STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:1712–1723
Published online 24 January 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2795

Statistical power of multiplicity adjustment strategies for correlated binary endpoints

Andrew C. Leon[1,2,*,†], Moonseong Heo[1], Jedediah J. Teres[1] and Toshihiko Morikawa[3]

[1] *Department of Psychiatry, Weill Cornell Medical College, Box 140, 525 East 68th Street, New York, NY 10021, U.S.A.*
[2] *Department of Public Health, Weill Cornell Medical College, Box 140, 525 East 68th Street, New York, NY 10021, U.S.A.*
[3] *Biostatistics Center, Kurume University, 67 Asahimachi, Kurume City 830-0011, Japan*

SUMMARY

There are numerous alternatives to the so-called Bonferroni adjustment to control for familywise Type I error among multiple tests. Yet, for the most part, these approaches disregard the correlation among endpoints. This can prove to be a conservative hypothesis testing strategy if the null hypothesis is false. The James procedure was proposed to account for the correlation structure among multiple continuous endpoints. Here, a simulation study evaluates the statistical power of the Hochberg and James adjustment strategies relative to that of the Bonferroni approach when used for multiple correlated binary variables. The simulations demonstrate that relative to the Bonferroni approach, neither alternative sacrifices power. The Hochberg approach has more statistical power for $\rho \leqslant 0.50$; whereas the James procedure provides more statistical power with higher ρ, the common correlation among the multiple outcomes. A study of gender differences in New York City homicides is used to illustrate the approaches. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: James adjustment; Bonferroni adjustment; multiplicity; multiple comparisons; statistical power; correlated binary endpoints

1. INTRODUCTION

The problem of multiplicity is encountered in randomized controlled clinical trials (RCT) and observational studies alike. This can arise from testing multiple hypotheses that reflect multiple primary outcome measures, multiple assessment time points, or multiple subgroups. The primary

[*] Correspondence to: Andrew C. Leon, Department of Psychiatry, Weill Cornell Medical College, Box 140, 525 East 68th Street, New York, NY 10021, U.S.A.

[†] E-mail: acleon@med.cornell.edu

Contract/grant sponsor: National Institute Health; contract/grant numbers: MH060447, MH068638, DA06534

Received 15 October 2006
Accepted 16 November 2006

concern is that multiple testing can elevate the risk of committing a Type I error. Specifically, the probability α_{FWE} of familywise Type I error (FWE) with k independent tests is as follows:

$$\alpha_{\mathrm{FWE}} = 1 - (1-\alpha)^k \tag{1}$$

where α is a pre-specified significance level for each individual null hypothesis, H_{0j}, $j = 1, \ldots, k$, i.e. $\alpha = P(\text{Reject } H_{0j} \mid H_{0j})$. For instance, the nominal α-level of 0.05 is clearly exceeded with $k = 4$ independent tests, $\alpha_{\mathrm{FWE}} = 1 - (1-0.05)^4 = 0.185$, where $\alpha_{\mathrm{FWE}} = P(\text{Reject } H_0 \mid H_0)$ in that H_0 is a 'global' null hypothesis such that H_0: H_{0j} is true for all j. Based on concerns about elevated Type I error, regulatory agencies [1], scientific journals and guidelines such as the CONSORT Statement [2] advocate the use of multiplicity adjustments.

Several such adjustments are based on the Bonferroni inequality, which in essence states that the probability of occurrence of at least one event among k events (E_j, $j = 1, \ldots, k$) is less than or equal to the sum of the probabilities of each of those individual events

$$p\left(\bigcup_{j=1}^{k} E_j\right) \leqslant \sum_{j=1}^{k} P(E_j)$$

The so-called Bonferroni adjustment (α_{B}) is based on this inequality. For instance for k tests, the Bonferroni adjustment partitions the nominal alpha-level (α) into k equal components and yields $\alpha_{\mathrm{B}} = \alpha/k$ such that $\alpha_{\mathrm{B}} = P(\text{Reject } H_{0j} \mid H_{0j})$ for $j = 1, \ldots, k$. Thus, for $k = 4$ tests and a nominal unadjusted $\alpha = 0.05$, $\alpha_{\mathrm{B}} = 0.0125$. As a result, the familywise error rate based on the Bonferroni adjustment becomes

$$\alpha_{\text{B-FWE}} = 1 - (1-\alpha_{\mathrm{B}})^4 = 1 - (1-0.0125)^4 = 0.049$$

