

Significance level adjustments for multiple testing in health studies: A case for false discovery rate control

Mark E. Glickman, Ph.D.<sup>1,2</sup>, Sowmya R. Rao, Ph.D.<sup>1,3</sup>, Mark R. Schultz, Ph.D.<sup>1</sup>

<sup>1</sup> Center for Health Quality, Outcomes, and Economics Research, Bedford VA Medical Center, Bedford, MA;

<sup>2</sup> Department of Health Policy and Management, Boston University School of Public Health, Boston, MA;

<sup>3</sup> Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA.

Corresponding Author:

Mark E. Glickman, Ph.D.

Center for Health Quality, Outcomes and Economics Research

200 Springs Road (152)

Bedford, MA 01730

Phone 781-687-2875; Fax 781-687-3106

Email [mg@bu.edu](mailto:mg@bu.edu)

Running Title: Significance level adjustments for multiple tests

Word Count: 3999

## ABSTRACT

Procedures for adjusting significance levels when performing many hypothesis tests are commonplace in health/medical studies. Such procedures, most notably the Bonferroni adjustment, control for study-wide false positive rates, and recognize that the probability of a single false positive result increases with the number of tests. In this paper we remind readers of the main arguments against significance level adjustments based on the number of tests. We argue that confusion may exist between an increased numbers of tests being performed and a high (prior) probability of each null hypothesis being true. This confusion may lead to the unwarranted multiplicity adjustment. We demonstrate how false discovery rate adjustments are a more principled approach to calibrating error rates of testing decisions in health and medical studies.

Key words: Multiple tests, multiple comparisons, p-value, Bonferroni, FWER

We are now in an age of scientific inquiry where health and medical studies are routinely collecting large amounts of data which typically result in the researcher attempting to draw many inferential conclusions through numerous hypothesis tests. To account for the increasing probability of reporting false positive results through multiple tests, researchers are typically advised to perform some type of multiple-test adjustment, often a significance level adjustment, which lowers the probability of falsely rejecting true null hypotheses. Arguments have been put forth over the years whether adjustments for multiple testing 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 for multiple testing are still common in health studies. An examination of recent issues of several highly cited medical and health journals (JAMA, NEJM, Annals of Internal Medicine, and Medical Care) reveals an abundant use of multiple-test adjustments. From January 2010 through May 2011, we found 118 articles that mentioned “multiple comparisons.” Of these, 59 made some adjustment for multiple testing. Most of the studies performed the Bonferroni adjustment. A relatively minor number of studies used other methods for the adjustment (Scheffe, Duncan, Lan-DeMets, O’Brien-Fleming). Even some of the studies that did not make any adjustment reported, almost apologetically, that they had not performed an adjustment, and some even reported consequences of not adjusting for multiple tests.

Despite the continued popularity of multiple test adjustments in health studies, we argue that instead of performing multiple test adjustments, controlling the false

discovery rate<sup>1</sup> (FDR) is a much more compelling approach to drawing statistical inferences when multiple tests are performed. FDR control has become increasingly standard practice in genomics studies where an abundance of testing occurs, but FDR control has yet to make serious in-roads into more general health studies.

This paper is intended to remind readers on the reasons to avoid multiple-test adjustment procedures, and explain the concept of FDR control and why it is highly relevant for improving inferential conclusions in health studies. The explanations we present to discourage use of adjustments based on multiplicity of tests are not new – the case has been made strongly in several articles over the past 10-20 years<sup>2-6</sup>. Arguments in favor of using FDR control have been made based on power considerations<sup>7,8</sup>, but we have not found explanations for the reasons FDR adjustments are more relevant and appropriate from a foundational standpoint for the problems with multiple testing. In this respect, the material that follows paints a more complete picture of the nature of the problem and a prescribed solution.

