Sometimes when I come across an interesting statistical question in daily work, I would like to take some time to do further research. But the research is always very quick and rough until I present it to the team. The presentation is a good opportunity for me to re-organize what I did in the past and enhance my understanding of the past topic. Today my topic is “Bonferroni or Holm or Hochberg? Multiplicity adjustment method choice based on simulation”.

As we know, the multiplicity issues in clinical trial arises from the multiple comparisons or multiple outcomes, for example, multiple doses or treatments comparison, multiple endpoints, interim analyses, subgroup analyses, use different statistical models. In these cases, type I error will inflate as multiple hypotheses are tested. As the below formulae shows, the probability of at least one positive hypothesis among the k hypotheses conditional on the all null hypotheses are true is larger than alpha level. This means the type I error inflates, just like p value fishing. To control type I error inflating in clinical trial, multiplicity adjustment needs to be considered when design the clinical trial.

Both FDA and EMA have released the draft guidance on multiplicity issues in clinical trial. In China, currently there is a statistical expert consensus paper on multiplicity issue. This paper is written by China clinical trial statistics group, the group members of CCTS includes NMPA reviewers, academic professors and pharmaceutical statisticians. Based on expert consensus and recent development of multiplicity, China NMPA is making the multiplicity guidance and will release it next year.

There are also some great review papers and tutorials on multiplicity, for example, the below three papers published in journal of “statistics in medicine” are valuable to read.

As we know, the reviewers in NMPA sometimes ask the sponsor some questions on the study design and statistical methods. In Feb this year, a China RM asked me to attend a CFDA meeting with her to answer the possible questions on statistical design/methods of a non-inferiority study. I read the protocol and CSR before the meeting. One issue attracted my attention when I read the trial design part. Bonferroni adjustment was used in the study to control familywise type I error rate when there may exist correlations between the two primary endpoints. With hindsight I think this Bonferroni adjustment is too conservative for our study although the primary outcomes both meet the positive criteria.

Let us look at the trial design of this study. The study aims to demonstrate of the efficacy and safety of 3D Digital Breast Tomosynthesis. There are three arms: DBT: 3D Digital Breast Tomosynthesis; MG: 2D Mammography; V-Preview: 2D reconstructed images. And there are three combinational treatments: DBT + MG, DBT + V-preview and MG. MG is a shared control. The study aims to compare the DBT+MG with MG and compare DBT+V-Preview with MG. The primary endpoint are the conspicuity (easily seen or noticed; readily visible or observable) scores collected for DBT + MG and DBT+V-Preview vs MG. (negative 2) -2 means DBT+MG/DBT+V-Preview was clearly less conspicuous than MG. -1 means DBT+MG/DBT+V-Preview was slightly less conspicuous than MG. 0 means DBT+MG/DBT+V-Preview and MG exhibited similar conspicuity. 1 means DBT+MG/DBT+V-Preview was slightly more conspicuous than MG Null Hypothesis. 2 means DBT+MG/DBT+V-Preview was clearly more conspicuous than MG. Null Hypothesis 1：Median of conspicuity score of DBT + MG vs MG <= 0.5 (non-inferiority margin) ; Null Hypothesis 2：Median of conspicuity score of DBT + V-Preview vs MG <= 0.5 (non-inferiority margin) ; The study will be declared to be successful if at least one null hypothesis is rejected, H1 rejected or H2 rejected. And Bonferroni method is used to control overall type I error rate. As we can see, the control is same for both two comparisons, and the DBT + MG and DBT + V-Preview both have the arm DBT. Thus, there may exist correlation between two primary outcomes. Is the Bonferroni adjustment is the most appropriate method for this trial design? If we can design the study again, do we still choose Bonferroni method to adjust multiplicity, or choose a more appropriate method?

As we know, the Bonferroni method tests each hypothesis at alpha/k level and it is appropriate if two primary outcomes or two hypotheses are independent. When there is correlation between two primary outcomes, Bonferroni tends to be conservative. Assume the two hypotheses represented by two circles, the intersect area indicates there are correlation between two hypotheses. The larger correlation is, the larger intersect area is, the more overall type I error is controlled. Take an extreme example, When the correlation between two primary outcomes is 1, the true overall type I error may decrease from 0.05 to approximately 0.025. This will negatively impact the power of the study. When choose the multiplicity adjustment method, ignorance of correlation of multiple outcomes will tend to lose a lot of power and inflate the type II error.

There are many alternatives to Bonferroni adjustment which are less conservative and consider the correlation. In this presentation, I focus on two classical adjustment methods, Holm method and Hochberg method.

