Closed Testing procedure for multiplicity control. An application on oxidative stress parameters in Hashimoto's thyroiditis

Angela Alibrandi [1]

(1) University of Messina, Department of Economics, Unit of Statistical and Mathematical Sciences, Messina, Italy

CORRESPONDING AUTHOR: *Angela Alibrandi, University of Messina, Department of Economics, Unit of Statistical and Mathematical Sciences, Via dei Verdi, 75 – 98122 Messina, Italy. Telephone: +39 090 6767108 Fax: +39 090 771005; e-mail: aalibrandi@unime.it*

DOI: ***10.2427/12056***
Accepted on September 10, 2016

ABSTRACT

Background: Closed Testing procedures represent an effective solution to the need to make inferences on multiple aspects at the same time, controlling the *Familywise Error Rate* (FWER), that is the error rate of the hierarchical family. Closed Testing procedures have a high degree of adaptability to a wide range of experimental situations, both in parametric than in non-parametric ambit.
Methods: The attention is focused on the Bonferroni-Holm method, frequently used to counteract the problem of multiple comparisons. The present paper aims to show an original application of the Closed Testing procedures for multiplicity control in medical research with reference to the oxidative stress; in particular the Min-P Bonferrroni-Holm method was applied to the p-value adjustment, related to three parameters (BAP, D-ROMS, AGEs) of oxidative stress in Hashimoto's thyroidytis.
Results: Comparisons between different patients are performed (cases vs controls, AbTg positive vs negative patients, AbTPO positive vs negative patients, normal vs high-normal TSH serum levels). Looking at the raw and the adjusted p-value, the Closed Testing procedure is slightly conservative in controlling type I error.
Conclusion: Closed Testing procedure checks the multiplicity, controlling type I error, increasing the probability of accepting the null hypothesis.

Key words: familywise error rate, multiplicity control, bonferroni-holm method, oxidative stress.

INTRODUCTION

In the study of very complex phenomena, the statistician often needs to make inferences about most aspects of a problem. In order to make inference on multiple aspects at the same time, the global error must be controlled. In this context there is a need for a procedure that allows a decision, through the joint use of univariate and multivariate tests. The CLOSED TESTING procedures, firstly proposed by Marcus [1], represent a simple and effective solution to this issue.

In [2] and [3] the above-mentioned data analysis procedure is thoroughly discussed; the authors emphasized that the Closed Testing methods are among the most powerful multiple inference methods and are quickly gaining significant acceptance and popularity. The article widely explains the methodology, the conditions on which it is based and the tests by which it can be carried out.

In [4] the authors underline the high degree of adaptability of the Closed Testing procedures to a wide range of experimental situations, both in parametric than in non-parametric ambit; peculiarly the authors introduced, in the Closed Testing procedure, non-parametric permutation methods [5], [6], [7], [8], [9] and made a comparison between the different methods, in terms of robustness and power. The authors also point out how the permutation methods are particularly suitable for use of Closed Testing procedures, for three particular reasons:

- a greater robustness of the partial and combined permutation tests, compared to the parametric tests, especially in conditions of non-normality;
- the opportunity to have combined tests which can take into account the structure of dependence between variables, without it being formally explicated;
- the possibility of evaluating systems by directional or not directional hypotheses, characterized by a large cardinality of the components hypotheses.

In [10] the attention is focused on multivariate multiple comparisons for multiplicity control in the non-parametric permutation context. In particular the authors illustrate the selection criteria of a function to combine the p-value associated with minimal hypothesis test.

Recent contribution is due to [11] that realize a systematic comparison of methods for combining p-values from independent tests.

The present paper aims to show an original application of the Closed Testing procedures for multiplicity control in medical research, with reference to the oxidative stress; in particular the Min-P Bonferroni-Holm method was applied to the p-value adjustment, related to three parameters (BAP, D-ROMS, AGEs) of oxidative stress in Hashimoto's Thyroiditis.