In contrast, the Dunn–Šidák adjustment [3], $\alpha_{\mathrm{DS}} = 1 - (1-\alpha)^{1/k}$, yields a familywise error of precisely α. In the case of $k = 4$

$$\alpha_{\mathrm{DS}} = 1 - (1-0.05)^{1/4} = 0.01274$$

which results in the familywise error rate

$$\alpha_{\text{DS-FWE}} = 1 - (1-\alpha_{\mathrm{DS}})^4 = 1 - (1-0.01274)^4 = 0.05$$

Alternatively, sequentially rejective procedures modify the alpha level with each of k successive tests. Holm's approach [4], for instance, tests each H_0 in *ascending* order of p-values. Successively larger p-values have less rigorous alpha, α_{Holm}: $\alpha/k, \alpha/(k-1), \ldots, \alpha/(k-(k-1))$. For $k = 4$, the successive levels of α_{Holm} are: 0.0125, 0.0167, 0.025, and 0.05. The sequential Holm tests terminate after the first non-significant test and no subsequent H_{0j} is rejected.

Hochberg's approach [5], in contrast, sequences the k tests in *descending* order of p-values. Each successive H_{0j} has a more rigorous α-threshold. For $k = 4$, the successive α-levels are $\alpha/(k-3)$, $\alpha/(k-2)$, $\alpha/(k-1)$, and α/k, or 0.05, 0.025, 0.0167, and 0.0125. The sequential tests terminate after the first significant test and each subsequent H_{0j} is rejected.

Each approach provides protection in the case of a true H_{0j}. Yet, if the H_{0j} proves to be false, there is a corresponding reduction in statistical power. In this regard, the 'global' alternative hypothesis (H_{a}) states that at least one H_{0j} is false or that the global null hypothesis H_0 is false. Thus, the statistical power is the probability of rejecting at least one null hypothesis when the

DOI: 10.1002/sim

global null hypothesis H_0 is false. In this paper, statistical power is operationalized in this way to reflect a protocol that identifies k primary endpoints and specifies that the two groups will be deemed significantly different if at least one H_{0j} is rejected. In this context, with alpha thresholds that vary, the Holm and Hochberg procedures can each provide more statistical power than a strict application of the Bonferroni adjustment. The statistical power of strategies for multiple tests of group means has been evaluated [6–11]. Yet, none of the aforementioned approaches incorporates the correlations between endpoints. The resulting adjustments tend to overcompensate for multiplicity when the correlation among endpoints exceeds 0.50 [12].

Here, we consider a multiplicity adjustment that was originally proposed by James [11] to incorporate correlations between normally distributed endpoints. Elsewhere we have shown that with a true H_{0j} and highly correlated binary endpoints (i.e. $0.60 \leqslant \rho \leqslant 0.90$, where ρ is the correlation among endpoints in population), this approach maintains a steady nominal alpha level unlike those discussed above [13].

The objective of this manuscript is to examine the statistical power of three procedures (Bonferroni, Hochberg, and James) when used for two-tailed χ^2 tests that compare two groups on correlated binary endpoints. The performance of the three approaches is examined with simulated data. The Bonferroni adjustment, the most commonly used adjustment, serves as the benchmark. The Hochberg approach was chosen over the Holm approach because by definition, it has statistical power that is greater than or equal to that of Holm. That is, if Holm rejects then Hochberg rejects, but the converse is not necessarily true. Initially, the three approaches to multiplicity are applied to a medical examiner study of gender differences in circumstances of homicides in New York City. The performance of each method is then compared in a simulation study that examines statistical power.

2. JAMES PROCEDURE: ADJUSTMENT FOR CORRELATED ENDPOINTS

The James p-value adjustment [11] is based on the standard multivariate normal. An equal pairwise sample correlation r among k variables is assumed here. An alternative approach described by James [11] relaxes that assumption by using the Armitage and Parmar [14] approximation (Note that for the application, we use the sample correlation r; whereas in the simulations, which involve population parameters we use ρ.) Adapting the James' notation, an adjusted p-value, p_{adj}, is estimated for each observed unadjusted p-value, p, as follows for two-tailed tests:

$$p_{\text{adj}} = 1 - D_1(1-r^2) - D_2 r^2 - D_3 r(1-r) - D_4(2 - 2(1-r)^{1/2} - r - r^2) \tag{2}$$

where

$$D_1 = (1-p)^k, \quad D_2 = 1-p, \quad D_3 = 0$$

and

$$\begin{aligned} D_4 &= k(k-1)\phi(b)\int_{-\infty}^{\infty} \Phi(z)^{k-2}\phi(z)^2 \, \mathrm{d}z \\ &= k(k-1)\phi(b)G(k) \end{aligned}$$

DOI: 10.1002/sim

Here $b=\Phi^{-1}(1-p/2)$, ϕ is the probability density function and Φ is the cumulative probability function of a standard normal variable; Φ^{-1} is the inverse function of Φ. As applied, each H_{0j} is rejected that corresponds to $p_{\text{adj}}<0.05$. James presented values for $G(k)$ in the Appendix. For example: $G(2)=0.28209479$, $G(3)=0.14104740$, $G(4)=0.08578128$, $G(5)=0.05814822$.

