International Journal of
Environmental Research and Public Health
ISSN 1661-7827
www.ijerph.org
© 2008 by MDPI

A Method to Compute Multiplicity Corrected Confidence Intervals for Odds Ratios and Other Relative Effect Estimates

Jimmy Thomas Efird[1*] and Susan Searles Nielsen[2]

[1]Division of Pediatric General and Thoracic Surgery, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, S.9.548 (MLC 7000), Cincinnati, Ohio 45229-3039, USA
[2]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Cancer Epidemiology Research Cooperative, POB 19024, 1100 Fairview Ave, North, MS M4-C308, Seattle, Washington 98109-1024, USA
[*]Correspondence to Dr. Jimmy Thomas Efird. E-mail: jimmy.efird@stanfordalumni.org

Received: 22 September 2008 / Accepted: 24 November 2008 / Published: 31 December 2008

Abstract: Epidemiological studies commonly test multiple null hypotheses. In some situations it may be appropriate to account for multiplicity using statistical methodology rather than simply interpreting results with greater caution as the number of comparisons increases. Given the one-to-one relationship that exists between confidence intervals and hypothesis tests, we derive a method based upon the Hochberg step-up procedure to obtain multiplicity corrected confidence intervals (CI) for odds ratios (OR) and by analogy for other relative effect estimates. In contrast to previously published methods that explicitly assume knowledge of P values, this method only requires that relative effect estimates and corresponding CI be known for each comparison to obtain multiplicity corrected CI.

Keywords: Multiplicity correction, Hochberg step-up procedure

Abbreviations: CI = confidence intervals; FWER = familywise error rate; HR = hazard ratios; LCI = lower confidence interval; OR = odds ratios; PFER = per family error rate; RR = relative risks; SE = standard error

Introduction

Testing the statistical significance of multiple null hypotheses is a routine practice in epidemiologic and other types of biomedical research. By chance, the probability of wrongly rejecting one or more null hypotheses increases in proportion to the number of comparisons tested [1]. This is referred to as "multiplicity bias."

Various methods have been presented in the literature for controlling the type I error in the context of multiple hypothesis testing. The classic Bonferroni inequality [2] provides a simple distribution-free method for multiplicity P value correction. Letting α_i denote the probability that hypothesis S_i is incorrect, the Bonferroni probability for the joint null hypothesis may be written as:

$$\mathrm{P}\left(\bigcap_{i=1}^{n} S_i\right) = \mathrm{P}\left[\left(\bigcup_{i=1}^{n} S_i^{C}\right)^{C}\right] \geq 1 - \sum_{i=1}^{n} \alpha_i . \qquad (1)$$

The Bonferroni method rejects the n set of null hypotheses if $p_i \leq \sum_{i=1}^{n} \alpha_i / n$ for at least one i, where p_i denotes the P value corresponding to the i^{th} null hypothesis. In the simple case, α is apportioned evenly among the tests. Although the family-wise (FWER) and per family (PFER) error rates are preserved at the α level of significance, the Bonferroni procedure is known to be conservative, especially for highly correlated test statistics (i.e., type I error probability is less than the nominal level of α). For example, in the case of a study of multiple genetic polymorphisms, the assumption is that all variants being tested have equal probability of being truly associated with the outcome of interest and leads to overcorrection.[3] The first order Bonferroni inequality may be improved upon given knowledge of the joint bivariate probabilities [2, 4, 5] or when the absolute value of the correlation coefficient is greater than 50% [2, 6]. However, these improvements have been limited in applied practice due to their restrictive nature. Several multiple testing procedures [7-9] based upon the "closure method"[10] and "Simes equality"[11] have been introduced and shown to be more powerful than the Bonferroni method for testing the intersection hypothesis [12-13]. Of the closure method based options, the Hochberg step-up multiple comparisons procedure [7] has gained popularity as being "easier to apply" than the more powerful procedures of Hommel [9] and Rom. The procedure also is uniformly more powerful than the Bonferroni-based, sequentially-rejective method of Holm [14] in many applied situations, e.g., when test statistics are uncorrelated, follow a multivariate normal or T^2 distribution, or are model independent [15-17]. Given an ordered set of P values, i.e., $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$, the Hochberg procedure rejects all hypothesis $H_{i \leq j}$ if $p_{(j)} < \alpha/(n-j+1)$ for any $j=1, \ldots, n$. P values are incrementally corrected in order from smallest to largest by multiplying $p_{(j)}$ by $(n-j+1)$, wherein the multiplicative factor for the largest P value is unity and thus remains the same after multiplicity correction.

