Controlling the False Discovery Rate in Two-Stage Combination Tests for Multiple Endpoints

Sanat K. Sarkar, Jingjing Chen and Wenge Guo

May 29, 2011

Sanat K. Sarkar is Professor and Senior Research Fellow, Department of Statistics, Temple University, Philadelphia, PA 19122 (Email: sanat@temple.edu). Jingjing Chen is PhD candidate, Department of Statistics, Temple University, Philadelphia, PA 19122 (Email: jjchen@temple.edu). Wenge Guo is Assistant Professor, Department of Mathematical Sciences New Jersey Institute of Technology, Newark, NJ 07102 (Email: wenge.guo@gmail.com). The research of Sarkar and Guo were supported by NSF Grants DMS-1006344 and DMS-1006021 respectively.

ABSTRACT

We consider the problem of simultaneous testing of null hypotheses associated with multiple endpoints in the setting of a two-stage adaptive design where the hypotheses are sequentially screened at the first stage as being rejected or accepted based on boundaries on the false discovery rate (FDR) and the remaining null hypotheses are tested at the second stage having combined their p-values from the two stages through some combination function. We propose two procedures to control the false discovery rate (FDR), extending the Benjamini-Hochberg (BH) procedure and its adaptive version incorporating an estimate of the number of true nulls from single-stage to two-stage setting. These procedures are theoretically proved to control the FDR under the assumption that the pairs of first- and second-stage p-values are independent and those corresponding to the null hypotheses are identically distributed as a pair (p_1, p_2) satisfying the p-clud property of Brannath et al. (2002). We consider two types of combination function, Fisher's and Simes', and present explicit formulas involving these functions

towards carrying out the proposed procedures based on pre-determined critical values or through estimated FDR's. Our simulation studies indicate that the proposed procedures can have significant power improvement over the single-stage BH procedure, and can continue to control the FDR under some dependence situations. Application of the proposed procedures to a real gene expression data set produces more discoveries compared to the single-stage BH procedure.

Keywords: Combination test; early rejection and acceptance boundaries; false discovery rate; multiple testing; stepwise multiple testing procedure; two-stage adaptive design.

1 INTRODUCTION

Gene association or expression studies that usually involve a large number of endpoints (i.e., genetic markers) are often quite expensive. Multi-stage adaptive design with its feature of being cost effective and efficient, since genes are being allowed to be screened in early stages and selected genes are being further investigated in later stages using additional observations, has become more and more attractive in such genetic studies. To address the multiplicity concern in simultaneous testing of the hypotheses associated with the endpoints, controlling the familywise error rate (FWER), the probability of at least one type I error among all hypotheses, is a commonly applied concept. However, these studies are often explorative, so controlling the false discovery rate (FDR), which is the expected proportion of type I errors among all rejected hypotheses, is more appropriate than controlling the FWER (Weller et al., 1998; Benjamini and Hochberg, 1995; and Storey and Tibshirani, 2003). Moreover, with large number of hypotheses typically being tested in these studies, better power can be achieved in a multiple testing method under the FDR framework than under the more conservative FWER framework.

Adaptive designs with multiple endpoints have been considered in the literature under both the FWER and FDR frameworks. Miller et al. (2001) suggested using a two-stage design in gene experiments, and proposed using the Bonferroni method to control the FWER in testing the hypotheses selected at the first stage, although only the second stage observations are used for this method. This was later improved by Satagopan and Elston (2003) by incorporating the first stage data through group sequential schemes

in the final Bonferroni test. Zehetmayer et al. (2005) considered a two-stage adaptive design where promising hypotheses are selected using a constant rejection threshold for each p-value at the first stage and an estimation based approach to controlling the FDR asymptotically (as the number of hypotheses goes to infinity) was taken (Storey, 2002; Storey et al., 2004) at the second stage to test the selected hypotheses using more observations. Zehetmayer et al. (2008) have extended this work from two-stage to multi-stage adaptive designs under both FDR and FWER frameworks, and provided useful insights into the power performance of optimized multi-stage adaptive designs with respect to the number of stages, and into the power difference between optimized integrated design and optimized pilot design. Posch et al. (2009) showed that a data-dependent sample size increase for all the hypotheses simultaneously in a multi-stage adaptive design has no effect on the asymptotic (as the number of hypotheses goes to infinity) control of the FDR if the hypotheses to be rejected are determined only by the test at the final interim analysis, under all scenarios except the global null hypothesis when all the null hypotheses are true.

Construction of methods with the FWER or FDR control in the setting of a two-stage adaptive design allowing reduction in the number of tested hypotheses at the interim analysis has been discussed, as a sperate issue from sample size adaptations, in Bauer and Kieser (1999) and Kieser, Bauer and Lehmacher (1999) , who presented methods with the FWER control, and in Victor and Hommel (2007) who focused on controlling the FDR in terms of a generalized global p-values. We revisit this issue in the present paper, but focusing primarily on the FDR control in a non-asymptotic setting (with the number of hypothesis not being infinitely large).

Our motivation behind this paper lies in the fact that the theory presented so far (see, for instance, Victor and Hommel, 2007) towards developing an FDR controlling procedure in the setting of a two-stage adaptive design with combinations tests does not seem to be as simple as one would hope for. Moreover, it does not allow setting boundaries on the FDR at the first stage and operate in a manner that would be a natural extension of standard single-stage FDR controlling methods, like the BH (Benjamini and Hochberg, 1995) or methods related to it, from a single-stage to a two-stage design setting. So, we consider the following to be our main problem in this paper:

To construct an FDR controlling procedure for simultaneous testing of the null hypotheses associated with multiple endpoints in the setting of a two-stage adaptive design where the hypotheses are sequentially screened at the first stage as being rejected or accepted based on prescribed thresholds on the FDR, and the null hypotheses that are left out at the first stage are again sequentially tested at the second stage having combined their p-values from the two stages through a combination function.

We propose two FDR controlling procedures, one extending the original single-stage BH procedure, which we call the BH-TSADC Procedure (BH type procedure for two-stage adaptive design with combination tests), and the other extending an adaptive version of the single-stage BH procedure incorporating an estimate of the number of true null hypotheses, which we call the Plug-In BH-TSADC Procedure, from single-stage to a two-stage setting. Let (p_{1i}, p_{2i}) be the pair of first- and second-stage p-values corresponding to the ith null hypothesis. We provide a theoretical proof of the FDR control of the proposed procedures under the assumption that the (p_{1i}, p_{2i})'s are independent and those corresponding to the true null hypotheses are identically distributed as (p_1, p_2) satisfying the p-clud property (Brannath et al., 2002), and some standard assumption on the combination function. We consider two special types of combination function, Fisher's and Simes', which are often used in multiple testing applications, and present explicit formulas for probabilities involving them that would be useful to carry out the proposed procedures at the second stage either using critical values that can be determined before observing the p-values or based on estimated FDR's that can be obtained after observing the p-values.

We carried out extensive simulations to see how well the proposed procedures control the FDR and perform in terms of power compared to the BH method based on the first-stage p-values under independence, and whether or not they can continue to control the FDR under the different types of (positive) dependence among the underlying test statistics we consider, such as equal, clumpy and AR(1) dependence. Our simulation studies indicate that between the two proposed procedures, the BH-TSADC seems to be the better choice in terms of controlling the FDR and power improvement over the single-stage BH procedure when π_0, the proportion of true nulls, is large. If π_0 is not

large, the Plug-In BH-TSADC procedure is better, but it might loose the FDR control when the p-values exhibit equal or AR(1) type dependence with a large equal- or auto-correlation.

We applied the proposed procedures to reanalyze the data on multiple myeloma considered before by Zehetmayer et al. (2008), of course, for a different purpose. The data consist of a set of 12625 gene expression measurements for each of 36 patients with bone lytic lesions and 36 patients in a control group without such lesions. We considered this data in a two-stage framework, with the first 18 subjects per group for Stage 1 and the next 18 per group for Stage 2. With some pre-chosen early rejection and acceptance boundaries, these procedures produce significantly more discoveries than single-stage BH procedure at the same FDR level.

The article is organized as follows. We review some basic results on the FDR control in a single-stage design in Section 2, present our proposed procedures in Section 3, discuss the results of simulations studies in Section 4, and illustrate the real data application in Section 5. We make some concluding remarks in Section 6 regarding the approach we have taken in this paper to construct our procedures, contrasting it with other possible approaches. We give proofs of theorems and propositions in Appendix.

