Journal of Clinical Epidemiology 67 (2014) 850–857

Journal of Clinical Epidemiology

REVIEW ARTICLES

False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies

Mark E. Glickman[a,b,*], Sowmya R. Rao[a,c], Mark R. Schultz[a]

[a]*Center for Health care Organization and Implementation Research, Bedford VA Medical Center, 200 Springs Road (152), Bedford, MA 01730, USA*
[b]*Department of Health Policy and Management, Boston University School of Public Health, 715 Albany Street, Talbot Building, Boston, MA 02118, USA*
[c]*Department of Quantitative Health Sciences, University of Massachusetts Medical School, 55 N. Lake Avenue, Worcester, MA 01655, USA*

Accepted 12 March 2014; Published online 13 May 2014

Abstract

Objectives: Procedures for controlling the false positive rate when performing many hypothesis tests are commonplace in health and medical studies. Such procedures, most notably the Bonferroni adjustment, suffer from the problem that error rate control cannot be localized to individual tests, and that these procedures do not distinguish between exploratory and/or data-driven testing vs. hypothesis-driven testing. Instead, procedures derived from limiting false discovery rates may be a more appealing method to control error rates in multiple tests.

Study Design and Setting: Controlling the false positive rate can lead to philosophical inconsistencies that can negatively impact the practice of reporting statistically significant findings. We demonstrate that the false discovery rate approach can overcome these inconsistencies and illustrate its benefit through an application to two recent health studies.

Results: The false discovery rate approach is more powerful than methods like the Bonferroni procedure that control false positive rates. Controlling the false discovery rate in a study that arguably consisted of scientifically driven hypotheses found nearly as many significant results as without any adjustment, whereas the Bonferroni procedure found no significant results.

Conclusion: Although still unfamiliar to many health researchers, the use of false discovery rate control in the context of multiple testing can provide a solid basis for drawing conclusions about statistical significance.

Keywords: False discovery rate; False positive rate; FWER; Multiple tests; *P*-value; Study-wide error rate

1. Introduction

We are now in an age of scientific inquiry where health and medical studies are routinely collecting large amounts of data. These studies typically involve the researcher attempting to draw many inferential conclusions through numerous hypothesis tests. Researchers are typically advised to perform some type of significance-level adjustment to account for the increased probability of reporting false positive results through multiple tests. Such adjustments are designed to control study-wide error rates and lower the probability of falsely rejecting true null hypotheses. The most commonly understood downside of these procedures is the loss of power to detect real effects. Arguments have been put forth over the years whether adjustments for controlling study-wide error rates should be made, with plenty of advocates on each side of the argument. It appears doubtful that researchers will coalesce behind a unified point of view any time soon.

Significance level adjustments that control study-wide error rates are still common in peer-reviewed health studies. An examination of recent issues of several highly cited medical and health journals (Journal of the American Medical Association, New England Journal of Medicine, Annals of Internal Medicine, and Medical Care) reveals an abundant use of multiple-test adjustments that control study-wide error rates: We found 191 articles published in 2012 to 2013 making some adjustment for multiple testing, with 102 (53.4%) performing the Bonferroni or another study-wide error adjustment. Some other studies reported explicitly, and almost apologetically, that they had not performed an adjustment, and some even reported consequences of not having adjusted for multiple tests.

Despite the continued popularity of multiple test adjustments in health studies that control false positive error rates, we argue that controlling the false discovery rate [1] is an attractive alternative. The false discovery rate

Conflict of interest: None.
Funding: None.
* Corresponding author. Tel.: 781-687-2875; fax: 781-687-3106.
E-mail address: mg@bu.edu (M.E. Glickman).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2014.03.012

What is new?

- Controlling the false positive rate to address multiplicity of tests in health studies can result in logical inconsistencies and opportunities for abuse.
- Errors in hypothesis test conclusions depend on the frequency of the truth of null hypotheses being tested.
- False discovery rate control procedures do not suffer from the philosophical challenges evident with Bonferroni-type procedures.
- Health researchers may benefit from relying on false discovery rate control in studies with multiple tests.

is the expected fraction of tests declared statistically significant in which the null hypothesis is actually true. The false discovery rate can be contrasted with the false positive rate, which is the expected fraction of tests with true null hypotheses that are mistakenly declared statistically significant. In other words, the false positive rate is the probability of rejecting a null hypothesis given that it is true, while the false discovery rate is the probability that a null hypothesis is true given that the null hypothesis has been rejected.

