Published in final edited form as:

J Stat Comput Simul. 2014 April 1; 84(4): 812–818. doi:10.1080/00949655.2012.729212.

Relations among Three Parametric Multiple Testing Methods for Correlated Tests

Changchun Xie^{*,†}

Division of Epidemiology and Biostatistics, Department of Environmental Health, University of Cincinnati; Center for Clinical and Translational Science and Training, University of Cincinnati, Ohio 45267, USA

SUMMARY

Multiple endpoints in clinical trials are usually correlated. To control the family-wise type I error rate, both Huque and Alosch's flexible fixed-sequence (FFS) testing method and Li and Mehrotra's adaptive alpha allocation approach (4A) have taken into account correlations among endpoints. I suggested a weighted multiple testing correction (WMTC) for correlated tests and compared it with FFS. However, the relationship between the 4A method and the FFS method or the relationship between the 4A method and the WMTC method has not been studied. In this paper, simulations are conducted to investigate these relationships. Tentative guidelines to help choosing an appropriate method are provided.

Keywords

Bonferroni correction; Fixed-sequence; Correlated endpoints; Multivariate normal distribution; Weighted multiple testing correction

1. Introduction

To understand the overall effect of a treatment, investigators often collect many medically related endpoints in their clinical trials and test the treatment effect for each endpoint. Without multiple testing adjustment, the probability of false positive results, or family-wise type I error rate, increases as the number of tests increases. When some hypotheses are more important than others, weighted multiple testing correction methods are required. Many weighted multiple testing correction methods have been proposed in the literature.

Holm [1] developed the weighted Holm-Bonferroni method: Let $p_1, \dots, p_m$ be the unadjusted p-values and $w_i > 0, i=1, \dots, m$ be the corresponding weights that add to 1. Without loss of generality, assume that $q_1 \leq q_2 \leq \dots \leq q_m$, where $q_i = p_i/w_i, i=1, \dots, m$. The adjusted p-value for the j th hypothesis is $P_{adj,j} = \min(1, q_j), j=1; P_{adj,j} = \min(1, \max(P_{adj,(j-1)}, (w_j + \dots + w_m)q_j)), j=2, \dots, m$. If the adjusted p-value is less than the family-wise error rate, α , the corresponding hypothesis will be rejected.

Moyé [2,3] presented the prospective alpha allocation scheme (PAAS) to allow different alpha levels to be prospectively assigned for the multiple endpoints. Wiens [4] proposed a Bonferroni fixed sequence (BFS) procedure. Like PAAS, BFS gives a certain significance level to each of the null hypotheses. Unlike PAAS, BFS requires a pre-specified testing order, which allows the significance level to accumulate for testing further down in this

^{*}Correspondence to: Changchun Xie, Ph.D., Department of Environmental Health, University of Cincinnati, Ohio.

[†]xiecn@UCMAIL.UC.EDU

order, as each null hypothesis is rejected. Wiens and Dmitrienko [5–6] suggested a fallback procedure which developed the original BFS further and uses more available alpha to provide a testing procedure with more power than the original BFS. All the methods that we have discussed so far do not take into account correlation among endpoints.

Westfall and Young [7] have shown permutation (or bootstrapping) method can provide substantial improvements over Bonferroni correction since the dependence structures of the multiple endpoints are automatically incorporated into the analysis. However, permutation (or bootstrapping) method will not be discussed further in this paper since permutation (or bootstrapping) is computationally intensive and is rarely used to set the alpha level for sample size calculation at the time of protocol development because the data required by permutation (or bootstrapping) have not been collected yet. Huque and Alosh [8] developed a flexible fixed-sequence (FFS) testing method to improve the BFS approach by considering correlation among endpoints, which can be viewed as a parametric version of the fallback procedure. Li and Mehrotra [9] suggested an adaptive alpha allocation approach (4A), where the first endpoint is tested at a pre-specified level, the second endpoint at an adaptive level based on the p-value for the first endpoint and the correlation between the first endpoint and the second endpoint. I [10] proposed a weighted multiple testing correction (WMTC) for correlated tests, assuming that test statistics are asymptotically distributed as multivariate normal with known correlation matrix, and compared it with FFS. However, the relationship between 4A and FFS or the relationship between 4A and WMTC has not been studied.