2 CONTROLLING THE FDR IN A SINGLE-STAGE DESIGN

Suppose that there are m endpoints and the corresponding null hypotheses H_i, $i = 1, \ldots, m$, are to be simultaneously tested based on their respective p-values p_i, $i = 1, \ldots, m$, obtained in a single-stage design. The FDR of a multiple testing method that rejects R and falsely rejects V null hypotheses is $E(\text{FDP})$, where $\text{FDP} = V/\max\{R, 1\}$ is the false discovery proportion. Multiple testing is often carried out using a stepwise procedure defined in terms of $p_{(1)} \leq \cdots \leq p_{(m)}$, the ordered p-values. With $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$, a stepup procedure with critical values $\gamma_1 \leq \cdots \leq \gamma_m$ rejects $H_{(i)}$ for all $i \leq k = \max\{j : p_{(j)} \leq \gamma_j\}$, provided the maximum exists; otherwise, it accepts all null hypotheses. A stepdown procedure, on the other hand, with these same critical values rejects $H_{(i)}$ for all $i \leq k = \max\{j : p_{(i)} \leq \gamma_i \text{ for all } i \leq j\}$, provided the maximum exists, otherwise, accepts all null hypotheses. The following are formulas

for the FDR's of a stepup or single-step procedure (when the critical values are same in a stepup procedure) and a stepdown procedure in a single-stage design, which can guide us in developing stepwise procedures controlling the FDR in a two-stage design. We will use the notation FDR_1 for the FDR of a procedure in a single-stage design.

RESULT 1. (Sarkar, 2008). Consider a stepup or stepdown method for testing m null hypotheses based on their p-values p_i, $i = 1, \ldots, m$, and critical values $\gamma_1 \leq \cdots \leq \gamma_m$ in a single-stage design. The FDR of this method is given by

$$\mathrm{FDR}_1 \leq \sum_{i \in J_0} E\left[\frac{I(p_i \leq \gamma_{R_{m-1}^{(-i)}(\gamma_2,\ldots,\gamma_m)+1})}{R_{m-1}^{(-i)}(\gamma_2, \ldots, \gamma_m) + 1}\right], \tag{1}$$

with the equality holding in the case of stepup method, where I is the indicator function, J_0 is the set of indices of the true null hypotheses, and $R_{m-1}^{(-i)}(\gamma_2, \ldots, \gamma_m)$ is the number of rejections in testing the $m-1$ null hypotheses other than H_i based on their p-values and using the same type of stepwise method with the critical values $\gamma_2 \leq \cdots \leq \gamma_m$.

With p_i having the cdf $F(u)$ when H_i is true, the FDR of a stepup or stepdown method with the thresholds γ_i, $i = 1, \ldots, m$, under independence of the p-values, satisfies the following:

$$\mathrm{FDR}_1 \leq \sum_{i \in J_0} E\left(\frac{F(\gamma_{R_{m-1}^{(-i)}(\gamma_2,\ldots,\gamma_m)+1})}{R_{m-1}^{(-i)}(\gamma_2, \ldots, \gamma_m) + 1}\right). \tag{2}$$

When F is the cdf of $U(0,1)$ and these thresholds are chosen as $\gamma_i = i\alpha/m$, $i = 1, \ldots, m$, the FDR equals $\pi_0\alpha$ for the stepup and is less than or equal to $\pi_0\alpha$ for the stepdwon method, where π_0 is the proportion of true nulls, and hence the FDR is controlled at α. This stepup method is the so called BH method (Benjamini and Hochberg, 1995), the most commonly used FDR controlling procedure in a single-stage deign.. The FDR is bounded above by $\pi_0\alpha$ for the BH as well as its stepdown analog under certain type of positive dependence condition among the p-values (Benjamini and Yekutieli, 2001; Sarkar, 2002, 2008).

The idea of improving the FDR control of the BH method by plugging into it a suitable estimate $\hat{\pi}_0$ of π_0, that is, by considering the modified p-values $\hat{\pi}_0 p_i$, rather than the original p-values, in the BH method, was introduced by Benjamini and Hochberg

(2000), which was later brought into the estimation based approach to controlling the FDR by Storey (2002). A number of such plugged-in versions of the BH method with proven and improved FDR control mostly under independence have been put forward based on different methods of estimating π_0 (for instance, Benjamini, Krieger, Yekutieli, 2006; Storey, Taylor and Siegmund, 2004; and Blanchard and Roquain, 2009).

3 CONTROLLING THE FDR IN A TWO-STAGE ADAPTIVE DESIGN

Now suppose that the m null hypotheses H_i, $i = 1, \ldots, m$, are to be simultaneously tested in a two-stage adaptive design setting. When testing a single hypothesis, say H_i, the theory of two-stage combination test can be described as follows: Given p_{1i}, the p-value available for H_i at the first stage, and two constants $\lambda < \lambda'$, make an early decision regarding the hypothesis by rejecting it if $p_{1i} \leq \lambda$, accepting it if $p_{1i} > \lambda'$, and continuing to test it at the second stage if $\lambda < p_{1i} \leq \lambda'$. At the second stage, combine p_{1i} with the additional p-value p_{2i} available for H_i using a combination function $C(p_{1i}, p_{2i})$ and reject H_i if $C(p_{1i}, p_{2i}) \leq \gamma$, for some constant γ. The constants λ, λ' and γ are determined subject to a control of the type I error rate by the test.

For simultaneous testing, we consider a natural extension of this theory from single to multiple testing. More specifically, given the first-stage p-value p_{1i} corresponding to H_i for $i = 1, \ldots, m$, we first determine two thresholds $0 \leq \hat{\lambda} < \hat{\lambda}' \leq 1$, stochastic or non-stochastic, and make an early decision regarding the hypotheses at this stage by rejecting H_i if $p_{1i} \leq \hat{\lambda}$, accepting H_i if $p_{1i} > \hat{\lambda}'$, and continuing to test H_i at the second stage if $\hat{\lambda} < p_{1i} \leq \hat{\lambda}'$. At the second stage, we use the additional p-value p_{2i} available for a follow-up hypothesis H_i and combine it with p_{1i} using the combination function $C(p_{1i}, p_{2i})$. The final decision is taken on the follow-up hypotheses at the second stage by determining another threshold $\hat{\gamma}$, again stochastic or non-stochastic, and by rejecting the follow-up hypothesis H_i if $C(p_{1i}, p_{2i}) \leq \hat{\gamma}$. Both first-stage and second-stage thresholds are to be determined in such a way that the overall FDR is controlled at the desired level α.

Let $p_{1(1)} \leq \cdots \leq p_{1(m)}$ be the ordered versions of the first-stage p-values, with $H_{(i)}$ being the null hypotheses corresponding to $p_{1(i)}$, $i = 1, \ldots, m$, and $q_i = C(p_{1i}, p_{2i})$. We

describe in the following a general multiple testing procedure based on the above theory, before proposing our FDR controlling procedures that will be of this type.

A General Stepwise Procedure.

1. For two non-decreasing sequences of constants $\lambda_1 \leq \cdots \leq \lambda_m$ and $\lambda'_1 \leq \cdots \leq \lambda'_m$, with $\lambda_i < \lambda'_i$ for all $i = 1, \ldots, m$, and the first-stage p-values p_{1i}, $i = 1, \ldots, m$, define two thresholds as follows: $R_1 = \max\{1 \leq i \leq m : p_{1(j)} \leq \lambda_j \text{ for all } j \leq i\}$ and $S_1 = \max\{1 \leq i \leq m : p_{1(i)} \leq \lambda'_i\}$, where $0 \leq R_1 \leq S_1 \leq m$ and R_1 or S_1 equals zero if the corresponding maximum does not exist. Reject $H_{(i)}$ for all $i \leq R_1$, accept $H_{(i)}$ for all $i > S_1$, and continue testing $H_{(i)}$ at the second stage for all i such that $R_1 < i \leq S_1$.

2. At the second stage, consider $q_{(i)}$, $i = 1, \ldots, S_1 - R_1$, the ordered versions of the combined p-values $q_i = C(p_{1i}, p_{2i})$, $i = 1, \ldots, S_1 - R_1$, for the follow-up null hypotheses, and find $R_2(R_1, S_1) = \max\{1 \leq i \leq S_1 - R_1 : q_{(i)} \leq \gamma_{R_1+i}\}$, given another non-decreasing sequence of constants $\gamma_{r_1+1}(r_1, s_1) \leq \cdots \leq \gamma_{s_1}(r_1, s_1)$, for every fixed $r_1 < s_1$. Reject the follow-up null hypothesis $H_{(i)}$ corresponding to $q_{(i)}$ for all $i \leq R_2$ if this maximum exists, otherwise, reject none of the follow-up null hypotheses.

Remark 1. We should point out that the above two-stage procedure screens out the null hypotheses at the first stage by accepting those with relatively large p-values through a stepup procedure and by rejecting those with relatively small p-values through a stepdown procedure. At the second stage, it applies a stepup procedure to the combined p-values. Conceptually, one could have used any type of multiple testing procedure to screen out the null hypotheses at the fisrt stage and to test the follow-up null hypotheses at the second stage. However, the particular types of stepwise procedure we have chosen at the two stages provide flexibility in terms of developing a formula for the FDR and eventually determining explicitly the thresholds we need to control the FDR at the desired level.

