high-risk Black women[26,27] could help to reduce the Black–White disparity in infant mortality rates. □

References

1. Yankauer A. The relationship of fetal and infant mortality to residential segregation. *Am Sociol Rev.* 1950;15:644–648.
2. LaVeist TA. Linking residential segregation to the infant-mortality race disparity in U.S. cities. *Social Soc Res.* 1989;73:90–94.
3. Polednak AP. Black-White differences in infant mortality in 38 standard metropolitan statistical areas. *Am J Public Health.* 1991;81:1480–1482.
4. Massey DS, Denton NA. *American Apartheid: Segregation and the Making of the Underclass.* Cambridge, Mass: Harvard University Press; 1993.
5. Jargowsky PA, Bane MJ. Ghetto poverty in the United States. In: Jencks C, Peterson PE, eds. *The Urban Underclass.* Washington, DC: The Brookings Institution; 1991: 235–273.
6. Farley R, Frey WH. *Changes in the Segregation of Whites from Blacks during the 1980's: Small Steps toward a Racially Integrated Society.* Ann Arbor, Mich: Population Studies Center, University of Michigan; 1992. Research report 92–257.
7. *Vital Statistics of the United States. Mortality, Part B.* Hyattsville, Md: US Dept of Health and Human Services; 1982–1990.
8. *Vital Statistics of the United States. Natality.* Hyattsville, Md: US Dept of Health and Human Services; 1982–1990.
9. National Center for Health Statistics. *Health United States, 1992 and Healthy People 2000 Review.* Hyattsville, Md: Public Health Service; 1993.
10. National Center for Health Statistics. *Health United States, 1993.* Hyattsville, Md: Public Health Service; 1994.
11. Collins JW, David RJ. Race and birthweight in biracial infants. *Am J Public Health.* 1993;83:1125–1129.
12. Haenszel W, Loveland DB, Sirken MG. Lung cancer mortality as related to residence and smoking histories. *JNCI.* 1962; 28:1000–1001.
13. Shiono PH, Klebanoff MA, Graubard BI, et al. Birth weight among women of different ethnic groups. *JAMA.* 1986;255: 48–52.
14. Lieberman E, Ryan KJ, Monson RR, Shoenbaum SC. Risk factors accounting for racial differences in the rate of premature birth. *N Engl J Med.* 1987;317:743–748.
15. Rowley DL, Tosteson H. *Racial Differences in Preterm Delivery.* New York, NY: Oxford University Press Inc; 1993.
16. Afifi AA, Clark V. *Computer-Aided Multivariate Analysis.* Belmont, Calif: Lifetime Learning Publications; 1984.
17. Wilkinson L. *SYSTAT: The System for Statistics.* Evanston, Ill: SYSTAT Inc; 1990.
18. Gates-Williams J, Jackson MN, Jenkins-Monroe V, Williams LR. The business of preventing African-American infant mortality. *West J Med.* 1992;157:350–356.
19. Plough A, Olafson F. Implementing the Boston Healthy Start initiative: a case study of community empowerment and public health. *Health Educ Q.* 1994;21:221–234.
20. LaVeist TA. The political empowerment and health status of African-Americans: mapping a new territory. *Am J Sociol.* 1992;97:1080–1095.
21. Centers for Disease Control. Differences in infant mortality between blacks and whites—United States, 1980–1991. *MMWR Morb Mortal Wkly Rep.* 1994;43:288–289.
22. Mayer SE, Jencks C. Growing up in poor neighborhoods: how much does it matter? *Science.* 1989;243:1441–1445.
23. Jencks C, Peterson PE. *The Urban Underclass.* Washington, DC: The Brookings Institution; 1991.
24. Rawlings JS, Rawlings VB, Read JA. Prevention of low birth weight and preterm delivery in relation to interval between pregnancies among white and black women. *N Engl J Med.* 1995;332:69–74.
25. Yankauer A. What infant mortality tells us. *Am J Public Health.* 1990;80:653–654.
26. Rafferty MP. The effects of WIC and Medicaid participation on pregnancy outcome. In: *Proceedings of the 1991 Public Health Conference on Health and Statistics.* Washington, DC: US Dept of Health and Human Services; 1991:162–167. DHHS publication PHS 92-1214.
27. Edwards CH, Knight EM, Johnson AA, et al. Multiple factors as mediators of the reduced incidence of low birth weight in an urban clinic population. *J Nutr.* 1994;124: 927S–935S.

