

# Intervention Effects in the Case of Heterogeneity between Three Subgroups

## Assessment within the Framework of Systematic Reviews

G. Skipka; R. Bender

Institute for Quality and Efficiency in Health Care, Cologne, Germany

### Keywords

Subgroups, heterogeneity, systematic review, interaction, meta-analysis

### Summary

**Background:** Usually, statistical tests for interactions are applied to investigate potential effect modifiers. If an effect modifier, consisting of three categories, is found to be statistically significant, the application of pairwise interaction tests is indicated. In this case, the problem of non-transitive relations may occur if the significance level is fixed at 0.05 for all tests.

**Objective:** To develop an algorithm for which non-transitive relations do not occur.

**Methods:** A hierarchical testing procedure is applied, based on the heterogeneity statistic  $Q$ . In a first step the interaction will be tested for the three trial subgroups altogether, applying the significance level  $\alpha = 0.05$  (global test). If a significant interaction is proven in the first step, pairwise tests for interaction will

be applied in a second step. Theoretical data scenarios will be considered and p-values will be calculated for the pairwise tests. Based on these results the significance level for pairwise testing will be determined.

**Results:** Fixing the significance level at 0.05 for all tests, the problem of non-transitive relations is mostly relevant, if the difference in the effects between the three trial subgroups is approximately 3.5 standard errors and the effect of the 'middle' trial subgroup is not close to one of the other two effects. This problem vanishes when the significance level is set to  $\alpha = 0.22$ . We propose to select  $\alpha = 0.20$  to get a more 'even' and simple value.

**Conclusions:** By increasing the significance level for the pairwise tests to 0.20, non-transitive relations are virtually avoidable. The proposed hierarchical testing procedure represents a clear practical guidance to perform subgroup analyses in the framework of systematic reviews.

In this paper we deal with the following situation. The aggregated estimates of an effect are present for several clinical trials comparing an experimental intervention to a control intervention. The trials can be divided into three distinct "trial subgroups" due to a potential effect modifier based upon subject-matter knowledge (e.g. high, medium and low dose for the experimental intervention). In this situation, effect modifiers are usually investigated by tests for heterogeneity within the framework of meta-analyses [1].

Generally, the existence of an effect modifier is proven if the interaction or heterogeneity is statistically significant applying 0.05 as the significance level, i.e. for the p-value  $p < 0.05$ . In the case of significant heterogeneity between three trial subgroups the following question arises: Are all the three effects of the trial subgroups substantially different from each other, or is only one trial subgroup substantially different from the other two trial subgroups? In the latter case the homogeneous results of the two trial subgroups can be summarized with the goal of providing more precise effect estimators. To our knowledge, no suggestion exists in the literature so far on how to deal with this question.

We introduce two examples to illustrate the situation considered here. The first example refers to a single study including three subgroups of patients. This setting is slightly different from that described above, where a set of trials consists of three trial subgroups. Regardless of this difference, the first example drew our attention to the problem considered in this paper. The second fictitious example illustrates the situation of a set of trials.

### Correspondence to:

G. Skipka  
IQWiG  
Dillenburger Str. 27  
51105 Köln  
Germany  
E-mail: guido.skipka@iqwig.de

Methods Inf Med 2010; 49: 613–617

doi: 10.3414/ME09-02-0054

received: December 8, 2009

accepted: April 14, 2010

prepublished: July 20, 2010

## 1. Introduction

Systematic reviews collate all the evidence on the effect of an intervention in health care. An important task within the framework of systematic reviews is the investigation of effect modifiers. The intervention effect may vary with different populations (e.g. age groups) or intervention character-

istics (e.g. drug doses). Such variation due to an effect modifier is known among statisticians as an interaction (between intervention and effect modifier). Usually subgroup analyses (for categorical variables) and meta-regressions are applied to investigate such interactions. If an effect modifier is present, heterogeneous effect sizes occur for the different levels of the effect modifier.