Let V_1 and V_2 denote the total numbers of falsely rejected among all the R_1 null hypotheses rejected at the first stage and the R_2 follow-up null hypotheses rejected at

the second stage, respectively, in the above procedure. Then, the overall FDR in this two-stage procedure is given by

$$\mathrm{FDR}_2 = E\left[\frac{V_1 + V_2}{\max\{R_1 + R_2, 1\}}\right].$$

The following theorem (to be proved in Appendix) will guide us in determining the first- and second-stage thresholds in the above procedure providing a control of FDR_2 at the desired level. This is the procedure that will be one of those we propose in this article.

THEOREM 1. The FDR of the above general multiple testing procedure satisfies the following inequality

$$\begin{aligned}\mathrm{FDR}_2 \;\le\; & \sum_{i\in J_0} E\left[\frac{I(p_{1i} \le \lambda_{R_1^{(-i)}+1})}{R_1^{(-i)}+1}\right] + \\ & \sum_{i\in J_0} E\left[\frac{I(\lambda_{\tilde{R}_1^{(-i)}+1} < p_{1i} \le \lambda'_{S_1^{(-i)}+1},\ q_i \le \gamma_{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1, S_1^{-i}+1})}{\tilde{R}_1^{(-i)} + R_2^{(-i)} + 1}\right],\end{aligned}$$

where $R_1^{(-i)}$ is defined as R_1 in terms of the $m-1$ first-stage p-values $\{p_{11}, \ldots, p_{1m}\}\backslash\{p_{1i}\}$ and the sequence of constants $\lambda_2 \le \cdots \le \lambda_m$, $\tilde{R}_1^{(-i)}$ and $S_1^{(-i)}$ are defined as R_1 and S_1, respectively, in terms of $\{p_{11}, \ldots, p_{1m}\} \setminus \{p_{1i}\}$ and the two sequences of constants $\lambda_1 \le \cdots \le \lambda_{m-1}$ and $\lambda'_2 \le \cdots \le \lambda'_m$, and $R_2^{(-i)}$ is defined as R_2 with R_1 replaced by $\tilde{R}_1^{(-i)}$ and S_1 replaced by $S_1^{(-i)} + 1$ and noting the number rejected follow-up null hypotheses based on all the combined p-values except the q_i and the critical values other than the first one; that is,

$$\begin{aligned} R_2^{(-i)} &\equiv R_2^{(-i)}(\tilde{R}_1^{(-i)}, S_1^{(-i)} + 1) \\ &= \max\{1 \le j(\ne i) \le S_1^{(-i)} - \tilde{R}_1^{(-i)} + 1 : q_{(j)}^{(-i)} \le \gamma_{\tilde{R}_1^{(-i)}+j+1}(\tilde{R}_1^{(-i)}, S_1^{(-i)} + 1)\},\end{aligned}$$

where $q_{(j)}^{(-i)}$'s are the ordered versions of the combined p-values for the follow-up null null hypotheses except the q_i.

3.1 BH Type Procedures

We are now ready to propose our FDR controlling multiple testing procedures in a two-stage adaptive design setting with combination function. Before that, let us state some assumptions we need.

ASSUMPTION 1. The combination function $C(p_1, p_2)$ is non-decreasing in both arguments.

ASSUMPTION 2. The pairs (p_{1i}, p_{2i}), $i = 1, \ldots, m$, are independently distributed and the pairs corresponding the null hypotheses are identically distributed as (p_1, p_2) with a joint distribution that satisfies the 'p-clud' property (Brannath et al., 2002), that is,

$$\Pr(p_1 \leq u) \leq u \text{ and } \Pr(p_2 \leq u \mid p_1) \leq u \text{ for all } 0 \leq u \leq 1. \tag{3}$$

Let us define

$$H(c; t, t') = \int_t^{t'} \int_0^1 I(C(u_1, u_2) \leq c) du_2 du_1.$$

Definition 1. (BH-TSADC Procedure: The BH type procedure for two-stage adaptive design with combination tests).

1. Given the level α at which the FDR is to be controlled, three sequences of constants $\lambda_i = i\lambda/m$, $i = 1, \ldots, m$, $\lambda'_i = i\lambda'/m$, $i = 1, \ldots, m$, for some prefixed $\lambda < \alpha < \lambda'$, and $\gamma_{r_1+1,s_1} \leq \cdots \leq \gamma_{s_1,s_1}$, satisfying

$$H(\gamma_{r_1+i,s_1}; \lambda_{r_1}, \lambda'_{s_1}) = \frac{(r_1 + i)(\alpha - \lambda)}{m}, \tag{4}$$

   $i = 1, \ldots, s_1 - r_1$, for every fixed $1 \leq r_1 < s_1 \leq m$, find $R_1 = \max\{1 \leq i \leq m : p_{1(j)} \leq \lambda_j \text{ for all } j \leq i\}$ and $S_1 = \max\{1 \leq i \leq m : p_{1(i)} \leq \lambda'_i\}$, with R_1 or S_1 being equal to zero if the corresponding maximum does not exist.

2. Reject $H_{(i)}$ for $i \leq R_1$; accept $H_{(i)}$ for $i > S_1$; and continue testing $H_{(i)}$ for $R_1 < i \leq S_1$ making use of the additional p-values p_{2i}'s available for all such follow-up hypotheses at the second stage.

3. At the second stage, consider the combined p-values $q_i = C(p_{1i}, p_{2i})$ for the follow-up null hypotheses. Let $q_{(i)}$, $i = 1, \ldots S_1 - R_1$, be their ordered versions. Reject

$H_{(i)}$ [the null hypothesis corresponding to $q_{(i)}$] for all $i \le R_2(R_1, S_1) = \max\{1 \le j \le S_1 - R_1 : q_{(j)} \le \gamma_{R_1+j,S_1}\}$, provided this maximum exists, otherwise, reject none of the follow-up null hypotheses.

PROPOSITION 1. Let π_0 be the proportion of true null hypotheses. Then, the FDR of the BH-TSADC method is less than or equal to $\pi_0\alpha$, and hence controlled at α, if Assumptions 1 and 2 hold.

The proposition is proved in Appendix.

The BH-TSADC procedure can be implemented alternatively, and often more conveniently, in terms of some FDR estimates at both stages. With $R^{(1)}(t) = \#\{i : p_{1i} \le t)$ and $R^{(2)}(c; t, t') = \#\{i : t < p_{1i} \le t', C(p_{1i}, p_{2i}) \le c\}$, let us define

$$\begin{aligned} \widehat{\mathrm{FDR}}_1(t) &= \begin{cases} \frac{mt}{R^{(1)}(t)} & \text{if } R^{(1)}(t) > 0 \\ 0 & \text{if } R^{(1)}(t) = 0, \end{cases} \\ \text{and } \widehat{\mathrm{FDR}}_2(c; t, t') &= \begin{cases} \frac{mH(c;t,t')}{R^{(1)}(t)+R^{(2)}(c;t,t')} & \text{if } R^{(2)}(c; t, t') > 0 \\ 0 & \text{if } R^{(2)}(c; t, t') = 0, \end{cases} \end{aligned} \tag{5}$$

Then, we have the following:

The BH-TSADC procedure: An alternative description. Reject $H_{(i)}$ for all $i \le R_1 = \max\{1 \le k \le m : \widehat{\mathrm{FDR}}_1(p_{1(j)}) \le \lambda \text{ for all } j \le k\}$; accept $H_{(i)}$ for all $i > S_1 = \max\{1 \le k \le m : \widehat{\mathrm{FDR}}_1(p_{1(k)}) \le \lambda'\}$; continue to test $H_{(i)}$ at the second stage for all i such that $R_1 < i \le S_1$. Reject $H_{(i)}$, the follow-up null hypothesis corresponding to $q_{(i)}$, at the second stage for all $i \le R_2(R_1, S_1) = \max\{1 \le k \le S_1 - R_1 : \widehat{\mathrm{FDR}}_2(q_{(k)}; R_1\lambda/m, S_1\lambda'/m) \le \alpha - \lambda\}$.

REMARK 1. The BH-TSADC procedure is an extension of the BH procedure, from a method of controlling the FDR in a single-stage design to that in a two-stage adaptive design with combination tests. When $\lambda = 0$ and $\lambda' = 1$, that is, when we have a single-stage design based on the combined p-values, this method reduces to the usual BH method. Notice that $\widehat{\mathrm{FDR}}_1(t)$ is a conservative estimate of the FDR of the single-step test with the rejection $p_i \le t$ for each H_i. So, the BH-TSADC procedure screens out those null hypotheses as being rejected (or accepted) at the first stage the estimated

FDR's at whose p-values are all less than or equal to λ (or greater than λ').