METHODS

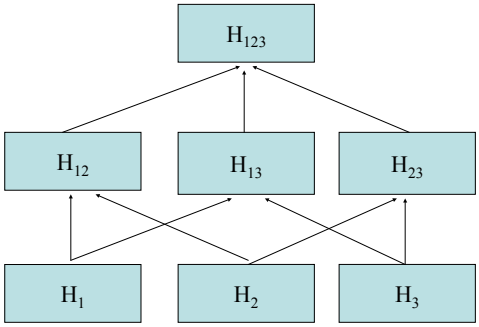
The Closed testing methodology

When multiple tests are used (when comparing two or more groups), in the context of univariate and multivariate distributions, consider a family of distinct hypoteses H_i: $\omega \in O_i$, $i \in I$, where O_i is a proper subset of O and I is the set of indices. The hypothesis $H_0 = \cap_i H_i$ Hi is defined *global hypothesis*. If H_i implies H_j, ($H_i \rightarrow H_j$), then H_j is a own component of H_i and an implication relation exists between H_i and H_j. The hypotheses that do not have own component are called *minimal*; they are referred to the pairwise comparisons; the hypotheses that contain own components are called *not minimal*.

The *Hierarchical Family* is a family of hypotheses, where at least one implication relationship exists.

Figure 1 shows the structure of the hierarchical hypotheses with three minimal hypotheses.

FIGURE 1. Structure of the hierarchical inclusions in Closed Testing procedure.



When the existence of significant differences between groups is assessed, it is necessary that inferences check, at the fixed α level, the value of *Familywise Error Rate* (FWER), that is the error rate of the hierarchical family. The FWER is the probability of making at least an univariate first type error or the probability of making a multivariate first type error; therefore, a multiple test procedure with $C_1,\ldots, C_k$ critical regions has to be applied in order to test the null hypotheses $H_{01},\ldots, H_{0k}$ in which the probability of first type error is less than or equal to α, so that controls the FWER (when $H_{01},\ldots,H_{0k}$ are true). The goal of Closed Testing methods is to create a procedure that is characterized by the properties of coherence and, possibly, of consonance and for which the experimental error does not exceed α. A multiple testing procedure for a hierarchical hypotheses family enjoys two important properties:

- **coherence properties**: if, given any pair of hypothesis (H_i, H_j), such that H_j is included in H_i, the acceptance of H_j implies the acceptance of H_i;
- **consonance properties**: if, when a non-minimal hypothesis H_i is refused, there is at least a minimal hypothesis that must be refused.

In Closed Testing procedures the coherence properties are required, whereas the consonance properties are desirable.

A fundamental characteristic of the Closed Testing is to refer to a set of statistical hypotheses that are closed with respect to the intersection and for which each test (associated to them) has α level. In fact, given a hypotheses family $\{H_i\ (1\leq i\leq k)\}$, the "closure" of the set refers to the set $H_p = \cap_i \in p\ H_i$, $p \in 1,\ldots,k$ of all non-empty intersections of H_i, with i=1,…,k.

In [1] the authors demonstrated that Closed Testing procedure controls FWER at fixed α level.

In Closed Testing the adjusted p-value, related to a certain hypothesis H_i is equal to the maximum of the p-values associated to hypotheses that include Hi [12]. In order to test composed hypotheses several methods were

proposed in literature. Certainly, the choice of adequate test for minimal and composed hypotheses influences the power procedure.

There is not a unique method that is the best in all situations; the choice of these tests depends on the nature of alternative hypothesis that has to be verified. The applicability of the Closed Testing is tied to the use of tests that have to be consistent and unbiased; among these, two tests have to be mentioned for the advantage of being released from the knowledge of the dependency structure between the minimal hypotheses tests: Bonferroni test and Simes test for composed hypotheses (see [13] for methodological deepening). In this perspective, Abdi [14] focus the attention on Bonferroni and Šidák corrections for multiple comparisons.