Adjusting for Multiple Testing When Reporting Research Results: The Bonferroni vs Holm Methods

Mikel Aickin, PhD, and Helen Gensler, PhD

ABSTRACT

Public health researchers are sometimes required to make adjustments for multiple testing in reporting their results, which reduces the apparent significance of effects and thus reduces statistical power. The Bonferroni procedure is the most widely recommended way of doing this, but another procedure, that of Holm, is uniformly better. Researchers may have neglected Holm's procedure because it has been framed in terms of hypothesis test rejection rather than in terms of *P* values. An adjustment to *P* values based on Holm's method is presented in order to promote the method's use in public health research. (*Am J Public Health.* 1996;86:726–728)

Introduction

It is well recognized that when one tests multiple hypotheses, all bearing on a single issue, the individual *P* values of the tests may not be an appropriate guide to actual statistical significance. Public health examples of this problem occur quite frequently. One is the attempt to characterize a new, ill-defined disease such as "sick building syndrome." If the investigator tabulates a long list of symptoms that might differentiate cases from controls, even if none of the symptoms are in fact related to the disease, some of the *P* values may fall below the customary .05 cutoff point. The argument advanced for adjusting the *P* values is that, without adjustment, the probability of declaring that some symptom is related to disease can be far higher than the nominal .05 level when none of the symptoms are actually related.

Another class of examples consists of assessing the effects of an intervention, such as a smoking cessation program, in different subpopulations determined by gender, age, social class, smoking inten-

The authors are with the Arizona Cancer Center, University of Arizona, Tucson.

Requests for reprints should be sent to Mikel Aickin, PhD, Biometry Program, Arizona Cancer Center, 1515 N Campbell, Tucson, AZ 85724.

This paper was accepted November 1, 1995.

Editor's Note. See related annotation by Levin (p 628) in this issue.

sity, and smoking duration. Even if the program is ineffective in all groups, the multiplicity of tests may lead to some groups showing nominal effects. Yet another category of multiple testing situations involves the fitting of linear models (such as logistic regression), in which a nominal P value of less than .05 for an individual coefficient may need to be interpreted in light of the fact that it is implicitly embedded in a series of significance statements about the other coefficients in the model.

Although the Bonferroni procedure is widely recommended as a general method of adjustment, a more powerful procedure has been known to biostatisticians for nearly 16 years. This method is virtually unknown among practitioners, and so the intent of this paper is to point out how simple adjusted P values that are always better than those adjusted by the Bonferroni method can be computed.

Methods

The Bonferroni procedure can be described very simply. When the tests involve null hypotheses H_i ($i = 1, \ldots n$), in order to maintain an overall type I error bound of α on all of them simultaneously, each of the corresponding P values P_i is compared with α/n instead of α. The argument runs as follows. Assuming that t of the n hypotheses are true, a type I error can occur only if one of the events $P_i \leq \alpha/n$ occurs for one of the true hypotheses. Since the Bonferroni inequality states that the probability of a union of events is less than or equal to the sum of the events' individual probabilities, the probability that any event $P_i \leq \alpha/n$ occurs (for a true hypothesis) is not greater than $t\alpha/n$, which is less than or equal to α.

The Bonferroni testing procedure is equivalent to an adjustment that replaces each P_i with nP_i (or 1, whichever is smaller) and compares these adjusted values with α. The values nP_i can be considered "Bonferronied" P values, in the sense that nP_i is the smallest overall significance level at which the individual hypotheses H_i would be rejected.

