

The False Discovery Rate: A Key Concept in Large-Scale Genetic Studies

James J. Chen, PhD, Paula K. Roberson, PhD, and Michael J. Schell, PhD

Background: In experimental research, a statistical test is often used for making decisions on a null hypothesis such as that the means of gene expression in the normal and tumor groups are equal. Typically, a test statistic and its corresponding *P* value are calculated to measure the extent of the difference between the two groups. The null hypothesis is rejected and a discovery is declared when the *P* value is less than a prespecified significance level. When more than one test is conducted, use of a significance level intended for use by a single test typically leads to a large chance of false-positive findings.

Methods: This paper presents an overview of the multiple testing framework and describes the false discovery rate (FDR) approach to determining the significance cutoff when a large number of tests are conducted.

Results: The FDR is the expected proportion of the null hypotheses that are falsely rejected divided by the total number of rejections. An FDR-controlling procedure is described and illustrated with a numerical example.

Conclusions: In multiple testing, a classical “family-wise error rate” (FWE) approach is commonly used when the number of tests is small. When a study involves a large number of tests, the FDR error measure is a more useful approach to determining a significance cutoff, as the FWE approach is too stringent. The FDR approach allows more claims of significant differences to be made, provided the investigator is willing to accept a small fraction of false-positive findings.

Introduction

Simultaneous considerations of a set of inferences are common in clinical or preclinical studies. Clinical trials frequently incorporate one or more of the following design features: multiple outcome measures,^{1,2} repeated tests of significance as the study progresses (interim analyses) to ensure early detection of effective treatments,^{3,4} subgroup analysis to address particular concerns on efficacy and safety of the drug in some specific patient subgroups,^{5,7} and various combinations of these features.^{1,8} In preclinical animal experiments for chronic effects, researchers routinely analyze approximately 10 to 30 tumor types and/or sites for screening of carcinogenicity of chemicals.⁹ Multiple testing refers to carrying out more than one statistical test for simultaneous inferences on a study. The best-known multiple tests are the analysis of variance (ANOVA) post-hoc tests after obtaining a significant omnibus *F* test¹⁰ when comparing means of several groups. A significant *F* test result sug-

gests rejecting the null hypothesis that the means are equal. Multiplicity tests are then used to determine which pairs of means are significantly different. Recently, analysis of genomic, proteomic, and metabolomic data for identifying potential molecular biomarkers from hundreds or thousands of variables has presented challenges in multiplicity testing.^{11,12} Appendix 1 contains a glossary of terms often used in statistical inference.

A “statistical hypothesis test” is a formal scientific method to examine the plausibility of a specific statement regarding the comparison of a parameter (or measurement) between one group and a fixed value or between two or more groups. The statement about the comparison is typically formulated as a “null hypothesis” that there is no difference in the values of the parameter between the groups. An “alternative hypothesis” is designed for the study objective to be proved, such as that the mean values differ between the groups or the mean of one specific group is greater than the other. The null hypothesis is used to derive the “null distribution” of a test statistic (such as a *T* statistic). This distribution serves as a reference to describe the variability of the test statistic due to chance. The test procedure compares the test statistic to the null distribution and computes a *P* value to summarize the test result. The *P* value is the probability of obtaining an experimental outcome as observed or more extreme if the null hypothesis is true. A small *P* value indicates that the test statistic lies in the extremities of the null distribution. This finding suggests that the null hypothesis does not accurately describe the situation. When the *P* value is less than a prespecified “level of significance,” such as .05, the null

From the Division of Personalized Nutrition and Medicine, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas (JJC), the Department of Biostatistics, University of Arkansas for Medical Sciences, Little Rock, Arkansas (PKR), and the Biostatistics Department, H. Lee Moffitt Cancer Center & Research Institute, Tampa, Florida (MJS).

Submitted May 27, 2009; accepted July 21, 2009.

Address correspondence to James J. Chen, PhD, Division of Personalized Nutrition and Medicine, National Center for Toxicological Research, Food and Drug Administration, HFF20, Jefferson, AR 72079. E-mail: JamesJ.Chen@fda.hhs.gov

The views presented in this paper are those of the authors and do not necessarily represent those of the US Food and Drug Administration.

