12. MISTAKES IN SUBGROUP ANALYSES

12.1. Overview

Consider a meta-analysis where the studies can be classified as belonging to two or more subgroups. We might want to compute the effect size within each subgroup, and then compare these different effect sizes. Immediately below, I present an example of such an analysis. In the balance of the chapter, I discuss various parts of the analysis, highlighting common mistakes.

12.1.1. Example | Drugs for weight-loss

Padwal, Li, and Lau (2003) performed an analysis of studies to assess the utility of drugs to help patients lose weight (see also (Borenstein & Higgins, 2013)). In each study, patients were randomly assigned to either the drug or placebo, and the researchers recorded the proportion of patients in each group who succeeded in meeting their weight-loss goal within the study's timeframe.

The effect size is the risk difference. A mean risk difference of 0.24 would tell us that the treatment increased the likelihood of success by 0.24. For example, in a population where the control group had a success rate of 0.40, the treated group would have a success rate of 0.64.

In Figure 68, the studies have been separated into two subgroups. The first set of studies compared Orlistat vs. placebo, and for this subgroup the mean effect size is a difference of 0.213 [B1]. The second set of studies compared Sibutramine vs. placebo, and for this subgroup the mean effect size is a difference of 0.320 [B2]. The difference between subgroups is 0.108 with a 95% confidence interval of 0.047 to 0.168 (see Appendix IX). In round numbers, the difference in group means is at least 0.05 and possibly as much as 0.17. A test of the null hypothesis (that the mean effect size is the same in both subgroups) yields a Q-value of 12.098 with 1 degree of freedom and a corresponding p-value of < 0.001.

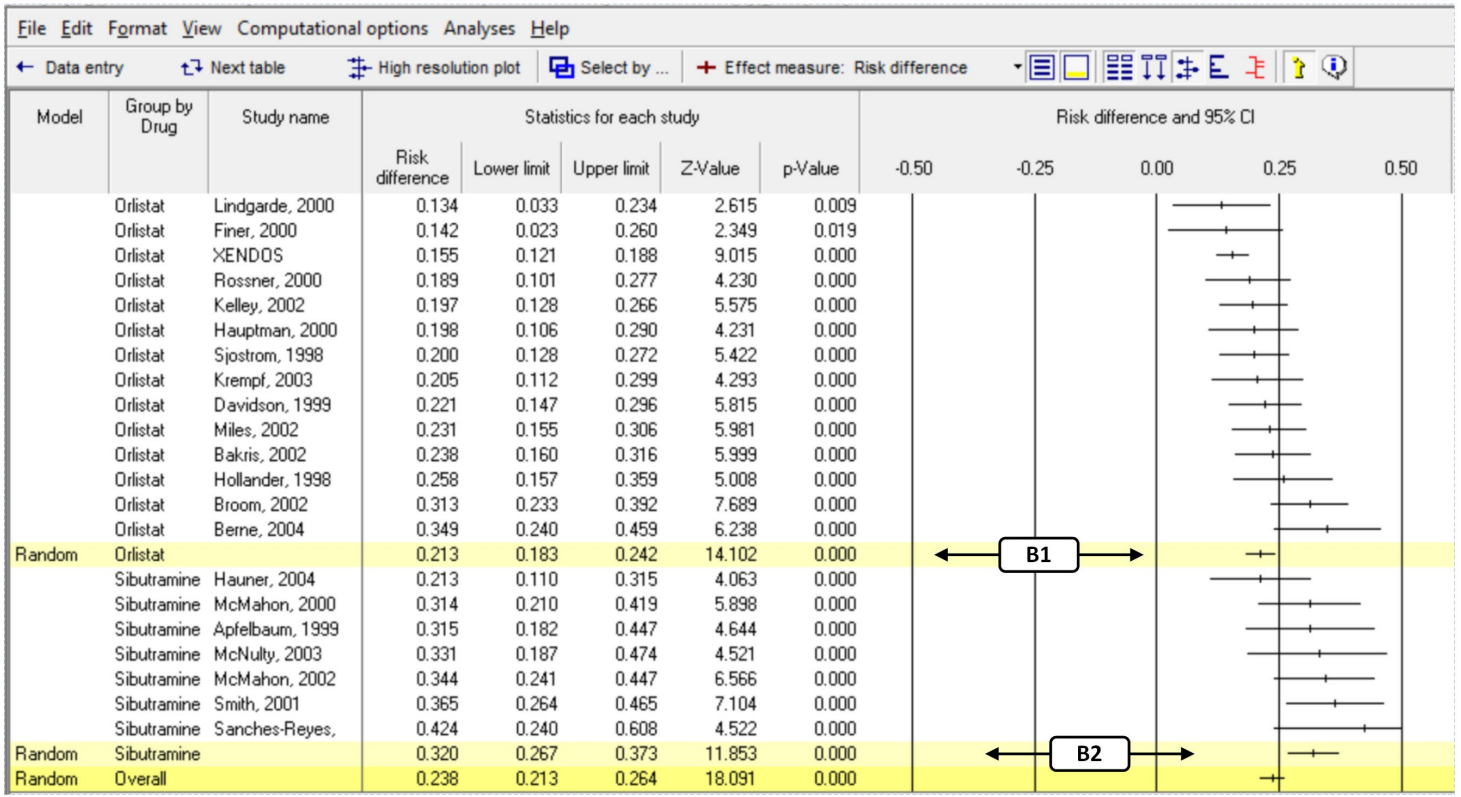


Figure 68 | Impact of drug on likelihood of success in losing weight | Risk difference > 0 favors drug

We can use this to assess the impact of Orlistat vs. placebo, to assess the impact of Sibutramine vs. placebo, and to compare the effect size in the first set of studies vs. the effect size in the second set of studies. However, the analysis must be performed correctly, and there are important limitations on what conclusions we can reach.

On the pages that follow, I address various issues including the following

- Researchers sometimes assume that a difference between subgroups is evidence of a causal link. In fact, the difference is observational, not causal.
- Researchers sometimes adopt the fixed-effect model for comparing subgroups. The correct model is the mixed-effects model.
- Researchers sometimes employ a separate estimate of T^2 for each subgroup. This is generally a mistake.

