False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters

Yoav BENJAMINI and Daniel YEKUTIELI

Often in applied research, confidence intervals (CIs) are constructed or reported only for parameters selected after viewing the data. We show that such selected intervals fail to provide the assumed coverage probability. By generalizing the false discovery rate (FDR) approach from multiple testing to selected multiple CIs, we suggest the false coverage-statement rate (FCR) as a measure of interval coverage following selection. A general procedure is then introduced, offering FCR control at level q under any selection rule. The procedure constructs a marginal CI for each selected parameter, but instead of the confidence level $1 - q$ being used marginally, q is divided by the number of parameters considered and multiplied by the number selected. If we further use the FDR controlling testing procedure of Benjamini and Hochberg for selecting the parameters, the newly suggested procedure offers CIs that are dual to the testing procedure and are shown to be optimal in the independent case. Under the positive regression dependency condition of Benjamini and Yekutieli, the FCR is controlled for one-sided tests and CIs, as well as for a modification for two-sided testing. Results for general dependency are also given. Finally, using the equivalence of the CIs to testing, we prove that the procedure of Benjamini and Hochberg offers directional FDR control as conjectured.

KEY WORDS: Directional decision; False discovery rate; Multiple comparison procedure; Positive regression dependency; Simultaneous confidence interval; Type III error.

1. INTRODUCTION

It is common practice to ignore the issue of selection and multiplicity when it comes to multiple confidence intervals (CIs), reporting a selected subset of intervals at their marginal (nominal, unadjusted) level. CIs are not corrected for multiplicity even when the only reported intervals, or those highlighted in the abstract, are those for the "statistically significant" parameters. As a concrete example of this practice, consider the study of Giovannucci et al. (1995), which we later discuss in some detail. That study examined relationships between about 100 types of food intake and the risk of prostate cancer; its abstract reported only the three 95% CIs for the odds ratio that do not cover 1.

In another highly publicized report, the long-range effects of hormone therapy in postmenopausal women were studied in a large randomized clinical trial (Rossouw, Anderson, Prentice, and LaCroix 2002). Many parameters were considered in that study, and Bonferroni-adjusted CIs were reported, with marginal CIs reported alongside. As so often occurs, the multiplicity-adjusted CIs and the marginal CIs had rather contradictory implications. The research team, including some prominent statisticians, discussed the discrepancy and chose to focus on the marginal CIs. These were also the only intervals reported in the abstract. Because of their clinical importance, affecting tens of millions of women, the results of the study were further highlighted and discussed in an editorial (Fletcher and Colditz 2002). The editorial addressed the issue of which CIs to use as follows: "The authors present both nominal and rarely used adjusted CIs to take into account multiple testing, thus widening the CIs. Whether such adjustment should be used has been questioned...." Even though this study is special in that the practice was discussed and defended in the report itself, it attests to the common practice described in our opening sentence. We return to these two studies later in this article.

Ignoring the multiplicity of intervals is generally more common than ignoring the problem of multiplicity in testing. One reason why unadjusted CIs seem more acceptable than unadjusted tests is that they give the right coverage on average; the proportion of 95% CIs covering their respective parameters out of the intervals constructed (namely, the number covering divided by the number of parameters m) is expected to be .95, and thus only .05 will not be covered. So why worry?

It is often argued against this sentiment that failing to adjust for multiplicity is harmful in that it does not offer *simultaneous coverage* at a 95% level for all of the parameters considered in the problem. The main thrust of the present article is that ignoring multiplicity is harmful even if simultaneous inference is not of direct concern to the researcher. The selection of the parameters for which CI estimates are constructed or highlighted tends to cause reduced average coverage, unless their level is adjusted.

It is well known that selection, which can be presented as conditioning on an event defined by the data, may affect the coverage probability of a CI for a single parameter. For example, suppose that we report a CI only if it does not cover 0. If the true value of the parameter is 0, then the coverage probability of the single conditional CI is obviously 0.

The same problem exists when dealing with multiple CIs that are constructed for multiple parameters after selection. If we select, as before, to report or highlight only those intervals that do not cover 0, then the average coverage property may deteriorate to 0, exactly as in the case of a single parameter, and will be a far cry from the desired .95.

Example 1: Unadjusted Selected Intervals. $T_j \sim \mathrm{N}(\theta_j, 1)$ are independently distributed estimators of θ_j, $j = 1, \ldots, 200$. For each simulation, $\theta_j \equiv \theta$ remained fixed. This is done for five values of θ: 0, .5, 1, 2, and 4. The 200 parameter estimates are first subjected to a selection criterion based on initial testing unadjusted for multiplicity: select θ_j only if $|T_j| \geq Z_{1-.05/2}$. Next, for every parameter selected, a marginal (unadjusted) CI is constructed, namely $T_j \pm Z_{1-.05/2}$. The conditional coverage probability—the number of times that a parameter is covered by the CI divided by the number of times that the parameter is selected—is 0, .60, .84, .95, and .97 for $\theta = 0$, .5, 1, 2, and 4 (standard error $\leq$.01).

Yoav Benjamini is Professor (E-mail: *ybenja@post.tau.ac.il*) and Daniel Yekutieli is Lecturer (E-mail: *yekutiel@post.tau.ac.il*), Department of Statistics and Operations Research, School of Mathematical Sciences, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. This research was supported by the FIRST Foundation of the Israeli Academy of Sciences and Humanities.

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000001907

Whereas without selection, a marginal CI would ensure a coverage probability of .95, following the marginal testing selection criterion, the conditional coverage probability ranges from 0 to .97. Thus, not only might selection dramatically reduce the coverage, but also the amount of reduction is a function of the unknown parameter θ.

As already noted, constructing simultaneous CIs is used to address the issue of such selective inference. According to the Bonferroni procedure for constructing simultaneous CIs on m parameters, each marginal CI constructed at the $1 - \alpha/m$ level. Without selection, these CIs offer simultaneous coverage, in the sense that the probability that all CIs cover their respective parameters is at least $1 - \alpha$. Unfortunately, even such a strong property does not ensure the conditional confidence property following selection, as the following example demonstrates.

Example 2: Bonferroni-Selected–Bonferroni-Adjusted Intervals. The setting is similar to that in Example 1, except that the 200 parameters were first subjected to a selection criterion with Bonferroni testing: selecting θ_j only if $|T_j| \geq Z_{1-.05/(2\cdot 200)}$. Next, for every selected parameter, a Bonferroni-adjusted CI is constructed, namely, $T_j \pm Z_{1-.05/(2\cdot 200)}$. The conditional coverage probability is 0, .82, .97, 1.0, and 1.0 for $\theta = 0$, .5, 1, 2, and 4 (standard error $\leq .01$).

Although better than before, the values for small θ, particularly the zero coverage at $\theta = 0$, are as troublesome here as in Example 1. Apparently, the goal of conditional coverage following any selection rule for any set of (unknown) values for the parameters is impossible to achieve. We propose settling for a somewhat weaker property when it comes to selective CIs.

For that purpose, we suggest a point of view that emphasizes the construction of a noncovering CI. In other words, the obstacle to avoid is that of making a *false coverage statement*. For a single parameter with no selection, this point of view offers nothing new; in repeated experimentation, if on average more than $1 - \alpha$ of the CIs (constructed) cover the parameter, then no more than α of the constructed CI fail to do so. However, when selection steps in, three outcomes are possible at each repetition; either a covering CI is constructed, a noncovering CI is constructed, or the interval is not constructed at all. Therefore, even though a $1 - \alpha$ CI does not offer selective (conditional) coverage, the probability of constructing a noncovering CI is at most α,

$$\Pr\{\theta \notin CI, CI \text{ constructed}\} \leq \Pr\{\theta \notin CI\} \leq \alpha. \qquad (1)$$

When inference about multiple parameters is needed in an experiment with no selection, the situation is again similar to that of the single-parameter case. The number of noncovering CIs is equal to the number of parameters minus the number of covering CIs. Thus constructing a marginal $1 - \alpha$ CI for each parameter ensures that the expected proportion of the CIs covering their respective parameters is $1 - \alpha$ and the expected proportion of noncovering CIs is α. However, when facing both multiplicity and selection, not only is the expected proportion of coverage over selected parameters at $1 - \alpha$ not equivalent to the expected proportion of noncoverage at α, but also the latter no longer can be ensured by constructing marginal CIs for each selected parameter, as the following example demonstrates.

Example 3: The False Coverage Rate for Unadjusted Selected Intervals. The setting is similar to Example 1, where selection is based on unadjusted individual testing and unadjusted CIs are constructed. At each simulated realization, the proportion of intervals failing to cover their respective parameters among the constructed CIs is calculated (setting the proportion to 0 when none are selected). Averaging the proportions over the simulation, we get 1.0, .40, .16, .05, and .03 for $\theta = 0$, .5, 1, 2, and 4 (standard error $\leq .01$).

Thus, using a marginal procedure for each parameter, we can no longer assure that, on average, the proportion of noncovering intervals is controlled. In fact, the procedure with no adjustment for multiplicity is as poor at giving average false coverage control as it is inadequate at controlling the conditional coverage.

At this stage, the similarity between a false coverage statement about a CI for a selected parameter and a false rejection of a true null hypothesis (a false discovery) should seem natural. In fact, the expectation studied by the simulation in Example 3, is equivalent to the *false discovery rate* (FDR) criterion in multiple testing, as presented by Benjamini and Hochberg (1995; hereafter denoted by BH). Thus, if we take seriously the concern about the average false coverage of CIs after selection, then we should define a criterion that is similar to the FDR in the context of selective CIs.

We present such a criterion in this article. We define the "confidence intervals FDR," as the expected proportion of parameters not covered by their CIs among the selected parameters, where the proportion is 0 if no parameter is selected. This *false coverage-statement rate* (FCR) is a property of any procedure that is defined by the way in which parameters are selected and the way in which the multiple intervals are constructed. We formally define the FCR (in Sec. 2), discuss its properties, and demonstrate that it is a reasonable and intuitive criterion.

Example 4: FCR for Bonferroni-Selected–Bonferroni-Adjusted Intervals. The setting is similar to that of Example 2, where selection is based on Bonferroni testing, and Bonferroni CIs are then constructed. The FCR is estimated as in Example 3. The values of FCR for the foregoing selective multiple CI procedure are .05, .03, .02, 0, and 0 for $\theta = 0$, .5, 1, 2, and 4 (standard error $\leq .01$).

Thus, although the Bonferroni–Bonferroni procedure cannot offer conditional coverage, it does control the FCR at $< .05$ (see details in Sec. 2). In fact it does so too well, in the sense that the FCR is much too close to 0 for large values of θ. In this article we present better procedures, in that they adhere better to the desired level of error.

We try to face the problem in its generality. Given any selection rule, and a family of marginal confidence intervals, can we find a method of specifying the confidence level for the CI constructed that controls the FCR? This can be done, and in Section 3 we present such a general FCR controlling procedure for the case where the estimators of the parameters are independent. Our method of constructing FCR-controlling CIs is directly linked to the FDR-controlling procedure of BH. In the BH procedure, after sorting the p values $P_{(1)} \leq \cdots \leq P_{(m)}$ and calculating $R = \max\{j : P_{(j)} \leq j \cdot q/m\}$, the R null hypotheses for which $P_{(i)} \leq R \cdot q/m$ are rejected. Our suggested method of adjusting for FCR at level q is, roughly stated, to construct

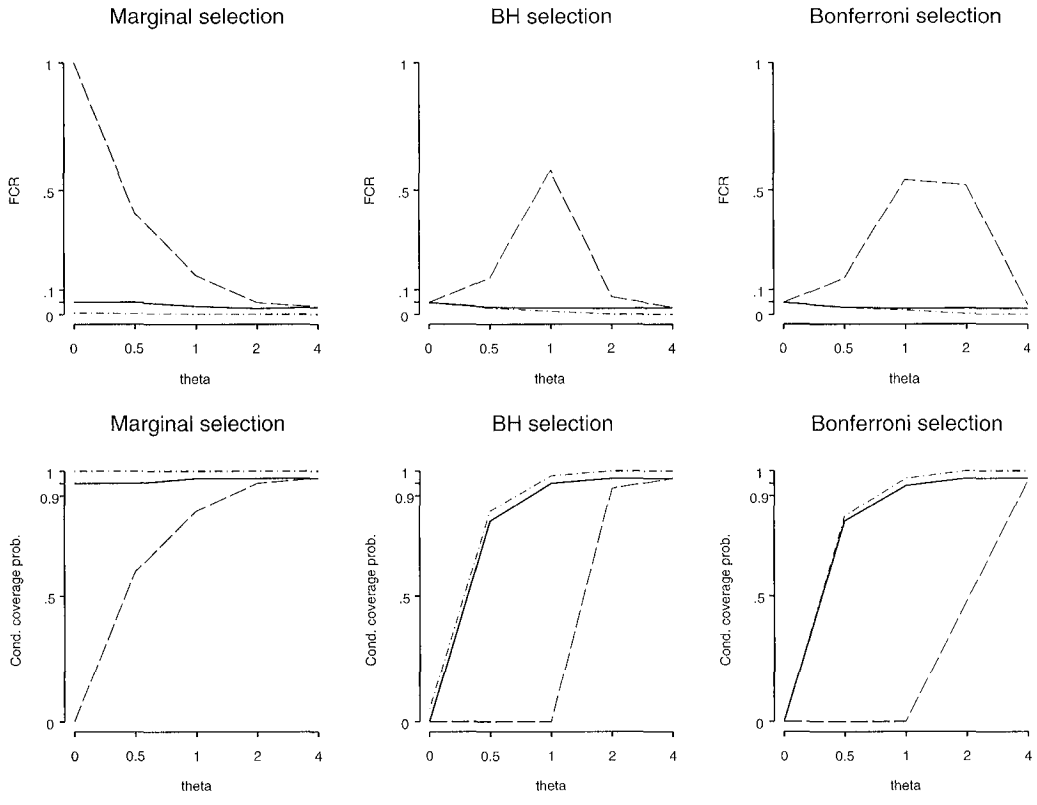

Figure 1. Simulation Based FCR and Conditional Coverage Probabilities of Marginal (- - - -), FCR-Adjusted (——), and Bonferroni (·—·—·) .95 CIs for the Marginal, BH, and Bonferroni Level .05 Selection Schemes.

a marginal CI with confidence level $1 - R \cdot q/m$ for the R parameters selected. We show that in some sense, this procedure is also the best possible general procedure.

In Section 4 we revert to the motivating problem, the construction of symmetric CIs for parameters selected by two-sided multiple-hypothesis testing procedures. Applying the general procedure allows us, as always, to control the FCR at level q. We show that if testing is done using the Bonferroni procedure, then the lower bound of the FCR may drop well below the desired level q, implying that the intervals are too long (see Fig. 1 for examples). In contrast, applying the following procedure, which combines the general procedure with the FDR controlling testing in the BH procedure, also yields a lower bound for the FCR, $q/2 \leq FCR$. This procedure is sharp in the sense that for some configurations, the FCR approaches q.

Definition 1: FCR-Adjusted BH-Selected CIs.

1. Sort the p values used for testing the m hypotheses regarding the parameters, $P_{(1)} \leq \cdots \leq P_{(m)}$.
2. Calculate $R = \max\{i : P_{(i)} \leq i \cdot q/m\}$.
3. Select the R parameters for which $P_{(i)} \leq R \cdot q/m$, corresponding to the rejected hypotheses.
4. Construct a $1 - R \cdot q/m$ CI for each parameter selected.

Thus the foregoing procedure complements the FDR controlling testing procedure of BH; all CIs constructed do not cover their null parameter values that have been rejected. Although the foregoing results hold under some assumptions about the pivotal statistics and under independence of the estimators of the parameters, some results are shown to hold under positive dependency as well. Others hold under the most general condition at the cost of inflating the FCR by a calculable constant that depends on the number of parameters only. We discuss these results in Section 5.

The connection between FDR testing and the foregoing CIs allows us to answer in the affirmative the question of whether the BH procedure controls the FDR of the directional errors as well. That means that if we also count as an error a correctly rejected two-sided hypothesis whose direction of deviation from the null hypothesis value is opposite to the direction declared, then the expected proportion of the so-defined errors is still controlled. The concern that this need not be the case has accompanied the FDR controlling procedure since the work of Shaffer (1995) and Williams, Jones, and Tukey (1999), and has been further addressed by Shaffer (2002).

Throughout this article, we make a distinction between adjusting for multiplicity to ensure simultaneous coverage and adjusting for multiplicity to avoid the selection effect. When only a single tool is available for both purposes, the discussion of the distinction makes little difference. The availability of different tools for different goals puts the choice in the hands of the researcher. In Section 7 we discuss guidelines for making this choice intelligently in more detail, although further discussions on this subject probably will ensue.

2. THE FALSE COVERAGE–STATEMENT RATE

Consider a procedure for constructing selective multiple CIs (selective CIs), based on a vector of m parameter estimators $\mathbf{T}$. The selection procedure is given by $\mathcal{S}(\mathbf{T}) \subseteq \{1, \ldots, m\}$ and is followed by the construction of some CI for each θ_i, $i \in \mathcal{S}(\mathbf{T})$.