If the researcher aims at evaluating the difference between two independent samples on *n* variables, n hypotheses are formulated, in the comparison between groups; he must verify the *n* minimal hypotheses at the significance α level by means of adequate tests, such as Student t-test or a non parametric test [15]; alternatively, non-parametric tests based on sampling of the permutation space can be used [8]; they offer the advantage of including the effects of the dependence structure between variables, without the need to directly estimate it. The "closed" set is created, i.e. the set of all possible composed hypotheses; each hypothesis is tested through an appropriate test. Simple H_i hypothesis is rejected if the simple test is significant and if the intersection of each test that includes H_i are significant.

After determining, in this way, the p-value of minimal and composed hypotheses, the Closed Testing adjustment can be made by considering, for a H_i hypothesis, the maximum among the p-value of the hypotheses that include H_i [16]. Among the procedures on several levels this paper focuses on the "Sequentially Rejective Bonferroni Procedure", known as "MinP Bonferroni-Holm Procedure", proposed by Holm [17].

Minp Bonferroni-Holm Method

The Bonferroni-Holm method is a method used to counteract the problem of multiple comparisons; it is intended to control the Familywise Error Rate and offers a simple test which is uniformly more powerful than the Bonferroni correction. It is one of the earliest usages of stepwise algorithms in simultaneous inference. It applies the Bonferroni method to generate a step-wise procedure, as follows:

1. for each single hypothesis H_k (k=1,…,K) the significance of a t-test for two independent samples is calculated and the vector of significance, arranged in increasing $p_{(1)},\ldots,p_{(k)}$ is thus determined;
2. if $p_{(k)} \geq \alpha/k$, we have to accept $H_1,\ldots,H_K$ and the algorithm stops; otherwise we have to reject the global hypotheses and proceed;
3. if $p_{(k-1)} \geq \alpha/(k-1)$, we have to accept $H_{(1)},\ldots,H_{(K-1)}$ and the algorithm stops, otherwise we have to reject $H_{(K-1)}$ and proceed;
4. the process is repeated as the previous step, verifying if $p_{(k-i)} \geq \alpha/(k-i)$ at each subsequent step. The algorithm stops at the first tested inequality. Similarly, the composed hypothesis is rejected when (K-i) $p_{(K-i)} \geq \alpha$.

This procedure requires to calculate only p-value associated with minimal assumptions.

Supposing we want to test hypotheses H_1, H_2, and H_3, the closed testing procedure works as follows:

1. Test each hypothesis H_1, H_2, H_3 using an appropriate α-level test.
2. Create the "closure" of the set, which is the set of all possible intersections among H_1, H_2, H_3, in this case the hypotheses H_{12}, H_{13}, H_23, and H_{123}.
3. Test each intersection using an appropriate α-level test.
4. We may reject any hypothesis Hi, with controls FWER, when the following conditions both hold:
   - the test of H_i itself yields a statistically significant result, and
   - the test of every intersection hypothesis that includes H_i is statistically significant.

The p-value associated to multivariate assumption H_{123} (resulting from the combination of the three investigated variables) is calculated multiplying the lesser of the p-value of minimal assumptions (for example, the one associated to the third variable) by the number of included hypotheses.

Let α=0.05, if the adjusted p-value is less than the fixed significance level, the H_{123} hypothesis must be rejected together with all hypotheses of which H_3 is component, i.e. H_3, H_{13} and H_{23}.

After adjustment for the Closed Testing, at H_3 minimal hypothesis we associated a significance of 3min*P*, where min*P* represents the minimum p-value considered in the combination. Subsequently, the intersection of hypotheses not yet rejected should be evacuate, using the above-described procedure, i.e. multiplying by 2 the low p-value of minimal assumptions, following the smallest ever (already used) … and so on.