## 1.1 Example 1

The randomized CURE study [2] compared two interventions in more than 12,000 patients with acute coronary syndrome without ST-segment elevation. Clopidogrel plus acetylsalicylic acid or acetylsalicylic acid plus placebo were given for a minimum of 3 and a maximum of 12 months (mean observation period: 9 months). The first primary endpoint was a composite endpoint of cardiovascular mortality, myocardial infarction, and stroke. Subgroup analyses were carried out for smoking status with three different categories: smokers, ex-smokers, and non-smokers. ▶ Figure 1 shows the effects on the composite endpoint for the three subgroups and the total population [3].

The CURE study shows a statistically significant effect in favor of the combination therapy for the total population ( $p < 0.05$ ). However, there is a statistically significant interaction between smoking status and therapy ( $p < 0.05$ ). The question of interest is how to further evaluate the results. Should the results be interpreted for each of the subgroups separately? Then the reduced risk of vascular events would be particularly pronounced in smokers and ex-smokers only. Or should the homogeneity of at least two of the subgroups be investigated as an additional step? Perhaps the effects of smokers and ex-smokers or ex-smokers and non-smokers are sufficiently similar to summarize the results of two subgroups. The advantage of the latter approach would

be more precise effect estimators in case of partly homogeneous subgroups. However, it remains unclear how to deal with the term 'sufficiently similar' in detail.

## 1.2 Example 2

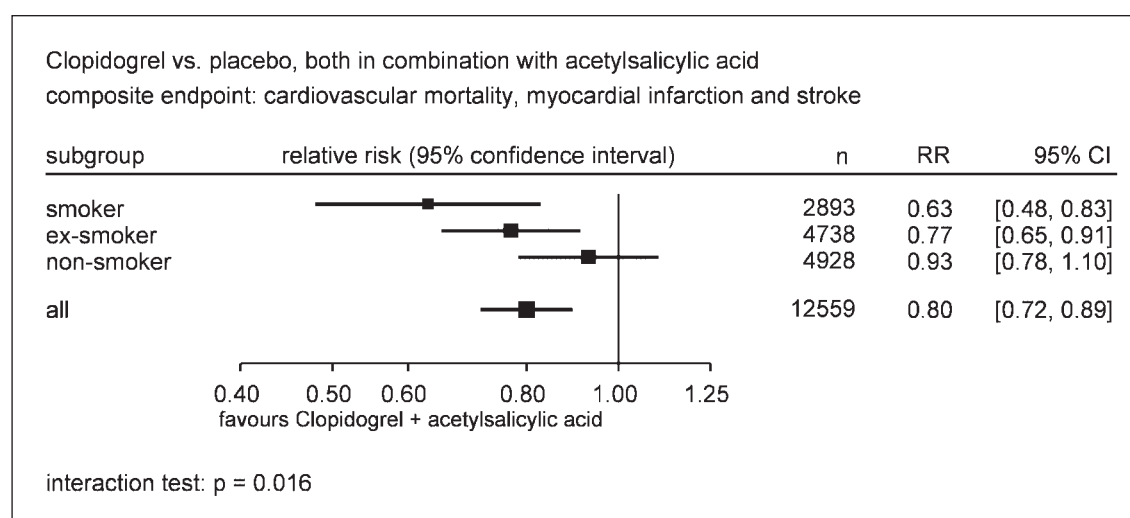
This fictitious example refers to the setting described in the introduction. The data closely follow Example 1. However, the dose level of an experimental drug represents the effect modifier. Suppose that seven clinical trials, comparing an experimental drug with control, are included in a systematic review. The trials used three different dose levels (low, medium, high) for the experimental drug. ▶ Figure 2 shows the results of meta-analyses summarizing the trials as a whole and separating the seven trials according to dose level.

Summarizing the trials as a whole, the meta-analysis shows a statistically significant effect in favor of the experimental drug ( $p < 0.05$ ). However, there is a statistically significant interaction (heterogeneity) between the three trial subgroups too ( $p < 0.05$ ), which proves different effects between the three dose levels. The decision whether two of the three trial subgroups are homogeneous enough for summarizing could be based on pairwise tests for interaction. The calculation of the pairwise tests gives the following results:

- high vs. medium dose:  $p = 0.230$
- high vs. low dose:  $p = 0.008$
- medium vs. low dose:  $p = 0.059$