Table 1 illustrates the distinction between the false positive and false discovery rates. Suppose that a set of tests can be cross-classified into a 2×2 table according to truth of the hypotheses (whether the null hypothesis is true or not), and the decision made based on the data (whether to reject the null hypothesis or not). Let a be the fraction of tests with true null hypotheses that are not rejected, b be the fraction of tests with true null hypotheses that are mistakenly rejected, c be the fraction of tests with false null hypotheses that are mistakenly not rejected, and d be the fraction of tests with false null hypotheses that are rejected. Assuming these fractions can be viewed as long-run rates, the false positive rate is computed as $b/(a + b)$, whereas the false discovery rate is computed as $b/(b + d)$. Although the numerators of these fractions are the same, the denominator of the false positive rate is the rate of encountering true null hypotheses, and the denominator of the false discovery rate is the overall rate of rejecting null hypotheses.

Conventional hypothesis testing, along with procedures to control study-wide error rates, are set up to limit false positive rates, but not false discovery rates. False discovery rate control has become increasingly standard practice in genomic studies and the analysis of micro-array data where an abundance of testing occurs. Several recent examples of false discovery rate control in health applications include provider profiling [2] and clinical adverse event rates [3], but false discovery rate control has yet to make serious in-roads into more general health studies. Of the 191 articles we found in highly cited journals that mention adjustments for multiple tests, only 14 (7.3%) include false discovery rate adjustments.

This article is intended to remind readers of the fundamental challenges of multiple-test adjustment procedures that control study-wide error rates and explain why false discovery rate control may be an appealing alternative for drawing statistical inferences in health studies. In doing so, we distinguish between tests that are exploratory and those that are hypothesis driven. The explanations we present to discourage use of adjustments based on study-wide error rate control are not new—the case has been made strongly over the past 10 to 20 years [4–9]. Arguments in favor of using false discovery rate control have been made based on power considerations [10–13], but we are unaware of explanations based on the distinction between exploratory and hypothesis-driven testing.

2. Adjustment for multiple testing through false positive rate control

The usual argument to convince researchers that adjustments are necessary when multiple tests are performed is to point out that, without adjustments, the probability of *at least one* null hypothesis being rejected is larger than acceptable levels. Suppose, for example, that a researcher performs 100 tests at the $\alpha = 0.05$ significance level in which the null hypothesis is true in every case. If all the tests are independent, then the probability that at least one test would be incorrectly rejected is $1 - (1 - 0.05)^{100} = 0.9941$, or 99.41%. In most studies, tests are not independent (eg, when tests share the same data), in which case, the probability of at least one incorrect rejection would not be quite so large, though likely large enough to be of some concern. Recognizing that the probability of at least one false positive may be unacceptably large, a common strategy is to adjust the significance level as a function of the number of tests performed.

One of the simplest and most commonly used approaches to adjusting significance levels is the Bonferroni procedure [14,15]. Letting n be the number of tests performed, and α the significance level one would normally use if performing only one test, the Bonferroni procedure involves rejecting null hypotheses whose P-values are less than α/n rather than α. For example, if a study involves 100 hypothesis tests and the researcher would ordinarily use $\alpha = 0.05$ as the significance level for a single test, then the Bonferroni procedure requires the researcher to compare each of the 100 P-values to $\alpha/n = 0.05/100 = 0.0005$. By invoking this procedure, the researcher is guaranteed that the probability of at least one false positive, regardless of the dependence among the tests, is no more than 0.05. Lowering the significance level in this manner requires the oft-acknowledged trade-off that the power to detect actual effects has been compromised, despite capping the probability of at least one false positive.

Table 1. Cross-classification of tests into the fraction with null hypotheses that are true vs. false, and those whose null hypotheses are rejected vs. not rejected

		Decision		
		Do not reject null hypothesis	Reject null hypothesis	
Truth	Null hypothesis true	a	b	$a + b$
	Null hypothesis false	c	d	$c + d$
		$a + c$	$b + d$	

The variables a, b, c, and d are the fractions of tests within each of the four cross-classified categories. The false positive rate is $b/(a + b)$, and the false discovery rate is $b/(b + d)$.