Many researchers and journal editors increasingly recognize confidence intervals (CI) as the preferred measure for conveying statistical uncertainty of effect size estimates such as odds ratios (OR), relative risks (RR), and hazard ratios (HR), as P values have been commonly misunderstood and misinterpreted in the literature [18-22]. Similar to hypothesis testing by way of P values, CI also may be corrected for multiplicity to minimize the risk of making false-positive inferences. Several authors have provided techniques to correct CI for multiple hypothesis testing [23-26]. However, most of the methods are computationally intensive or mathematically complex, and more importantly, none provide a way to correct CI when corresponding P values are not provided for the individual hypothesis tests.

Below, we present a method to compute multiplicity corrected CI for OR and by analogy for other measures of relative risk, when no P values have been explicitly provided. This computationally simple method based upon the Hochberg step-up procedure only requires knowledge of individual test OR and CI, and the number of comparisons being tested.

Methodology

The derivation of multiplicity corrected confidence intervals for a set of n OR involves expressing the standard error (SE) for the logarithm of OR_i (i=1 to n) in terms of the lower confidence interval (LCI) for OR_i. Letting:

$$LCI\left(OR_i\right) = e^{\left\{\log\left(OR_i\right) - z(1-\alpha/2) \cdot SE\left[\log\left(OR_i\right)\right]\right\}}, \qquad (2)$$

where $z_{(1-\alpha/2)}$ is the 100% x (1-α/2) percentile of a standard normal distribution, and solving for $SE[\log(OR_i)]$ we see that:

$$SE\left[\log\left(OR_i\right)\right] = -\left\{\log\left[LCI\left(OR_i\right)\right] - \log\left(OR_i\right)\right\} / z(1-\alpha/2). \qquad (3)$$

Substituting the right hand side of (3) into the equation for the 2-tailed z test statistic gives:

$$z_i = \left|\frac{\log\left(OR_i\right)}{SE\left[\log\left(OR_i\right)\right]}\right| = \left|\frac{\log\left(OR_i\right)}{-\left\{\log\left[LCI\left(OR_i\right)\right] - \log\left(OR_i\right)\right\} / z(1-\alpha/2)}\right|. \qquad (4)$$

The corresponding P value is computed as:

$$p_i = 2 \cdot \left[1 - \Phi\left(z_i\right)\right], \qquad (5)$$

Where:

$$\Phi\left(z_i\right) = \int_{-\infty}^{z_i} \frac{e^{-x^2/2}}{\sqrt{2 \cdot \pi}} dx. \qquad (6)$$

Ordering the P values (p_i's) from the lowest to highest values i.e., $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$ (with arbitrary ordering in the case of ties), the Hochberg multiplicity corrected P values denoted by "*" are computed as:

$$p^*_{(j)} = p_{(j)} \cdot \left(n - j + 1\right), \qquad (7)$$

where j ranges from 1 to n in a 1:1 identity mapping with the i values and $p^*_{(j)}$ is bounded by unity. Rearranging (5) and solving for $SE^*\left[\log\left(OR_i\right)\right]$ in the equation:

$$\Phi\left[\frac{\log\left(OR_i\right)}{SE^*\left[\log\left(OR_i\right)\right]}\right] + \left(\frac{p^*_{(j)}}{2}\right) - 1 = 0 \qquad (8)$$

gives the Hochberg corrected standard error for the logarithm of OR_i, i.e.:

$$SE^*\left[\log\left(OR_i\right)\right] = \frac{\log\left(OR_i\right)}{\Phi^{-1}\left(1 - \frac{p^*_{(j)}}{2}\right)}. \qquad (9)$$

The multiplicity corrected (1-α/2) x 100% CI for $OR_{(i)}$ based upon the Hochberg step-up procedure can then be computed by substituting the above standard error from eq. 9 into the following basic equation:

$$CI_{(1-\alpha/2)} = e^{\left\{\log\left(OR_i\right) \pm z(1-\alpha/2) \cdot SE^*\left[\log\left(OR_i\right)\right]\right\}}. \qquad (10)$$

By analogy, replacing OR in the above equations with other relative effect estimates such as RR or HR gives the corresponding multiplicity corrected CI for these measures. When P values are directly available for the individual hypothesis tests, the Hochberg multiplicity corrected CI may be computed directly beginning with eq.