Holm–Bonferroni method is uniformly more powerful than the classic Bonferroni correction. There are other methods for controlling the family-wise error rate that are more powerful than Holm-Bonferroni. Among those we have to cite the Hochberg and Hommel procedures [18]. However, the Hochberg procedure requires the hypotheses to be independent or under certain forms of positive dependence, whereas Holm-Bonferroni can be applied with no further assumptions on the data.

RESULTS

The above-mentioned multivariate methodology was applied to the oxidative stress parameters. Oxidative stress, which occurs as a result of an imbalance between free radicals production and antioxidant defence mechanisms, has been implicated in the pathogenesis of several autoimmune disorders, including thyroid diseases.

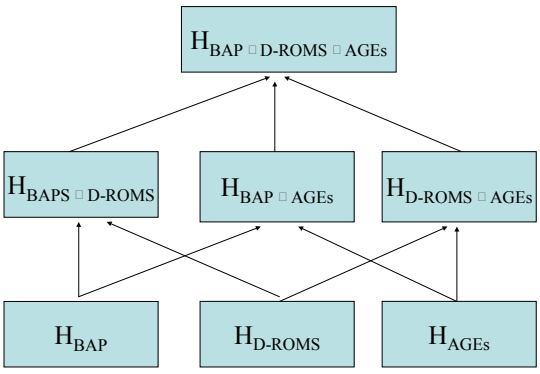
Using the dataset previously analysed by [19], three markers of oxidative stress in Hashimoto's Thyroiditis (HT) were examined:

1. Biological Antioxidant Potential (BAP)
2. Reactive Oxygen Metabolites (D-ROMs)
3. Advanced Glycation End Products (AGEs).

In the study 134 euthyroid subjects were included: 71 newly diagnosed HT patients (8 Male e 63 Female; mean age 38±13 yr) and 63 age and sex-matched healthy controls.

Figure 2 shows the reference scheme for the application of Closed Testing procedure to oxidative stress, with reference to the three above-mentioned variables.

FIGURE 2. Structure of the hierarchical inclusions for oxidative stress



The Closed testing procedure was applied using the "MINp Bonferroni Holm" method for multiplicity control.

In the following illustrative chart, realized for each comparison, the minimal hypothesis 1 refers to BAP, the minimal hypothesis 2 to D-ROMS and finally the minimal hypothesis 3 to AGEs. The figures show the p-value associated with each hypothesis (minimal, not minimal and multivariate) and the adjusted p-value (denoted by Adj.) after correction by Closed Testing procedure, for comparison between:

- cases and controls (Figure 3);
- AbTg positive and negative patients (Figure 4);
- AbTPO positive and negative patients (Figure 5);
- normal and high-normal TSH serum levels (Figure 6) setting a cut-off value for TSH of 4.2 mIU/L, according to Mayo Clinic (one of the leading global research institutions).

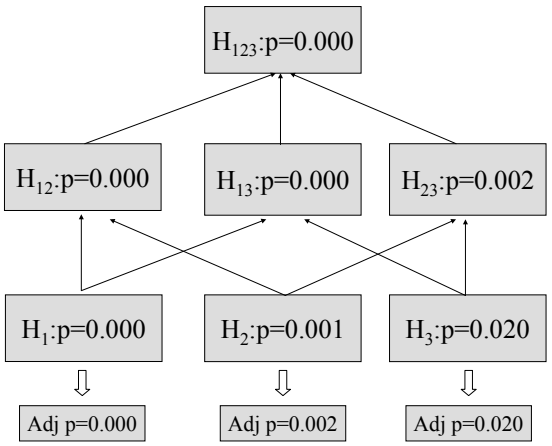
Table 1 shows Mean ±SD of the three variables, in different patient groups.

In order to test each minimal hypothesis H_1, H_2, H_3 the NPC test (based on permutation solution) was applied, for the optimal properties which characterize it [8]. In Figure 3 we reported the results of Closed Testing procedure to compare cases and controls.

TABLE 1. Mean ±SD of oxidative stress parameters according to the group.