Clearly, the BH-TSADC procedure can potentially be improved in terms of having a tighter control over its FDR at α by plugging a suitable estimate of π_0 into it while choosing the second-stage thresholds, similar to what is done for the BH method in a single-stage design. As said in Section 2, there are different ways of estimating π_0, each of which has been shown to provide the ultimate control of the FDR, of course when the p-values are independent, by the resulting plugged-in version of the single-stage BH method (see, e.g., Sarkar, 2008). However, we will consider the following estimate of π_0, which is of the type considered in Storey, Taylor and Siegmund (2004) and seems natural in the context of the present adaptive design setting where $m - S_1$ of the null hypotheses are accepted as being true at the first stage:

$$\hat{\pi}_0 = \frac{m - S_1 + 1}{m(1 - \lambda')}. \tag{6}$$

The following theorem gives a modified version of the the BH-TSADC procedure using this estimate.

Definition 2. (Plug-In BH-TSADC Procedure: A plug-in version of the BH-TSADC procedure).

Consider the BH-TSADC procedure with the sequences of contacts $\lambda_i = i\lambda/m$, $i = 1, \ldots, m$, and $\lambda'_i = i\lambda'/m$, $i = 1, \ldots, m$, given $0 \leq \lambda < \lambda' \leq 1$, providing the early decision thresholds R_1 and S_1, and the second-stage critical values $\gamma^*_{R_1+i}$, $i = 1, \ldots, S_1 - R_1$, satisfying

$$H(\gamma^*_{R_1+i,S_1}; \lambda_{R_1}, \lambda'_{S_1}) = \frac{(R_1 + i)(\alpha - \lambda)}{m\hat{\pi}_0}. \tag{7}$$

for $i = 1, ..., S_1 - R_1$.

Proposition 2. The FDR of the Plug-In BH-TSADC method is less than or equal to α if Assumptions 1 and 2 hold.

A proof of this proposition is given in Appendix.

As in the BH-TSADC procedure, the Plug-In BH-TSADC procedure can also be

described alternatively using estimated FDR's at both stages. Let

$$\widehat{\mathrm{FDR}}_2^*(c;t,t') = \begin{cases} \frac{m\hat{\pi}_0 H(c;t,t')}{R^{(1)}(t)+R^{(2)}(c;t,t')} & \text{if } R^{(2)}(c;t,t') > 0 \\ 0 & \text{if } R^{(2)}(c;t,t') = 0, \end{cases} \tag{8}$$

Then, we have the following:

The Plug-In BH-TSADC procedure: An alternative description. At the first stage, decide the null hypotheses to be rejected, accepted, or continued to be tested at the second stage based on $\widehat{\mathrm{FDR}}_1$, as in (the alternative description of) the BH-TSADC procedure. At the second stage, reject $H_{(i)}$, the follow-up null hypothesis corresponding to $q_{(i)}$, for all $i \le R_2^*(R_1,S_1) = \max\{1 \le k \le S_1 - R_1 : \widehat{\mathrm{FDR}}_2^*(q_{(k)}; R_1\lambda/m, S_1\lambda'/m) \le \alpha - \lambda\}$.

3.2 Two Special Combination Functions

We now present explicit formulas of $H(c;t,t')$ for two special combination functions - Fisher's and Simes' - often used in multiple testing applications.

Fisher's combination function: $C(p_1,p_2) = p_1p_2$.

$$\begin{aligned} H_{Fisher}(c;t,t') &= \int_t^{t'}\int_0^1 I(C(u_1,u_2) \le c)\, du_2 du_1 \\ &= \begin{cases} c\ln\left(\frac{t'}{t}\right) & \text{if} \quad c < t \\ c - t + c\ln\left(\frac{t'}{c}\right) & \text{if} \quad t \le c < t' \\ t' - t & \text{if} \quad c \ge t' \ , \end{cases} \end{aligned} \tag{9}$$

for $c \in (0,1)$.

Simes' combination function: $C(p_1, p_2) = \min\{2\min(p_1, p_2), \max(p_1, p_2)\}$.

$$
\begin{aligned}
H_{Simes}(c; t, t') &= \int_t^{t'} \int_0^1 I(C(u_1, u_2) \le c) du_2 du_1 \\
&= \begin{cases}
\frac{c}{2}(t' - t) & \text{if} \quad c \le t \\
c(\frac{t'}{2} - t) + \frac{c^2}{2} & \text{if} \quad t < c \le \min(2t, t') \\
c(t' - t) & \text{if} \quad t' < c \le 2t \\
\frac{c}{2}(1 + t') - t & \text{if} \quad 2t < c \le t' \\
\frac{c}{2}(1 + 2t') - \frac{c^2}{2} - t & \text{if} \quad \max(2\mathrm{t}, \mathrm{t}') \le \mathrm{c} \le 2\mathrm{t}' \\
t' - t & \text{if} \quad c \ge 2t',
\end{cases}
\end{aligned} \tag{10}
$$

for $c \in (0, 1)$.

These formulas are given in Chen, Sarkar and Bretz (2010) and can be used to determine the critical values γ_i's before observing the combined p-values or to estimate the FDR after observing the combined p-values at the second stage in the BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions. Of course, for large values of m, it is numerically more challenging to determine the γ_i's than estimating the FDR at the second stage, and so in that case we would recommend using the alternative versions of these procedures.

To illustrate what or how to determine the critical values at the two stages in the BH-TSADC procedure for relatively small values of m, and also to see how the second stage critical values compare between Fisher's and Simes combination functions, we consider testing $m = 5$ null hypotheses in a two-stage adaptive design setting using these combination functions with $\lambda = 0.025$, $\lambda' = 0.5$, and $\alpha = 0.05$. The first stage critical values in these procedures for the stepdown test are $\lambda_{r_1} = 0.005 r_1$, $r_1 = 0, 1, \ldots, 5$, and for the stepup test are $\lambda'_{s_1} = 0.1 s_1$, $s_1 = 1, \ldots, 5$. The values of γ_{r_1+i, s_1}, $i = 1, \ldots, s_1 - r_1$, satisfying the equation $H_{Fisher}(\gamma_{r_1+i,s_1}; 0.005 r_1, 0.1 s_1) = (r_1 + i) 0.025/m$ for Fisher's combination function and the equation $H_{Simes}(\gamma_{r_1+i,s_1}; 0.005 r_1, 0.1 s_1) = (r_1 + i) 0.025/m$ for Simes combination function, for different pairs (r_1, s_1), where $r_1 < s_1 = 0, 1, \ldots, 5$, are presented in Table 1.

It is to be noted that the combined p-value based on Fisher's combination function is stochastically smaller than that based on Simes' combination function, and so in Table 1 the second stage critical values corresponding to Fisher's combination function are

always seen to be smaller than those corresponding to Simes' combination function. Of course, it does not necessarily mean that Simes' combination function is always a better choice in our proposed procedures.

4 SIMULATION STUDIES

This section presents the results of simulation studies we conducted to investigate the following two questions related to the proposed procedures:

Q1. How well do the proposed BH-TSADC and Plug-In BH-TSADC procedures perform under independence compared to the single-stage BH procedure in terms of FDR control and power?

Q2. Can the proposed BH-TSADC and Plug-In BH-TSADC procedures continue to control the FDR for dependent p-values?

To investigate Q_1, (i) we generated two independent sets of m uncorrelated random variables $Z_i \sim N(\mu_i, 1)$, $i = 1, \ldots, m$, one for Stage 1 and the other for Stage 2, having set $m\pi_0$ of these μ_i's at zero and the rest at 2, (ii) tested $H_i : \mu_i = 0$ against $K_i : \mu_i > 0$, simultaneously for $i = 1, \ldots, m$, by applying the (alternative versions of) BH-TSADC and Plug-In BH-TSADC procedures at level α with both Fisher's and Simes combination functions, $\lambda = 0.025$, and $\lambda' = 0.5$ to the generated data for both stages and the level α BH procedure to the data for the first stage, and (iii) noted the false discovery proportion and the proportion of false nulls that are rejected. We repeated steps (i)-(iii) 1000 times and averaged out the above proportions over these 1000 runs to obtain the final simulated values of FDR and average power (the expected proportion of false nulls that are rejected) for each of these procedures.

The simulated FDR's and average powers for these three procedures have been graphically displayed in Figures 1 and 2. Figure 1 compares the proposed BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes combination functions with those of the BH procedure for different values of π_0, $\alpha = 0.05$, and $m = 10, 100$, and 1000, in terms of the simulated FDR, while Figure 2 does the same in terms of the simulated average power.

In our simulation study to investigate Q_2, we considered three different scenarios for dependent p-values. In particular, we generated two independent sets of $m = 100$ correlated normal random variables $Z_i \sim N(\mu_i, 1)$, $i = 1, \ldots, m$, one for Stage 1 and the other for Stage 2, with $m\pi_0$ of the μ_i's being equal to 0 and the rest being equal to 2, and a correlation matrix exhibiting one of three different types of dependence - equal, clumpy and autoregressive of order one [AR(1)] dependence. In other words, the Z_i's were assumed to have a common, non-negative correlation ρ in case of equal dependence, were broken up into ten independent groups with 10 of the Z_i's within each group having a common, non-negative correlation ρ in case of clumpy dependence, and were assumed to have correlations $\rho_{ij} = \text{Cor}(Z_i, Z_j)$ of the form $\rho_{ij} = \rho^{|i-j|}$ for all $i \neq j = 1, \ldots, m$, and some non-negative ρ in case of AR(1) dependence. We then applied the (alternative versions of) BH-TSADC and Plug-In BH-TSADC procedures at level $\alpha = 0.05$ with both Fisher's and Simes combination functions, $\lambda = 0.025$, and $\lambda' = 0.5$ to these data sets. These two steps were repeated 1000 times before obtaining the simulated FDR's and average powers for these procedures, as in our study related to Q_1.