Assuming the usual significance level  $\alpha = 0.05$ , a difference in effect sizes is shown comparing high with low dose. However, the effect sizes do not differ statistically significantly for high vs. medium dose and medium vs. low dose, respectively. Therefore, using  $\alpha = 0.05$  for the pairwise comparisons, a problem in interpreting the results arises. This relation is called non-transitive in mathematics. The question of interest is how to avoid the appearance of non-transitive relations in this context. A proposal for handling this problem is to increase the significance level for the pairwise tests for interaction. The objective of this paper is to propose an algorithm to derive evidence of effects in the situation of three trial subgroups. It should be noted that the general conditions and problems of conducting subgroup analyses are outside the scope of this paper (for details see e.g. [4]).

It should be noticed that the problem considered here is present in both examples. Regardless of considering one single clinical trial (Example 1) or a meta-analysis of several clinical trials (Example 2), the following proposed procedure is applicable. Nevertheless, the focus of the paper is the framework of systematic reviews. The investigation of effect modifiers is a major task in systematic reviews, since heterogeneous populations and intervention characteristics are routinely considered. Although the exploration of effect modifiers is also important in case of one clinical trial, homogeneous populations and intervention characteristics are more common within one trial.



**Fig. 1**

Results from the CURE study [3]. Effects on the composite endpoint for three subgroups (smoker, ex-smoker, non-smoker) and the total population (RR = relative risk, CI = confidence interval)