GROUPS	BAP	D-ROMS	AGES
Controls	3380,3±873,4	267,3±69,6	189,6±72,1
Cases	2496,5±774,8	339,2±92,6	223,2±86,8
Neg.Abtg	3029,5±933,1	293,2±79,7	206,8±81,6
Pos.Abtg	2545,1±877,5	343,8±101,9	204,4±78,5
Neg.Abtpo	3246,8±833,5	281,6±83,5	194,3±73,0
Pos.Abtpo	2420,6±823,9	340,8±89,0	225,2±89,8
Normal TSH	2905,7±931,8	304,7±89,0	202,6±79,8
High TSH	2589,0±875,3	348,3±98,1	234,5±86,4

FIGURE 3. Closed Testing procedure for comparison between cases and controls



In the comparison between cases and controls, the minimal hypotheses (related to the three examined variables) and the multivariate hypotheses H_{123} are rejected at the fixed significance level. Statistically significant differences exist between cases and controls for the three parameters of oxidative stress Bap, D-ROMS and AGEs, even after correction using Closed Testing. Figure 4 shows the results of Closed Testing for the comparison between Abtg positive and negative patients.

In the comparison between AbTg positive and negative patients, the minimal hypotheses $H_{1(BAP)}$ ed $H_{2(dROMS)}$ are rejected, revealing the existence of significant differences between the groups, while H_3 hypothesis concerning AGEs is accepted. The multivariate hypothesis H_{123} is rejected, since the p-value is significant, even when adjusted by Closed Testing procedure.

Figure 5 illustrates the results of Closed Testing to control multiplicity in the comparison between AbTPO positive and negative patients.

In the comparison between AbTPO positive and negative patients, all minimal hypothesis (related to the three examined variables) and the multivariate hypotheses H_{123} are rejected at the significance level α=0.05, indicating the existence of significant differences between groups.

In the comparison between patients with high-normal (TSH+) and normal (TSH-) thyroid-stimulating hormone serum levels (Figure 6), all minimal hypotheses and the multivariate hypothesis H_{123} are accepted. For minimal hypothesis H_2, related to D-ROMS, we note a dissimilarity between the raw p-value (that results significant) and the adjusted p-value (that is not significant), highlighting the low degree of conservativeness of which the Closed Testing procedure is characterized.

From an endocrinological point of view this result is not surprising, considering that all patients are euthyroid, with similar levels of TSH.