Figures 3-5 graphically display the simulated FDR's of these procedures for different values of π_0 and types of dependent p-values considered.

As seen from all these figures, the proposed procedures with Fisher's combination function seem to have slight edge over the corresponding ones with Simes' combination function in terms of FDR control and power. Between these two procedures, whether it's based on Fisher's or Simes' combination function, the BH-TSADC appears to be the better choice when π_0 is large, which is often the case in practice. It controls the FDR not only under independence, which is theoretically known, but also the FDR control seems to be maintained even under different types of positive dependence. Also, it provides a better power improvement over the single-stage BH procedure. If, however, π_0 is not large, the Plug-In BH-TSADC procedure provides a better control of the FDR and its power improvement over the single-stage BH procedure seems more significant than the BH-TSDADC procedure; of course, it may loose the FDR control when the p-values exhibit equal or AR(1) type dependence with a moderately large equal- or auto-correlation.

5 A REAL DATA APPLICATION

To illustrate how the proposed procedures can be implemented in practice, we reanalyzed a dataset taken from an experiment by Tian et al. (2003) and post-processed by Jeffery et al. (2006). Zehetmayer et al. (2008) considered this data for a different purpose. In this data set, multiple myeloma samples were generated with Affymetrix Human U95A chips, each consisting 12625 probe sets. The samples were split into two groups based on the presence or absence of focal lesions of bone.

The original dataset contains gene expression measurements of 36 patients without and 137 patients with bone lytic lesions, However, in our reanalysis, we used the gene expression measurements of 36 patients with bone lytic lesions and a control group of the same sample size without such lesions. We considered this data in a two-stage framework, with the first 18 subjects per group for Stage 1 and the next 18 subjects per group for Stage 2. We prefixed the Stage 1 early rejection boundary λ at 0.025 and the early acceptance boundary λ' at 0.5, and applied the proposed (alternatives versions of) BH-TSADC and plug-in BH-TSADC procedures at the overall FDR level 0.05.

In particular, we considered all $m = 12625$ probe set gene expression measurements at Stage 1 and analyzed them based on a stepdown procedure with the critical values $\lambda_i = i0.025/m$, $i = 1, \ldots, m$, and a steup procedure with the critical values $\lambda'_i = i0.5/m$, $i = 1, \ldots, m$, using the corresponding p-values generated from one-sided t-tests. We noted the probe sets that were rejected by the stepdown procedure and those that were accepted by the stepup procedure. With these numbers being r_1 and $m - s_1$, respectively, we took the probe sets that were neither rejected by the stepdown procedure nor accepted by the stepup procedure, that is, the probe sets with the first-stage p-values more than $r_1\lambda/m$ but less than or equal to $s_1\lambda'/m$, for further analysis using estimated FDR based on their first-stage and second-stage p-values combined through Fisher's and Simes' combination functions, as described in the alternative versions of the BH-TSADC and plug-in BH-TSADC procedures.

The results of this analysis are reported in Table 2. As seen from this table, the BH-TSADC procedure with Fisher's combination function and its plug-in version produce 144 and 93 discoveries, respectively; whereas, these numbers are 40 and 32, respectively, for the Simes' combination function. These numbers are significantly much larger than

18, the number of discoveries made by the single-stage BH procedure.

6 CONCLUDING REMARKS

Our main goal in this article has been to construct a two-stage multiple testing procedure that allows making early decisions on the null hypotheses, in terms of rejection, acceptance or continuation to the second stage for further testing with more observations, and eventually controls the FDR. Such two-stage formulation of multiple testing is of practical importance in many statistical investigations; nevertheless, generalizations of the classical BH type methods from single-stage to the present two-stage setting, which seem to be the most natural procedures to consider, have not been put forward until the present work. We have been able to construct two such generalizations and provided proofs of their FDR control and simulation and practical examples of their improved power performances compared to the corresponding single-stage BH type methods under independence. We also have presented numerical evidence that they can maintain a control over the FDR even under some dependence situations.

It is important to emphasize that the theory behind the developments of our proposed two-stage FDR controlling methods has been driven by the idea of setting the early decision boundaries $\lambda < \lambda'$ on the FDR of the first-stage p-values, rather than on these p-values themselves. In other words, we reject (or accept) those null hypotheses at the first stage at whose p-values the estimated FDR's are all less than or equal to λ (or greater than λ'); see Remark 1. This, we would argue, is often practical and meaningful when we are testing multiple hypotheses in two-stages with a view to controlling the overall FDR.

Brannath et al. (2002) have defined a global p-value $\tilde{p}(p_1, p_2)$ for testing a single hypothesis in a two-stage adaptive design with combination function $C(p_1, p_2)$. With the boundaries $\lambda < \lambda'$ set on each p_{1i}, the global p-value for each H_i is defined by

$$\tilde{p}_i \equiv \tilde{p}_i(p_{1i}, p_{2i}) = \begin{cases} p_{1i} & \text{if} \quad p_{1i} \le \lambda \text{ or } p_{1i} > \lambda' \\ \lambda + H(C(p_{1i}, p_{2i}); \lambda, \lambda') & \text{if} \quad \lambda < p_{1i} \le \lambda' \ . \end{cases} \tag{11}$$

They have shown that each $\tilde{p}_i$ is stochastically larger than or equal to $U(0,1)$ when (p_{1i}, p_{2i}) satisfies the p-clud property, and the equality holds when p_{1i} and p_{2i} are

independently distributed as $U(0,1)$. So, one may consider the BH method based on the $\tilde{p}_i$'s. This would control the overall FDR under the assumptions considered in the paper, maybe under some positive dependence conditions as well. However, it does not set the early decision boundaries on the FDR.

One may consider taking an estimation based approach to the present problem as follows. An estimated global FDR at $(p_1, p_2) = (t, t')$ when testing multiple hypotheses in a two-stage adaptive design with combination function $C(p_1, p_2)$ can be defined in the spirit of the above global p-value by

$$\widehat{\mathrm{FDR}}_{12}(c,t,t') = \begin{cases} \frac{m\hat{\pi}_0 c}{R^{(1)}(c)\vee 1} & \text{if} \quad c \le t \text{ or } c > t' \\ \frac{m\hat{\pi}_0[t+H(c;t,t')]}{[R^{(1)}(t)+R^{(2)}(c;t,t')]\vee 1} & \text{if} \quad t < c \le t' \ , \end{cases} \tag{12}$$

for some estimate $\hat{\pi}_0$ of π_0. Let

$$\begin{aligned} \hat{t}_\lambda &= \sup\{c \le t : \ \widehat{\mathrm{FDR}}_{12}(c',t,t') \le \lambda \text{ for all } c' \le c\} \\ \hat{t}_{\lambda'} &= \inf\{c > t' : \ \widehat{\mathrm{FDR}}_{12}(c',t,t') > \lambda' \text{ for all } c' > c\}, \\ \text{and} \qquad \hat{c}_\alpha(\lambda,\lambda') &= \sup\{\hat{t}_\lambda < c \le \hat{t}_{\lambda'} : \ \widehat{\mathrm{FDR}}_{12}(c,\hat{t}_\lambda,\hat{t}_{\lambda'}) \le \alpha\}. \end{aligned} \tag{13}$$

Then, reject H_i if $p_{1i} \le \hat{t}_\lambda$ or if $C(p_{1i}, p_{2i}) \le \hat{c}_\alpha(\lambda,\lambda')$ and $\hat{t}_\lambda < p_{1i} \le \hat{t}_{\lambda'}$. This may control the overall FDR asymptotically under the weak dependence condition and the consistency property of $\hat{\pi}_0$ (as in Storey et al., 2004).