12.2. Assuming a causal relationship

12.2.1. Mistake

In the weight-loss example, the difference between subgroups is both clinically important and statistically significant. On that basis, some would conclude that Sibutramine is more effective than Orlistat. In fact, that conclusion is not supported by the data.

12.2.2. Details

In the weight-loss example, the difference between subgroups is both clinically important and statistically significant. On that basis,

A. We *can* conclude that effect of treatment was larger in the Sibutramine studies than in the Orlistat studies.
B. We *cannot* conclude that Sibutramine is more effective than placebo.

The difference between A and B may not be obvious, but is nevertheless critically important. It reflects the fact that the relationship between subgroup membership and effect size is observational, not causal.

The logic here is the same as it would be in a primary study. In a primary study, if we *randomize* patients to be treated with either Sibutramine or Orlistat and then find the difference between groups is statistically significant in favor of Sibutramine, we would conclude that Sibutramine was more effective. By contrast, if we located people who had elected to take either drug, and then found that those taking Sibutramine had a better outcome, we would recognize that the difference could be due to the drug but could also be due to a confound. For example, people who elected to use Sibutramine might tend to be younger, and the reason that the effect is larger in the Sibutramine subgroup could be the fact that younger people are more likely to lose weight than older people. We chose to label the groups as Sibutramine vs. Orlistat, but the more relevant label could be Young vs. Old.

The same idea applies to the comparison of subgroups in a meta-analysis. The test for statistical significance tells us that the Sibutramine subgroup did better than the Orlistat subgroup. The difference could be due to the fact that Sibutramine is more effective than Orlistat, but could also be due to a confound. For example, the studies that compared Sibutramine vs. placebo might have enrolled younger patients, and age could be an important factor.

We chose to label the subgroups as Sibutramine vs. Orlistat, but the more relevant label could be Young vs. Old.

Critically, this is true even though every one of the studies was a randomized controlled trial. This is because the randomization took place *within* studies, not *between* studies. We can conclude that the drug is more effective than placebo because patients were randomized to either drug or a placebo. However, the choice of Orlistat vs. Sibutramine was not based on random assignment, and therefore the better performance in the second subgroup does not prove that Sibutramine is more effective than Orlistat. The only exception would be if all the studies had been set up in advance by a consortium of researchers, and each study site was randomly assigned to use one drug or the other.

An additional example should help to illustrate this point.

12.2.3. Example | Impact of caffeine on pain

Derry, Derry, and Moore (2014) looked at the use of caffeine to relieve certain types of pain. In all studies, patients were treated with a pain medication and then (additionally) randomly assigned to receive either caffeine or a placebo (Figure 69). Researchers recorded the proportion of patients in each group who felt some improvement within thirty minutes. The mean risk ratio over all studies is 1.127, which tells us that (on average) caffeine increased the likelihood of a response by around 13%.

As noted, the caffeine (or placebo) was administered *in addition* to a pain medication. In one subgroup of studies the pain medication was Ibuprofen (name brands include Advil, Motrin, and Nuprin). In another subgroup of studies, the pain medication was Paracetamol (name brands include Tylenol and Panadol).

In the Ibuprofen subgroup, caffeine increased the likelihood of response by 29% as compared with placebo [B1]. By contrast, in the Paracetamol subgroup, caffeine increased the likelihood of response by only 11% as compared with placebo [B2]. The ratio of the two effects (1.111/1.294) is 0.859 with a 95% confidence interval of 0.737 to 1.000 (see Appendix IX). We estimate that caffeine is 14% less effective in the paracetamol subgroup as compared with the ibuprofen subgroup, but the actual difference could be as high as 26% or as low as 0% (see Appendix IX). Alternatively, we can test the null hypothesis that the impact of caffeine is identical in the two subgroups. A test of this null hypothesis yields a Q- value of 3.819, with 1 degree of freedom and a corresponding p-value of 0.051. In this discussion, I will consider this to be statistically significant.