3. APPLICATION

The Office of Chief Medical Examiner of New York City is responsible for investigating each death believed to be a homicide, suicide, or accident and instances of death unattended by a physician. Our study of the medical consequences of substance abuse reviewed the details about each homicide and suicide that was certified by the Office of Chief Medical Examiner of New York City in the 1990s [15–17]. Here, we consider gender differences in the circumstances of homicides in 1997. Location and method of homicide, and serum toxicology are examined. The analyses are limited to victims who had undergone systematic toxicologic analyses for cocaine, opiates, and ethanol. Based on the elimination half life of ethanol, analyses were restricted to victims who had lived less than two hours after injury.

Males and females ($X_i=1$ for males and $X_i=2$ for females) were compared on five binary endpoints, Y_j. These include place of injury (victim's home *versus* elsewhere), firearm death (yes/no), and three variables determined by toxicologic analyses: presence of ethanol, cocaine, and opiates. A χ^2 test was used for each of the respective H_{0j}: $\pi_{1j}=\pi_{2j}$, $j=1,\ldots,5$. Initially, an unadjusted two-tailed α-level of 0.05 was used for each χ^2 test. These results are presented in Table I. Four of the five unadjusted tests are statistically significant, indicating that there were gender differences in each circumstance of homicides except for the presence of opiates. Female victims were more likely to be murdered at home, whereas males were more likely to be murdered with a firearm, and were more likely to have cocaine and ethanol in their bodies at the time of autopsy. However, with 5 endpoints, the probability of FWE is 22.6 per cent (i.e. $1-(1-0.05)^5=0.226$), if the five binary variables are independent. Hence, an adjustment procedure is clearly indicated. The Bonferroni adjustment and two alternatives are also applied in Table I.

With the Bonferroni adjusted α-level, two null hypotheses are rejected: victim's home and firearms. In applying the Hochberg's approach, the $k=5$ tests were sequenced in *descending* order of p-values: opiates, ethanol, cocaine, firearms, and victim's home. The initial α-threshold is 0.05 and, thus, opiates are not found to differ across genders. However, with the Hochberg approach the sequential tests terminate after the first significant test, ethanol in this case, and all subsequent tests are statistically significant. The Hochberg approach detected the gender differences in victim's home, firearms, cocaine, and ethanol.

Finally, the unadjusted p-values were adjusted with the James procedure as described above. The mean correlation among the five endpoints, $r=0.11$, was incorporated in those calculations. The adjusted p-values for victim's home, firearms, and cocaine are each less than the alpha level 0.05 and thus each of those variables is statistically significant. It is clear that the choice of adjustment procedure has bearing on the results and interpretation of these data. In practice, the adjustment procedure must be identified prior to data analysis, preferably at the time of protocol development. The performance of these procedures is now examined in a simulation study.

DOI: 10.1002/sim

Table I. Male and female victims are compared on circumstances of their homicides: an examination of three multiplicity adjustment procedures for χ^2 tests of correlated binary endpoints.

						Adjustment procedure			
Variable	Female ($N = 45$) (per cent positive)	Male ($N = 386$) (per cent positive)	χ^2	df	Unadjusted p-value	Unadjusted α	Bonferroni α_B	Hochberg* α	James adjusted p-value†
Victim's home	73.3	28.2	37.10	1	<0.0001	0.05‡	0.01‡	‡	< 0.001*
Firearms	51.1	75.9	12.67	1	0.0004	0.05‡	0.01‡	‡	0.002*
Cocaine	4.4	20.2	6.62	1	0.0101	0.05‡	0.01	‡	0.049*
Ethanol	20.0	37.0	5.13	1	0.0235	0.05‡	0.01	0.0250‡	0.111
Opiates	6.7	5.4	0.12	1	0.7343	0.05	0.01	0.05	0.999

*Using Hochberg approach, the sequential tests terminate after the first significant test and all subsequent tests are deemed significant.

†Calculations of James-adjusted p-values incorporate the mean correlation among the 5 endpoints, $r = 0.1105$.

‡Indicates statistically significant with the alpha-level of the corresponding adjustment procedure.

