Methods

When to Combine Hypotheses and Adjust for Multiple Tests

Peter J. Veazie

Objective. To provide guidelines for identifying composite hypotheses and addressing the probability of false rejection for multiple hypotheses.

Data Sources and Study Setting. Examples from the literature in health services research are used to motivate the discussion of composite hypothesis tests and multiple hypotheses.

Methods. This article is a didactic presentation.

Principal Findings. It is not rare to find mistaken inferences in health services research because of inattention to appropriate hypothesis generation and multiple hypotheses testing. Guidelines are presented to help researchers identify composite hypotheses and set significance levels to account for multiple tests.

Conclusions. It is important for the quality of scholarship that inferences are valid: properly identifying composite hypotheses and accounting for multiple tests provides some assurance in this regard.

Key Words. Composite hypotheses, joint tests, multiple hypotheses, significance level adjustment

Recent issues of *Health Services Research (HSR)*, the *Journal of Health Economics*, and *Medical Care* each contain articles that lack attention to the requirements of multiple hypotheses. The problems with multiple hypotheses are well known and often addressed in textbooks on research methods under the topics of joint tests (e.g., Greene 2003; Kennedy 2003) and significance level adjustment (e.g., Kleinbaum et al. 1998; Rothman and Greenland 1998; Portney and Watkins 2000; Myers and Well 2003; Stock and Watson 2003); yet, a look at applied journals in health services research quickly reveals that attention to the issue is not universal.

This paper has two goals: to remind researchers of issues regarding multiple hypotheses and to provide a few helpful guidelines. I first discuss when to combine hypotheses into a composite for a joint test; I then discuss the adjustment of test criterion for sets of hypotheses. Although often treated in

statistics as two solutions to the same problem (Johnson and Wichern 1992), here I treat them as separate tasks with distinct motivations.

In this paper I focus on Neyman–Pearson testing using Fisher’s p -value as the interpretational quantity. Classically, a test compares an observed value of a statistic with a specified region of the statistic’s range; if the value falls in the region, the data are considered not likely to have been generated given the hypothesis is true, and the hypothesis is rejected. However, it is common practice to instead compare a p -value to a significance level, rejecting the hypothesis if the p -value is smaller than the significance level. Because most tests are based on tail areas of distributions, this is a distinction without a difference for the purpose of this paper, and so I will use the p -value and significance-level terms in this discussion.

Of greater import is the requirement that hypotheses are stated *a priori*. A test is based on the prior assertion that if a given hypothesis is true, the data generating process will produce a value of the selected statistic that falls into the rejection region with probability equal to the corresponding significance level, which typically corresponds to a p -value smaller than the significance level. Setting hypotheses *a priori* is important in order to avoid a combinatorial explosion of error. For example, in a multiple regression model the *a posteriori* interpretation of regression coefficients in the absence of prior hypotheses does not account for the fact that the *pattern* of coefficients may be generated by chance. The important distinction is between the *a priori* hypothesis “the coefficient estimates for these particular variables in the data *will be* significant” and the *a posteriori* observation that “the coefficient estimates for these particular variables *are* significant.” In the first case, even if variables other than those identified in the hypothesis do not have statistically significant coefficients, the hypothesis is rejected nonetheless. In the second case, the observation applies to any set of variables that happen to have “statistically significant” coefficients. Hence, it is the probability that *any* set of variables have resultant “significant” statistics that drives the *a posteriori* case. As the investigator will interpret *any* number of significant coefficients that happen to result, the probability of significant results, given that no relationships actually exist, is the probability of getting *any* pattern of significance across the set of explanatory variables. This is different from a specific *a priori* case in which the pattern is preestablished by the explicit hypotheses. See the literatures on

Address correspondence to Peter J. Veazie, Ph.D., Department of Community and Preventive Medicine, University of Rochester, 601 Elmwood Ave., Box 644, Rochester, NY 14642.

