1, 0)$ representing an effect in the subgroup with FAST score less than 7. The standardized effect size was set to 0.8 again. The TSR curve represents the rate of correctly choosing the subgroup effect model while the FPR curve represents the rate of incorrectly concluding the overall effect model M_{01}. Note that 1−(TSR+FPR) is the rate of incorrectly accepting the overall null or concluding a different subgroup model.

Figure 3 addresses issues similar to those in traditional power curves, plotting rates against various standardized effect sizes. The TIE was fixed at 0.05 by choosing $c_0 = 0.9$ and $c_1 = 0.7$. In Figure 3(a) the true model is the overall effect model M_{01}. As in Figure 2(a), TPR and FSR are plotted along with the traditional power curve TPR0 for comparison. The curves in Figure 3(b) are computed under the same true subgroup model as in Figure 2(b), i.e., $M_{11} = (1, 0)$.



(a) Simulation under M_{01} (b) Simulation under M_{11}, $\boldsymbol{\gamma} = (1, 0)$

Figure 3: Rates vs. Effect Size, STI Example

6.3 Results for STI Example

Returning to the STI study, we used the experimental data and independent $Beta(.5, .5)$ priors for p and q. Using simulation, the cut-off value c_0 for the posterior probability was determined to be 0.9 to correspond to an overall Type-I error rate of 0.05. To address the FSR, one would need to study it as a function of the effect size. However, it is possible to define and control the value of average FSR, averaged over a reasonable effect size distribution. We used the normal distribution for the effect size with mean 0 and standard deviation equal to that of the data. Then, two values of c_1 were chosen to correspond to the average FSR values of 0.06 and 0.05. Thus the tuning parameters in the procedure were chosen to control the probability of incorrectly rejecting the overall null and the probability of incorrectly picking a subgroup model when the overall effect model is true.

Implementation of the stepwise procedure resulted in the posterior probabilities as given in Table 4, and selection of the models, as given in Table 5. The result shows that STI has an overall effect, when the two error rates are set at 0.05. When the second error rate is set at 0.06, there is evidence of a subgroup effect for MVOCAL in the sense that treatment effect is absent when MVOCAL=0 and it is present when MVOCAL=1.

Model Space	$\mathcal{M}_0$				
Models	M_{00}				M_{01}
Posterior Probability	0.009				0.991
Model Space	$\mathcal{M}_1$				
Models	M_{10}	M_{11}	M_{12}	M_{13}	M_{14}
Posterior Probability	0.0078	0.0025	0.2273	0.3774	0.3850
Model Space	$\mathcal{M}_2$				
Models	M_{20}	M_{21}	M_{22}	M_{23}	M_{24}
Posterior Probability	0.0003	0.0002	0.6026	0.3039	0.0930

Table 4: Posterior Probabilities of overall and subgroup effects Models for Kovach(2006) data. Model spaces $\mathcal{M}_1$ and $\mathcal{M}_2$ correspond to the covariates FAST and VBS, respectively.

c_0	c_1	Model Selected	Average FSR
0.90	0.89	Overall Effect Model, M_{01}	0.05
0.90	0.80	Subgroup Effect Model, M_{22}	0.06

Table 5: Overall or Subgroup effect model as selected by the procedure in Section 5 for Kovach et al.(2006) data. The cut-off value c_0 corresponds to an overall Type-I error rate of 5%.

7 Discussion

We have presented a method for doing subgroup analysis that takes account some of the major concerns and recommendations. A Bayesian model selection approach is used to determine the presence (or absence) of treatment-subgroup interactions. For the subgroups defined by each covariate, we represented the treatment-subgroup interactions of interest by a set of models, and calculated their posterior probabilities. Determination of the presence (or absence) of a treatment-subgroup interaction is done based on some pre-specified threshold values for the posterior probabilities. These thresholds are set to yield a specified overall Type-I error rate. Adjustment for multiplicity is thus achieved by controlling the overall Type-I error rate. We limit the number of subgroup effects tested by focusing on the covariates one at a time in the order of their importance as determined by the investigator, and by not looking further when a treatment-subgroup interaction for a covariate is deemed as present. Falsely finding a subgroup effect, when the primary hypothesis of overall effect is true, is also a concern. The approach proposed can safeguard against this by setting a suitably high threshold value for the posterior probability used in this comparison, which may be set to correspond to a low value of a weighted average of this error rate, as illustrated in the example.

The proposed method requires the investigator(s) to rank-order the covariates according to their clinical importance. While this is helpful in curbing the number of subgroup effects tested, one may possibly find two or more covariates as equally important, such as Age and Gender. In such instances, our recommendation is to consider the two different model spaces, one corresponding to each covariate, at the same step. This would mean comparing the models in each space separately, and continuing to the next step (only) when no subgroup effects found for either covariate. This approach can thus yield conclusions such as there is a subgroup effect due to

Age and Gender. If, on the other hand, Age-Gender combinations are also interest, one may form a single covariate using these combinations and proceed as before.

Acknowledgement

We thank Christine Kovach, PhD, RN of the University of Wisconsin-Milwaukee and Brent Logan, PhD of the Medical College of Wisconsin for providing advice and the data from their study. This research was initiated during a program at SAMSI (Statistical and Applied Mathematical Sciences Institute, NC).

References

[1] Blackwell, D. and MacQueen, J. B. Ferguson Distributions via Polya urn schemes. Ann. Stat. 1:353-355, 1973.

[2] Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. Med J Aust. 2004 Mar 15;180(6):289-91.

[3] Collins R, Gray R, Godwin J, Peto R. Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews. Stat Med. 1987 6(3):245254.AprMay

[4] Dixon DO, Simon R. Bayesian subset analysis. Biometrics 1991; 47:871882

[5] Kovach CR , Logan BR ,Noonan PE ,Schlidt AM ,Smerz J ,Simpson M ,Wells T. Effects of the Serial Trial Intervention on discomfort and behavior of nursing home residents with dementia. American Journal of Alzheimer's Disease and Other Dementias 2006; Vol. 21, No. 3, 147-155.

[6] Liang F, Paulo R, Molina G, Clyde MA, and Berger JO. Mixtures of g-priors for Bayesian Variable Selection. ISDS Discussion Paper 2005-12.

[7] Pitman, J. Some Developments of the Blacwell-MacQueen Urn Scheme. Statistics, and Probability and Game Theory. IMS Lecture notes 2006, Vol. 30.

[8] Pocock SJ, Assmann SF, Enos LE et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. Statist. Med. 2002; 21:29172930

[9] Rothwell, PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005; 365: 17686

[10] Sargent, DJ. and Hodges, JS. Smoothed ANOVA with application to Subgroup Analysis. Research Reprt rr2005-018, Department of Biostatistics, University of Minnesota.

[11] Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. Statistics in Medicine 2002; 21:29092916.