In this paper, simulations are conducted to investigate the relationships among those parametric multiple testing methods for correlated tests. Tentative guidelines to help choosing an appropriate method are provided. The outline of this paper is as follows. The reviews of the FFS, 4A and WMTC are presented in Section 2. In Section 3, simulations are conducted to estimate the family-wise type I error rate of the three methods and compare the power and the significance levels of the three methods. Examples are given in Section 4 to illustrate the three methods and compare them. Section 5 provides the tentative guidelines to help choosing an appropriate method.

2. Reviews of the FFS, 4A and WMTC

In order to let the readers understand the FFS, 4A and WMTC methods without referring to the original papers, simple reviews of the three methods for two endpoints are given in this section. For details, see [8,9,10].

2.1. The FFS method

Suppose that T_1 and T_2 are test statistics for hypothesis $H_0^{(1)}$ and $H_0^{(2)}$ respectively. The FFS procedure is as follows:

- a. Test $H_0^{(1)}$ at the significance level α_1 ($< \alpha$) and reject $H_0^{(1)}$ if $T_1 \geq C_{1;\alpha_1}$, where $C_{1;\alpha_1}$ satisfies $P(T_1 \geq C_{1;\alpha_1} | H_0^{(1)}) = \alpha_1$.
- b. If $H_0^{(1)}$ is rejected, Test $H_0^{(2)}$ at the significance level α . If $H_0^{(1)}$ is not rejected, Test $H_0^{(2)}$ at the significance level α_2 and reject $H_0^{(2)}$ if $T_2 \geq C_{2;\alpha_2}$, where $C_{2;\alpha_2}$ satisfies $P(T_1 < C_{1;\alpha_1}, T_2 \geq C_{2;\alpha_2} | H_0^{(1)}, H_0^{(2)}) = \alpha - \alpha_1$.

2.2. The 4A method

Assume P_1 and P_2 are the p-value for the first endpoint and the second endpoint respectively. Let ρ is the correlation between the two endpoints. The 4A method tests the

null hypothesis for the first endpoint at the prespecified level $\alpha_1 (< \alpha)$ and tests the null hypothesis for the second endpoint at the adaptive level $\alpha_2 = \alpha$ if $P_1 \leq \alpha_1$ or $\alpha_2 = \min(\alpha_1, \lambda \alpha_t / P_1^2)$ if $P_1 > \alpha_1$, where λ is the largest constant such that $P(P_1 > \alpha_1, P_2 \leq \alpha_2 | P_1, \alpha_1, \alpha, \rho) \leq \alpha - \alpha_1$ and $\alpha_t = \alpha_1 (a - \alpha_1) / (1 - \alpha_1)$ if $\alpha_1 + \alpha_1^2 - \alpha_1^3 > \alpha$, otherwise, $\alpha_t = \alpha_1 \{1 - \sqrt{(2\alpha_1 - \alpha - \alpha_1^2) / \alpha_1}\}^2$.

2.3. The WMTC method

Let p_1, p_2 be the observed p-values for null hypotheses $H_0^{(1)}, H_0^{(2)}$ respectively and $w_i > 0$, $i=1, 2$ be the weight for null hypothesis $H_0^{(i)}$. For each $i=1, 2$, let $q_i = p_i / w_i$. Then the adjusted p-value for the null hypothesis $H_0^{(i)}$ is

$$\begin{aligned} P_{adj_i} &= P(\min_j q_j \leq q_i) \\ &= 1 - P(\text{all } q_j > q_i) \\ &= 1 - P(\text{all } p_j / w_j > p_i / w_i) \quad (1) \\ &= 1 - P(\text{all } p_j > p_i w_j / w_i) \\ &= 1 - P\left(\bigcap_{j=1}^m a_j \leq X_j \leq b_j\right), \end{aligned}$$

where $X_j, j=1, 2$ are standardized multivariate normal with correlation matrix Σ and

$$a_j = \Phi^{-1}\left(\frac{p_i w_j}{2w_i}\right), \quad b_j = \Phi^{-1}\left(1 - \frac{p_i w_j}{2w_i}\right) \quad (2)$$

for the two-sided case.

If $P_{adj_i} \leq \alpha$, we will reject the corresponding null hypothesis $H_0^{(i)}$. If only one null hypotheses has been rejected, we can reject the other hypothesis if its observed p-value $\leq \alpha$.