7 Appendix

Proof of Theorem 1.

$$\begin{aligned} \mathrm{FDR}_2 &= E\left[\frac{V_1+V_2}{\max\{R_1+R_2,1\}}\right] \\ &\le E\left[\frac{V_1}{\max\{R_1,1\}}\right] + E\left[\frac{V_2}{\max\{R_1+R_2,1\}}\right]. \end{aligned}$$

Now,

$$E\left[\frac{V_1}{\max\{R_1,1\}}\right] = \sum_{i\in J_0} E\left[\frac{I(p_{1i}\le\lambda_{R_1})}{\max\{R_1,1\}}\right]$$
$$= \sum_{i\in J_0} E\left[\frac{I(p_{1i}\le\lambda_{R_1})}{\max\{R_1,1\}}\right] \le \sum_{i\in J_0} E\left[\frac{I(p_{1i}\le\lambda_{R_1^{(-i)}+1})}{R_1^{(-i)}+1}\right];$$

(as shown in Sarkar, 2008; see also Result 1). And,

$$E\left[\frac{V_2}{\max\{R_1+R_2,1\}}\right]$$
$$= \sum_{i\in J_0} E\left[\frac{I(\lambda_{R_1+1}<p_{1i}\le\lambda'_{S_1},\ q_i\le\gamma_{R_1+R_2,S_1}, S_1>0, S_1>R_1, R_2>0)}{R_1+R_2}\right]. \quad (14)$$

Writing R_2 more explicitly in terms of R_1 and S_1, we see that the expression in (14) is equal to

$$\sum_{i\in J_0}\sum_{s_1=1}^{m}\sum_{r_1=0}^{s_1-1}\sum_{r_2=1}^{s_1-r_1}$$
$$E\left[\frac{I(\lambda_{r_1+1}<p_{1i}\le\lambda'_{s_1},\ q_i\le\gamma_{r_1+r_2}, R_1=r_1, S_1=s_1, R_2(r_1,s_1)=r_2)}{r_1+r_2}\right]$$
$$= \sum_{i\in J_0}\sum_{s_1=1}^{m}\sum_{r_1=0}^{s_1-1}\sum_{r_2=1}^{s_1-r_1}$$
$$E\left[\frac{I(\lambda_{r_1+1}<p_{1i}\le\lambda'_{s_1},\ q_i\le\gamma_{r_1+r_2}, \tilde{R}_1^{(-i)}=r_1, S_1^{(-i)}=s_1-1, R_2^{(-i)}(r_1,s_1)=r_2-1)}{r_1+r_2}\right]$$
$$= \sum_{i\in J_0}\sum_{s_1=0}^{m-1}\sum_{r_1=0}^{s_1}\sum_{r_2=0}^{s_1-r_1}$$
$$E\left[\frac{I(\lambda_{r_1+1}<p_{1i}\le\lambda'_{s_1+1},\ q_i\le\gamma_{r_1+r_2+1,s_1+1}, \tilde{R}_1^{(-i)}=r_1, S_1^{(-i)}=s_1, R_2^{(-i)}(r_1,s_1+1)=r_2)}{r_1+r_2+1}\right]$$
$$= \sum_{i\in J_0} E\left[\frac{I(\lambda_{\tilde{R}_1^{(-i)}+1}<p_{1i}\le\lambda'_{S_1^{(-i)}+1},\ q_i\le\gamma_{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1,S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1}\right]. \quad (15)$$

Thus, the theorem is proved.

Proof of proposition 1. As seen from Theorem 1 and the assumptions made in

the proposition,

$$\begin{aligned}
\mathrm{FDR}_2 &\leq \sum_{i\in J_0} E\left[\frac{Pr_H(p_1 \leq \lambda_{R_1^{(-i)}+1})}{R_1^{(-i)}+1}\right] + \\
&\quad \sum_{i\in J_0} E\left[\frac{Pr_H(\lambda_{\tilde{R}_1^{(-i)}+1} < p_1 \leq \lambda'_{S_1^{(-i)}+1},\ C(p_1,p_2) \leq \gamma_{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1,S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1}\right] \\
&\leq \sum_{i\in J_0} E\left[\frac{\lambda_{R_1^{(-i)}+1})}{R_1^{(-i)}+1}\right] + \\
&\quad \sum_{i\in J_0} E\left[\frac{Pr(\lambda_{\tilde{R}_1^{(-i)}+1} < u_1 \leq \lambda'_{S_1^{(-i)}+1},\ C(u_1,u_2) \leq \gamma_{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1,S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1}\right].
\end{aligned} \tag{16}$$

Now, note that the first sum in (16) is less than or equal to $\pi_0\lambda$, since $\lambda_{R_1^{(-i)}+1} = [R_1^{(-i)}+1]\lambda/m$, and the second summation is less than or equal to $\pi_0(\alpha-\lambda)$, since the probability in the numerator in this summation is equal to

$$\begin{aligned}
&H(\gamma_{\tilde{R_1}^{(-i)}+R_2^{(-i)}+1,S_1^{(-i)}+1}; \lambda_{\tilde{R_1}^{(-i)}+1}, \lambda'_{S_1^{(-i)}+1}) \\
= &\quad \frac{\left[\tilde{R}_1^{(-i)}+1+R_2^{(-i)}\right](\alpha-\lambda)}{m}.
\end{aligned} \tag{17}$$

Thus, the theorem is proved.

Proof of Proposition 2. This can be proved as in Proposition 1. More specifically, first note that the FDR here, which we call the FDR_2^*, satisfies the following:

$$\begin{aligned}
FDR_2^* &\leq \sum_{i\in J_0} E\left[\frac{I(p_{1i} \leq \lambda_{R_1^{(-i)}+1})}{R_1^{(-i)}+1}\right] + \\
&\quad \sum_{i\in J_0} E\left[\frac{I(\lambda_{\tilde{R}_1^{(-i)}+1} \leq p_{1i} \leq \lambda'_{S_1^{(-i)}+1}, q_i \leq \gamma^*_{\tilde{R}_1^{(-i)}+R_2^{*(-i)}+1,S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)}+R_2^{*(-i)}+1}\right],
\end{aligned} \tag{18}$$

where

$$\begin{aligned}
R_2^{*(-i)} &\equiv R_2^{*(-i)}(\tilde{R}_1^{(-i)}, S_1^{(-i)}+1) \\
&= \max\{1 \leq j(\neq i) \leq S_1^{(-i)} - \tilde{R}_1^{(-i)} + 1 : q_{(j)}^{(-i)} \leq \gamma^*_{\tilde{R}_1^{(-i)}+j+1,S_1^{(-i)}+1}\},
\end{aligned}$$

with $q_{(j)}^{(-i)}$ being the ordered versions of the combined p-values except the q_i. As in Proposition 1, the first sum is less than or equal to $\pi_0\lambda$, while the second sum is less than or equal to

$$\begin{aligned} & \sum_{i\in J_0} E\left[\frac{H(\gamma_{\tilde{R}_1^{(-i)}+R_2^{*(-i)}+1,S_1^{(-i)}+1};\lambda_{\tilde{R}_1^{(-i)}+1},\lambda'_{S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)}+R_2^{*(-i)}+1}\right] \\ = \; & (\alpha-\lambda)\sum_{i\in J_0} E\left[\frac{1-\lambda'}{m-S_1^{(-i)}}\right] \leq \pi_0(\alpha-\lambda), \end{aligned}$$

since $E[\frac{1-\lambda'}{m-S_1^{(-i)}}] \leq \frac{1-\lambda'm\pi_0}{m}$; see, for instance, Sarkar (2008).

References

[1] P. Bauer and M. Kieser. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine*, 18:1833–1848, 1999.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

[3] Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25:6083, 2000.

[4] Y. Benjamini, A.M. Krieger, and D. Yekutieli. Adaptive linear step-up false discovery rate controlling procedures. *Biometrika*, 93(3):491–507, 2006.

[5] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.

[6] G. Blanchard and E. Roquain. Adaptive fdr control under independence and dependence. *Journal of Machine Learning Research*, 10:2837–2871, 2009.

[7] W. Brannath, M. Posch, and P. Bauer. Recursive combination tests. *Journal of the American Statistical Association*, 97(457):236–244, 2002.

[8] J. Chen, S.K. Sarkar, and F. Bretz. Finding critical values with prefixed early stopping boundaries and controlled type I error for two-stage combination test. 2010. submitted for publication.

[9] I.B. Jeffery, D.G. Higgins, and A.C. Culhance. Comparison and evaluation of methods for generating differentiall expressed genes lists from microarray data. *BMC Bioinformatics*, 7:359–375, 2006.

[10] M. Kieser, P. Bauer, and W. Lehmacher. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal*, 41:261–277, 1999.

[11] R.A. Miller, A. Galecki, and R.J. Shmookler-Reis. Interpretation, design, and analysis of gene array expression experiments. *J Gerontol A-Biol*, 56:B52–B57, 2001.

[12] M. Posch, S. Zehetmayer, and P. Bauer. Hunting for significance with the false discovery rate. *Journal of the American Statistical Association*, 104(486):832–840, 2009.

[13] S.K. Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, 0:239–257, 2002.

[14] S.K. Sarkar. On methods controlling the false discovery rate. *Sankhya: The Indian Journal of Statistics*, 70-A(part 2):135–168, 2008.

[15] J.M. Satagopan and R.C. Elston. Optimal two-stage genotyping in population-based association studies. *Genetic Epidemiology*, 25:149–157, 2003.

[16] J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*, 64:479–498, 2002.

[17] J.D. Storey, J. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B*, 66:187–205, 2004.