False Discovery Rate (e.g., Benjamini and Hochberg 1995; Benjamini and Liu 1999; Yekutieli and Benjamini 1999; Kwong, Holland, and Cheung 2002; Sarkar 2004; Ghosh, Chen, and Raghunathan 2005) and Empirical Bayes (Efron et al. 2001; Cox and Wong 2004) for methods appropriate for a posteriori investigation.

WHEN TO COMBINE HYPOTHESES

What is achieved by testing an individual a priori hypothesis in the presence of multiple hypotheses? The answer to this question provides guidance for determining when a composite hypothesis (i.e., a composition of single hypotheses) is warranted. The significance level used for an individual test is the *marginal* probability of falsely rejecting the hypothesis: the probability of falsely rejecting the hypothesis regardless of whether the remaining hypotheses are rejected (see online-only appendix for details). The implied indifference to the status of the remaining hypotheses, however, is indefensible if the conclusions require a specific result from other hypotheses. This point underlies a guiding principle:

Guideline 1: *A joint test of a composite hypothesis ought to be used if an inference or conclusion requires multiple hypotheses to be simultaneously true.*

The guideline is motivated by the logic of the inference or conclusion and is independent of significance levels. Examples from the literature can prove helpful in understanding the application of this guideline. Because it is unnecessarily pejorative to reference specific studies, the following discussion will only identify the nature of the problem in selected articles but not the articles themselves (the editors of HSR and the reviewers of this paper were provided the explicit references), but the general character of the examples ought to be familiar to most researchers.

Example 1—Polynomials

Two recent articles in the *Journal of Health Economics* each regressed a dependent variable on, among other variables, a second order polynomial—a practice used to capture nonlinear relationships. The null hypothesis for each coefficient of the polynomial was rejected according to its individual *t*-statistic. It was concluded that the explanatory variable had a parabolic relationship with the dependent variable, suggesting the authors rejected the hypotheses that *both* coefficients were simultaneously zero: the joint hypothesis regarding

both coefficients is the relevant one. This is different from a researcher testing second-order nonlinearity (as opposed to testing the parabolic shape); in this case an individual test of the coefficient on the second-order term (i.e., the coefficient on the squared variable) is appropriate because the value of the first order term is not influential in this judgment of nonlinearity.

Example 2—Set of Dummy Variables

A recent article in *Medical Care* categorized a count variable into three size-groups and used a corresponding set of dummy variables to represent the two largest (the smallest group being the reference category); based on the individual significance of the two dummy variables they rejected the hypothesis that both coefficients were zero and concluded that the dependent variables was related to being larger on the underlying concept. In this conclusion, they collapsed two categories into a single statement about being larger on the underlying variable. Yet, if the authors meant that both categories are larger than the reference group, then it is a test of both coefficients being simultaneously zero that is relevant. A similar example using dummy variables is if we have an a priori hypothesis that the utilization of emergency services is not greater for blacks than whites, and another a priori hypothesis stating that utilization is not greater for Native Americans than whites. We may be justified in testing each coefficient if our interest in each minority group is independent of the other. However, a claim that “blacks and Native Americans both do not differ from whites in their utilization” makes sense only if both coefficients are simultaneously zero. Again, a joint test is indicated.

Example 3—Combined Effects across Independent Variables

Recent articles in *HSR* and the *Journal of Health Economics*, developed and tested a priori hypotheses regarding individual model parameters. So far, so good; but it was then inferred that the expected value of the dependent variable would differ between groups defined by different *profiles* of the explanatory variables. Here again the conclusion requires rejecting that the coefficients are simultaneously zero. For example, suppose we reject the hypothesis that age does not differentiate health care utilization and we reject the hypothesis that wealth does not differentiate health care utilization. These individual hypothesis tests do not warrant claims regarding wealthy elderly, poor youth, or other combinations. The coefficients for the age and wealth variables must both be nonzero, if such claims are to be made.

