Tests for Interaction in Clinical Studies

Chul Ahn[1], Mourad Atlas[1]
[1]Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993

Abstract

Tests for interaction are used in clinical trials to find a treatment effect that differs by subgroup. In this paper, some of the statistical issues surrounding interaction tests will be discussed. It will include the subgroup-specific test, quantitative and qualitative interaction, and statistical power. Regarding statistical power on interaction test, we will discuss the misconception that there is always less statistical power for interactions than main effects. We will also discuss the misconception about the interpretability of main effects when there is an interaction.

Key Words: Interaction, quantitative, qualitative, power, effect size

1. Introduction

We are often interested in finding out whether there exists any difference in treatment effects between gender and how to detect such difference statistically if it exists. Test for interaction is an appropriate statistical method in understanding the difference in treatment effects between men and women. . The example below will illustrate the importance of having adequate numbers of both men and women being included in the trial as well as the fallacy of subgroup-specific test in addressing interaction.

1.1 Example: Results stratified by Gender

A medical device ABC is an ophthalmic device to treat glaucoma subjects. The goal of the study was to assess the safety and effectiveness of the device in lowering intraocular pressure (IOP) in glaucomatous eyes in conjunction with cataract surgery, as compared to eyes treated with cataract surgery alone. The sponsor conducted a randomized clinical trial with one of the primary endpoint being the percent of subjects with at least 20% reduction in IOP at 24 month from baseline. Each group received cataract surgery, and upon completion of uncomplicated cataract removal IOL implantation, they were randomized to receive either the ABC or no device (i.e., cataract-only group). A subject's outcome is considered a success, responder, if 24-month IOP decreases by at least 20% from baseline. The results are shown in the table below. Using Fisher's Exact Test, there was a significant treatment difference (p-value = 0.0005). In the following table IOPR stands for IOP reduction at 24 months.

	≥ 20% IOPR	< 20% IOPR	Total	Success %	p-value
ABC	148	133	281	53%	0.0005
Control	92	150	240	38%	
Total	240	283	521		

However, if we break it down by gender, we can see some interesting results. For the male group, there was a significant treatment effect (p-value=0.006), however no significant effect was seen for the female group (p-value=0.1). Does it imply that the treatment works for male, but not for female?

For Male Group:

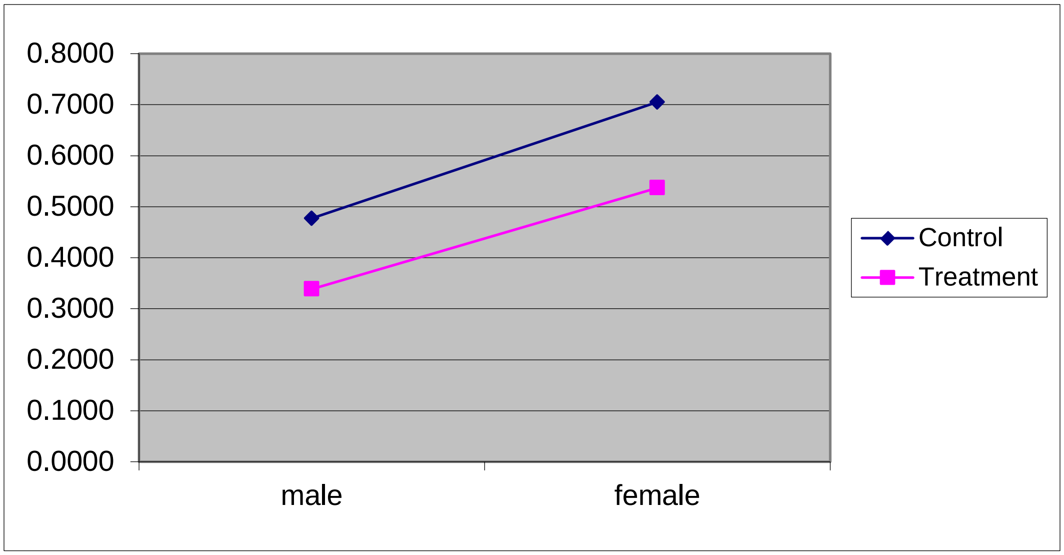
	≥ 20% IOPR	< 20% IOPR	Total	Success %	p-value
ABC	105	115	220	48%	0.006
Control	63	123	186	34%	
Total	168	228	406		

For Female Group:

	≥ 20% IOPR	< 20% IOPR	Total	Success %	p-value
ABC	43	18	61	70%	0.1
Control	29	25	54	54%	
Total	72	43	115		

Let's look at two tables more closely. The observed treatment differences are about the same between two subgroups. The observed treatment difference for male is 14% (48%-34%), and that for female is 16% (70%-54%). In fact, the larger effect was observed for female. It can also be seen in the graph below.