DOI: 10.1002/sim

4. SIMULATION STUDY

Simulation specifications: The simulation study was designed to examine the performance of three multiplicity adjustment procedures under a range of conditions. In each case, a two-tailed χ^2 test is used to test each hypothesis H_{0j}: $\pi_{1j} = \pi_{2j}$. The global null hypothesis is: H_0: $\pi_{1j} = \pi_{2j}$ for all j. The global alternative hypothesis H_a is simply that H_0 is false. The following specifications varied across the simulations:

(1) Rate of the binary endpoints (e.g. response rates) for all j.

   a. $\pi_{1j} = 0.15, 0.35$.
   b. $\pi_{2j} = \pi_{1j} + 0.05, \pi_{1j} + 0.15, \pi_{1j} + 0.25$.

(2) Number of endpoints to be compared across groups (i.e. # tests), $k = 2, 3, 4$.
(3) Correlation among the binary endpoints $\rho = 0, 0.10, 0.20, \ldots, 0.90$.
(4) Adjustment procedure.

   a. Bonferroni adjusted alpha.
   b. Hochberg's approach (α adjusted as described above).
   c. James' p-value adjustment (2).

4.1. Generation of random correlated binary variables

The algorithm proposed by Park *et al.* [18] was used to generate multivariate binary variables with a compound symmetry correlation structure. Specifically, it was assumed that the rate parameter (or 'success' probability) π is the same for each of 4 binary variables and that the pairwise correlation ρ is constant over the 6 combinations of the pairs. Initially, a Poisson random variable W was generated with an intensity parameter $\lambda_1 = \log(1+\rho(1-\pi)\pi^{-1})$. Another Poisson random variable $X_i(X_i,\ i = 1, 2, 3, 4)$ was generated with an intensity parameter $\lambda_2 = \log \pi^{-1} - \lambda_1$ independently from each other and also from W. Finally, the correlated binary random variables Y_i were generated as $Y_i = 1(Z_i = 0)$, where $Z_i = W + X_i$ and $1(c)$ is an indicator function, which returns 1 if the condition 'c' is met, and 0 otherwise. In this way the random variables Y_i are correlated through the common addend W. This procedure was applied to each group so that the correlation of the binary outcomes equals the pre-specified ρ in each group. We then randomly selected the k binary variables when the number of variables under consideration is k. Specifically, we took a random vector $(Y_1, \ldots, Y_{k'})$ when $k = k'$. The correlated multivariate binary variables were generated independently between subjects. However, across the number of variables k, the multivariate binary variables are not independent. For example, random vectors (Y_1, Y_2, Y_3) for $k = 3$ and (Y_1, Y_2, Y_3, Y_4) for $k = 4$ would not be independent, but instead common elements are contained in the vectors for $k = 3$ and 4.

The sample size per group varied across the specified values of π_1 and π_2. It was based on the N required for statistical power of 0.80 with $\alpha = 0.05$ for $k = 1$ and estimated using the algorithm presented by Fleiss [19] for chi square tests without a continuity correction. The estimated sample sizes were each rounded up to assure that power for $k = 1$ was not less than 0.80. This approach was used because it reflects sample size determination strategies often used in clinical trial and observational study design.

DOI: 10.1002/sim

For each of 60 combinations of simulation specifications ($2\pi_1 \times 3\pi_2 \times 10\rho$), 10 000 data sets were generated. Two-tailed χ^2 tests were used to compare the two groups on each of k binary variables ($k = 2, 3, 4$). For each test, the unadjusted α-level was 0.05. The Bonferroni, Hochberg and James procedures were then each used to evaluate the statistical significance of each χ^2 test.

4.2. Evaluation of adjustment procedures

Statistical power was defined as the proportion of the 10 000 data sets per simulation specification in which at least one of the k null hypotheses was rejected. Power was operationalized in this way to reflect a protocol that identifies k primary endpoints and specifies that the two groups will be deemed significantly different if at least one H_0 is rejected. The power of the Hochberg and James procedures is presented relative to the power of the Bonferroni approach. The relative power ratio, $\text{Power}_{\text{James}}/\text{Power}_{\text{Bonferroni}}$, for example, quantifies the gain (>1) or loss (<1) in power of the two alternatives to the Bonferroni approach. SAS release 8.02 and S-PLUS 6.2 were used for all computations.