Assume there are two hypotheses: hypothesis H1 and hypothesis H2; Assume the two p values, P1 smaller than P2. Assume the pre-specified type I error is 0.05. As the table shows, the rejection rule for Bonferroni is testing each test independently: If p1 smaller than 0.025, reject H1; If p2 smaller than 0.025, reject H2; Holm adjustment begins with the smallest p-value. If p smaller than 0.025, reject H1 and continue; otherwise stop and accept all hypotheses; If p2<=0.05, reject H2; Otherwise, accept H2; Hochberg method begins with the largest p value. If p2<=0.05, accept all hypotheses; Otherwise continue to P2; If p2<=0.05/2, reject H1; As shown in the flowchart, Bonferroni adjustment independently test two hypotheses and do not consider the correlation. In Holm method, the hypotheses with larger p-value is conditional on whether the hypothesis with smaller p-value is rejected or not. Holm method tends to be more powerful than Bonferroni method. Hochberg method begins with largest p-value. Whether the hypothesis with larger p-value is rejected or not, the subsequent test keeps going on. Thus, Hochberg method tends to be more powerful than both Hochberg and Bonferroni method. Take two examples to compare the power of three methods. First example: p1 is 0.03 and P2 is 0.04. Bonferroni accept H1 and H2; Holm accept H1 and H2; Hochberg reject H1 and H2. In this example, the result is same between Bonferroni and Holm and Hochberg is more powerful. Second example: p1 is 0.02 and p2 is 0.04. Bonferroni accept H2 and reject H1. Holm reject H1 and H2. Hochberg reject H1 and H2. In this example, Hochberg and holm are same and they are more powerful than Bonferroni. In summary, Hochberg is more powerful than Holm, Holm is more powerful than Bonferroni. Can we use Hochberg directly if we can design the DBT trial again?

To answer this question, we need to consider the correlation degree and know if Hochberg control the overall type I error to smaller than 0.05. It is difficult to calculate the overall type I error and power when consider the correlation because we can hardly know the analytic mathematical formulae. So I will demonstrate the impact of the correlation degree on the type I error and power of three methods based on simulation. The two primary outcome values range from -2 to 2. In the SAP, there is a assumption that the true central location of score is zero and standard deviation is 2. Since we do not know the correlation degree, let us assume correlation ranges from 0 to 1 by interval 0.05. How to simulate a bivariate outcome sequences with correlation, like -2, -1, 0, 1, 2. I do not know how to simulate this kind of data. I simulate two correlated primary outcomes from a bivariate normal distribution with the same mean values and standard deviation to demonstrate the impact of correlation on type I error control. The simulation setting is as below: Set the seed as 1234567 and set simulation times as 100, 000. Sample size is set as 200, same with the sample size in SAP. Assume the null hypothesis 1 is mean1 equals to 0; the null hypothesis 2 is mean2 equals to 0; Positive correlation is set from 0 to 1 by interval 0.05. Mean vector is set as 0 and 0 to demonstrate the overall type I error and set as 0.4 and 0.4 to demonstrate the power. Standard deviation is set as 2 so the variance / covariance matrix is shown as below.

The successful criteria is at least one hypothesis is rejected.

As the plot shows, the red line is the changing curve of overall type I error of Hochberg adjustment, and the green line is the changing curve of overall type I error of both Bonferroni adjustment and Holm adjustment. In the sense of overall type I error for two primary outcomes, Holm adjustment is same with Bonferroni adjustment. The overall type I error for Holm and Bonferroni adjustment decreases more sharply than Hochberg adjustment. The whole red line is above the green line which means that Hochberg adjustment is more powerful than Bonferroni and Holm method. Let us look at the right table, when the correlation coefficient is 0.7, the type I error is 0.0438, far smaller than 0.05. The type I error decrease more sharply when correlation coefficient is larger than 0.7. From this table, Hochberg seems to control the overall type I error around 0.05, not far from alpha level. However, when the correlation ranges from 0 to 0.3, the type I error is larger than 0.05. It means that the Hochberg procedure is not guaranteed to control the overall type I error rate to alpha level for positively-correlated endpoints. As mentioned in FDA guideline, the Hochberg procedure is known to provide adequate overall alpha-control for independent endpoint tests and also for two positively-correlated dependent tests with standard test statistics, such as the normal Z, student’s t, and 1 degree of freedom chi-square. Various simulation experiments indicate that the Hochberg procedure usually will, but is not guaranteed to, control the overall Type I error for positively-correlated endpoints. Therefore, beyond the aforementioned cases where the Hochberg procedure is known to be valid, its use is generally not recommended for the primary comparisons of confirmatory clinical trials unless it can be shown that adequate control of Type I error rate is provided.

Let us look at the power of three adjustment. As shown in left plot and right table, Bonferroni and Holm method have the same power, and the Hochberg is more powerful than them.

Let us look at another type of power, that is, the probability of two hypotheses rejected when the two null hypotheses are not true. The red line is the power of Hochberg method, the blue line is power of Holm method, and the green line is the power of Bonferroni method. As shown in the plot and table, Holm and Hochberg method are more powerful than Bonferroni. It means that Holm method can reject more individual hypotheses than Bonferroni method, although they can both achieve the at-least-one win criteria.

So the answer seems to be clear now. If we can re-design the study, Holm adjustment may be the most appropriate method for this study. If the correlation coefficient is larger than 0.35, we can choose the Hochberg adjustment instead of Holm adjustment. If the correlation is smaller than 0.35, Hochberg adjustment is not recommended as the type I error is bigger than 0.05. However, I do not know how to simulate the bivariate distribution data like (-2, -1, 0, 1, 2). So in the real example, we need to simulate the same distribution data, and can not use the other distribution data to estimate the impact. Simulation is a valuable tool to help our biostatistician to understand the advantages and disadvantages of different methods when it is difficult to know the analytic mathematical formulae. When design a clinical trial, simulation can help our biostatistician to choose a more powerful method to increase the successful probability of the trial under the type I error controlled at alpha level.