Duke GLOBAL HEALTH INSTITUTE | Research Design & Analysis Core

Core Guide: Multiple Testing, Part 1

Part of a series addressing common issues in statistical and epidemiological design and analysis

Background

An important feature in the planning of any quantitative research study is the determination of the required sample size necessary to achieve sufficient statistical power. Power is defined with respect to a specific hypothesis, or set of hypotheses, of interest. Typically, sample size calculations are performed under the assumption that we would like to have at least 80% power (equivalently, a **Type II error** rate of 20%) to reject the null hypothesis with a false positive (i.e., **Type I error**) rate fixed at 5% (see Box A).

It is becoming increasingly common for modern global health research studies to utilize sophisticated designs with complex sets of objectives. Such studies may involve multiple sets of endpoints at multiple times, as well as several different exposure groups from different target populations. Each of these factors alone may necessitate the use of multiple statistical tests simultaneously to achieve the study's objectives. Failure to take into account these multiple tests can lead to an increase in the rate of Type I errors and may also affect power. This phenomenon is sometimes alternatively referred to as a "multiplicity" or "multiple testing" problem.

Box A. Types of statistical error

In its simplest form, hypothesis testing focuses on a single comparison (i.e. null vs. alternative hypothesis). In this case, we make a decision about whether to reject the null hypothesis based on the results of a statistical test. Statistical error refers to the risk of making an incorrect decision. Rejecting the null hypothesis when it is actually true is a **false positive**, or a **Type I error.** Failing to reject the null hypothesis when it is actually false is a **false negative**, or a **Type II error.** False positive rates are typically denoted by **α**, while false negative rates are typically denoted by **β.**

Decision about null hypothesis	**Null hypothesis is**	
	True	**False**
Reject	False positive (**α**)	True positive (1- **β**)
Fail to reject	True negative (1- **α**)	False negative (**β**)

When designing quantitative research studies, we are usually interested in controlling the false positive rate **α** while still maintaining adequate statistical power, which is defined as **1-β**. We often refer to the level of **α** control as the significance level.

There remains considerable debate in the biomedical literature about the need for multiple testing adjustments and the appropriate methods to implement these adjustments in different contexts (Bender & Lange, 2001; Rothman, 1990). The issue of multiple testing and how best to deal with it in a specific study can be daunting. As a result, there is a strong need for specific, practical guidance to help investigators understand which situations call for such adjustments to be made, which adjustment methods best match certain study designs and questions, and the degree to which making (or failing to make) these adjustments impacts interpretation of the results. The purpose of this Core Guide is to provide a brief, non-technical overview of the multiple testing problem, and to summarize how and when adjustments should be made. A separate Core Guide, Multiple Testing, Part 2, focuses on specific multiple testing adjustment methods, and guidance on which method to use in specific contexts.

Overview of the Multiple Testing Problem

The primary purpose of multiple testing procedures is to control the rate of Type I errors when performing a number of statistical tests on the same sample. Most statistical tests used for deriving inference from hypothesis-driven studies (e.g. randomized controlled trials) produce a p-value, which we consider 'significant' if it is less than some pre-determined threshold, **α**. Alternatively, inference may instead be based around examining the width of a confidence interval on a particular estimate of interest. The width of a confidence interval is typically linked to the threshold **α**, and thus the same multiple testing procedures we will discuss for adjusting p-values are also relevant for inference based on confidence intervals. Since using the confidence interval to make inference on whether two population parameters (e.g. risk) overlap is akin to the inference made from interpretation of the p-value, this method does not avoid the multiple testing problem (Lash, 2017).

The p-value for a single hypothesis test (i.e. null vs. an alternative hypothesis) is the probability that a *test statistic* (i.e. some statistical summary of the data, like the mean difference between two groups) would be equal to or more extreme than its observed value, assuming that the null hypothesis is true. This means that p-values are interpreted with respect to the null hypothesis of the test used to generate them. If the p-value is as small or smaller than **α**, then we deem it **'**significant' and we reject the null hypothesis. Traditionally, the value of **α** is set to 0.05, corresponding to an 'acceptable' false positive rate of 5%. That is to say that if the null hypothesis was actually true and we repeated the experiment 100 times, then we would expect to falsely reject the null hypothesis, on average, 5 out of those 100 repetitions simply due to random chance.