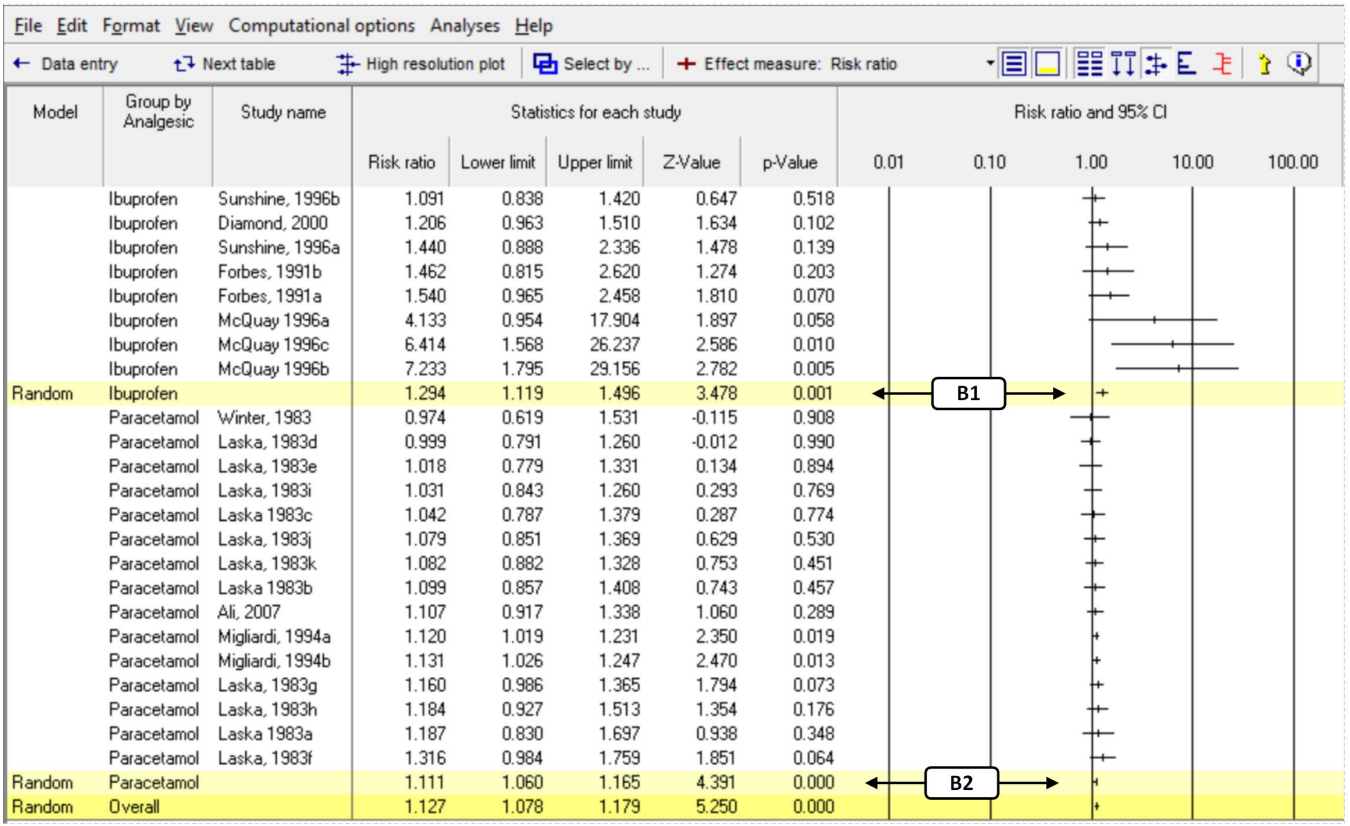


Figure 69 | Impact of caffeine on pain relief | Risk ratio > 1 favors caffeine

Based on the logic outlined above, we report that the caffeine is more effective in the first subgroup. However, we cannot conclude that this is a causal relationship based on the use of Paracetamol vs. Ibuprofen. Rather, the fact that the effect is stronger in one subgroup could be due to some other difference between the subgroups.

Indeed, the data suggest two potential confounds. In the Paracetamol subgroup, three of the eight studies were performed by McQuay, and these studies yielded the highest effect sizes. In the Ibuprofen subgroup, eleven of the fifteen studies were performed by Laska. It is possible that McQuay tended to enroll patients who were more likely to benefit from caffeine, while Laska was less likely to enroll these patients. We chose to call the subgroups Ibuprofen and Paracetamol, but the more relevant labels might be McQuay and Laska or *Likely to Benefit* and *Unlikely to Benefit*.

Confounds can exist in any analysis, but are especially worrisome when we are dealing with small numbers of studies. When there are *many* studies in each subgroup, confounds tend to be systematic – for example, it might be that studies with older patients tend to use one drug rather than another. In this case we may be able to identify potential confounds and look for them using regression or subgroup analyses. By contrast, when there are only a *few* studies in the analysis we need to be concerned about systematic confounds and also random confounds. If a subgroup labeled "Moderate Dose" includes only one or two studies, it is possible (indeed, likely) that these studies differ from studies in other subgroups in other ways, in addition to dose. Typically, we will not know to look for these confounds, and in any event will not have sufficient data to use subgroup analysis or regression to assess their potential impact.

Summary

If all studies in the analysis are randomized controlled trials, the difference between the treated and control conditions can be attributed to the treatment. However, the difference in effect size between subgroups cannot be attributed to any specific factor (such as the fact that each subgroup employed a different drug). The difference between subgroups is observational, not causal.

12.3. Choosing a statistical model

12.3.1. Mistake

When working with subgroups, researchers sometimes apply the fixed-effect model within subgroups. This is a mistake.

12.3.2. Details

In section 7, I discussed three statistical models for a simple analysis. The random-effects model applies when the studies in the analysis will be used to make an inference to the universe of comparable studies. The fixed-effect (singular) model applies when all studies are based on one population, and the results will apply only to this population. The fixed-effects (plural) model applies when the studies are based on different populations, and our goal is to report the mean effect for these studies, but not to generalize to any wider universe.

When we are working with subgroups, the choice of statistical models becomes a little more complicated. Here, we must choose which statistical model applies *within* subgroups, and also which statistical model applies *across* subgroups. The same criteria outlined above, apply here as well.

Consider the weight-loss analysis shown in Figure 70. There are fourteen studies that compared Orlistat vs. placebo and another eight that compare Sibutramine vs. placebo.

If we see the fourteen Orlistat studies as representative of a universe of comparable studies, and we want to make an inference to that universe, then we should be using the random-effects model to estimate the mean effect size for this set. Similarly, if we see the seven Sibutramine studies as representative of a universe of comparable studies, and we want to make an inference to that universe, then we should be using the random-effects model to estimate the mean effect size for this set.

If we use the random-effects model, we will get an estimate of the mean effect size for the universe of studies that compare Orlistat vs. placebo (at the top) and we will get an estimate of the mean effect size for the universe of studies that compare Sibutramine vs. placebo (at the bottom). When we compare the two means, this comparison will address the difference in the two universes.