Figure 1: The treatment and control effect stratified by gender.



The reason why we see a non-significant result from female group, even with larger treatment effect observed compared to male group, is due to its sample size. The sample size for female is much smaller than that for male (115 versus 406). If we double the sample size for female to 230, the treatment effect becomes significant (p-value=0.009 < 0.05). This example illustrates why it is important to have adequate numbers of both men and women included in the trial.

1.2 Subgroup specific test

Often the investigators use the subgroup-specific test in determining whether there exists a differential treatment effect between men and women. In other words, they construct separate hypothesis in each subgroup with the null hypothesis being that there is no treatment effect for a particular subgroup. In the previous example, hypothesis was constructed separately in each subgroup. Below, the first set of hypotheses in (I) is for the male subgroup and the second set of hypotheses in (II) for the female group.

(I) H_{0m}: $P_{mA} = P_{mC}$ vs. H_{1m}: $P_{mA} \neq P_{mC}$ for male subgroup

(II) H_{0f}: $P_{fA} = P_{fC}$ vs. H_{1f}: $P_{fA} \neq P_{fC}$ for female subgroup

In the mathematical expression of the null and alternative hypotheses (H_{0m} and H_{1m}) for the male subgroup in (I), P_{mA} denotes the percentage of male subjects assigned to the ABC group who achieve at least 20% reduction in IOP, and P_{mC} , the percentage of male subjects assigned to the Control group who achieve at least 20% reduction in IOP. Likewise, in (II), P_{fA} denotes the percentage of female subjects assigned to the ABC group who achieve at least 20% reduction in IOP, and P_{fC} , the percentage of female patients assigned to the Control group who achieve at least 20% reduction in IOP.

The p-values from statistical hypothesis testing of (I) and (II) were reported as 0.006 and 0.1 for male and female subgroup groups, respectively. The investigators may use these subgroup p-values in determining whether there exists a differential treatment effect between female and male. They may argue that there is a difference between genders because there is a treatment effect in men, but not in women and conclude that the treatment works for men, but not for women. This kind of misleading conclusion (which is to claim heterogeneity on the basis of separate tests of treatment effects within each of the levels of the baseline variable) is commonly seen in clinical trials.

If the overall result is significant, almost inevitably some subgroups will and some will not show significant differences depending on chance. Therefore, investigators should be cautious when undertaking subgroup analyses. This subgroup-specific test may result in inappropriate subgroup claims as it can be seen in the previous example where the investigators may claim that the treatment only works for men, but not for women. The appropriate statistical analysis to determine whether there exist a differential treatment effect between men and women should include the interaction of treatment by gender, which will be discussed below.

2. Interaction test

Interaction between treatment and gender (or interaction of treatment by gender) may be defined as the difference in treatment effects between men and women. The correct assessment of the treatment by gender interaction may require the appropriate statistical test and adequate sample sizes for the two subgroups. The appropriate statistical test to determine whether there exists a differential treatment effect between men and women is a test of interaction. In the previous example, the difference in the proportions of subjects achieving more than 20% IOP reduction between ABC group and Control group in men, or simply the treatment effect in men would be expressed as follows:

Treatment effect in men = $P_{mA} - P_{mC}$

Likewise, the treatment effect in women would be expressed as follows:

Treatment effect in women = $P_{fA} - P_{fC}$

A test of interaction is to determine whether these two treatment effects are the same or not, and it may be expressed as follows:

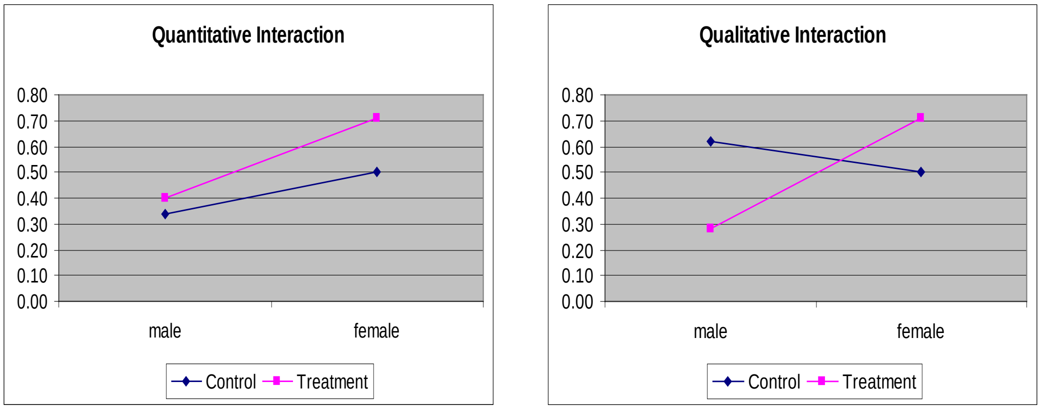
$$H_0\text{: } P_{mA} - P_{mC} = P_{fA} - P_{fC} \quad \text{vs.} \quad H_1\text{: } P_{mA} - P_{mC} \neq P_{fA} - P_{fC}\ .$$