As the number of hypothesis tests performed increases, so too does the expected false positive rate. If for a single hypothesis test with threshold **α** the probability of *not* making a type I error is (1-**α**), the probability of not making a type I error for *N* hypothesis tests is equal to $(1-\boldsymbol{\alpha})^N$. Thus, the probability of a type I error for *N* hypothesis tests is equal to $1 - (1-\boldsymbol{\alpha})^N$. When *N=1*, this reduces to **α**. For an **α** of 0.05 and *N*=2, this probability is 9.75%. That is, if we are performing two hypothesis tests (each with **α** set to 0.05), for each of the 100 repetitions of the experiment, instead of 5 false rejections as above (when we were performing only a single hypothesis test per repetition of the experiment), we would expect to see almost 10 false rejections (i.e. on 10 out of the 100 repetitions of the experiment we get a false positive

finding on at least one of our two hypotheses). When *N*=10, the probability of at least one type I error across our ten hypothesis tests is 40.1% (i.e. on about 40 out of our 100 repetitions of the experiment we get a false positive finding on at least one of our ten hypotheses). It is important to emphasize that, in the frequentist paradigm, type I error rates are interpreted with respect to multiple hypothetical repetitions of the experiment, where *within each repetition* we are performing *N* hypothesis tests.

When performing a series of hypothesis tests, instead of thinking about the type I error rate of each test in isolation, we should be thinking about this *overall* type I error rate across this "family" of *N* related tests. This is called the **family-wise error rate (FWER)**. Instead of saying that our 'acceptable' false positive threshold is **α** for each test, we would say that our 'acceptable' FWER threshold is **α.** Most common multiple testing procedures are designed to control the FWER. Essentially, when controlling the FWER, we are protecting ourselves against the possibility that the observed results of the experiment we conducted were one of the chance false positive repetitions (e.g. one of the 40 out of 100 hypothetical repetitions of an experiment with 10 hypothesis tests).

Before proceeding, it is also important to recognize that there is great debate within the statistical community about the merit of p-values and of hypothesis testing itself. For example, a recent statement from the American Statistical Association on the topic notes that scientific conclusions should not be based only on whether a p-value passes a specific threshold(Wasserstein & Lazar, 2016), and a recent publication in Nature Human Behavior calls for the standard significance level of 5%, widely accepted by scientific journals and taught in most introductory courses, be revised to 0.005% (Benjamin et al., 2017). Due in part to the limitations and misuse of p-values in scientific research, the journals *Epidemiology*, *The American Journal of Epidemiology*, *Basic and Applied Social Psychology*, and several other biomedical journals, now have highly restrictive guidelines for the reporting of p-values in published articles. Further, Bayesian methods of statistical inference, which are not based on p-values or confidence intervals, are an increasingly common alternative to traditional hypothesis testing, but there is some debate on the necessity of multiple testing adjustments in that context. Bayesian statistical theory is beyond the scope of this Core Guide, but see Berry & Hochberg (1999) and Gelman et al. (2012) for more discussion on this topic. Nevertheless, given the ubiquity of statistical testing and the use of p-values in biomedical and global health research, it is important that practitioners understand the issues at stake with multiple testing in the context of using p-values for statistical inference.

Example: Pairwise Group Differences in a 3-Arm Trial

Imagine that a randomized controlled trial (RCT) was conducted in a malaria endemic region of rural Tanzania, designed to assess the efficacy of two different vector (i.e. mosquito) management strategies. 300 individuals were randomized in a 1:1:1 ratio to one of three treatment arms: individuals in the first treatment arm received insecticide-treated bednets (ITN) to protect its residents from mosquito bites while they sleep; the living areas for individuals in the second treatment arm were treated with indoor residual spraying (IRS) to kill mosquitoes; and the final treatment arm was the control. The objective of the study was to assess which of these interventions was most successful in reducing malaria transmission, by measuring the prevalence of malaria 12-months after randomization. The results of the

study are shown in **Table 1**. At 12-months, eight of the individuals in the ITN arm had malaria, 19 of the individuals in the IRS arm had malaria, and 33 of the individuals in the control arm had malaria.

