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,

$$\Pr\{\theta \notin CI, CI \text{ constructed}\} \leq \Pr\{\theta \notin CI\} \leq \alpha. \tag{1}$$

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

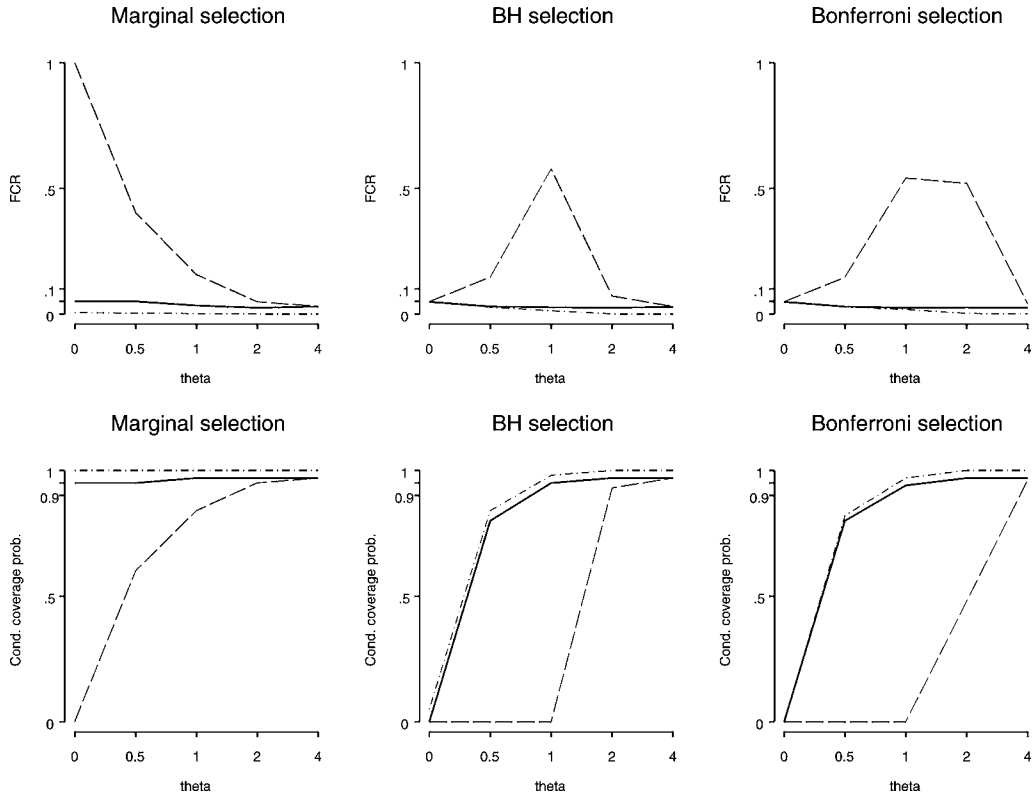

Figure 1. Simulation Based FCR and Conditional Coverage Probabilities of Marginal (- - - -), FCR-Adjusted (——), and Bonferroni (·—·—·—) .95 CIs for the Marginal, BH, and Bonferroni Level .05 Selection Schemes.

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 $\mathcal{S}(\mathbf{T})$, 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
$$\begin{aligned} &R_{\min}\big(\mathbf{T}^{(i)}\big) \\ &\quad := \min_t \big\{ \big|\mathcal{S}\big(\mathbf{T}^{(i)}, T_i = t\big)\big| : i \in \mathcal{S}\big(\mathbf{T}^{(i)}, T_i = t\big) \big\}. \quad (2) \end{aligned}$$
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 $\mathrm{U}[\theta_i, \theta_i + 1]$ 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,

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

Therefore, V_{CI} can be expressed as $V_{CI} = \min(k, V^*)$, where $V^* \sim \mathrm{Binom}(m, q')$. 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 $P_S(\mathbf{P})$,

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

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 $R_{CI} > \sqrt{m}$,

$$FCR = E\frac{V''}{R_{CI}} + E\frac{V'}{R_{CI}} \le E_{R_{CI}}\left\{ E_{V''|R_{CI}}\left(\frac{V''}{R_{CI}}\right)\right\} + E\frac{V'}{\sqrt{m}}$$
$$= E_{R_{CI}}\frac{\sqrt{m} \cdot R_{CI} \cdot q/m}{R_{CI}} + \frac{(m - \sqrt{m}) \cdot q/m}{\sqrt{m}} < \frac{2 \cdot q}{\sqrt{m}},$$

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, $FCR \ge q/2$. 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

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

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 $|T_i - \theta_i^0| \ge \mathcal{T}_i^{1-|S| \cdot q/(2m)}$, then $i \in S$.

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

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

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

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

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

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

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

$$E_{\mathbf{T}}(Q_{CI})$$
$$= \sum_{i=1}^{m}\sum_{k=1}^{m}\frac{1}{k} \cdot \Pr_{\mathbf{T}}\left\{C_k^{(i)}, i \in \mathcal{S}, \theta_i \notin CI_i\left(\frac{k \cdot q}{m}\right)\right\} \tag{9}$$
$$\ge \sum_{i=1}^{m}\sum_{k=1}^{m}\frac{1}{k} \cdot \Pr_{\mathbf{T}}\{C_k^{(i)}, |T_i - \theta_i^0| \ge \mathcal{T}_i^{1-k \cdot q/(2 \cdot m)},$$
$$|T_i - \theta_i| \ge \mathcal{T}_i^{1-k \cdot q/(2 \cdot m)}\} \tag{10}$$

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

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

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

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

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

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,

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

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 $\mathcal{S}(\mathbf{T})$, 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

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

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