If we use Breslow-Day test (one method of testing interaction) for the previous example, we obtain the p-value>0.2, and can conclude that the two treatment effects are the same between men and women, which leads to the correct conclusion. This example shows that why test of interaction is appropriate in understanding the difference in treatment effects between men and women. In summary, Interaction test asks if there are any differences in treatment effect between subgroups (appropriate when making inferences from subgroup analyses). Subgroup-specific test asks if there is any treatment effect within each subgroup. Pocock et al (2002) stated that statistical tests for interaction are the most appropriate methods for making subgroup inferences, but are often not used.

3. Qualitative and quantitative interaction

There are two different kinds of interaction - “quantitative interaction” and “qualitative interaction”. Quantitative interaction means that the treatment is effective in both men and women (or treatment effects are in the same direction), but the magnitude of the effect is different. Qualitative interaction means that the treatment is effective in one gender but ineffective or harmful in the other. Quantitative interaction is model dependent and sometimes it is possible to remove them by a transformation of the variable. However, qualitative interaction is model independent and may not be removed by transformation or any other modeling. The following graphs show these two types of interactions.

Figure 2: Quantitative and qualitative interactions



4. Power of main effect and interaction

One of the misconceptions about interaction is that there is always less statistical power for interactions than for main effects. In this section, we will examine whether this is true. Suppose that we have two treatments (T and C), two subgroups (M and F), and let's assume normality and homogeneous variances. There are four population means, μ_{MT} , μ_{MC} , μ_{FT} , μ_{FC}. The expression $\mu_{FT} - \mu_{FC}$ will denote the treatment effect in females and $\mu_{MT} - \mu_{MC}$ the treatment effect in males. Then, the overall main effect can be written as $\theta_1 = \{ (\mu_{FT} - \mu_{FC}) + (\mu_{MT} - \mu_{MC}) \}/2$ and the interaction effects can be written as
$\theta_2 = \{ (\mu_{FT} - \mu_{FC}) - (\mu_{MT} - \mu_{MC}) \}/2$.

If we let $\bar{X}_{MT}, \bar{X}_{MC}, \bar{X}_{FT}$, and $\bar{X}_{FC}$ denote the point estimates of four population means observed in the trial, then the main and interaction effects can be estimated as follows:
$\hat{\theta}_1 = \{(\bar{X}_{FT} - \bar{X}_{FC}) + (\bar{X}_{MT} - \bar{X}_{MC})\}/2$

$\hat{\theta}_2 = \{(\bar{X}_{FT} - \bar{X}_{FC}) - (\bar{X}_{MT} - \bar{X}_{MC})\}/2.$

We can immediately see that $\hat{\theta}_1$ and $\hat{\theta}_2$ have the same standard error. Let's calculate the power of main effect and interaction, and compare:

$H_0 : \theta_1 = 0$ vs. $H_1 : \theta_1 \neq 0$ at significance level α

$Power$ for main effect $= P($reject H_0 when H_1 is true$)$
$= 2P(Z > z_\alpha - \frac{\theta_1}{s.e.(\hat{\theta}_1)}),$

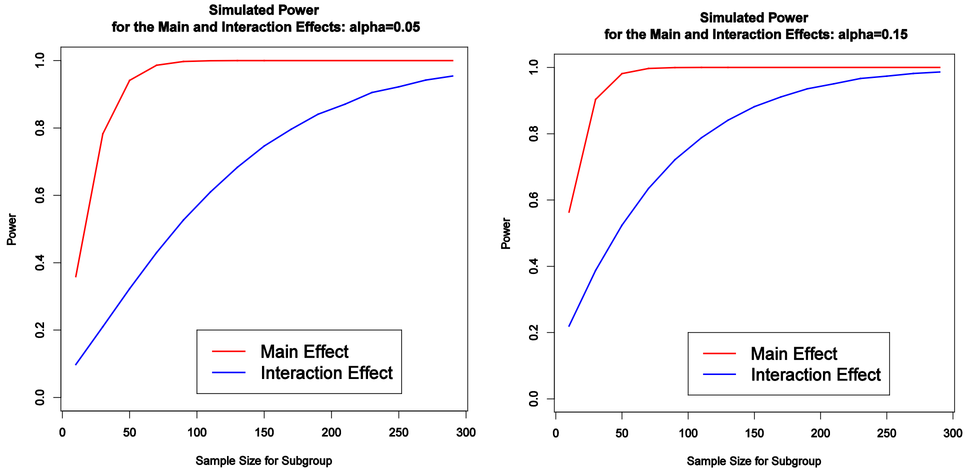
Similarly, power for interaction effect = $2P(Z > z_\alpha - \frac{\theta_2}{s.e.(\hat{\theta}_2)})$

We notice that the power for main effect and interaction are equal as long as $\theta_1 = \theta_2$. Therefore, mathematically speaking, we cannot say that there is always less statistical power for interactions than for main effects.