3. Comparing the three parametric methods

Both the FFS and the 4A methods require the pre-specified testing sequence, while the WMTC method does not. Both the FFS and the 4A method have the property that the rejection of earlier hypothesis (usually more important hypothesis) does not depend on the rejection of the later hypothesis (less important hypothesis), while the WMTC method does not have this property. Both the FFS and the 4A method can only handle a very limited number of endpoints. The FFS method faces a computational difficulty when there are four or more endpoints, and the 4A method faces a computational difficulty when there are three or more endpoints. The WMTC method can handle up to a thousand endpoints, using the package “mvtnorm” in R [11,12]. In this section, simulations are conducted to compare the power, the significance level, and the family-wise type I error rate of the three parametric multiple testing methods.

We considered trials with two endpoints. Each trial has 240 individuals. Each individual had probability 0.5 to receive the active treatment and probability 0.5 to receive placebo. The endpoints were generated from a bivariate normal distribution with the correlation between the two endpoints, ρ chosen as 0.0, 0.3, 0.5, 0.7 and 0.9. The treatment effect size (per unit standard deviation) was assumed as 0.0, 0.1 and 0.4. The weights for the two endpoints were (4, 1), which correspond to alpha allocations (0.04, 0.01). The observed p-values were calculated using two-sided t-tests for the coefficient of the treatment, $\beta = 0$, in linear

regressions. Since the critical values and significance levels in the FFS and 4A methods might not be available, the known correlation was used. For the WMTC method, the estimated correlation (from simulated data) was used. The simulation results are shown in Table 1. From these simulations, we can conclude:

1. All the three methods have the simulated family-wise type I error rate at 5.0% (the first part of Table 1). However, the significance level for the first hypothesis in the FFS and 4A does not change, while the significance levels for all the hypotheses in the WMTC method increase, when the correlation between the two endpoints, ρ increases. This may give insight to why the WMTC procedure results in increasing the power of the tests for all endpoints as ρ increases. The significance levels for the second hypothesis in the FFS and 4A methods are similar. They increase as ρ increases and increase a bit faster than the significance levels for the second hypothesis in the WMTC method.
2. The FFS and 4A methods have the same power for testing the first hypothesis in the testing sequence and do not change when ρ increases. The WMTC method has higher power for testing the first hypothesis than the FFS and 4A method. In the FFS and 4A method, when the first and second endpoints have the same effect size, the power for testing the first hypothesis can be lower than the power for the second hypothesis although the α allocation given to the first hypothesis is higher (see the last part of Table 1). Given that hypotheses are ranked in order of importance, this reversal of desired power for the two hypotheses is inappropriate. The relative importance of the two hypotheses is clear in the WMTC method, whereby the hypothesis with larger weight always has higher power when the endpoints have the same effect size.
3. The 4A method was originally proposed to handle the situation, where the first endpoint is adequately powered and the second endpoint is potentially underpowered. Based on our simulations (assuming the effect size is 0.4, 0.1 for the first, second endpoint respectively), the 4A method does show higher power for the second endpoint compared to the FFS and WMTC method, when the correlation between the two endpoint is low, however, when the correlation between the two endpoints is high, the 4A method does not show any advantage on power for the second endpoint. When the second endpoint is adequately powered and the first endpoint is underpowered, as expected, our simulations (assuming the effect size is 0.1, 0.4 for the first, second endpoint respectively) show the 4A method performs badly compared to the FFS and WMTC method, especially when the correlation between the two endpoints is high. The WMTC method has higher power for the first endpoint than the FFS method, however the FFS method has higher power for the second endpoints and higher power to reject at least one null hypothesis than the WMTC method. When the first and the second endpoints have the same effect size, the 4A method has higher power for the second endpoint and higher power to reject at least one null hypothesis compared to the FFS and WMTC method.

4. Examples

In this section, examples have been provided to illustrate the performances of the three methods for different scenarios.

4.1. Example 1

Assume a two-arm trial with two-sided $\alpha = 0.05$, two endpoints, the correlation between the two endpoints, $\rho = 0.7$, a corresponding α allocations (0.04, 0.01), and the observed p-values: 0.043, 0.016.