Example 4—Combined Effects across Dependent Variables

Recent articles in *Medical Care*, *HSR*, and the *Journal of Health Economics* each included analyses in which the same set independent variables were regressed on a number of dependent variables. Individual independent variables were considered regarding their influence across the various dependent variables. If an independent variable is considered to be simultaneously related to a number of dependent variables, then a joint test of a composite hypothesis is warranted. For example, suppose I wish to test a proposition that after controlling for age, health care utilization does not differ by sex. Suppose I use a two-part model (one part models the probability of any utilization, the other part models positive utilization given some utilization).¹ In this case I have two dependent variables (an indicator of any utilization and another variable measuring how much utilization gives positive utilization). If my proposition is correct then the coefficients on sex across both models should be simultaneously zero: a joint test is appropriate. If instead I test the two sex coefficients separately, I will implicitly be testing the hypotheses that (1) sex does not differentiate any utilization *whether or not it differentiates positive utilization* and (2) sex does not differentiate positive utilization *whether or not it differentiates any utilization*, which statistically does not address the original proposition. One might suppose if the dependent variables were conditionally independent from each other the joint test would provide similar results as the two individual hypotheses, not so. The type 1 error rate when using the individual tests is too large, unless the corresponding significance levels are divided by the number of hypotheses (see the section on adjusting for multiple tests below), in which case this type of adjustment is sufficient for independent tests.

Alternatively, suppose I wish to consider the effects of using nurse educators regarding diabetes care on health outcomes (e.g., A1c levels) and on patients' satisfaction with their health care organization, but my interest in these effects are independent of each other. In this case I am interested in two separate hypotheses, say for example (1) there is no effect on health outcomes regardless of the effect on satisfaction and (2) there is no effect on satisfaction regardless of any effect on outcomes. So long as I do not interpret these as a test that both effects are simultaneously zero, I can legitimately consider each hypothesis separately. But if each individual test does not reject the null, I should not infer that both effects are zero in the population (even with appropriate power) as this would require a joint test.

The preceding examples are in terms of individual model parameters. Guideline 1, however, applies to any set of hypotheses regardless of their

complexity. In general, if a researcher desires to test a theory with multiple implications that must simultaneously hold for the theory to survive the test, then the failure of a single implication (as an independent hypothesis) defeats the theory. A joint hypothesis test is indicated.

The following guideline presents another heuristic to distinguish the need for joint versus separate tests.

Guideline 2: *If a conclusion would follow from a single hypothesis fully developed, tested, and reported in isolation from other hypotheses, then a single hypothesis test is warranted.*

Guideline 2 asks whether a paper written about a given inference or conclusion would be coherent if based solely on the result of a single hypothesis. If so, then a single hypothesis test is warranted; if not, then consideration should be given to the possibility of a composite hypothesis. One could not support a claim that wealthy elderly use more services than poor youth, based solely on the hypothesis relating wealth and utilization, information regarding age is required.

Unfortunately, joint tests have a limitation that must be kept in mind, particularly when the hypothesis being tested is not the hypothesis of interest (which is often the case with null hypotheses). Rejecting a joint test of a composite hypothesis does not tell us which specific alternative case is warranted. Remember that a joint test of N hypotheses has $2^N - 1$ possible alternatives (in terms of the patterns of possible true and false hypotheses); for example, a joint test of two hypotheses (say, h_1 and h_2) has three possible alternatives (h_1 true and h_2 false; h_1 false and h_2 true; and both h_1 and h_2 false); a joint test of five hypotheses has 31 possible alternatives. If your interest is in a specific alternative (e.g., all hypotheses are false, which is common and is the case in many of the examples discussed above), the rejection of a joint test does not provide unique support.