Table 1. Prevalence 12-months after randomization for each of the three treatment arms in the malaria RCT example.

Treatment Arm	Prevalence (π)
ITN	$\frac{8}{100} = 0.08$
IRS	$\frac{19}{100} = 0.19$
Control	$\frac{33}{100} = 0.33$

The researchers were interested in three different pairwise comparisons: each of the two treatment arms with the control, and then the two treatment arms with each other. This resulted in three null hypotheses, each of which was evaluated using a Z-test for equality of proportions. The hypotheses and the p-values resulting from each test are shown in **Table 2.** For each test, the null hypothesis states that the prevalence of malaria is equal in the two groups being compared; this is equivalent to testing that the difference between the prevalences between the two groups is equal to 0. Thus, the null hypothesis can be evaluated by calculating a test statistic that is basically the difference in prevalences between the two groups (with scaling by the standard error of that difference). A p-value is then calculated by comparing the observed values of this test statistic with the sampling distribution of that statistic under the null hypothesis. All three tests returned p-values below the 0.05 threshold, and thus were deemed significant.

Table 2. Null hypotheses, test statistics, and p-values for each of the 3 comparisons in the malaria RCT example.

Comparison	Null Hypothesis	Difference in prevalences	P-value
Control vs. ITN	$\pi_{Control} - \pi_{ITN} = 0$	0.33 – 0.08	<0.001
Control vs IRS	$\pi_{Control} - \pi_{IRN} = 0$	0.33 – 0.19	0.036
IRS vs ITN	$\pi_{IRN} - \pi_{ITN} = 0$	0.19 – 0.08	0.039

Therefore, from these results, we would conclude that both vector management interventions succeeded in significantly reducing malaria prevalence, and that bednets were more successful at this than indoor spraying.

However, in interpreting the results this way, we are implicitly assuming that each of these p-values and null hypotheses are independent of one another. In fact, this assumption is not true: all 3 comparisons are related since they are using the same data (i.e. the same groups of people) to inform the results of the hypothesis test. As mentioned previously, our significant threshold, **α,** is typically interpreted as the "acceptable" type I error rate for a given hypothesis test. Instead of thinking about the type I error rate of each test in isolation, however, we should be thinking about the FWER.

In order to control the FWER, we need to implement some sort of adjustment procedure to control for the fact that we have made multiple hypothesis tests.

Types of Multiple Testing Adjustment Procedures

In general, there are two broad approaches towards interpreting the results of a series of hypothesis tests, to inform a "global" null hypothesis. The first is the **union-intersection ("at-least-one") testing approach**; here, we reject our global null hypothesis if *at least one* of the individual hypotheses in our family of tests is rejected (Dmitrienko & D'Agostino, 2013). For example, in our malaria example, we would take the rejection of the null hypothesis for *any* of the 3 pairwise group comparisons as evidence against the "global" null hypothesis (i.e. that none of the interventions reduce malaria prevalence). The alternative is the **intersection-union ("all-or-nothing") testing** approach, whereby we only reject the global null hypothesis if *all* of the individual null hypotheses in the family of tests are rejected. That is, we would only consider our treatment efficacious if we observe a significant effect for *all* 3 pairwise group comparisons. In most research context, the union-intersection (i.e. "at-least-one") approach is of interest, and thus the FWER must be controlled; the intersection-union approach, while it avoids type I error inflation, comes at the cost of decreased power, and thus is not widely used.