Duality of Interest: None declared.

Table 1. — True State and Decision From a Hypothesis Testing

True State of Nature	Decision From Hypothesis Test	
	Null hypothesis rejected (significant)	Null hypothesis not rejected (not significant)
Null True	False-positive type I error	True negative
Alternative True	True positive	False-negative type II error

hypothesis is rejected. This result is described as “statistically significant” or simply as “significant.” Conversely, if the *P* value is above the threshold, the null hypothesis is not rejected since the results are not inconsistent with the null hypothesis; the result is described as “non-significant.” A rejection is referred to as a “statistical discovery.”

When a statistical test is performed, depending on whether the null hypothesis is true or false and whether the statistical test rejects or does not reject the null hypothesis, one of four outcomes will occur: (1) the procedure rejects a true null hypothesis (a false positive type I error), (2) the procedure does not reject a true null hypothesis (a true negative), (3) the procedure rejects a false null hypothesis (a true positive), or (4) the procedure does not reject a false null hypothesis (a false-negative type II error). The true state and the decision to accept or reject a null hypothesis can be summarized into a 2 × 2 table (Table 1).

The *P* value of a test can be simply interpreted as the probability of a false positive. When the *P* value is less than the level of significance, the null hypothesis is rejected. The level of significance is usually defined under a single test. When there is only one hypothesis, the probability of making a false-positive error is controlled at the level of significance. If more than one test is conducted, the *P* value of an individual test can no longer be interpreted as the probability of false positive of the overall test. Consider an experiment to compare treatment vs control groups separately for measurements made on the samples. If there are 10,000 such measurements (such as gene expressions), then up to 10,000 tests on individual measurements can be performed. With just one test, performing at the usual 5% significance level, there is a 5% chance of incorrectly rejecting the null hypothesis. However, with 20 tests in which all null hypotheses are true, the expected number of such false rejections (positives) is 1 (5% of 20). If 10,000 tests are performed, the expected number of false positives is 500 (5% of 10,000). In this instance, at least one null hypothesis (and likely many) will be rejected incorrectly. Unfortunately, we will not know which are correct or incorrect.

Many statistical techniques have been developed to control the false-positive error rate associated with making multiple statistical comparisons.¹³⁻¹⁷ One of the simplest methods to account for multiple testing is to adjust the significance level to account for the number of tests. This is achieved by dividing the significance level for each test by the number of tests performed. To insure

that the false-positive error rate is no greater than 5% when performing 20 simultaneous tests, each individual is significant only if its *P* value is less than $(.05/20) = .0025$. This approach is known as the Bonferroni adjustment. The Bonferroni adjustment has the advantage that it can easily be used in any multiple testing application. One major disadvantage to the Bonferroni adjustment is that it is not an exact procedure in the sense that it over-adjusts. For 20 independent tests, the actual level of significance required is .00256 instead of .0025. That is, the Bonferroni adjustment gives a smaller significance level than the true significance level. It adjusts more than necessary, which results in reducing the power to detect a difference. Stepwise methods (step-down¹⁴ and step-up¹⁵ methods) and resampling methods¹⁶ have been developed as alternatives to the Bonferroni approach. However, the Bonferroni and its improvement methods, which control the family-wise error rate (FWE), are not ideal when the number of tests is large, such as microarray gene expression experiments that often involve thousands of tests, as the level of significance is too stringent. This paper presents an overview of multiple testing and describes the false discovery rate (FDR) approach when the number of tests is large.

Multiple Testing Framework

In a large-scale study, such as microarray experiments, the FWE approach tends to screen out all but a few genes that show extreme differential expressions. Benjamini and Hochberg¹⁸ introduced the concept of the FDR as an alternative to the FWE for the multiple testing problems. Their article represents one of the most highly cited statistical reports in the past 25 years.

Consider testing *m* hypotheses. Assume that m_0 are from the true null population and $m_1 = (m - m_0)$ are from the true alternative population. A statistical test is performed for each hypothesis; the null is either rejected (significant) or not rejected (not significant). According to the true state of nature, either the null or the alternative is true. Based on the decision to reject the null hypothesis or not, the results from *m* tests can be summarized as a 2 × 2 table (Table 2).