To answer the question of why the joint hypothesis was rejected, it can be useful to switch from the testing paradigm to a classic p -value paradigm by inspecting the relative “level of evidence” the data provides regarding each alternative case. Here p -values are used to suggest individual components of the composite hypothesis that are relatively not well supported by the data, providing a starting point for further theory development. In this exercise, the individual p -values of the component hypotheses are compared relative to each other; they are not interpreted in terms of significance. For example, if a two component composite null hypothesis is rejected but the individual p -values are .45 and .15, the “nonsignificance” of the p -values is irrelevant, it

is the observation that one of the p -values is greatly smaller than others that provides a hint regarding why the joint hypothesis was rejected. This is theory and model building, not testing; hence, analyzing the joint test by inspecting the marginal p -values associated with its individual components is warranted as a useful heuristic—but admittedly not very satisfactory.

Alternatively, because identifying reasons for failure of a joint hypothesis is an a posteriori exercise, one could apply the methods of False Discovery Rate (Benjamini and Hochberg 1995; Benjamini and Liu 1999; Yekutieli and Benjamini 1999; Kwong, Holland, and Cheung 2002; Sarkar 2004; Ghosh, Chen, and Raghunathan 2005) or Empirical Bayes Factors (Efron et al. 2001; Cox and Wong 2004) to identify reasons for failure of the joint test (i.e., the individual hypotheses that are more likely to be false).

WHEN TO ADJUST SIGNIFICANCE LEVELS

In this section I use the phrase “significance level” to mean the criterion used in a test (commonly termed the “ α level”); I use the phrase “probability of false rejection,” denoted by pfr , to refer to the probability of falsely rejecting one or more hypotheses. A significance level is an operational part of a test (denoting the probability associated with the test’s rejection region) whereas a probability of false rejection is a theoretical result of a test or grouping of tests. I use the modifier “acceptable” in conjunction with pfr to mean a probability of false rejection deemed the largest tolerable risk. I use the modifier “implied” in conjunction with pfr to mean the probability of false rejection resulting from the application of a test or group of tests. An acceptable pfr is subjective and set by the researcher, whereas an implied pfr is objective and calculated by the researcher. Suppose I wish to test three hypotheses, and I consider a 0.1 or less probability of falsely rejecting at least one of the hypotheses as acceptable across the tests; the acceptable pfr is 0.1. If I set the significance level for each test to 0.05 (thereby determining to reject a hypothesis if the p -value of its corresponding statistic is less than .05), the probability of false rejection is $1 - (1 - 0.05)^3 = 0.143$; this is the implied pfr of the analysis associated with the hypothesis testing strategy. In this case, my strategy has an implied pfr value (0.143) that exceeds my acceptable pfr value (0.1); by this accounting, my strategy is unacceptable in terms of the risk of falsely rejecting hypotheses.

The preceding section on joint hypothesis tests presents guidance for identifying appropriate individual and composite hypotheses. Once a set of hypotheses is identified for testing, significance levels for each test must be set;

or more generally, the rejection regions of the statistic must be selected. This task requires setting acceptable *pfr*'s; that is determining the acceptable risk of rejecting a true hypothesis.

A *pfr* can be associated with any group of tests. Typically no more than three levels are considered: individual hypotheses, mutually exclusive families of hypotheses, and the full analysis-wide set of hypotheses. Although common, it is not required to use the same acceptable *pfr* for each test. Some hypotheses may have stronger prior evidence or different goals than others, warranting different test-specific acceptable *pfr*'s. A family of hypotheses is a subset of the hypotheses in the analysis. They are grouped as a family explicitly because the researcher wishes to control the probability of false rejection among those particular hypotheses. For example, a researcher may be investigating two specific health outcomes and have a set of hypotheses for each; the hypotheses associated with each outcome may be considered a family, and the researcher may desire that the *pfr* for each family be constrained to some level. An acceptable analysis-wide *pfr* reflects the researcher's willingness to argue their study remains useful in the face of criticisms such as "Given your hypotheses are correct, the probability of reporting one or more false rejections is P " or "Given your hypotheses are correct, the expected number of false rejections is N ."