Holm[1] provided a method that applies in the same cases as the Bonferroni procedure but is uniformly more powerful. His method is accomplished as follows. First, the P_i values are placed in increasing order. For the purpose of exposition, one can resubscript them so that they are already in increasing order; P_1 is the smallest, and P_n is the largest. Second, each P_i is compared with $\alpha/(n - i + 1)$ for rejection. Third, starting with the smallest P value, one continues applying these comparisons (from $i = 1$ and proceeding in order) until the first nonrejection. Thus, the rejected hypotheses H_i are those for which $P_j \leq \alpha/(n - j + 1)$ for all $j \leq i$. Because the divisors are $n - j + 1$ instead of n, Holm's procedure never rejects fewer hypotheses than the Bonferroni procedure does.

That Holm's procedure also bounds the type I error at α can be seen as follows. Assume that t of the hypotheses are true. Let P denote the minimum of the P values of these true hypotheses. A type I error occurs only if P first causes rejection at some stage i, which entails $P \leq \alpha/(n - i + 1)$. However, since this is the first rejection caused by a true hypothesis, there must be at least $i - 1$ false hypotheses (which had to cause rejections at earlier stages than i), and so $i - 1 \leq n - t$. Thus, $\alpha/(n - i + 1) \leq \alpha/t$, and so a type I error implies that $P \leq \alpha/t$. The Bonferroni inequality applied to P now proves the result.

From the description of Holm's procedure, it follows immediately that if one defines $q_i = \max (n - j + 1) P_j$ or 1, whichever is smaller ($1 \leq j \leq i$), then Holm's method rejects those and only those H_i for which $q_i \leq \alpha$. Thus, one can state that each q_i is the "Holmed" P value for its corresponding hypothesis, in the sense that it is the smallest overall significance level at which the individual H_i would be rejected.

TABLE 1—Bonferroni- and Holm-Adjusted *P* Values in a Comparison of Skin Levels of Vitamin E

Comparison	Mean Difference	Bonferroni-Adjusted *P*	Holm-Adjusted *P*
(UVB + E) − C	18.408	.001	.001
UVB − C	9.368	.070	.047
(UVB + E) − UVB	9.040	.082	.047

Note. UVB = ultraviolet light; E = vitamin E supplementation; C = control.

Results

A practical example is shown in Table 1. The setting was a preclinical study of potential agents for chemoprevention of skin cancer. The design specified three groups of mice to be used in studying the effect of exposure to ultraviolet light and vitamin E supplementation on skin levels of vitamin E. The groups were as follows: (1) control (no ultraviolet light and no vitamin E), (2) ultraviolet light but no vitamin E, (3) and ultraviolet light and vitamin E. The Bonferronied P values indicated that the ultraviolet light and vitamin E group had higher levels than the control group, but the other two comparisons would not have met the overall .05 level criterion. Conversely, the Holmed P values indicate that all three comparisons are significant at the overall .05 level.

This example illustrates several points. The first arises because this research was part of a program investigating the preventive (or potentially harmful) effects of α-tocopherol conjugates in commercially marketed sunscreens, which might have substantial implications for the prevention of skin cancers. It is important to substantiate basic biochemical mechanisms in animals before moving on to human studies, and yet funding restrictions often compel the use of a smaller number of animals (five per group in this case) than one might like. It is therefore extremely important to maximize statistical power, and in this regard Holm's procedure is uniformly better than Bonferroni's.

The second point is that it is traditional to approach data such as these using analysis of variance. However, in the traditional F test, only the null hypothesis of no differences among means is tested. The F test gives no guidance concerning which groups are different when the F statistic is significant and provides little power to detect a small number of differences when most means coincide. For this reason, if individual group differences are to be interpreted, there is no reason to perform the analysis of variance; it is better to proceed directly to Holm's procedure.

Discussion

There is substantial debate in the biostatistical and epidemiologic literature concerning when (if ever) adjustment for multiple testing is warranted. It is not our

aim to contribute to this discussion; rather, we want to emphasize that once one has made the decision to adjust, one is obligated to use the most powerful method available.