There are a wide variety of methods used to control the FWER at a pre-specified level by taking multiple testing into account. These methods are typically based around changing either the p-values and/or the threshold level **α** according to certain sets of rules. See **Appendix 2** for an overview of some of the most common procedures. Broadly speaking, these methods can all be grouped into some general categories based on the type of approach they take for controlling the type I error rate:

- **Strong FWER control** refers to methods that ensure the total FWER across all tests is never to exceed **α** under any conditions. This is often based around some method of *alpha allocation*, whereby the FWER is set to **α**, with each individual test in the family being allocated some fraction of **α** as its individual significance threshold.
    - Single-step procedures use alpha allocation to set the individual threshold for each test at a level that ensures strict control of the overall FWER. The **Bonferroni correction**, for example, distributes the threshold evenly by dividing the overall FWER threshold by the number of tests. For example, if the number of tests performed is two, then the

Bonferroni FWER threshold is **α**/2 (e.g., 0.05/2= 0.025). A pre-specified alpha allocation strategy may weight hypotheses by their relative clinical importance; e.g. in a family of three hypotheses tests, the "most important" hypothesis might be given a threshold of **α**/2 (e.g., 0.05/2= 0.025), with the other two tests each given a threshold of **α**/4 (e.g., 0.05/4= 0.0125) so that the overall FWER is **α.**

  - Adaptive procedures are designed such that the **α** level of each test is dependent in some fashion on whether one or more null hypotheses in the family have already been rejected. This is typically done by sequentially testing the hypotheses according to some hierarchy, with the result of a given test determining the threshold used on the subsequent test. The hierarchy may be data-driven (e.g. with **Holm's step-down procedure**, which orders the sequence of hypothesis tests from those with the lowest p-value to those with the highest) or pre-specified (e.g. the sequence is selected according to substantive knowledge and/or clinical importance).

- **Weak FWER control** refers to methods that allow the FWER to exceed **α** under certain defined conditions; specifically, the FWER is controlled at **α** only when *all* null hypotheses are *true*. These methods are designed to increase the power of the tests relative to strong FWER control methods, at the cost of potentially inflated type I error rates. As with strong FWER control, these methods may be either single-step or adaptive.
  - The most common method for ensuring weak FWER control is by using some threshold *q* to control the **false discovery rate** (FDR). Whereas FWER-based methods are aimed at controlling the probability of *at least one* type I error across the family of tests, FDR-based methods are aimed at controlling the expected proportion of rejected null hypotheses that are false positives.
- Empirical methods that control Type I error rates adaptively based on distributional characteristics of the observed data and the correlations between tests in the family. These are typically based on bootstrap or permutation re-sampling (see Westfall & Young, 1993). These methods may be either single-step or adaptive, and provide either weak or strong control, depending on the exact implementation. These methods are not widely implemented, since they are only guaranteed to control the FWER for large samples, though they are flexible enough to handle more complex multiple testing scenarios.

To illustrate these methods, we apply several procedures to our malaria example and compare how they impact the results (**Table 3)**. Using the *p.adjust* function in R, we applied the Bonferroni correction (*method="bonferroni")* and Holms' step-down procedure (*method="holm"*) for strong control of the FWER and the **Benjamini-Hochberg** (BH) procedure for weak control of the FWER (*method="BH"*). Details on these methods can be found in **Appendix 2**. This table denotes how the p-values for each of the three comparisons change after implementation of each procedure controlling for multiple testing.

Table 3. Updated results of the malaria RCT example, adjusting for multiple testing using three different common adjustment procedures: Bonferroni correction, Holms' step-down procedure, and Benjamini-Hochberg procedure.