4.3. Simulation results

By design, or more specifically, the choice of sample size for each pair of π_1 and π_2, the absolute level of statistical power exceeded 0.80. In fact, power was inversely related to ρ, ranging from about 0.80 ($\rho = 0.90$) to 0.92 ($\rho = 0$). Although not shown here due to space limitations, this applies to for all simulation specifications except for the Bonferroni method with $\rho = 0.90$, where power was 0.77. The results of the simulation study focus on relative power and are presented in Figures 1(a)–(c) and 2(a)–(c) for group 1 binary endpoint rates (π_1) of 0.15 and 0.35, respectively. These results indicate that neither the James nor Hochberg approach sacrifices statistical power relative to the Bonferroni method for any specification examined. In fact, the Hochberg approach appears to be slightly advantageous with $\rho \leqslant 0.50$, particularly for $k = 4$. However, when ρ exceeds 0.50 in the simulations, the James procedure has relatively more statistical power and the phenomenon becomes more pronounced with increases in ρ. These trends are apparent for both $\pi_1 = 0.15$ and $\pi_1 = 0.35$. The pattern indicates that the James strategy, the one multiplicity adjustment examined here that incorporates the correlation (ρ) among endpoints in the adjustment, achieves its objective.

5. DISCUSSION

This simulation study examined the statistical power of three multiplicity adjustment procedures for χ^2 tests that compare two groups on multiple correlated binary endpoints. The Hochberg approach has more statistical power for $\rho \leqslant 0.50$; whereas, the James procedure provides more statistical power with higher ρ. The increased relative power of the James adjustment is more pronounced with a large group difference in proportions ($\pi_1 - \pi_2$) for relatively lower correlations ($\rho \cong 0.50$); whereas for a small group difference, the increased relative power is apparent with larger correlation ($\rho \cong 0.70$), as seen in Figures 2(c) and (a), respectively. This is consistent with theory in that the James' threshold is a function of the correlation and, thus, is expected to be sensitive to higher correlations. Heuristically, the increasing power of the James adjustment with increasing correlation is due to the fact that the James adjusted p-values lie in a continuum between the Dunn–Šidák adjusted p-value ($p_{\text{adj}} = 1 - (1 - p)^k$ when $\rho = 0$) and unadjusted p-value ($p_{\text{adj}} = p$ when $\rho = 1$). Relative to the Bonferroni approach, neither alternative sacrifices power.