Let R_{CI} be the number of CIs constructed, which is the size of $S(\mathbf{T})$, and let V_{CI} be the number of constructed CIs not covering their respective parameters.

Definition 2. The FCR of a selective CI procedure is $FCR = E_{\mathbf{T}}(Q_{CI})$, where Q_{CI} is defined as

$$Q_{CI} := \begin{cases} V_{CI}/R_{CI} & \text{if } R_{CI} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

For a single parameter ($m = 1$), the FCR equals the probability of constructing a noncovering CI. Therefore, according to (1), a $1 - q$ CI has $FCR \le q$. We now show that some of the commonly used methods of constructing multiple CIs also control the FCR.

1. *Constructing a marginal (unadjusted) $1 - q$ confidence interval for all parameters.* In this case $R_{CI} = m$. The distribution of V_{CI} is determined by the joint distribution of the estimators, but $E(V_{CI}) \le m \cdot q$. Therefore,
$$E(Q_{CI}) = E(V_{CI})/m \le q.$$
2. *$1 - q$ confidence region.* Suppose that we have a procedure yielding a $1 - q$ confidence region $CR(\mathbf{T})$ for a multidimensional parameter $\boldsymbol{\theta}$, meaning that $P\{\boldsymbol{\theta} \in CR(\mathbf{T})\} \ge 1 - q$. One approach is to view $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$ as a single multidimensional parameter, that is, $R_{CI} = 1$, if the confidence region is reported; $V_{CI} = 1$ if $R_{CI} = 1$ and $\boldsymbol{\theta} \notin CR(\mathbf{T})$. Thus
$$E(Q_{CI}) = \Pr\{V_{CI} = 1\} \le \Pr\{\boldsymbol{\theta} \notin CR(\mathbf{T})\} \le q.$$
3. *Projecting a $1 - q$ confidence region.* Another use of $CR(\mathbf{T})$, more relevant to our discussion, is to project it onto the coordinates, thereby deriving a marginal confidence interval $CI_i(\mathbf{T})$ for each θ_i. A Bonferroni confidence region is a special case in which $CR(\mathbf{T})$ is a cross-product of CI_i, where each CI_i is a $1 - q/m$ marginal CI. As $CR \subseteq \{\boldsymbol{\theta} : \theta_i \in CI_i\}$, for *any* selection procedure $\mathcal{S}$, the probability of constructing at least one noncovering CI_i is also $\le q$, that is,
$$\begin{aligned} \Pr(V_{CI} > 0) &= \Pr\{\exists\, \theta_i : i \in \mathcal{S}, \theta_i \notin CI_i\} \\ &\le \Pr\{\boldsymbol{\theta} \notin CR(\mathbf{T})\} \le q. \end{aligned}$$
The property $\Pr(V_{CI} > 0) \le q$ is an extension of the familywise error (FWE) rate in multiple testing. Finally, as $\Pr(V_{CI} > 0) \ge E(Q_{CI})$, $FCR \le q$.
4. *Constructing a $1 - q$ interval for independently selected parameters.* Here we mean that the selection criterion is independent of the data from which the CIs are estimated. An obvious example is when the identity of the parameters for which the CI is constructed is determined before the data are available. Such a procedure takes us back to case 1. A less obvious example is the use of a training set, $\mathbf{T}_1$, to select the R_{CI} parameters and an independent testing set, $\mathbf{T}_2$, to construct the CIs. Under such circumstances,
$$\begin{aligned} E_{\mathbf{T}_1,\mathbf{T}_2}(Q_{CI}) &= E_{\mathbf{T}_1}\left\{ I(R_{CI} > 1) \cdot \frac{1}{R_{CI}} \cdot E_{\mathbf{T}_2} V_{CI} \right\} \\ &= E_{\mathbf{T}_1}\left\{ I(R_{CI} > 1) \cdot \frac{R_{CI} \cdot q}{R_{CI}} \right\} \le q. \end{aligned}$$

Example 5 is another case in which inference is needed for a set of CIs after a selection process. In this example, a false confidence statement can be made not only because the selected CI does not cover the parameter, but also because the decision to make the statement is false as no parameter to be covered exists.

Example 5: Search for Quantitative Trait Loci—Genetic Loci Affecting Quantitative Traits. In quantitative trait loci (QTL) analysis, the effort is to locate genes on the chromosome that partially affect the level of a quantitative property of interest. The log-odds (LOD) score is used to test for linkage between a series of genetic markers located densely over the chromosomes and several quantitative traits, in order to pinpoint a QTL. A discovery of a QTL is reported if the LOD score exceeds some threshold. The reported result is a genomic region enclosing the discovery that is suspected of covering the QTL. Considerable effort was invested in methods for finding a genomic region with a .95 probability of containing the QTL (see, e.g., Mangin, Goffinet, and Rebai 1994). Nevertheless, suppose that a quantitative trait with no genetic background, and thus no QTL, is considered. Then any genomic region reported cannot contain the QTL, and in particular, no method can provide a .95 probability of covering the parameter. Under such circumstances, the mere decision to make a confidence statement is false.

Adopting the new framework for providing inference for selected multiple CIs, a possible solution is to control the FCR—the proportion of noncovering genomic regions out of the total number of regions reported. Interestingly, addressing multiplicity is considered essential in determining the LOD threshold for QTL discovery, either by controlling the FWE in multiple testing (Lander and Kruglyak 1995) or by controlling the FDR (Weller, Song, Heyen, Lewin, and Ron 1998), but is ignored when the genomic regions are reported.

3. FALSE COVERAGE–STATEMENT RATE ADJUSTMENT FOR SELECTIVE CONFIDENCE INTERVALS

We now introduce a general method for adjusting the marginal levels of the CIs of the selected parameters, so that the corresponding selective CI procedure controls the FCR. We assume that we have at our disposal a procedure for constructing marginal CIs at any desired level. That is, for $i = 1, \ldots, m$ and each $\alpha \in [0, 1]$, $CI_i(\alpha)$ is a marginal $1 - \alpha$ CI for θ_i, $\Pr_{\theta_i}\{\theta_i \in CI_i(\alpha)\} \ge 1 - \alpha$. We further assume that the foregoing CI procedure is monotone in the confidence level: $\alpha \ge \alpha'$ implies that $CI_i(\alpha) \subseteq CI_i(\alpha')$. Recall that the selection procedure is given by $\mathcal{S}(\mathbf{T})$, and the number selected is $|\mathcal{S}(\mathbf{T})|$.

Definition 3: Level-q FCR-Adjusted Selective CIs.

1. Apply the selection criterion $\mathcal{S}$ to $\mathbf{T}$, yielding the selected set of parameters $\mathcal{S}(\mathbf{T})$.
2. For each selected parameter θ_i, $i \in \mathcal{S}(\mathbf{T})$, partition $\mathbf{T}$ into T_i and $\mathbf{T}^{(i)}$ ($\mathbf{T}$ without T_i) and find
$$\begin{aligned} &R_{\min}\left(\mathbf{T}^{(i)}\right) \\ &\quad := \min_t \left\{ \left| \mathcal{S}\left(\mathbf{T}^{(i)}, T_i = t\right) \right| : i \in \mathcal{S}\left(\mathbf{T}^{(i)}, T_i = t\right) \right\}. \quad (2) \end{aligned}$$
3. For each selected parameter θ_i, $i \in \mathcal{S}(\mathbf{T})$, construct the following CI:
$$CI_i\left(\frac{R_{\min}(\mathbf{T}^{(i)}) \cdot q}{m} \right).$$

Remark 1. For many plausible selection criteria, including selection by unadjusted testing, by Bonferroni testing, and by BH testing, $R_{\min}(\mathbf{T}^{(i)})$ can be substituted by R_{CI} in Definition 3. The reason for this is that for each $i = 1, \ldots, m$ given $\mathbf{T}^{(i)}$ for values $T_i = t$ such that θ_i is selected, $|\mathcal{S}(\mathbf{T}^{(i)}, t)|$ assumes a single value. Notable exceptions are adaptive FDR procedures (Benjamini and Hochberg 2000; Benjamini, Krieger, and Yekutieli 2003; Storey, Taylor, and Seigmund 2004), where some values of $\mathbf{T}^{(i)}$ yield $R_{\min}(\mathbf{T}^{(i)})$, which is less than R_{CI}.

Incorporating R_{CI} into Definition 3, the FCR adjustment takes on a very simple form. To ensure an FCR level q, multiply q by the number of parameters selected, divide by the size of the pool of candidates from which the selection is made and construct the marginal intervals at the adjusted level for the selected parameters. The length of the constructed CIs increases as the number of parameters considered increases, but decreases as the number of selected parameters increases. Their length may vary from that of the unadjusted to that of the Bonferroni-adjusted, depending on the extent of the selection process.

Theorem 1. If the components of $\mathbf{T}$ are independent, then for any selection procedure $\mathcal{S}(\mathbf{T})$, the FCR-adjusted selective CIs in Definition 3 enjoy $FCR \le q$.

Proof. For $r > 1$, let $A_{v,r}$ denote the following event: r CIs are constructed, and v of these CIs do not cover the corresponding parameter. Let NCI_i denote the event that a noncovering CI interval is constructed for θ_i.

Lemma 1.

$$\Pr_{\mathbf{T}}(A_{v,r}) = \frac{1}{v} \cdot \sum_{i=1}^{m} \Pr_{\mathbf{T}}\{A_{v,r}, NCI_i\}.$$

Proof. Let $A^{\omega}_{v,r}$ denote the event that the subset of parameters for which a noncovering CI is constructed is $\omega \subseteq \{1, \ldots, m\}$, where $|\omega| = v$. If $i \in \omega$, then $\Pr_{\mathbf{T}}\{A^{\omega}_{v,r}, NCI_i\} = \Pr_{\mathbf{T}}(A^{\omega}_{v,r})$; however, if $i \notin \omega$, then $\Pr_{\mathbf{T}}\{A^{\omega}_{v,r}, NCI_i\} = 0$. Then

$$\begin{aligned}\sum_{i=1}^{m} \Pr_{\mathbf{T}}\{A_{v,r}, NCI_i\} &= \sum_{\omega} \sum_{i=1}^{m} \Pr_{\mathbf{T}}\{A^{\omega}_{v,r}, NCI_i\} \\ &= \sum_{\omega} \sum_{i=1}^{m} I(i \in \omega) \cdot \Pr_{\mathbf{T}}\{A^{\omega}_{v,r}\} \\ &= \sum_{\omega} v \cdot \Pr_{\mathbf{T}}\{A^{\omega}_{v,r}\} = v \cdot \Pr_{\mathbf{T}}\{A_{v,r}\}.\end{aligned}$$

Because $\bigcup_{v=1}^{r} A_{v,r}$ is a disjoint union of events that equals the event $|\mathcal{S}| = r$, incorporating Lemma 1 into the definition of the FCR yields

$$\begin{aligned}E_{\mathbf{T}}(Q_{CI}) &= \sum_{r=1}^{m} \sum_{v=1}^{r} \frac{v}{r} \cdot \Pr_{\mathbf{T}}\{A_{v,r}\} \\ &= \sum_{r=1}^{m} \sum_{i=1}^{m} \frac{1}{r} \cdot \Pr_{\mathbf{T}}\{|\mathcal{S}| = r, NCI_i\}. \qquad (3)\end{aligned}$$

For $i = 1, \ldots, m$ and $k = 1, \ldots, m$, we define the following series of events:

$$C_k^{(i)} := \left\{\mathbf{T}^{(i)} : R_{\min}\left(\mathbf{T}^{(i)}\right) = k\right\}.$$

According to (2), for each value of $\mathbf{T}^{(i)}$ and $T_i = t_i$ such that θ_i is selected, $R_{\min} \le |\mathcal{S}(\mathbf{T}^{(i)}, t_i)|$. Therefore, (3) is less than or equal to (4),

$$\le \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \cdot \Pr_{\mathbf{T}}\left\{C_k^{(i)}, i \in \mathcal{S}, \theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right\}, \qquad (4)$$

$$\le \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \cdot \Pr_{\mathbf{T}}\left\{C_k^{(i)}, \theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right\}, \qquad (5)$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \cdot \Pr_{\mathbf{T}^{(i)}}\left\{C_k^{(i)}\right\} \cdot \Pr_{T_i}\left\{\theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right\}, \qquad (6)$$

$$\le \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \cdot \Pr_{\mathbf{T}^{(i)}}\left\{C_k^{(i)}\right\} \cdot \frac{k \cdot q}{m} = q. \qquad (7)$$

Inequality (5) follows from dropping the condition $i \in \mathcal{S}$. Equality (6) is due to the independence of $\mathbf{T}^{(i)}$ and T_i. The inequality in (7) is due to the marginal coverage property of the CIs, $CI_i(\cdot)$.

Theorem 1 demonstrates that the increase in the marginal coverage probability as dictated in Definition 3 is sufficient to ensure FCR control at level q. We now show that this increase is necessary, at least in some specific setting.

Example 6. T_i are independently distributed $U[\theta_i, \theta_i + 1]$ random variables. The marginal $1 - \alpha$ CI constructed for each θ_i is of the form $CI_i(\alpha) = [T_i - (1 - \alpha), T_i]$. The selection criterion is to choose the k parameters corresponding to the k largest parameter estimators. It is clear that this is one of the selection rules for which $R_{\min}(\mathbf{T}^{(i)}) \equiv k = R_{CI}$, so the FCR-adjusted selective CIs are of the form $CI(\frac{k \cdot q}{m})$. We further assume that all $\theta_i = \theta$, and for each of the k parameters selected, we construct a CI with confidence level $1 - q'$. In this example,

$$V_{CI} = \#\{j : \text{rank}(T_j) \ge m - k + 1, \theta_j < T_j - (1 - q')\}.$$

Therefore, V_{CI} can be expressed as $V_{CI} = \min(k, V^*)$, where $V^* \sim \text{Binom}(m, q')$. This yields an upper bound for the FCR,

$$FCR = E\frac{V_{CI}}{R_{CI}} = E\frac{V_{CI}}{k} \le E\frac{V^*}{k} = \frac{m \cdot q'}{k}.$$

The goal is small FCR values, typically $FCR = .05$, so we need values of q' such that $k \gg m \cdot q'$, thereby implying that $\Pr(V^* > k) \approx 0$. Because under the foregoing conditions, the FCR is approximately $\frac{m \cdot q'}{k}$, to control the FCR at level q, we must set $q' = k \cdot q/m$.

Example 7: The Selective CIs in Practice. Giovannucci et al. (1995) studied the relationship between the intake of carotenoids and retinol and the risk of prostate cancer, a study that received wide nonscientific press coverage. That study's findings suggest that the intake of lycopene or other compounds in tomatoes may reduce prostate cancer risk, but that other measured carotenoids are unrelated to risk. It further recommends increasing consumption of the first. Only three 95% CIs for the estimated relative risks (RRs) are reported in the abstract (that carries the foregoing recommendation)—none covers one, of course; the CI furthest away from 1 is (.44, .95), with the point estimate of RR = .65. A closer look at that article reveals

that some 131 parameters regarding various foods and beverages were inspected, at least by one count. Unfortunately, in contrast to the way it should be, the family of hypotheses tested is not well defined, and the exact count is somewhat difficult to get from the reading of the paper. Thus we do not repeat the modified calculation exactly. Nevertheless, even if we settle for a minimal count of $m = 30$ hypotheses from which the three were selected, $R/m = 3/30$, and the length of the intervals on the log scale should be inflated by >40%. For the aforementioned CI, the corresponding selective CI is (.37, 1.17). With the other two CIs also covering the value 1 for the RR, it is clear that the message conveyed in the abstract should be very different from that published. We thank Professor Kafadar for bringing the multiplicity problem in this study to our attention.

4. SELECTION VIA MULTIPLE HYPOTHESIS TESTING

In the study described in Example 7, although not stated explicitly, it seems that the selection criterion was to report only the parameters that were significantly different from 1 (marginally). The fact is that even though any selection criterion can be used in selective CIs, the practice of basing parameter selection on testing is very common.

In this section, we assume that the distribution of $T_i - \theta_i$ has a symmetric distribution independent of θ_i, $F_{\mathcal{T}_i}$, where θ_i is associated with a null value θ_i^0 and the set of parameters selected corresponds to the set of rejected null hypotheses $H_i^0 : \theta_i = \theta_i^0$ tested versus $\theta_i \neq \theta_i^0$. Testing is conducted using the two-sided p values $P_i = 2 \cdot (1 - F_{\mathcal{T}_i}(|T_i - \theta_i^0|))$, and the rejection region is specified by a critical p value $P_S(\mathbf{P})$,

$$S(\mathbf{T}, \boldsymbol{\theta}^0) = \{\theta_{(i)} : P_{(i)} \leq P_S(\mathbf{P})\}.$$

FCR-adjusted selective CIs provide the desired FCR control for selection based on testing as well, but may offer too much protection at the undesirable cost of too-wide confidence intervals. Thus in this section we study the effect of the testing procedure used for selection on the FCR-adjusted selective CIs. The fact that the selection rule has direct implication for the FCR-adjusted selective CIs, with a lower FCR associated with a stricter selection criterion, is intuitively clear from the extreme case, where if $|S(\mathbf{T})| \equiv 0$, then, trivially, $FCR = 0$. Example 8 demonstrates the foregoing phenomenon in a more realistic setting, where the Bonferroni procedure is used for testing.