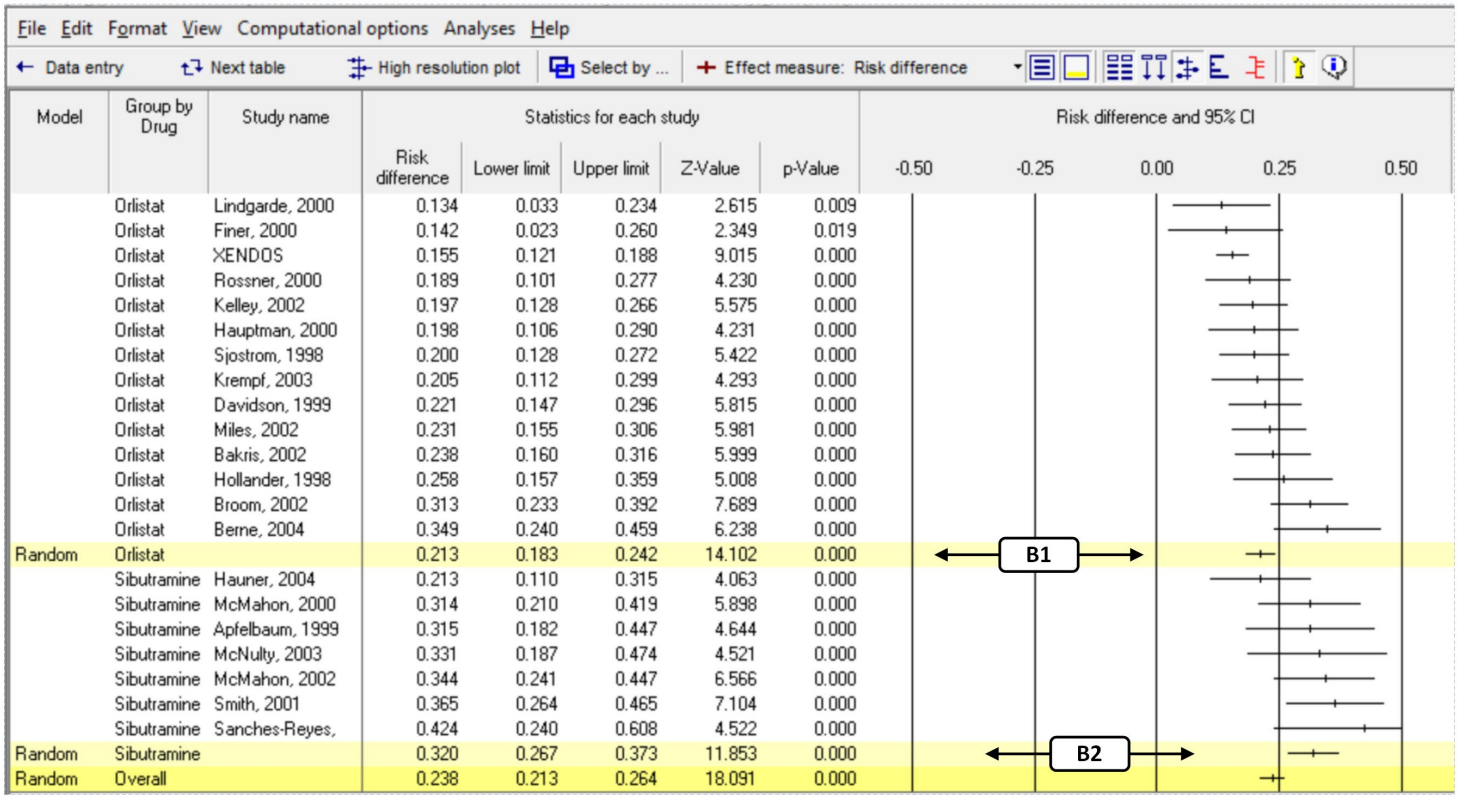


Figure 70 | Weight loss by subgroup

By contrast, if we use the fixed-effects model, we will get an estimate of the mean effect size for the specific 14 studies that compare Orlistat vs. placebo (at the top) and we will get an estimate of the mean effect size for the specific 8 studies that compare Sibutramine vs. placebo (at the bottom). When we compare the two means, this comparison will address the difference in these two specific sets of studies.

When we are pulling studies from the literature, our intent is almost invariably the first rather than the second, and therefore we should be using the random-effects model within subgroups.

12.3.3. Mixed-effects

The studies *within* a subgroup are treated as random, but the subgroups are treated as fixed. The use of the fixed model here is a source of confusion since most researchers assume that this model requires that all subgroups are identical to each other. However, as explained earlier (section 1), we need to distinguish between the fixed-effect (singular) model and the fixed-effects (plural) model. The former applies when all subgroups are identical to each other, which is obviously not the case here. The latter applies when each subgroup is unique, and we care only about the subgroups included in the analysis. This is the model that applies here. That is, we will not be generalizing from Orlistat and Sibutramine to all possible drugs. Rather, the results will apply to these two drugs specifically. Since the model is random at one level and fixed at the other, it is called a mixed-effects model.

In Figure 71, there are two sets of results. Those at the top refer to the fixed-effect model while those at the bottom refer to the mixed-effects model.

In the fixed-effect section the mean effect for the Orlistat subgroup is 0.200 with a standard error of 0.010 and the mean effect for the Sibutramine subgroup is 0.319 with a standard error of 0.022. The line labeled *Total Between* addresses the difference between these two means. The *Q*-value is 23.532 with 1 degree of freedom, and a *p*-value of < 0.001.

In the mixed-effects section the mean effect for the Orlistat subgroup is 0.213 with a standard error of 0.015 and the mean effect for the Sibutramine subgroup is 0.320 with a standard error of 0.027. The line labeled *Total Between* addresses the difference between these two means. The *Q*-value is 12.098 with 1 degree of freedom, and a *p*-value of 0.001 [A].

If our interest was limited to the twenty-one studies included in the analysis, we would use the fixed-effect section. On the other hand, if we intend to generalize from these studies to all comparable studies (which we do) we should be using the mixed-effects section. Note that the standard errors

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity		
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value
Fixed effect analysis											
Orlistat	14	0.200	0.010	0.000	0.180	0.219	20.236	0.000	27.560	13	0.010
Sibutramine	7	0.319	0.022	0.001	0.275	0.363	14.173	0.000	6.454	6	0.374
Total within									34.014	19	0.018
Total between									23.532	1	0.000
Overall	21	0.219	0.009	0.000	0.201	0.236	24.225	0.000	57.546	20	0.000
Mixed effects analysis											
Orlistat	14	0.213	0.015	0.000	0.183	0.242	14.102	0.000			
Sibutramine	7	0.320	0.027	0.001	0.267	0.373	11.853	0.000			
Total between									12.098	1	0.001
Overall	21	0.238	0.013	0.000	0.213	0.264	18.091	0.000			

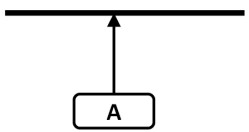


Figure 71 | Fixed-effects analysis at top | Mixed-effects analysis at bottom

for the mixed-effects model are larger than those for fixed-effects model. Therefore, under the mixed-effects model, the confidence interval for the difference in means will be wider and the difference in means is less likely to be statistically significant.