Table 2. — Four Possible Outcomes When Testing *m* Hypotheses

True State of Nature	Null Hypothesis Rejected	Null Hypothesis Not Rejected	Total
Null True	V	S	m_0
Alternative True	U	T	m_1
Total	R	$m - R$	m

m_0 = the number of true null hypotheses and m_1 = the number of true alternative hypotheses, $m_0 + m_1 = m$.
 V = the number of the true null hypotheses that are falsely rejected.
 U = the number of times the null hypothesis is correctly rejected because the alternative is true.
 S = the number of true null hypotheses that are correctly not rejected.
 T = the number of times null hypothesis is incorrectly not rejected (alternative is true).
 R = the total number of null hypotheses rejected among the *m* tests.

For a given decision rule, the total number of rejections (significances declared) (R) and the total number of nonrejections (insignificances declared) ($m - R$) are observable. However, V , U , S , and T are unknown (unobservable). The proportion of the null hypotheses that are falsely rejected over the total number of rejections is V/R . The proportion of the alternative hypotheses that are falsely declared not significant is $T/(m - R)$. The proportions V/R and $T/(m - R)$ refer to the false discovery ratio and false nondiscovery ratio, respectively. Similarly, the proportions U/R and $S/(m - R)$ refer to the true discovery ratio and true nondiscovery ratio, respectively. Furthermore, the proportions V/m_0 and U/m_1 can be called the false-positive ratio and true-positive ratio (or power), respectively. In the context of medical diagnostic test summary, U/m_1 and S/m_0 are referred to as sensitivity and specificity, respectively.

In testing m hypotheses, the FWE approach is commonly used in testing multiple clinical points. The FWE is defined as the probability of rejecting at least one true null hypothesis. An FWE-controlled procedure guarantees that the probability of one or more false positives is not greater than a predetermined level, regardless of how many genes are tested. In other words, FWE is the probability that the number of false rejections (V) is greater than 0. In mathematical notation, $FWE = \Pr(V > 0)$. When m is large, this probability $\Pr(V > 0)$ can be large. Assume the m tests are independent. Using 5% as the level of significance for each individual test, the probability of at least one false rejection can be calculated as $\Pr(V > 0) = [1 - (1 - 0.05)^{20}] \sim 0.64$ when $m = 20$. This probability becomes approximately .99 when $m = 100$. To control the FWE at the .05 level (to ensure that the probability of making at least one false rejection is no greater than .05), the level of significance for an individual test, denoted by α , can be computed using the Bonferroni adjustment: $\alpha = .05/m$. For $m = 20$, $\alpha = .0025$ and for $m = 100$, $\alpha = .0005$. When 20 tests are performed, setting the significance level at .0025 for each individual test will ensure the FWE is $\leq .05$. For $m = 100$, the significance level for each individual test needs to be .0005 to ensure controlling the FWE at .05. In summary, the FWE approach is not practical when the number of tests is large.

False Discovery Rate

Benjamini and Hochberg¹⁸ defined the FDR as the expectation of V/R ; in notation, $FDR = E(V/R)$ when $R > 0$, and $FDR = 0$ when $R = 0$ since no hypothesis is rejected. If $m_0 = m$ (all null hypotheses are true), both FDR and FWE are 1 when there is any rejection. In this case, FWE and FDR are equivalent, ie, $\Pr(V > 0) = E(V/R)$. If $m_0 < m$, it can be shown that $\Pr(V > 0) > E(V/R)$.¹⁸ Thus, if the FWE is less than the level of significance α , so is the FDR.

The FDR approach allows the findings to be made, provided that the investigator is willing to accept a small fraction of false-positive findings. It is worth mentioning that in the context of FDR, false rejection of 2 out of 10 rejections ($FDR = .20$) is more serious than

false rejection of 4 null hypotheses out of 100 rejections ($FDR = .04$). The FDR approach can be used to either control FDR¹⁸⁻²² or estimate FDR.²³⁻²⁶ A typical FDR approach estimates the cutoff for the significant hypotheses so that the FDR is controlled at the desired level, say, 5%. On the other hand, for the specified rejection region (either using the P value or the number of rejections as a cutoff criterion), an FDR error probability can be calculated. These two approaches would lead to the same conclusion once the P values are calculated. Below we describe the use of an FDR approach in analyzing a microarray gene expression experiment.