DOI: 10.1002/sim

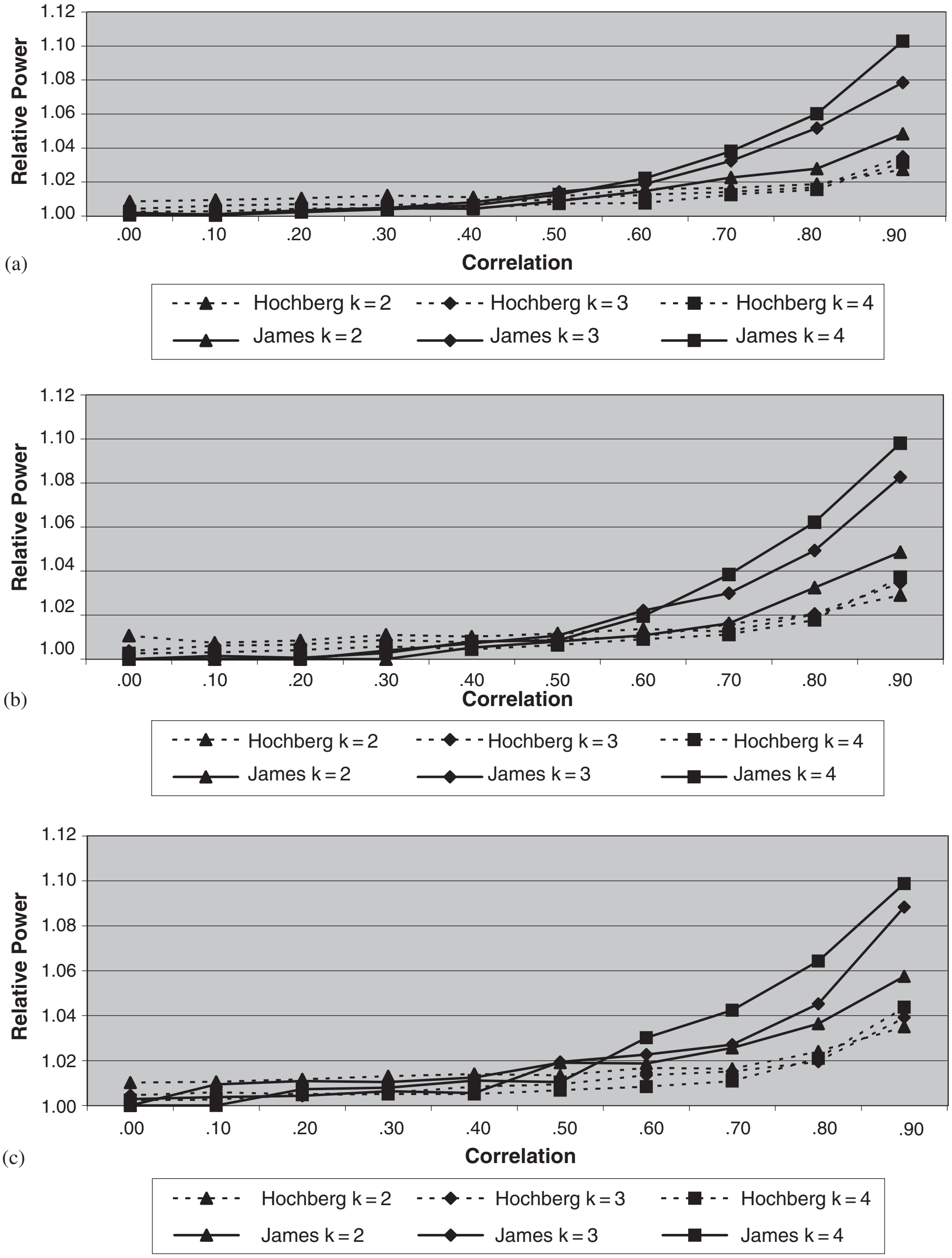

Figure 1. Statistical power of the James and Hochberg approaches relative to the Bonferroni method: (a) simulation study results for $\pi_1 = 0.15$ and $\pi_2 = 0.20$ with $N = 906$/group; (b) simulation study results for $\pi_1 = 0.15$ and $\pi_2 = 0.30$ with $N = 121$/group; and (c) simulation study results for $\pi_1 = 0.15$ and $\pi_2 = 0.40$ with $N = 49$/group.