The usefulness of setting *pfr*'s depends on one's perspective. From one view, we might contend that the information content of investigating 10 hypotheses should not change depending on whether we pursue a single study comprising all 10 hypotheses or we pursue 10 studies each containing one of the hypotheses; yet if we apply an analysis-wide *pfr* to the study with 10 hypotheses, we expect to falsely reject fewer hypotheses than we expect if we tested each hypothesis in a separate study.² If the hypotheses are independent such that the 10 independent repetitions of the data generating process do not in themselves accrue a benefit, there is merit to this observation and we might suppose that an analysis-wide *pfr* is no more warranted than a program-wide *pfr* (i.e., across multiple studies).

Our judgment might change, however, if we take a different view of the problem. Suppose I designed a study comprising 10 hypotheses that has an implied *pfr* corresponding to an expected false rejection of 9 of the 10 hypotheses. Should I pursue this study of 10 hypotheses for which I expect to falsely reject 90 percent of them if my theory is correct? Moreover, is it likely the study would be funded? I suggest the answers are both no. What if instead we expect 80 percent false rejections, or 40 percent, or 10 percent? The question naturally arises, what implied *pfr* is sufficient to warrant pursuit of the

study? To answer that question is to provide an acceptable *pfr*. Once an acceptable *pfr* is established it seems prudent to check whether the design can achieve it, and if it cannot, to make adjustments. Clearly, this motivation applies to any family of hypotheses within a study as well. From this perspective, the use of *pfr*'s in the design of a single study is warranted.

This is not to suggest that researchers ought to always concern themselves with analysis-wide *pfr*'s in their most expansive sense; only that such considerations can be warranted. Research is often more complex than the preceding discussion implies. For example, it is common practice to report a table of descriptive statistics and nuisance parameters (e.g., parameters on control variables) as background along side the core hypotheses of a study. A researcher may legitimately decide that controlling the Type 1 error across these statistics is unimportant and focus on an acceptable *pfr* only for the family of core hypotheses. In this case, however, a more elegant approach is to provide interval estimates for the descriptive statistics and nuisance parameters without the pretense of "testing," particularly as a priori hypotheses about background descriptive statistics are not often developed, thereby precluding them from the present consideration.

In setting acceptable *pfr*'s, the researcher should keep in mind that the probability of mistakenly rejecting hypotheses increases with the number of hypothesis tests. For example, if a researcher has settled on 10 tests for their analysis, the probability of mistakenly rejecting one or more of the 10 hypotheses at a significance level of 0.05 is approximately 0.4. Is it acceptable to engage a set of tests when the probability of falsely rejecting one or more of them is 40 percent? The answer is a matter of judgment depending on the level of risk a researcher, reviewer, editor, or reader is willing to take regarding the reported findings.

Being arbitrary, the designation of an acceptable *pfr* is not likely to garner universal support. Authors in some research fields, with the goal of minimizing false reports in the literature, recommend adjusting the significance levels for tests to restrict the overall analysis-wide error rate to 0.05 (Maxwell and Delaney 2000). However, when there are numerous tests, this rule can dramatically shrink the tests' significance levels and either require a considerably larger sample or severely diminish power. A recent article in *Health Services Research* reported results of 74 tests using a significance level of 0.05: there is a 98 percent chance of one or more false rejections across the analysis, a 10 percent chance of six or more, and a 5 percent chance of between seven and eight or more. The expected number of false rejections in the analysis is approximately four. The analysis-wide *pfr* can be restricted to

less than 0.05 as Maxwell and Delaney (2000) suggest by setting the significance levels to 0.00068 (the process of adjustment is explained below). This recommended significance level is two orders of magnitude smaller than 0.05. If an analysis-wide *pfr* of 0.98 (associated with the significance levels of 0.05) is deemed unacceptably high and an analysis-wide *pfr* of 0.05 (associated with the significance levels of 0.00068) is deemed too strict, the researchers may settle on a reasoned intermediate value. For example, to obtain a just-better-than-even odds *against* a false rejection across the full analysis of 74 tests (e.g., setting the *pfr* to 0.499), the significance levels would have been adjusted to 0.0093. Alternatively, the researchers might desire to control the expected number of false rejections across the full analysis, which can be calculated as the sum of the individual significance levels. For example, setting significance levels to 0.01351 provides an expectation of one false rejection among the 74 tests rather than the expected four associated with the original 0.05 significance levels. The adjusted significance levels in this example are less than the original significance level of 0.05, and they vary in their magnitude (and therefore power to discern effects) depending on their underlying reasoning.