[18] J.D. Storey and R. Tibshirani. Statistical significance in genomewide studies. *Proceedings of the National Academy of Science USA*, 100:9440–9445, 2003.

[19] E. Tian, F. Zhan, R. Walker, E. Rasmussen, Y. Ma, B. Barlogie, and J.D. Shaughnessy. The role of the wnt-signaling antagonist dkki in the development of osteolytic lesions n multple myeloma. *New England journal of Medicine*, 349:2438–2494, 2003.

[20] A. Victor and G. Hommel. Combining adaptive design with control of the false discovery rate - a generalized definition for a global p-value. *Biometrical Journal*, 49:94–106, 2007.

[21] J.I. Weller, J.Z. Song, D.W. Heyen, H.A. Lewin, and M. Ron. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics*, 150:1699–1706, 1998.

[22] S. Zehetmayer, P. Bauer, and M. Posch. Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics*, 21:3771–3777, 2005.

[23] S. Zehetmayer, P. Bauer, and M. Posch. Optimized multi-stage designs controlling the false discovery or the family-wise error rate. *Statistics in Medicine*, 27:4145–4160, 2008.

Table 1: The second stage critical values γ_{r_1+i,s_1}, $i = 1, \ldots, s_1 - r_1$, for different $r_1 < s_1$ in testing $m = 5$ null hypotheses using the BH-TSADC procedure based on Fisher's (Simes') combination function with prefixed early stopping boundaries $\lambda = 0.025$ and $\lambda' = 0.5$ and the FDR level $\alpha = 0.05$.

r_1	s_1	γ_{r_1+1}	γ_{r_1+2}	γ_{r_1+3}	γ_{r_1+4}	γ_{r_1+5}
0	1	0.0009				
		(0.0091)				
	2	0.0008	0.0017			
		(0.0083)	(0.0167)			
	3	0.0007	0.0016	0.0026		
		(0.0077)	(0.0154)	(0.0231)		
	4	0.0007	0.0015	0.0025	0.0035	
		(0.0071)	(0.0143)	(0.0214)	(0.0286)	
	5	0.0006	0.0015	0.0024	0.0033	0.0044
		(0.0067)	(0.0130)	(0.0200)	(0.0267)	(0.0333)
1	2	0.0027				
		(0.0250)				
	3	0.0024	0.0037			
		(0.0231)	(0.0308)			
	4	0.0023	0.0034	0.0046		
		(0.0214)	(0.0286)	(0.0357)		
	5	0.0022	0.0033	0.0043	0.0054	
		(0.0200)	(0.0267)	(0.0333)	(0.0400)	
2	3	0.0044				
		(0.0385)				
	4	0.0041	0.0054			
		(0.0357)	(0.0429)			
	5	0.0038	0.0051	0.0064		
		(0.0333)	(0.0400)	(0.0467)		
3	4	0.0061				
		(0.0500)				
	5	0.0057	0.0071			
		(0.0467)	(0.0533)			
4	5	0.0078				
		(0.0600)				

Table 2: The results of two-stage combination tests with Fisher's and Simes' combination functions with prefixed early stopping boundaries $\lambda = 0.025$ and $\lambda' = 0.5$ and the FDR level $\alpha = 0.05$ of 12625 probe sets in the Affymetrix Human U95A Chips data taken from Tian et al. (2003).

		Fisher's		Simes'		
		BH-TSADC	Plug-in BH-TSADC	BH-TSADC	Plug-in BH-TSADC	Single-Stage BH
Stage 1	Rejection	4	4	4	4	18
	Acceptance	10520	10520	10520	10520	12607
Stage 2	Rejection	140	89	36	28	NA
	Acceptance	1961	2012	2065	2073	NA
Total	Rejection	144	93	40	32	18

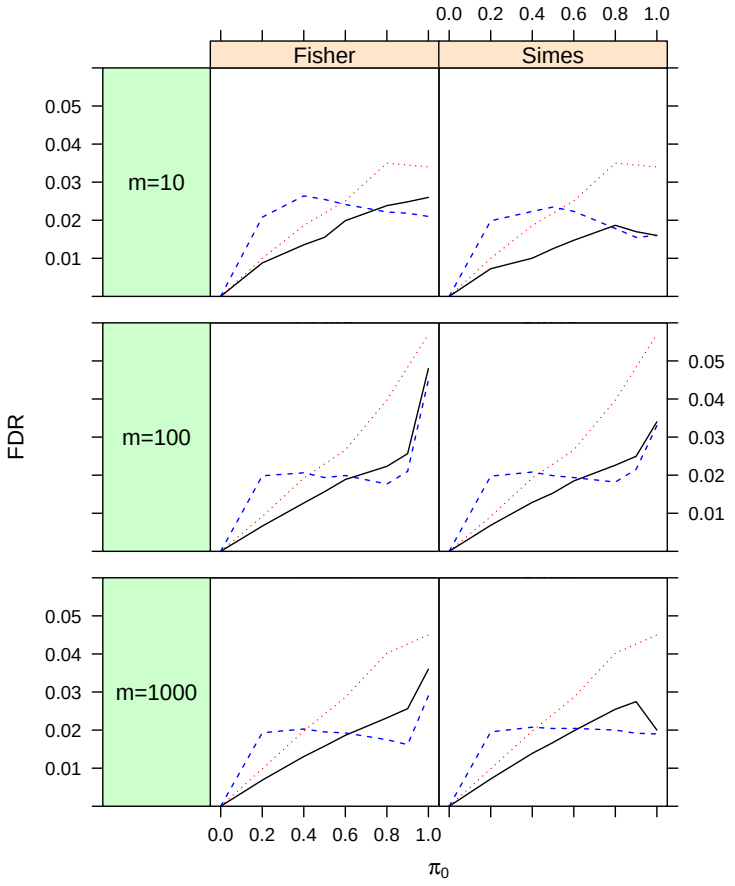

Figure 1: Comparison of simulated FDR's of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions, prefixed early stopping boundaries $\lambda = 0.025$ and $\lambda' = 0.5$ with simulated FDR of single-stage BH procedure, for $m = 10, 100$, and 1000, and $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dashed line, and BH: dotted line.]

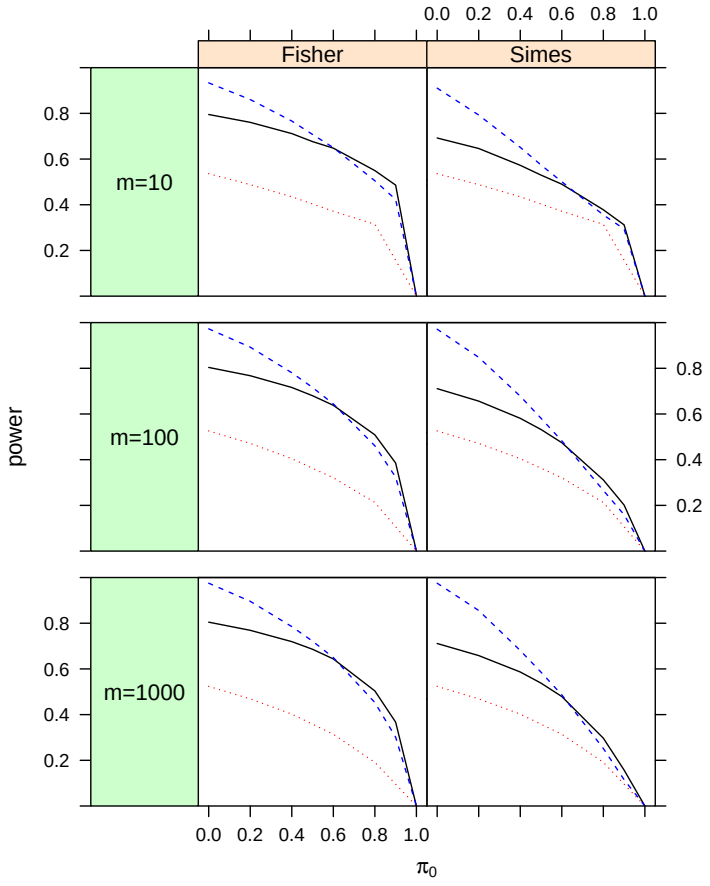

Figure 2: Comparison of simulated average power of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions, prefixed early stopping boundaries $\lambda = 0.025$ and $\lambda' = 0.5$ with the simulated average power of single-stage BH procedure, for $m = 10, 100$, and 1000, and $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dashed line, and BH: the dotted.]

