Effect of Misspecified Correlations in Parametric Multiple Testing Methods for Correlated Tests

Changchun Xie[1,2*†] , Xuewen Lu[3], Radhey Singh[4] and Ding-Geng (Din) Chen[5,6,7]

[1]Division of Epidemiology and Biostatistics, Department of Environmental Health, [2]Center for Clinical and Translational Science and Training, University of Cincinnati, Ohio, USA
[3]Department of Mathematics and Statistics, University of Calgary, Alberta, Canada
[4]Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada
[5]School of Nursing, [6]Department of Biostatistics and Computational Biology, [7]Department of Orthopeadics and Rehabilitation, School of Medicine and Dentistry, University of Rochester Medical Center, Rochester, New York, USA

SUMMARY

In clinical trials, multiple endpoints are usually correlated. However, many commonly used multiple testing correction methods proposed to control family-wise type I errors disregard the correlation among the endpoints, for example, the Bonferroni correction and Holm procedure. Recently, some parametric multiple testing methods have been proposed to take into account correlations among endpoints. However, the exact correlations among endpoints are usually unknown. If the correlations are misspecified, how robust are these parametric multiple testing methods in controlling family-wise type I errors? In this paper, simulations are conducted to study the effect of misspecified correlations in these parametric multiple testing methods along with an example to address this question.

KEY WORDS: Bonferroni correction; Correlated endpoints; Multivariate normal distribution; Weighted multiple testing correction

[*]Correspondence to: Changchun Xie, Ph.D., Department of Environmental Health, University of Cincinnati, Ohio. [†]E-mail: xiecn@UCMAIL.UC.EDU

1. Introduction

In a clinical trial with multiple endpoints, multiple tests are performed. Multiple testing correction methods are usually used to control the family-wise Type I error rate (FWER). When some hypotheses are more important than others, weighted multiple testing correction methods are required. However, many weighted methods proposed to control the FWER often disregard the correlations among the endpoints. Recently, some weighted parametric multiple testing methods have been suggested to take into account the correlations among the endpoints, such as Huque and Alosh's (2008) flexible fixed-sequence (FFS) testing method and Li and Mehrotra's (2008) adaptive α allocation approach (4A), Xie's (2012) weighted multiple testing correction (WMTC), Bretz *et al.* 's (2011) graphical approaches and Millen and Dmitrienko's (2011) chain procedures. The chain procedure gives weights on test statistics, while FFs, 4A and WMTC give weights on p-values. Xie (2012) compared FFs, 4A and WMTC and provided guidelines to help choosing an appropriate method. All these weighted parametric multiple testing methods assume the correlation among endpoints are known or can be correctly estimated from the data. However, the exact correlations among endpoints are usually unknown. If the correlations are misspecified, the type I error rate can be out of control. We usually think the type I error rate will be inflated if the correlations are overestimated (Wiens and Dmitrienko, 2010). From our simulations below, we will see this is not true for 4A method. When the inflation occurs and how large the inflation can be might depend on the method and the magnitude of the misspecification of the correlations among the endpoints. So far, from our best knowledge, the exact effect of the misspecification of the correlations has not been studied yet. To fill this gap, simulations are conducted in this paper to study the effect of misspecified correlations in these weighted parametric multiple testing methods. The graphical approach and chain procedures are not included in the comparisons since for the time being the implement of graphical approach (gMCP) is based on one-sided test only and the chain procedure gives weights on test statistics instead of P-values, which change the interpretation due to a nonlinear relationship between P-values and test statistics. 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 when the correlations are misspecified. Examples are given in Section 4 to illustrate the effect of misspecified correlations in these parametric multiple testing methods. Finally, some discussions and concluding remarks are made in Section 5.

2. Reviews of the FFS, 4A and WMTC

In this section, the simple reviews of the three methods (FFS, 4A and WMTC) for two endpoints are given. For details, see (Huque and Alosh, 2008; Li and Mehrotra, 2008; Xie, 2012).