Whatever rational determines the acceptable *pfr*'s for the analysis, the significance levels must be set to assure these *pfr*'s are not exceeded at any level for which they are set. Guideline 3 presents one procedure to accomplish this task.

Guideline 3: *A five-step procedure for setting significance levels.*

Step 1. Determine the set of hypotheses to be tested (applying Guidelines 1 and 2 to identify any joint hypotheses), assign an acceptable *pfr* to each hypothesis, and set the significance levels equal to these *pfr*'s.

Step 2. Determine families of hypotheses, if any, within which the probability of false rejection is to be controlled, and assign each family an acceptable *pfr*.

Step 3. Assign an acceptable analysis-wide *pfr* if desired.

Step 4. For each family, compare the implied family *pfr* with the acceptable family *pfr*. If the implied *pfr* is greater than the acceptable *pfr*, adjust the significance levels (see the following discussion on adjustment) so the implied *pfr* based on the adjusted significance levels is no greater than the acceptable *pfr*.

Step 5. If an analysis-wide acceptable *pfr* is set, calculate the analysis-wide *pfr* implied by the significance levels from *Step 4*. If the implied *pfr* exceeds the acceptable analysis-wide *pfr*, then adjust the test-specific significance levels such that the implied *pfr* does not exceed the acceptable *pfr*.

By this procedure the resulting significance levels will assure that the acceptable pfr at each level (hypothesis, family, and analysis) is not exceeded. The resulting significance levels are governed by the strictest pfr 's. Ignoring a level is implicitly setting its pfr to the sum of the associated significance levels.

Steps 4 and 5 of Guideline 3 require the adjustment of significance levels. One approach to making such adjustments is the ad hoc reassessment of the acceptable hypothesis-specific pfr 's such that they are smaller. By this approach, the researcher reconsiders her acceptable pfr for each hypothesis and recalculates the comparisons with the higher level pfr 's. Of course, the outside observer could rightfully wonder how well reasoned these decisions were to begin with if they are so conveniently modified. A common alternative is to leave all pfr 's as they were originally designated and use a Bonferroni-type adjustment (or other available adjustment method). To preserve the relative importance indicated by the relative magnitudes among the acceptable pfr 's, a researcher transforms the current significance levels into normalized weights and sets the new significance levels as the weight multiplied by the higher-level pfr . For example, if a family of three hypotheses has significance levels of 0.05, 0.025, and 0.01, the implied family-level pfr is 0.083. If the acceptable family pfr is 0.05, then the implied pfr is greater than the acceptable pfr and adjustment is indicated. Weights are constructed from the significance levels as $w_1 = 0.05/(0.05+0.025+0.01)$, $w_2 = 0.025/(0.05+0.025+0.01)$, and $w_3 = 0.01/(0.05+0.025+0.01)$. The adjusted significance levels are then calculated as $0.029 = w_1 \times 0.05$, $0.015 = w_2 \times 0.05$, and $0.006 = w_3 \times 0.05$, which have an implied family-level $pfr = 0.049$ meeting our requirement that it not exceed the acceptable pfr of 0.05.