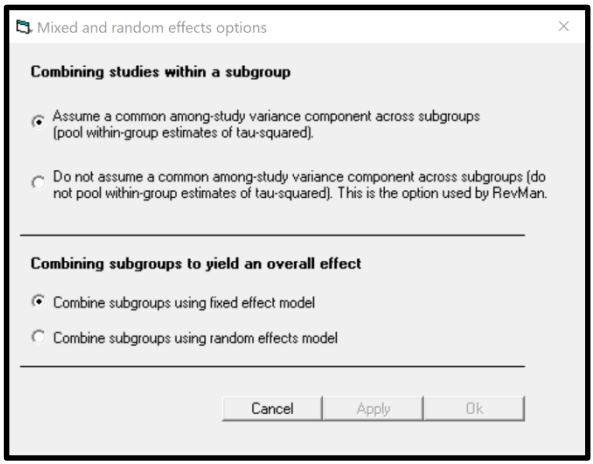


Figure 72 | Recommended options in CMA

In CMA ™ (Comprehensive Meta-Analysis) the option to use Mixed-effects is set on the Analysis screen. Choose Computational options > Mixed and random-effects options, to open a Dialog box (Figure 72). At the bottom, select [Combine subgroups using fixed-effect model]. This refers to the fact that we intend to make an inference to these two subgroups (Orlistat and Sibutramine) only, and not generalize to a universe of other possible drugs.

Summary

When we perform a subgroups analysis, we should almost invariably be using a mixed-effects model. This means that the subgroups are fixed, in the sense that these are the only subgroups of interest. By contrast, the studies *within* a subgroup are random, in the sense that these studies are a random sample of comparable studies (at least in theory), and we will be generalizing to those studies. Since the model is fixed at one level and random at the other, it is referred to as a mixed-effects model.

12.4. Mistakes in estimating τ^2

12.4.1. Mistake

When we are working with subgroups, the between-study variance (τ^2) must be computed within subgroups. Then, we have the option of using each estimate of τ^2 for the corresponding subgroup, or of pooling all the estimates and using the pooled value for all subgroups. Researchers sometimes choose the first option, but that is generally a bad idea.

12.4.2. Details

When we are working with subgroups, the between-study variance (τ^2) must be computed within subgroups. The reason is that τ^2 represents the between-study variance in the universe of interest, and this is the universe within the subgroup. Another way to say this, is that τ^2 is defined as the *unexplained* variance. The variance *between* subgroups *is* explained by subgroup membership, and therefore must be excluded from the computation of T^2. It is only the variance *within* subgroups that remains unexplained.

After we have estimated τ^2 within subgroups, we have two options. One is to use the estimate computed within each subgroup for that subgroup. The other is to pool the estimates, and use the pooled value for all subgroups.

In the weight-loss example (Figure 73), τ^2 is estimated as 0.0017 [D1] for the Orlistat studies, and as 0.0003 [D2] for the Sibutramine studies. The pooled estimate is 0.0014. If we use the separate estimates, we are saying that the between-study variance in the first set of studies is 0.0017 while the between-study variance in the second set of studies is 0.0003. By contrast, if we use the pooled estimate, we are saying that the between-study variance is 0.0014 in each set of studies. The formula for pooling estimates is given in Appendix X.

The argument for using separate estimates is that the variance between studies is unique to each subgroup. However, if we want to estimate τ^2 within subgroups we need to have a reasonable number of studies within each subgroup. When we use a unique estimate for each subgroup based on a small number of studies, each estimate will be very unreliable. The damage caused by estimating τ^2 based on a small number of studies is likely to be much larger than the damage caused by pooling estimates when the underlying values are different.

