PhUSE 2016

Paper SP04

Multiplicity Issues in Clinical Trials-Look at it with a magnifying glass!

Madhusudhan Bandi, Hyderabad, India

ABSTRACT

With recent trends in the drug development process, modern clinical trials have gone a step ahead in evaluating safety and efficacy of new treatments with inclusion of multiple objectives in a trial and its interpretation. Multiplicity arises when a trial departs from simple design to complex design because of multiple tests, when one considers a set of statistical inferences simultaneously. Multiplicity issues arise in various situations (multiple endpoints, sub-group analysis, multiple doses or treatments, interim monitoring and with repeated measures data), in dealing with these issues, one need to control error rate by controlling FWER (Family Wise Error rate) or FDR (False Discovery Rate), eventually there will be loss of power. Often more complex and consistent procedures are required to deal with both multiplicity and power. Here we are going to review multiplicity and different procedures to deal with these issues, with strong control on α (alpha) with minimal loss of power.

INTRODUCTION

We have well known study design principles like randomization, which helps in avoiding selection bias in treatment. Stratification helps in grouping patients into sub groups for better power with similar variability for treatment differences. The third principle is blinding; which helps in avoiding assignment of a better treatment to some group of subjects at patient, investigator or sponsor level. Beyond these principles we need to focus on 'Multiplicity' before planning a study. Multiplicity (large number) of inferences is present in almost all clinical trials.

Sources of multiplicity are

1. Multiple endpoints - different criteria in judging efficacy of a therapy.
2. Interim analysis - examining the data part-way of the study and at end of the study.
3. Subgroups - evaluating one or more subgroups in addition to all patients and when one considers many treatment groups for comparing which treatment is better than another, in clinical trial.
4. Repeated measures.

With increase in technology, the availability of data has become easy in many ways. New technologies that are leading to this increase in data range from next generation sequencing machines in molecular biology to imaging of patients in clinical studies or electronic medical records. So everyone is planning to explore so many possibilities out for a single study by considering different hypothesis at a time. With increase of possibilities on the other side of the coin declaring a significant difference when there is no difference is also increasing. However all statistical tests that we do, will run into risk of making mistakes and declaring that a real difference exists when in fact the observed difference is by chance. The risk (type I error) is controlled for each individual single test and that is precisely what is meant by the significance level of the test. Our usual cut off for committing a type I error is 0.05.Suppose if we perform 5 hypothesis tests at a level of 0.05 level of significance for each test we are having total chance for committing an error is 0.226(1-(1-0.05)5) but we are having only 0.05 chance of committing an error for all the 5 tests so we need to explore new possibilities considering whether a test is significant or not. One must consider and address multiplicity issues to have our statistical analysis operate rigorously at designated significance level.

FUNDAMENTALS

Suppose if we are testing hypothesis that a parameter β is equal to 0 versus the alternative hypothesis that β not equal to zero. There is a possible chance of doing two types of errors.

1. Type I error - incorrect rejection of a true null hypothesis.
2. Type II error - failing to reject a false null hypothesis.

Test Result	Truth	
	β = 0	β ≠ 0
β = 0 (Not significant)	Correct Decision (U)	Type II error (T)
β ≠ 0 (Significant)	Type I error(α) (V)	Correct Decision (S)

U, V, T, S are counts of decisions.

ERROR RATES

As we all know that Type I error is more serious in nature, there are different approaches to control Type I error. Some of them are listed below which we are going to dealt with in detail.

False Positive rate: The rate at which false results are called significant E (V/(V+U)).

Family wise error rate (FWER): The probability of at least one false positive P (V >=1).

False discovery rate (FDR): The rate at which claims of significance are false E (V/(V+S)). FDR is designed to control the proportion of false positives (Type I errors) among the set of rejected hypotheses (V+S).

As we have these many types of error measures, in order to get effective results we need to control at least one of these error measures. Here we are going to look at different procedures that will control these error measures and have a glance on those procedures that are useful in different situations.

1.CONTROLLING FALSE POSITIVE RATE

If the p-values are correctly calculated then calling all p-values < α will control the false positive rate at the level of α. Here there is a problem, suppose if we perform 1000 hypothesis tests and all of them are significant (i.e. you are rejecting the entire number null hypothesis that β equal to 0). Here there is a chance of incorrectly rejecting the null hypothesis (expected value of false positives) at the level of 0.05 significance is 50(1000*0.05). Out of 1000 results we are rejecting 50 hypotheses. By controlling the false positive rate we can't control these many false positives so we need to look at FWER which controls for at least one false positive.