In the preceding example the adjusted significance levels implied a pfr that was less than the acceptable pfr (i.e., $0.049 < 0.05$). A Bonferroni-type adjustment assures that the implied pfr is less than or equal to the acceptable pfr , consequently it is conservative and may unnecessarily diminish power by setting overly strict significance levels. An adjustment with better power, while not exceeding the acceptable pfr , can be attained by inflating the Bonferroni adjusted significance levels by a constant factor until the implied pfr is equal to the acceptable pfr . Although perhaps trivial in the present example, inflating each Bonferroni-adjusted significance level by a factor of 1.014 yields significance levels with an implied family-level pfr of 0.05, exactly that of the acceptable pfr .

Adjusting significance levels by reweighting produces the distribution of a higher-level pfr according to the relative importance implied by the initial significance levels. The final adjusted significance levels are a redistribution of

the strictest pfr ; therefore, adjusted significance levels no longer represent the acceptable hypothesis-level pfr 's. However, the implied hypothesis-level pfr 's will be less than the acceptable hypothesis-level pfr 's thereby meeting the requirement that the probability of false rejection is satisfactory at all levels. Because, when the inflation factor of the preceding paragraph is not used, the adjusted significance levels sum to the strictest pfr , this pfr can be interpreted as the expected number of false rejections. If the inflation factor is used, then of course the adjusted significance levels will be larger and the expected number of rejections will be larger than the pfr .

Sample size and power calculations should be based on the final adjusted significance levels. If the corresponding sample size is infeasible or the power unacceptable, then reconsideration of the study design, sampling strategy or estimators is warranted. If no additional efficiency is forthcoming, then a re-evaluation of the project goals may save the proposed study. For example, it may be that changing the study goals from guiding policy decisions to furthering theory will warrant greater leniency in the acceptable pfr 's.

CONCLUSIONS

This paper is not intended as a tutorial on the statistical procedures for joint tests and significance level adjustment: there is considerable information in the statistics and research methods literatures regarding these details. F -statistics and χ^2 statistics are commonly available for testing sets of hypotheses expressed as functions of jointly estimated model parameters, including both single and multiple equation models (see, e.g., Greene 2003; Kennedy 2003). More generally, there are joint tests available for sets of hypotheses generated from separately estimated models (see the routine *Seemingly Unrelated Estimation*, based on White's sandwich variance estimator, in STATA 2003); for example, hypotheses comparing functions of parameters from a linear regression, a logistic regression, and a Poisson model can be jointly tested. If these tests are not applicable, tests based on bootstrapped data sets can often be successfully used (Efron and Tibshirani 1993). Regarding basic descriptions of Bonferroni adjustment, see Johnson and Wichern (1992), Harris (2001), and Portney and Watkins (2000), among others.

The preceding section on when to adjust significance levels implies such adjustments are warranted. This is not a universally accepted view; indeed, the use of adjustments for multiple tests has been the focus of considerable debate

(see e.g., Rothman 1990; Saville 1990; Savitz and Olshan 1995, 1998; Goodman 1998; Thompson 1998a, b). When reviewing this debate, or considering the merit of multiple testing adjustment, the distinction between a priori hypotheses and a posteriori observations is important, a distinction carefully drawn by Thompson (1998a).

Although didactic in nature, I do not presume the ideas presented here are new to the majority of health services researchers; however, reading the journals in our field suggests that we may sometimes forget to apply what we know. The two goals of this paper were to remind researchers to consider their approach to multiple hypotheses and to provide some guidelines. Whether researchers use these guidelines or others, it is important for the quality of scholarship that we draw valid inferences from the evidence we consider: properly identifying composite hypotheses and accounting for multiple tests provides some assurance in this regard.

NOTES

1. Appreciation to an anonymous reviewer for suggesting the two-part model as an example.
2. Appreciation to an anonymous reviewer for pointing out the 10 hypotheses/10 studies example.