### **Significance level adjustment for multiple testing**

The usual argument to convince researchers that multiple-test 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 increased beyond 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 (e.g., if some 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 approaches to adjusting significance levels, albeit an extremely conservative one, is the Bonferroni procedure<sup>9,10</sup>. Letting  $n$  be the number of tests performed, and  $\alpha$  the significance level one would use if performing only one test, the Bonferroni procedure involves rejecting null hypotheses whose p-values are less than  $\alpha/n$  rather than  $\alpha$ . 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 has the benefit of capping the probability of at least one false positive, though the oft-acknowledged trade-off is that the power to detect actual effects is being compromised.

While the justification for using procedures like the Bonferroni adjustment may seem reasonable, two philosophical problems exist. The first is that procedures like the Bonferroni adjustment are tests of a composite or a “universal” null hypothesis against an omnibus alternative hypothesis<sup>2,5,11</sup>, so that 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.

Arguably one is more interested in testing individual hypotheses, and that the composite hypothesis is rarely of scientific concern. A second more subtle problem is that the probability of a false positive result cannot be localized to a specific set of tests. Put another way, one can arbitrarily choose the tests over which a significance level adjustment is applied, and this arbitrary choice can lead to inconsistent conclusions<sup>2</sup>.

Suppose, for example, that two researchers independently analyze the same data set. The first researcher performs 20 hypothesis tests, all of which result 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, but an additional 80 as well. Suppose that these latter 80 tests 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, therefore, is that even though both researchers performed 20 of the same tests, the second researcher could not conclude significance on any of them simply by virtue of 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 above, 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. 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 multiple-test adjustments recognizing the above-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<sup>12-14</sup>. An FWER is typically asserted at the start of a study, and the collection of tests is subject to what usually amounts to a somewhat conservative adjustment. This approach acknowledges that fewer tests may have been performed than the FWER accounts for, but in truth the problem with localizing false positive rates still remains. 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 (i.e., *post-hoc* tests). When considering a multiple-test adjustment, a common recommendation is to apply the adjustment to unplanned tests, but use ordinary significance levels for planned tests. The problem with this strategy also relates to the inability to localize the false positive rate: One researcher may perform 20 tests in which 10 are unplanned, while another researcher may perform the same 20

tests with all of them being planned (and therefore performs no significance level adjustment). Once again, despite performing the same tests, the inferential conclusions may differ.

### **Disentangling the confusion with significance level adjustments**

One way to understand the reasons many researchers have a proclivity to perform multiple-test adjustments is to appreciate more clearly the thought process that usually leads to declaring a result significant. Most researchers understand that a p-value is the probability of observing data at least as surprising as what was seen, given that the null hypothesis is true. A small p-value is therefore evidence that the null hypothesis must not have been true in the first place, and the significance level sets the threshold to decide whether a p-value is small enough to declare a test statistically significant. It therefore stands to reason that if a result has been declared significant at the  $\alpha=0.05$  significance level, the researcher can have some assurance that the data were too surprising to be consistent with the null hypothesis, and this led to its rejection. The mistake in logic is then to conclude that a statistically significant result therefore most likely corresponds to a false null hypothesis. Such a conclusion is an incorrect interpretation of a statistically significant result, mainly because the probability a statistically significant result indicates a false null hypothesis depends critically on the probability the null hypothesis was false prior to observing the data, as we describe below.

To appreciate this idea in the context of multiple testing, suppose that a researcher performs 500 tests at the  $\alpha=0.05$  significance level in which the null

hypothesis is false in every case, and that 400 of the tests are declared significant, a situation in which the power is 80%. For this situation, the probability a significant result indicates a false null hypothesis is 100%. Meanwhile, suppose another researcher performs a different set of 500 tests at the  $\alpha=0.05$  significance level in which the null hypothesis is true in every case. Here, suppose 25 tests are declared significant, which is the number that would be expected by chance. In this case, among the 25 significant results, the probability a significant result indicates a false null hypothesis is 0%. Any rule to adjust the significance level as a function of the number of tests performed (500 in each situation) will still result in all correct null hypothesis rejections in the first case, and no correct null hypothesis rejections in the second case. This is a clear case where a multiple test adjustment is inappropriate.