7. Furthermore, if the hypothesis test is 1-sided, then α must be multiplied by 2 in the above equations.

Example

Table 1 below presents OR from a case-control study for a hypothetical disease (D) and exposure to 3 dichotomously coded environmental risk factors. The OR and 95% CI for (D) uncorrected for multiplicity (*n*=3 factors) are shown in Columns 3 and 4: Factor 1 (OR=1.652, 95% CI=0.551-4.953); Factor 2 (OR=1.151, 95% CI=0.142-9.324); and Factor 3 (OR=6.509, 95% CI=1.646-25.743). Applying equations 7, 9 and 10 gives the corresponding multiplicity corrected P values (0.740, 0.895, 0.024; not shown in table), standard errors for the logarithm of the OR (1.513, 1.068, 0.830), and 95% CI (0.09-32, 0.14-9.3, 1.3-33). The multiplicity corrected CI for Factor 1 and Factor 3 are considerably wider than the corresponding uncorrected intervals, thus indicating a greater degree of variability for the estimated OR. In the case of Factor 2, the uncorrected and corrected CI is the same since Factor 2 had the highest P value of the 3 comparisons when applying the Hochberg algorithm.

In this example, the conclusions regarding the association (or lack thereof) of (D) and the exposure do not substantively change after correction for multiplicity, thus lending weight to what otherwise might be only cautious interpretation referencing the possibility of a chance observation due to multiple comparisons. However, in other situations where CI is close to containing unity, a null hypothesis might no longer be rejected at least in strict statistical terms after correction for multiplicity.

Discussion

Confidence intervals for OR, RR and other relative effect estimates are commonly reported in epidemiologic and public health literature without correction for multiple hypothesis testing. The failure to account for multiplicity may lead to inflation of type I error and over interpretation of any apparently "positive" findings. In the current paper, we show how CI for relative effect size estimates such as OR may be corrected for multiplicity by use of the Hochberg step-up procedure, a "closed-testing" method for protecting against making excessive false-positive inferences due to multiple comparisons.

Our method has several strengths. The corrected CI are simple to compute in standard statistical software packages that have function routines for determining percentiles and areas under a curve for a normal distribution. Since P values are not required for the original hypothesis tests, multiplicity corrected CI may be computed *post hoc* (when estimates are reported with sufficient precision) from publications that only report values for effect size estimates and corresponding CI. When the test statistics are uncorrelated, the family-wise type I error probability is theoretically guaranteed by the Hochberg step-up procedure. Simulation results also show that the Hochberg step-up procedure holds for many commonly encountered dependent test statistics [27].

Table 1: Odds ratios (OR) and 95% confidence intervals (CI) for a hypothetical disease (D) and exposure to 3 dichotomously coded environmental risk factors, uncorrected and corrected for multiplicity

Variable	*Cases/Control*	*Odds Ratio*[a]	*Uncorrected for Multiplicity*	*Corrected for Multiplicity*[b]	
			95% CI (OR)	*SE*[*] *[log(OR)]*	*95% CI*[*] *(OR)*
Factor 1					
Non-Exposed	587 / 2143	1.0	Referent	1.513	Referent
Exposed	5 / 10	1.652	[0.551-4.953]		[0.09-32]
Factor 2					
Non-Exposed	246 / 2143	1.0	Referent	1.068	Referent
Exposed	1 / 10	1.151	[0.142-9.324]		[0.14-9.3][c]
Factor 3					
Non-Exposed	141 / 2143	1.0	Referent	0.830	Referent
Exposed	3 / 10	6.509	[1.646-25.743]		[1.3-33]

[a] Adjusted for age and sex.

[b] Using Hochberg step-up procedure.

[c] Note: The multiplicity adjusted and unadjusted 95% CI will be equal in this case since the corresponding unadjusted P value for the Factor 2 comparison was the highest of the 3 comparisons and thus the multiplicative factor for $p_{(j)}$ in equation (7) will be equal to 1.

[*] Multiplicity adjusted estimates.