REFERENCES

- Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society B* 57 (1): 289–300.
- Benjamini, Y., and W. Liu. 1999. "A Step-Down Multiple Hypotheses Testing Procedure That Controls the False Discovery Rate under Independence." *Journal of Statistical Planning and Inference* 82 (1–2): 163–70.
- Cox, D. R., and M. Y. Wong. 2004. "A Simple Procedure for the Selection of Significant Effects." *Journal of the Royal Statistical Society B* 66 (2): 395–400.
- Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher. 2001. "Empirical Bayes Analysis of a Microarray Experiment." *Journal of the American Statistical Association* 96 (456): 1151–60.
- Ghosh, D., W. Chen, and T. Raghunathan. 2005. "The False Discovery Rate: A Variable Selection Perspective." *Journal of Statistical Planning and Inference*, in press. Available at <http://www.sciencedirect.com>

- Goodman, S. N. 1998. "Multiple Comparisons, Explained." *American Journal of Epidemiology* 147 (9): 807–12.
- Greene, W. H. 2003. *Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall.
- Harris, R. J. 2001. *A Primer of Multivariate Statistics*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Johnson, R. A., and D. W. Wichern. 1992. *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall.
- Kennedy, P. 2003. *A Guide to Econometrics*. Cambridge, MA: MIT Press.
- Kleinbaum, D. G., L. L. Kupper, K. E. Muller, and A. Nizan. 1998. *Applied Regression Analysis and Other Multivariate Methods*. Boston: Duxbury Press.
- Kwong, K. S., B. Holland, and S. H. Cheung. 2002. "A Modified Benjamini-Hochberg Multiple Comparisons Procedure for Controlling the False Discovery Rate." *Journal of Statistical Planning and Inference* 104 (2): 351–62.
- Maxwell, S. E., and H. D. Delaney. 2000. *Designing Experiments and Analyzing Data. A Model Comparison Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Myers, J. L., and A. D. Well. 2003. *Research Design and Statistical Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Portney, L. G., and M. P. Watkins. 2000. *Foundations of Clinical Research: Applications to Practice*. Upper Saddle River, NJ: Prentice Hall Health.
- Rothman, K. J. 1990. "No Adjustments Are Needed for Multiple Comparisons." *Epidemiology* 1 (1): 43–6.
- Rothman, K. J., and S. Greenland. 1998. *Modern Epidemiology*. New York: Lippincott Williams & Wilkins.
- Sarkar, S. K. 2004. "FDR-Controlling Stepwise Procedures and Their False Negatives Rates." *Journal of Statistical Planning and Inference* 125 (1–2): 119–37.
- Saville, D. J. 1990. "Multiple Comparison Procedures: The Practical Solution." *American Statistician* 44 (2): 174–80.
- Savitz, D. A., and A. F. Olshan. 1995. "Multiple Comparisons and Related Issues in the Interpretation of Epidemiologic Data." *American Journal of Epidemiology* 142: 904–8.
- . 1998. "Describing Data Requires No Adjustment for Multiple Comparisons: A Reply from Savitz and Olshan." *American Journal of Epidemiology* 147 (9): 813–4.
- STATA. 2003. "Seemingly Unrelated Estimation—Suest." In *STATA 8 Reference S-Z*, pp. 126–47. College Station, TX: Stata Press.
- Stock, J. H., and M. W. Watson. 2003. *Introduction to Econometrics*. Boston: Addison Wesley.
- Thompson, J. R. 1998a. "Invited Commentary Re: Multiple Comparisons and Related Issues in the Interpretation of Epidemiologic Data." *American Journal of Epidemiology* 147 (9): 801–6.
- . 1998b. "A Response to 'Describing Data Requires No Adjustment for Multiple Comparisons'." *American Journal of Epidemiology* 147 (9): 915.
- Yekutieli, D., and Y. Benjamini. 1999. "Resampling-Based False Discovery Rate Controlling Multiple Test Procedures for Correlated Test Statistics." *Journal of Statistical Planning and Inference* 82 (1–2): 171–96.

SUPPLEMENTARY MATERIAL

The following supplementary material for this article is available online:

APPENDIX. When to Combine Hypothesis and Adjust for Multiple Tests.