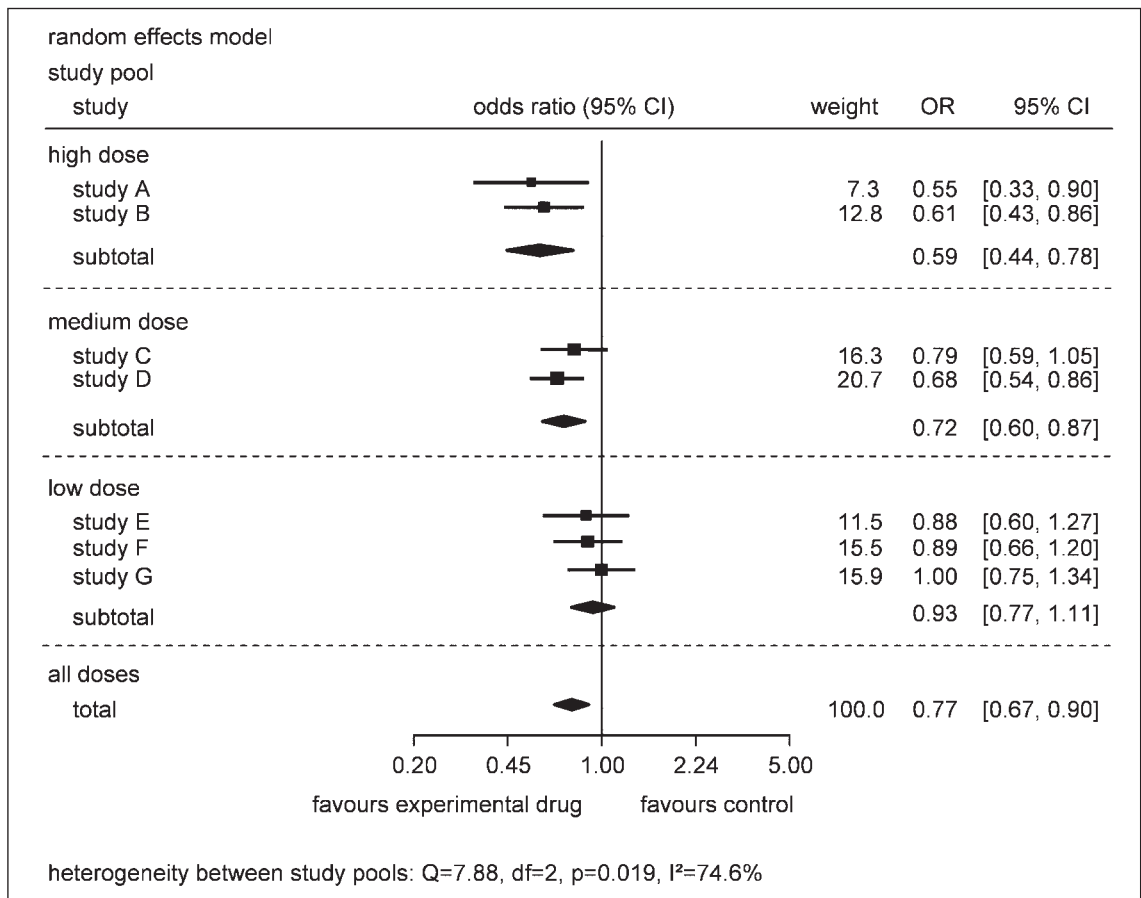


Fig. 2

Fictitious results from a meta-analysis including 7 trials with 3 different dose levels (OR = odds ratio, CI = confidence interval, df = degrees of freedom,  $p$  =  $p$ -value)

## 2. Methods

We consider the situation of Example 2 described in the Introduction. We consider the situation in which statistical tests for interaction between intervention and the (potential) effect modifier are applied to explore differences between the effect in the trial subgroups. These tests are based on Cochran's  $Q$  statistic. This heterogeneity statistic may be written as

$$Q = \sum_{i=1}^3 w_i (\theta_i - \theta)^2,$$

with the pooled effect estimate

$$\theta = \frac{\sum_{i=1}^3 w_i \theta_i}{\sum_{i=1}^3 w_i},$$

where  $w_i = 1/SE_i^2$  and  $\theta_i, SE_i$  ( $i = 1, 2, 3$ ) are the effect estimates and standard errors for each trial subgroup, respectively [1]. Assuming that there are no differences in effects among the three independent trial subgroups the statistic  $Q$  is approximately  $\chi^2$ -distributed with two degrees of freedom [5].

We apply a hierarchical testing procedure: In a first step the interaction will be tested for the three trial subgroups altogether, applying the significance level  $\alpha = 0.05$  (global test). If a significant interaction is proven in the first step, pairwise tests for interaction will be applied in a second step, based on the significance level  $\alpha^*$ . Otherwise, no further tests will be applied.

The goal is to specify the significance level  $\alpha^*$  for pairwise testing, such that the appearance of non-transitive relations will not occur (cf. Example 2). We will consider theoretical data scenarios and calculate  $p$ -values for the pairwise tests for interaction. Based on these results we will determine the significance level  $\alpha^*$ .

## 3. Results

The theoretical data scenario is specified as follows. Three trial subgroups ( $TSG_1, TSG_2, TSG_3$ ) are given. Furthermore, effect estimates ( $\theta_1, \theta_2, \theta_3$ ) are calculated on the basis

of meta-analyses for each trial subgroup. The precision (standard error,  $SE$ ) of these three estimates are assumed to be equal to keep the calculations simple. Subsequently, we will address the case of unequal standard errors. Without loss of generality we set  $SE = 1$  and the effect estimate for the first trial subgroup  $\theta_1 = 0$ . Further, we assume  $\theta_1 \leq \theta_2 \leq \theta_3$ .

As a first step we determine the minimum heterogeneity required for the three effect estimates, such that the global interaction test for the three trial subgroups results exactly in  $p = 0.05$ . In other words, the difference between the smallest  $\theta_1$  and the largest  $\theta_3$  effect estimate is set to the smallest value for which the following condition holds: varying the 'middle' effect estimate  $\theta_2$ , the maximum  $p$ -value of the interaction test is not greater than 0.05. With the assumptions made here, the test statistic  $Q$  reduces to the sum of  $(\theta_i - \theta_a)^2$ ,  $i = 1, 2, 3$ , with  $\theta_a$  as the arithmetic mean of the three effect estimates. Assuming homogeneity between the trial subgroups,  $Q$  is  $\chi^2$ -distributed with two

degrees of freedom (for details see [1]). It is easy to verify that if  $\theta_2 = (\theta_1 + \theta_3)/2$  the statistic  $Q$  reaches its minimum and the p-value therefore reaches its maximum. The 95%-quantile of the  $\chi^2$ -distribution with two degrees of freedom is  $\chi^2_{2;0.95} = 5.99$ . As a result, by setting  $Q = 5.99$  and  $\theta_1 = 0$  the effect estimate for the third trial subgroup  $\theta_3$  has to be the square root of  $2 \cdot 5.99$ , which gives approximately 3.46. ▶ Figure 3 illustrates this situation.