Although using procedures like the Bonferroni adjustment may seem statistically justifiable, two philosophical problems exist. The first is that procedures like the Bonferroni adjustment are actually simultaneous tests of a composite or a "universal" null hypothesis against an omnibus alternative hypothesis [5,8,16]. Rejecting the null hypothesis in favor of the alternative is merely a statement that at least one of the components that make up the composite null hypothesis is rejected, but without being able to specify which one. Controlling the probability that at least one component is rejected is usually too restrictive and rarely of interest to the researcher. As an example, suppose a researcher is interested in whether a health education program to reduce stress among participants of a randomized controlled study is effective, and separately tests the program's effectiveness within 20 different sociodemographic subgroups at the $\alpha = 0.05$ significance level. The Bonferroni procedure involves comparing the P-value for each of the 20 tests to $0.05/20 = 0.0025$. If any of the P-values is less than 0.0025, the conclusion of the Bonferroni procedure is that the composite null hypothesis, that the health education program is not effective for all the subgroups, is rejected. Arguably, a researcher is more interested in learning the significance of individual component hypotheses (ie, which subgroups evidence effectiveness). The Bonferroni procedure is often misapplied by singling out individual tests that are significant according to the above criterion. The procedure was developed as a solution to a problem in simultaneous inference and was not constructed for application to single tests [6,17–19].

A second more subtle problem is that the probability of a false positive result cannot be localized to specific tests. In other words, one can arbitrarily choose the tests over which a significance level adjustment is applied, and this arbitrary choice can lead to inconsistent conclusions [5].

As an illustration of the second problem, suppose that two researchers independently analyze the same data set. The first researcher performs 20 hypothesis tests, all resulting in P-values of 0.001. Using a Bonferroni procedure and assuming a single-test significance level of $\alpha = 0.05$, the adjusted significance level for each of the 20 tests is $0.05/20 = 0.0025$. The researcher would therefore conclude that all the tests are significant. Meanwhile, the second researcher performs the same 20 hypothesis tests and an additional 80 that also result in P-values of 0.001. For the combined 100 tests, the second researcher applies the Bonferroni adjustment and uses a significance level of $0.05/100 = 0.0005$ for each of the 100 tests. For this researcher, none of the tests are significant. The curious conclusion is that although both researchers performed 20 of the same tests, the second researcher could not conclude significance on any of them simply by performing additional tests.

A more common type of example occurs when a researcher chooses to divide a collection of tests into smaller groupings. Suppose a researcher performs 100 tests and obtains P-values of 0.001 for every test. As described previously, the Bonferroni-adjusted significance level is 0.0005, and none of the tests would be declared significant. But if the researcher decided to partition the 100 tests into five sets of 20 each, with the intention, for instance, of publishing each set of 20 in its own manuscript, then the researcher might perform the Bonferroni adjustment based on 20 tests on five separate occasions. In this latter situation, the Bonferroni-adjusted significance level in each of the five sets of 20 tests would be 0.0025, and every test would therefore be declared significant because $0.001 < 0.0025$. Again, the inferential conclusions depend solely on whether and how the tests are divided into groups.

We are aware of two serious attempts at justifying the use of Bonferroni-type multiple-test adjustments recognizing the previously mentioned difficulties. First, some researchers suggest asserting a maximum study-wise or family-wise error rate (FWER) that accounts for the largest number of tests one could conceivably perform in a study [20–22]. Despite specifying an FWER up front, a researcher with a penchant for data analysis may still perform more tests than the pre-specified FWER accounted for, rendering the purpose of the FWER adjustment moot. A second approach divides tests into those that are planned, and those that are unplanned (ie, *post-hoc* tests). When considering a multiple-test adjustment, a common recommendation is to apply the adjustment to unplanned tests, but use unadjusted significance levels for planned tests [5,23,24]. The problem with this strategy relates to the inability to localize the false positive rate: One researcher may perform 20 tests of which 10 are unplanned, whereas another researcher may perform the same 20 tests with all of them planned (and therefore performs no significance level adjustment). Once again, despite performing the same tests, the inferential conclusions may differ.

3. Toward an alternative criterion for assessing significance

To appreciate the basis for the difficulties associated with Bonferroni-type adjustments, we first remind the reader of the process by which null hypotheses are rejected. With conventional hypothesis testing:

1. A significance level (eg, $\alpha = 0.05$) is asserted.
2. A *P*-value is computed from the data.
3. If the *P*-value is less than the significance level, the result is declared statistically significant, and the null hypothesis is rejected.
4. The researcher then concludes that the null hypothesis is a false statement (or can be treated as such).

The researcher might therefore expect that the hypothesis-testing procedure provides some assurance that only a small fraction of rejected tests correspond to true null hypotheses. However, this is not necessarily the case.