A similar argument is evident in the following example in diagnostic screening. Suppose a screening tool for a disease has a 95% probability of correctly diagnosing someone with the disease, and an 80% probability of correctly diagnosing someone without the disease. These probabilities are chosen to mimic the usual true positive and true negative rates associated with hypothesis testing (5% significance level and 80% power). Consider first screening a high-risk population: A random sample is selected from a population whose prevalence of the disease is 90%. It is a straightforward probability calculation to show that among those with a positive screen, 97.7% have the disease. Meanwhile, suppose a random sample is obtained in a general population in which only 10% have the disease. In this case, the probability of having the disease among those with a positive screen is only 34.5%. The key point is that the error probability of positive screens depends crucially on the population prevalence. In the

context of multiple hypothesis testing, the error probability of a significant result depends critically on the frequency of true null hypotheses in the first place, and not on the number of tests performed. If a clinician were to perform the diagnostic screening as described above to a high-risk population, it would be illogical to argue that more evidence is required that a patient has the disease if screening a large number of patients, and yet this is precisely what researchers do when performing multiple test adjustments.

The confusion in multiple-test adjustments is arguably manifested in researchers not distinguishing between performing many tests in which the null hypotheses frequently may be false, and performing many tests based on limited prior information in which the null hypothesis is likely to be true a large proportion of the time. The conventional wisdom in data mining and exploratory hypothesis testing is that continual and unrestricted testing of hypotheses will eventually produce spurious results. As the screening example illustrates, the problem is not the multiplicity of the analyses, but the overwhelming fraction of analyses that may be haphazard, not entirely scientifically motivated, and generally based on limited scientific knowledge. It would be a mistake, for example, to restrict the ability to declare significant results based on the number of tests, or for any other reason, if it were known in advance, or could be inferred, that tests being performed are likely motivated by the existence of real effects. As argued in recent papers<sup>15,16</sup>, the pre-study odds of the hypotheses affects the interpretation of statistical significance.

It may seem that distinguishing between settings in which null hypotheses tend to be true versus settings in which null hypotheses tend to be false is a hopeless task;

after all, the operating assumption of hypothesis testing is that the researcher is unaware about the truth of the stated hypotheses. Fortunately, the distribution of observed p-values within a study provides important information. This is because the distribution of p-values is a mixture of two distributions: The distribution of p-values for true null hypotheses, which is necessarily uniformly distributed between 0 and 1, and the distribution of p-values for false null hypotheses, which is right-skewed<sup>17</sup>.

To see the distinction, consider the distribution of p-values from two different studies. The first study involves the 28 p-values from Table 6 of Marx et al.<sup>18</sup> 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.<sup>19</sup> that summarize differences in patient characteristics across two mental health conditions. Figure 1 displays the distribution of p-values for each study. In the first study (represented by the top histogram of p-values), the distribution of p-values is roughly uniform, which is consistent with the null hypothesis being true for every test. In such a situation, the tests producing small p-values, even 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 a significance level adjustment 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 Figure 1, have 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 significance level adjustment would inappropriately discount the small p-values, which are indicative of truly significant 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 limited knowledge or unmotivated null hypotheses in which the null hypotheses are generally true, and that some studies tend to involve scientifically-driven hypotheses in which the null hypotheses are generally false. In the case of haphazard testing, one would expect that greater numbers of tests should result in more pronounced significance level adjustments, but when tests are scientifically motivated the larger number of tests ought not to decrease the significance level cutoff. The method, by itself, should not be simply a function of the number of tests performed.

### **False discovery rate control**

A principled way to implement such a process is by controlling the false discovery rate.<sup>1</sup> Many health researchers are unaware of the FDR, even though it is a natural concept and one that has important utility for calibrating error rates in hypothesis tests. The FDR is defined as follows: Among tests that are declared significant in a study, the FDR is the fraction of those tests in which the null hypothesis is true. In fact, this is the reverse conditional probability of the false positive rate, i.e., the probability that a test is declared significant among true null hypotheses. The main goal of FDR control is to determine a rule for setting significance levels, which may vary across the collection of tests, that lead to making the proportion of false positives out of the total number of tests declared significant as small as possible.