The historical reluctance to use Holm's method may have several bases. For one, Holm's proof of the correctness of his method is considerably more advanced than the simple textbook argument that can be used to demonstrate the Bonferroni procedure. To address this, we have provided an argument that should be comprehensible to anyone familiar with the elementary facts about probabilities. However, another reason for the reluctance may be that many researchers understandably want to present *P* values rather than simplistic reject/confirm decisions. As we have shown, Holmed *P* values are easy to compute. Consequently, there does not appear to be any valid reason to continue using the Bonferroni procedure. □

Acknowledgments

This research was supported under a grant from the National Cancer Institute (CA25702).

We would like to thank Bruce Levin for comments that materially improved our presentation.

Reference

1. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Statistics.* 1979;6:65–70.

Tuberculosis Surveillance in the United States: Case Definitions Used by State Health Departments

Scott B. McCombs, MPH, Ida M. Onorato, MD, Eugene McCray, MD, and Kenneth G. Castro, MD

ABSTRACT

Health departments in all 53 reporting areas in the United States were asked to submit the case definition they used for tuberculosis surveillance. Sixteen areas used the 1990 case definition; two areas sent 1977 guidelines; and 34 areas sent other definitions. Case reports sent to the Centers for Disease Control and Prevention (CDC) in 1992 were analyzed; 4% of cases did not meet the 1990 definition. Tuberculosis case reporting criteria are not uniformly applied in the United States. CDC, in collaboration with state and local health officials, is evaluating the current definition and will implement uniform national criteria for tuberculosis surveillance. (*Am J Public Health.* 1996;86:728–731)

Introduction

National reporting of tuberculosis began in 1953. After decades of decline, the number of tuberculosis cases reported in the United States increased 14% between 1985 and 1993, from 22 201 to 25 287.[1] On review of recent national surveillance data, we found that criteria used to verify cases for reporting to the Centers for Disease Control and Prevention (CDC) appeared to vary by reporting area. This study was undertaken to determine tuberculosis case definitions used by reporting areas and to describe the extent to which the current (1990) definition published by the Council of State and Territorial Epidemiologists and CDC is used.

Methods

The 1990 surveillance definition for tuberculosis[2] has three components: (1) culture-positive cases in which *Mycobacterium tuberculosis* is isolated from a clinical specimen; (2) cases in which there is demonstration of acid-fast bacilli in a clinical specimen when a culture has not been or cannot be obtained; and (3) clinically diagnosed cases, which require all four of the following criteria: (a) a positive tuberculin skin test; (b) signs and symptoms compatible with tuberculosis, such as an abnormal and unstable (i.e., worsening or improving) chest radiograph, or clinical evidence of current disease; (c) treatment with two or more antituberculosis medications; and (d) a completed diagnostic evaluation. CDC has traditionally included in national morbidity reports all cases that are considered verified by the reporting areas without requiring that the cases meet the published case definition.

In January 1993, a copy of the criteria used to verify cases of tuberculosis for reporting to CDC was requested from each tuberculosis control officer at the health department in the 53 reporting areas (50 states, District of Columbia, New York City, and Puerto Rico). Tuberculosis control officers were also asked to submit written documentation of other criteria used to verify tuberculosis in children or in patients infected with the human immunodeficiency virus (HIV).

The authors are with the Division of Tuberculosis Elimination, National Center for HIV, STD, and TB Prevention (pending organizational approval), Centers for Disease Control and Prevention, Atlanta, Ga.

Requests for reprints should be sent to Scott B. McCombs, MPH, Surveillance and Epidemiologic Investigations Branch, Division of Tuberculosis Elimination, National Center for HIV, STD, and TB Prevention, Centers for Disease Control and Prevention, 1600 Clifton Rd, NE, Mailstop E-10, Atlanta, GA 30333.

This paper was accepted November 16, 1995.