Example 8. Numerous two-sided hypotheses are tested using the Bonferroni procedure at level q. Of the m tested hypotheses, $\sqrt{m}$ are false null hypotheses with $|\theta_i - \theta_i^0| \to \infty$. The remaining $m - \sqrt{m}$ hypotheses are true null hypotheses. In this case all false null hypotheses are correctly rejected, and the number of true null hypotheses rejected is $V' \sim \text{Binom}(m - \sqrt{m}, q/m)$. Thus $R_{CI} = \sqrt{m} + V'$. Given R_{CI}, for each rejected parameter, the following CI is constructed: $T_j \pm T_j^{1-R_{CI} \cdot q/(2 \cdot m)}$. Thus V_{CI} equals the V' null parameters selected plus the number of nonnull parameters not covered by their respective CIs $V'' \sim \text{Binom}(\sqrt{m}, R_{CI} \cdot q/m)$. As $R_{CI} > \sqrt{m}$,

$$FCR = E\frac{V''}{R_{CI}} + E\frac{V'}{R_{CI}} \leq E_{R_{CI}}\left\{ E_{V''|R_{CI}}\left(\frac{V''}{R_{CI}}\right)\right\} + E\frac{V'}{\sqrt{m}}$$

$$= E_{R_{CI}}\frac{\sqrt{m} \cdot R_{CI} \cdot q/m}{R_{CI}} + \frac{(m - \sqrt{m}) \cdot q/m}{\sqrt{m}} < \frac{2 \cdot q}{\sqrt{m}},$$

and as $m \to \infty$, $FCR \to 0$.

Next, we show that if the multiple-testing procedure used for selection is more liberal than the FDR-controlling test of BH at level q, then for any θ, $FCR \geq q/2$. This result, proven in Theorem 2, means that the intervals are not excessively long for any possible values of the parameters. Moreover, we then show in Corollary 1 that for some values of θ, the FCR even approaches q. Thus the FCR-adjusted BH-selected CIs described in Definition 1 yields FCRs that range from $q/2$ to q, and in some cases $FCR \approx q$.

For the aforementioned results, we need a few more conditions: (a) The components of $\mathbf{T}$ are independently distributed; (b) the testing procedure satisfies $R_{\min}(\mathbf{T}^{(i)}) = R_{CI}$ in Definition 3 (see Remark 1); and (c) denoting by $\mathcal{T}_i^{\alpha}$ the α quantile of $F_{\mathcal{T}_i}$, the marginal CI are of the form

$$CI_i(\alpha) = \{\theta_i : |T_i - \theta_i| \leq \mathcal{T}^{1-\alpha/2}\}.$$

Theorem 2. Consider an FCR-adjusted selective CI procedure under the foregoing conditions (a)–(c). If its selection is based on a multiple testing procedure which is more liberal than the procedure in BH at level q, its FCR is always greater than or equal to $q/2$.

Before we prove Theorem 2, note the following characterization of a multiple-testing procedure $S(\mathbf{T})$ that is more liberal than the procedure of BH.

Lemma 2. $S(\mathbf{T}) \supseteq S_{BH}(\mathbf{T}; q)$ implies that if $|T_i - \theta_i^0| \geq \mathcal{T}_i^{1-|S| \cdot q/(2m)}$, then $i \in S$.

Proof. The condition in the lemma can be expressed as $P_i \leq \frac{|S| \cdot q}{m}$. Recall that the number of hypotheses in $S_{BH}(\mathbf{T}; q)$ is defined as

$$|S_{BH}| = \max\left\{k : P_{(k)} \leq \frac{k \cdot q}{m}\right\}. \tag{8}$$

Thus for $S \equiv S_{BH}$, we get

$$S = \left\{i : P_i \leq \frac{|S| \cdot q}{m}\right\}.$$

For a strictly more liberal $S \supsetneq S_{BH}$, according to (8), $\frac{|S| \cdot q}{m} < P_{(|S|)}$. Thus we get

$$S \equiv \{\theta_i : P_i \leq P_{(|S|)}\} \supsetneq \left\{\theta_i : P_i \leq \frac{|S| \cdot q}{m}\right\}.$$

Proof of Theorem 2. The beginning of the proof of Theorem 2 is identical to that of Theorem 1 up to expression (3). Recall that event $C_k^{(i)}$ is defined according to $R_{\min}$. Because $R_{\min}$ now can be substituted by the number of parameters selected, the inequality in expression (4) in the proof of Theorem 1 can be replaced by an equality in expression (9) in the current proof. Thus

$$E_{\mathbf{T}}(Q_{CI})$$

$$= \sum_{i=1}^{m}\sum_{k=1}^{m}\frac{1}{k} \cdot \Pr_{\mathbf{T}}\left\{C_k^{(i)}, i \in \mathcal{S}, \theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right\} \tag{9}$$

$$\geq \sum_{i=1}^{m}\sum_{k=1}^{m}\frac{1}{k} \cdot \Pr_{\mathbf{T}}\{C_k^{(i)}, |T_i - \theta_i^0| \geq \mathcal{T}_i^{1-k \cdot q/(2 \cdot m)},$$

$$|T_i - \theta_i| \geq \mathcal{T}_i^{1-k \cdot q/(2 \cdot m)}\} \tag{10}$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \cdot \Pr_{\mathbf{T}}\{C_k^{(i)}\} \times \Pr\{|T_i - \theta_i^0| \geq \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}, |T_i - \theta_i| \geq \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}\} \quad (11)$$

$$> \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \cdot \Pr_{\mathbf{T}}\{C_k^{(i)}\} \cdot \Pr\{T_i \geq \theta_i + \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}\}$$

$$= \frac{m \cdot q}{2 \cdot m}. \quad (12)$$

Inequality (10) is due to the result of Lemma 2. The inequality in (12) is true because for $\theta_i \geq \theta_i^0$,

$$\{|T_i - \theta_i^0| \geq \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}, |T_i - \theta_i| \geq \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}\} \supseteq \{T_i \geq \theta_i + \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}\},$$

and for $\theta_i \leq \theta_i^0$,

$$\{|T_i - \theta_i^0| \geq \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}, |T_i - \theta_i| \geq \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}\} \supseteq \{T_i \leq \theta_i - \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}\}.$$

Notice that if $|\theta_i - \theta_i^0| \to 0$ or $|\theta_i - \theta_i^0| \to \infty$, then

$$\Pr\{|T_i - \theta_i^0| \geq \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}, |T_i - \theta_i| \geq \mathcal{T}_i^{1-k\cdot q/(2\cdot m)}\} \to q/m.$$

Therefore, if for all θ_i either condition holds, then (11) in the proof of Theorem 2 approaches q. Combining this and the result of Theorem 1, we get the following:

Corollary 1. Under the conditions of Theorem 2, if for all $i = 1, \ldots, m$, $|\theta_i - \theta_i^0| \to 0$ or $|\theta_i - \theta_i^0| \to \infty$, then the FCR of the FCR-adjusted CIs approaches q.

Theorem 2 and Corollary 1 emphasize the advantages of selection via the BH procedure or less conservative multiple-testing procedures, in that they do not control the FCR at an excessively low level. But there is a clear advantage to selection with the BH procedure, because it preserves the usual duality between CIs and testing. Using it as the testing procedure, any choice of parameter values covered by the CIs will not be rejected by the multiple-testing procedure, while the other parameters for which CIs are not constructed remain at their null values. That is, for any θ^* satisfying $\theta_i^* \in CI_i$ for some $i \in S$ and $\theta_i^* = \theta_i^0$ otherwise, the BH procedure will not reject $\theta_i^* \in CI_i$. In the other direction, for any θ^* satisfying $\theta_i^* \notin CI_i$ for all $i \in S$ and $\theta_i^* = \theta_i^0$ otherwise, the BH procedure will reject all θ_i^*'s for $i \in S$. In contrast, using a less conservative testing procedure than the BH procedure, a parameter can be selected after deciding that $\theta_i \neq \theta_i^0$, yet θ_i^0 is included in the CI constructed, CI_i. Thus, under the conditions of Theorem 2, the recommended procedure is the FCR-adjusted BH-selected CIs given in Definition 1, enjoying $q/2 \leq FCR \leq q$, and for some configurations of the parameters approaching q.

Figure 1 presents the results of a simulation study that demonstrates the extent of this phenomenon. The setting is as described in Example 1. Unadjusted, BH, and Bonferroni selection is applied at $q = .05$, and three types of marginal CIs are constructed, also at level $q = .05$. The three panels at the bottom show that for values θ close to 0, the conditional coverage property cannot be controlled by any of the CI schemes. The three top panels show that unadjusted marginal intervals fail to control the FCR, whereas the FCR of Bonferroni intervals approaches 0 in many cases. In comparison, the FCR of FCR-adjusted intervals is very close to .05.

Tukey (1995) was the first to search for multiple CIs dual to the BH procedure. He considered constructing CIs of the foregoing form, because they reflected the rejection decisions reached by the FDR-controlling procedure of BH. However, his construction included CIs for *all* parameters, and so he could not come up with any explicit statement about some joint coverage property of his proposed procedure. To arrive at some coverage property, Tukey (1995) tried to resort to hybrid CIs, replacing the CIs for the nonrejected parameters with Bonferroni. He later gave up (Tukey 1996), and that suggestion disappeared from his subsequent publications. Realizing that the fundamental problem is that of setting CIs for selected parameters and defining the FCR as the relevant measure of error involved, we were able to derive the relevant coverage properties. Admittedly, we gained further insight into the problem once we had to face extremely large problems in genetic research, encompassing thousands of parameters, in which interest and inference are focused on the selected parameters only. Such encounters were rare 10 years ago.

5. FALSE COVERAGE–STATEMENT RATE–ADJUSTED SELECTIVE CONFIDENCE INTERVALS UNDER DEPENDENCY

5.1 Positive Regression Dependency

The general result in Theorem 1 holds for independent parameter estimators. We now discuss parameter estimators possessing the positive regression dependent on a subset (PRDS) property.

Definition 4 (Benjamini and Yekutieli 2001). The components of $\mathbf{X}$ are PRDS on I_0 if for any increasing set D (where $x \in D$ and $y \geq x$ implies that $y \in D$) and for each $i \in I_0$, $\Pr(\mathbf{X} \in D | X_i = x)$ is nondecreasing in x.

If $\mathbf{X}$ is PRDS on any subset, then we denote it simply as PRDS. We further require that the selection criterion and the CIs be concordant, in the following sense.

Definition 5. A procedure for selective CIs is concordant if for all values of $\boldsymbol{\theta}$, for all $0 < \alpha < 1$, and for $i = 1, \ldots, m$, $k = 1, \ldots, m$, both $\{\mathbf{T}^{(i)} : k \leq R_{\min}(\mathbf{T}^{(i)})\}$ and $\{T_i : \theta_i \notin CI(\alpha)\}$ are either increasing or decreasing sets.

An example of a concordant selective CI is selection via a multiple-hypothesis procedure of tests with one-sided alternatives, $H_j^1 : \theta_j^0 < \theta_j$, and one-sided confidence intervals, $CI_j(\alpha) = \{\theta_j : \theta_j \geq T_j + \mathcal{T}^\alpha\}$.

Theorem 3. If the components of $\mathbf{T}$ are PRDS and the selection criterion and the CIs are concordant, then the FCR-adjusted selective CIs in Definition 3 enjoy $FCR \leq q$.

Proof. Without loss of generality, let us assume that the two sets in Definition 5 are increasing. Then $D_k^{(i)} = \bigcup_{j=1}^{k} C_k^{(i)}$, which can be expressed as $\{\mathbf{T}^{(i)} : R_{\min}(\mathbf{T}^{(i)}) < k + 1\}$, is a

decreasing set. Furthermore, for $\alpha \le \alpha'$, we can express $\{T_i : \theta_i \notin CI(\alpha)\} = \{T_i : t \le T_i\}$ and $\{T_i : \theta_i \notin CI_i(\alpha')\} = \{T_i : t' \le T_i\}$ with $t \le t'$. Thus the PRDS condition then implies that

$$\Pr\left(D_k^{(i)} \middle| \theta_i \notin CI(\alpha)\right) \le \Pr\left(D_k^{(i)} \middle| \theta_i \notin CI_i(\alpha')\right). \tag{13}$$

Hence for $k = 1, \ldots, m$, we get

$$\begin{aligned}
&\Pr\left(D_k^{(i)} \middle| \theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right) \\
&\quad + \Pr\left(C_{k+1}^{(i)} \middle| \theta_i \notin CI_i\left(\frac{(k+1) \cdot q}{m}\right)\right) \\
&\le \Pr\left(D_k^{(i)} \middle| \theta_i \notin CI_i\left(\frac{(k+1) \cdot q}{m}\right)\right) \\
&\quad + \Pr\left(C_{k+1}^{(i)} \middle| \theta_i \notin CI_i\left(\frac{(k+1) \cdot q}{m}\right)\right) \\
&= \Pr\left(D_{k+1}^{(i)} \middle| \theta_i \notin CI_i\left(\frac{(k+1) \cdot q}{m}\right)\right).
\end{aligned} \tag{14}$$

As defined, the event $D_m^{(i)}$ is the entire sample space. Therefore, repeatedly applying inequality (14) for $k = 1, \ldots, m$, we get

$$\begin{aligned}
\sum_{k=1}^{m} \Pr\left(C_k^{(i)} \middle| \theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right) &\le \Pr\left(D_m^{(i)} \middle| \theta_i \notin CI_i\left(\frac{m \cdot q}{m}\right)\right) \\
&= 1.
\end{aligned} \tag{15}$$

To complete the proof, we proceed from inequality (5) in the proof of Theorem 1,

$$\begin{aligned}
&E_{\mathbf{T}}(Q_{CI}) \\
&\le \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \cdot \Pr\left\{C_k^{(i)}, \theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right\} \\
&= \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \cdot \Pr\left\{C_k^{(i)} \middle| \theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right\} \\
&\qquad \cdot \Pr\left\{\theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right\} \\
&\le \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \cdot \Pr\left\{C_k^{(i)} \middle| \theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right\} \cdot \frac{k \cdot q}{m} \le q.
\end{aligned} \tag{16}$$

The first inequality in (16) is due to the coverage property of CIs, and the second inequality is due to (15).

5.2 General Dependency

Theorem 4. For any monotone marginal CIs, any selection procedure $S(\mathbf{T})$, and any dependency structure of the test statistics, the FCR of the FCR-adjusted selective CIs is bounded by $q \cdot \sum_{j=1}^{m} \frac{1}{j}$.

The immediate corollary is that FCR-adjusted selective CIs at level $q / \sum_{j=1}^{m} \frac{1}{j}$ ensure that $FCR \le q$ for all distributions of $\mathbf{T}$.

Proof of Theorem 4. The proof is based on the proof of theorem 1.3 of Benjamini and Yekutieli (2001). Whereas the proof of Benjamini and Yekutieli (2001) unnecessarily uses the assumption that $\Pr\{P_i \in [\frac{j-1}{m} q, \frac{j}{m} q]\} = \frac{q}{m}$, we only assume here that the CIs are monotone.

For each $i = 1, \ldots, m$, we define the random variable I_i. $I_i = 1$ is the event $\theta_i \notin CI_i(\frac{q}{m})$; for $j = 2, \ldots, m$, $I_i = j$ is the intersection of $\theta_i \in CI_i(\frac{j-1}{m} q)$ and $\theta_i \notin CI_i(\frac{j}{m} q)$; $I_i = m + 1$ is the event $\theta_i \in CI_i(q)$. Because the CIs $CI(\alpha)$ are monotone for $1 \le j \le m$,

$$\theta_i \notin CI_i\left(\frac{k}{q} m\right) = \bigcup_{j=1}^{k} \{T_i : I_i = j\}. \tag{17}$$

Let I_{unif} denote the following random variable: for $j = 1, \ldots, m$, $I_{\text{unif}} = j$ with probability $\frac{q}{m}$ and $I_{\text{unif}} = m + 1$ with probability $1 - q$. Finally, let j^{rec} define the following decreasing function: $j^{\text{rec}}(j) = \frac{1}{j}$ for $j = 1, \ldots, m$ and $j^{\text{rec}}(m + 1) = 0$. The validity of $CI_i(\cdot)$ implies that all I_i's are stochastically greater than I_{unif}, and thus, because j^{rec} is a decreasing function,

$$\begin{aligned}
&\sum_{j=1}^{m} \frac{1}{j} \cdot \Pr\{I_i = j\} \\
&\quad = \sum_{j=1}^{m+1} j^{\text{rec}}(j) \cdot \Pr\{I_i = j\} \\
&\quad \le \sum_{j=1}^{m+1} j^{\text{rec}}(j) \cdot \Pr\{I_{\text{unif}} = j\} = \frac{q}{m} \sum_{j=1}^{m} \frac{1}{j}.
\end{aligned} \tag{18}$$

Incorporating (17) into (5) yields

$$\begin{aligned}
FCR &\le \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \sum_{j=1}^{k} \Pr_{\mathbf{T}}\left\{C_k^{(i)}, I_i = j\right\} \\
&\le \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{1}{j} \sum_{k=j}^{m} \Pr_{\mathbf{T}}\left\{C_k^{(i)}, I_i = j\right\} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{1}{j} \Pr_{\mathbf{T}}\{I_i = j\} \le \sum_{i=1}^{m} \frac{q}{m} \sum_{j=1}^{m} \frac{1}{j}.
\end{aligned} \tag{19}$$

The inequality in (19) is due to (18).