FIGURE 4. Closed Testing procedure for comparison between AbTg+ and AbTg-

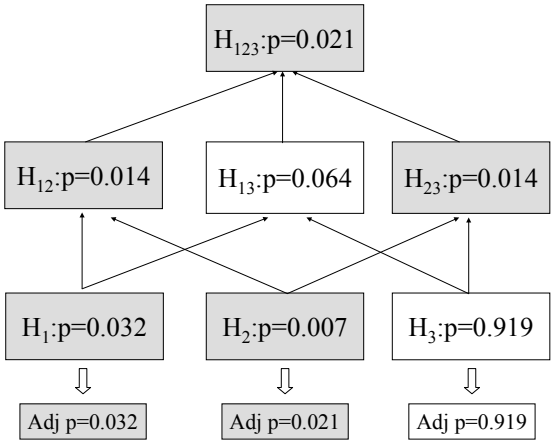


FIGURE 5. Closed Testing procedure for comparison between AbTPO+ and AbTPO-

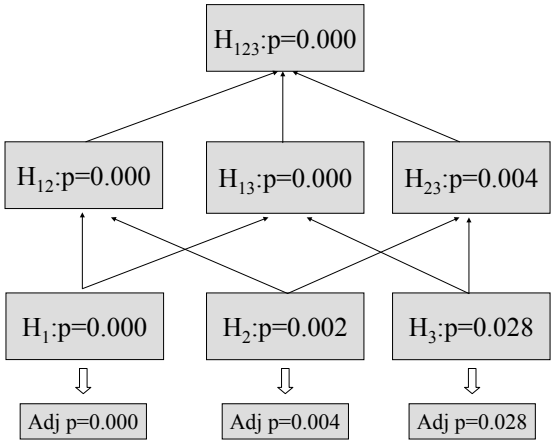
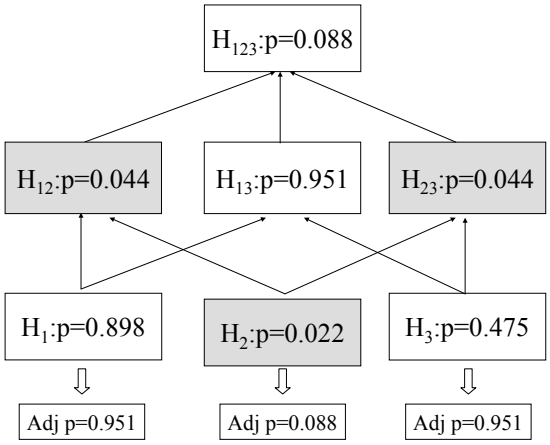


FIGURE 6. Closed Testing procedure for comparison between TSH+ and TSH-



DISCUSSION

The analysis of multiple outcomes is becoming increasingly common in biomedical research. For this, it's needed taking into account the dependence structure among outcomes, that affects the difference between "unadjusted" and "adjusted" p values. Unlike raw (or unadjusted) p-values, adjusted p-values (deriving from combined tests) incorporate the underlying correlation structure among variables, without it being made formally explicit. In particular, «Bonferroni - Holm minp test is very conservative, especially when the correlation structure among variables is strong » ([9] p.185).

Very frequently in biomedical research we need to compare different groups of patients with reference to variables related each other; for example total cholesterol, HDL, LDL, etc; here the multiplicity control becomes essential because raw p-value and adjusted p-value could lead to different decisions: the raw p-value generally leads to reject a null hypothesis, adjusted p-value leads to accept it.

In addition, adjustment procedure guarantees the possibility to evaluate system made up hypotheses characterized by a large cardinality of the component hypotheses, too; therefore, in presence of a lot of variables the adjustment is necessary, because it allow to identify the only really significant variables; so, when we are often in presence of a high number of variables in medical studies, the statistician has to apply an adequate procedure for multiplicity control (such as the Sequentially Rejective Multiple Test [17]), that allows to determine which of the partial tests are effectively significant, providing a more reliable interpretation of the results. In this regard, we cite [20]; here the authors propose, as an optimal solution to the multiplicity problem, the Closed

Testing procedure that maintains fixed the α, as multiple error level.

CONCLUSION

This paper proposes an original application of the Closed testing procedure (by use of min-p Bonferroni-Holm method) to three parameters of oxidative stress in a population of patients affected by Hashimoto's Thyroiditis. Comparisons between different patients are performed and, for each of them, the raw and the adjusted p-values are shown. The results allow to highlight the utility of Closed testing procedure in controlling type I error. Looking at the raw and the adjusted p-value, we can note that the Closed Testing procedure is slightly conservative, because it leads to accept the null hypothesis. In fact in the comparison between normal and high-normal TSH serum levels, for only D-ROMS, the raw p-value is significant, but the adjusted p-value is not statistically significant at the fixed α level. Generalizing, Closed Testing procedure checks the multiplicity, controlling type I error, since it increases the probability of accepting the null hypothesis, such as showed in this application.

Closed Testing procedures offers strong control of FWER. For the final inferences, an elementary null hypothesis Hi is rejected if, and only if, its corresponding test is significant at α level, and every other hypothesis in the family that implies it is rejected by its α level test.

Finally, this article aims at encouraging the use of the Closed Testing procedure in medical research since it is preferable to other correction procedures, because it controls the multiplicity, very often recurrent in medicine [21], [22], [23], ensuring the observance of global α level.