2.1. The FFS method

Let T_1 and T_2 be test statistics for hypothesis $H_0^{(1)}$ and $H_0^{(2)}$ respectively, the FFS procedure can be stated as follows:

i) Test $H_0^{(1)}$ at the significance level $\alpha_1(<\alpha)$ and reject $H_0^{(1)}$ if $T_1 \ge C_{1;\alpha_1}$, where $C_{1;\alpha_1}$ satisfies $P(T_1 \ge C_{1;\alpha_1} \mid H_0^{(1)}) = \alpha_1$.

ii) 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 $T_2 \ge C_{2;\alpha_2}$,

where $C_{2;\alpha_2}$ satisfies $P(T_1 < C_{1;\alpha_1}, T_2 \ge C_{2;\alpha_2} \mid H_0^{(1)}, H_0^{(2)}) = \alpha - \alpha_1$.

2.2. The 4A method

Let p_1 and p_2 be the p-value for the first endpoint and the second endpoint respectively and let ρ be the correlation between the two endpoints, the 4A method can be stated as follows:

i) Test the null hypothesis for the first endpoint at the level $\alpha_1 (< \alpha)$.

ii) If $p_1 \le \alpha_1$, test the null hypothesis for the second endpoint at the level $\alpha_2 = \alpha$.

If $p_1 > \alpha_1$, test the null hypothesis for the second endpoint at the adaptive level $\alpha_2 = \min(\alpha_1, \lambda\alpha_t / {p_1}^2)$, where λ is the largest constant such that $P(p_1 > \alpha_1, p_2 \le \alpha_2(p_1, \alpha_1, \alpha, \rho)) \le \alpha - \alpha_1$ and $\alpha_t = \alpha_1(\alpha - \alpha_1)/(1 - \alpha_1)$ 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, $w_i (> 0)$ be the weight for null hypothesis $H_0^{(i)}$, $i = 1,2$ and let $q_i = p_i / w_i$, $i = 1,2$. (Note p_i and q_i are realizations of random variables Q_i and P_i respectively, $i = 1,2$). The adjusted p-value for the null hypothesis $H_0^{(i)}$ is

$$\begin{aligned} P_{adj\_i} &= P(\min_j Q_j \le q_i) \\ &= 1 - P(all \quad Q_j > q_i) \\ &= 1 - P(all \quad P_j / w_j > p_i / w_i) \end{aligned}$$

$$= 1 - P(all \quad P_j > p_i w_j / w_i)$$

$$= 1 - P\left(\bigcap_{j=1}^{m} a_j \le X_j \le b_j \right), \qquad (1)$$

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

$$a_j = \Phi^{-1}(p_i w_j / (2w_i)), \qquad b_j = \Phi^{-1}(1 - p_i w_j / (2w_i)) \qquad (2)$$

for the two-sided case.

If $P_{adj\_i} \le \alpha$, we will reject the corresponding null hypothesis $H_0^{(i)}$. If only one null hypotheses has been rejected, we test the other hypothesis at level α and reject the hypothesis if its observed p-value $\le \alpha$. The package mvtnorm (Genz, Bretz and Hothorn, 2010) in the R software environment (http://www.r-project.org/) is used to compute of the adjusted p-values in (1) and (2).

3. Simulation

In this section, simulations are conducted to compare the family-wise type I error rate of the three parametric multiple testing methods when the correlations are misspecified. 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, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9. The treatment effect size (per unit standard deviation) was assumed as 0.0. 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. For each correlation given in data generation, the specified correlations, 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 were used in multiple testing adjustments. 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% if the correlation is correctly specified. 2) Both WMTC and FFS have inflated family-wise type I error rate if the correlation among the endpoints is overspecified and the family-wise type I error rate increases with the magnitude of the misspecification of the correlations. Family-wise type I error rate of FFS increases faster than that of WMTC. Both WMTC and FFS have deflated family-wise type I error rate if the correlation among the endpoints is underspecified and the family-wise type I error rate decreases with the magnitude of the misspecification of the correlations. Family-wise type I error rate of FFS decreases a little bit faster than that of WMTC. 3) Unlike WMTC and FFS, 4A has deflated family-wise type I error rate if the correlation among the endpoints is overspecified and the family-wise type I error rate decreases with the magnitude of the misspecification of the correlations. 4A has inlated family-wise type I error rate if the correlation among the endpoints is underspecified and the family-wise type I error rate increases with the magnitude of the misspecification of the correlations. 4) When the true correlation among the endpoints > 0.5, the effect of the misspecified correlation in both WMTC and FFS methods can be large even the magnitude of the misspecification of the correlations is 0.1. However, the effect of the misspecified correlation in 4A method will not be serious if the magnitude of the misspecification of the correlations ≤ 0.1.