Several limitations must be observed when applying our procedure for computing CI. The technique is not applicable when "exact sampling distribution" methods have been used to make statistical inferences. The Hochberg multiplicity correction also will inflate P values and related CI when one or more of the hypothesis tests involve a multi-level, logically related categorical variable (e.g., current smoker, former smoker, never smoker). In this case, it is unnecessary to correct CI for multiplicity for a logically related variable in multivariate space. The computed multiplicity corrected CI will be an approximate solution when the decimal accuracy is limited for the original OR and CI values. Accordingly, it is generally recommended that at least 2 or 3 significant digits of accuracy are available for published estimates when using this method in a *post hoc* manner to compute multiplicity corrected confidence intervals. Additionally, the rule for computing $\overset{*}{p}_{(j)}$ (eq. 7) in rare cases may lead to an anomaly wherein $\overset{*}{p}_{(j)}$ but not $\overset{*}{p}_{(j-1)}$ will achieve statistical significance. In this situation, one might apply the *de facto* variation of multiplying $p_{(j)}$ and lesser ranked P values by *j* to obtain the corresponding Hochberg corrected P values.[28] And finally, the method should not be used if the logarithm of the effect estimate does not follow a normal distribution, or if the underlying observations are not independent and identically distributed.

It also is important to note that correction for multiplicity may not be necessary or even desirable in some situations [29-33]. For example, correction for multiplicity may be unnecessary when an *a priori* biologic mechanism of action exists for an independent variable that manifests a linear dose response in relationship to the outcome variable. Similarly, multiplicity correction may not be desirable when attempting to control type II errors as the latter will be inflated by virtue of decreasing type I errors [31]. Furthermore, multiplicity correction based upon the "universal null hypothesis," which tests that two groups are identical for all comparisons between variables, fails to take into account which and how many variables differ if the joint hypothesis is rejected [31]. Methods to correct for multiplicity also do not account for the inclusion of hypotheses that are biologically improbable or otherwise indefensible, which unnecessarily inflate the probability of incorrectly rejecting the joint null hypotheses [18, 29]. Philosophically, some researchers believe that the "primary" purpose for CI are to indicate a range of parameter values consistent with the data rather than for *de facto* hypothesis testing based on whether or not they include 1.0. Another salient concern regarding the appropriateness of multiplicity correction techniques is "how does one choose the universe for the number of comparisons." Clearly, multiplicity adjustment remains a debated topic with diverse opinions presented in the literature [34-35].

In the early days of the development of stepwise and closed tests for the control of type I error in multiple hypothesis testing, epidemiologists and statisticians commonly believed that joint CI could not be constructed for these procedures. However, it has been shown since that standard methods for constructing CI also readily apply to common stepwise multiplicity procedures.[23-24] Here, we have expanded on the seminal work of these researchers to develop a simple method for computing multiplicity corrected CI for standard estimates of effect size. Although our derivation has focused on the case of binary predictor variables, it is possible that similar principles might be developed and applied to obtain joint confidence sets in the more complex case of multilevel categorical variables.

Conclusions

Although the most effective strategy to minimize type I error related to multiple comparisons is to simply reduce the number of comparisons, this in effect penalizes the researcher for conducting a more informative multivariable study [32]. Statistical correction for multiple comparisons is not a substitution for the parsimonious and epidemiologically prudent selection - during the design phase of a study - of hypotheses to test. Nor should it be used in lieu of careful and informed interpretation of the results, taking into account biological plausibility (or lack thereof) and the results of prior studies. However, when statistical correction for multiple comparisons is appropriate, as is the case in many but not all situations, the method we present may have application as a supportive measure. A key advantage of this method is its correspondence with CI, which are typically more informative, and potentially more readily available, than P values.

Acknowledgements: This manuscript was made possible by a grant from NIH (1U01DK072493-01) entitled "Adolescent Bariatrics – Assessing Health Benefits and Risks." Amy Lehman of the Biostatistics Center at The Ohio State University offered valuable comments during the writing of this manuscript and her knowledge and insight was greatly appreciated. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of any funding agency.

References:

1. Hsu, J.: Multiple Comparisons: Theory and Methods. Boca Raton: Chapman & Hall/CRC, **1996**.
2. Hochberg, Y.; Tamhane, A.: *Multiple Comparison Procedures.* New York: Wiley, **1987**.
3. Campbell, H.; Rudal, I.: Interpretation of genetic association studies in complex disease. *Pharmacogenomics J,* **2002**; *2*:349-360.
4. Kounias E.: Bounds for the probability of a union of events, with applications. *Ann Math Statist,* **1968**, *39*:2154-2158.