The reason that a small false positive rate (eg, 0.05) does not translate to a small probability of true null hypotheses among rejected results (the false discovery rate) is that the latter probability depends on the frequency of tests in which the null hypotheses are true. In the most extreme case, if a researcher tests *only* true null hypotheses, 5% of which on average will be declared significant at the $\alpha = 0.05$ level, then *all* of the null hypotheses among the significant results will be true. In a less extreme but more realistic setting, genomic researchers who perform multitudes of tests for the significance of single-nucleotide polymorphisms on, say, phenotypic outcomes are in a situation where an enormous fraction of the tests involve true (or difficult to disprove) null hypotheses. By simply comparing *P*-values to a 0.05 significance level in this situation, most of the tests declared significant will involve true null hypotheses. On the other hand, if a researcher is performing many tests that are expected to involve mostly false null hypotheses, then the frequency of false null hypotheses among the tests declared significant will be high.

A common set of situations in which one might expect differences in the rates of true null hypotheses are tests that are hypothesis driven vs. tests that are exploratory or data driven. Hypothesis-driven testing, arguably the gold standard for scientific research, presumes that prior knowledge and scientific background to a problem motivates the tests to be performed. The researcher expects that null hypotheses are usually false in hypothesis-driven tests. Exploratory testing, which is common in data mining scenarios where the researcher is ''hunting'' for significant results, is not ordinarily motivated by scientific knowledge of a problem. For such tests, one can expect null hypotheses to be frequently true.

The impact of testing null hypotheses with different truth frequencies can be determined as a simple application of Bayes rule. Consider two researchers, each of whom rejects true null hypotheses at a 0.05 significance level, and who reject false null hypotheses at a 0.80 probability (ie, the power to detect the alternative hypothesis is 80%). Suppose the first researcher is primarily data driven, and performs tests in which the null hypothesis is true 99% of the time. Suppose that the second researcher is primarily hypothesis driven, and performs tests in which the null hypothesis is true only 1% of the time. For the first researcher, among results declared statistically significant at the 0.05 level, the probability the alternative hypothesis is true is given by Bayes rule as $0.01 \times 0.8/(0.01 \times 0.8 + 0.99 \times 0.05) = 0.139$, or about 14%. For the second researcher, the analogous computation yields $0.99 \times 0.8/(0.99 \times 0.8 + 0.01 \times 0.05) = 0.999$, or 99.9%. Thus, for the researcher who performs tests mostly with true null hypotheses, statistically significant results are very likely to be errors. Meanwhile, the researcher who performs tests with mostly true alternative hypotheses is nearly guaranteed to be correct when declaring statistical significance.

Ideally, if some assurance were desired that statistically significant results corresponded to true alternative hypotheses, one would want to know the a priori probability of the null hypotheses being true before setting a significance level cutoff. In settings with exploratory or data-driven hypotheses, the significance level cutoff would need to be small to maintain a low frequency of true null hypotheses among significant results. In settings with hypothesis-driven tests, the significance level cutoff should be higher. Procedures based on controlling the false positive rate (eg, the Bonferroni procedure) do not recognize this principle.

4. Inferring the probability of a true null hypothesis

When performing a single hypothesis test, it is nearly impossible to infer anything meaningful about the probability the null hypothesis is true. However, when performing multiple tests in a study, the distribution of *P*-values provides information relevant to inferring the frequency of true null hypotheses. This is because the distribution of *P*-values is a mixture of two components: The distribution of *P*-values for true null hypotheses, which by construction is uniformly distributed between 0 and 1, and the distribution of *P*-values for false null hypotheses, which is right skewed [25].

To see the distinction, consider the distribution of *P*-values from two different studies. We do not know the truth of the hypotheses being tested in either study, but based on the distribution of *P*-values we can make inferences about the truth. The first study involves the 28 *P*-values from Table 6 of Marx et al. [26] which summarize the effects of predictors in four multiple regressions on neuropsychological performance measures, and the second study involves 55 *P*-values from Table 1 of Bombardier et al. [27] that summarize differences in patient