Comparison	Difference in prevalences	Unadjusted p-value	Adjusted p-values		
			Bonferroni	Holms'	Benjamini-Hochberg
Control vs. ITN	0.33 – 0.08	<0.001	<0.001	<0.001	<0.001
Control vs IRS	0.33 – 0.19	0.036	0.108	0.072	0.039
IRS vs ITN	0.19 – 0.08	0.038	0.116	0.072	0.039

As we can see, each adjustment method leads to different interpretations of the results. In all cases, regardless of method, the prevalence in the ITN group is significantly reduced compared to control. This would give us confidence that this result is not a type I error. However, the results for the other two comparisons (control vs. IRS, and IRS vs. ITN) are more ambiguous. Using the Bonferroni correction or the Holms' procedure, when each adjusted p-value is compared to the α=0.05 level (which, because of the multiple testing procedure, is now equivalent to controlling the overall α-level at α=0.05 level), we find no significant difference between the control the IRS arm, or between the two treatment arms. Using the Benjamini-Hochberg procedure, however, all three of our comparisons remain statistically significant at the overall FWER of α=0.05.

The Bonferroni correction is the most conservative of the strong FWER control methods; meaning that while it is guaranteed to control for type I error, it may lead to increased type II error (and thus reduced power). As a weak FWER control method, Benjamini-Hochberg protects against type II error (and thus preserves power), but is not as stringent at controlling type I errors. The Holms' procedure lies somewhere between the two in terms of both type I error control and power (i.e. it is a strong FWER control method that is less conservative than Bonferroni).

So, how does one choose the most appropriate result?

Recommendations

Overall, there is still considerable debate as to when and how to adjust for multiple testing. The simplest way of avoiding the multiple testing problem is to limit the number of tests being performed, and the best way to do this is by planning the analysis in advance. Hypothesis testing should be restricted to as small a set of *a priori* comparisons as possible to achieve the objectives of the study. Note, however, that simply because the multiple tests are planned *a priori* does not automatically preclude the possibility of type I error inflation (Frane, 2015). For exploratory analyses, or for novel research programs where the pre-test probability of rejecting the null hypothesis is quite low, the type I error rate will already be very high. In these situations, null hypothesis testing, either by generation of p-values, or by assessing overlap of 95% confidence intervals, may not be appropriate (Lash, 2017). In contrast, in hypothesis-driven RCTs, researchers commonly get around the question of multiple testing

by pre-specifying the primary goal of the RCT to be focused on a single outcome at a single time point. As a consequence, it may be argued that there is no need to adjust for multiple testing.

One may also collapse the individual outcomes into a single *composite outcome*, and rely on a single significance test on this composite. While these methods avoid any inflation of type I error as a result of multiple testing, they are rarely of practical interest, since we may often wish to interpret the results differently based on which combinations of individual tests are, or are not, significant. For more discussion on statistical methods used in these settings, see Offen et al. (2007), Huque et al. (2011), and Tamhane & Dmitrienko (2009).

In the case that adjustment for multiple testing will be used, a variety of procedures have been proposed in the literature. It is difficult to provide an exhaustive list of all procedures and their benefits/drawbacks. In the absence of such an exhaustive list, we provide recommendations to guide the decision based on the general class of multiple testing procedure.

Most researchers are advised to utilize methods for strong FWER control. While some authors advocate for broader use of weak FWER control methods due to their increased power (Glickman et al., 2014), it is not generally recommended to use these methods except in the context of high dimensional data (e.g. having thousands of gene expression levels collected on a small number of patients) where strong FWER control methods may be too conservative (Storey, 2010).

The choice of *which* strong FWER control method to apply may be driven by a variety of factors that are highly context dependent. Broadly speaking, the decision should be driven by *a priori* knowledge of how the null hypotheses being tested relate to each other and to the underlying phenomenon under investigation. Pre-specified, adaptive alpha allocation methods (e.g. the **fixed-sequence procedure** or **chain procedures**) provide an optimal balance between statistical power and error rate control; however, proper implementation of these methods requires the explicit specification of a hierarchy of importance of the hypotheses. This can only be done using substantive (non-statistical) understanding of the subject matter.

In the lack of such a hierarchy of importance, the decision on the appropriate adjustment method is dependent on the degree of correlation you expect to see between the hypothesis tests. If the hypotheses are not expected to be correlated with one another, data-driven alpha allocation based methods are more powerful than the single-step procedures while still maintaining strong FWER control. Examples of these approaches include **Holm's step-down procedure**, the **Hochberg step-up procedure**, and the **Hommel procedure**. The latter two are more powerful under the assumption that the family of test statistics is multivariate normal and the correlation between them is positive; thus, if these assumptions are not warranted, the more conservative Holm's procedure is recommended. If those assumptions *are* warranted, the Hommel procedure is more powerful than the Hochberg procedure.