DOI: 10.1002/sim

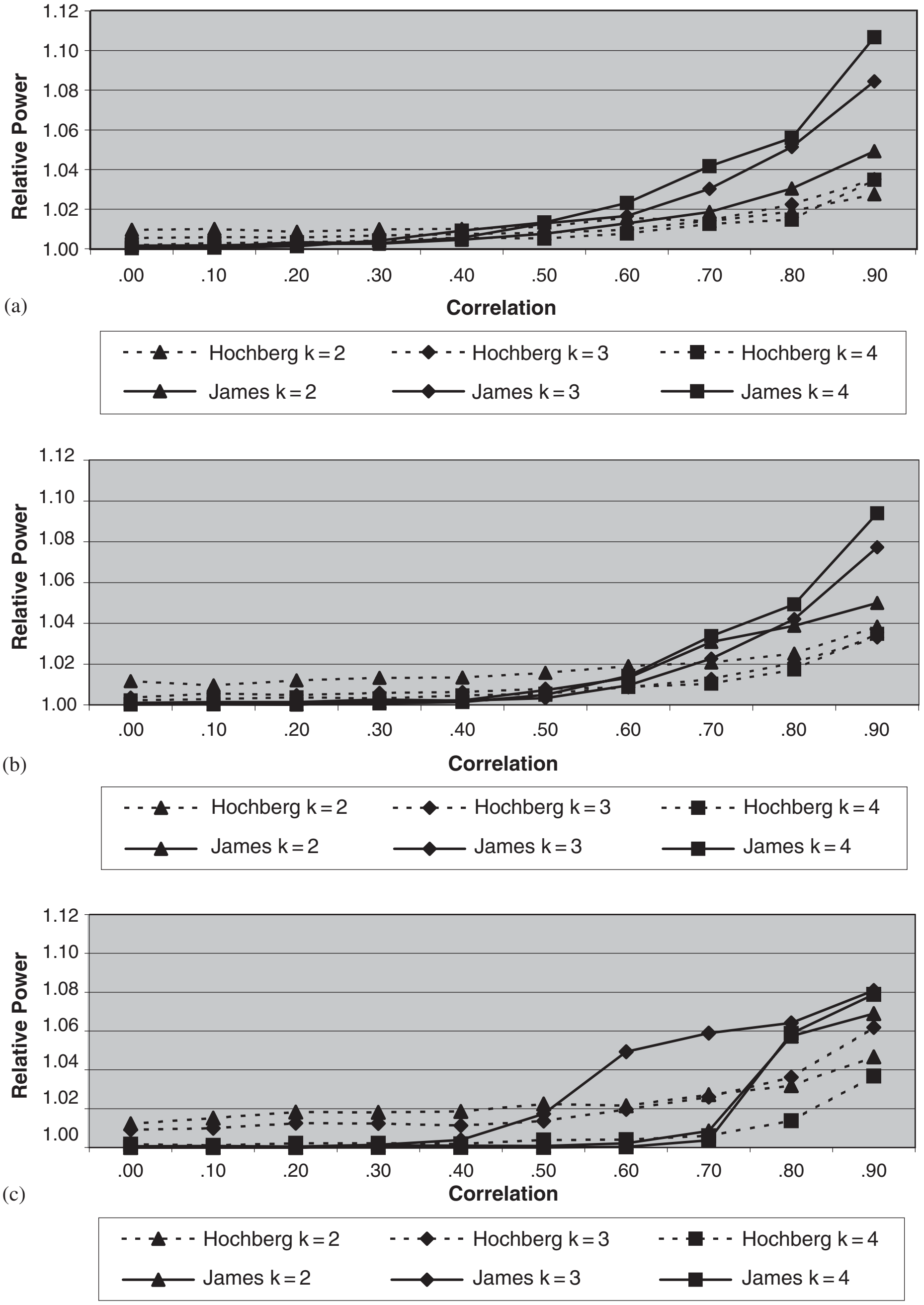

Figure 2. Statistical power of the James and Hochberg approaches relative to the Bonferroni method: (a) simulation study results for $\pi_1 = 0.35$ and $\pi_2 = 0.40$ with $N = 1471/\text{group}$; (b) simulation study results for $\pi_1 = 0.35$ and $\pi_2 = 0.50$ with $N = 170/\text{group}$; and (c) simulation study results for $\pi_1 = 0.35$ and $\pi_2 = 0.60$ with $N = 62/\text{group}$.

Each of these multiplicity adjustment procedures has been shown to protect against the inflation of familywise error seen in unadjusted analyses [8]. Elsewhere, we have shown that when the correlation, ρ, among endpoints exceeded 0.60, familywise error was unnecessarily conservative with the Bonferroni and Hochberg approaches, typically ranging from 0.02 to 0.04. The James procedure, in contrast, was less sensitive to ρ, not diverging from the nominal α-level, even with $\rho = 0.70$, 0.80, and 0.90 [13]. A theoretical approximation for an adjusted alpha level was presented by James [11, p. 1130].

The results of the three adjustment procedures varied with the homicide data, ranging from 2 to 4 statistically significant tests. Hochberg's sequentially rejective procedure detected four gender differences in the analyses of circumstances of homicides. The appeal of the James approach, of course, is that it accounts for the correlation among endpoints. However, with these data the correlation was only 0.11. Had it been higher, the results of the James approach might have been salient.

James [11] proposed her multiplicity adjustment to incorporate correlations between normally distributed endpoints. A sequentially rejective adaptation of the James procedure for continuous data was examined by Arani and Chen [10]. We have applied the James approach here for correlated binary endpoints, but did not use a sequentially rejective adaptation. Alternative strategies for binary data include a modified Bonferroni approach that Tarone [20] proposed for discrete data and applied to animal carcinogenicity studies. That approach did not include all endpoints in the multiplicity adjustment, but only those which occurred with at least some minimal prevalence. Bootstrap methods [21, 22] have also been proposed.

The choice of adjustment strategies in applied settings, of course, cannot be guided by the approach, among many, that yields the most favourable results. Instead one multiplicity adjustment procedure should be identified in the study protocol before the data are analysed, whether for multiple-dependent variables or multiple subgroups [23]. In some cases it would be reasonable to describe in the protocol an empirically based selection criteria, if that choice were based on the correlation among endpoints and not the magnitude or frequency of apparent groups differences. For example, based on the simulation results presented here, one could specify that the Hochberg approach will be used with observed mean $r \leqslant 0.50$, otherwise James approach will be used. Nonetheless, the relative power of the James procedure was acceptable across the range of correlations among endpoints. In addition to simulation results presented here, we also examined $\pi_1 = 0.05$ and 0.25. The pattern of relative power seen in those results was quite similar to that presented above.

There are noteworthy limitations inherent in the design of the simulation study and thus the simulation results that were presented. An equicorrelation structure among the binary endpoints was assumed and data were generated accordingly. This conforms to one of two approaches that James proposed in which the mean r is incorporated in the p-value adjustment. James also described an application of an algorithm that does not assume equicorrelation, but instead is based on the Armitage and Parmar [14] approximation. When this approximation was applied to the homicide data, the results were nearly identical to those reported in Table I that assumed equicorrelation. The plausibility of the equicorrelation assumption will vary across the substantive research contexts. It is unclear how the approaches would perform when observed data differ substantially from the specifications used in these simulations. For instance, we did not examine autocorrelation structures, unequal sample sizes, or multiple groups. Furthermore, multiplicity-adjusted sample sizes [24] were not used in the simulations, where sample size requirements increase with k. Instead the sample size was fixed for each pair of π_1 and π_2. Finally, we used only one definition

DOI: 10.1002/sim

of statistical power, the probability of rejecting at least one H_{0j}. In contrast, Arani and Chen [10] also examine probability of rejecting all k H_{0j}'s.