6. CONNECTIONS BETWEEN THE FALSE COVERAGE–STATEMENT RATE AND THE FALSE DISCOVERY RATE

In this section we express the FDR and the directional FDR (Benjamini, Hochberg, and Kling 1993) as the FCR of selective CIs. This way, we are able to prove the validity of the BH procedure as a corollary of Theorem 3. More important, we use this same argument to prove that the BH procedure offers directional FDR control.

6.1 The BH Procedure Controls the False Discovery Rate

For $i = 1, \ldots, m$, let P_i be a p value for testing $H_i^0 : \theta_i \in \Theta_i^0$ versus the alternative hypothesis $\theta_i \in \mathbb{R} - \Theta_i^0$. Thus for each $0 < \alpha < 1$, $\Pr_{\theta_i \in \Theta_i^0}(P_i \le \alpha) \le \alpha$.

$\mathbf{P} = (P_1, P_2, \ldots, P_m)$ is used to define selective CIs. The selection criterion, $\mathcal{S}(\mathbf{P})$, is given by the level-q BH procedure. For each $i \in \mathcal{S}(\mathbf{P})$, the $1 - \alpha$ CI constructed is

$$CI_i(\alpha) = \begin{cases} \mathbb{R} - \Theta_i^0 & \text{if } P_i \le \alpha \\ \mathbb{R} & \text{if } P_i > \alpha. \end{cases} \tag{20}$$

In this setting the test statistic is the p value and not the parameter estimator, but the CI in (20) remains a valid, albeit somewhat wasteful, marginal CI. Furthermore, it is easy to verify that this selective CI procedure is concordant in $\mathbf{P}$.

The next step is to apply a level-q FCR adjustment to the selective CIs. Then, according to Theorem 3, if the components of $\mathbf{P}$ are positive regression dependent on any subset, $FCR \le q$.

As all $i \in \mathcal{S}(\mathbf{P})$ have $P_i \le \frac{R_{CI} \cdot q}{m}$, applying the FCR adjustment implies that all CI_i's constructed are $\mathbb{R} - \Theta_i^0$. Therefore, V_{CI} is the number of $i \in \mathcal{S}(\mathbf{P})$ for which $\theta_i \in \Theta_i^0$, that is, the number of true null hypotheses rejected by the BH procedure. Hence the FCR equals the FDR of the BH procedure, and the latter is therefore $\le q$.

The preceding result can be improved. The event $\theta_i \notin CI_i(\alpha)$ can only occur for $\theta \in \Theta_i^0$. Therefore, we can alter the summation in the proof of Theorem 3 from summation over $i = 1, \ldots, m$ to summation over the m_0 true null hypotheses. This also means that positive regression dependent on any subset is no longer needed, because positive regression dependent on the subset of true null hypotheses is sufficient. The foregoing is an alternative proof to the result of Benjamini and Yekutieli (2001).

Corollary 2. If $\mathbf{P}$ is PRDS on the subset of p values corresponding to the true null hypotheses, then the FDR of the procedure in BH is less than or equal to $m_0 \cdot q/m$.

6.2 Directional False Discovery Rate Control Under Independence

We now address in much the same way the problem of determining whether the parameter $\delta_i = \theta_i - \theta_i^0$ is positive or negative. A discovery is declaring δ_i to be either positive or negative, but there is of course the possibility of making no discovery. Making a false statement about the sign of δ_i is termed a *directional error*, or a type III error. Williams et al. (1999), Benjamini and Hochberg (2000), and Shaffer (2002) all conjectured that the BH procedure can also offer control over type III errors. Shaffer (2002) also gave some theoretical support at extreme configurations of the parameters.

To address the problem of directional errors within the FDR framework, Benjamini et al. (1993) introduced two variants of directional FDR. In *pure directional FDR*, the expected proportion of discoveries in which a positive parameter is declared negative or a negative parameter is declared positive. In *mixed directional FDR*, the expected proportion of discoveries in which a nonnegative parameter is declared negative or a nonpositive parameter is declared positive. Obviously, the pure directional FDR is always smaller than the mixed directional FDR, so the following results on the control of the mixed directional FDR hold for the pure directional FDR as well.

We assume that the distribution of the parameter estimator $D_i = T_i - \theta_i^0$ increases stochastically with δ_i, and that the cdf of D_i given $\delta_i = 0$, $F_i(D_i)$ is known. The one-sided p value is $P_i = 1 - F_i(D_i)$, and the two-sided p value is $P_{|i|} = 2 \cdot \min(P_i, 1 - P_i)$.

Definition 6: The Level-q BH Directional FDR Procedure.

1. Test the set of m two-sided hypotheses with the two-sided p values using the BH procedure at level q.
2. Let R denote the number of discoveries made.
3. If $P_{|i|} \le \frac{Rq}{m}$ and $D_i > 0$, then declare $\delta_i > 0$.
4. If $P_{|i|} \le \frac{Rq}{m}$ and $D_i < 0$, then declare $\delta_i < 0$.

We now define the selective CIs. The selection criterion is the BH procedure using the m two-sided p values. The marginal CIs are of the form,

$$CI_i(\alpha) = \begin{cases} (0, \infty) & \text{if } P_i \le \alpha/2 \\ (-\infty, \infty) & \text{if } \alpha/2 < P_i < 1 - \alpha/2 \\ (-\infty, 0) & \text{if } 1 - \alpha/2 \le P_i. \end{cases} \tag{21}$$

Applying the level-q FCR adjustment to the foregoing specific CIs, all of the CI_i constructed are either $(0, \infty)$ for $D_i > 0$ or $(-\infty, 0)$ for $D_i < 0$. Hence the FCR equals the mixed directional FDR of the level-q BH directional FDR procedure. Therefore, Theorem 1 implies that if the components of $\mathbf{D}$ are independent, then the mixed directional FDR is bounded by q.

Now take a closer look at $CI_i(\alpha)$. If $\delta_i = 0$, then $\Pr(\delta_i \notin CI_i(\alpha)) = \alpha$, whereas for $\delta_i \ne 0$, $\Pr(\delta_i \notin CI_i(\alpha)) < \alpha/2$. Modifying the summation of i in the proof of Theorem 1 from summation over all m parameters to separate summation over the m_+ indices $\{i : \delta_i > 0\}$, the m_- indices $\{i : \delta_i < 0\}$, and the m_0 indices $\{i : \delta_i = 0\}$, we get the following.

Corollary 3. If the components of D_i are independent, then the mixed directional FDR of Definition 6 is

$$\le q/2 \cdot \frac{m_+ + m_-}{m} + q \cdot \frac{m_0}{m} = q/2 \cdot \left(1 + \frac{m_0}{m}\right).$$

6.3 Directional False Discovery Rate Control Under Positive Regression Dependency

We now assume that $\mathbf{D}$ is PRDS dependent. This does not imply that the vector of two-sided p values is PRDS, but it does imply that any order-preserving transformation of $\mathbf{D}$—in this case the vector of m one-sided p values—retains the PRDS property.

Thus, rather than simultaneously testing m two-sided hypotheses, we suggest separately testing each vector of m one-sided hypotheses: (a) Using the m one-sided p values, P_i, to test the m null hypotheses $H_i^{0+} : \delta_i \le 0$, the number of true null hypotheses is $m_+ + m_0$; and (b) using the m one-sided p values, $1 - P_i$, to test the m null hypotheses $H_i^{0-} : \delta_i \ge 0$, the number of true null hypotheses is now $m_- + m_0$. Corollary 2 implies the following.

Corollary 4. If $\mathbf{D}$ is PRDS on $\{D_i : \delta_i \le 0\}$, then the mixed directional FDR of the level-q BH procedure of $\{H_i^{0+}\}_{i=1}^m$ is less than or equal to $\frac{(m_+ + m_0) \cdot q}{m}$.

Corollary 5. If $\mathbf{D}$ is PRDS on $\{D_i : \delta_i \ge 0\}$, then the mixed directional FDR of the level-q BH procedure of $\{H_i^{0-}\}_{i=1}^m$ is less than or equal to $\frac{(m_- + m_0) \cdot q}{m}$.

According to Benjamini and Yekutieli (2001), it is easy to verify that a given vector of one-sided test statistics is PRDS. For example, positive correlated multivariate normal test statistics are PRDS. However, it is much harder to show that two-sided test statistics are PRDS. For example, absolute values of

positive correlated multivariate normals are not PRDS. The following procedure ensures FDR control for two-sided inference even if the two-sided test statistics are not PRDS.

Definition 7: The Level-q Modified BH Procedure for Two-Sided Inference.

1. Using P_i, test $\{H_i^{0+}\}_{i=1}^m$ using the BH procedure at level $q/2$; let I^+ denote the set of rejected one-sided null hypotheses.
2. Using $1 - P_i$, test $\{H_i^{0-}\}_{i=1}^m$ using the BH procedure at level $q/2$; let I^- denote the set of rejected one-sided null hypotheses.
3. Reject the set of null hypotheses, $I_1 = I^+ \cup I^-$.

Let V^+, V^-, V, R^+, R^-, and R denote the number of false discoveries and total number of discoveries at stages 1, 2, and 3 of the modified BH procedure. According to Corollaries 4 and 5, and because

$$E\frac{V^+}{R^+} + E\frac{V^-}{R^-} \geq E\frac{V}{R},$$

we get the following.

Corollary 6. If the vector of parameter estimators is PRDS, then the mixed directional FDR of the modified BH procedure for two-sided inference is less than or equal to $q \cdot \frac{2 \cdot m_0 + m_+ + m_-}{2m}$.

It is easy to verify that Definition 6 is equivalent to simultaneously testing all $2 \cdot m$ one-sided null hypotheses using the BH procedure at level q. This implies that Definition 7 is less powerful than Definition 6. On the other hand it has the advantage that the FDR is controlled separately both for both the positive and the negative differences. This may be a desirable property in some applications, such as multiple endpoints in clinical trials or overexpression and underexpression of genes in microarray analysis.

It is often argued that in reality, an exact null hypothesis is never true (see Williams et al. 1999); that is, $m_0 = 0$, in which case Definitions 6 and 7 at level $2 \cdot q$ have directional $FDR \leq q$.

7. DISCUSSION

The term "simultaneous and selective inference" was repeatedly used by Yosef Hochberg as a synonym for "multiple comparisons" when he delivered the National Science Foundation regional workshop held at Temple University in the summer of 2001. Hochberg attributed the concern about selective inference when faced with multiplicity to an unpublished work by Yosef Putter. Accepting the foregoing point of view, we offer formulation and procedures that address this concern while giving up on simultaneous inference. We argue that in many situations, the selection effect is the more pressing reason why the marginal level of multiple CIs should be adjusted.

Yet this is certainly not always the case. Simultaneous coverage is essential if one wants to be able to, for example, consider functions of all of the parameters. Simultaneous coverage is also needed when an action is to be taken based on the value of all of the parameters. Thus comparing primary endpoints between two treatments in a clinical trial is likely to involve the inspection of all of them, whether they are significantly different or not. This is a clear situation where simultaneous coverage is needed. Looking at a list of secondary endpoints, it is more likely that only significant differences will be relevant. Here the selection of the improved endpoints may be followed by FCR-adjusted CIs, to assess the size of the improvement.

The offering of tools for selective inference allows researchers to judge whether they need simultaneous or selective CIs and choose accordingly. As an example of the confusion that may otherwise arise, let us return in more detail to the study of the failure of preventive hormone therapy in postmenopausal women, as mentioned in Section 1. There were three preselected major outcomes in this study: breast cancer (primary adverse outcome), coronary heart disease (primary outcome), and an index of global outcomes. There were seven other major outcomes, other related outcomes, and composite outcomes (e.g., total cancer). The authors defended using the unadjusted intervals for the three major endpoints by emphasizing that they were designated to serve as such in the monitoring plan. Thus the revealed concern of the researchers is on the effect of selection, not about simultaneous coverage, because preselection does not ensure simultaneous coverage. The foregoing justification for the choice is reiterated in the editorial. If that is the case for the primary outcomes, then it is only natural to assume that the researchers would be satisfied with average coverage for the other outcomes as well. Nevertheless, the researchers did state that the reason why they should report the Bonferroni intervals is because the marginal ones fail to offer simultaneous coverage.

If the researchers could have stated that they are only concerned about the selection effect, then their choice as to what set of intervals to emphasize would have been almost right. For the three preselected parameters, the marginal intervals are appropriate. They also reported *all* intervals for the other (major) outcomes, so the unadjusted intervals give the right coverage. However, they did emphasize significant findings in their discussion, suggesting that using FCR-adjusted intervals is even more appropriate. Using the selective procedure of this article, they should have reported the $1 - .05 \cdot 5/7$ level CIs. Although these CIs are always wider than the marginal intervals, they are closer to the marginal ones than to the Bonferroni-adjusted ones. In retrospect, the researchers were justified in hesitating to use the simultaneous CIs. It may even be argued that although protection against the effect of selection is sufficient for the other outcomes, simultaneous coverage may be needed for the three primary outcomes, one of which is an adverse outcome, because the decision from the trial will ultimately be taken on observing them jointly.

Offering control of FCR rather than simultaneous coverage may run the risk of being misused where stricter control is more appropriate. We do not believe that the response to such a risk should be to always insist on simultaneous coverage as a protection. The danger of such overprotection is that even careful scientists will refrain from following it and use no protection at all, as is currently the case. The decisions as to what statistical criterion best fits the actual problem are admittedly difficult, and we hope that many statisticians will participate in shaping them and will not leave them solely to the users. Similar participation in designing strategies regarding multiple inference in clinical trials has been going on for years, with very productive results.

Here we suggest a modest first step. A practical distinction between situations where simultaneous coverage is needed and

those where selective CIs suffice lies in how the list of unselected parameters is treated. If the identity of the unselected parameters is ignored, not reported, or even set aside in a website, then it is unlikely that simultaneous coverage is needed. These situations indicate that selective coverage should offer sufficient control. In microarray analysis, for example, when searching for those few tens of genes that are differentially expressed among tens of thousands of genes, no one cares about the identity of the undiscovered genes. Nor is the situation different in the QTL analysis discussed earlier. In these cases, reporting the FCR-adjusted selective CIs should go a long way toward addressing the issue of multiplicity. It is quite safe to say that when the size of the problem increases into the hundreds, it is unlikely that the values of *all* of the parameters are needed for the decision making. Although one can find exceptions to the foregoing rule of thumb, it is a reasonable guideline.

Returning to hypothesis testing, some debate has taken place between those advocating the FDR concept and those advocating the pFDR. In the latter, the expectation of the proportion of false discoveries is conditioned on having made some discovery. The pFDR concept, when translated into CIs, is equivalent to the conditional coverage property discussed in Section 1. As shown in Examples 1 and 2, it is impossible to ensure such conditional coverage with either an unadjusted procedure or Bonferroni-selected–Bonferroni-adjusted intervals. In contrast, the FCR that captures the FDR concept for selected CIs can (and should) be controlled. This is a strong argument in favor of using the original FDR. Nevertheless, when m is large, and the proportion of parameters for which CIs are constructed is away from 0, the two concepts are the same, so the Bayesian interpretation offered by Storey (2002) to the pFDR remains relevant to the FDR. When these conditions do not necessarily hold, the FDR concept is the relevant one.

Finally, the problem of inference on the selected set is not unique to frequentist intervals. We believe that if Bayesian-credible CIs are set for all parameters, but only a handful of interesting parameters are selected for reporting, say the ones with posterior modes furthest away from 0, then the current practice of Bayesians to ignore multiplicity is questionable. This discussion removes us far away from our original purpose, and we merely raise it as a question.

[Received October 2002. Revised May 2004.]

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

——— (2000), "On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics," *Journal of Education and Behavioral Statistics*, 25, 60–83.

Benjamini, Y., Hochberg, Y., and Kling, Y. (1993), "False Discovery Rate Control in Pairwise Comparisons," Working Paper 93-2, Tel Aviv University, Dept. of Statistics and Operations Research.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2003), "Adaptive Linear Step-Up Procedures That Control the False Discovery Rate," unpublished manuscript.

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188.

Fletcher, S. W., and Colditz, G. A. (2002), "Failure of Estrogen Plus Progestin Therapy for Prevention," *Journal of the American Medical Association*, 288, 366–369.

Giovannucci, E., Ascherio, A., Rimm, E. B., Stampfer, M. J., Colditz, G. A., and Willett, W. C. (1995), "Intake of Cartenoids and Retinol in Relation to Risk of Prostate Cancer," *Journal of the National Cancer Institute*, 87, 1767–1776.