4. Example

Assume a two group trial with two-sided $\alpha = 0.05$, two endpoints with true correlation between them, $\rho = 0.7$, the observed p-values: 0.11, 0.02 and the corresponding α allocations: (0.04, 0.01).

The FFS method cannot reject any of the two null hypotheses of no treatment difference since 0.11>0.04 and 0.02>0.017, based on the significance level provided by Huque and Alosh (2008). The 4A method cannot reject any of the two null hypotheses of no treatment difference since 0.11>0.04 and 0.02>0.019, based on the significance level provided by Li and Mehrotra (2008). The WMTC method cannot reject any of the two null hypotheses of no treatment difference either since both "working" adjusted p-values (0.121 and 0.089 respectively) >0.05.

If we over-specify the correlation between the two endpoints, for example, $\rho = 0.9$, the FFS method can reject the second null hypothesis of no treatment difference for the second endpoint since 0.02<0.029, based on the significance level provided by Huque and Alosh [1]. The 4A method cannot reject any of the two null hypotheses of no treatment difference since 0.11>0.04 and 0.02>0.018, based on the significance level provided by Li and Mehrotra [2]. The WMTC method cannot reject any of the two null hypotheses of no treatment difference either since both "working" adjusted p-values (0.112 and 0.082 respectively) >0.05.

If we under-specify the correlation between the two endpoints, for example, $\rho = 0.5$, The 4A method can reject the second null hypothesis of no treatment difference for the second endpoint since 0.02<0.032, based on the significance level provided by Li and Mehrotra (2008). As expected, the FFS method cannot reject any of the two null hypotheses of no treatment difference since 0.11>0.04 and 0.02>0.013, and the WMTC method cannot reject any of the two null hypotheses of no treatment difference either since both “working” adjusted p-values (0.128 and 0.094 respectively) >0.05.

5. Concluding remarks and discussions

A multiple testing method is often thought to be conservative if it ignores the correlation among the endpoints (that is, it underestimates the correlation). This is not true for 4A method. The FWER in 4A method can be inflated if the correlation is underestimated. The 4A method is conservative if the correlation is overestimated. As usual, both WMTC and FFS are conservative if the correlation among the endpoints is underestimated and the FWER can be inflated if the correlation is overestimated. If the magnitude of the misspecification of the correlations is 0.1 or less, the effect of the misspecified correlation in 4A method will not be serious, but it can be serious for WMTC method and more serious for FFS method if the true correlation among the endpoints > 0.5. Developing a new method, which is more robust than FFS, WMTC and 4A methods and is more powerful than nonparametric methods such as the Holm and the alpha-exhaustive fallback procedures will be our future research.

Acknowledgement

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. Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W. and Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal* **53**, 894-913.
2. Genz, A., Bretz, F., Hothorn, T. (2010). mvtnorm: multivariate normal and t distribution. R package version 2.12.0. (available at http://cran.r-project.org/web/packages/mvtnorm/index.html)
3. Huque, M. F. and Alosh, M. (2008). A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference* **138,** 321-335.
4. Li, J. and Mehrotra, D. V. (2008) An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in Medicine* **27**, 5377-5391.
5. Millen, B. A. and Dmitrienko, A. (2011). Chain procedures: A class of flexible closed testing procedures with clinical trial applications. *Statistics in Biopharmaceutical Research* **3**, 14–30.
6. R Development Core Team. (2010). R: a language and environment for statistical computing. R Foundation for Statistical Computing. (available at http://www.r-project.org/)
7. Wiens, B. L., and Dmitrienko, A. (2010). On selecting a multiple comparison procedure for analysis of a clinical trial: fallback, fixed sequence, and related procedures. *Statistics in Biopharmaceutical Research* **2,** 22-32.