The reason FDR control is such a compelling concept is that it explicitly controls the error rate of test conclusions among significant results. If, for example, a journal had a policy to enforce FDR control rates of 5%, which would mean that no more than 5% of significant results were true null hypotheses, then readers could be assured that at least 95% of the significant findings across all research manuscripts in the journal (at the FDR of 5%) could be “trusted,” as they would correspond to correctly rejected null hypotheses. Results that are significant at the 5% significance level do not have this property, and suffer from the difficulty in interpretations described earlier.

Procedures for implementing FDR control is an ongoing area of statistical research, though the original FDR control method developed in the foundational paper by Benjamini and Hochberg<sup>1</sup>, henceforth BH, is still useful and appropriate for many applications. If a researcher wants to enforce FDR control for a study with  $N$  tests with maximum FDR rate  $d$  (often 0.05), then the procedure is carried out as follows:

1. Sort the  $N$  p-values in ascending order; label these  $p_1, p_2, \dots, p_N$
2. For  $k=1, \dots, N$ , declare test  $k$  significant at the false discovery rate  $d$  if  $p_k \leq dk/N$ .

The derivation and rationale for this method is described elsewhere<sup>1,19</sup>. Notice that if  $d=0.05$ , the rejection criterion in Step 2 ensures that most of the p-values must be quite a bit less than 0.05 in order for a test to be declared significant because  $k/N$  takes on the values  $1/N, 2/N, \dots, (N-1)/N, 1$ . 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  $0.05/N$ .

To illustrate the FDR control procedure, we re-examined the p-values from Marx et al. and Bombardier et al. In Table 1, we display the number of significant tests at the  $\alpha=0.05$  level without adjusting the significance level, using a Bonferroni adjustment, and using a BH adjustment at the 0.05 FDR level. Table 2 provides summaries using a significance level of  $\alpha=0.20$ , and 0.2 FDR level. For the Marx et al. 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. Intuitively, 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. 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. study is large, the BH procedure recognizes that almost no adjustment is needed to the significance level.

Unlike the Bonferroni and other multiple-testing adjustment procedures, the BH and other FDR control procedures are scalable in the sense that the procedure works equally well with an increasing number of tests performed. With the Bonferroni procedure, in particular, a large number of tests reduces the single-test significance level to such an extent that any 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 and true alternative hypotheses should be maintained. For the BH procedure, specifically, this turns out not to be a crucial assumption because the BH computation incorporates a conservative approximation that assumes every null hypothesis is true. Thus, at worst, if most or every null hypothesis being tested is false, then the BH procedure is conservative but no worse than the usual multiple testing procedures. The second more crucial assumption is that the distribution of p-values for tests with true alternative hypotheses be maintained. 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, including performing additional tests with larger sample sizes, and tests that are likely to have larger (or smaller) effect sizes than the ones already examined.

## **Conclusions**

Despite the commonplace use of multiple test significance level adjustments, such as the Bonferroni procedure, to address the increased probability of mistakenly rejecting true null hypotheses, we argue as others have that such adjustments are inappropriate and not defensible on foundational grounds. Furthermore, adjustments based on the multiplicity of tests do not directly address the real source of concern, namely that collections of tests may be addressing null hypotheses that are likely to be true (or at least unsubstantiated with available data) because they are purely exploratory and speculative in nature. A more principled approach to address such a concern 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. FDR 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 FDR control as a means to make reliable, well-calibrated inferences from their studies.

Developmental work in FDR control methods since the original BH method has mostly centered on relaxing the assumption that all null hypotheses are true, and instead either estimate the fraction of true null hypotheses<sup>20-22</sup>, or model the probability any specific null hypothesis is true versus false through mixture modeling<sup>23</sup>. The result of such work has been more powerful FDR control methods, though greater care is required to assess the differential probabilities of true null hypotheses.