Lander, E. S., and Kruglyak, L. (1995), "Genetic Dissection of Complex Traits: Guidelines for Interpreting and Reporting Linkage Results," *Nature Genetics*, 11, 241–247.

Mangin, B., Goffinet, B., and Rebai, A. (1994), "Constructing Confidence Intervals for QTL Location," *Genetics*, 138, 1301–1308.

Rossouw, J. E., Anderson, G. L., Prentice, R. L., and LaCroix, A. Z. (2002), "Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative Randomized Controlled Trial," *Journal of the American Medical Association*, 288, 321–333.

Shaffer, J. P. (1995), "Multiple Hypothesis Testing," *Annual Review of Psychology*, 46, 561–584.

——— (2002), "Multiplicity, Directional (Type III) Errors, and the Null Hypothesis," *Psychological Methods*, 7, 356–369.

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Ser. B, 64, 479–498.

Storey, J. D., Taylor, J. E., and Seigmund, D. (2004), "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society*, Ser. B, 66, 187–205.

Tukey, J. W. (1995), "Perspectives on Statistics for Educational Research: Proceedings of a Workshop," eds. V. S. L. Williams, L. V. Jones, and I. Olkin, Technical Report 35, National Institute of Statistical Sciences.

——— (1996), "The Practice of Data Analysis," in *Essays in Honor of J. W. Tukey*, eds. D. R. Brillinger, L. T. Fernholz, and S. Morgentaler, Princeton, NY: Princeton University Press.

Weller, J. I., Song, J. Z., Heyen, D. W., Lewin, H. A., and Ron, M. (1998), "A New Approach to the Problem of Multiple Comparisons in the Genetic Dissection of Complex Traits," *Genetics*, 150, 1699–1706.

Williams, V. S. L., Jones, L. V., and Tukey, J. W. (1999), "Controlling Error in Multiple Comaprisons, With Examples From State-to-State Differences in Education Achievment," *Journal of Educational and Behavioral Statistics*, 24, 42–69.

Comment

Don EDWARDS

In this offering, Benjamini and Yekutieli introduce a new error concept for the construction of multiple confidence intervals (CIs), which they call false coverage-statement rate (FCR) control. FCR is the interval-estimation counterpart to the false discovery rate (FDR) concept for multiple hypothesis tests. When a great many tests are to be done, the FDR (or some alternate form, such as the pFDR mentioned in sec. 7) represents a promising alternative between comparisonwise error (CWE) protection, often considered to be too liberal, and familywise error (FWE) protection, often considered to be too conserv-

Don Edwards is Professor, Department of Statistics, University of South Carolina, Columbia, SC 29208 (E-mail: *edwards@stat.sc.edu*).

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000001943

ative. Such compromises are needed, especially in situations where there are literally thousands of tests to be performed, as in genomics and image processing applications (see, e.g., Efron 2004, combining FDR and empirical Bayes ideas).

The duality between FCR and FDR is in itself an adequate motivation to study FCR (in some form). Unfortunately, the authors attempt to motivate FCR as a cure for selective reporting, the practice of performing many hypothesis tests and then reporting only those that are statistically significant. The authors allege that this is common practice and provide what they consider to be two examples. Without reducing the discussion to nitpicking, and based only on the abstracts of those two articles, I do not believe there is gross abuse in either case. In fact, the second article (Rossouw et al. 2002), does not seem to be an example of selective reporting at all, as those authors report several CIs for hazard ratios that include 1. An accomplished colleague of mine (Robert Best, Professor and Director of the Division of Genetics in South Carolina's Department of Obstetrics and Gynecology) has stated of selective reporting that "I don't believe any researcher worth their salt would do that, and I am reasonably confident that it is not common practice." My own consulting background is primarily with ecological and environmental scientists, who are very aware (at times too much so) of multiple-testing issues.

Not only have the authors not made a convincing argument that selective reporting abuses are commonplace, but it is also questionable whether FCR control as defined here would improve the situation. The error rate under control using a q-FCR procedure as defined by the authors is not the error rate that one would really want to control in a formalized selective inference. According to the authors' definition 2, $FCR = E_{\mathbf{T}}(Q_{CI})$ for

$$Q_{CI} := \begin{cases} V_{CI}/R_{CI} & \text{if } R_{CI} > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where R_{CI} is the number of reported CIs and V_{CI} is the number of reported CIs that do not include their respective parameters. Under this definition, a procedure that never reports any CIs achieves a perfect FCR score. More seriously, a procedure that rarely reports CIs could have a rather high error rate among the reported CIs and still achieve $FCR < q$ for specified q. If selective inference is to be done at all, perhaps two error rates should be studied (after all, these procedures are inherently two-stage procedures): (1) the rate of CIs falsely constructed, $E_{\mathbf{T}}(R_{CI}/m|\theta = 0)$, and (2) among those CIs correctly constructed (i.e., having $\theta \neq 0$), the proportion that do not include θ.

My own preference, however, is not to invent new procedures to encourage researchers' bad habits. Researchers should estimate meaningful quantities, and report objective measures of accuracy for these quantities. Of the articles that the authors discuss, it is at least a little encouraging that both provided interval estimates for *some* of the tested quantities; this shows that the readership of these journals is at least thinking about effect sizes instead of just statistical significance. The next step would be to provide interval estimates for *all* of the examined parameters, even those that are not statistically significant; the Roussouw et al. (2002) abstract does this. After all, a CI of (.99, 1.01) for (say) an odds ratio tells a very different story than the CI (.50, 1.50). Whether these CIs should be simultaneous or if some other correction for "multiple looks" should be made will depend on the setting, but we will not decide that here. Even the most famous statisticians vary widely in their attitudes about this.

ADDITIONAL REFERENCE

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104.

Comment

Juliet Popper SHAFFER

1. INTRODUCTION

Statistical tests concentrate on specific values of a parameter, whereas confidence intervals (CIs) treat all values equally. This leads to different issues in testing and interval estimation even though they are closely related. The article by Benjamini and Yekutieli (BY) offers a new approach to that relationship.

If individual $1 - \alpha$ confidence intervals are calculated for all parameters, then the coverage probability of each interval is $1 - \alpha$, by definition. As BY point out, however, often confidence intervals are calculated only for selected parameters. If the selection is unrelated to the data, the coverage probabilities for the selected intervals are unaffected. Usually, however, the parameters are selected based on the results of a preliminary test. In that case, the coverage probabilities conditional on selection may be very different from the nominal $1 - \alpha$. (BY use the symbol q rather than α; the more familiar α is used here.)

BY give examples of the common practice of testing multiple hypotheses, but calculating CIs only for parameters that differ significantly from 0 at some specified (single or multiple) level α according to some test procedure. In that case it is well known that when the true parameter value is close to 0, the resulting CI has coverage probability less than the nominal probability $1 - \alpha$ (Olshen 1973; Scheffé 1977), or noncoverage probability greater than α. For intervals of fixed size, as in BY's examples, the noncovering CIs are likely to contain values more different from 0 than the true value. For intervals of randomly varying sizes, as in the case of t-test–derived CIs,

Juliet Popper Shaffer is Senior Lecturer Emerita, Department of Statistics, University of California, Berkeley, CA 94720 (E-mail: *shaffer@stat.berkeley.edu*).

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000001952

the noncovering CIs are also likely to be shorter than the expected size. These problems can obviously lead to distortions in decision making that depends on the assumed parameter values. When parameter values are sufficiently far from 0 to lead to almost-certain rejection of the null hypothesis, the CIs will have approximately the correct nominal coverage. In between, the coverage probability can be greater than the nominal, as illustrated in BY's example 1.

As BY point out, there is no way of correcting for these conditional CI distortions, because the amount and direction depend on the unknown true parameter values. The conditional noncoverage probabilities, of course, equal 1 minus the conditional coverage probabilities. BY propose consideration of the noncoverage probabilities rather than the coverage probabilities and of the marginal rather than the conditional noncoverage probabilities of the selected parameter intervals, by which they mean the joint probability of at least one parameter being selected and the noncoverage probabilities of the CIs. BY call the expected proportion of noncovering intervals, or the average joint noncoverage probability, the *false coverage rate* (FCR), because these intervals cover the false but not the true parameter values.

There are many issues related to the type of error control appropriate in testing and in confidence estimation, and whether the approaches should be consistent in the two areas. The testing issues have been discussed extensively, with less discussion in the confidence area, as BY note. The comments here are directed not to these issues, but rather to two other issues: (a) the interpretation of these intervals and (b) their utility given a different type of selection.

2. INTERPRETATION OF BY–PROPOSED CONFIDENCE INTERVALS

To examine the properties of the BY CI procedure, it is illuminating to consider relation between the BY's examples 1 and 3 of marginal CIs for 200 selected parameters, and the conditional CI of a single selected parameter.

Consider BY's example 1. They consider 200 estimates of a parameter θ based on independent test statistics each distributed $N(\theta, 1)$. Five simulations use five different values of θ: 0, .5, 1, 2, and 4, with θ constant within each simulation. Marginal confidence interval coverage proportions (example 1) and noncoverage proportions (example 3) are given for parameters selected by individual normal-theory tests, each at $\alpha = .05$. The latter are the FCR values for this procedure for constructing individual $1 - \alpha$ CIs, ignoring multiplicity. These values are 0, .60, .84, .95, and .97 for coverage and 1.0, .40, .16, .05, and .03 for noncoverage. (Note that the latter values are one minus the former values, to two decimal places, in this application.)

Suppose now that a single test statistic distributed $N(\theta, 1)$ is available, and that a $1 - \alpha$ CI is constructed only if the associated test leads to rejection at level .05. Assume that the true parameter value is either 0, .5, 1, 2, or 4, as in BY's example. Then the conditional probabilities of CI coverage and noncoverage are identical, to two decimal places, to the values given by BY when the number of tests $m = 200$.

The reason for the correspondence is that if 200 independent individual level-α tests are conducted, the probability that at least one parameter is selected (i.e., one test value is significant) when $\theta = 0$ is $1 - .95^{200} = .999965$, and it is higher for nonzero values of θ. Thus in the reported simulations in example 1, selection is a virtual certainty. In that case the conditional noncoverage probability and the marginal noncoverage probability are equal. Because all of the parameter values are the same, this is the probability that a single test is significant, explaining the correspondence with the conditional case for $m = 1$. Yet for $m = 1$, the FCR must be $\leq .05$, according to BY's results.

The explanation for the difference between the conditional CI noncoverage and the marginal noncoverage (FCR) for $m = 1$ points up some problematic aspects of the FCR concept. According to BY's definition of the FCR, the noncoverage of selected parameters is 0 if no parameters are selected. Thus, for example, for $\theta = 0$, the marginal noncoverage probability (FCR) is $(.95)(0) + (.05)(1) = .05$. For $m = 1$, the marginal probabilities of interval noncoverage for $\theta = 0$, .5, 1, 2, and 4 are .05, .03, .03, .03, and .03. For $\theta = 5$ and 6, they both are .05, and they are also .05 for all larger values of θ, because θ would be selected with probability close to 1 for these values, and the conditional and unconditional probabilities of noncoverage would be equal to two decimal places.

The change from considering conditional noncoverage to marginal noncoverage, treating no selections as zero noncoverage, is an ingenious idea—the heart of the article—and clearly related to the similar treatment of the probability of rejecting no hypotheses under the FDR criterion. It is also a somewhat problematic aspect of the procedure. BY point out that it is appropriate to treat the noncoverage probability as 0 when a parameter is not selected, because no CI is constructed. However, suppose that we look at the *coverage* probability for unselected parameters. It would seem just as reasonable to treat the coverage probability as 0 when no CI is constructed, but if it is so treated, then the marginal coverage probability is not 1 minus the marginal noncoverage probability. For $m = 1$, these marginal coverage probabilities for $\theta = 0$, .5, 1, 2, 4, 5, and 6 are 0, .05, .14, .49, .95, .95, and .95.

In fact, the coverage probability for unselected parameters must be implicitly treated as unity for the two marginal probabilities to be consistent. So the BY criterion represents avoidence of error, but not success. It is hard to think in this way; note the comment of Storey, Taylor, and Siegmund (2004, p. 1) that "this is useful in exploratory analyses, where *we are more concerned with having mostly true findings among several, rather than guarding against one or more false-positive results.*"

If the probability of no selection is very small, then this difficulty in interpretation is not a problem. In that case, marginal and conditional probabilities are approximately equal. This is the case in problems involving large numbers of tests, where it can be safely assumed that many hypotheses are false, as in microarray and wavelet analyses. Analyses in those areas almost always result in some rejections, no matter which multiple-testing procedures are used. These are the areas in which this article makes a real contribution. A number of adaptive procedures have been proposed based on the FDR and the pFDR (Storey 2002), which is approximately equivalent to the FDR criterion under these circumstances. (For recent work on adaptive procedures, see Black 2004; Cox and Wong 2004; Storey et al. 2004.)

3. OTHER TYPES OF SELECTION

Sometimes, usually in studies testing a small number of hypotheses, there is an interest in CIs for the accepted hypotheses, or separately for the accepted and rejected hypotheses. There has been a concern, especially in the social sciences, with the inadequate power of most research, stemming from the work of Cohen (1962), who was the first to estimate the typical power of psychological research studies. In these cases, often the magnitude of a departure from the null value is not of special interest, as long as the null is rejected, and CIs are calculated for accepted hypotheses to give an indication of the range of plausible parameter values, in view of the presumed low power due to practical constraints. Given acceptance, parameter values close to the null are more likely to be included in the interval than when the null hypothesis is rejected, and the CIs have conditional coverage probabilities greater than the nominal probability $1 - \alpha$ or noncoverage probabilities less than α, whereas for parameter values far from the null, the conditional confidence coverage probabilities approach 0 and noncoverage probabilities approach 1. For $m = 1$, with the test statistic distributed $N(\theta, 1)$ and for $\theta = 0, .5, 1, 2, 4, 5$, and 6, when CIs are calculated only if the hypothesis is accepted, the conditional noncoverage probabilities are 0, .02, .03, .05, 1, 1, and 1, whereas the marginal noncoverage probabilities are 0, .02, .02, .02, .02, 0, and 0. These values suggest that the FCR approach may be less useful for parameters selected when hypotheses are not rejected, because the CIs may be too wide for useful inferences.

4. CONCLUSIONS

Although it is a good idea for researchers to be aware of problems with conditional coverage of CIs, there is not much that can be done to address them. Because the true values of the relevant parameters are unknown, there is no way of adjusting for the conditional coverage probabilities of the associated CIs, given parameters selected on the basis of the data. What BY show is that the joint probability of some parameters being selected and parameter noncoverage rates can be controlled at a level smaller than a specified α, for independent tests and some types of positively dependent tests, regardless of the selection method used. A simple method of guaranteeing this maximum noncoverage probability, the FCR, is to test the selected hypotheses at level $R\alpha/m$, where R is the number selected and m is the total number, although improvements are possible using adaptive methods. This approach is useful when there are large numbers of hypotheses and many hypotheses are expected to be false, with CIs desired for rejected hypotheses. The article makes a valuable contribution to analysis in such situations. The methodology appears to be less useful with small numbers of hypotheses and in studies with low power to reject any hypotheses.

ADDITIONAL REFERENCES

Black, M. A. (2004), "A Note on the Adaptive Control of False Discovery Rates," *Journal of the Royal Statistical Society*, Ser. B, 66, 297–304.

Cohen, J. (1962), "The Statistical Power of Abnormal Social Psychological Research: A Review," *Journal of Abnormal and Social Psychology*, 65, 145–153.

Cox, D. R., and Wong, M. Y. (2004), "A Simple Procedure for the Selection of Significant Effects," *Journal of the Royal Statistical Society*, Ser. B, 66, 395–400.

Olshen, R. A. (1973), "The Conditional Level of the *F*-Test," *Journal of the American Statistical Association*, 68, 692–698.

Scheffé, H. (1977), "A Note on a Reformulation of the *S*-Method of Multiple Comparison," *Journal of the American Statistical Association*, 72, 143–146.

Comment

Ajit C. TAMHANE

I congratulate the authors for providing a solution to the vexing problem of constructing multiple confidence intervals (CIs) with controlled error rate for parameters selected by a multiple-testing procedure. There are a number of new important ideas in the article, a thorough discussion of which would require much additional work. I am sure that there will be many follow-up articles that will explore these ideas in detail; here I restrict my comments to only a few basic points.

The authors begin by demonstrating that unadjusted and Bonferroni-adjusted procedures do not ensure prescribed conditional coverage probability if CIs are computed only for those means for which the null hypothesis that the mean equals 0 is rejected (the so-called "discoveries"). For each such discovery, the set of "acceptable" values of the mean is used as its CI, which is therefore dual to the corresponding significance test; in particular, it excludes 0. This obviously makes the conditional coverage probability equal to 0 when the null hypothesis holds. For small nonzero means, the conditional coverage probability still falls below the nominal confidence level. One reason for this phenomenon is that the estimates of the selected means are highly biased (except when the true mean is 0, in which case the estimate is unbiased). As a result, the intervals are incorrectly centered at these biased estimates. Would it be possible to use shrinkage estimates instead, although the resulting intervals will not be duals of the corresponding significance tests?