A microarray experiment is conducted to study the changes in gene expressions under different experimental conditions of interest (eg, with or without exposure to a specific drug or toxic compound). The FDR approach is commonly applied to identifying genes that show differences in expression between experimental conditions (differentially expressed genes). Assume an experiment that consists of m genes. For each gene, an appropriate statistical test is performed to determine if its mean expressions are different between experimental conditions. Let $p_{(r)}$ denote the r -th smallest P value. For the desired FDR level q^* , an FDR-controlled procedure¹⁸ is to find the largest $p_{(r)}$ such that $(m \times p_{(r)})/r \leq q^*$. Those r genes with P values $\leq p_{(r)}$ are then selected as differentially expressed genes. Conversely, if r genes are selected, then an empirical FDR estimate is $(m \times p_{(r)})/r$, which is called the q value.

Table 3 provides a numerical example to illustrate the FWE and FDR approaches with $m = 20$. At the 5% significance level, the FWE approach (row 2 of Table 3) identifies three significant genes (genes 1, 2, and 3) and the FDR approach (row 3 of Table 3) identifies five significant genes (genes 1 through 5).

If the number of true null hypotheses, m_0 , were known, an improved FDR estimate of $(m_0 \times p_{(r)})/r$ could be used. Since m_0 is not typically known, m is used in practice to ensure proper control of error rates. Several statistical methods for estimation of the number of true null hypotheses have been considered by Hsueh et al,²⁷ which could be used to gain additional rejections.

Both the FDR and FWE approaches emphasize controlling the false-positive error. When the experimental objective is to develop genetic profiles, many genes that are involved in the complex functional relationship with other genes might merely have moderate differences in expressions between experimental conditions. Because of the large number of genes and an effort to maintain a low significance level, an FDR approach often results in a short list of significant genes. In the development of a classification algorithm aiming at therapeutic predictions, some genes that have good prediction powers might not be captured in the list. In these genomic/genetic applications, both the false-positive error and false-negative error are of concern. Such applications require a procedure that is capable of selecting a large number of potentially differentially expressed genes. Delongchamp et al²⁸ proposed a receiver operating

Table 3. — Numerical Example to Illustrate the FWE and FDR Approaches With $m = 20$

Gene	1	2	3	4	5	6	7	8	9	10
$p_{(r)}$.0002	.0004	.0016	.0057	.0091	.0187	.0225	.0364	.0441	.0473
$p_{(r)}^{adj}$.0040	.0080	.0320	.1140	.1820	.3740	.4500	.7280	.8820	.9460
$q_{(r)}$.0040	.0040	.0107	.0285	.0364	.0623	.0643	.0910	.0946 ^a	.0946 ^a

Gene	11	12	13	14	15	16	17	18	19	20
$p_{(r)}$.0536	.0779	.0862	.1081	.2341	.3570	.4682	.6420	.6833	.8248
$p_{(r)}^{adj}$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$q_{(r)}$.0975	.1298	.1326	.1545	.3121	.4463	.5508	.7133	.7193	.8248

The Table shows observed P values ($p_{(r)}$), Bonferroni-adjusted FWE P values ($p_{(r)}^{adj}$), and FDR q values: where $p_{(r)}^{adj}$, the Bonferroni-adjusted P value, is the observed P value times the number of tests bounded above by 1 (ie, $p_{(r)}^{adj} = \min\{1, m \times p_{(r)}\}$), and the q value is $(m \times p_{(r)})/r$.

^a $q_{(9)} = (m \times p_{(9)})/r = .098$, and $q_{(10)} = .0946$; due to monotonicity constraints, $q_{(9)} = q_{(10)}$.

The P values in bold are the right-most ones in the row $< .05$ and indicate the number of significant genes identified by the two multiple comparison methods.

characteristic (ROC) approach in determining an optimal cutoff based on minimizing the total cost from making false-positive and false-negative errors. This approach requires estimates of m_0 and m_1 .