For smaller differences between the trial subgroups 1 and 3, the maximum p-value of the interaction test becomes greater than 0.05, for larger differences the maximum p-value becomes lower than 0.05.

As a next step we determine the p-values of the pairwise interaction tests for this

scenario. Of special interest are the two comparisons of adjacent trial subgroups, i.e.  $TSG_1$  vs.  $TSG_2$  and  $TSG_2$  vs.  $TSG_3$ . For the comparison of the trial subgroups 1 with 3, the p-value is always smaller than the p-value of the global interaction test. ▶ Figure 4 shows the p-values of the two pairwise interaction tests of adjacent trial subgroups as a function of the 'middle' effect estimate  $\theta_2$ .

Applying the usual significance level  $\alpha = 0.05$  for the pairwise tests here means that for a broad range of values of  $\theta_2$  the problem of a non-transitive relation occurs (both p-values are greater than 0.05 if  $0.69 < \theta_2 < 2.77$ ). In other words, if the difference in the effects between  $TSG_1$  and  $TSG_3$  is approximately 3.46 standard errors

and the effect of  $TSG_2$  is not close to one of the other two effects, there is a problem in interpreting the results.

At the point of intersection of the curves the p-value equals  $1 - F(0.25 \cdot \chi^2_{2;0.95}) \approx 0.22$ . Therefore, the problem of non-transitive relations vanishes when the significance level is set to this value. On the one hand, setting  $\alpha = 0.22$  would solve the problem. On the other hand, 0.22 is a little 'odd'. We propose to select  $\alpha = 0.20$  to get a more 'even' and simple value. Moreover, the use of the significance level  $\alpha = 0.2$  for heterogeneity tests in the framework of meta-analyses is already proposed in the scientific literature [6]. As ▶ Figure 4 shows, the risk of occurrence of a non-transitive relation is very small for  $\alpha = 0.20$ . This case only occurs if  $\theta_2$  is more or less exactly in the middle of  $\theta_1$  and  $\theta_3$ .

The question remains as to what happens if the difference in the effects between  $TSG_1$  and  $TSG_3$  is smaller or larger than 3.46 standard errors. In the case of a larger difference, the problem of non-transitive relations cannot occur, since the p-values for the pairwise interaction tests are smaller than illustrated in ▶ Figure 4. In the case of a smaller difference in the effects between  $TSG_1$  and  $TSG_3$  a significant interaction (global test) appears only if the effect of  $TSG_2$  comes close to one of the other trial subgroup effects. A non-transitive relation

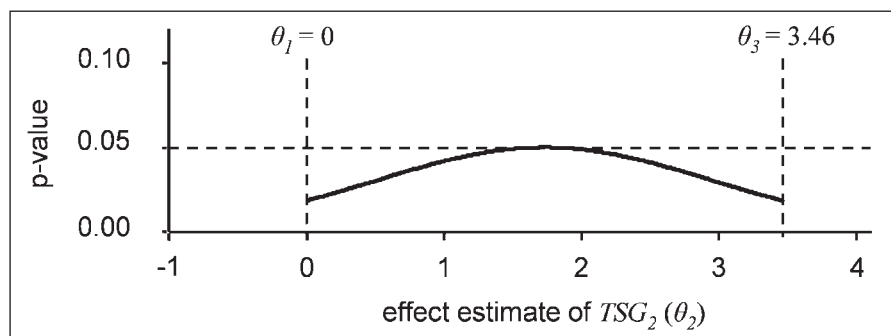


Fig. 3 P-value of the statistical test for interaction as a function of the effect estimate for trial subgroup 2