To give an idea of the bias involved in selected means, consider independent $T_j \sim N(\theta_j, 1)$, $j = 1, 2, \ldots, m$. A "nominal" $(1 - \alpha)$ marginal or simultaneous CI, $T_j \pm c$, for θ_j is computed conditional on an α-level test of $\theta_j = 0$ rejecting when

Ajit C. Tamhane is Professor and Chairman, Department of Industrial Engineering and Management Sciences (IE/MS) and Professor of Statistics, Northwestern University, Evanston, IL 60208 (E-mail: *ajit@iems.northwestern.edu*). The author thanks Dingxi Qiu, a graduate student in the IE/MS Department, for providing computational help and useful comments.

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000001925

Table 1. Bias in T_j Conditional on $|T_j| > c$ for $\alpha = .05$

θ	Unadjusted test	Bonferroni-adjusted test
.5	1.4927	3.2798
1.0	1.4503	2.9680
2.0	.7722	2.0778
4.0	.0509	.5961

NOTE: The unadjusted test uses $c = Z_{.975} = 1.96$, whereas the Bonferroni-adjusted test uses $c = Z_{.999875} = 3.6623$.

$|T_j| > c$. Here $c = Z_{1-\alpha/2}$ for an unadjusted test coupled with a marginal CI and $c = Z_{1-\alpha/2m}$ for the Bonferroni-adjusted test coupled with a simultaneous CI. Assume that $\theta_j = \theta$ for all $j = 1, 2, \ldots, m$. It is easily shown that the conditional expectation of T_j, conditioned on $|T_j| > c$, is given by

$$E(T_j||T_j| > c) = \frac{\theta - \int_{-c}^{c} t\phi(t-\theta)\,dt}{\Phi(\theta - c) - \Phi(-\theta - c)} = \theta + \frac{\phi(\theta - c) - \phi(-\theta - c)}{\Phi(\theta - c) + \Phi(-\theta - c)},$$

where ϕ and Φ are the pdf and cdf of the standard normal distribution. The second term gives the bias, which has the same sign as θ. Table 1 gives the bias values for selected θ for both unadjusted and Bonferroni-adjusted procedures when $\alpha = .05$ and $m = 200$. We see that the bias is quite large for small values of θ and decreases with θ.

Some readers may be confused, as indeed I was, by the fact that the estimated FCRs for the unadjusted procedure in example 3 equal exactly 1 minus the corresponding conditional coverage probabilities from example 1 (in particular, the FCR equals 1 when $\theta = 0$), whereas this relation does not hold (in particular, the FCR does not equal 1, but equals .05 when $\theta = 0$) for the Bonferroni-adjusted procedure in example 4. The reason for this is that the ratio V_{CI}/R_{CI} is defined as 0 when $R_{CI} = 0$; hence the FCR can be expressed as

$$FCR = E\left(\frac{V_{CI}}{R_{CI}}\middle| R_{CI} > 0\right)P(R_{CI} > 0).$$

If $R_{CI} > 0$ when $\theta = 0$, then $V_{CI}/R_{CI} \equiv 1$ for both the unadjusted and Bonferroni-adjusted procedures. Therefore, $FCR = P(R_{CI} > 0)$. For the unadjusted procedure,

$$P(R_{CI} > 0) = 1 - (.95)^{200} \approx 1,$$

and hence $FCR \approx 1$. In contrast, for the Bonferroni-adjusted procedure,

$$P(R_{CI} > 0) = 1 - (.99975)^{200} \approx .05,$$

and hence $FCR \approx .05$.

The foregoing explanation demonstrates that the FCR is controlled for the Bonferroni-adjusted procedure at the .05 level even for $\theta = 0$, because CIs are computed in only 5% of the cases, although all of them miss the true means. To me, this does not provide the necessary security about the accuracy of the CIs, and suggests that the positive FCR,

$$pFCR = E\left(\frac{V_{CI}}{R_{CI}}\middle| R_{CI} > 0\right),$$

may be a more appropriate criterion. I recognize, as the authors note, that the pFCR is equivalent to the conditional coverage probability and cannot be controlled for all parameter values. However, there are other criteria that could be used instead. In summary, I think that the debate on the choice between

$$\frac{FDR}{FCR} \quad \text{versus} \quad \frac{pFDR}{pFCR}$$

is far from over.

As an aside, I note that it is not necessary to estimate the quantities in examples 1–4 by simulation, because the following exact expressions for them can be readily derived. First, the conditional coverage probability is given by

$$P(\theta \in [T_j - c, T_j + c]||T_j| > c) = \frac{\Phi[\min(c, \theta - c)] - \Phi(-c)}{\Phi(\theta - c) + \Phi(-\theta - c)}.$$

Next, the FCR is given by

$$\begin{aligned} FCR &= E\left(\frac{V_{CI}}{R_{CI}}\middle| R_{CI} > 0\right)P(R_{CI} > 0) \\ &= P(\theta \notin [T_j - c, T_j + c]||T_j| > c) \\ &\quad \times \{1 - [P\{-c \le T_j \le c\}]^m\} \\ &= \left\{1 - \frac{\Phi[\min(c, \theta - c)] - \Phi(-c)}{\Phi(\theta - c) + \Phi(-\theta - c)}\right\} \\ &\quad \times \{1 - [\Phi(-\theta + c) - \Phi(-\theta - c)]^m\}. \end{aligned}$$

This last expression holds only when $\theta_j = \theta$ for all $j = 1, 2, \ldots, m$.

In closing, I congratulate the authors once again for a thought-provoking article, and I thank the editor for giving me an opportunity for contributing to its discussion.

Comment

Peter H. WESTFALL

1. INTRODUCTION

Benjamini and Yekutieli (BY) solve important problems in false discovery rate–controlling multiple-comparison procedures (FDRMCPs), thus increasing their utility and applicability. Familywise error rate–controlling multiple-comparison procedures (FWEMCPs) have historically been interval-based

Peter H. Westfall is Horn Professor of Statistics, Department of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, TX 79409 (E-mail: *peter.westfall@ttu.edu*).

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000001916

as often as testing-based. FDRMCPs, on the other hand, have been exclusively testing-based; BY fill an important gap by providing confidence intervals (CIs).

Mathematical development of FDRMCPs is often more difficult than that of FWEMCPs; thus it is surprising that problems still unsolved in FWEMCPs—control of directional errors with stepwise FWEMCPs and CI correspondence with stepwise testing-based FWEMCPs—should have such a simple solution in the false discovery rate (FDR) case. I think that BY's results concerning directional FDR control are both interesting and useful. I also think the false coverage-statement rate (FCR) CIs are interesting, but reserve judgment as to their utility. My first comments concern optimality properties of BY's directional determinations. I then offer critiques of FCR intervals in terms of practical interpretations, empirical and theoretical comparisons with interval-based FWEMCPs, Bayesian correspondence, and regression to the mean after selection.

2. DECISION THEORY: CLASSIFYING DIRECTIONS

FDRMCPs adapt to underlying structure of the data; this is clearly their strength. I now show that the directional FDR adapts to the data to produce a nearly optimal decision rule for classifying signs of the parameters.

2.1 Approximate Critical Values for Test Statistics

Suppose, as in BY's examples, that $T_j|\theta_j \sim \mathrm{N}(\theta_j, 1)$. If two-sided tests are performed for each hypothesis, then the FDR critical value for $H_{(i)}$ is $Z_{1-iq/2m}$. Further, if the θ_j are a random sample from a distribution F_θ, then each T_j is (marginally) a sample from the convolution distribution $F_{T+\theta}$; for example, if $F_\theta = \mathrm{N}(0, \sigma_\theta^2)$, then $F_{T+\theta} = \mathrm{N}(0, \sigma_\theta^2 + 1)$. Assuming that $T_j|\theta_j$ are independent, $T_1, \ldots, T_m$ is a random sample from $F_{T+\theta}$.

As noted by Genovese and Wasserman (2002), the critical threshold q^* for p values using FDR is asymptotically the crossing point of two cumulative p value distributions; translating this theory from p values to Z values gives the critical values of BY's FCR procedure. As $m \to \infty$, the cumulative distribution of the critical values $Z_{1-iq/2m}$ converges deterministically to $F_{FCR}(z) = 0$, $2\{1 - \Phi(z)\} > q$ and to $1 - 2\{1 - \Phi(z)\}/q$ otherwise. By the Glivenko–Cantelli theorem, the cdf of $|T_j|$ converges a.s. to $F_{|T|}(z) = 1 - 2\{1 - \Phi(z\sqrt{1-\rho^2})\}$, where $\rho^2 = \sigma_\theta^2/(1+\sigma_\theta^2)$. The FCR critical value is obtained by solving $F_{FCR}(z) = F_{|T|}(z)$ for Z_{FCR}. Using $1 - \Phi(z) \approx \phi(z)/z$ and $\ln(q) \approx \ln\{2\phi(Z_{1-q/2})/Z_{1-q/2}\}$, we have

$$Z_{FCR}^2 \approx (1/\rho^2)\{Z_{1-q/2}^2 + \ln(Z_{1-q/2}^2) + \ln(\pi/2) + \ln(1-\rho^2)\}. \quad (1)$$

(When $\rho^2 = .5$, this approximation yields $Z_{FCR} \approx 3.15$ for $q = .05$; the actual value is $Z_{FCR} = 3.29$.) Thus Z_{FCR} adapts to ρ^2 in a very natural way; with smaller ρ^2, T_j is not as reliable an estimate of θ_j, and more caution is needed when determining θ_j's sign, thereby justifying the larger critical value Z_{FCR}. What is especially attractive about Z_{FCR} is that ρ^2 need not be prespecified; it is determined implicitly from the data.

2.2 Decision Theory Correspondence

Recently, Shaffer (1999) and Lewis and Thayer (2004; hereafter denoted by LT) developed decision-theoretic multiple-comparision procedures and showed that FDRMCPs correspond in various ways. BY find that directional errors are controlled using FDR methods; directional determination is fundamentally a problem in decision theory. This section shows that BY's directional determination corresponds well with the optimal procedure of LT.

Consider again $T_j|\theta_j \sim \mathrm{N}(\theta_j, 1)$, with $\theta_j \sim \mathrm{N}(0, \sigma_\theta^2)$, with the goal of classifying each θ_j as either above 0 (AZ), below 0 (BZ), or not significantly different from 0 (NZ). LT suggested loss functions $L_{AZ}(\theta) = 1$, for $\theta < 0$ and 0 otherwise; $L_{BZ}(\theta) = 1$ for $\theta > 0$ and 0 otherwise; and $L_{NZ}(\theta) = A$ for $\theta \neq 0$ and 0 otherwise. In the framework of Waller and Duncan (1969), $1/A$ is the "k ratio" measuring the severity of type I error relative to type II error. The resulting optimal decision rule classifies θ_j as AZ if $T_j > (1/\rho)Z_{1-A}$, as BZ if $T_j < -(1/\rho)Z_{1-A}$, and as NZ otherwise. Comparing (1) with the optimal squared critical point $Z_{OPT}^2 = (1/\rho^2)Z_{1-A}^2$ shows that BY's procedure is nearly optimal. The implied k-ratio can be determined as a function of ρ^2 by equating $(1/\rho)Z_{1-A}$ to Z_{FCR} and solving for $1/A$. For $\rho^2 = 1/6$, $1/3$, $1/2$, $2/3$, and $5/6$, the k-ratios implicit in BY's procedure are 126, 113, 100, 86, and 70. It is reassuring that the implied k-ratios of BY's procedure are near the commonly used default value 100.

3. PRACTICAL INTERPRETATION OF THE FALSE COVERAGE–STATEMENT RATE CRITERION

Objecting to CIs is like objecting to motherhood and apple pie. However, when the goal is to identify a small percentage (of a large m) of important effects for further experimentation, CIs can be unnecessary baggage. CIs are more interesting with small m where investigators carefully scrutinize the values of a few parameters, such as effects on multiple endpoints in a clinical trial. But in small m cases, the FCR criterion is difficult to interpret.

Suppose that a client has a study in which five such CIs are selected and constructed using definition 1 of BY, with $q = .05$. The client asks whether the inferences of the selected intervals are correct. To answer, the statistician asks the client to imagine the following:

1. Her study is one of a sequence $S_1, S_2, S_3, S_4, S_5, \ldots$ of studies; hers is S_i.
2. Each study results in a number of intervals that are selected for scrutiny, for example, $10, 0, 0, 4, 23, \ldots$ in studies $S_1, S_2, S_3, S_4, S_5, \ldots$ (five are selected in S_i).
3. In each study, there will be a certain number of intervals that do not contain their respective parameter values, say $1, 0, 0, 0, 4, \ldots$; these numbers are unknown.
4. The FCR, $q = .05$, that applies to the client's study S_i, is an upper bound on the long-run average of the values $1/10, 0, 0, 0, 4/23, \ldots$.

At this stage, the client is undoubtedly baffled as to what all this has to do with her study! Interval-based FWEMCPs more clearly pertain to the client's study: all intervals in her study are correct unless her study is statistically rare.

4. COMPARING FALSE COVERAGE–STATEMENT RATE WITH INTERVAL–BASED FAMILYWISE ERROR RATE–CONTROLLING MULTIPLE COMPARISON PROCEDURES: A STUDY WITH ELECTROENCEPHALOGRAM FUNCTIONAL DATA

In this section I compare FCR with familywise error (FWE) intervals, offering two more critiques in the context of a real example: (1) intervals that account for correlations are readily available for FWE control, not so for FCR control, and (2) FCR inferences can possibly mislead compared to FWE inferences, even with very large m.

An experiment by Dr. Rockefeller Young of the Texas Tech University Health Sciences Center involved locating the portion of the brain responsible for distinguishing color. The five treatments were green light at 60%, 80%, and 100% intensities and red light at 90% and 100% intensities. The goal was to compare red100% versus green100%; comparisons between intensities were needed to establish the sensitivity of the experiment. Electroencephalogram (EEG) data were collected using 43 time series responses (~2 milliseconds apart) with electrodes at 62 scalp locations. The experiment was repeated 70 times per treatment group, yielding 350 independent response vectors, $\mathbf{Y}_{ij}$, $i = 1, \ldots, 5$, $j = 1, \ldots, 70$, each containing $43 \times 62 = 2{,}666$ spatiotemporal EEG responses (data provided on request). A model is $\mathbf{Y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij}$, where the $\boldsymbol{\varepsilon}_{ij}$ are iid with mean $\mathbf{0}$ and unstructured $(2{,}666 \times 2{,}666)$-dimensional covariance matrix $\boldsymbol{\Sigma}$. Simultaneous CIs for all components of $\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}$, $1 \le i < i' \le 5$, entail $m = 10 \times 2{,}666 = 26{,}660$ comparisons. FWE-controlling methods accounting for spatiotemporal correlations as well as nonnormal distributional characteristics are readily available using the "maxT" method (e.g., Dudoit, Shaffer, and Boldrick 2003). The single-step maxT method adjusts p values for testing H_l using $\tilde{p}_l = P(\max_{1 \le j \le 26{,}660} T_j \ge t_l | H_0)$, calculated where the T_j are absolute values of ANOVA-based test statistics, and using bootstrap sampling of residual vectors $\hat{\boldsymbol{\varepsilon}}_{ij} = \mathbf{Y}_{ij} - \hat{\boldsymbol{\mu}}_i$ to estimate the $\tilde{p}_l$. Critical values for simultaneous CIs follow simply as $t_q^{\mathrm{BOOT}} = \min(t_l : \tilde{p}_l \le q)$; using these, we have the approximate $100(1-q)\%$ simultaneous intervals $\hat{\mu}_{ik} - \hat{\mu}_{i'k} \pm t_q^{\mathrm{BOOT}} \mathrm{s.e.}(\hat{\mu}_{ik} - \hat{\mu}_{i'k})$ for all i, i' and $1 \le k \le 2{,}666$. Westfall and Young (1993, pp. 125–126) provided details for these CIs and directional error control. Troendle, Korn, and McShane (2004) noted that whereas some bootstrap methods fail in high-dimensional cases, the Westfall–Young method works reasonably well.

Using PROC MULTTEST of SAS/STAT, $t_{.05}^{\mathrm{BOOT}} = 4.53$; the Bonferroni critical value is the $1 - .05/(2 \times 26{,}660)$ quantile of the t_{350-5} distribution, $t_{.05}^{\mathrm{BON}} = 4.85$. Thus the bootstrap maxT method incorporates correlations and is less conservative. For these data, there are 5,397 FDR significances at $q = .05$, so the FCR critical value is the $1 - (5{,}397/26{,}660).05/2$ quantile of the t_{350-5} distribution, $t_{.05}^{\mathrm{FCR}} = 2.585$. Presumably, $t_{.05}^{\mathrm{FCR}}$ would be reduced if correlation were accommodated.