8. Xie, C. (2012). Weighted multiple testing correction for correlated tests. *Statistics in Medicine* **31**, 341-352.

9. Xie, C. (2012). Relations among Three Parametric Multiple Testing Methods for Correlated Tests. *Journal of Statistical Computation and Simulation*, DOI:10.1080/00949655.2012.729212

Table 1. Two endpoints: Simulated family-wise type I error rate (%) based on 1,000,000 runs for WMTC, FFS and 4A when different correlations are specified in multiple testing adjustments. The total sample size is 240. α allocation is (0.04, 0.01) or weight is (4, 1).

True correlations	Methods	Specified correlations in multiple testing adjustments									
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	WMTC	5.0	5.0	5.0	5.1	5.1	5.2	5.3	5.5	5.7	6.0
	FFS	5.0	5.0	5.0	5.1	5.1	5.2	5.4	5.6	6.0	6.8
	4A	5.0	5.0	4.9	4.8	4.7	4.6	4.5	4.4	4.4	4.4
0.1	WMTC	5.0	5.0	5.0	5.0	5.1	5.2	5.3	5.5	5.7	6.0
	FFS	5.0	5.0	5.0	5.0	5.1	5.2	5.4	5.6	6.0	6.8
	4A	5.0	5.0	4.9	4.8	4.7	4.6	4.5	4.4	4.4	4.4
0.2	WMTC	5.0	5.0	5.0	5.0	5.1	5.2	5.3	5.5	5.7	6.0
	FFS	5.0	5.0	5.0	5.0	5.1	5.2	5.4	5.6	6.0	6.7
	4A	5.1	5.1	5.0	4.9	4.8	4.7	4.6	4.5	4.5	4.5
0.3	WMTC	4.9	4.9	5.0	5.0	5.1	5.1	5.3	5.4	5.6	5.9
	FFS	4.9	4.9	5.0	5.0	5.1	5.1	5.3	5.5	5.9	6.7
	4A	5.2	5.1	5.1	5.0	4.9	4.8	4.6	4.6	4.5	4.5
0.4	WMTC	4.9	4.9	4.9	5.0	5.0	5.1	5.2	5.4	5.6	5.8
	FFS	4.9	4.9	4.9	4.9	5.0	5.1	5.2	5.5	5.8	6.5
	4A	5.3	5.3	5.2	5.1	5.0	4.9	4.8	4.7	4.6	4.6
0.5	WMTC	4.8	4.8	4.8	4.9	4.9	5.0	5.1	5.3	5.5	5.7

	FFS	4.8	4.8	4.8	4.9	4.9	5.0	5.1	5.4	5.7	6.4
	4A	5.5	5.4	5.4	5.3	5.1	5.0	4.9	4.8	4.7	4.8
0.6	WMTC	4.7	4.7	4.7	4.8	4.8	4.9	5.0	5.2	5.4	5.6
	FFS	4.7	4.7	4.7	4.7	4.8	4.9	5.0	5.2	5.5	6.2
	4A	5.6	5.6	5.5	5.4	5.3	5.1	5.0	4.9	4.8	4.8
0.7	WMTC	4.6	4.6	4.6	4.6	4.7	4.8	4.9	5.0	5.2	5.5
	FFS	4.6	4.6	4.6	4.6	4.7	4.7	4.8	5.0	5.3	5.9
	4A	5.7	5.7	5.6	5.5	5.4	5.3	5.1	5.0	4.9	4.9
0.8	WMTC	4.4	4.4	4.4	4.5	4.5	4.6	4.7	4.8	5.0	5.3
	FFS	4.4	4.4	4.4	4.4	4.5	4.5	4.6	4.8	5.0	5.6
	4A	5.7	5.7	5.7	5.6	5.5	5.3	5.2	5.1	5.0	5.0
0.9	WMTC	4.2	4.2	4.2	4.2	4.3	4.4	4.5	4.6	4.8	5.0
	FFS	4.2	4.2	4.2	4.2	4.2	4.2	4.3	4.4	4.6	5.0
	4A	5.5	5.5	5.5	5.4	5.4	5.3	5.2	5.1	5.0	5.0