FDR control, while currently unfamiliar to many health researchers, is an important concept to appreciate, especially in light of the inappropriate tendency to use significance level adjustments based on the multiplicity of tests. Aside from the philosophical reasons to use FDR control adjustments over multiple-test based 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 FDR adjustments, multiple-test based adjustments may eventually be relegated as a statistical curiosity.

## REFERENCES

1. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Statist Soc B*. 1995;57(1):289–300.
2. O'Keefe DJ. Should familywise alpha be adjusted? *Hum Commun Res*. 2003; 29(3), 431-447.
3. Rothman KJ. No Adjustments Are Needed for Multiple Comparisons. *Epidemiology*. 1990;1(1):43–46.
4. Perneger TV. What's wrong with Bonferroni adjustments? *BMJ*. 1998; 316,1236-1238.
5. Savitz DA, Olshan AF. Multiple Comparisons and Related Issues in the Interpretation of Epidemiologic Data. *Am J of Epidemiol*. 1995;142:904–908.
6. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. *JREE*. 2011; In press.
7. Noble WS. How does multiple testing correction work? *Nat Biotechnol*. 2009; 27(12):1135-1137.
8. Verhoeven KJF, Simonsen KL, McIntyre LM. Implementing false discovery rate control: increasing your power. *Oikos*. 2005; 108:643-647.
9. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. Upper Saddle River, NJ: Prentice Hall Health; 2000.
10. Abdi H. Bonferroni and Šidák corrections for multiple comparisons. In N.J. Salkind (ed.). *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage; 2007:103-107.

11. Schulz KF, Grimes DA. Multiplicity in randomized trials I: Endpoints and treatments. *Lancet*. 2005; 365:1591-1595.
12. Thompson JR. Invited Commentary Re: Multiple Comparisons and Related Issues in the Interpretation of Epidemiologic Data. *Am J of Epidemiol*. 1998;147(9):801–806.
13. Veazie PJ. When to Combine Hypotheses and Adjust for Multiple Tests. *Health Serv Res*. 2006; 41(3):804-818.
14. Bender R, Lange S. Adjusting for multiple testing – when and how? *JCE*. 2001; 54:343-349.
15. Ioannidis JPA. Why most published research findings are false. *PLoS Medicine*. 2005; 2(8):696-701.
16. Wacholder S, Chanock S, Garcia-Closas M, et al. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J Natl Cancer Inst*. 2004; 96:434-442.
17. Schweder T, Spjøtvoll E. Plots of p-values to evaluate many tests simultaneously. *Biometrika*. 1982;69:493-502.
18. Marx BP, Brailey K, Proctor S, et al. Association of time since deployment, combat intensity, and post traumatic stress symptoms with neuropsychological outcomes following Iraq war deployment. *Arch Gen Psychiatry*. 2009; 66(9): 996-1004.
19. Bombardier CH, Fann JR, Temkin NR, et al. Rates of Major Depressive Disorder and Clinical Outcomes Following Traumatic Brain Injury. *JAMA*. 2010;303:1938-1945.

20. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci*. 2003; 100:9440-9445.
21. Cox DR, Wong MY. A Simple Procedure for the Selection of Significant Effects. *J R Statist Soc B*. 2004;66(2):395–400.
22. Efron B, Tibshirani R, Storey JD, et al. Empirical Bayes analysis of a microarray experiment. *JASA*. 2001; 96:1151-1160.
23. Allison DB, Gadbury GL, Heo M, et al. A mixture model approach for the analysis of microarray gene expression data. *Comput Stat Data An*. 2002; 39:1-20.