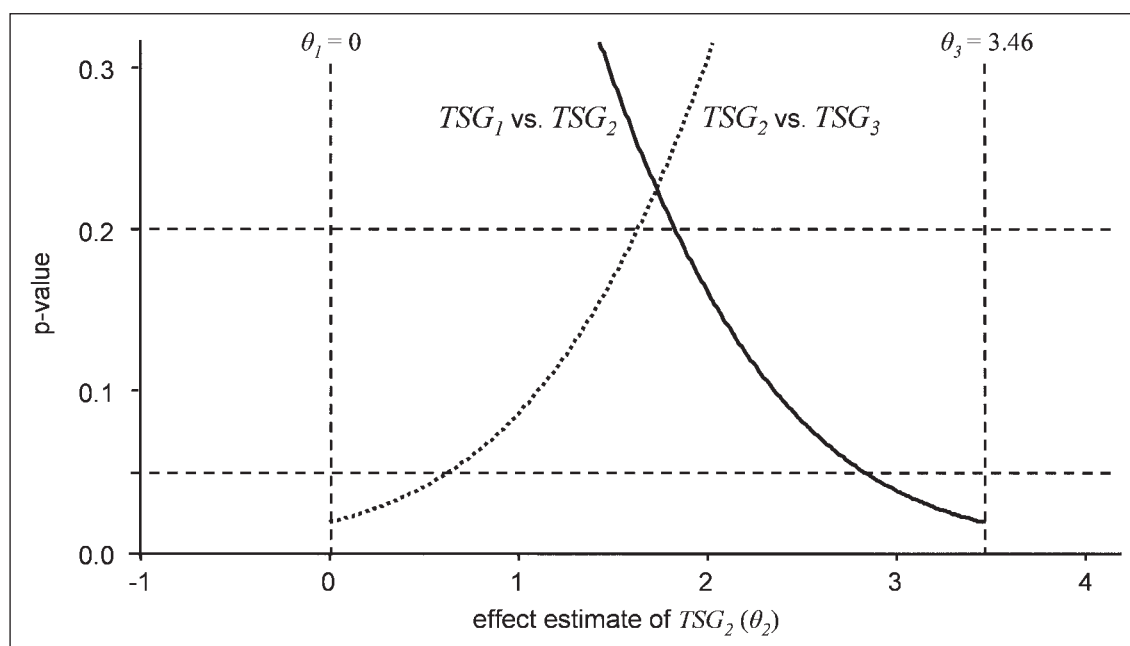


Fig. 4 P-value of the pairwise statistical tests for interaction as a function of the effect estimate for trial subgroup 2

cannot occur even in this case, since the p-values for the pairwise comparison would not be simultaneously greater than 0.20.

To keep the calculations simple we assumed the three standard errors to be equal. In case of unequal standard errors the point of intersection of the p-value curves (► Fig. 4) depends on the ratios  $SE_2/SE_1$  and  $SE_3/SE_1$ . E.g., if the precision of  $\theta_1$  and  $\theta_2$  is equal and  $SE_3$  is greater ( $SE_1 = SE_2 < SE_3$ ), the p-value at the point of intersection gets smaller than 0.22. The same situation appears in case of  $SE_1 = SE_3 > SE_2$ . On the other hand the 'intersection p-value' is greater than 0.22 if  $SE_1 = SE_2 > SE_3$  or  $SE_1 = SE_3 < SE_2$ . In these situations the problem of non-transitive relations may occur more likely. This problem could be solved by adjusting the significance level by the ratios of standard errors. We did not pursue this approach. First of all, the calculations would get too complicated. Second of all, the investigation of heterogeneity gets more complex in case of strongly different standard errors. In this situation, the specific data constellation should be taken into account, whether non-transitive relations occur or not.

Based upon these results, we propose the following hierarchical algorithm to derive evidence of effects in the situation of three trial subgroups.

1. Apply an interaction test including the three trial subgroups altogether with the significance level  $\alpha = 0.05$  (global test).
  - a) If  $p \geq 0.05$ , there is no evidence for the existence of an effect modification and the assessment of the effect is based on the summary of all trials as a whole. The algorithm stops here.
  - b) If  $p < 0.05$ , the existence of an effect modification is proven, i.e., the effects of at least two of the three trial subgroups differ. Proceed with step 2.
2. Apply pairwise interaction tests with a significance level  $\alpha = 0.20$  for both comparisons of adjacent trial subgroups, i.e.  $TSG_1$  vs.  $TSG_2$  and  $TSG_2$  vs.  $TSG_3$ .
  - a) If both p-values are smaller than 0.20, the assessment of the effect is performed separately in the three trial subgroups (applying a significance level of  $\alpha = 0.05$  for the test of treatment effect).