Several alternative multiplicity adjustment procedures were not evaluated here. It is not clear how permutation methods [25–27], bootstrap methods [21, 22], or Tarone's [20] modified Bonferroni approach would compare to those evaluated in our simulation study.

In conclusion, the James procedure is an appropriate choice to propose when designing a study that anticipates highly correlated endpoint insofar as detecting at least one false null hypothesis is concerned. In such situations, it affords appropriate Type I error protection in the case of the true global H_0 [13] and provides relatively more statistical power than the other procedures examined here. The approach can be applied to RCTs and observational studies alike.

ACKNOWLEDGEMENTS

The authors are grateful to Ken Tardiff, MD and Peter Marzuk, MD for providing the homicide data.

REFERENCES

1. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. *Guidance for Industry*: *E9 Statistical Principles for Clinical Trials*, 1998.
2. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T for the CONSORT Group. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Annals of Internal Medicine* 2001; **134**:663–694.
3. Ury HK. A comparison of four procedures for multiple comparisons among means pairwise contrasts; for arbitrary sample sizes. *Technometrics* 1976; **18**:89–97.
4. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.
5. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**:800–803.
6. Thomas DAH. Error rates in multiple comparisons among means—results of a simulation exercise. *Applied Statistics* 1974; **23**:284–294.
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B* 1995; **57**:289–300.
8. Morikawa T, Terao A, Iwasaki M. Power evaluation of various modified Bonferroni procedures by a Monte Carlo study. *Journal of Biopharmaceutical Statistics* 1996; **6**:343–359.
9. Brown BW, Russell K. Methods correcting for multiple testing: operating characteristics. *Statistics in Medicine* 1997; **16**:2511–2528.
10. Arani RB, Chen JJ. A power study of a sequential method of p-value adjustment for correlated continuous endpoints. *Journal of Biopharmaceutical Statistics* 1998; **8**:585–598.
11. James S. The approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials. *Statistics in Medicine* 1991; **10**:1123–1135.
12. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **43**:487–498.
13. Leon AC, Heo M. Comparison of multiplicity adjustment strategies for correlated binary endpoints. *Journal of Biopharmaceutical Statistics* 2005; **15**:839–855.
14. Armitage P, Parmar M. Some approaches to the problem of multiplicity in clinical trials. *Proceedings of the XIIIth International Biometric Conference*, Biometric Society, Seattle, 1986.
15. Tardiff KT, Marzuk PM, Leon AC, Hirsch CS, Stajic M, Portera L, Hartwell N. Homicide in New York City: cocaine use and firearms. *Journal of the American Medical Association* 1994; **272**:43–46.
16. Tardiff K, Marzuk PM, Leon AC, Hirsch CS, Stajic M, Portera L, Hartwell N. Cocaine, opiates, and ethanol in homicides in New York City: 1990 and 1991. *Journal of Forensic Sciences* 1995; **40**:387–390.
17. Marzuk PM, Tardiff K, Leon AC, Hirsch CS, Stajic M, Portera L, Hartwell N. Use of prescription psychotropic drugs among suicide victims in New York City. *American Journal of Psychiatry* 1995; **152**:1520–1522.

DOI: 10.1002/sim

18. Park CG, Park T, Shin DW. A simple method for generating correlated binary variates. *American Statistician* 1996; **50**:306–310.
19. Fleiss JL. *Statistical Methods for Rates and Proportions* (2nd edn). Wiley: New York, 1981.
20. Tarone RE. A modified Bonferroni method for discrete data. *Biometrics* 1990; **46**:515–522.
21. Westfall PH, Young SS. p value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association* 1989; **84**:780–786.
22. Westfall PH, Wolfinger RD. Multiple tests with discrete distributions. *American Statistician* 1997; **51**:3–8.
23. Lagakos SW. The challenge of subgroup analyses—reporting without distorting. *NEJM* 2006; **354**:1667–1669.
24. Leon AC. Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power when using the Bonferroni adjustment. *Journal of Clinical Psychiatry* 2004; **65**:1511–1514.
25. Brown CC, Fears TR. Exact significance levels for multiple binomial testing with application to carcinogenicity screens. *Biometrics* 1981; **37**:763–774.
26. Farrar DB, Crump KS. Exact statistical tests for any carcinogenic effect in animal bioassays. *Fundamental and Applied Toxicology* 1988; **11**:652–663.
27. Heyse JF, Rom D. Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity studies. *Biometrical Journal* 1988; **30**:883–896.

DOI: 10.1002/sim