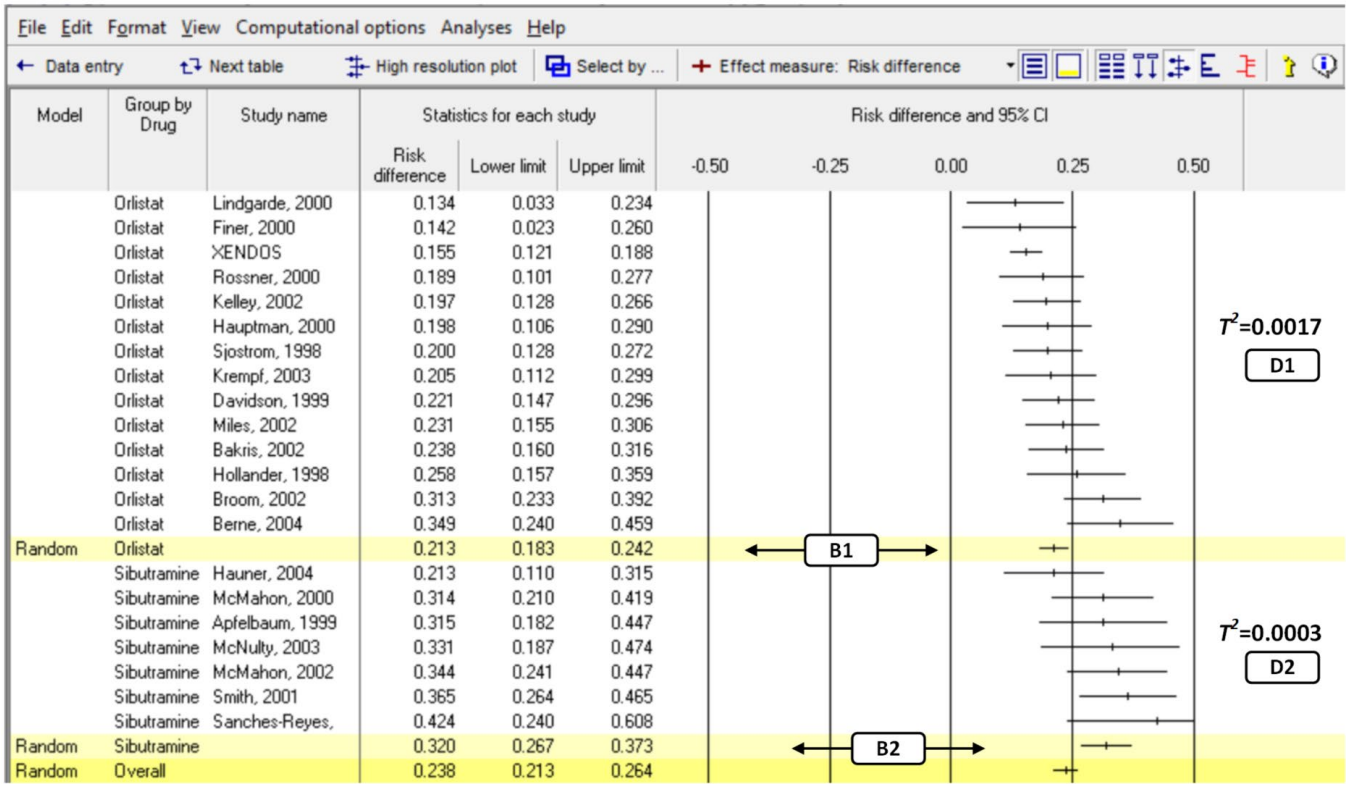


Figure 73 | Weight loss by subgroup

The extreme example is the case where a subgroup includes only one study. Without pooling, the between-study variance for these subgroups is estimated as zero, which is obviously incorrect. If one then proceeds to compare subgroups, the incorrect standard error will be used in those comparisons, and these comparisons will be incorrect.

The default position should be to pool the estimates of τ^2 across subgroups. The option to use separate estimates should only be considered if there is good reason to believe that τ^2 differs substantially from one subgroup to the next, and additionally we have a good number of studies within each subgroup.

I have deliberately been vague about how many studies we need to get a reliable estimate of τ^2, since (a) there are no standards and (b) the number will vary from one analysis to the next. I would suggest that one should never consider using separate estimates unless we have at least ten studies in each subgroup, and that twenty studies would be a better minimum. To be clear, there is no consensus on these numbers, and I am only trying to provide a sense of scale.

In CMA ™ (Comprehensive Meta-Analysis) this option is set on the Analysis screen. Choose Computational options > Mixed and random-effects options, to open a Dialog box (Figure 74). At the top, select [Assume a common among-study variance component across subgroups (pool within-group estimates of tau-squared)].

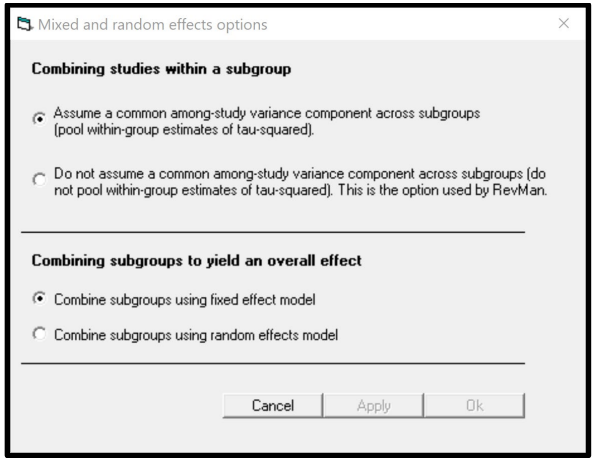


Figure 74 | Combining studies within a subgroup (Top) Option to pool estimates of T^2

Note.

In this discussion I have assumed that the analyst computes T^2 within subgroups, and explored the option of pooling (or not). Some papers have published analyses where T^2 is computed across all studies (not within subgroups). It would be hard to justify this approach.

Summary

When we work with subgroups, the estimate of T^2 should always be computed within subgroups. In the vast majority of cases, it should then be pooled across subgroups, with the pooled estimate applied to all subgroups.

12.5. Comparing the effect size in different subgroups

12.5.1. Mistake

When we are working with two (or more) subgroups, we generally want to see if the effect size is larger in some subgroups than in others. The correct way to approach this is to perform a significance test and/or estimate the magnitude of the difference with its confidence interval. Researchers sometimes use other approaches, which is a mistake.

12.5.2. Details

If the effect size is statistically significant for one subgroup but not another, researchers sometimes assume that the difference between groups must be statistically significant. This is incorrect. It is possible for the effect size to be identical in both subgroups, but statistically significant in one subgroup but not the other because one subgroup includes more studies, larger studies, and/or a more homogeneous set of effects.

Conversely, if the confidence interval for the mean in one subgroup overlaps the confidence interval for the other subgroup, researchers sometimes assume that the difference between groups cannot be statistically significant. Again, this is a mistake. It is possible for the difference between subgroups to be statistically significant even if the confidence intervals have some overlap with each other.

Therefore, in both cases, we need to perform the appropriate analysis to determine whether the difference between subgroups is statistically significant.

There are two basic approaches we can take here. One is to test the difference for statistical significance. The other is to compute the difference and its confidence interval. The advantage and disadvantage of either approach are the same here as discussed for the main effect in section 10.1.2 and section 10.1.3. The significance test should be used when our goal is to test the null hypothesis, which may be the case in a legal context. By contrast, effect size estimation provides an estimate of the difference between groups, which is generally what we care about.

Note that these analyses may have low statistical power. Therefore, if the difference between subgroups is not statistically significant (and/or the confidence interval for the difference includes zero), we *cannot* conclude that

the effect size in the different subgroups is comparable (Hedges & Pigott, 2001, 2004)

Below, I provide examples of both approaches.

12.5.3. Weight loss

In the weight-loss example (Figure 75) the difference in effect sizes is displayed in the section labeled "Mixed-effects analysis" on the line labeled "Total between". The *Q*-value for the difference is 12.098 with 1 degree of freedom and a *p*-value of 0.001 [A]. We conclude that the effect size for the Sibutramine studies (0.320) is significantly higher than the effect size for the Orlistat studies (0.213).

