

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

Madhusudhan Bandi, Cytel

Agenda

- What is multiplicity in clinical trials?
- Why multiplicity matters?
- Dealing with multiplicity
- Different sources of multiplicity and remedies
- Conclusion

What is Multiplicity?

Multiplicity - considering large number of questions to be answered at a time.

Sources of multiplicity:

- Multiple endpoints
- Interim analysis
- Subgroup analysis
- Repeated measures

Why Multiplicity matters?

- Consider 5 hypothesis tests, performed at 0.05 level of significance.

$$P(\text{Making an error}) = 0.05$$

$$P(\text{Not making an error}) = 1 - 0.05 = 0.95$$

$$P(\text{Not making an error in 5 tests}) = (1 - 0.05)^5 = 0.774$$

$$P(\text{Making at least 1 error in 5 tests}) = 1 - (1 - 0.05)^5 = 0.226$$

Types of errors

Test Result	Truth	
	H_0 True	H_0 False
H_0 Accepted	Correct Decision (U)	Type II error (T)
H_0 Rejected	Type I error(α) (V)	Correct Decision (S)

$$P(\text{Type I error}) = \alpha$$

U, V, T, S are counts of decisions

Approaches to control Type I error

- **False positive rate:** The rate at which false results are called significant $[V/(V+U)]$.
- **Family wise error rate(FWER):** The probability of at least one type I error $P(V \geq 1)$.
- **False discovery rate(FDR):** The rate at which claims of significance are false $[V/(V+S)]$. FDR is designed to control the proportion of false positives(Type I errors) among the set of rejected hypotheses $(V+S)$.

Controlling False positive rate

- Suppose if we perform 5 hypothesis tests at level of 0.05 significance. Here there is a chance of 0.226 doing an error.
- For these many tests, if we control false positive rate it would not be helpful as we are chance of doing an error is 0.226 in all 5 tests.

Controlling Family Wise Error Rate (FWER)



- In FWER, we are controlling for at least one Type I error to happen by chance.

Ways to control FWER

1. **Single step:** adjustments made to each p-value
2. **Sequential:** adaptive adjustments made to each p-value

Single step: Bonferroni correction

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

Call significant: Adjusted p-value = $p\text{-value} * n < \alpha$

Where 'n' is number of tests performed.

Single step: Bonferroni correction

Let $n=5$ and $\alpha=0.05$

Test	Unadjusted p-value	Adjusted p-value p-value*n	Significant?
A	0.001	$0.001*5=0.005 < 0.05$	Yes
B	0.003	$0.003*5=0.0125 < 0.05$	Yes
C	0.029	$0.029*5=0.145 > 0.05$	No
D	0.045	$0.045*5=0.225 > 0.05$	No
E	0.060	$0.060*5=0.30 > 0.05$	No

Sequential: Holm's method

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

Call significant: Adjusted p-value = $p\text{-value} * (n - (r_i - 1)) < \alpha$

Where 'n' is number of tests performed and 'r_i' is rank of corresponding p-value.

Sequential: Holm's method

Let $n=5$ and $\alpha=0.05$

Test	Unadjusted p-value	Rank	Adjusted p-value $p\text{-value} \times (n - (r_i - 1))$	Significant?
A	0.001	1	$0.001 \times 5 = 0.005 < 0.05$	Yes
B	0.003	2	$0.003 \times 4 = 0.012 < 0.05$	Yes
C	0.029	3	$0.029 \times 3 = 0.087 > 0.05$	No
D	0.045	4	$0.045 \times 2 = 0.090 > 0.05$	No
E	0.060	5	$0.060 \times 1 = 0.360 > 0.05$	No

Who Cares about not making ANY Type I errors?

- FWER - To guard against ANY false positives
- Genomics and Microarray studies we can live with a certain number of false positives
- False discovery rate (FDR) is designed to control the proportion of false positives(V) among the set of rejected hypotheses ($V+S$).

Benjamini & Hochberg (BH)

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

Call significant: Adjusted p-value = $p\text{-value} * (n/r_i) < \alpha$

Where 'n' is number of tests performed and 'r_i' is rank of corresponding p-value.

Benjamini & Hochberg (BH)

Let $n=5$ and $\alpha=0.05$

Test	Unadjusted p-value	Rank	Adjusted p-value p-value*(n/r_i)	Significant?
E	0.060	5	$0.060*(5/5)=0.06 > 0.05$	No
D	0.045	4	$0.045*(5/4)=0.056 > 0.05$	No
C	0.029	3	$0.029*(5/3)=0.048 < 0.05$	Yes
B	0.003	2	$0.003*(5/2)=0.0075 < 0.05$	Yes
A	0.001	1	$0.001*(5/1)=0.005 < 0.05$	Yes

Multiple endpoints

Remedy:

- Bonferroni correction
- Choose primary outcome measure
- Multivariate analysis
- Splitting α value among multiple endpoints
 - **Gate keeping procedure**
 - **Fixed sequence procedure**
 - **Fall back procedure**

Subgroup analysis

Remedy:

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

Repeated measures

Remedy:

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

Interim analysis

- Group sequential designs
- As Bonferroni is highly conservative for accumulated data for each interim analysis we need a different procedures this is explained by Pocock (1983)

# of interim analysis	$\alpha = 0.05$	$\alpha = 0.01$
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

Conclusion

- FWER - afraid of getting stuff on your significant list.
- FDR - afraid of missing out an interesting stuff.
- 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.

Thank you