Figure 1 shows the results of the bootstrap FWE and FCR simultaneous confidence bands for red100% versus green100% treatment differences at scalp location 35. The FWE interpretation is clear; all intervals are correct for this study unless the study itself is unusual. Although the FCR intervals show significance for early time points, there is concern that they are errors;

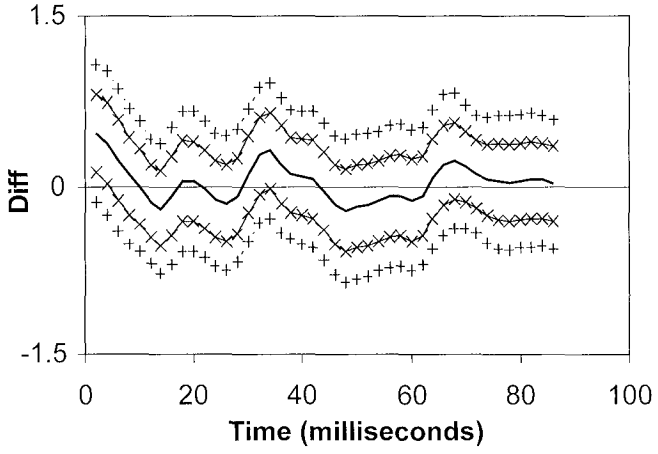

Figure 1. Estimated red100% versus green100% Differences Over Time (——) for Scalp Location 35, Along With Simultaneous 95% FCR Intervals (×'s) and Simultaneous 95% FWE Intervals (+'s), Among $m = 26{,}660$ Inferences.

Figure 2(a) shows that there are strong effects of intensity (here red90% vs. red100%), but the red100% versus green100% differences are essentially null [Fig. 2(b)]. Thus it appears that the appropriate conclusion from this example is obtained using FWE; FCR is too liberal. Indeed, scientific theory suggests that there should be no difference in red100% versus green100% (Young, personal communication, 2004).

This example might be considered somewhat unfair, in that the data should be collapsed over spatiotemporal dimensions, perhaps using principal components. Nevertheless, the current climate of data mining, with the increasing use of FDRMCPs for such applications, makes the example relevant.

5. BAYESIAN CORRESPONDENCES

Methods that have reasonable interpretations from several perspectives are most likely to be considered useful. Bayesian correspondences of testing-based FWEMCPs exist and have been discussed by Jeffreys (1961, pp. 253–255) and Westfall, Johnson, and Utts (1997), but they require strong prior assumptions. Testing-based FDRMCPs have a nice correspondence with empirical Bayes methods, requiring weaker prior assumptions (Efron, Tibshirani, Storey, and Tusher 2001).

Interval-based FWEMCPs have a straightforward Bayesian connection through improper priors. For instance, in the normal homoscedastic linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the pivotal vector $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\hat{\sigma}$ has the same multivariate t-distribution whether considered from frequentist or from improper Bayesian standpoints; thus the exact FWE-controlling CIs defined by $I_j(\hat{\boldsymbol{\beta}}, \hat{\sigma}) = \{\mathbf{c}_j'\hat{\boldsymbol{\beta}} \pm t_q^{\mathrm{EXACT}} \mathrm{s.e.}(\mathbf{c}_j'\hat{\boldsymbol{\beta}})\}$ have a simple Bayesian correspondence, $P(\mathbf{c}_j'\boldsymbol{\beta} \in I_j(\hat{\boldsymbol{\beta}}, \hat{\sigma}),$ for all $j | \mathbf{Y}) = 1 - q$. Further details and software (both Bayesian and frequentist) have been given by Westfall, Tobias, Rom, Wolfinger, and Hochberg (1999). Unlike FCR CIs, the frequentist confidence statement is exact without requiring independence or conditions on the θ_j. Thus, relative disadvantages of FCR CIs are inexactness and lack of Bayesian correspondence.

With proper priors, there can be no correspondence, because the parameter estimates themselves must be shrunk toward the prior mean. Suppose again that $T_j | \theta_j \sim \mathrm{N}(\theta_j, 1)$, with

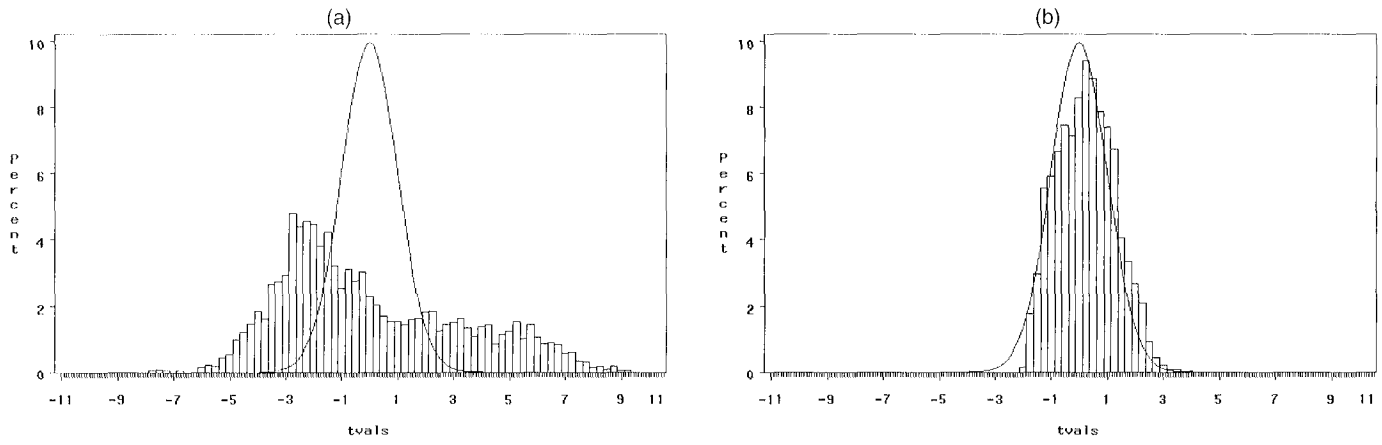

Figure 2. Histograms of t-Statistics for Comparisons at 2,666 Spatiotemporal Locations, With N(0, 1) Overlay. (a) t-statistics comparing red90% versus red100%. (b) t-statistics comparing red100% versus green100%.

$\theta_j \sim \mathrm{N}(0, \sigma_\theta^2)$, a proper prior. The need for shrinkage is demonstrated by $\theta_j|T_j \sim \mathrm{N}(\rho^2 T_j, \rho^2)$. Figure 3 shows the standardized pivots $\mathrm{sign}(T_j) \times (\theta_j - T_j)/1$ and $\mathrm{sign}(T_j) \times (\theta_j - \hat{\rho}^2 T_j)/\hat{\rho}$ for FDR-selected intervals in 400 simulated studies, each having $m = 2{,}000$ and $\rho^2 = .5$. The former pivot tells us roughly where we can locate θ_j within an FCR interval after FDR selection; the latter, where we can locate θ_j within an empirical Bayes interval [using $\hat{\rho}^2 = \hat{\sigma}_\theta^2/(1 + \hat{\sigma}_\theta^2)$, where $\hat{\sigma}_\theta^2$ is a method of moments estimate] after FDR selection. When $T_j > 0$, the value of θ_j tends to lie to the left of T_j in the selected FCR intervals; conversely, when $T_j < 0$, θ_j tends to lie to its right. This is an illustration of regression to the mean after selection. In contrast, θ_j is properly centered within the selected empirical Bayes intervals, as shown in Figure 3(b).

Although regression to the mean following selection also affects FWE-controlling CIs, the problem is more directly relevant with FCR-controlling CIs, because selection is their primary motivation.

6. CONCLUSION

The directional determinations shown by BY are quite useful, with approximate decision-theoretic optimality. But are their FCR-controlling intervals viable? For the reasons stated herein, I am not yet convinced.

ADDITIONAL REFERENCES

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Genovese, C. R., and Wasserman, L. (2002), "Operating Characteristics and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society*, Ser. B, 64, 499–518.

Jeffreys, H. (1961), *Theory of Probability*, Oxford, U.K.: Oxford University Press.

Lewis, C., and Thayer, D. T. (2004), "A Loss Function Related to the FDR for Random Effects Multiple Comparisons," *Journal of Statistical Planning and Inference*, 125, 49–58.

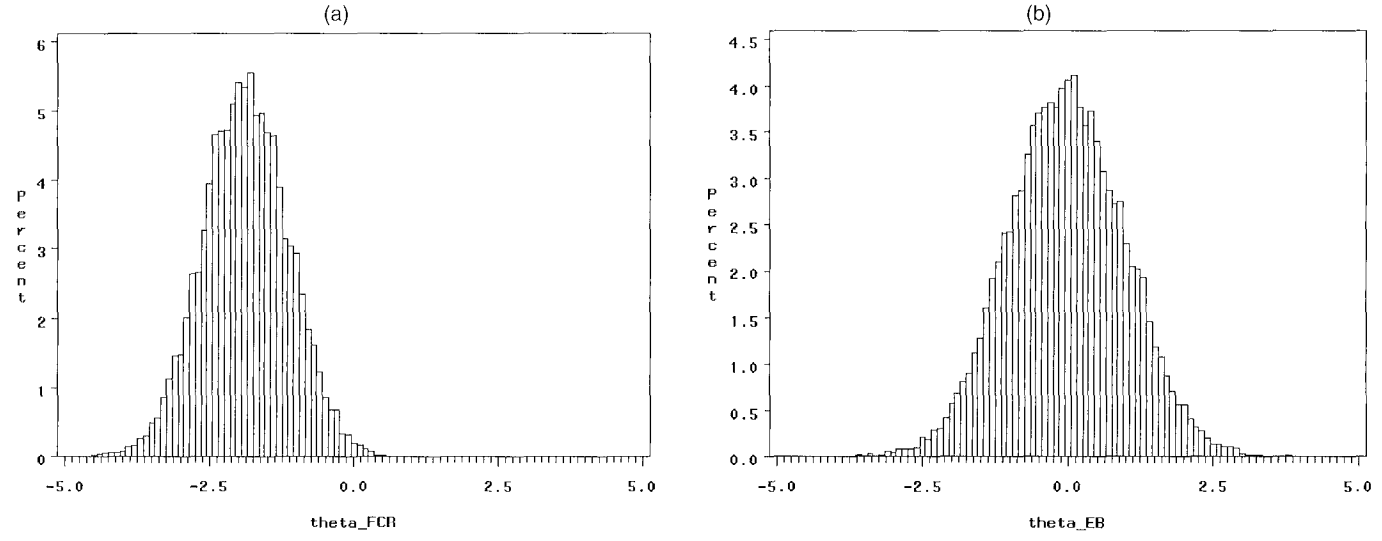

Figure 3. Histograms of Pivots $\mathrm{sign}(T_j) \times (\theta_j - T_j)/1$ Corresponding to FCR Intervals (a) and Pivots $\mathrm{sign}(T_j) \times (\theta_j - \hat{\rho}^2 T_j)/\hat{\rho}$ Corresponding to Empirical Bayes Intervals (b). The 0 of the horizontal axis indicates that θ_j lies in the center of the confidence interval; values to the left of 0 indicate that θ_j tends to be closer to 0 than its estimate T_j.

Shaffer, J. P. (1999), "A Semi-Bayesian Study of Duncan's Bayesian Multiple Comparison Procedure," *Journal of Statistical Planning and Inference*, 82, 197–213.

Troendle, J. F., Korn, E. L., and McShane, L. M. (2004), "An Example of Slow Bootstrap Convergence in High Dimensions," *The American Statistician*, 58, 25–29.

Waller, R. A., and Duncan, D. B. (1969), "A Bayes Rule for the Symmetric Multiple Comparisons Problem," *Journal of the American Statistical Association*, 64, 1484–1503.

Westfall, P. H., Johnson, W. O., and Utts, J. M. (1997), "A Bayesian Perspective on the Bonferroni Adjustment," *Biometrika*, 84, 419–427.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., and Hochberg, Y. (1999), *Multiple Comparisons and Multiple Tests Using SAS®*, Cary, NC: SAS Institute Inc.

Westfall, P. H., and Young, S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*, New York: Wiley.

Young, R. (2004), Personal communication.

Comment

Burt HOLLAND

1. INTRODUCTION

Benjamini and Hochberg's (1995) promotion of the false discovery rate (FDR) error criterion and the Benjamini–Hochberg (BH) testing procedure for controlling FDR, as well as subsequent modifications of the BH procedure by these and other authors, arrived along with the advent of massive datasets resulting from data mining, including those resulting from experiments involving gene expression microarrays, brain imaging, and so on. As Benjamini and Yekutieli point out, although in most of the literature adequate attention is routinely given to multiplicity in testing, the control of multiplicity is commonly ignored when reporting simultaneous confidence intervals (CIs) except for special cases, such as the comparisons of all pairs among a small number of population means.

Most multiple-testing procedures developed over the past 20 years are conducted in a stepwise fashion. As such, it is difficult to invert or dualize them to analogous simultaneous CIs. A further complication that arises when inverting the BH procedure is its control of the FDR rather than the more familiar familywise error rate.

The inversion of the BH procedure to simultaneous CI error control was unsuccessfully attempted by Tukey (1995, 1996), a leading pioneer developer of multiple-comparison procedures. On this basis, the successful inversion of the BH procedure by Benjamini and Yekutieli is a substantial technical achievement, as well as an important contribution to scientific inquiry. In addition, this new procedure for simultaneous CIs is as easy to implement and understand as the analogous BH FDR-controlling testing procedure.

The authors' key insight into developing the inversion is the segregation of the parameters under consideration into two groups: those *selected* for study and those set aside. The examples that they report in section 1 clearly demonstrate the requirement that in many contexts, analysts should consider and justify a distinction between selected and unselected parameters. Indeed, it is surprising that this dichotomization apparently has not been proposed prior to this article. The authors convincingly argue that in situations where simultaneous coverage is impractical and there are parameters that can be classified as unselected, the false coverage-statement rate (FCR)-adjusted control discussed here is strongly preferred over unadjusted intervals.

Burt Holland is Professor of Statistics, Temple University, Philadelphia, PA 19122 (E-mail: *bholland@temple.edu*).

2. SOME HISTORY AND ITS IMPLICATION

Recently, the Institute for Scientific Information (2004) reported that "*JASA* was the most highly cited journal in the mathematical sciences in 1991–2001, with 16,457 citations, more than 50% more than the next most highly cited journals." The article by Benjamini and Hochberg (1995) did not appear in *JASA*, to which I understand it was initially submitted. If it had been published in *JASA*, the application of FDR and the BH procedure might not have been delayed. In the early 1990s, relatively few statisticians anticipated today's prevalence of massive datasets for which familywise error control is impractical. Initial resistance to the use and control of FDR rather than the familywise error rate was evidently caused by a concern that casual investigators would invoke the BH procedure rather than a familywise error rate controlling procedure to rationalize additional hypothesis rejections (discoveries) attainable with the BH procedure.

Figure 1 plots the total annual number of citations of Benjamini and Hochberg (1995) in the *Science Citation Index* and *Social Science Citation Index* for publications appearing during the years 1996–2003. Citations in the first several months of 2004 have exceeded the annual rate attained in 2003. This figure illustrates an initial resistance to the BH procedure followed by its eventual widespread approval.

The first discussion of FDR in *JASA* did not occur until the article by Efron, Tibshirani, Storey, and Tusher (2001). I believe that this article gave FDR an imprimatur and respectability that it did not previously have, and has driven the burgeoning use depicted in Figure 1.

By appearing in *JASA*, the present article should quickly receive the attention and dissemination it deserves, and avoid the delayed recognition and implementation of its dual, the article by Benjamini and Hochberg (1995).

3. ADDITIONAL NEEDED WORK

Three results in this article discuss conditions for the joint distribution of the pivotal statistics under which the FCR-adjusted selective CIs dualized from the BH procedure may be used. Theorem 1 says that in the case of independent statistics, these confidence intervals control the FCR below a designated

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000001934

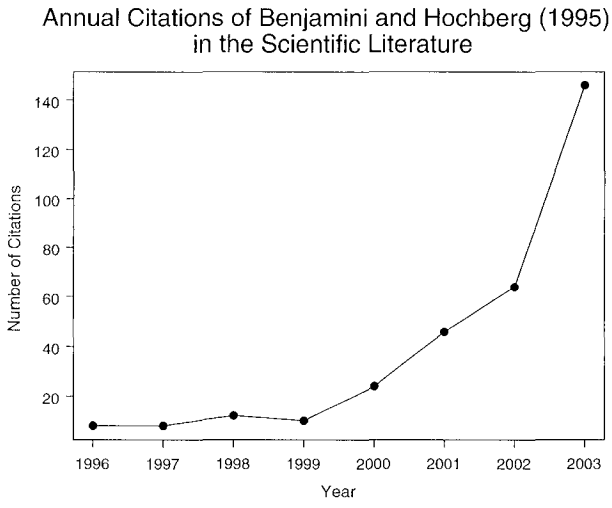

Figure 1. Number of Citations Appearing in the Science Citation Index and Social Science Citation Index.

value q. Theorem 4 indicates that if the dependence structure of the joint distribution is unknown, FCR control is still maintained but at a level somewhat higher than the FDR control q of the analogous BH multiple-testing procedure, where the level increases with the number of parameters under consideration. This may lead to unattractively wide confidence intervals. In practice, the most useful result is likely to be theorem 3, which says that under mild additional conditions, FCR control is maintained below q if the joint distribution of the pivotal statistics meets the positive regression dependent on a subset (PRDS) condition. The circumstances in which FCR control will be used will typically be the same as when FDR control for multiple testing has become widely accepted—problems with a large to massive number of simultaneous inferences. Storey et al. (2004) suggested that the typical joint distribution of the many pivotal statistics in these situations can be characterized as *weak dependence*, a condition that these authors formally defined. To facilitate comfortable use of FCR control, it is desirable to develop linkages between this weak dependence notion with the PRDS condition. A useful reference for this task is the article by Sarkar (2002), where the PRDS condition is discussed in detail.