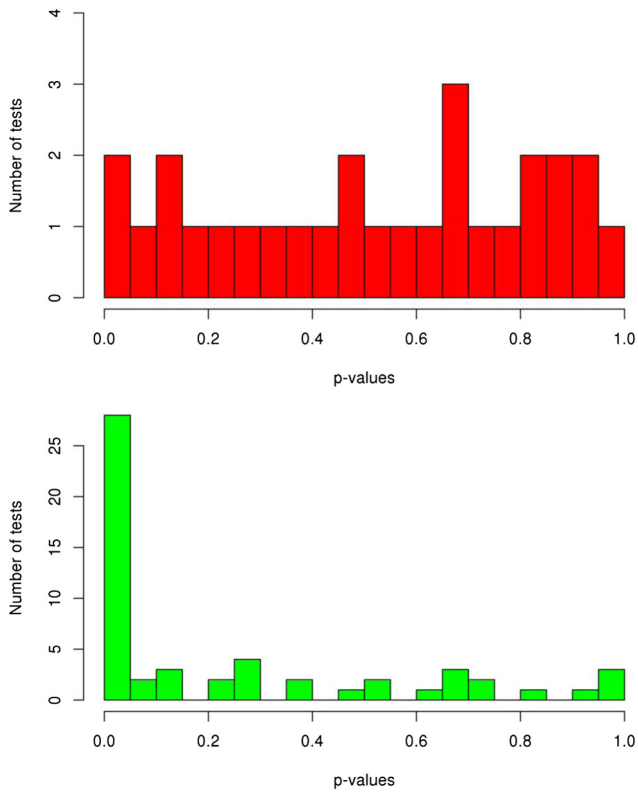

Fig. 1. Distribution of *P*-values for two studies. Top: Distribution of 28 *P*-values from Table 6 of Marx et al. [26]. Bottom: Distribution of 55 *P*-values from Table 1 of Bombardier et al. [27].

characteristics across two mental health conditions. Fig. 1 displays the distribution of *P*-values for each study. In the first study (represented by the top histogram), the distribution of *P*-values is roughly uniform, which is consistent with the null hypothesis being true for every test. Specifically, the *P*-values that are less than 0.05 in the presence of the remaining *P*-values are what would be expected if all the null hypotheses were true. In such a situation, the tests producing small *P*-values, including *P*-values below a significance level of 0.05, intuitively should not be declared significant because they are consistent with a uniform distribution of *P*-values. This study would be a likely candidate for setting a very low significance level cutoff for declaring significant results because we should not believe that the small *P*-values are indicative of false null hypotheses.

The second study, whose *P*-values are represented in the bottom histogram of Fig. 1, has a greater proportion of very small *P*-values than would be consistent with a uniform distribution. This lends support to the notion that many of the small *P*-values are instances of false null hypotheses. For this study, a very low significance level cutoff would inappropriately discount the small *P*-values, which intuitively indicate true positive results.

The problem, therefore, is to have a method to determine an appropriate significance level cutoff for *P*-values which recognizes that some studies tend to involve data-driven or exploratory null hypotheses that are generally true, and that some studies tend to involve scientifically driven hypotheses in which the null hypotheses are generally false. The method, by itself, should not be simply a function of the number of tests performed, as is the case with the Bonferroni procedure.

5. False discovery rate control

One way to implement such a process is by controlling the false discovery rate [1]. Many health researchers are unaware of the false discovery rate, although it is a natural concept, and one that has important utility for calibrating error rates in hypothesis tests. Among tests that are declared significant in a study, the false discovery rate is the expected fraction of those tests in which the null hypothesis is true. The main goal of false discovery rate control is to set significance levels for a collection of tests in such a way that among tests declared significant; the proportion of true null hypotheses is lower than a specified threshold. For example, if a false discovery rate procedure is applied to a set of 100 tests, and 20 are declared significant by the procedure, then the expected fraction of the 20 significant results that have true null hypotheses is lower than some pre-determined threshold (often chosen to be 0.05).

The reason false discovery rate control is an attractive alternative to false positive rate adjustments is that it explicitly controls the error rate of test conclusions among significant results. If, for example, a collection of published study results reported statistical significance at a false discovery rate of 10%, then the reader would have some assurance that at most 10% of the significant findings were mistakenly concluded to be true positives. Results that are significant at the 10% significance level do not have this property and suffer from the interpretive difficulties described earlier.

The original false discovery rate control method developed in the landmark article by Benjamini and Hochberg, henceforth BH, is still useful and appropriate for many applications. If a researcher wants to enforce false discovery rate control for a study with n tests with maximum false discovery rate d (often 0.05, but this is not to be viewed as a default choice), then the procedure is carried out as follows:

1. Sort the n *P*-values in ascending order; label these p_1, p_2, …, p_n.
2. Let k denote the largest index i for which $p_i \leq d \times i/n$, for all i.
3. Declare all tests with *P*-values p_1, p_2, …, p_k significant.