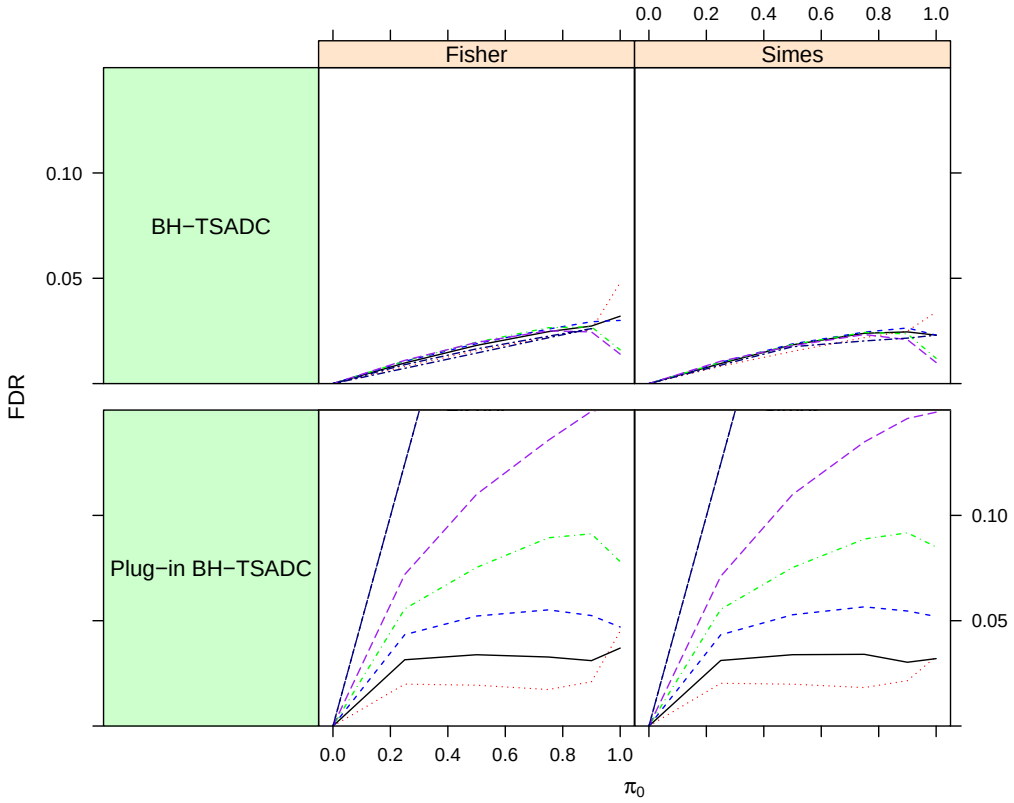

Figure 3: Comparison of simulated FDR of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions under equal dependence with early stopping boundaries $\lambda = 0.025$ and $\lambda' = 0.5$, $m = 100$, and $\alpha = 0.05$. [Solid line: $\rho = 0$; dash line: $\rho = 0.2$; dotted line: $\rho = 0.4$; dotdash line: $\rho = 0.6$; long-dash line: $\rho = 0.8$; and two-dash line: $\rho = 1$.]

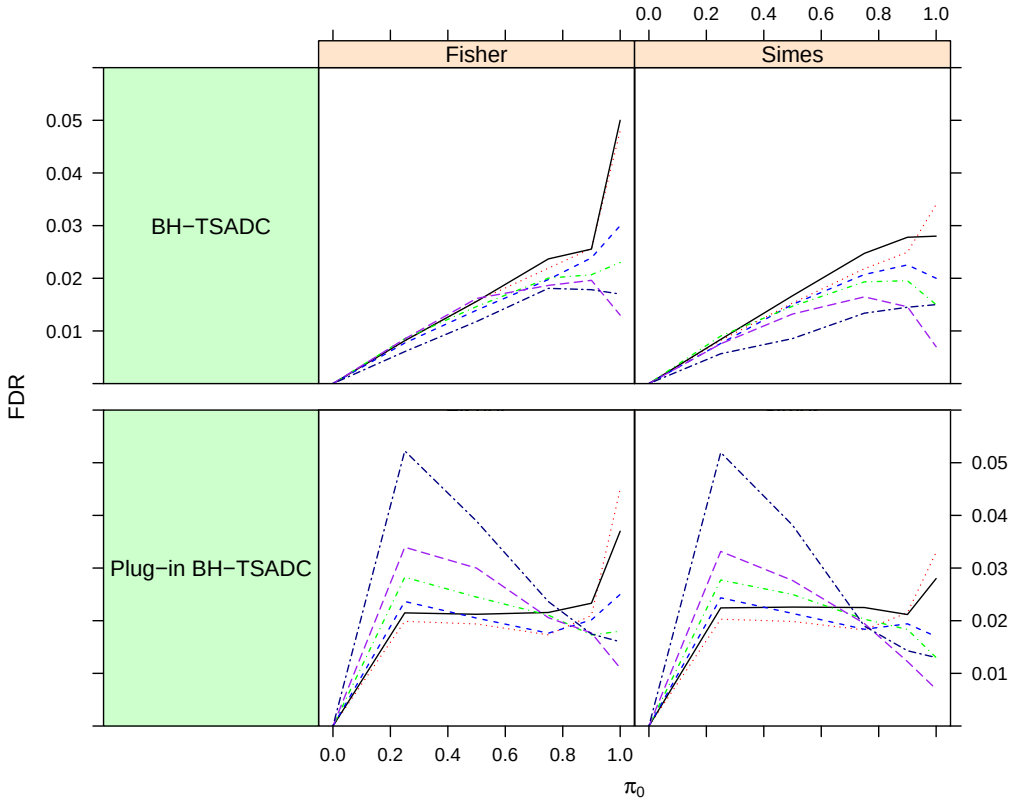

Figure 4: Comparison of simulated FDR of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions under clumpy dependence with early stopping boundaries $\lambda = 0.025$ and $\lambda' = 0.5$, $m = 100$, and $\alpha = 0.05$. [Solid line: $\rho = 0$; dash line: $\rho = 0.2$; dotted line: $\rho = 0.4$; dotdash line: $\rho = 0.6$; long-dash line: $\rho = 0.8$; and two-dash line: $\rho = 1$.]

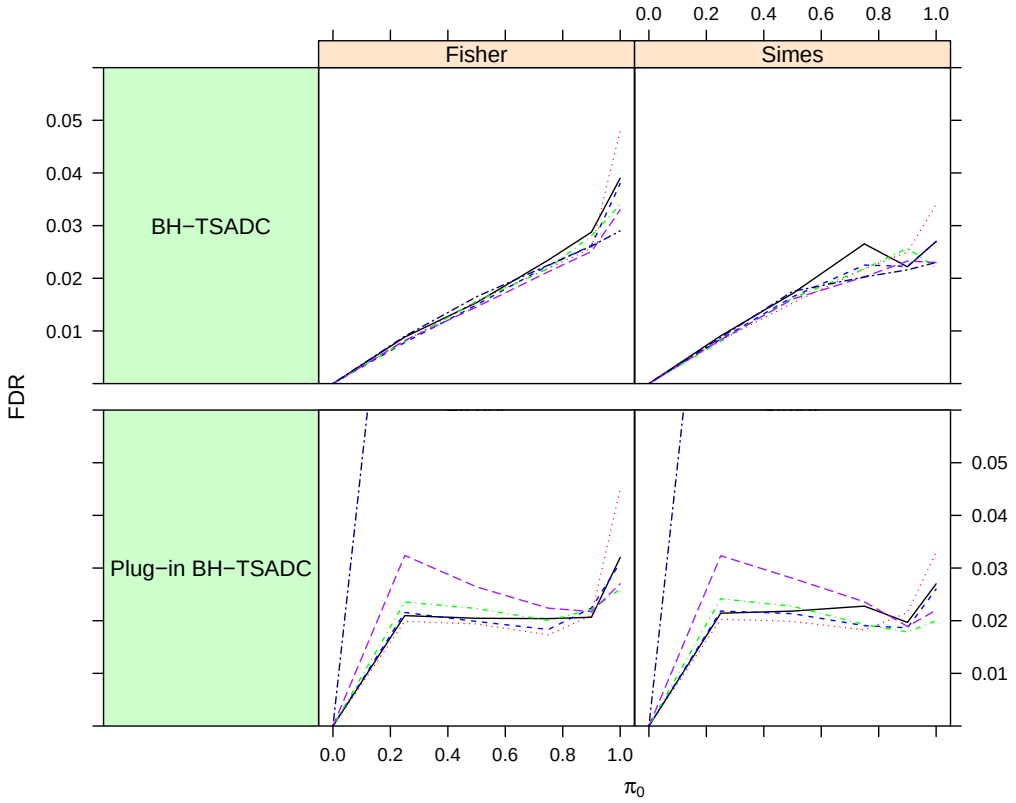

Figure 5: Comparison of simulated FDR of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions under AR(1) dependence with early stopping boundaries $\lambda = 0.025$ and $\lambda' = 0.5$, $m = 100$, and $\alpha = 0.05$. [Solid line: $\rho = 0$; dash line: $\rho = 0.2$; dotted line: $\rho = 0.4$; dotdash line: $\rho = 0.6$; long-dash line: $\rho = 0.8$; and two-dash line: $\rho = 1$.]