However, in the real situation, it is less likely that $\theta_1 = \theta_2$. Let's consider the situation where the treatment effects for females and males are 10 and 4, respectively. In this situation, the main effect will be 7 since it is the average of the treatment effects for females and males. And, the interaction effect will be 10/2 – 4/2 = 3. The only way the interaction term equals 7 is if you get all the treatment effect in one subgroup and none in the other: 14/2 – 0/2 = 7, which is unlikely in practice.

We have used simulation to illustrate the difference in statistical power between the main and interaction effects. The following graph shows that the simulated power for the main and interaction power when the significance level is 0.05 (the left graph) and 0.15 (the right graph). These two graphs also show that the interaction effect has less statistical power compared to the main effect.

Figure 3: Simulated power for the main and interaction power when the significance level is 0.05 (the left graph) and 0.15 (the right graph)



5. Effect and Dummy Coding

In this section, we will discuss the misconception about the interpretability of main effects when there is an interaction. The question is whether interaction always makes main effect uninterpretable. This problem is related to the coding of the categorical variables, that is, effect and dummy coding.

A designed experiment is orthogonal if the effects of any factor balance out (sum to zero) across the effects of the other factors. Effect coding for orthogonality guarantees that the effect of one factor or interaction can be estimated separately from the effect of any other factor or interaction in the model. With dummy coding the estimate of the interaction is fine but main effects are not "true" main effects but rather what are called simple effects, i.e., the treatment effect at one level of the subgroup. With an interaction of two categorical variables, effect coding provides some benefits. The primary benefit is that you get reasonable estimates of both the main effects and interaction using effect coding. Let’s illustrate this using the simple example. Suppose we have a following model with two main effects and its interaction:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon,$$

where X_1 and X_2 denote gender and treatment.

Let’s consider an effect coding with 0.5 and -0.5, that is male=0.5, female=-0.5, treatment=0.5, CNT=-0.5). The treatment effect for female can be written as:

$$\mathrm{E}(Y|X_1\text{=-0.5, } X_2\text{=0.5)-E(}Y|X_1\text{=-0.5, } X_2\text{=-0.5)=} \beta_2 \text{ - 0.5 } \beta_{12},$$

and the treatment effect for male can be written as:
$\mathrm{E}(Y|X_1\text{=0.5, } X_2\text{=0.5)-E(}Y|X_1\text{=0.5, } X_2\text{=-0.5)=} \beta_2 + 0.5\, \beta_{12}.$

In this example, the main effect will be β_2, and the interaction effect will be $-\beta_{12}/2$.

With a dummy coding with 0 and 1, that is, female=0, male=1, control=0, and treatment=1, the main effect can be written as:

$$(\beta_2 + (\beta_2 + \beta_{12}))/2 = \beta_2 + \beta_{12}/2,$$

and the interaction effect as - $\beta_{12}/2$.

In summary, with effect coding, the estimates of the main effects are "true" main effects regardless of the presence of interaction, and interaction can be estimated separately from the main effects. However, that is not true with dummy coding.

References

Assmann, S., Pocock, S., Enos, L., and Kasten, L. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials, LANCET 355; 1064-1069.

Bulpitt, Christopher J. (1988). Subgroup analysis. Lancet 2:31-34.

Cox, D. R.(1984). Interaction, International Statistical Review, 52, 1-31 (1984).

Gail, M. and Simon, R. (1985): Testing for qualitative interactions between treatment effects and patient subsets. Biometrics 41, 362-372.

Pocock, S.J., Assmann, S.E., Enos, L.E. and Kasten, L.E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting. Statistics in Medicine 21: 2917-2930.

Russek-Cohen, E. and Simon, R. (1993) Qualitative Interactions in Multifactor Studies, Biometrics 49, 467-477.

Russek-Cohen, E. and Simon R. (1997) Evaluating treatments when a gender by treatment interaction may exist. Statistics in Medicine 16: 455-464

Scott, Pamela and Campbell, Gregory (1998). Interpretation of subgroup analyses in medical device clinical trials. Drug Information Journal 32:213-220.