The derivation and rationale for this method is described elsewhere [1,2,28]. Notice that if $d = 0.05$, the rejection criterion in Step 3 ensures that most of the *P*-values must

Table 2. Worked-out example of the Benjamini–Hochberg procedure

Sorted *P*-values	0.0001	0.0002	0.01	0.013	0.03	0.04	0.07	0.15	0.26	0.52
$d \times i/n$	0.005	0.01	0.015	0.02	0.025	0.03	0.035	0.04	0.045	0.05
Is $P \leq d \times i/n$?	Yes	Yes	Yes	Yes	No	No	No	No	No	No

The upper row of the table is ten *P*-values in ascending order, and the bottom row is the comparison values in the procedure to determine which *P*-values to declare significant.

be quite a bit less than 0.05 for a test to be declared significant because each p_i is being compared with a fraction of 0.05 (ie, k/n of 0.05 where $k = 1, 2, \ldots, n$). In this sense, the BH procedure is much more conservative than simply rejecting tests by comparing *P*-values to a nominal level such as 0.05, but more powerful than the Bonferroni procedure which would compare all *P*-values to $0.05/n$. By comparison to the Bonferroni approach, the BH procedure compares only the smallest *P*-value to 0.05/n.

Although the BH procedure is implemented in many computer statistics packages, including SAS and R, the calculations are straightforward to illustrate manually. Suppose a researcher performs $n = 10$ hypothesis tests in a study resulting in the *P*-values 0.52, 0.07, 0.013, 0.0001, 0.26, 0.04, 0.01, 0.15, 0.03, and 0.0002. Furthermore, suppose that the researcher wants to implement the BH procedure with a maximum false discovery rate of $d = 0.05$. The first step is to sort the *P*-values in ascending order; these appear on the top row of Table 2. The second step is to list the values of $d \times i/n$, for $i = 1,\ldots,10$. With $d = 0.05$ and $n = 10$, these values are listed in the second row of Table 2. Finally, we note whether the values in the first row are less than or equal to the values in the second row. The third row of Table 2 indicates the comparison results. Because the four lowest *P*-values are less than their corresponding $d \times i/n$, these four tests are significant at the 0.05 false discovery rate, and the other six are not significant. If a Bonferroni adjustment were performed instead of a false discovery rate adjustment (misapplied by interpreting individual test results), only the first two lowest *P*-values would be significant, as all *P*-values would be compared with 0.005. Without any adjustment, the tests corresponding to the six lowest *P*-values would be declared significant.

The BH procedure overcomes the philosophical difficulties pointed out earlier with the Bonferroni procedure. First, the BH procedure is not a test of a composite null hypothesis against an omnibus alternative. Instead, the results of the BH procedure identify the individual tests that are to be declared significant. Second, the types of inconsistencies evident with the Bonferroni procedure do not occur with the BH procedure. In the earlier example in which two researchers independently analyze the same data, with the first performing 20 tests and the second performing an additional 80 tests, all with *P*-values of 0.001, all the tests are significant at the 0.05 false discovery rate level in each situation using the BH procedure; in fact, increasing the number of tests (assuming the *P*-values remain 0.001) will always result in every test significant at the 0.05 false discovery rate level because $i = n$ is the largest index for which $p_i \leq d \times i/n$ (ie, $0.001 \leq 0.05\ n/n = 0.05$).

This above example illustrates that the BH procedure is "scalable" as a function of the number of tests. Unlike the Bonferroni and other multiple-testing adjustment procedures, the BH and other false discovery rate control procedures work equally well with an increasing number of tests performed. With the Bonferroni procedure in particular, a sufficiently large number of tests can reduce the single-test significance level to such an extent that the possibility of rejecting any null hypothesis is all but prevented. It should be noted, however, that the validity in applying the BH procedure in a scenario where one performs the adjustment on an initial set of tests, and then again on a batch of additional tests, relies on two assumptions being met. First, the relative frequency of true null to true alternative hypotheses is assumed constant with the addition of new tests. The second more crucial assumption is that the distribution of *P*-values for tests with true alternative hypotheses is maintained with the addition of new tests. Intuitively, this means that the evidence for true alternative hypotheses in the additional tests should be equally strong (or weak), on average, to the ones initially studied. This assumption could be violated in a number of ways. For example, if the additional tests involved larger sample sizes, or if the additional tests were more likely to have larger (or smaller) effect sizes than the ones initially examined, then the distribution of *P*-values for true alternative hypotheses would likely be different from the initial set. In this case, the BH procedure would fail to scale properly with the addition of new results.