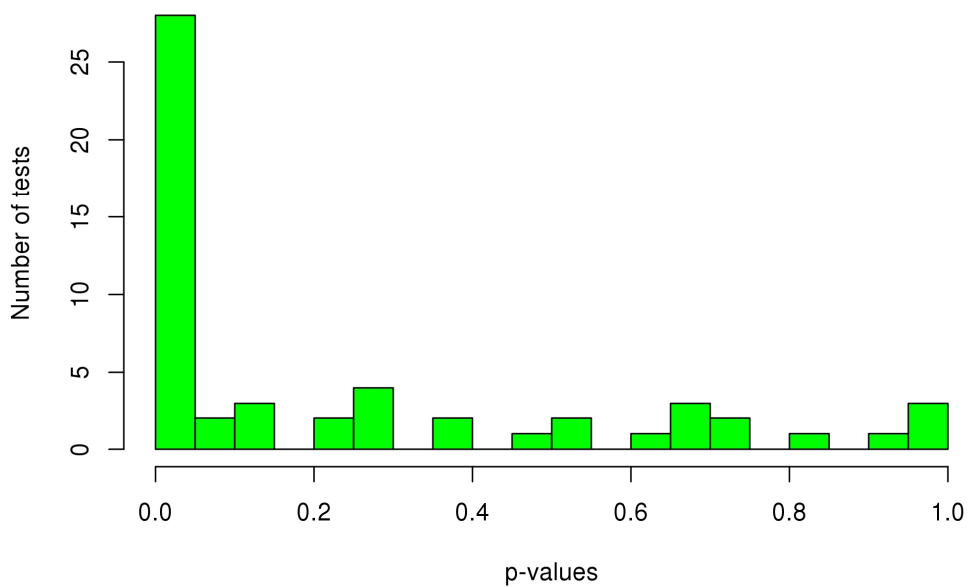
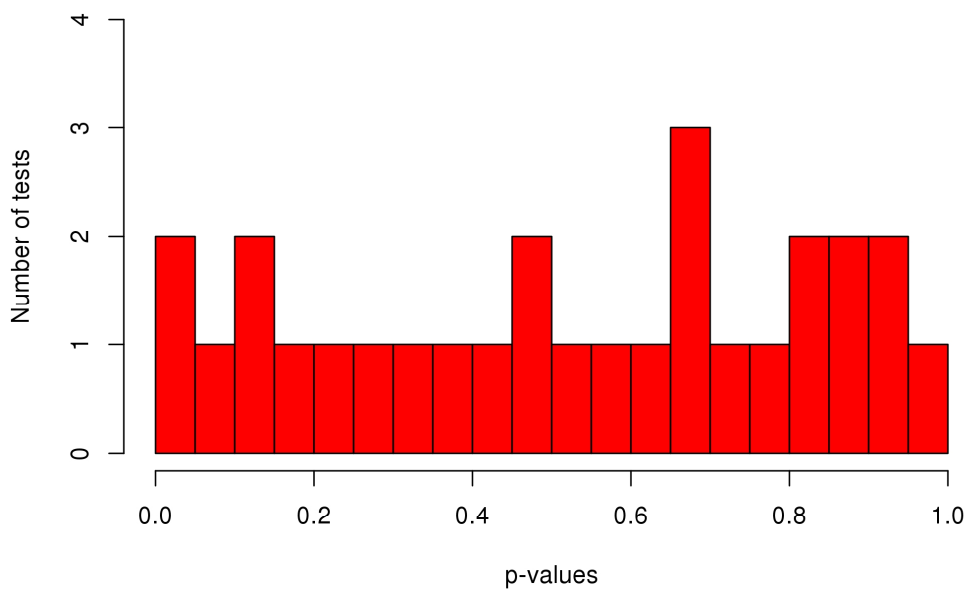


Figure 1: Distribution of p-values for two studies. Top: Distribution of 28 p-values from Table 6 of Marx et al. (2009). Bottom: Distribution of 55 p-values from Table 1 of Bombardier et al. (2010).

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

*Table 1:* 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 FDR control procedure.

	$N$	Unadjusted	Bonferroni	Benjamini-Hochberg
Marx et al. (Table 6)	28	6	1	1
Bombardier et al. (Table 1)	55	33	17	30

*Table 2:* 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 FDR control procedure.