- b) If only one of the p-values is 0.20 or greater, the assessment of the effect includes two parts. Summarize these two trial subgroups which are more homogeneous and take a separate look at the third trial subgroup (applying a significance level of  $\alpha = 0.05$  for the test of treatment effect).
- c) If both p-values are 0.20 or greater, the assessment of the effect is difficult and the specific data constellation should be taken into account. However, as described above, this case will occur very rarely in case of comparable standard errors.

### 3.1 Example 2 (continued)

Applying the proposed algorithm to Example 2 we get the following information. The global interaction test ( $p = 0.019$ ) shows that the dose of the experimental drug is an effect modifier. Therefore, pairwise interaction tests with  $\alpha = 0.20$  are indicated. The p-value for the comparison of high and medium dose is  $p = 0.230$ , so both trial subgroups are homogeneous enough for summarizing the results within a meta-analysis. The pooled effect is statistically significant ( $p < 0.05$ , data not shown). On the other hand, the low-dose trial subgroup has to be considered separately, since the p-value for the comparison of medium and low dose is smaller than 0.20 ( $p = 0.059$ ). As shown in ► Figure 2, the effect for this trial subgroup is not statistically significant ( $p > 0.05$ ). As a summary, there is evidence for a benefit of the experimental drug in comparison to the control intervention. However the benefit applies to the high and medium dose only.

## 4. Discussion and Conclusion

Despite the increasing importance of systematic reviews and meta-analyses there is currently no clear guidance available how subgroup analyses in the framework of systematic reviews should be performed. Usually, statistical tests for interactions are applied to investigate potential effect modifiers. If an effect modifier, consisting of three

categories, is found to be statistically significant, the application of pairwise interaction tests is indicated. In this case, the problem of non-transitive relations may occur if the significance level is fixed for all tests at  $\alpha = 0.05$ . In this paper it was demonstrated that the increase of the significance level for the pairwise tests to  $\alpha = 0.20$ , non-transitive relations can be avoided in most practical situations. The proposed hierarchical algorithm does not hold the multiple significance level in a strong confirmatory sense. However, this is impossible in the framework of systematic reviews in any case. The proposed increased significance level of  $\alpha = 0.20$  for the decision to combine subgroups or not is in agreement with previous suggestions concerning the application of heterogeneity tests in meta-analyses [6].

In conclusion, the proposed hierarchical testing procedure represents a useful practical guidance to perform subgroup analyses in the framework of systematic reviews that avoids non-transitive relations.

### Acknowledgments

We thank Marian Cairns for editorial support. We also thank the anonymous reviewers for constructive comments on the manuscript.

### References

1. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G, Altman DG, editors. *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd edition). London (UK): BMJ Publication Group; 2001.
2. Yusuf S, Zhao F, Mehta SR, Chrolavicius S, Tognoni G, Fox KK. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. *N Engl J Med* 2001; 345 (7): 494–502.
3. Institute for Quality and Efficiency in Health Care. Clopidogrel plus acetylsalicylic acid in acute coronary syndrome: final report; commission A04-01B (Internet). Cologne: IQWiG; Jan 28, 2009 (cited Nov 5, 2009). Available from: [http://www.iqwig.de/download/A04-01B\\_AB\\_Clopidogrel\\_plus\\_ASS\\_bei\\_akutem\\_Koronarsyndrom.pdf](http://www.iqwig.de/download/A04-01B_AB_Clopidogrel_plus_ASS_bei_akutem_Koronarsyndrom.pdf).
4. Alosch M, Huque MF. A flexible strategy for testing subgroups and overall population. *Stat Med* 2009; 28 (1): 3–23.
5. Cochran WG: The Combination of Estimates from Different Experiments. *Biometrics* 1954; 10: 101–129.
6. Koch A, Ziegler S. Metaanalyse als Werkzeug zum Erkenntnisgewinn. *Med Klin* 2000; 95, 109–116.