Most applications of false discovery rate control have been in situations where tens of thousands of tests (or more)

Table 3. Number of tests in each study, followed by the number of significant tests at the $\alpha = 0.05$ level when the significance level is unadjusted, Bonferroni-adjusted, and adjusted using the Benjamini–Hochberg false discovery rate control procedure

	N	Unadjusted	Bonferroni	Benjamini–Hochberg
Marx et al. (Table 6) [26]	28	2	0	0
Bombardier et al. (Table 1) [27]	55	27	0	26

Table 4. Number of tests in each study, followed by the number of significant tests at the $\alpha = 0.20$ level when the significance level is unadjusted, Bonferroni-adjusted, and adjusted using the Benjamini–Hochberg false discovery rate control procedure

	N	Unadjusted	Bonferroni	Benjamini–Hochberg
Marx et al. (Table 6) [26]	28	6	1	1
Bombardier et al. (Table 1) [27]	55	33	17	30

are performed, but the procedures work reliably in smaller numbers of tests. Simulation analyses [1,12,29] have indicated that false discovery rate control has uniformly better power than other competitor methods (including FWER control), and the fraction of false positives is about what would be expected, even in small to moderate numbers of simultaneous tests. Thus, false discovery rate control has application in smaller studies, though the advantages are more pronounced with larger numbers of tests.

To illustrate false discovery rate control in a real example, the BH procedure can be applied to the *P*-values from Marx et al. [26] and Bombardier et al. [27]. In Table 3, we display the number of significant tests at the $\alpha = 0.05$ level without adjusting the significance level, using a Bonferroni adjustment and BH adjustment at the 0.05 false discovery rate level. Table 4 provides summaries using a significance level of $\alpha = 0.20$, and 0.2 false discovery rate level. For the Marx et al. [26] study, in which the distribution of *P*-values is roughly uniform, both the Bonferroni and BH adjustments result in very few tests declared significant, both at the 0.05 and 0.20 levels. The low *P*-values likely correspond to true null hypotheses that had low *P*-values by chance, so it is reasonable that they should not be declared significant results. By contrast, for the Bombardier et al. [27] study, the Bonferroni adjustment produces a low number of significant tests at both the 0.05 and 0.20 levels, but the BH procedure results in nearly the same number of significant results as without any significance level adjustment. Because the frequency of low *P*-values for the Bombardier et al. [27] study is large, the BH procedure recognizes that almost no adjustment is needed to the significance level.

Although more powerful than the Bonferroni procedure, the BH procedure acts as though every null hypothesis were true when estimating the number of falsely rejected null hypotheses. This results in an inflated estimate of the false discovery rate. A popular refinement to the BH procedure [28] estimates the frequency of true null hypotheses from the distribution of *P*-values. To illustrate the logic, the 22 *P*-values uniformly distributed to the right of 0.2 (eg,) in the bottom histogram of Fig. 1 arguably correspond to true null hypotheses. This implies that approximately $22/0.8 = 27.5$ *P*-values total, including those to the left of 0.2, are true null hypotheses. Thus, of the 55 *P*-values in the study, approximately $27.5/55 = 50\%$ of the *P*-values correspond to true null hypotheses. Using 50% as the true null hypothesis frequency instead of 100% can sharpen conclusions and increase power. Other refinements have been proposed, including other methods to estimate the fraction of true null hypotheses [30,31], and mixture models for the probability any specific null hypothesis is true [32].

6. Conclusion

Despite the commonplace use of Bonferroni-type significance level adjustments to address the increased probability of mistakenly rejecting true null hypotheses, we argue that such adjustments are difficult to justify on philosophical grounds. Furthermore, if researchers are concerned about being unable to limit the probability of mistaken conclusions among statistically significant results, then using Bonferroni-type adjustments based on the multiplicity of tests do not directly address this concern. An alternative approach is to implement false discovery rate control, an adjustment method that has a solid foothold in areas of data mining large data sets, especially in the context of genomic data research. False discovery rate control is used much less frequently in health studies. But as health research continues to expand into areas requiring the mining of large databases or exploring highly detailed health information, researchers need to be aware of false discovery rate control as a means to make reliable, well-calibrated inferences from their studies.

The principle of false discovery rate control is not limited to multiple tests, but is much more difficult to implement when evaluating single tests. If a researcher had an estimate of the probability that a null hypothesis were true before performing a test, then with the help of Bayes theorem false discovery rate control could be applied by determining an appropriate significance level cutoff. Some authors [5,33] have advocated no significance level adjustments in multiple testing and, by implication, single testing. Although choosing not to adjust significance levels is justifiable, a disadvantage is that error rates among significant results cannot be properly calibrated if this were a goal of interest. However, in scenarios such as single tests where it is difficult to assess the probability that a null hypothesis is true, performing no significance level adjustment may be the only objective course of action.

False discovery rate control is unfamiliar to many health researchers, but is an important concept to appreciate especially in light of the common tendency to use significance level adjustments based on controlling study-wide error rates. Aside from the philosophical appeal to use false discovery rate control adjustments, one main practical benefit is the increased power, which researchers, no doubt, will come to recognize once they work with large databases and need to perform many tests. As scientific work

continues to see greater use of false discovery rate control, adjustments based on controlling false positive rates may become increasingly more difficult to justify.