5. Stoline M.: The Hunter method of simultaneous inference and its recommended use for applications having large known correlation structures. *JASA* **1983**, *78*:366-370.
6. Hunter D.: An upper bound for the probability of a union. *J Appl Prob* 1976, *13*:597-603.
7. Hochberg Y.: A sharper Bonferroni procedure for multiple tests of significance. *Biometrika,* **1988**; *75:*800-802.
8. Rom D.: A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **1990**, *77*:663-665.
9. Hommel G.: A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika,* **1988**, *75*:383-386.
10. Marcus, R.; Peritz, E.; Gabriel, K.: On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **1976**; *63*:655-660.
11. Simes, R.: An improved Bonferroni procedure for multiple tests of significance. *Biometrika,* **1986**, *73:751*-754.
12. Denne, J.; Koch, G.: A sequential procedure for studies comparing multiple doses. *Pharmaceut Statist,* **2002**, *1*:107-118.
13. Liu, W.: Multiple tests of a non-hierarchical finite family of hypotheses. *J R Statist Soc B,* **1996**, *58*:455-461.
14. Holm, S.: A simple sequentially rejective multiple test procedure. *Scand J Stat*, **1979**, *6:* 65-70.
15. Sarkar, S.: Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture. *Ann Statist,* **1998**, *26*:494-504.
16. Sarkar, K.; Chang, C.: The Simes method for multiple hypothesis testing with positively dependent test statistics. *JASA,* **1997**, *92*:1601-1608.
17. Huang, Y; Hsu, J.: Hochberg's step-up method: cutting corners off Holm's step-down method. *Biometrika,* **2007***,* *94*:965-975.
18. Walker, A.: Reporting the results of epidemiologic studies. *Am J Public Health,* **1986**, *76*:556-668.
19. Cooper, R.; Wears, R.; Schriger, D.: Reporting research results: recommendations for improving communication. *Ann Emerg Med,* **2003**; *41:*561-564.
20. Connor J.: The value of a p-valueless paper. *Am J Gastroenterol,* **2004**, *99:*1638-1640.
21. Altman D.: Why we need confidence intervals. *World J Surg,* **2005**, *29*: 554-556.
22. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *Med Educ,* **1999**, *33*:66-78.
23. Stefannsson, G.; Kim, W.; Hsu J.: On confidence sets in multiple comparisons. In SS Gupta, J O Berger, editors, *Statistical Decision Theory and Related Topics IV,* Springer-Verlag, *New York*, **1988**; *2*:89-104.
24. Hayter, A.; Hsu, J.: On the relationship between stepwise decision procedures and confidence sets. *JASA,* **1994**; *89*: 128-136.
25. Holm, S.: Multiple confidence sets based on stagewise tests. *JASA,* **1999**; *94*: 489-495.
26. Ludbrook, J.: Multiple inferences using confidence intervals. *Clin Exp Pharmacol P,* **2000**, *27*:212-215.
27. Shaffer, J.: Multiple hypothesis testing. *Annu Rev Psychol*, **1995**, *46*:561-584.
28. Rochon, J.: Update on repeated measurements (Short course notes - Roche Global Development, Nutley, NJ, **1996**).
29. Perry, J.: Multiple-comparison procedures: A dissenting view. *J Econ Entomol,* **1986**, *79*:1149-1155.
30. Rothman, K.: No adjustments are needed for multiple comparisons. *Epidemiology,* **1990**, *1:*43-46.
31. Perneger, T.: What's wrong with Bonferroni adjustments? *BMJ,* **1998**, *316*: 1236-1238.
32. Berry, D.; Hochberg, Y.: Bayesian perspectives on multiple comparisons. *J Statistical Plann Inference* **1999**, *82*:215-227.
33. Savitz, D.; Olshan, A.: Multiple comparisons and related issues in the interpretation of epidemiologic data. *Am J Epidemiol,* **1995**, *142*:904-908.
34. Bender, R.; Lange, S.: Multiple test procedures other than Bonferroni's deserve wider use. *BMJ* **1999**, *318*:600.
35. Aickin, M.: Other method for adjustment of multiple testing exists. *BMJ,* **1999**, *318*:127.