2.CONTROLLING FAMILY WISE ERROR RATE

The probability of at least one false positive is the FWER. Suppose if we do m hypothesis tests for which we want to control the FWER at the level of α then we have α/m as a new significance level for calling all the p-values to be significant otherwise not.

Rather than adjusting α for each and every test, it is easy to adjust the p-values which we have calculated. We have direct calculations available to adjust p-values in various statistical tools like R. After adjusting the p-values, they are no longer called as classically defined p-values. Adjusted p-values don't have the properties of classical p-values but they can used to control error measures directly without doing modifications to α.

General approaches to control FWER are

Single step: Adjustments are made to each p-value.

BONFERRONI CORRECTION

1. Calculate the p-values normally for each test.
2. Multiply each p-value by number of tests performed.
3. Call test as significant if adjusted p-value is less than α.

Where adjusted p-value= p-value*n < α (0.05), 'n' is number of tests performed.

Sequential: adaptive adjustments are made to each p-value.

HOLM'S METHOD

1. Calculate the p-values and rank them from smallest to largest.
2. Multiply the p-values by ($n - (r_i - 1)$) from smallest to largest. If adjusted p- value is less than α call significant.

Where adjusted p-value= p-value*($n - (r_i - 1)$) < α (0.05), 'n' is number of tests performed and 'r_i' is rank of corresponding p-value.

Example:
Let n=1000 and α=0.05

Test	Unadjusted p-value	Rank	Adjusted p-value	Significant?
A	0.00001	1	0.00001*1000=0.01 <0.05	Yes
B	0.00003	2	0.00003*999=0.02997 <0.05	Yes
C	0.00005	3	0.00005*998=0.0499 <0.05	Yes
D	0.00006	4	0.00006*997=0.05982 >0.05	No
E	0.00008	5	0.00008*996=0.07968 >0.05	No

Follow that sequence until a test is found to be not significant. Those p-values which found to be significant before a non-significant result found are significant no matter regular p-values (unadjusted p-values) found to be less than α (0.05) value.

WESTFALL AND YOUNG PERMUTATION METHOD

Westfall and Young permutation is a step-down procedure similar to the Holm method, combined with a bootstrapping method to compute the p-value.

Another group of methods that control FWER are Tuckey-Kramer method which is best for possible pair-wise comparisons when sample sizes are unequal and Dunnet's method which is used when we want to compare one group (mainly 'control') with other groups.

FWER is appropriate when you want to guard against any false positives. However, in many cases (particularly in genomics) we can live with a certain number of false positives. As number of tests is increasing the most suitable and relevant quantity is to control FDR.

3.CONTROLLING FDR

Suppose we perform m hypothesis tests, we have calculated the p-values normally. Now if we want to control FDR, order the p-values of hypothesis tests in an ascending order then call p-values which are less than α*(i/m) as significant. For controlling FDR we have BH Procedure.

BENJAMINI & HOCHBERG (BH)

1. Calculate the p-values and rank them from smallest to largest.
2. Multiply the largest p-value by number of tests divided by its rank. If adjusted p-value is less than α call significant.

Where adjusted p-value= p-value*n/r_i < α (0.05), 'n' is number of tests performed and 'r_i' is rank of corresponding p-value.

Example:
Let n=1000 and α=0.05

Test	Unadjusted p-value	Rank	Adjusted p-value	Significant?
A	0.4	1000	0.4*(1000/1000)= 0.4 > 0.05	No
B	0.06	999	0.06*(1000/999)= 0.060 > 0.05	No
C	0.004	998	0.004*(1000/998)= 0.0040 < 0.05	Yes
D	0.0002	997	0.0002*(1000/997)=0.0002 < 0.05	Yes
E	0.0001	996	0.0001*(1000/996)=0.0001 < 0.05	Yes

Start from largest to smallest and call the tests with p-values (adjusted) less than largest significant value as significant, even though the p-values (unadjusted) are not less than significant value (0.05).