Acknowledgments

This article is the result of work supported with resources and the use of facilities at the Bedford VA Medical Center, Bedford, MA, USA.

References

[1] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995;57:289–300.
[2] Jones HE, Ohlssen DI, Spiegelhalter DJ. Use of the false discovery rate when comparing multiple health care providers. J Clin Epidemiol 2008;61:232–40.
[3] Mehrotra DV, Heyse JF. Use of the false discovery rate for evaluating clinical safety data. Stat Methods Med Res 2004;13:227–38.
[4] Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. J Res Educ Eff 2012;5:189–211.
[5] O'Keefe D. Should familywise alpha be adjusted? Hum Commun Res 2003;29:431–47.
[6] Perneger T. What's wrong with Bonferroni adjustments? BMJ 1998; 316:1236–8.
[7] Rothman K. Adjustments are needed for multiple comparisons. Epidemiology 1990;1:43–6.
[8] Savitz D, Olshan A. Multiple comparisons and related issues in the interpretation of epidemiologic data. Amer J Epidemiol 1995;142: 904–8.
[9] Nakagawa S. A farewell to Bonferroni: the problems of low statistical power and publication bias. Behav Ecol 2004;15:1044–5.
[10] Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. Am J Public Health 1996;86:726–8.
[11] Noble W. How does multiple testing correction work? Nat Biotechnol 2009;27:1135–7.
[12] Verhoeven K, Simonsen K, McIntyre L. Implementing false discovery rate control: increasing your power. Oikos 2005;108:643–7.
[13] Lazar N. The big picture: multiplicity control in large data sets presents new challenges and opportunities. Chance 2012;25:37–40.
[14] Abdi H. Bonferroni and Šidák corrections for multiple comparisons. In: Salkind N, editor. Encyclopedia of measurement and statistics. Thousand Oaks, CA: Sage; 2007:103–7.
[15] Portney L, Watkins M. Foundations of clinical research: applications to practice. Upper Saddle River, NJ: Prentice Hall Health; 2000.
[16] Schulz K, Grimes D. Multiplicity in randomized trials I: endpoints and treatments. Lancet 2005;367:1591–5.
[17] Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986;73:751–4.
[18] Miller RG. Simultaneous statistical inference. New York, NY: Springer-Verlag; 1981.
[19] Dunn OJ. Multiple comparisons among means. J Am Stat Assoc 1961; 56:52.
[20] Bender R, Lange S. Adjusting for multiple testing—when and how? J Clin Epidemiol 2001;54:343–9.
[21] Thompson J. Invited commentary: Re: multiple comparisons and related issues in the interpretation of epidemiologic data. Am J Epidemiol 1998;147:801–6.
[22] Veazie P. When to combine hypotheses and adjust for multiple tests. Health Serv Res 2006;41:804–18.
[23] Keppel G, Zedeck S. Data analysis for research designs: analysis-of-variance and multiple regression/correlation approaches. New York, NY: W.H. Freeman; 1989.
[24] Ruxton GD, Beauchamp G. Time for some a priori thinking about post hoc testing. Behav Ecol 2008;19:690–3.
[25] Schweder T, Spjøtvoll E. Plots of p-values to evaluate many tests simultaneously. Biometrika 1982;69:493–502.
[26] Marx B, Brailey K, Proctor S, Macdonald H, Graefe A, Amoroso P, et al. Association of time since deployment, combat intensity, and posttraumatic stress symptoms with neuropsychological outcomes following Iraq war deployment. Arch Gen Psychiatry 2009;66: 996–1004.
[27] Bombardier C, Fann J, Temkin N, Esselman P, Barber J, Dikmen S. Rates of major depressive disorder and clinical outcomes following traumatic brain injury. J Am Med Assoc 2010; 303:1938–45.
[28] Storey J, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100:9440–5.
[29] Williams V, Jones L, Tukey J. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. J Educ Behav Stat 1999;24:42–69.
[30] Cox D, Wong M. Simple procedure for the selection of significant effects. J R Stat Soc Ser B 2004;66:395–400.
[31] Efron B, Tibshirani R, Storey J, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 2001;96:1151–60.
[32] Allison D, Gadbury G, Heo M, Fernández J, Lee C, Prolla T, et al. A mixture model approach for the analysis of microarray gene expression data. Comput Stat Data Anal 2002;39:1–20.
[33] Cook R, Farewell V. Multiplicity considerations in the design and analysis of clinical trials. J R Stat Soc Ser A 1996;159:93–110.