The FFS method rejects the second null hypothesis of no treatment difference for the second endpoint since $0.016 < 0.0171479$, based on the significance level provided by Huque and Alosch [8]. The 4A method rejects the second null hypothesis of no treatment difference for the second endpoint since $0.016 < 0.04$, based on the significance level provided by Li and Mehrotra [9]. However, both the FFS and the 4A methods cannot reject the first null hypothesis of no treatment difference for this first and most important endpoint since $0.043 > 0.04$. In fact, based on these two methods, the power for the first endpoint is independent of the correlation between the endpoints. Clearly, both the FFS and the 4A methods cannot reject the first null hypothesis, even as ρ increases to 0.99. The WMTC method, which does not require the testing sequence, using $\rho = 0.7$ and the weight (4,1), which corresponds to α allocations (0.04, 0.01) gives the “working” adjusted p-values: 0.0485 and 0.0715 respectively. The first null hypotheses can be rejected. After removing the rejected null hypotheses, only one test (for the second endpoint) remains. Since $0.016 < 0.05$, the second null hypothesis is rejected.

4.2. Example 2

This example is the same as Example 1 except that the correlation between the two endpoints, ρ is 0.2.

The FFS method cannot reject the first and the second null hypotheses of no treatment difference since $0.043 > 0.04$ and $0.016 > 0.0107472$. The 4A method cannot reject the first null hypothesis, but can reject the second null hypothesis since $0.016 < 0.04$. The WMTC method using $\rho = 0.2$ and the weight (4,1) gives the “working” adjusted p-values: 0.0530 and 0.0784 respectively. Since all the adjusted p-values > 0.05 , the WMTC cannot reject both the first and the second null hypotheses.

4.3. Example 3

This example is the same as Example 1 except that the observed p-values for the first and the second endpoints are 0.5 and 0.016, respectively.

As in the example1, the FFS method cannot reject the first null hypothesis, but can reject the second null hypothesis. The 4A method cannot reject the first and the second null hypotheses since $0.05 > 0.04$ and $0.016 > 0.0009$. The WMTC method using $\rho = 0.7$ and the weight (4,1) gives the “working” adjusted p-values: 0.0562 and 0.0715, respectively. Since all the adjusted p-values > 0.05 , the WMTC cannot reject both the first and the second null hypotheses.

4.4. Example 4

This example is the same as Example 1 except that the observed p-values for the first and the second endpoints are 0.043 and 0.042, respectively.

The FFS method cannot reject the first and the second null hypotheses of no treatment difference, since $0.043 > 0.04$ and $0.042 > 0.0171479$. The 4A method cannot reject the first and the second null hypotheses either since $0.043 > 0.04$ and $0.042 > 0.04$. The WMTC method, using $\rho = 0.7$ and the weight (4, 1) gives the “working” adjusted p-values: 0.0485 and 0.1825, respectively. The first null hypotheses can be rejected. After removing the rejected null hypotheses, only one test (for the second endpoint) remains. Since $0.042 < 0.05$, the second null hypothesis is rejected. The WMTC can reject both null hypotheses, which cannot be rejected by the FFS and the 4A method.

5. Discussion

This paper compared the three parametric multiple testing methods (FFS, 4A and WMTC). The following is the tentative guidelines to help choosing an appropriate method:

The WMTC method may be preferred among the three methods compared when one of the following conditions is satisfied:

1. There are more than three endpoints.
2. The testing sequence is not easy to be specified.
3. The endpoint with the most weight (usually the first one) has the highest power.
4. The power for all endpoints increases as the correlation between endpoint increases.
5. The endpoint with the more weight potentially has larger effect size and the correlation between endpoints is high.

The WMTC method should not be chosen if the investigator needs the property that the rejection of earlier hypothesis (usually more important hypothesis) does not depend on the rejection of the later hypothesis (less important hypothesis). The FFS method or the 4A method might be the choice.

The FFS method may be preferred among the three methods compared if the investigator's primary interest is to reject any of the two null hypotheses when the second endpoint is adequately powered and the first endpoint is underpowered.

The 4A method may be preferred among the three methods compared if the investigator's primary interest is to reject any of the two null hypotheses when the first and the second endpoints have the same effect size.