Example

Alon et al²⁹ presented an analysis of gene expression in 40 tumor and 22 normal colon tissue samples with 2,000 human genes. This data set is used to illustrate an analysis of identifying differentially expressed genes with 2,000 tests performed. The normal and tumor sample groups were compared using a permutation t test with unequal variances. The permutation test was used since the procedure did not assume the data were normally distributed. The P values were computed based on 500,000 permutations.

Using the Bonferroni approach to identifying differentially expressed genes at the significance level FWE = .05, the corresponding significance level for the individual hypothesis is $\alpha = .05 / 2000 = 2.5 \times 10^{-5}$. The number of genes selected is 17. With the FDR approach using the significance level of $q^* = .05$, the Benjamini and Hochberg¹⁸ procedure searched for the largest r such that $p_{(r)} \leq r \times .05 / 2000$. The number of genes selected is 110. The FDR approach can identify much more genes than the FWE approach. For those 17 genes identified by the FWE approach, the probability that each of the 17 genes is a false positive is $< .05$. On the other hand, among the 110 genes identified by the FDR approach, the expected number of false positives is about $110 \times .05 = 6$. Unfortunately, the genes that are false positives cannot be identified. Finally, simply using the $\alpha = .01$ as the level of significance, the number of significances is 203. In this case, the expected number of false positives is 20, which corresponds to an FDR of .099.

Statistical analysis of gene expression data often involves identifications of a subset of genes that are differentially expressed among different sample groups, eg, between drug treatment and no drug treatment. To

select differentially expressed genes, an investigator often uses a multiple testing criterion to decide a cutoff threshold, which depends on the sample size and the study purpose. One important application of microarray experiments is the development of biomarker classifiers (prediction model) for safety assessment, disease diagnostics and prognostics, and prediction of response for patient assignment. Because of heterogeneity in patient populations and complexity of the disease and genetic and genomic factors, multiple genomic markers often are needed to capture complex relationships with clinical (biological) outcomes. For prediction purposes in genomic/genetic profiling studies, the omission of informative genes in the development of a classifier generally has a more serious consequence on predictive accuracy than the inclusion of noninformative genes. In such cases, the stringent FWE approach to controlling any false-positive error is not essential. The FDR approach, which allows more discoveries of differences to be made, is more useful.

Conclusions

Multiple testing refers to conducting more than one hypothesis test simultaneously to make inferences on different aspects of a problem in a study experiment.

The significance level of a P value is defined under a single test. When more than one test is conducted, simple use of the significance level for each individual test leads to a probability of false-positive findings that is greater than the stated α level.

Two commonly used approaches for choosing a significance level in multiple testing are the FWE and the FDR. The FWE approach controls the probability of making one or more false positives. The FDR approach considers the proportion of significant results that are expected to be false positives. When a study involves a large number of tests, the FDR error measure is a more useful approach in determining a significance cutoff because the FWE approach is too stringent. The FDR approach allows more claims of significant differences

to be made, provided the investigator is willing to accept a small fraction of false-positive findings.

References

1. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics*. 1984;40(4):1079-1087.
2. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics*. 1987;43(3):487-498.
3. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35(3):549-556.
4. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977;64(2):191-199.
5. Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. *Am J Epidemiol*. 1998;147(7):615-619.
6. Assmann SF, Pocock SJ, Enos LE, et al. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064-1069.
7. Cui L, Hung HM, Wang SJ, et al. Issues related to subgroup analysis in clinical trials. *J Biopharm Stat*. 2002;12(3):347-358.
8. Tang DI, Geller NL, Pocock SJ. On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*. 1993;49(1):23-30.
9. Chen JJ, Lin KK, Huque M, et al. Weighted p-value adjustments for animal carcinogenicity trend test. *Biometrics*. 2000;56(2):586-592.
10. Kutner MH, Nachtsheim CJ, Neter J, et al, eds. *Applied Linear Statistical Models*. 5th ed. Boston, MA: McGraw-Hill Irwin; 2004.
11. Dudoit S, Yang YH, Callow MJ, et al. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*. 2002;12:111-139.
12. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci*. 2003;18(1):71-103.
13. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6:65-70.
14. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75(4):800-802.
15. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. New York, NY: John Wiley & Sons Inc; 1987.