In the case that you expect the hypotheses to be correlated with one another, but you lack an explicit hierarchy of importance to inform the adaptive alpha allocation procedures mentioned above, the single-step procedures (e.g. the **Bonferroni correction** or the **Šidák procedure**) are preferable to data-

driven approaches. The Bonferroni correction is always an attractive option since it is highly conservative and simple to implement. The Šidák procedure is more powerful than the Bonferroni correction for independent or positively correlated hypotheses; however, for negatively correlated hypotheses it may inflate the type I error rate.

Overall, there is no "one-size–fits-all" approach to determining when and how to adjust for multiple testing in quantitative research studies in global health. It is important to consider a range of options and how they may best address the research question of interest. There should be a clear procedure in place before the data is even collected as to which hypothesis tests will be conducted and how the results will be interpreted given any pattern of 'significant' or 'not significant' results across tests. Ultimately, the goal is to find a balance between statistical power and the risk for error, which can only be done through substantive understanding of the phenomenon under investigation and how the different statistical hypotheses being conducted relate to this phenomenon.

Prepared by: Ryan Simmons, MSc
Reviewed by: Duke Global Health Institute | Research Design & Analysis Core
Version: 1.0, last updated December 29, 2017

Appendix 1: Glossary of Terms

Type I error rate (α): the probability of rejecting a true null hypothesis when the null hypothesis is true; a false positive.

Type II error rate (β): the probability of accepting the null hypothesis when the alternative hypothesis is true; a false negative.

Power (1- β): the probability of rejecting the null hypothesis when it is actually false; a true positive.

Test statistic: a statistical summary of a dataset used to evaluate a null hypothesis within the confines of some specific statistical test.

Family-wise error rate (FWER): the probability of making at least one type I error when performing multiple hypothesis tests.

False discovery rate (FDR): the expected proportion of rejected null hypotheses that are type I errors.

Strong FWER control: the total FWER across all hypothesis tests is never to exceed some threshold **α.**

Weak FWER control: if all null hypotheses are actually true, the total FWER across all hypothesis tests is never to exceed some threshold **α.** Controlling the FDR is equivalent to weak FWER control.

Appendix 2: Common Multiple Testing Adjustment Procedures

More detailed descriptions of these procedures, and discussions of their properties, can be found in Dmitrienko et al. (2012), Dmitrienko & D'Agostino (2013), and the recent Food and Drug Administration (FDA) industry guidance on multiple endpoints in clinical trials (*Multiple Endpoints in Clinical Trials: Guidance for Industry*, 2017).

STRONG FWER CONTROL: these procedures ensure that the total FWER across all *m* hypothesis tests is never to exceed some threshold **α.**

SINGLE-STEP PROCEDURES: these procedures adjust the significance level of each hypothesis test to a *fixed* threshold <**α**, such that the total FWER is strongly controlled at **α**.

-The **Bonferroni correction** divides the total FWER evenly across all tests by setting the significance level of each test to **α/m.**

-The **Šidák procedure** divides the total FWER evenly across all tests by setting the significance level of each test to $\mathbf{1-(1-\alpha)^{1/m}}$.

- **α allocation** may be used in a single-step fashion as a modified Bonferroni correction. For example, in a family of three hypothesis tests, the "most important" hypothesis may be allocated **α/2**, to increase the chance of a significant finding, while the other two tests are given **α/4**.