In general, users of FDR testing methodology should be required to address two issues before applying it to their investigation:

- Justify their choice of the family of m related inferences.
- Clarify why FDR is a more appropriate error control concept than familywise error, based on the definitions of these criteria.

Instances of misuse of the FDR methodology occasionally arise. Consider, for example, the conclusion of Thissen, Steinberg, and Kuang (2002):

> Given its easy implementation, it is feasible to include the BH procedure in introductory instruction in inferential statistics, augmenting or replacing the Bonferroni technique. Students trained with this more powerful technique should be less likely to use the nearly powerless Bonferroni procedure, or to eschew correction for multiple comparisons entirely, due to a perceived loss of power.

Thissen et al. (2002) provided no rationale other than power for using FDR rather than familywise error as an error criterion. If power is the only standard for multiple testing, it follows from their conclusion that investigators should reject each and every hypothesis they encounter.

This example illustrates the need to ensure that new enhancements of older procedures be used with adequate care. Along these lines, it is the responsibility of statisticians to promote and assure the appropriate use of the FCR-adjusted selective CIs introduced in this article.

Analogous to advice for determining the family of related parameters over which the familywise error rate or FDR should be controlled, Benjamini and Yekutieli give this succinct guideline for deeming parameters to be *unselected*: "[parameters that are] ignored, not reported or even set aside in a website." To promote their new methodology, the authors should consider providing a more expansive set of guidelines to which researchers in any discipline can refer, comparable with the family choice guidelines offered by Westfall and Young (1993).

ADDITIONAL REFERENCES

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Institute for Scientific Information (2004), *Science Citation Index*, Philadelphia, PA: Author.

Sarkar, S. K. (2002), "Some Results on the False Discovery Rate in Stepwise Multiple Testing Procedures," *The Annals of Statistics*, 30, 239–257.

Thissen, D., Steinberg, L., and Kuang, D. (2002), "Quick and Easy Implementation of the Benjamini–Hochberg Procedure for Controlling the False-Positive Rate in Multiple Comparisons," *Journal of Educational and Behavioral Statistics*, 27, 77–83.

Westfall, P., and Young, S. S. (1993), *Resampling-Based Multiple Testing*, New York: Wiley.

Rejoinder

Yoav BENJAMINI and Daniel YEKUTIELI

We thank the five discussants for their careful reading of our article, and for contemplating the ideas expressed there. We are very grateful for their complimentary and illuminating comments, as well as for their thought-provoking critical ones. Rather than responding to each discussant separately, we have grouped their concerns into three major themes.

1. THE EXTENT OF THE SELECTION PROBLEM

Professor Edwards states that the selection problem is not commonplace in science, and that the proposed criterion

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000001961

might encourage such behavior. These are the only statements made with which we totally disagree. We hope, together with Edwards and Dr. Best, that the issue of bluntly reporting only significant results is rare in all branches of science. Nevertheless, the methodology of a recent medical article that we stumbled on is as follows: The variables, whose number was not reported, were screened for their effect on time to event, and only 20 individually significant ones at the .07 level were retained for further analysis. These variables were included in a survival model and underwent further screening, until three variables remained in the multivariate model, each significant at the .05 level (marginally). The abstract of the article merely reports the estimators and marginal 95% confidence intervals (CIs) for these three variables, as well as for the individual effect of six variables that were individually significant at the .05 level. In summary, this is a clear-cut incidence of abused selective inference. Still, the false coverage-statement rate (FCR) approach has nothing to offer here, unless the number of variables originally scanned is disclosed—in which case the inference addresses the selection effect. Therefore, the offering of FCR does not encourage "bad habits."

Much more important, commonplace selection is subtler, but may be as harmful. The results of medical research are mostly communicated by the abstracts of the articles. This fact is well recognized by the editors, and in many journals each abstract is required to be a self-contained summary of the goals, methods, and results of the reported research. By design, the abstract carries only selected results from the body of the article. As we noted in our introduction, the results in the article are usually not adjusted for simultaneous inference, so when results are selected for inclusion in the abstract, there may be a strong selection effect.

The two examples we gave in the article were in the category of such "selection by highlighting." In both studies, all CIs appear in the body of the article. But in the article by Giovannucci et al. (1995), the abstract communicates CIs for the three significant findings only, thereby grossly misrepresenting the results of the study. Adjusting to control the FCR could avoid the pitfall, because the results emphasized in the abstract are the ones making the headlines.

In the article by Rossouw et al. (2002) the abstract includes most of the findings: three out of three in one family of hypotheses, five out of seven in the other (mostly, but not only, significant ones, as correctly noted by Edwards). There was very little selection in highlighting these outcomes, and the FCR-adjusted CIs were close to the unadjusted CIs, adapting to the results of the selection process. Our point is that the researchers were unduly worried about the selection effect in their results, and their approach of ignoring of the solution offered by simultaneous inference not only is intuitive, but also can be formally justified from the FCR perspective.

2. THE INTERPRETATION AND APPROPRIATENESS OF THE FALSE COVERAGE–STATEMENT RATE CRITERION

Professors Shaffer, Tamhane, and Westfall indicate apparent difficulties with the interpretation of FCR when small numbers of parameters are involved or when the probability of no selection at all is high. In that respect, Shaffer describes our approach as "treating no selections as zero noncoverage," from which she concludes that we implicitly treat the unselected parameters as being covered with probability 1. Although we treat "no selections at all" as we treat "all selected parameters are covered"—both are equally harmless, it is not the case for the implication—in the FCR approach we simply ignore the coverage of a nonselected parameter, so it is irrelevant whether the coverage is 0 or 1. The following description may help with the interpretation: The FCR approach treats "a parameter not selected" as less harmful than a "parameter selected and not covered" and more harmful than "a parameter selected and covered"; the first does not change the Q_{CI} at all, whereas the second increases it and the third decreases it.

Westfall argues that the false discovery rate (FDR) is difficult to interpret, yet finds it easy to explain simultaneous coverage. It is even easier to explain conditional coverage. Actually, the FCR (just as the FDR) moves adaptively between the foregoing two interpretations. Facing a situation where in almost all realizations some parameter will be selected, the FCR can be interpreted as offering conditional coverage (as emphasized by Shaffer and Tamhane). However, when facing a situation where "there is nothing worth selecting," it behaves very much like simultaneous coverage. Moving smoothly between these two extremes can be perceived as follows: If any CI constructed involves further cost to the experimenter (say, for a follow-up study), then each experiment ends with a proportion of wasted money. The experimenter has an interest in keeping the long-run average of this proportion bounded at some low value, even by occasionally not following up on any finding from an experiment.

Westfall further argues that FCR inferences can be misleading even with very large m, using a real data example. He points at the two parameters in figure 1 whose FCR CIs do not cover 0, arguing that according to existing biological theory, they are likely to be two errors. This should come as no surprise; a researcher using the FCR or FDR must realize that some small proportion of the discoveries made may be false. In applications where two errors in a thousand discoveries is as harmful as two out of two, the margin for error is narrow, and the use of the familywise error rate (FWE) should be the rule. Thus if Westfall is concerned that two false discoveries (or even 20) out of a total of 5,397 discoveries will invalidate the conclusions from the entire study, then he should use FWE control as he advocates.

A related fundamental issue is the choice of the relevant family of hypotheses. Westfall states that the goal of the experiment was to compare red100% and green100%, with the other comparisons serving to indicate sensitivity. Therefore, the relevant family is the family of 2,666 tests displayed in the appropriate side of figure 2, from which the two seeming errors are selected and where they are in fact extreme. On the other hand, all discoveries that red80% differs from red100% are expected by theory. So why combine the $9 \times 2{,}666$ such pairwise comparisons to the family of 2,666 ones of primary interest? The price for the simultaneous approach taken over too large a family is that even if the difference that we see in figure 1 is real (a small and decaying difference in the beginning of the measured time period), it will not be detected. Imposing control over a larger-than-needed family of hypotheses makes it too easy to show equivalence using FWE, just as it might make it too easy to

discover differences using the FDR. The relevant issue is the proper choice of both the criterion and the family.

We want to use this example further to emphasize the role of the appropriate graphical display as an aid in interpretation. The FCR intervals presented in figure 1 are FCR-adjusted Benjamini–Hochberg (BH)-selected CIs. FCR is a property of the set of intervals not covering 0, and does not involve intervals for the non selected. To emphasize this point, only the FCR confidence bands not covering 0 should be drawn. Alternatively, one can draw the FCR confidence band for all parameters, but apply special visual impact to the selected parameters, in the form of heavy versus thin lines, or black versus gray symbols. Such a band will give FCR control of at most .05 for all parameters (because it offers marginal coverage of at least .95), while emphasizing visually the primary subset of the selected parameters for which the FCR property is critical.

3. SELECTIVE INFERENCE AS A CONTINUING RESEARCH CHALLENGE

All of the discussants raise important questions worth further attention and research efforts.

3.1 Dependency

Both Westfall and Holland raise the issue of FCR control under dependency. Westfall critically notes that there is no method for addressing correlated data within the FCR framework, in contrast to the Westfall and Young resampling method, which is always available for FWE control (as demonstrated in his example). In section 5.2 we gave a general FCR controlling procedure that is valid under any type of dependency. Using this general procedure on the same data, we calculate the inflating factor: $1 + 1/2 + \cdots + 1/26600 = 10.766$. The critical value at .05 is thus $t^{G-FCR} = 3.336$, corresponding to the tail probability of $1 - (5{,}397/(2 \times 26{,}600)) \times (.05/10.766)$. It is larger than the original $t^{FCR} = 2.585$, but still much smaller than $t^{BOOT} = 4.53$ or $t^{BON} = 4.85$.

Although the general solution is available and viable, in many cases it will not be needed. A practical answer to the dependency challenge may simply be in FCR-adjusted BH-selected CIs. Note that we address positive dependency in one-sided CIs using the regular FCR adjustment, and its properties are similar to FDR control of the BH procedure. Thus we conjecture that the FCR of two-sided CIs will be as close to q as the FDR of the procedure in BH is close to qm_0/m in two-sided testing problems.

As to the latter, the analytical results available for the BH procedure give an upper bound for normally distributed, positively correlated, test statistics—$FDR \leq qm_0/m$. However, simulations further reveal that for correlated normal test statistics, and two-sided tests, the FDR is actually very close to qm_0/m. Furthermore, in earlier work (Yekutieli and Benjamini 1999), we extended the methodology of Westfall and Young to produce resampling-based FDR controlling testing procedures. Interestingly enough, our working experience reveals that the BH procedure is as effective as the resampling-based FDR in analyzing highly correlated data (e.g., Benjamini and Yekutieli 2004). This is mostly because the FDR criterion is less sensitive to dependency than the FWE criterion. Taking the approach suggested by Holland may be a fruitful step toward establishing this property asymptotically.

3.2 Other Selection Rules

Shaffer demonstrates that if the selection procedure is "select the nonrejected" then the general adjustment procedure is too conservative for some parameter values. She is right. However, the fault is not in the concept of FCR, as implied, but rather in the general procedure. As always, the more general a procedure, the more likely it is to be dominated at some specific setting by a more specifically tailored procedure. We hope that research questions regarding selection rules of practical importance will receive attention, and eventually be answered by specific, and thus more powerful, selective inference procedures.

It is important to emphasize that if the interest lies in both the rejected and the nonrejected hypotheses, as discussed by Shaffer, then simultaneous inference is needed, and the FCR apparently is not the quantity of interest. Simultaneous CIs can be further tailored to satisfy goals beyond coverage, which are relevant in a research problem. Such nonequivariant simultaneous CIs were suggested to address questions regarding minimal length at some given parameter value, bioequivalence, or increased power in sign determination (e.g., Pratt 1961; Brown, Cassela, and Hwang 1995; Benjamini and Stark 1998).

3.3 Selection, Conditional Coverage Probability, and the False Coverage-Statement Rate

We emphasized the difficulties in achieving conditional coverage as exemplified in selection rules of the form $\mathcal{S}(\mathbf{T}) = \{i|\mu_i \notin CI_i(\mathbf{T})\}$. Notice, however, that for a given selection criterion, it might be possible to construct a CI offering $1 - \alpha$ conditional coverage probability. For example, if one uses the "reject by testing at level α" selection rule, then $1 - \alpha^2$ marginal CIs offer conditional coverage probability $\geq 1 - \alpha$. Professor L. Brown (personal correspondence) has suggested CIs based on the inversion of acceptance regions computed according to the conditional distribution of $T_i||T_i| \geq z_{1-\alpha/2}$ that are even shorter than that. Assuring conditional coverage probability for such a specific selection rule, where the selection of parameter i depends only on its estimator, implies control over the FCR.

3.4 The Effect of Selection on Estimators' Bias

Tamhane points at the bias of the estimators after selection as another problem and demonstrates its extent when the selection is based on testing a null value. Westfall demonstrates that when parameter values are normally distributed and a normal error is added to each, shrinkage estimators can provide unbiased estimation after selection via the BH procedure. Westfall further suggests constructing empirical Bayes CIs centered at the shrunk estimators of the BH-selected parameters. Although the bias is corrected, we are still in the dark about two important questions: (1) whether the CIs offer quantifiable coverage after the FDR selection, and (2) how the lengths of the empirical Bayes CIs compare with those of the FCR-adjusted CIs. Hopefully, these questions will be answered. In fact, it may well be that in the same way that adaptive FDR procedures found their interpretation in the empirical Bayes framework, so will the FCR. The foregoing setting may be the appropriate one in which to explore this important question.

Of course, it should be noted that the good performance of the specific estimators (and CIs) used by Westfall depends on the assumed model under which the estimators were derived. They will not necessarily retain their desirable properties under the mixture model used in microarray analysis. Thus the foregoing questions should be answered in different settings as well, opening up many more research questions. In that respect, it is interesting to note the performance of the FDR "testimator," where the parameters selected by the testing procedure in BH are kept as is, and the other estimators are shrunk all the way to 0. This testimator has asymptotically (in the number of parameters) minimax performance over sparse bodies of parameters (Abramovich, Benjamini, Donoho, and Johnstone 2000).

4. FINAL WORDS

We opened our rejoinder with one of Edwards's comments, and we will end with his closing remark "even the most famous statisticians vary widely in their attitude" toward the need to address multiplicity. This is a correct description, as vividly illustrated in a recent interview with Dennis Lindley (*Significance*, 2004, 73–74). We humbly think that multiplicity matters, and this fact is becoming clearer to researchers and decision makers as they face larger and more complex problems. Because we also agree with Edwards that CIs are more informative than just statistical significance, and they are not merely baggage, we tried in this article to contribute to the theory and practice of multiple CIs. Not only did we offer CIs to accompany FDR methodology, we also made an effort to illuminate the difference between simultaneous coverage and selective coverage, and offered a framework and procedures for addressing the latter.

If the results reported in abstracts of medical journals, for example, would all be adjusted for the selection effect using the FCR criterion, then it would be a great step forward toward addressing multiplicity effects. The protection offered by FCR will not always be enough, and simultaneous coverage may be needed. It is part of our responsibility to identify where one is more appropriate than the other, or even where neither is needed. As may be clear by now, our article is not the end of the story, but rather is much closer to the beginning. Still, we hope that enough has been said to make it not only thought-provoking, but also useful.

ADDITIONAL REFERENCES

Abramovich, F., Benjamini, Y., Donoho, D., and Johnstone, I. (2000), "Adapting to Unknown Sparsity by Controlling the False Discovery Rate," Technical Report 2000-19, Stanford University, Dept. of Statistics.

Benjamini, Y., and Stark, P. (1998), "Nonequivariant Simultaneous Confidence Less Likely to Contain Zero," *Journal of the American Statistical Association*, 91, 329–337.

Benjamini, Y., and Yekutieli, D. (2004), "Quantitative Trait Loci Analysis Using the False Discovery Rate," *Genetics*, under revision.

Brown, L. D., Cassela, G., and Hwang, J. T. G. (1995), "Optimal Confidence Sets, Bioequivalence, and the Limacon of Pascal," *Journal of the American Statistical Association*, 90, 880–889.

Pratt, J. W. (1961), "Length of Confidence Intervals," *Journal of the American Statistical Association*, 56, 549–567.

Yekutieli, D., and Benjamini, Y. (1999), "Resampling-Based False Discovery Rate Controlling Procedure for Dependent Test Statistics," *Journal of Statistical Planning and Inference*, 82, 171–119.