The line "Total between" is an omnibus test that asks if there are any differences among the subgroups. In this example there are only two subgroups, so the "omnibus" test is also a pairwise test. If there were three or more groups, we would also do pairwise comparisons to test the difference between any two groups (see Appendix IX).

In addition to testing the difference between subgroups for statistical significance, it is also important to report the magnitude of the difference along with the corresponding confidence interval.

For Orlistat vs. placebo, the mean effect size is 0.213. For Sibutramine vs. placebo the mean effect size is 0.320. When we are working with a risk difference (as we are here) the difference between subgroups is simply the difference in effect sizes. Here, that difference is 0.108 with a 95% confidence interval of 0.047 to 0.168. In round numbers, the difference in the effect size is at least 0.05 and possibly as much as 0.17. The formula for computing the confidence interval is given in Appendix IX.

12.5.4. Caffeine

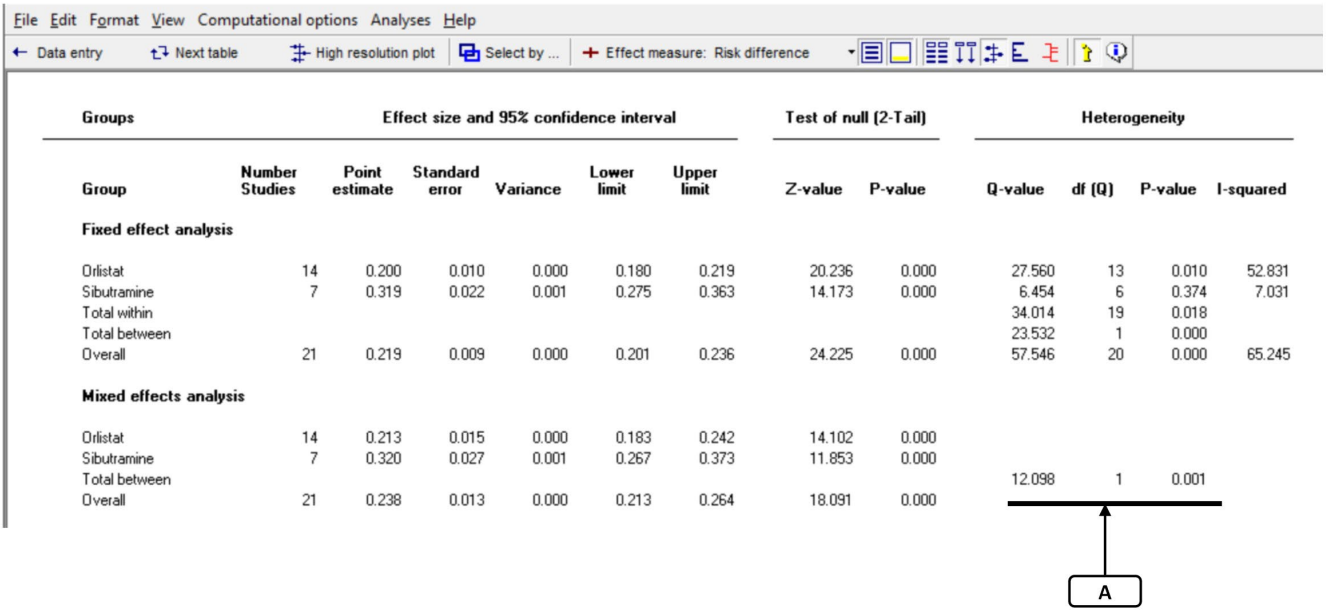
In the caffeine example (Figure 76) the difference in effect sizes is displayed in the section labeled "Mixed-effects analysis" on the line labeled "Total between". The *Q*-value for the difference is 3.819 with 1 degree of freedom and a *p*-value of 0.051 [A]. If we accept 0.051 as meeting the criterion for statistical significance, we conclude that the effect size for the Ibuprofen studies (1.294) is significantly higher than the effect size for the Paracetamol studies (1.111).

The line "Total between" is an omnibus test that asks if there are any differences among the subgroups. In this example there are only two subgroups, so the "omnibus" test is also a pairwise test. If there were three or

more groups, we would also do pairwise comparisons to test the difference between any two groups (see Appendix IX).

In addition to testing the difference between subgroups for statistical significance, it is also important to report the magnitude of the difference along with the corresponding confidence interval.

In the Ibuprofen subgroup, caffeine increased the likelihood of response by 29% as compared with placebo [B1]. In the Paracetamol subgroup, caffeine increased the likelihood of response by only 11% as compared with placebo [B2]. When we are working with a risk ratio (as we are here) the magnitude of the group difference is the ratio of the two effects. The ratio of the two effects (1.11/1.29) is 0.859 with a 95% confidence interval of 0.737 to 1.000. We estimate that caffeine is 14% less effective in the paracetamol subgroup as compared with the ibuprofen subgroup, but the actual mean ratio could be as high as 26% or as low as 0%. The formula for computing the confidence interval is given in Appendix IX.



Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis												
Orlistat	14	0.200	0.010	0.000	0.180	0.219	20.236	0.000	27.560	13	0.010	52.831
Sibutramine	7	0.319	0.022	0.001	0.275	0.363	14.173	0.000	6.454	6	0.374	7.031
Total within									34.014	19	0.018	
Total between									23.532	1	0.000	
Overall	21	0.219	0.009	0.000	0.201	0.236	24.225	0.000	57.546	20	0.000	65.245
Mixed effects analysis												
Orlistat	14	0.213	0.015	0.000	0.183	0.242	14.102	0.000				
Sibutramine	7	0.320	0.027	0.001	0.267	0.373	11.853	0.000				
Total between									12.098	1	0.001	
Overall	21	0.238	0.013	0.000	0.213	0.264	18.091	0.000				

Figure 75 | Weight-loss by drug type / Fixed-effects analysis at top | Mixed-effects analysis at bottom

File Edit Format View Computational options Analyses Help

Data entry | Next table | High resolution plot | Select by ... | Effect measure: Risk ratio