ADAPTIVE PROCEDURES: these procedures adjust the significance level of each hypothesis test to some threshold <α and tests the hypotheses iteratively, such that the significance level for subsequent tests depends on whether you reject or fail to reject the preceding null hypotheses. The order the tests are conducted can be based on a pre-specified hierarchy or may be data-driven (e.g. beginning with the smallest p-value). The following common procedures are all data-driven:

-**Holm's step-down procedure** orders the p-values from smallest to largest and sets the threshold for each test to **α/(m-i-1)**, where i is the ordered index of the p-value. For example, the smallest p-value (i=1) is given the threshold **α/m**; if this hypothesis is rejected, the adjusted significance level for the second smallest p-value (i=2) is **α/(m-1)**, which corresponds to the Bonferroni correction applied to the remaining **m-1** hypotheses; and so on until a hypothesis is not rejected.

-The **Hochberg step-up procedure** orders the p-values from largest to smallest, and sets the threshold for each test to **α/i**, where i is the ordered index of the p-value. For example, the largest p-value is tested at the threshold **α;** if it is rejected, all null hypotheses are rejected. If it is not rejected, the second largest p-value is tested at the threshold **α/2**. If it is rejected, all remaining null hypothesis are rejected. If it is not

rejected, the procedure continues with each successive test using the threshold **α/i**, until a test is rejected.

-The **Hommel procedure** rejects all p-values at the threshold **α/j**, where j is the largest integer for which $p_{m-i+k} > k\alpha/i$ for all values of k from 1 to m. The nuances of the algorithm are complex, see Wright (1992) for more details.

Appendix 3: R code for malaria RCT example.

```
# Test 1: Control vs. ITN
test1 <- prop.test(x = c(33, 8), n = c(100, 100))

# Test 2: Control vs. IRS
test2 <- prop.test(x = c(33, 19), n = c(100, 100))

# Test 3: IRS vs. ITN
test3 <- prop.test(x = c(19, 8), n = c(100, 100))

# P-values
pval <- c(test1$p.value,test2$p.value,test3$p.value)


# Bonferroni correction
pval_bonferroni <- p.adjust(pval,method="bonferroni")


# Holmes step-down procedure
pval_holmes <- p.adjust(pval,method="holm")


# Benjamini-Hochberg procedure
pval_benjamini <- p.adjust(pval,method="BH")
```

References

Bender, R., & Lange, S. (2001). Adjusting for multiple testing - when and how? *Journal of Clinical Epidemiology, 54*, 343-349.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., . . . Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behavior, 2*, 6-10.

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Jounral of Statistical Planning and Infernece, 82*(1-2), 215-227.

Dmitrienko, A., & D'Agostino, R. (2013). Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine, 32*(29), 5172-5218.

Dmitrienko, A., D'Agostino, R. B., & Huque, M. F. (2012). Key multiplicity issues in clinical drug development. *Statistics in Medicine, 32*, 1079-1111.

Frane, A. V. (2015). Planned Hypothesis Tests Are Not Necessarily Exempt From Multiplicity Adjustment. *Journal of Research Practice, 11*(1).

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness, 5*, 189-211.

Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology, 67*, 850-857.

Huque, M., Alosh, M., & Bhore, R. (2011). **Addressing multiplicity issues of a composite endpoint and its components in clinical trials**. *J Biopharm Stat, 21*(4), 610-634.

Lash, T. (2017). The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing. *Am J Epidemiol., 186*(6), 627-635.

Multiple Endpoints in Clinical Trials: Guidance for Industry. (Docket No. FDA-2016-D-4460). (2017).

Offen, W., Chuang-Stein, C., Dmitrienko, A., Littman, G., Maca, J., Meyerson, L., . . . Yeh, C.-H. (2007). Multiple Co-primary Endpoints: medical and Statistical Solutions: A Report from the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of America. *Drug Information Journal, 41*, 31-46.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1*(1), 43-46.

Storey, J. D. (2010). False discovery rates. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science*.

Tamhane, A. C., & Dmitrienko, A. (2009). Analysis of Multiple Endpoints in Clinical Trials *Multiple Testing Problems in Pharmaceutical Statistics* (pp. 131-164): Chapman & Hall.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statment on p-Values: Context, Process, and Purpose. *The American Statistician, 70*(2), 129-133.

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