16. Westfall PH, Young SS. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York, NY: John Wiley & Sons Inc; 1993.
17. Saville DJ. Multiple comparison procedures: the practical solution. *Am Stat*. 1990;44(2):174-180.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289-300.
19. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J Stat Plan Inference*. 1999;82(1-2):171-196.
20. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J Educ Behav Stat*. 2000;25(1):60-83.
21. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4):1165-1188.
22. Storey JD. A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol*. 2002;64(Pt 3):479-498.
23. Tsai CA, Hsueh HM, Chen JJ. Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics*. 2003;59(4):1071-1081.
24. Pounds S, Cheng C. Improving false discovery rate estimation. *Bioinformatics*. 2004;20(11):1737-1745.
25. Cheng C, Pounds SB, Boyett JM, et al. Statistical significance threshold criteria for analysis of microarray gene expression data. *Stat Appl Genet Mol Biol*. 2004;3(1):Article 36.
26. Allison DB, Gadbury GL, Heo M, et al. A mixture model approach for the analysis of microarray gene expression data. *Comput Stat Data Anal*. 2002;39(1):1-20.
27. Hsueh HM, Chen JJ, Kodell RL. Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *J Biopharm Stat*. 2003;13(4):675-689.
28. Delongchamp RR, Bowyer JF, Chen JJ, et al. Multiple-testing strategy for analyzing cDNA array data on gene expression. *Biometrics*. 2004;60(3):774-782.
29. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*. 1999;96(12):6745-6750.

Appendix 1. — A Glossary of Key Terms Used for Statistical Inference

GLOSSARY

Alternative hypothesis: a hypothesis alternative to the null hypothesis; the study objective, to be proved or disproved, is generally designated as the alternative hypothesis.

Classification: a procedure to discriminate individual class's membership based on inherent information on one or more characteristics.

Correlation coefficient: a measure of interdependency between two variables.

Effect size: the targeted difference between the parameter of two populations designed to detect.

Expectation: the mean value in repeated sampling.

False discovery rate (FDR): the expected proportion of false rejections among the rejected null hypothesis.

Family-wise error rate (FWE): the probability of making at least one erroneous rejection among all hypotheses tested.

Hypothesis test: a scientific method driven by a data-based rule to determine whether to accept the null hypothesis or to reject it in favor of the alternative hypothesis.

Level of significance: the upper bound on probability of type I error, which is usually a small number, eg, .01, .05.

Multiplicity test: multiple hypotheses tested simultaneously in an experiment.

Null distribution: the probability distribution of the test statistic when the null hypothesis is true.

Null hypothesis: the particular hypothesis under test.

P value: the observed significance level.

Parameter: a numerical characteristic of a population.

Permutation test: A statistical significance test in which the null distribution is obtained by calculating possible values of the test statistic by rearrangements of the group labels of the data.

tion is obtained by calculating possible values of the test statistic by rearrangements of the group labels of the data.

Power of a test: the probability of rejecting null hypothesis when it is in fact false.

Random variable: a quantity that may take any of the values of a specified set with a specified relative frequency or probability.

Receiver operating characteristic (ROC): a graphical plot of the sensitivity vs (1 – specificity) by varying the level of significance.

Sample size: the number of experiment units (typically, the number of biological samples in the experiment).

Sensitivity: the proportion of correct positive predictions out of the number of true-positive samples.

Specificity: the proportion of correct negative predictions out of the number of true-negative samples.

Standardized variable: a variable transformed to have mean 0 and standard deviation 1.

Statistic: a summary numerical characteristic calculated from a sample that is used to infer the target parameters.

Statistical hypothesis: an assertion or conjecture about the probability distribution for the designated population and its relationship to study objective.

Test statistic: a statistic derived from the data to decide to either accept or reject the null hypothesis based on null distribution.

Type I error: an error incurred by rejecting a null hypothesis when it is in fact true.

Type II error: an error incurred by accepting a null hypothesis when the alternative hypothesis is in fact true.