Acknowledgments

This study was not possible without the clinical endocrinology scientific skillness and the unvaluable and friendly cooperation of Dr. Rosaria Maddalena Ruggeri, MD, PhD who permitted the access to her database.

REFERENCES

1. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 1976; 63, 655-60.
2. Westfall PH, Young SS. Resampling Based Multiple Testing. Examples and Methods for p-value Adjustment. 1993; Wiley and Sons, INC.
3. Westfall PH, Wolfinger RD. Closed Multiple Testing Procedures and PROC MULTTEST. 2000; SAS institute Inc.
4. Finos L, Pesarin F, Salmaso L. Test combinati per il controllo della molteplicità mediante procedure di Closed Testing, Statistica Applicata 2003; 15, 2, 301-29.
5. Pesarin F. A resampling procedure for nonparametric combination of several dependent tests. 1992; Journal Ital. Statist. Soc.
6. Pesarin F. Permutation testing of multidimensional Hypotheses. 1997; Cleup Editrice, Padova.
7. Pesarin F. Multidimensional testing of non parametric hypoteses by conditional resampling tecniques}. 1999; Cleup Editrice.
8. Pesarin F. Multivariate Permutation Test. 2001; Wiley, Chichester.
9. Pesarin F, Salmaso L. Permutation Tests for Complex Data. Theory, Applications and Software. 2010; Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester, UK.
10. Finos L, Pesarin F, Salmaso L. Confronti multipli tramite metodi di permutazione, Statistica Applicata 2003; 15(2): 275-300.
11. Loughin TM. A systematic comparison of methods for combining p-values from independent tests. Computational Statistics & Data Analysis 2004; 47(3): 467–85.
12. Rom DM, Holland B. A new closed multiple testing procedure for hierarchical families of hypotheses, Journal of Statistical Planning and Inference n°3, 1995; 265-75.
13. Simes RJ. An improved Bonferroni procedure for multiple test of significance, Biometrika, 1986; 73 (3): 751-4.
14. Abdi, H. Bonferroni and Šidák corrections for multiple comparisons. In Salkind, N. J.Encyclopedia of Measurement and Statistics (PDF). 2007; Thousand Oaks, CA: Sage.
15. Blair RC, Higgins JJ, Karniski W, Kromrey JD. (1994), A study of multivariate permutation test which may replace Hotelling's T2 test in prescribed circumstances, Multivariate Behavioral Research 1994; 29:141-63.
16. Calinski T, Lejeune M. Dimensionality in MANOVA tested by a closed testing procedure, Journal of Multivariate Analysis, 1998; 181-94.
17. Holm S. A simple sequentially rejective multiple testing procedure, Scandinavian Journal of Statistics 1979: 6(2): 65–70.
18. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 1988; 75(2):383–386. doi:10.1093/ biomet75.2.383. ISSN 0006-3444.
19. Ruggeri RM, Vicchio TM, Cristani M et al. Oxidative stress and advanced glycation end products (ages) in Hashimoto's Thyroiditis. Thyroid 2016; 26: 504-11.
20. Alt R, Fortin J, Weinberger S. The-Day-of-the-Week Effect Revisited: an Alternative Testing Approach, Reihe Okonomie Economic Series, Institute for Advanced Studies, Vienna, 2002.
21. Westfall PH, Bretz F. Multiplicity in clinical trials. In Chow, S. C., editor, Encyclopedia of Biopharmaceutical Statistics 2010; 2: 889–96. Informa Healthcare, New York, 3rd edition.
22. Hommel G, Bretz F, Maurer W. Multiple hypotheses testing based on ordered p values–ahistorical survey with applications to medical research. Journal of Biopharmaceutical Statistics 2011; 21(4):595–609.
23. Goeman J, Solari A. Multiple hypothesis testing in genomics. Statistics in Medicine 2014; 33 (11), DOI: 10.1002/sim.6082