Acknowledgments

The author wishes to thank the editor and the referee for their insightful and constructive comments that have greatly improved this paper. This project was partially supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 8 UL1 TR000077-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 1979; 6:65–70.
2. Moyé LA. Alpha calculus in clinical trials: considerations and commentary for the new millennium. *Statistics in Medicine*. 2000; 19:767–779. [PubMed: 10734280]
3. Moyé LA. P-value interpretation and alpha allocation in clinical trials. *Annals of Epidemiology*. 1998; 8:351–357. [PubMed: 9708870]
4. Wiens BL. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*. 2003; 2:211–215.
5. Wiens BL, Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics*. 2005; 15:929–942. [PubMed: 16279352]
6. Wiens BL, Dmitrienko A. On selecting a multiple comparison procedure for analysis of a clinical trial: fallback, fixed sequence, and related procedures. *Statistics in Biopharmaceutical Research*. 2010; 2:22–32.
7. Westfall, PH.; Young, SS. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley; New York: 1993.

8. Huque MF, Alos M. A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference*. 2008; 138:321–335.
9. Li J, Mehrotra DV. An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in Medicine*. 2008; 27:5377–5391. [PubMed: 18759248]
10. Xie C. Weighted multiple testing correction for correlated tests. *Statistics in Medicine*. 2012; 31:341–352. [PubMed: 22114000]
11. Genz, A.; Bretz, F.; Hothorn, T. mvtnorm: multivariate normal and t distribution. R package version 2.12.0. 2010. (available at <http://cran.r-project.org/web/packages/mvtnorm/index.html>)
12. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2010. (available at <http://www.r-project.org/>)

Table 1

Two endpoints: Simulated power (%), the significance level (%) and the family-wise type I error rate (%) based on 10,000,000 runs for selected treatment differences for WMTC, FFS and 4A. The total sample size is 240. α allocation is (0.04, 0.01) or weight is (4, 1).

Effect size	ρ	WMTC	FFS	4A
0.0, 0.0	0.0	4.0, 1.2	(5.0)	4.0, 1.2 (5.0)
	0.3	4.1, 1.3	(5.0)	4.0, 1.4 (5.0)
	0.5	4.3, 1.6	(5.0)	4.0, 1.8 (5.0)
	0.7	4.5, 2.1	(5.0)	4.0, 2.4 (5.0)
	0.9	4.9, 3.1	(5.0)	4.0, 3.6 (5.0)
0.1, 0.1	0.0	10.3, 4.5	(13.5)	10.2, 4.8 (13.8)
	0.3	10.5, 5.2	(13.1)	10.2, 6.0 (13.7)
	0.5	10.7, 5.9	(12.8)	10.2, 6.7 (13.2)
	0.7	11.1, 6.8	(12.4)	10.2, 7.7 (12.8)
	0.9	11.8, 8.6	(12.1)	10.2, 9.7 (12.3)
0.4, 0.1	0.0	84.9, 10.7	(85.4)	84.7, 11.5 (86.0)
	0.3	85.0, 11.4	(85.1)	84.7, 11.7 (85.2)
	0.5	85.2, 11.7	(85.3)	84.7, 11.8 (84.9)
	0.7	85.7, 11.7	(85.8)	84.7, 11.8 (84.8)
	0.9	86.6, 11.8	(86.6)	84.7, 11.9 (84.9)
0.1, 0.4	0.0	11.5, 70.9	(72.2)	10.2, 71.2 (72.6)
	0.3	11.7, 70.3	(70.9)	10.2, 71.3 (71.9)
	0.5	11.8, 70.2	(70.4)	10.2, 72.6 (72.9)
	0.7	11.9, 70.4	(70.6)	10.2, 75.5 (75.8)
	0.9	12.0, 71.3	(71.6)	10.2, 81.5 (81.7)
0.4, 0.4	0.0	86.2, 84.1	(95.3)	84.7, 84.2 (95.3)
	0.3	86.0, 83.3	(92.7)	84.7, 83.4 (92.8)
	0.5	86.0, 83.0	(90.9)	84.7, 83.3 (91.1)
	0.7	86.2, 82.9	(89.1)	84.8, 83.6 (89.5)
	0.9	86.6, 83.6	(87.3)	84.7, 84.8 (87.8)

The first cell entry is for the first endpoint and the second entry is for the second endpoint. The probability (%) that at least one hypothesis is rejected is given in brackets.