Consider a simulated data using R statistical software, to see how FWER and FDR will be controlled using different procedures especially Bonferroni which controls FWER and Benjamini and Hochberg procedure which controls FDR.

CASE STUDY I: without any true positives

Here we have simulated random variables x and y which are independent of each other.

```
set.seed(12345)  ##Setting seed for replication of the test result
pValues <-rep(NA,1000)  ##Creating dummy dataset with NA values
for (i in 1:1000){
  y <- rnorm(20)
  x <- rnorm(20)  ##Generating random normal samples using rnorm function.
  pValues[i] <- summary(lm(y ~x))$coeff[2,4]  ##Building a model y on x and assigning p-
                                                values to the dummy dataset.
}
```

```
sum(pValues < 0.05) ##Counting the p-values which are less than cut-off value.
```

```
[1] 39
```

As x and y are independent, we did not expect any p-value to be less than cut-off value (0.05). But we can see above that the count of p-values less than 0.05 are 39 which are false positives.

Now we will control FWER by Bonferroni method and FDR by BH method and count the number of p-values less than the cut-off after adjustment.

```
sum(p.adjust(pValues,method="bonferroni") < 0.05) ##Controling FWER by bonferroni method.
```

```
[1] 0
```

```
sum(p.adjust(pValues,method="BH") < 0.05)  ##Controling FDR by BH method.
```

```
[1] 0
```

From above we can see no result is significant after controlling for FWER and FDR, indicates false positives has been controlled by both procedures.

Now we will consider another simulated data which we voluntarily simulate the dependency between y and x variables with half of the data with true positives.

CASE STUDY II: with 50 % true positives

```
set.seed(12345)
pValues <-rep(NA,1000)
for (i in 1:1000){
    x <- rnorm(20)
    if(i <= 500){y <- rnorm(20)}
    else{y<-rnorm(20,mean=2*x)}  ##First 500 beta=0,last 500 beta=2
    pValues[i] <- summary(lm(y ~x))$coeff[2,4]
}

trueStatus <- rep(c("beta eq zero", "beta ne zero"),each=500)  ##Creating another dummy
dataset with two factor values ("beta eq zero", "beta ne zero") of 500 each in 1000
observations data.

sum(pValues < 0.05)  ##Counting the p-values less than the cut-off.
```

```
[1] 518
```

```
table(pValues < 0.05,trueStatus) ##Creating logical table for the two factors.
```

```
       trueStatus
        beta eq zero beta ne zero
FALSE          482            0
TRUE            18          500
```

In the above table 518 p-values are found to be significant. But we have simulated our data for only 500 p-values to be significant. 18 p-values are found to be false positives in this study for our significance level which will bias our inference.

Now we will control FWER and FDR using the same methods as earlier

```
table(p.adjust(pValues,method="bonferroni") < 0.05,trueStatus) ##Controlling  FWER by
bonferroni method
```

```
       trueStatus
        beta eq zero beta ne zero
FALSE          500           18
TRUE             0          482
```

Out of 518 p-values found to be significant, after adjusting with Bonferroni we can see that has been reduced to 482 only but we have 500 significant p-values for which we have simulated our data. This procedure is very stringent and conservative that it is not allowing for even one false positive.

```
table(p.adjust(pValues,method="BH") < 0.05,trueStatus) ##Controlling FDR by BH method
```

```
      trueStatus
       beta eq zero beta ne zero
FALSE          494             0
TRUE             6           500
```

After controlling for FDR we can see 506 are found to be significant, along with 500 significant results this BH procedure predicting 6 false positives results.BH procedure is less conservative and powerful when compared to other procedures which controls FWER in situations where we can live with some false positives along with actual results.

Now we will review sources of multiplicity issues that can happen in a clinical trial and possible recommendations.

MULTIPLE ENDPOINTS

When different outcome measures are used to measure the efficacy of therapy the problem of multiple testing arise. It is rare that only a single measure is used ('once you have got hold of the subject then measure everything in sight') to tell about the efficacy of the therapy given. For example, for a hypertensive study it's routine to record systolic and diastolic blood pressure in sitting, supine, standing and pulse rate. However, separate significance tests on each separate end-point comparison increases the chance of some false positives.

Remedy:

- Bonferroni correction
- Choose primary outcome measure
- Multivariate analysis
- Splitting α value among multiple endpoints

Multiplicity adjustment can be made when there are multiple endpoints by using three different procedures.