Groups		Effect size and 95% interval			Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis										
Ibuprofen	8	1.294	1.119	1.496	3.478	0.001	16.101	7	0.024	56.524
Paracetamol	15	1.111	1.060	1.165	4.391	0.000	4.537	14	0.991	0.000
Total within							20.638	21	0.481	
Total between							3.819	1	0.051	
Overall	23	1.127	1.078	1.179	5.250	0.000	24.457	22	0.324	10.046
Mixed effects analysis										
Ibuprofen	8	1.294	1.119	1.496	3.478	0.001				
Paracetamol	15	1.111	1.060	1.165	4.391	0.000				
Total between							3.819	1	0.051	
Overall	23	1.127	1.078	1.179	5.250	0.000				

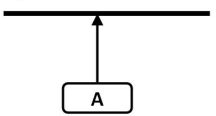


Figure 76 | Impact of caffeine by analgesic | Fixed-effects analysis at top | Mixed-effects analysis at bottom

Summary

If we want to compare the effect size in two subgroups, we need to actually run a test of statistical significance. We should also estimate the magnitude of the difference, with the corresponding confidence interval. These analyses should be based on the mixed-effects model rather than the fixed-effects model.

12.6. Reporting an overall effect size in the presence of subgroups

12.6.1. Mistake

When the analysis includes two or more subgroups, researchers typically report the mean effect size for each subgroup, and also the mean effect size across subgroups. We need to understand what the overall mean represents. Sometimes, it will not be appropriate to report this statistic.

12.6.2. Details

The overall mean is a weighted mean of the effect size in all subgroups. As such, more weight will be assigned to subgroups that have a more precise estimate of the mean effect size. In general, a subgroup with more studies will tend to get more weight in the estimate of the overall mean.

In our analysis there are more studies that used Orlistat and fewer that used Sibutramine. When we compute the overall effect size, the Orlistat studies get 76% of the weight while the Sibutramine studies get 24% of the weight. The mean effect size for Orlistat was 0.213 and the mean effect size for Sibutramine was 0.320. The overall effect size is reported as 0.238, and as such is clearly being pulled by the Orlistat studies.

What does this overall mean represent? It tells us the mean effect size in this sample of studies, but that is only relevant if this sample is representative of some universe. For example, if Orlistat was the drug of choice for 76% of all hospitals, and Sibutramine was the drug of choice for 24% of all hospitals, then the overall effect size in our analysis would reflect the overall effect size for the universe of all hospitals. However, that is probably not the case. The ratio of 76 to 24 probably does *not* represent a meaningful universe, and so the number is not terribly informative. Therefore, we may choose not to report the overall mean. If we do report it, we should be clear that the estimate is based on a weighted mean of the subgroups.

It should be noted that this is simply a special case of the issue outlined earlier for simple analyses (7.4.6). That is, if we had simply reported the mean for all studies as a simple analysis (without subgroups) that mean would *also* reflect the mix of populations in the analysis, which favors the Orlistat studies.

The difference between the simple case and this one is that if we want to predict the mean in the simple case, the best prediction we can make is the

overall mean. By contrast, when we have subgroups, we can predict the effect size for any study more accurately by using the mean for that subgroup.

Summary

When the analysis includes subgroups, the overall mean is a weighted combination of the subgroup means, and may be dominated by subgroups with more studies. Therefore, we may elect to report the means for the subgroups only, and hide the overall mean.

This is a special case of a more general issue. The mean effect will always depend on the particular mix of studies included in that analysis. If most studies come from one subgroup, that subgroup will dominate the analysis even if we perform a simple analysis without subgrouping.

12.7. Putting it all together

One of the key strengths of a meta-analysis is that it enables us to see how the effect size differs from one subgroup of studies to the next. However, it is imperative that we understand the limits of this tool and employ the correct statistical formulas.

To report that the mean effect size is higher in one subgroup of studies than another, it is not sufficient to know that the main effect met the criterion for statistical significance in one subgroup and not the other. Rather, we need to conduct a test of statistical significance for the difference in effects, and/or compute the difference in effects with confidence intervals.

If the difference between subgroups is statistically significant, we need to understand that this difference is observational, not causal. We can say that the mean effect size is higher in one subgroup, but we cannot say that the difference is due to the variable that we have used to name the subgroup, such as "Drug-A" vs. "Drug-B". While it is *possible* that the named variable is responsible for the difference, it is also possible that the difference is due to some other variable. For example, the researchers who tested Drug-A may have enrolled primarily younger patients, and it may be the impact of age, rather than drug, which is (primarily) responsible for the effect size in this subgroup. Therefore, the results of this kind of analysis should not be seen as definitive. Rather, they could be used to design additional primary studies, where the impact of drug can be tested properly.

The problem of potential confounds is present even when we have a substantial number of studies in each subgroup, since there may be a systematic relationship between the confound and our variable of interest. For example, if researchers who test Drug-A tend to enroll younger patients, this confound will exist even if we have many studies within each subgroup. When we have only a few studies within subgroups, we need to be concerned not only with systematic confounds, but also random confounds – the studies within a subgroup might differ from those in other subgroups on some important factors simply by chance.

When the studies in the analysis are being pulled from the literature, the correct statistical model is almost invariably a mixed-effects model. Concretely, studies within a subgroup will be used to make an inference to the universe of comparable studies, so we use the *random-effects* model for studies *within* subgroups. However, we care only about the specific subgroups in the analysis and will not make an inference to other subgroups, so we use the *fixed-effects* model for subgroups. Since the model is random at one level and fixed at the other, it is called a mixed-effects model.

Since we are using the random-effects model at one level and the fixed-effects model at another level, we are said to be using a *mixed-effects* model.

When we are working with multiple subgroups, we need to estimate the value of τ^2 within subgroups, but we then have the option of applying each estimate of τ^2 to the corresponding subgroup or pooling the estimates and applying the pooled estimate to all subgroups. One should always pool the estimates unless each subgroup has a substantial number of studies. In the extreme case, when some subgroups have only one or two studies, this approach is imperative.