1. Gate keeping procedure: This procedure is used when multiple endpoints are grouped into different families based on clinical relevance. In gate keeping procedure the families are tested in a sequential manner for significance at full α (0.05).If primary endpoints are significance then checks for the next family of endpoints (secondary) at full α.

2. **Fixed sequence procedure:** All the endpoints are tested in a fixed sequence order at full α. If any endpoint is not significance in the prior order then the subsequent endpoints are not tested for significance.

3. **Fall back procedure:** In this procedure full α will be split among different pre ordered endpoints (based on clinical relevance).The late ordered end points can be tested if the pre ordered endpoints are not significant.

SUBGROUP ANALYSES

Multiplicity issues arise when we compare different subgroups with in each group of patients, for example when sample of subjects are subdivided on baseline factors, e.g. on age and gender, resulting in four subgroups. Just as with multiple end-points, the chance of picking up an effect when none exists increases with the number of subdivisions.

If the overall results are non-significant and so subgroup analysis is to be done to retrieve significant result out of the study. This would help in planning a hypothesis for future study.

Remedy:

- Bonferroni adjustments
- Analysis of Variance
- Follow-up tests for multiple comparisons

INTERIM ANALYSIS

Mainly interim analysis is to be done to check for protocol compliance, to detect deleterious side effects, to keep interest on a trial and to avoid unnecessary continuation once the treatment differences are obvious. As data will be accumulated periodically it may be desirable to analyse the data to check for the above aspects and again multiple testing problems will arises with repeated tests on interim data which are not independent.

Remedy:

- Group sequential designs
- As Bonferroni is highly conservative procedure and data is getting accumulated for each interim analysis we need a

different procedures to analyse accumulated data. This was explained by Pocock(1983),details are summarized here.

# of interim analysis	α = 0.05	α = 0.01
2	0.029	0.0056
3	0.022	0.0041
4	0.018	0.0033
5	0.016	0.0028
10	0.0106	0.0018
15	0.0086	0.0015
20	0.0075	0.0013

Significance levels required for repeated two-sided significance testing for different α levels.

REPEATED MEASURES

When we measure an outcome (e.g. blood concentration of a metabolite) repeatedly at different time intervals like 2, 4, 6, 8, 24 hours of post dosing repeated measure data will arise. If we have two treatment groups for which this repeated measures has to be evaluated, it's tempting to use two sample t-tests at each time point. By doing multiple t-tests on different time points will give uncertain results without certain adjustments.

Remedy:

- Bonferroni adjustments – these adjustments may be very conservative as the tests will be highly correlated.
- Multivariate analysis for repeated measures

SUMMARY AND CONCLUSION

Bonferroni and Holm's method controls FWER did not account for dependency across tests, thus tend to have lower statistical power than that of bootstrap or permutation resampling methods like Westfall and Young method which accounts dependency across the correlated tests and controls FWER.BH procedure allows more Type I errors than that of the methods that control FWER. So statistical power is more for FDR controlled procedures than that of procedure that control FWER.

FWER is appropriate when you want to guard against ANY false positives. However, in many cases (particularly as n increases like gene expression studies) we can live with a certain number of false positives. In these cases, the more relevant quantity to control is the false discovery rate (FDR).We need to select the required control against the test which better suites your problem.

If you are most afraid of getting stuff on your significant list that should not have been there then controlling FWER is better option. If you are most afraid of missing out on interesting stuff then controlling FDR is better option. The choice of whether the study aims to control the FDR or FWER is an important design issue that needs to be specified prior to the data analysis.

REFERENCES

1. www.nickfieller.staff.shef.ac.uk/sheff-only/clinical.pdf
2. jtleek.com/genstats_site/lecture_notes/03_13_Multiple-testing.pdf
3. www.coursera.com
4. Statistical considerations for multiplicity in confirmatory protocols.pdf

ACKNOWLEDGEMENTS

I would like to acknowledge my colleagues who have helped me in write up and Cytel for Financial support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Madhusudhan Bandi
Cytel Statistical software & services pvt. Ltd,
Level-3, Wave rock, Nanakramguda,
Hyderabad, Telangana – 500008.
Ph.: +91 40 6635 0449
E-mail:madhusudhan.bandi@cytel.com